FRAUDBENCH: A MULTIMODAL BENCHMARK FOR DETECTING AI-GENERATED FRAUDULENT REFUND EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Artificial Intelligence (AI)-generated images have become increasingly realistic and readily adaptable to real-world claims, creating new challenges for verifying visual evidence. A concrete emerging risk is AI-generated refund fraud, in which manipulated or synthetic images are used to support claims about damaged products, poor delivery conditions, or service-related defects. Existing AI-generated image detection benchmarks mainly evaluate standalone authenticity classification, cross-generator transfer, or forensic localization, leaving claim-conditioned fraudulent evidence detection underexplored. To bridge this gap, we introduce FRAUDBENCH[1], a multimodal benchmark for detecting AI-generated fraudulent refund evidence. FRAUDBENCH is constructed from real-world user-review evidence across e-commerce, food delivery, and travel-service scenarios. We curate real evidence images together with their associated review and product metadata, identify genuine damaged and undamaged evidence through MLLM-assisted filtering and human annotation, and synthesize fake-damaged evidence from genuine undamaged reference images using 12 state-of-the-art image editing and generation models. Using FRAUDBENCH, we evaluate MLLMs, specialized AI-generated image detectors, and human evaluators on the same evaluation images. Experiments show that current MLLMs often recognize real-damaged evidence but fail on many fake-damaged subsets, with fake-damage detection rates (TPR) far below the 50% baseline on most generator subsets. Specialized detectors generally perform better but remain inconsistent across generators and can produce false positives on real-damaged samples, revealing a clear gap between generic AI image detection and reliable claim-conditioned refund-evidence verification.

Note: This paper is intended solely for academic research purposes. The goal is not to facilitate refund fraud or provide actionable guidance for misuse.

1 INTRODUCTION

The rapid evolution of generative artificial intelligence (AI) has reached a critical milestone: the visual quality of synthesized and edited images has crossed the threshold of human perception, making them nearly indistinguishable from authentic photographs (Rombach et al., 2022; Saharia et al., 2022). While approaching or surpassing human-level realism represents a triumph in deep learning, it simultaneously introduces unprecedented vulnerabilities into trust-based digital ecosystems. A concrete and rapidly emerging threat is AI-generated refund fraud. On transaction-oriented platforms, from e-commerce to food delivery and travel services (e.g., Taobao, Lazada, and Amazon), malicious actors now leverage state-of-the-art image generation and editing models to fabricate highly realistic visual evidence. By seamlessly injecting claim-specific defects (e.g., shattered screens, spoiled food, or stained fabrics) into genuine product images, these manipulated visuals easily pass manual review and exploit dispute-resolution systems for illicit financial gain (Li, 2025a;b). This transforms the generic "deepfake" problem into an attack on platform integrity, where AI is weaponized to generate context-aware, fraudulent evidence.

Despite the growing urgency of this threat, existing detection benchmarks are not well-suited to evaluate it. Many established datasets were constructed before the recent leaps in generative capa-

[1]An anonymized version of FRAUDBENCH is available at `https://dataverse.harvard.edu/previewurl.xhtml?token=0bf91c2e-daeb-408f-b08d-0826200fcc4d`.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

bilities, focusing largely on low-level artifacts from earlier models rather than the highly realistic edits prevalent today (Rossler et al., 2019; Zhu et al., 2023). Beyond image quality, there is also a structural mismatch between standard benchmark settings and how real-world evidence is presented. Most benchmarks treat images as isolated, single-view samples for generic authenticity classification (Yan et al., 2024; Wang et al., 2025). In practice, however, refund evidence is rarely a standalone image. Users typically submit multiple photographs from different angles to demonstrate a specific defect, often accompanied by a textual claim or review. As a result, detecting fraudulent evidence goes beyond spotting pixel-level anomalies in a single picture; it requires verifying multi-view visual consistency and ensuring the images align with the user's stated claim. To properly evaluate this capability, the community needs to move beyond generic synthetic datasets and focus on realistic, transaction-oriented scenarios.

To bridge this gap, we introduce **FRAUDBENCH**, a multimodal benchmark specifically designed for detecting AI-generated fraudulent refund evidence. Rather than collecting isolated images, we construct **FRAUDBENCH** from 822 real-world user reviews across 29 diverse categories in e-commerce, food delivery, and travel services, yielding a total of 12,868 evaluation images. This comprehensive foundation naturally provides realistic, multi-view evidence paired with textual claims. To ensure the benchmark evaluates genuine verification capabilities rather than relying on "shortcut learning" (e.g., detecting generic AI styles or resolution drops), we employ a rigorous, controlled synthesis pipeline. We curate real, undamaged multi-view images as references and utilize 12 state-of-the-art image generation and editing models to synthesize *fake-damaged* evidence. By applying category-specific damage patterns to authentic user photos, we preserve the original object, scene context, and lighting. This forces recognition models to look beyond superficial artifacts and perform rigorous cross-modal and multi-view consistency checks.

Using **FRAUDBENCH**, we conduct a comprehensive evaluation of 30 Multimodal Large Language Models (MLLMs), four specialized detectors with eight checkpoints, and human evaluators. Our findings reveal a critical dilemma in current detection paradigms when faced with the asymmetric risks of real-world dispute resolution. In practice, false positives (wrongly accusing an honest customer of fraud) can trigger severe trust crises, while false negatives (failing to detect AI evidence) lead to direct financial losses. We observe that current models struggle to safely navigate this trade-off. MLLMs exhibit a tendency toward ***over-credulity***; while they rarely misjudge genuine images, their detection rates for AI-manipulated damage are low, often falling below random guessing. Conversely, specialized detectors tend toward ***over-sensitivity***; although they can identify synthetic manipulations, several checkpoints suffer from unacceptable false positive rates, frequently flagging genuine, unedited images of damaged goods as fake. Notably, while human evaluators outperform all evaluated MLLMs, they still exhibit non-negligible error rates, confirming that verifying these highly realistic manipulations remains a formidable challenge even for manual inspection. This collective failure underscores the unique difficulty of **FRAUDBENCH** and highlights a profound gap in current multimodal verification capabilities. Our contributions are summarized as follows:

- We introduce **FRAUDBENCH**, to the best of our knowledge, the first multimodal benchmark dedicated to detecting AI-generated fraudulent refund evidence. Featuring 822 review samples and 12,868 evaluation images across 29 categories, it shifts the evaluation paradigm from standalone authenticity classification to claim-conditioned, multi-view evidence verification.
- We design a rigorous construction pipeline that synthesizes fake-damaged evidence from real-world references using 12 state-of-the-art generative models, ensuring the benchmark reflects realistic adversarial scenarios.
- We benchmark a diverse set of 30 MLLMs, four specialized detectors with eight checkpoints, and human evaluators. Our results expose a critical "over-credulity vs. over-sensitivity" dilemma, highlighting the urgent need for models capable of managing asymmetric risks in high-stakes applications.

2 RELATED WORK

Benchmarks for AI-Generated Image Detection. Benchmarks for AI-generated image detection are essential for measuring whether detectors generalize across generators, visual domains, and deployment conditions. Recent efforts have moved beyond early settings that evaluate detectors on a narrow set of face manipulations or a small number of generators. Specifically, DF40 expands face

deepfake evaluation to diverse manipulation techniques and evaluation settings (Yan et al., 2024). Community Forensics scales the generator space, highlighting generator diversity as a major bottleneck for cross-model generalization (Park & Owens, 2025), while RRBench studies robustness under transmission and re-digitization (Li et al., 2025). DFBench evaluates the deepfake detection capability of large multimodal models (Wang et al., 2025), and ForensicHub provides a unified benchmark and codebase across multiple image forensic domains (Du et al., 2025). Despite these advances, existing benchmarks mainly focus on generic authenticity classification, cross-generator transfer, or broad forensic localization. They do not directly evaluate refund-related fraudulent evidence, where an image must be assessed in relation to the specific claim it is intended to support.

AI-Generated Image Detectors. Detecting AI-generated images has become increasingly important as modern generative models can produce realistic images tailored for specific instructions and application scenarios. Existing approaches broadly fall into two categories. The first line of work develops specialized detectors for generalizable real/fake classification, from early artifact-based detection (Wang et al., 2020) to universal detectors for cross-generator generalization (Ojha et al., 2023). Recent methods improve generalization by combining foundation-model representations with local or frequency-level artifacts (Yan et al., 2025a), or by integrating semantic and pixel-level features (Cheng et al., 2025). Structured and adaptive feature-learning methods further improve robustness across unseen generators and post-processing conditions (Yan et al., 2025b; Li et al., 2026), yet these approaches remain optimized as task-specific binary classifiers without broader claim-level reasoning. A more recent direction explores multimodal large-model-based detection. Rather than only predicting real or fake labels, recent studies use MLLMs to provide artifact explanations, visual grounding, segmentation, or human-aligned reasoning (Wen et al., 2025; Zhou et al., 2025; Kang et al., 2025; Ji et al., 2026). This direction makes AI-generated detection more interpretable and closer to user-facing authenticity assessment. However, both specialized detectors and MLLM-based detectors are primarily evaluated on generic synthetic images or broad forensic benchmarks, leaving their effectiveness for claim-conditioned fraudulent evidence detection unclear.

3 Our Benchmark: FraudBench

3.1 Overview of FraudBench

FraudBench is a multimodal benchmark for evaluating AI-generated fraudulent refund evidence detection under realistic transaction settings. It is constructed from a task-relevant subset of Amazon Reviews 2023 (Hou et al., 2024), self-collected review evidence from Trip.com and GrabFood, and a small set of in-house captured real-world samples. In total, **FraudBench** contains **822** real review samples and **13,856** images. The 2,000 real images consist of 988 real-undamaged images and 1,012 real-damaged images. Using the 988 real-undamaged images as generation references, we further synthesize 11,856 fake-damaged images. Real-undamaged images serve solely as synthesis sources and are excluded from evaluation. The benchmark covers **four** data sources, **six** real-world transaction scenarios, and **29** product/service categories. The collected data span e-commerce, food delivery, trip service, offline shopping, dine-in & pickup, and express delivery, supplemented by self-collected evidence. The detailed construction process is shown in Fig. 1.

Unlike generic AI-generated image detection benchmarks, **FraudBench** focuses on refund-evidence verification, where detectors must judge whether visual evidence is authentic and whether it is consistent with the associated review context. Each real review sample is organized as a multimodal evidence instance containing one or more images, review text, or a transaction-specific description, category metadata, and authenticity labels. For synthetic evidence, fake-damaged images are generated from retained real-undamaged images using **12** state-of-the-art image editing and generation models, together with image-specific prompts and corresponding fake review comments.

As illustrated in Fig. 2 and detailed in Sec. 3.3, the benchmark is designed around five evaluation dimensions that reflect practical sources of uncertainty in refund-evidence verification. The input-modality dimension examines whether detectors remain reliable for either single images or multiple related images from the same review, while the contextual-information dimension tests whether review text helps distinguish genuine damage from synthetic or inconsistent evidence beyond visual cues alone. Meanwhile, the multi-step reasoning evaluates whether image-level decomposition improves claim-level decisions, and the prompt-sensitivity dimension measures the stability of detector outputs under semantically equivalent verification prompts. Finally, real-image preservation assesses whether models can detect synthetic refund evidence without over-flagging genuine user submissions, a prerequisite for trustworthy refund adjudication.

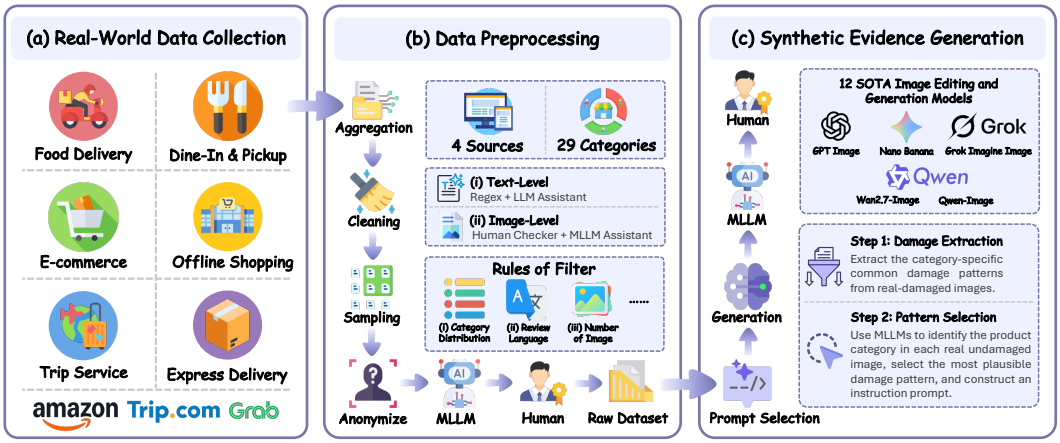

Figure 1: **Construction pipeline of our proposed FRAUDBENCH.** The benchmark is constructed in three stages: **(1) Data Collection**: We collect real-world refund-related review data from four data sources and consolidate them into 29 categories. **(2) Data Preprocessing**: We perform multimodal aggregation, two-level cleaning, rule-based filtering, representative sampling, and anonymization, with screening and human verification for quality control. **(3) Synthetic Evidence Generation**: We construct image-specific prompts via a two-step procedure. SOTA models are then used to generate fraudulent refund evidence, which is further verified by human annotators.

3.2 BENCHMARK CONSTRUCTION

Data Collection. To evaluate AI-generated refund-evidence detection under realistic transaction settings, we construct the benchmark from four complementary data sources. The primary e-commerce source is Amazon Reviews 2023 (Hou et al., 2024), a large-scale review dataset with user reviews, item metadata, and image-related resources, from which we curate **590** task-relevant review samples across **27** product categories. Because Amazon Reviews 2023 was not designed for refund-evidence verification, repurposing it for this setting requires non-trivial review–image–metadata alignment, task-specific filtering, and new claim-conditioned damage annotations. Moreover, service-oriented refund scenarios remain outside its e-commerce scope.

We therefore add **225** non-Amazon review evidence from Trip.com and GrabFood, covering service-oriented scenarios such as travel services, food delivery, and dine-in & pickup. Finally, we include a small set of in-house captured real evidence images collected under everyday conditions to improve coverage of practical visual acquisition settings. These sources provide real user-submitted or naturally captured images paired with review text or transaction-specific descriptions, allowing the benchmark to span diverse product categories, service contexts, image conditions, and damage or complaint patterns.

Data Preprocessing. After aggregating data from the four sources, we apply a multimodal preprocessing pipeline covering text cleaning, image cleaning and anonymization, image–review alignment filtering, distribution-aware sampling, and quality control. We use an MLLM as a first-stage screening model for both textual and visual content. For text cleaning, we remove empty, duplicated, corrupted, or structurally invalid reviews, and assess whether each review is relevant to the associated product, service, or complaint. For image cleaning and anonymization, the screening model flags invalid or risky samples, including privacy-sensitive images containing clear human faces or personal identifiers, technically invalid images such as corrupted files, duplicated images, screenshots, advertisements, or unreadable images, and task-irrelevant images that are unrelated to the target item or service. We further discard samples with weak image–review alignment, since review-conditioned detection requires consistency between textual context and visual evidence.

We then conduct distribution-aware sampling to preserve category diversity and evaluation usability. At the category level, we retain a post-filtering distribution that remains as close as possible to the distribution before sampling, avoiding over-representation of a small number of frequent categories while preserving the natural long-tail structure of the collected data. Within each category, we further control the image-count distribution across single-image and multi-image reviews, so that the benchmark supports both single-image evaluation and different levels of multi-image context. We

also retain multilingual review text when available, rather than translating all reviews into English, to reflect real-world platform usage and evaluate robustness under multilingual user contexts. Finally, samples flagged during MLLM-based screening are verified by human annotators, who make the final inclusion decision before the samples are retained in the benchmark.

Synthetic Evidence Generation. We generate synthetic fraudulent refund evidence from retained real-undamaged images through an image-specific prompt construction pipeline. We use 12 state-of-the-art image editing and generation models that are accessible through mainstream AI platforms or official API services, reflecting a realistic threat model for low-barrier synthetic evidence creation. The pipeline has two stages. First, we mine real damaged samples within each category to construct category-specific candidate damage lists, covering visually verifiable defects such as scratches, dents, stains, leakage, breakage, and related damage patterns. Second, for each target image, an MLLM analyzes the visible product, packaging, material, and physical attributes, and combines this visual analysis with product metadata, including the title, category, description, and features. The corresponding category-specific damage list is then retrieved, and the most plausible damage pattern is selected according to the object, material, scene, and metadata. When no category-specific candidate is suitable, the selection falls back to a universal damage vocabulary. The selected damage pattern is finally converted into an image-specific editing prompt.

Each prompt is shared across the 12 image editing and generation models, ensuring that the generated outputs are compared under the same editing intent. The prompt specifies that the modification should be visible, photorealistic, physically plausible, and localized to the target object, while preserving the original product identity, viewpoint, lighting, background, surrounding objects, and overall composition. We also generate a fake review comment for each synthetic sample for review-conditioned evaluation. Quality control verifies that generated comments are stylistically consistent with genuine user reviews and do not contain artifact-level linguistic cues that would permit reliable text-only classification, ensuring that claim verification requires joint reasoning over the visual evidence and the stated textual claim. This setup allows the benchmark to evaluate both visual authenticity detection and image–text consistency in refund-claim contexts. Generated samples are subsequently checked using the same quality-control procedure and regenerated when implausible damage, category mismatch, prompt misalignment, or trivial generation artifacts were found.

Quality Control. We apply quality control after both raw-data preprocessing and synthetic image generation. Details of the screening procedure are provided in Appx. C.5.

3.3 Evaluation Protocol

Detecting AI-generated review images in refund scenarios requires more than a single fixed evaluation setting. A reliable detector should operate across both single-image and multi-image inputs, use review text when it is available, reason over image sequences in a structured manner, remain stable under prompt phrasing changes, and, crucially, avoid misclassifying genuine user-submitted evidence as forged. We therefore evaluate each detector along five complementary dimensions: **(1) Input Modality**, **(2) Contextual Information**, **(3) Multi-Step Reasoning**, **(4) Prompt Sensitivity**, and **(5) Real Image Preservation**.

This design reflects the deployment conditions of refund-evidence verification. User reviews may contain one or multiple images, may or may not include informative text, and may be processed by downstream systems using different prompt templates or interaction settings. At the same time, the error costs are asymmetric: aggressively flagging all images can improve fake-image recall, but it also harms honest users by falsely rejecting genuine evidence. Evaluating only one setting, or reporting results only on synthetic images, would therefore obscure important failure modes. Fig. 2 provides qualitative examples for the main settings, while the corresponding quantitative results are reported in Sec. 4.2. We describe each evaluation dimension below.

Dimension 1: Input Modality. Real reviews from e-commerce and service platforms often include multiple images of the same product, order, or service experience. We evaluate two input settings. In the **Single-Image** setting, each image is judged independently. In the **Multi-Image** setting, all images from the same review are provided jointly in one inference call. The multi-image setting allows the detector to use cross-image cues such as consistent lighting, viewpoint, background, object geometry, and damage appearance. To avoid a degenerate comparison, this setting is applied only to reviews containing at least two images.

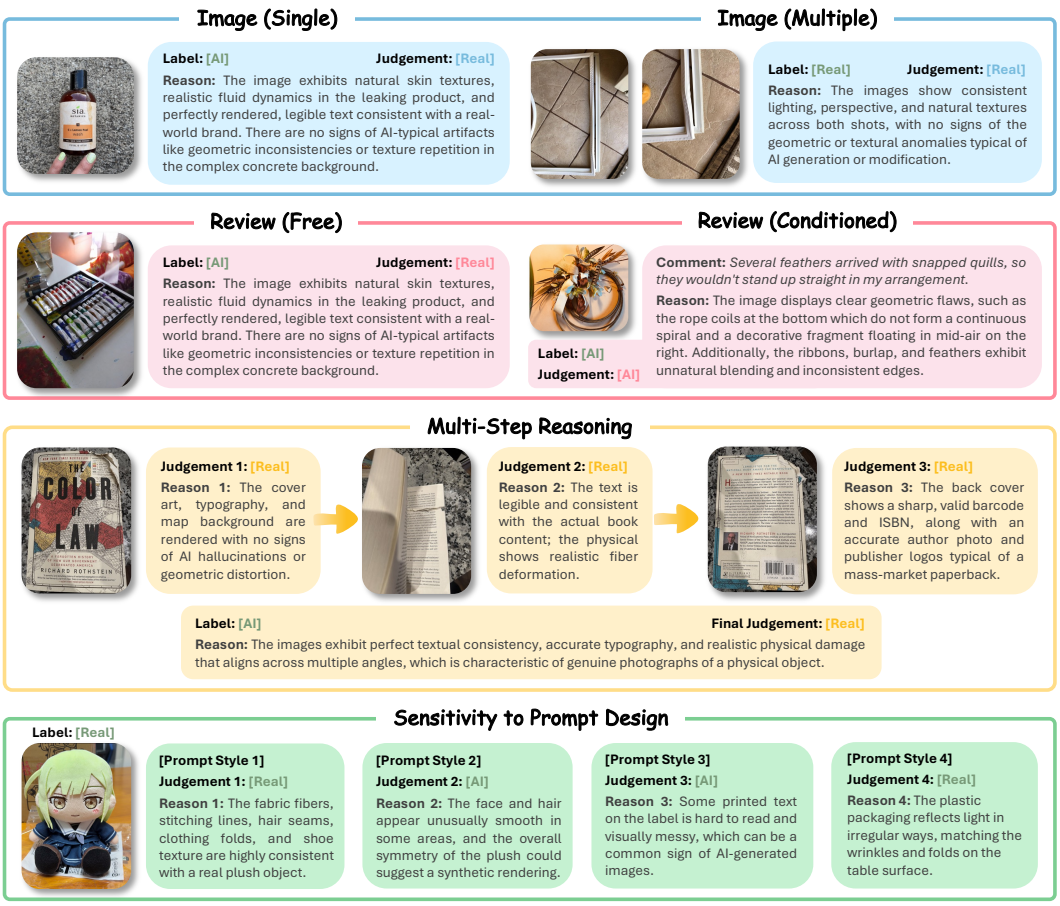

Figure 2: **Representative examples in FRAUDBENCH.** We illustrate examples covering the main evaluation settings: input modality, contextual information, multi-step reasoning, and prompt sensitivity. Text prompts, rationales, and intermediate judgments are abbreviated for visualization.

Dimension 2: Contextual Information. Review text can provide evidence that changes the interpretation of an image. For example, a complaint about a broken or leaking item supplies a ready explanation for visual irregularities, which may otherwise have been flagged as generation artifacts. We compare two settings: **Review-Free**, where the detector receives only the image input, and **Review-Conditioned**, where the associated user review text is also provided. This dimension measures whether textual context improves detection quality or instead introduces misleading shortcuts.

Dimension 3: Multi-Step Reasoning. When a review contains multiple images, a detector may benefit from examining each image individually before reaching a review-level decision. We therefore introduce a **Multi-Step** setting, in which the images are presented one per turn in a multi-turn conversation: the model gives a short observation for each image and then issues a final verdict for the entire review. We compare this setting against the Multi-Image setting, where the model jointly infers a single verdict from all images in one pass. This comparison tests whether structured per-image decomposition improves reliability, particularly when different images provide complementary or conflicting evidence.

Dimension 4: Prompt Sensitivity. In practical systems, prompt templates are often revised for clarity, length, or integration with downstream pipelines. A robust detector should not change its decision solely because the same task is phrased differently. We measure prompt sensitivity by evaluating six MLLMs with five prompt variants that progressively remove task-specific guidance, on a stratified 10% sample of the 27 Amazon categories, while holding the image input and ground-truth label fixed. Large variation across prompt styles indicates that the detector is responding to prompt wording rather than stable visual evidence.

Dimension 5: Real Image Preservation. Real image preservation evaluates whether a detector correctly retains genuine user uploads. This dimension is central to refund-evidence verification because false accusations of forgery can directly harm user experience and platform trust. Unlike the first four dimensions, which define setting variations, this dimension is evaluated across all settings using a shared real-image subset. It separates detectors that achieve high fake-image recall by over-flagging from detectors that detect synthetic evidence while preserving authentic content.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Metrics. We formulate the task as binary verification, where fake-damaged images are treated as the positive class, and genuine, real-damaged images are treated as the negative class. For each generator subset, we report Recall, i.e., True Positive Rate (TPR), to measure the model's ability to detect synthetic damage, using the shared real-damaged subset as the negative reference set. On the real-damaged subset, we report the True Negative Rate (TNR) to assess whether authentic damaged evidence is correctly retained rather than falsely rejected as synthetic. We further report the mean confidence on correct predictions for both generated and real-damaged images. For overall comparison, we report Balanced Accuracy and the F1 score of the fake-damaged (positive) class. All overall metrics are macro-averaged over product categories.

Models. We evaluate a broad set of state-of-the-art MLLMs from major providers, including proprietary models accessed through official APIs and open-weight models, and spanning general-purpose, vision-language, and omni-modal variants. We further compare them with four specialized AI-generated image detectors: CO-SPY (Cheng et al., 2025), ForgeLens (Chen et al., 2025), Effort (Yan et al., 2025b), and IAPL (Li et al., 2026), each evaluated with checkpoints trained on different datasets. The full list of evaluated MLLMs is provided in Appx. D.1.1.

Implementation Settings. For all MLLM settings, the model returns a binary prediction, a confidence score in $[0, 1]$, and a free-text rationale. In multi-image settings, the model additionally returns a review-level verdict and per-image predictions. For proprietary models, we obtain outputs directly through their respective official API services using default inference settings unless otherwise specified. Specialized AI-generated image detectors are evaluated on a cloud server with one NVIDIA A100 80GB GPU, using officially released implementations, released checkpoints, and default inference settings. Additional details, including prompts and hyperparameter settings, are provided in Appx. D.1.2. To interpret model performance, we conduct a small-scale human evaluation and report accuracy as a reference baseline, with details provided in Appx. D.1.3.

4.2 OVERALL RESULTS

Tables 1 and 2 present the primary evaluation results. We summarize the main findings below.

MLLM Failures Are Driven by Synthetic-Damage Under-Detection. In the single-image no-review setting, current MLLMs preserve genuine evidence more reliably than they detect synthetic damage. Across MLLMs, the average real-image TNR is 0.938, whereas the average TPR over fake-damaged subsets is only 0.230. This indicates that the dominant failure mode is not excessive rejection of genuine evidence, but systematic under-detection of synthetic refund evidence. GPT-6 Luna is the only exception, flagging most genuine evidence as fake with a TNR of 0.206. The category-level breakdown supports this interpretation: category-wise balanced accuracy is much more strongly associated with fake-damage recall than with real-image retention.

Failures Are Category-Dependent. The failure pattern is not uniform across categories. MLLMs are weakest on everyday-product and service-oriented categories such as Baby Products, Tools & Home Improvement, and Food Delivery, which form the lower end of the category-wise breakdown. These categories often involve heterogeneous objects, unconstrained capture conditions, cluttered backgrounds, and non-standard evidence styles, making them less aligned with clean forensic authenticity cues. By contrast, categories with more regular object structure or visually salient physical defects tend to provide stronger cues for detection. This category-level variation shows that **FRAUDBENCH** not only measures generator artifacts but also probes whether models can verify plausible damage under realistic product and service contexts. We provide detailed analysis in Appx. D.4.1.

Table 1: **Performance comparison of representative MLLMs and specialized AI-generated image detectors on FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context, which evaluates the input-modality dimension. Full results for all 30 evaluated MLLMs are provided in Tab. 12.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.084	0.693	0.171	0.701	0.213	0.690	0.255	0.700	0.357	0.710	0.423	0.721	0.601	0.391	0.952	0.764
Gemini 3.1 Pro	0.091	0.714	0.122	0.712	0.167	0.712	0.260	0.719	0.327	0.715	0.519	0.717	0.619	0.391	0.990	0.704
Gemini 3.5 Flash-Lite	0.050	0.965	0.066	0.958	0.085	0.964	0.119	0.968	0.178	0.965	0.244	0.968	0.560	0.212	0.996	0.950
Gemini 3.8 Flash	0.025	0.703	0.013	0.708	0.066	0.703	0.102	0.708	0.157	0.707	0.199	0.710	0.547	0.165	**1.000**	0.658
GPT-6 Astra	0.108	0.783	0.107	0.795	0.212	0.815	0.400	0.837	0.453	0.830	0.530	0.840	0.650	0.450	0.998	0.930
GPT-6 Sol	0.089	0.771	0.126	0.769	0.202	0.794	0.312	0.838	0.322	0.815	0.465	0.838	0.623	0.390	0.993	0.913
GPT-6 Luna	**0.761**	0.896	**0.910**	0.919	**0.934**	0.924	**0.928**	0.927	**0.845**	0.902	**0.951**	0.934	0.547	**0.873**	0.206	0.763
Grok 4.6	0.050	0.829	0.047	0.806	0.131	0.823	0.173	0.835	0.198	0.818	0.250	0.814	0.568	0.237	0.995	0.843
Kimi K3	0.015	0.848	0.043	0.718	0.029	0.711	0.127	0.808	0.152	0.742	0.276	0.773	0.541	0.189	0.975	0.866
Qwen3-VL-Flash	0.043	0.936	0.090	0.938	0.172	0.938	0.247	0.939	0.278	0.936	0.364	0.938	0.592	0.321	0.986	0.973
Qwen3-VL-Plus	0.002	0.920	0.006	0.932	0.024	0.942	0.050	0.945	0.049	0.952	0.108	0.947	0.518	0.075	0.997	0.977
Qwen3.5-Omni-Flash	0.013	0.911	0.050	0.910	0.083	0.910	0.147	0.910	0.152	0.911	0.185	0.911	0.551	0.183	0.996	0.952
Qwen3.5-Omni-Plus	0.015	0.928	0.026	0.928	0.055	0.928	0.109	0.931	0.121	0.933	0.180	0.936	0.542	0.151	**1.000**	0.971
Qwen3.8-27B	0.104	0.821	0.090	0.820	0.163	0.828	0.218	0.834	0.207	0.815	0.320	0.839	0.583	0.305	0.982	0.870
Qwen3.8-Flash	0.095	0.896	0.253	0.905	0.295	0.904	0.412	0.891	0.390	0.903	0.561	0.904	0.641	0.490	0.947	0.940
Qwen3.8-Max	0.112	0.789	0.180	0.806	0.234	0.811	0.414	0.828	0.452	0.829	0.619	0.850	**0.664**	0.490	0.993	0.864
Specialized Detectors																
CO-SPY [ProGAN]	**0.977**	0.991	0.099	0.903	0.312	0.903	0.218	0.869	0.894	0.972	**0.996**	0.997	0.782	0.733	**0.982**	0.998
CO-SPY [SD-v1.4]	0.360	0.767	0.310	0.743	0.282	0.750	0.360	0.759	0.339	0.750	0.349	0.769	0.509	0.455	0.685	0.844
ForgeLens [ProGAN]	0.700	0.955	0.217	0.869	0.516	0.905	0.823	0.949	0.859	0.966	0.925	0.973	0.815	0.798	0.958	0.968
ForgeLens [GenImage]	0.967	0.888	0.158	0.749	**1.000**	0.940	0.990	0.895	**1.000**	0.925	0.989	0.910	**0.904**	0.915	0.957	0.875
Effort [SD-v1.4]	0.710	0.816	0.546	0.763	0.658	0.804	0.755	0.827	0.858	0.874	0.890	0.885	0.795	0.828	0.853	0.823
Effort [Chameleon]	0.828	0.715	0.824	0.716	0.868	0.730	0.882	0.756	0.849	0.734	0.872	0.736	0.452	0.838	0.051	0.575
IAPL [ProGAN]	0.255	0.874	0.295	0.886	0.140	0.866	0.122	0.869	0.398	0.920	0.568	0.927	0.628	0.438	0.960	0.984
IAPL [SD-v1.4]	0.973	0.883	**0.895**	0.855	0.916	0.872	**0.996**	0.932	0.955	0.882	0.994	0.901	0.727	**0.932**	0.499	0.804
Baselines																
Random Guessing	0.500	—	0.500	—	0.500	—	0.500	—	0.500	—	0.500	—	0.500	—	0.500	—
Human Reference	0.420	—	0.509	—	0.539	—	0.598	—	0.697	—	0.709	—	0.686	0.704	0.793	—

Damage Type Further Shapes Detection Difficulty. The damage-type breakdown in Tab. 50 shows that MLLM failures also depend on the visual form of the claimed defect. More structural or visually salient defects are generally easier to detect: averaged over MLLMs, cracked, leaking, shattered/broken, and melted samples obtain TPRs of 0.331, 0.288, 0.282, and 0.266. By contrast, packaging-damaged, torn/ripped, mouldy/spoiled, and dented evidence remains substantially harder, with average TPRs of 0.144, 0.152, 0.163, and 0.192. Packaging-damaged evidence is especially difficult, with several MLLMs close to zero TPR. This suggests that current MLLMs are more sensitive to obvious object-level breakage or fluid-like damage than to subtle deformation, packaging-level defects, or surface-level complaints. Further analysis is provided in Appx. D.4.2.

Generator Effects Remain Strong Across Evaluation Settings. Detection difficulty varies substantially across generators. In the single-image no-review setting, the average MLLM TPR is only 0.104 on GPT Image 2 and 0.144 on Grok Imagine Image, but rises to 0.387 on Qwen-Image-Edit-Max. This ordering remains visible under the review-conditioned single-image, multi-image, and multi-step settings, indicating that added context does not eliminate generator-dependent failure modes. This pattern is consistent with the difference between frontier text-to-image generation and image-editing pipelines: GPT Image 2 and Grok Imagine Image produce more visually integrated fake-damaged evidence, while our visual inspection suggests that Qwen-Image-Edit-Max more often leaves localized edit cues, such as sharper boundaries or material-transition artifacts.

Specialized Detectors Are Stronger, But Less Reliable. Specialized AI-generated image detectors generally outperform MLLMs in the single-image no-review setting. ForgeLens, pre-trained on GenImage, achieves the highest overall balanced accuracy among the listed detectors, reaching 0.904 balanced accuracy. However, these detectors are highly checkpoint- and generator-dependent. Some methods achieve high TPR on generated images but low TNR on real evidence, indicating a tendency to over-reject authentic damaged evidence. For example, Effort pretrained on Chameleon obtains high TPR across several generator subsets, with an average fake-damage TPR of 0.854, but has very low TNR on real images at 0.051, making it unsuitable for refund verification, where falsely rejecting genuine customer evidence is costly.

Context Helps Only When Models Can Aggregate Evidence. Additional context does not translate into uniform gains. Review text alone has a limited effect, with the average balanced accuracy of the six MLLMs in Tab. 2 increasing only modestly from 0.591 to 0.619 in the single-image setting, and GPT-5.6 Terra even declining from 0.626 to 0.608, suggesting that models do not consistently ground customer comments in visual evidence. Multi-image evidence is more informative: all six models improve when related images from the same review are provided jointly, with Gemini 3.8

Table 2: **Performance comparison of representative MLLMs on FRAUDBENCH under review-conditioned evaluation modes.** Results are reported across six AI image editing and generation subsets for single-image, multi-image, and multi-step settings with review context. A representative subset of six MLLMs is shown. Specialized detectors are omitted as they do not support review text, multi-image inputs, or multi-step interaction.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Single Image with Review																
DeepSeek-V4.1-Flash	**0.261**	0.727	**0.334**	0.689	**0.397**	0.689	**0.454**	0.724	**0.519**	0.720	0.612	0.721	0.658	**0.588**	0.887	0.782
Gemini 3.8 Flash	0.119	0.697	0.106	0.703	0.201	0.703	0.274	0.710	0.349	0.710	0.467	0.712	0.626	0.385	**1.000**	0.673
GPT-5.6 Terra	0.069	0.780	0.086	0.801	0.119	0.807	0.295	0.870	0.275	0.843	0.458	0.864	0.608	0.346	0.999	0.964
Grok 4.6	0.037	0.854	0.041	0.821	0.110	0.828	0.189	0.842	0.200	0.824	0.273	0.828	0.570	0.241	0.998	0.865
Kimi K3	0.036	0.714	0.082	0.713	0.076	0.719	0.202	0.766	0.215	0.739	0.357	0.762	0.581	0.271	**1.000**	0.891
Qwen3.8-Max	0.115	0.781	0.198	0.798	0.257	0.811	0.414	0.831	0.452	0.827	**0.638**	0.848	**0.672**	0.502	0.999	0.888
Multi Image with Review																
DeepSeek-V4.1-Flash	**0.327**	0.707	**0.364**	0.735	**0.438**	0.727	0.572	0.739	**0.633**	0.759	**0.789**	0.759	**0.748**	**0.671**	0.980	0.803
Gemini 3.8 Flash	0.321	0.719	0.278	0.717	0.411	0.720	0.550	0.715	0.561	0.718	0.782	0.721	0.742	0.633	**1.000**	0.687
GPT-5.6 Terra	0.152	0.899	0.170	0.897	0.276	0.902	**0.591**	0.944	0.434	0.929	0.699	0.945	0.688	0.545	**1.000**	0.967
Grok 4.6	0.068	0.851	0.037	0.814	0.124	0.841	0.294	0.855	0.232	0.862	0.366	0.849	0.591	0.303	**1.000**	0.898
Kimi K3	0.094	0.746	0.154	0.731	0.222	0.778	0.372	0.769	0.361	0.780	0.582	0.789	0.648	0.444	**1.000**	0.900
Qwen3.8-Max	0.108	0.824	0.180	0.817	0.238	0.828	0.472	0.853	0.521	0.861	0.709	0.890	0.689	0.523	**1.000**	0.915
Multi Step with Review																
DeepSeek-V4.1-Flash	**0.513**	0.788	**0.737**	0.792	**0.773**	0.816	0.751	0.830	**0.867**	0.835	**0.934**	0.851	0.816	**0.849**	0.878	0.788
Gemini 3.8 Flash	0.436	0.764	0.400	0.778	0.544	0.769	0.694	0.748	0.733	0.765	0.869	0.748	0.806	0.747	**1.000**	0.869
GPT-5.6 Terra	0.050	0.836	0.068	0.850	0.132	0.875	0.418	0.887	0.317	0.877	0.531	0.911	0.625	0.391	**1.000**	0.927
Grok 4.6	0.218	0.805	0.184	0.797	0.348	0.814	0.555	0.856	0.564	0.826	0.655	0.842	0.707	0.579	0.997	0.867
Kimi K3	0.177	0.738	0.240	0.721	0.322	0.780	0.493	0.790	0.509	0.816	0.795	0.818	0.713	0.580	**1.000**	0.902
Qwen3.8-Max	0.389	0.841	0.526	0.823	0.619	0.850	**0.797**	0.879	0.756	0.882	0.883	0.905	**0.835**	0.785	0.997	0.903

Flash rising from 0.626 to 0.742 over the single-image with-review setting. Sequential multi-step prompting further extends this gain for five of the six models, with Qwen3.8-Max improving from 0.689 to 0.835, whereas GPT-5.6 Terra falls from 0.688 to 0.625. Thus, FRAUDBENCH exposes a distinction between having more context and reliably aggregating it: cross-image cues consistently help verification, but the benefit of stateful evidence integration remains model-dependent. To further probe whether review-conditioned performance reflects genuine cross-modal reasoning or weaker semantic shortcuts, we conduct an image-text mismatch experiment in Appx. D.4.3.

4.3 SENSITIVITY TO PROMPT DESIGN

Prompting techniques are widely adopted to improve the reasoning and problem-solving capabilities of large models. We therefore examine whether prompt design can improve the visual and forensic reasoning abilities of MLLMs on FRAUDBENCH. We compare five prompt variants with progressively reduced task-specific guidance: **(1) Baseline** (full forensic prompt), **(2) Merged** (system and user instructions combined), **(3) No Checklist** (artifact checklist removed), **(4) Generic** (generic assistant identity), and **(5) Minimal** (one-sentence instruction only).

Table 3: Effect of prompt design on AI-generated evidence detection, measured by F1.

Prompt	DeepSeek	Gemini	GPT	Grok	Kimi	Qwen
Baseline	0.402	**0.150**	0.400	**0.222**	0.171	0.485
Merged	0.526	0.106	0.292	0.107	**0.263**	**0.514**
NoChk.	0.494	0.139	**0.416**	0.170	0.159	0.489
Gen.	0.507	0.098	0.207	0.095	0.106	0.433
Min.	**0.532**	0.094	0.312	0.119	0.207	0.452

As shown in Tab. 3, no single prompt is best for all models. The full forensic prompt ranks first or second in F1 for Gemini 3.8 Flash, Grok 4.6, and GPT-5.6 Terra, while a generic assistant identity lowers F1 for five of the six models. DeepSeek-V4.1-Flash gains F1 under Merged and Minimal only by sacrificing real-image TNR. This indicates that detailed forensic instructions can help MLLMs attend to relevant visual cues, yet prompt engineering alone remains insufficient for robust refund-evidence detection. Extended results and discussion appear in Appx. D.4.4.

5 CONCLUSION

AI-generated image editing has created a new platform-risk setting in which synthetic visuals can be tailored to concrete refund claims. To systematically study this threat, we introduced FRAUDBENCH, a multimodal benchmark for detecting AI-generated fraudulent refund evidence, evaluating models along five practical dimensions. Our experiments expose a fundamental "over-credulity vs. over-sensitivity" dilemma: current MLLMs preserve genuine evidence reliably but fail to detect most synthetic damage claims, often falling below random guessing, while specialized detectors achieve stronger fake-image recall but vary widely across checkpoints, with some producing unacceptable false positive rates on authentic submissions, revealing a clear gap between generic AI-generated image detection and claim-conditioned refund-evidence verification. We hope FRAUDBENCH will inspire more robust and transparent verification methods that can protect platform integrity without harming legitimate users.

AI USE STATEMENT

In this work, we used generative AI tools for data curation and for producing the benchmark artifacts themselves. MLLMs are the object of evaluation in this study, and are also used for review-media screening, privacy-oriented filtering, damage-pattern mining, and prompt construction. Image editing and generation models synthesize the fake-damaged evidence, and LLMs generate the accompanying review text. Model outputs in every construction step are treated as candidates, and human annotators make the final inclusion decision. Generative AI tools were not used to originate the research problem or to design the benchmark and its evaluation protocol.

Additionally, we used generative AI tools for coding assistance, for language and structural editing of the manuscript, and for retrieving and verifying references. Generative AI tools were also used to propose initial outlines for sections of the manuscript. The authors revised all outlines and verified the edited text against the cited sources before rewriting it. All AI-assisted code was read and tested by the authors, the reported numbers were regenerated from the released evaluation harness rather than copied from model output, and every citation was checked against its primary source. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The human reference in Tab. 1 was produced by three evaluators. They had no role in generating or selecting the synthetic evidence and saw only the image, with labels, generator identities, model predictions, and source metadata withheld during evaluation. Under our institutional policy this evaluation did not require IRB approval, and we collected no personal data from the evaluators.

The real evidence is derived from publicly accessible review data. We did not contact or attempt to identify the review authors, as the records were public and contacting them at scale was not feasible. Every item passed automated screening followed by human inspection, and items showing faces, license plates, addresses, or other directly identifying content were removed. The benchmark contains synthetic evidence of the kind used in the threat it studies, which could help an attacker refine such evidence. We release aggregate failure modes rather than guidance for constructing successful claims. Intended use, restrictions, and further discussion are set out in Appx. H and I.

REPRODUCIBILITY STATEMENT

An anonymized version of the benchmark is available at `https://dataverse.harvard.edu/previewurl.xhtml?token=0bf91c2e-daeb-408f-b08d-0826200fcc4d`, and the evaluation code at `https://anonymous.4open.science/r/FraudBench-ICLR`. The construction pipeline and all prompts are specified in Appx. C. Model identifiers, API endpoints, checkpoints, and the hardware and software environment are given in Appx. D.1.1–D.1.2, and metric definitions in Sec. 4.1 and Appx. D.2. Per-category, per-damage-type, per-prompt-variant, and mismatched-review results for the evaluated MLLMs are reported in Appx. D.3, from which every number in Tab. 1–3 can be recomputed. Compute usage is reported in Appx. G.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yingjian Chen, Lei Zhang, and Yakun Niu. ForgeLens: Data-efficient forgery focus for generalizable forgery image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16270–16280, 2025.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. CO-SPY: Combining semantic and pixel features to detect synthetic images by AI. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13455–13465, 2025.

Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. MVSS-Net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2022.

Bo Du, Xuekang Zhu, Xiaochen Ma, Chenfan Qu, Kaiwen Feng, Zhe Yang, Chi-Man Pun, Jian Liu, and Ji-Zhe Zhou. ForensicHub: A unified benchmark & codebase for all-domain fake image detection and localization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5d8a47fc1cb0a143f149ee07321621c9-Paper-Datasets_and_Benchmarks_Track.pdf.

Google. Gemini 3 Flash: frontier intelligence built for speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash, December 2025. Accessed: 2026-05-01.

Google DeepMind. Gemini 3.1 Pro, February 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Google DeepMind. Gemini 3.5 Flash-Lite, July 2026b. URL https://deepmind.google/models/model-cards/gemini-3-5-flash-lite/.

Google DeepMind. Gemini 3.8 Flash, September 2026c. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 20606–20615. IEEE, 2023.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

Yikun Ji, Yan Hong, Qi Fan, Jun Lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang. FakeXplain: AI-generated image detection via human-aligned grounded reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, et al. LEGION: Learning to ground and explain for synthetic image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18937–18947, 2025.

Chunxiao Li, Xiaoxiao Wang, Meiling Li, Boming Miao, Peng Sun, Yunjian Zhang, Xiangyang Ji, and Yao Zhu. Bridging the gap between ideal and real-world evaluation: Benchmarking AI-generated image detection in challenging scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20379–20389, 2025.

Hongyang Li. AI-driven refund scams spur calls for stronger e-platform protection. https://global.chinadaily.com.cn/a/202512/12/WS693b6f42a310d6866eb2e42f.html, December 2025a. China Daily. Accessed: 2026-05-01.

Lei Li. Police punish shopper after AI video used in fake crab claim. https://global.chinadaily.com.cn/a/202512/04/WS693171e0a310d6866eb2cee8.html, December 2025b. China Daily. Accessed: 2026-05-01.

Yiheng Li, Zichang Tan, Guoqing Xu, Zhen Lei, Xu Zhou, and Yang Yang. Towards generalizable AI-generated image detection via image-adaptive prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21262–21272, 2026.

Moonshot AI. Kimi K2.6: From Code to Creation, From One to Many. https://www.kimi.com/ai-models/kimi-k2-6, 2026. Accessed: 2026-05-01.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

OpenAI. Introducing GPT-5.4 mini and nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano, March 2026a. Accessed: 2026-05-01.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, July 2026b. URL https://openai.com/index/gpt-5-6/.

OpenAI. GPT-6 Astra: A new generation of intelligence, September 2026c. URL https://openai.com/index/gpt-6-astra/.

OpenAI. Introducing GPT-6 Sol and Luna, September 2026d. URL https://openai.com/index/introducing-gpt-6-sol-and-luna/.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8245–8257, 2025.

Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, et al. On the design of Qwen3.8-Next architecture: Evaluation, efficiency, and training stability. *arXiv preprint arXiv:2608.30320*, 2026.

Qwen Team. QVQ: To See the World with Wisdom. https://qwen.ai/blog?id=qvq-72b-preview, December 2024. Accessed: 2026-05-01.

Qwen Team. Qwen3.6-35B-A3B: Agentic Coding Power, Now Open to All. https://qwen.ai/blog?id=qwen3.6-35b-a3b, April 2026a. Accessed: 2026-05-01.

Qwen Team. Qwen3.6-Plus: Towards Real World Agents. https://qwen.ai/blog?id=qwen3.6, April 2026b. Accessed: 2026-05-01.

Qwen Team. Qwen3.7: The agent frontier, May 2026c. URL https://qwen.ai/blog?id=qwen3.7.

Qwen Team. Qwen3.7-Plus: Multimodal agent intelligence, 2026d. URL https://qwen.ai/blog?id=qwen3.7-plus.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026e. URL https://qwen.ai/blog?id=qwen3.8.

Qwen Team. Qwen3.5-Omni Technical Report. *arXiv preprint arXiv:2604.15804*, 2026f.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, 2019.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

SpaceXAI. Introducing Grok 4.6, August 2026. URL https://x.ai/news/grok-4-6.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi K3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Jiarui Wang, Huiyu Duan, Juntong Wang, Ziheng Jia, Woo Yi Yang, Xiaorong Zhu, Yu Zhao, Jiaying Qian, Yuke Xing, Guangtao Zhai, et al. DFBench: Benchmarking deepfake image detection capability of large multimodal models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12666–12673, 2025.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8695–8704, 2020.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 58972–59005. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/551d275ae5cb1bb1ebe6e37fbe201db8-Paper-Conference.pdf.

xAI. Grok 4.1 Fast and Agent Tools API. https://x.ai/news/grok-4-1-fast, November 2025. Accessed: 2026-05-01.

xAI. Grok 4.20. xAI API model documentation, 2026. URL https://docs.x.ai/developers/models/grok-4.20-reasoning. Accessed: 2026-05-01.

Anyi Xu, B Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, CC Yu, et al. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. *arXiv preprint arXiv:2609.19969*, 2026.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. FakeShield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations*, volume 2025, pp. 31186–31216, 2025.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b0303773962ea1b5394c3a83cc7dd066-Paper-Conference.pdf.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. DF40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, 37:29387–29434, 2024.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *International Conference on Machine Learning*, pp. 70268–70288. PMLR, 2025b.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. AIGI-Holmes: Towards explainable and generalizable AI-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18746–18758, 2025.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting AI-generated image. *Advances in Neural Information Processing Systems*, 36:77771–77782, 2023.

A Appendix Overview

This appendix provides supplementary material for FraudBench, organized into the following sections.

Appendix B: Benchmark Schema. Describes the dataset structure and statistics of FraudBench. Appendix B.1 introduces the overall directory schema. Appendix B.2 reports per-category review and image counts. Appendix B.3 covers image-count and language distribution. Appendix B.4 presents rating and purchase-verification statistics. Appendix B.5 shows qualitative pairs of real-undamaged reference images and the fake-damaged evidence synthesized from them.

Appendix C: Benchmark Construction. Details the full construction pipeline. Appendix C.1 describes data collection from Amazon, Trip.com, GrabFood, and self-collected sources. Appendix C.2 covers data preprocessing, including MLLM-assisted image–review alignment cleaning, distribution-aware sampling, and human verification. Appendix C.3 documents the damage pattern mining, synthetic evidence prompt construction, and multi-model image generation pipeline. Appendix C.5 describes the two-stage quality control procedure, and Appendix C.4 describes region-level annotations of the edited area.

Appendix D: Benchmark Evaluation. Provides full evaluation details. Appendix D.1 documents implementation settings for MLLMs (Appx. D.1.1), specialized detectors and image forgery detection and localization methods (Appx. D.1.2), and human evaluations (Appx. D.1.3), together with the ablation study settings (Appx. D.1.4) and the localization evaluation (Appx. D.1.5). Appendix D.2 defines all reported metrics. Appendix D.3 reports complete numerical results for the evaluated MLLMs in every setting (30 in the single-image review-free setting, 18 in the review-conditioned single-image setting, eight on all 29 categories and 18 on the 27 Amazon categories in the multi-image settings, and six in the multi-step settings). Appendix D.4 presents supplementary analyses, including category-level detection difficulty (Appx. D.4.1), damage-type sensitivity (Appx. D.4.2), image-text mismatch (Appx. D.4.3), prompt sensitivity (Appx. D.4.4), threshold-free evaluation (Appx. D.4.5), image forgery detection and localization baselines (Appx. D.4.6), localization of the manipulated region (Appx. D.4.7), and the complementarity of MLLMs and specialized detectors (Appx. D.4.8).

Appendix E: Scope of FraudBench and the Role of MLLMs. Discusses the capabilities that FraudBench measures, where current models fall short, and the value and risk of MLLM rationales.

Appendix F: Limitations. Discusses the limitations of FraudBench and directions for future work.

Appendix G: Compute Reporting. Reports computational resources used for data collection, benchmark construction, and evaluation experiments.

Appendix H: Impact Statement. Discusses the broader societal impact and dual-use considerations of this work.

Appendix I: Responsible Research Practices. Documents data privacy measures, anonymization procedures, licensing considerations, and responsible-use constraints associated with FraudBench.

B BENCHMARK SCHEMA

B.1 OVERVIEW OF FRAUDBENCH

FRAUDBENCH is organized as a category-rooted directory tree. The top level contains one folder per product or service category, covering all 29 categories: 27 Amazon product categories and two service categories (*Hotels & Accommodations* from Trip.com and *Delivery, Pickup & Dine-Out* from GrabFood). Within each category, three subdirectories partition the data by subset and purpose: `Positive`, `Negative`, and `DeepFake`. The overall structure is illustrated in Fig. 3.

Benchmark Directory Structure (Abbreviated)

```
<Category>/
+-- Positive/                      # Real-undamaged images (synthesis
    sources)
|   +-- Review_XXX/
|       +-- Image_XXX_YY.jpg
|       +-- MetaReview_XXX.json
+-- Negative/                      # Real-damaged images (evaluation
    negatives)
|   +-- Review_XXX/
|       +-- Image_XXX_YY.jpg
|       +-- MetaReview_XXX.json
+-- DeepFake/                      # Synthetic fake-damaged images
    +-- <model-name>/              # One subfolder per generation model
    |   +-- Review_XXX/
    |       +-- Image_XXX_YY.jpg
    +-- Metadata/
       +-- Edit_XXX.json          # Generation metadata per review
```

Figure 3: Top-level directory structure of FRAUDBENCH, illustrated for a single product or service category. The structure is identical across all 29 categories.

Positive and Negative Subsets. The `Positive` directory contains real-undamaged review records, which serve as generation references for synthetic evidence and are excluded from evaluation. The `Negative` directory contains real-damaged review records, which form the negative class in the binary verification task. Both subsets follow the same review-level organization: each `Review_XXX` folder contains one or more image files named `Image_XXX_YY.jpg` and a corresponding `MetaReview_XXX.json` file. The metadata file stores the review identifier, rating, review text, local image filenames, product or service category, and enriched product or hotel metadata including titles, descriptions, and features, as well as the user identifier provided by the source platform and other platform-specific fields.

DeepFake Subset. The `DeepFake` directory is organized by generation model. Each of the 12 image editing and generation models has a dedicated subfolder, containing one `Review_XXX` folder per source review with the corresponding synthesized images. Synthetic images retain the original filenames of their source images to facilitate alignment across models and with the source `Positive` records. Generation metadata is stored in the shared `Metadata` subfolder as `Edit_XXX.json` files, one per review. Each metadata file records the selected damage type, the targeted surface or component, the full edit instruction submitted to the generation models, a short rationale, and the generated review comment used as the fake review text in review-conditioned evaluation settings.

B.2 DATASET STATISTICS

FRAUDBENCH contains **822** real review samples and **13,856** images across 29 product and service categories. The real-image pool consists of 988 real-undamaged images from 422 positive reviews and 1,012 real-damaged images from 400 negative reviews. Each of the 988 real-undamaged images

is processed by all 12 generation models, yielding 11,856 synthetic fake-damaged images in total. Table 4 reports the per-category review and image counts.

Table 4: **Per-category review and image counts in FRAUDBENCH.** Pos. and Neg. denote positive (real-undamaged) and negative (real-damaged) subsets respectively. DeepFake counts equal 12 times the corresponding Pos. image count, as each real-undamaged image is processed by all 12 generation models.

Category	Pos. Reviews	Neg. Reviews	Pos. Images	Neg. Images	DeepFake Images
All Beauty	11	11	28	34	336
Amazon Fashion	11	9	35	17	420
Appliances	12	14	26	39	312
Arts, Crafts & Sewing	7	12	20	31	240
Automotive	8	9	29	30	348
Baby Products	11	17	31	56	372
Beauty & Personal Care	14	9	40	20	480
Books	10	13	15	33	180
CDs & Vinyl	10	15	19	44	228
Cell Phones & Accessories	17	15	46	42	552
Clothing, Shoes & Jewelry	13	11	47	23	564
Delivery, Pickup & Dine-Out	78	63	100	100	1,200
Electronics	19	15	52	38	624
Grocery & Gourmet Food	15	20	34	65	408
Handmade Products	16	8	50	36	600
Health & Household	15	6	35	15	420
Health & Personal Care	19	6	49	17	588
Home & Kitchen	11	18	29	48	348
Hotels & Accommodations	45	46	90	110	1,080
Industrial & Scientific	12	13	24	30	288
Magazine Subscriptions	5	2	7	2	84
Musical Instruments	7	10	16	25	192
Office Products	11	11	37	27	444
Patio, Lawn & Garden	6	7	19	25	228
Pet Supplies	3	5	12	8	144
Sports & Outdoors	17	8	40	24	480
Tools & Home Improvement	4	9	13	27	156
Toys & Games	6	10	16	28	192
Video Games	9	8	29	18	348
Total	**422**	**400**	**988**	**1,012**	**11,856**

B.3 IMAGE-COUNT AND LANGUAGE DISTRIBUTION

Table 5 reports the distribution of reviews by the number of attached images. Over 70% of reviews contain one to three images, while reviews with seven or more images account for fewer than 4%, reflecting the long-tail image-count structure typical of real platform submissions. This distribution motivates the single-image and multi-image evaluation dimensions in FRAUDBENCH, as most reviews provide limited visual context while a non-trivial minority supply richer cross-image evidence.

Table 5: **Review distribution by image-count bin.** Counts are reported separately for positive and negative subsets.

Image-Count Bin	Pos. Reviews	Neg. Reviews	Total	Proportion
1	183	166	349	42.5%
2–3	137	120	257	31.3%
4–6	90	95	185	22.5%
7–10	12	19	31	3.8%
Total	**422**	**400**	**822**	100%

Table 6 reports the review language distribution. English dominates at 88.9%, consistent with the primary Amazon e-commerce source. Chinese-language reviews account for 6.9% of the corpus:

Simplified Chinese reviews derive from the GrabFood food-delivery source in Singapore, while Traditional Chinese reviews originate predominantly from Trip.com hotel submissions. A small proportion of reviews are multilingual, with label counts summing to 837 across 822 reviews. FRAUDBENCH retains multilingual text as collected rather than normalizing to English, enabling robustness evaluation under realistic multilingual user contexts.

Table 6: **Review language distribution.** Label counts exceed the total review count because a small number of reviews contain text in more than one language.

Language	Label Count	Proportion
English	744	88.9%
Simplified Chinese	27	3.2%
Traditional Chinese	31	3.7%
Other	35	4.2%
Total label counts	**837**	—

B.4 Rating and Verification Statistics

Table 7 reports the rating distribution across all 822 reviews. The distribution is strongly bimodal: 1-star reviews account for 45.3% of the corpus and 5-star reviews for 49.8%, while 3-star neutral reviews are entirely absent. This U-shaped pattern reflects the rating-derived subset assignment used during data collection, where high-star reviews are grouped as positive (real-undamaged) evidence and low-star reviews as negative (real-damaged) evidence.

Table 7: **Rating distribution across all 822 reviews.** The bimodal concentration at 1-star and 5-star reflects the rating-derived positive/negative subset assignment.

Star Rating	Count	Proportion
1	372	45.3%
2	28	3.4%
3	0	0.0%
4	13	1.6%
5	409	49.8%
Total	**822**	100%

Among the 590 Amazon reviews, 390 (66.1%) carry a verified-purchase label and 200 (33.9%) do not. Verified-purchase status is preserved in the review metadata but is not used as an evaluation signal in FRAUDBENCH, as the benchmark focuses on visual and multimodal evidence verification rather than purchase authenticity.

B.5 Source–Synthesis Pairs

Figures 4–12 show examples of source–synthesis pairs. Each example shows a real-undamaged reference image, the edit prompt constructed for it, and the fake-damaged evidence synthesized from it by each of the 12 image editing and generation models, and the caption names the product or service category, the damage type, and the edited object. Edit prompts are truncated for brevity. In these examples, the selected damage is localized to the target object, while the original product identity, viewpoint, lighting, and background are preserved.



Figure 4: **Source–synthesis example 1: All Beauty (leaking).** A leak at the flip-top cap and upper shoulder of a cleanser bottle.



Figure 5: **Source–synthesis example 2: Amazon Fashion (torn/ripped).** A tear along the wrist cuff seam of a gardening glove.

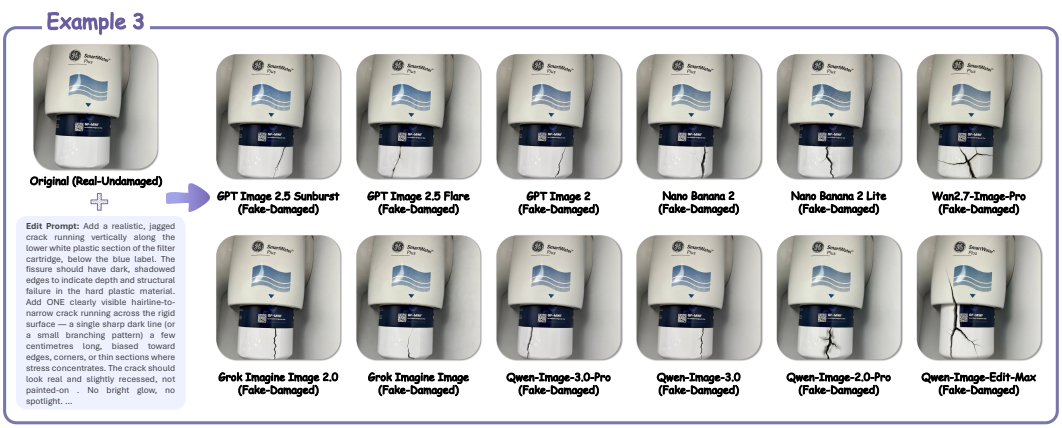

Figure 6: **Source–synthesis example 3: Appliances (cracked).** A crack in the plastic body of a refrigerator water filter cartridge.

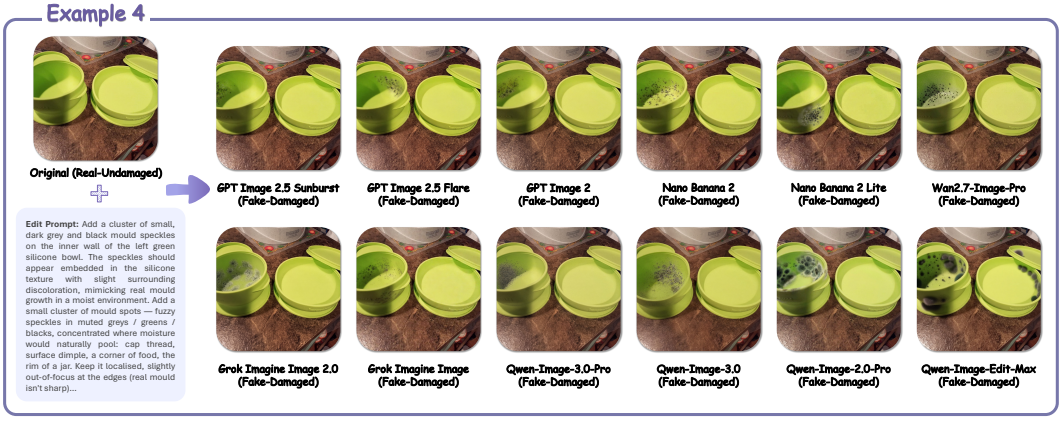

Figure 7: **Source–synthesis example 4: Baby Products (mouldy/spoiled).** Mould on the inner surface of a silicone baby bowl.

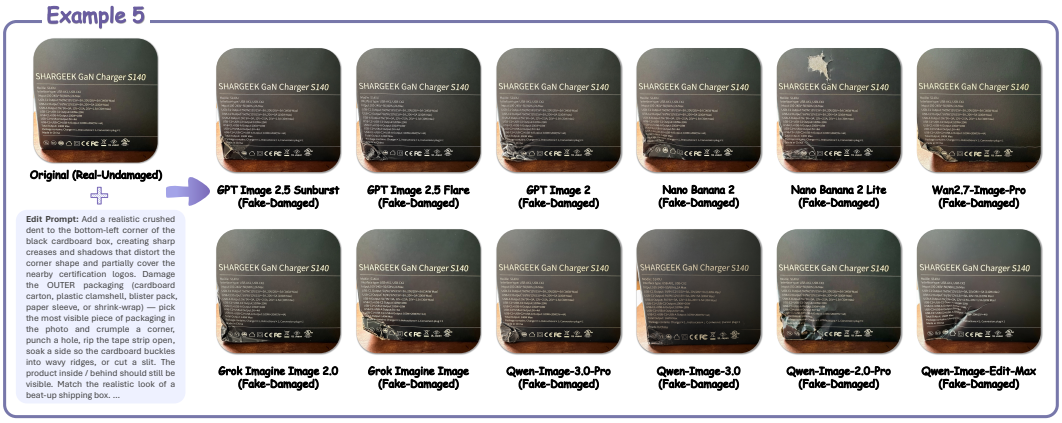

Figure 8: **Source–synthesis example 5: Cell Phones & Accessories (packaging damaged).** Damage to a corner of the retail box of a charger.

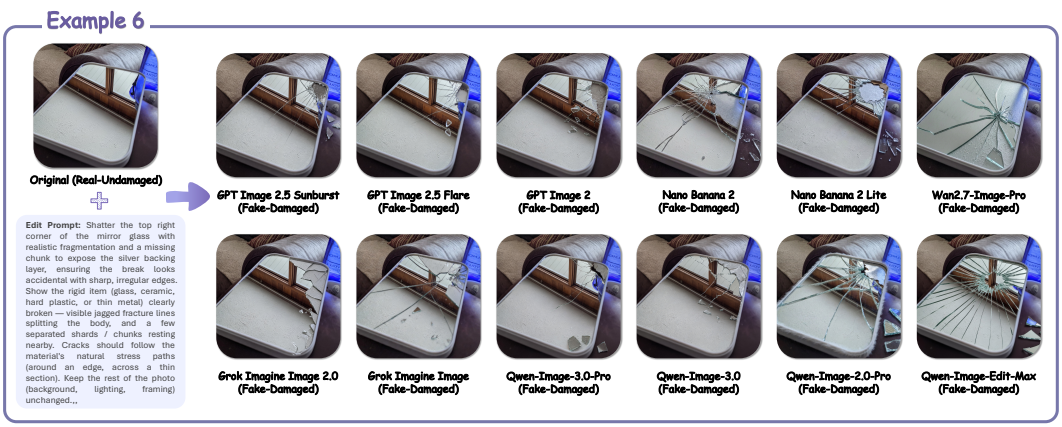

Figure 9: **Source–synthesis example 6: Beauty & Personal Care (shattered/broken).** A shattered corner of the glass of a makeup mirror.

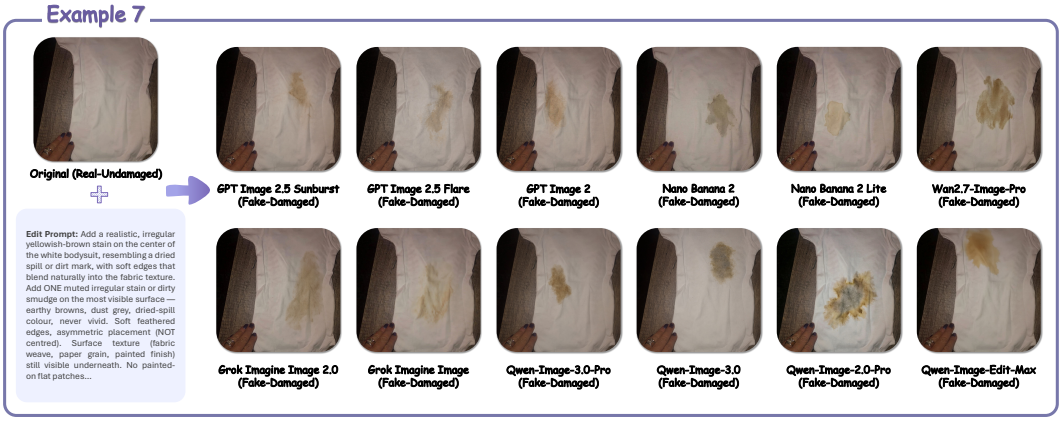

Figure 10: **Source–synthesis example 7: Clothing, Shoes & Jewelry (stained/dirty).** A stain on the front of a white infant bodysuit.

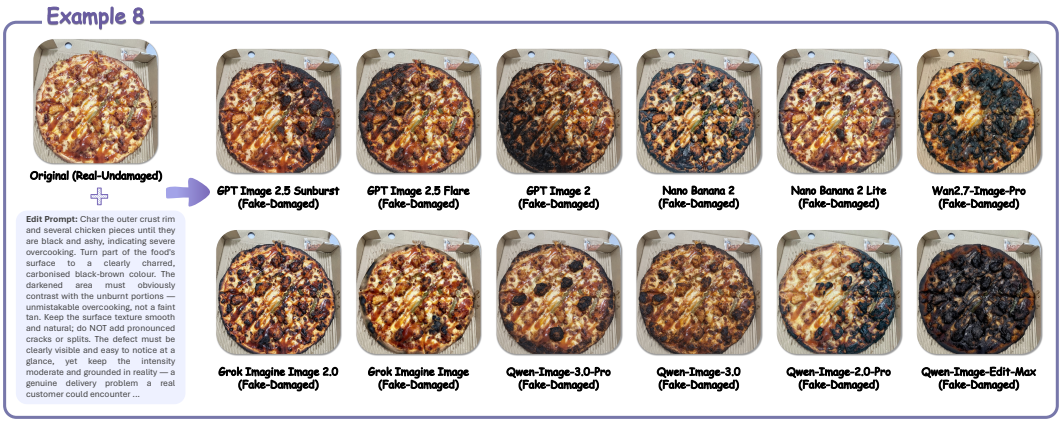

Figure 11: **Source–synthesis example 8: Delivery, Pickup & Dine-Out (burnt).** A burnt chicken pizza.

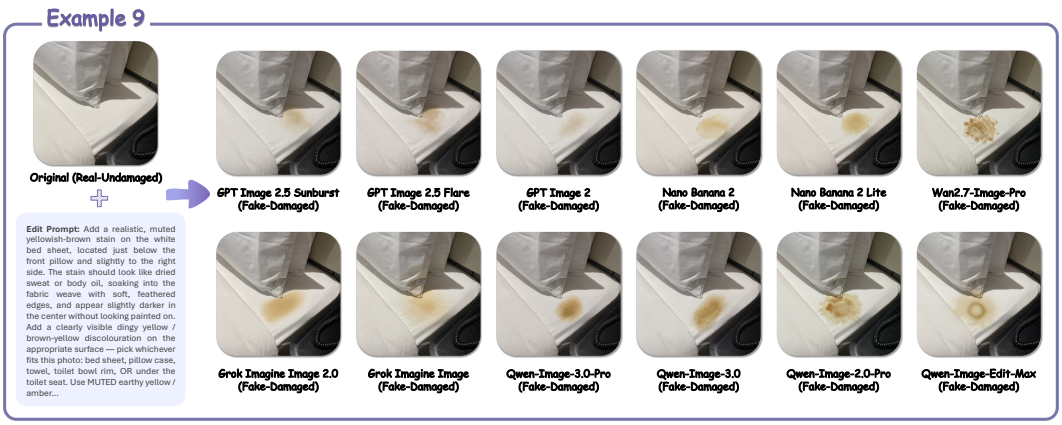

Figure 12: **Source–synthesis example 9: Hotels & Accommodations (yellow stains).** Yellow stains on a white bed sheet.

C BENCHMARK CONSTRUCTION

C.1 DATA COLLECTION

C.1.1 AMAZON

For the e-commerce source, we use the User Reviews and Item Metadata files from Amazon Reviews 2023 (Hou et al., 2024), a large-scale corpus containing user reviews, item metadata, and user–item links. The User Reviews files provide ratings, review text, helpfulness votes, timestamps, user and item identifiers, verified-purchase labels, and attached review images, while the Item Metadata files provide product titles, category hierarchies, descriptions, prices, store information, product images, and other product-level details. Representative records are shown in Fig. 13 and 14.

To collect the raw data, we parse the User Reviews JSONL files and retain records with image attachments. For each attached image, we extract the `large_image_url` field and download the corresponding high-resolution image. We keep only images satisfying a minimum resolution requirement of 1024×1024 pixels, since low-resolution images are less suitable for fine-grained evidence verification and subsequent synthetic damage generation.

For each retained review, we preserve the original review metadata, including rating, title, review text, timestamp, helpfulness votes, verified-purchase status, `asin`, `parent_asin`, and anonymized user identifier. We then enrich each review by matching `asin` and `parent_asin` to the Item Metadata files. The enriched metadata includes the product title, main category, category hierarchy, description, features, price, store information, product images, and available product details. This enrichment provides product-level context for later image–review alignment filtering and image-specific fake-damage prompt construction.

Example: Amazon User Review Record


```
{
  "sort_timestamp": 1634275259292,
  "rating": 3.0,
  "helpful_votes": 0,
  "title": "Meh",
  "text": "These were lightweight and soft but much too small for my
    liking. I would have preferred two of these together to make one
    loc. For that reason I will not be repurchasing.",
  "images": [
    {
      "small_image_url": "https://m.media-amazon.com/images/I/81
    FN4c0VHzL._SL256_.jpg",
      "medium_image_url": "https://m.media-amazon.com/images/I/81
    FN4c0VHzL._SL800_.jpg",
      "large_image_url": "https://m.media-amazon.com/images/I/81
    FN4c0VHzL._SL1600_.jpg",
      "attachment_type": "IMAGE"
    }
  ],
  "asin": "B088SZDGXG",
  "verified_purchase": true,
  "parent_asin": "B08BBQ29N5",
  "user_id": "AEYORY2AVPMCPDV57CE337YU5LXA"
}
```


Figure 13: Structure of an Amazon user review record.

After downloading and enriching the review records, we organize the retained Amazon data by category and rating-derived subset. Each product category is stored as a top-level directory. Within each category, reviews are divided into `Positive` and `Negative` subsets. The `Positive` subset contains high-rating reviews with no apparent damage evidence, corresponding to genuine undamaged evidence. The `Negative` subset contains low-rating reviews with damage- or complaint-related

evidence, corresponding to genuine damaged or problematic evidence. The rating is used only as an initial grouping signal; subsequent textual and visual filtering with human verification remove ambiguous or misaligned cases.

We then normalize the storage format at the review level. Each retained review is assigned a sequential identifier and stored as `Review_ID` under its corresponding category and subset. The associated review images are renamed as `Image_ID_SUBID.jpg`, where `ID` denotes the review index and `SUBID` denotes the image index within that review. The enriched metadata are saved as `MetaReview_ID.json`, and the local image filenames are written back into the JSON file under the `image_files` field. The resulting structure is illustrated in Fig. 15.

C.1.2 TRIP.COM

For the travel and hospitality source, we collect hotel review data from Trip.com, a large-scale online travel platform that provides hotel listings, hotel-level information, guest ratings, textual reviews, and user-uploaded review media. We use Trip.com as a complementary non-e-commerce source because hotel reviews contain multimodal evidence about room conditions, facilities, cleanliness, service quality, and other stay-related experiences. These reviews include both normal-condition evidence and genuine problematic evidence, making them suitable for constructing fraudulent refund or compensation evidence scenarios in the travel domain.

To collect hotel candidates, we define target cities across different geographic regions. In our implementation, the selected cities include Tokyo, Seoul, Paris, London, Cairo, Rio de Janeiro, and Melbourne, which ***increase geographic and cultural diversity***. For each city, we use the corresponding Trip.com city hotel-list page, where the URL is constructed from the city slug and city identifier. We parse the listing page to extract hotel identifiers, hotel names, hotel URLs, representative hotel images, rating values, review counts, and other available listing-level fields. Hotels are retained according to the default recommended order returned by Trip.com, and the top-ranked hotels in each city are selected as candidate hotels. This preserves the platform's ranking signal while yielding a diverse set of hotel properties.

For each selected hotel, we collect guest reviews by programmatically querying the Trip.com hotel comment interface. Each request specifies the hotel identifier, page index, page size, locale, currency, check-in date, check-out date, and review filter options. The interface returns structured review records, including review text, rating information, user-related fields, hotel reply information, and attached image or video URLs. Since our benchmark focuses on multimodal evidence rather than text-only complaints, we retain only reviews containing at least one image or video attachment. We use the image-video review filter to prioritize reviews with visual evidence, and for negative reviews we additionally apply the negative-review filter provided by the interface. Pagination continues until the target number of reviews is reached or the maximum page limit is exhausted.

We organize the retained Trip.com reviews into two rating-derived subsets. The `positive` subset contains 5-star reviews and represents normal or satisfactory hotel conditions. The `negative` subset contains low-rating reviews, with ratings from 1.0 to 2.0, and represents genuine problematic hotel evidence, such as dirty rooms, broken facilities, poor maintenance, pests, stains, mold, plumbing issues, or unclean bathrooms. Ratings are used only as an initial grouping signal; subsequent image–review relevance filtering and manual inspection remove irrelevant, ambiguous, or visually uninformative cases.

For each retained review, we preserve the available review-level metadata, including the review identifier, hotel identifier, hotel name, city, review URL, rating, star value, username, user avatar URL, travel type, room type, review text, creation date, check-in date, useful-vote count, hotel reply, and the original URLs of attached images and videos. Fig. 16 shows a representative retained Trip.com review-level record after metadata parsing and media URL extraction.

For hotel-level context, we also preserve the corresponding hotel metadata including the hotel identifier, hotel name, hotel URL, rating value, review count, representative image, city, and the source field indicating how the hotel was discovered. For each attached media item, we download the corresponding image or video into the review folder. Since Trip.com image URLs may contain thumbnail-resolution suffixes, we normalize image URLs where possible before downloading the media. Each downloaded file is assigned a deterministic local filename based on the original URL and written back into the review metadata. This keeps each review folder self-contained and aligns

Example: Amazon Item Metadata Record


```
{
  "main_category": "All Beauty",
  "title": "Lurrose 100Pcs Full Cover Fake Toenails Artificial
    Transparent Nail Tips Nail Art for DIY",
  "average_rating": 3.7,
  "rating_number": 35,
  "features": [
    "The false toenails are durable with perfect length. You have the
    option to wear them long or clip them short, easy to trim and file
     them to in any length and shape you like.",
    "ABS is kind of green enviromental material, and makes the nails
    durable, breathable, light even no pressure on your own nails.",
    ......
  ],
  "description": [
    "Description",
    "The false toenails are durable with perfect length. You have the
    option to wear them long or clip them short, easy to trim and file
     them to in any length and shape you like. Plus, ABS is kind of
    green enviromental material, and makes the nails durable,
    breathable, light even no pressure on your own toenails. Fit well
    to your natural toenails. Non toxic, no smell, no harm to your
    health.",
    "Feature",
    "- Color: As Shown.- Material: ABS.- Size: 14.3 x 7.2 x 1cm.",
    "Package Including",
    "100 x Pieces fake toenails"
  ],
  "price": 6.99,
  "images": [
    {
      "hi_res": "https://m.media-amazon.com/images/I/41a1Sj7Q20L.
    _SL1005_.jpg",
      "thumb": "https://m.media-amazon.com/images/I/31dlCd7tHSL._SS40_
    .jpg",
      "large": "https://m.media-amazon.com/images/I/31dlCd7tHSL.jpg",
      "variant": "MAIN"
    },
    ......
  ],
  "videos": [],
  "bought_together": null,
  "store": "Lurrose",
  "categories": [],
  "details": {
    "Color": "As Shown",
    "Size": "Large",
    "Material": "Acrylonitrile Butadiene Styrene (ABS)",
    "Brand": "Lurrose",
    "Style": "French",
    "Product Dimensions": "5.63 x 2.83 x 0.39 inches; 1.9 Ounces",
    "UPC": "799768026253",
    "Manufacturer": "Lurrose"
  },
  "parent_asin": "B07G9GWFSM"
}
```


Figure 14: Structure of an Amazon item metadata record.

Example: Amazon Enriched Review-Level Record


```
{
  "rating": 1.0,
  "title": "Damaged",
  "text": "Product arrived shattered",
  "images": [
    {
      "small_image_url": "https://m.media-amazon.com/images/I/71+
    Jx6Xx2GL._SL256_.jpg",
      "medium_image_url": "https://m.media-amazon.com/images/I/71+
    Jx6Xx2GL._SL800_.jpg",
      "large_image_url": "https://m.media-amazon.com/images/I/71+
    Jx6Xx2GL._SL1600_.jpg",
      "attachment_type": "IMAGE"
    }
  ],
  "asin": "B07QLN67ZB",
  "parent_asin": "B07QLN67ZB",
  "user_id": "AEYT2WPKKFXMRKEBZ27KV445J5ZA",
  "timestamp": 1632690103586,
  "helpful_vote": 0,
  "verified_purchase": true,
  "review_id": "1632690103586_AEYT2WPKKFXMRKEBZ27KV445J5ZA",
  "date": "2021-09-27",
  "image_files": [
    "Image_001_01.jpg"
  ],
  "product_meta": {
    "main_category": "All Beauty",
    "title": "Bzbuy 4 Nail Art Acrylic Liquid Powder Dappen Dish Glass
     Crystal Cup Glassware Tools",
    "average_rating": 4.2,
    "rating_number": 217,
    "features": [],
    "description": [
      "Bzbuy 4 Pieces Nail Art Acrylic Liquid Powder Dappen Dish Glass
     Crystal Cup Glassware Tools"
    ],
    "price": 6.0,
    "images": [
      {
        "thumb": "https://m.media-amazon.com/images/I/41iGFLpfJeL.
    _SS40_.jpg",
        "large": "https://m.media-amazon.com/images/I/41iGFLpfJeL.jpg
    ",
        "variant": "MAIN",
        "hi_res": "https://m.media-amazon.com/images/I/81TaeLP9CfL.
    _SL1500_.jpg"
      },
      ......
    ],
    "videos": [],
    "store": "Bzbuy",
    "categories": [],
    "details": {
      "Product Dimensions": "0.89 x 0.55 x 0.55 inches; 2.88 Ounces",
      "UPC": "619317532250"
    },
    "parent_asin": "B07QLN67ZB",
    "bought_together": null
  }
}
```


Figure 15: Structure of an enriched Amazon review-level metadata record after matching a retained user review with its corresponding item metadata.

the review text, rating information, hotel context, and visual evidence for later image–text consistency checking, fake-evidence prompt construction, and benchmark evaluation.

Finally, we normalize the Trip.com storage format at the city, hotel, and review levels. Each city is stored as a top-level directory, and each selected hotel is stored under a directory named by its hotel identifier and sanitized hotel name. Hotel-level metadata are saved as `Hotel_Info.json`. Under each hotel directory, retained reviews are divided into `Positive` and `Negative` folders. Each review is stored as `Review_ID`, containing the downloaded media files and a `MetaReview_ID.json` file with parsed metadata, original media URLs, and local media paths. Reviews requiring additional inspection are saved into `Manual_MetaReview.jsonl`.

Example: Trip.com User Review Record


```
{
  "review_id": "1547772078",
  "hotel_id": 2197402,
  "hotel_name": "Travelodge Hotel Melbourne Docklands",
  "city": "Melbourne",
  "review_url": "https://www.trip.com/hotels/detail/?hotelId=2197402",
  "rating": 1.0,
  "stars": 1,
  "username": "[anonymized]",
  "user_avatar": "[omitted]",
  "travel_type": "Family",
  "room_type": "Guest King Room",
  "content": "Terrible experience - not worth \$250 at all!\n\nThis
    was one of the worst stays I've ever had. The room is small, dark,
     and outdated. They joined two single mattresses and called it a
    king - horrible idea. It was so uncomfortable I woke up with back
    pain. The blanket was old, both bedside lamps didn't work, and
    there was thick dust under the bed, clearly not cleaned properly.\
    n\nThe bathroom was tiny with cheap fittings, and the toilet paper
     holder kept falling off during use. They also charged me \$1
    extra for a drink compared to the price on the menu - sneaky and
    dishonest.\n\nThis place seriously needs an upgrade. I won't come
    back, and I absolutely do not recommend it. Even budget motels are
     better than this!",
  "create_date": "2025-06-08 21:47:31",
  "check_in_date": "2025-06-01 00:00:00",
  "images": [
    {
      "original_url": "https://ak-d.tripcdn.com/images/0232
    w224x9227yyg2ED86.jpg"
    }
  ],
  "videos": [],
  "useful_count": 0,
  "reply": ""
}
```


Figure 16: Structure of a retained Trip.com user review record. User names and avatar links are redacted in this figure.

C.1.3 GRABFOOD

For the food delivery source, we collect restaurant review data from GrabFood, a large-scale food delivery platform that provides restaurant listings, merchant-level information, ratings, textual reviews, and user-uploaded review media. We use GrabFood as a complementary food-service source because food delivery reviews contain multimodal evidence about food quality, packaging condition, hygiene, freshness, and order presentation. Such reviews include both normal-condition food

evidence and genuine problematic food evidence, making them suitable for constructing fraudulent refund or complaint evidence scenarios in the food delivery domain.

To collect restaurant candidates, we use Singapore as the target city, with the corresponding latitude, longitude, country code, GrabFood country path, and language configuration. We query the GrabFood guest category interface using the city coordinates and the food-category shortcut identifier. The returned listing records are parsed to extract restaurant identifiers, restaurant names, cuisine tags, ratings, addresses, coordinates, representative image URLs, price levels, opening status, and raw listing metadata. We further filter out non-restaurant businesses, such as grocery stores, convenience shops, supermarkets, pharmacies, pet stores, flower shops, health and beauty stores, and electronics shops. The retained stores are used as restaurant-level collection targets.

For each selected restaurant, we collect user reviews through an Apify GrabFood review actor, which returns structured review records from GrabFood. Reviews are fetched page by page using the restaurant identifier, language setting, and pagination token. Pagination continues until the target numbers of `positive` and `negative` reviews are reached, the maximum page limit is exhausted, or no further review token is returned. Since our benchmark focuses on multimodal evidence rather than text-only feedback, we retain only reviews with non-empty review text and at least one image or video attachment.

We organize the retained GrabFood reviews into two rating-derived subsets. The `positive` subset contains high-rating reviews, with 5-star reviews used as the primary source and 4-star reviews used as a fallback when needed. This subset represents normal or satisfactory delivered food evidence. The `negative` subset contains low-rating reviews, with 1-star reviews used as the primary source and 2-star reviews used as a fallback. This subset represents genuine problematic food delivery evidence. Ratings are used only as an initial grouping signal; subsequent multimodal review filtering removes irrelevant, subjective, or visually uninformative cases.

For each retained review, we preserve the available review-level metadata, including the review identifier, restaurant identifier, restaurant name, city, reviewer name, rating, review type, review text, posting time, scraping time, order identifier, and the original URLs of attached photos and videos. For restaurant-level context, we also preserve store metadata, including the store identifier, store name, city, cuisine tags, rating, address, latitude, longitude, representative image URL, price level, and opening status. This aligns the visual evidence with restaurant context for later image–text consistency checking and fake-evidence prompt construction.

For each attached media item, we download the corresponding image or video into the review folder. The downloader supports both image and video content types, assigns deterministic local filenames, and writes these filenames back into the review metadata. Failed downloads are skipped after exhausting retry attempts, while successfully downloaded files are stored together with the corresponding review JSON. This makes each review folder self-contained for benchmark construction and evaluation.

Finally, we normalize the GrabFood storage format at the city, restaurant, and review levels. Each city is stored as a top-level directory, and each selected restaurant is stored under a directory named by its sanitized restaurant name and restaurant identifier. Restaurant-level metadata are saved as `Store_Info.json`. Under each restaurant directory, retained reviews are divided into `Positive` and `Negative` folders. Each review is stored as `Review_ID`, containing the downloaded media files and a `MetaReview_ID.json` file with parsed metadata, original media URLs, and local media paths.

C.1.4 Self-Collected Real-World Evidence

In addition to the platform-derived data, we include a small set of self-collected images to improve coverage of common real-world refund-evidence scenarios that are underrepresented in the raw sources. This subset contains 24 images from seven review-level records (22 real-undamaged and two real-damaged images), all assigned to the Hotels & Accommodations category according to their visible content; the real-undamaged images account for about 2.2% of the `Positive` images. For each selected image, we constructed a small number of review-level metadata records by following the schema of existing `MetaReview` JSON files from the same category. These manually constructed records include category information, review text, rating-derived subset label, local image filename, and other fields required by the downstream data-loading and prompt-generation pipeline.

Example: GrabFood User Review Record


```
{
  "review_id": "297f4d91c9528a6a0f4719af3c2d80b0",
  "store_name": "Alt Pizza - Balestier",
  "store_id": "4-CYVAPGJZJ63GEX",
  "city": "Singapore",
  "reviewer_name": "[anonymized]",
  "rating": 5,
  "review_type": "positive",
  "review_text": "Love the pizza so much! Taste great!",
  "posted_at": "2025-08-24T09:57:50Z",
  "scraped_at": "2026-04-09T08:54:10.427802+00:00",
  "order_id": "",
  "photos": [
    {
      "id": "2czN2QzMxMTN1IjN1kTOykTM",
      ......
    },
    ......
  ],
  "videos": [],
  "has_media": true,
  "media_count": 3
}
```


Figure 17: Structure of a retained GrabFood user review record. User name is redacted in this figure.

The self-collected subset is used only as a minor supplement to the platform-derived data, providing additional coverage of realistic user-submitted evidence patterns and is separately documented for provenance tracking.

C.1.5 Data Usage and Licensing

The benchmark is constructed from publicly accessible review data across the above sources. Amazon Reviews 2023 is used in accordance with its publicly documented access terms (Hou et al., 2024). Trip.com and GrabFood review records were collected from publicly accessible platform pages via API-based requests and direct collection. Because the copyright and licensing status of user-generated review media is not uniform across source platforms, we do not assert that all third-party review text, images, or metadata are covered by a single standardized open license. The released benchmark is restricted to research use, and users are responsible for complying with the applicable terms of the original platforms and data sources. Further details on privacy handling and responsible-use constraints are provided in Appx. I.

C.2 Data Preprocessing

We apply a multimodal preprocessing pipeline before constructing the benchmark. The pipeline covers MLLM-assisted image–review alignment filtering, human-conducted text and image cleaning, privacy-oriented anonymization, and human-verified quality control.

Image–Review Alignment Cleaning. To improve data quality, we apply platform-specific MLLM-based image relevance filters before finalizing each review record. In our implementation, each filter is built with `Qwen3.6-Plus`, accessed through the official API with default inference parameters, and returns a JSON object with a binary label and a short reason. The MLLM serves only as a preliminary screener; all final inclusion decisions are made by human annotators.

For Trip.com hotel reviews, the system prompt (Fig. 18) defines the model as a hotel review image relevance classifier. For `positive` reviews, the filter retains images showing hotel-related content, such as rooms, bathrooms, beds, lobbies, restaurants, pools, gyms, or amenities, and rejects

images dominated by unrelated scenery, tourist attractions, selfies, or unrecognizable content. For `negative` reviews, the filter retains images showing visible hotel problems, such as stains, dirt, damaged facilities, pests, mold, poor maintenance, plumbing issues, or unclean bathrooms.

Prompt: Trip.com Image Relevance Filtering Prompt

```
You are a hotel review image relevance classifier. You will receive
one or more images
from a hotel review together with the review text and the sentiment
label (positive or
negative). Your job is to decide whether the images are RELEVANT to
the hotel review
context.

Respond ONLY with a JSON object: {"relevant": true/false, "reason":
"..."}
Do NOT include any other text.
```

Figure 18: System prompt used for MLLM-based image relevance filtering of Trip.com hotel review media.

For GrabFood food delivery reviews, the system prompts (Figs. 19 and 20) define separate criteria for each sentiment. For `positive` reviews, the filter retains samples whose images show the purchased food or drink item and removes non-food content such as app screenshots, delivery-rider photos, packaging-only photos, selfies, or unrecognizable images. For `negative` reviews, the filter retains images showing visible food-quality or packaging problems, such as spoilage, mold, foreign objects, undercooked food, spilled or crushed packaging, leaking or dirty containers, or hygiene concerns. Samples describing only subjective opinions, price complaints, delivery speed, or vague dissatisfaction are removed unless the attached media shows a verifiable quality issue.

Prompt: GrabFood Positive Review Filtering Prompts

```
You are a food delivery review classifier. You will receive the
review text AND the
attached photo(s).

Given a single positive review with its photo(s), decide whether to
**keep** or
**remove** it.

## KEEP (label = "keep")
The photo shows the actual purchased food or drink item.

## REMOVE (label = "remove")
The photo does NOT show food/drink, for example:
- Screenshot of the app or order page
- Photo of the delivery rider or packaging only
- Selfie, receipt, or any non-food image
- Blurry or unrecognizable image

Respond with a JSON object:
{"label": "keep" or "remove", "reason": "one-sentence explanation, 15
words max"}

Return ONLY the JSON object, no markdown fences, no extra text.
```

Figure 19: System prompts used for MLLM-based filtering of GrabFood positive review media, retaining reviews whose images show the purchased food or drink.

Prompt: GrabFood Negative Review Filtering Prompts

```
You are a food delivery review classifier. You will receive the
review text AND the
attached photo(s).

Given a single negative review with its photo(s), decide whether to
**keep** or
**remove** it.

## KEEP (label = "keep")
The photo shows a visually observable food quality issue:
- Food spoilage or mold
- Foreign objects found in food (hair, insects, plastic, metal, etc.)
- Undercooked or raw
- Packaging problems: spilled, crushed, leaking, or dirty container
- Hygiene or cleanliness concerns

## REMOVE (label = "remove")
- The photo does not show any visible food quality issue
- False advertising: item received looks significantly different from
the menu photo, or advertised ingredients are missing
- Purely subjective taste opinions ("not tasty", "too sweet", "too
salty", "bland", "I didn't like the flavor")
- Price or value complaints only ("too expensive", "small portion for
the price")
- Delivery speed complaints only ("took too long"), unless the delay
caused food quality degradation
- Missing items or wrong items delivered
- Allergic reactions caused by undisclosed ingredients
- Vague dissatisfaction with no specific quality issue
("disappointing", "not worth it", "won't order again") unless
combined with a keep-worthy reason
- Rating-only reviews with no actionable quality detail

Respond with a JSON object:
{"label": "keep" or "remove", "reason": "one-sentence explanation, 15
words max"}

Return ONLY the JSON object, no markdown fences, no extra text.
```

Figure 20: System prompts used for MLLM-based filtering of GrabFood negative review media, retaining reviews whose images show visible food-quality or packaging issues.

Distribution-Aware Sampling. After filtering, we apply distribution-aware sampling to preserve structural diversity and evaluation usability across three dimensions. First, all 29 product and service categories are retained in the final benchmark. At the category level, the sampled distribution is kept as close as possible to the post-filtering distribution, avoiding over-representation of frequent categories while preserving the natural long-tail structure of the collected data. Second, multilingual review text is retained as collected rather than normalized to English, reflecting the natural language distribution of the source platforms and enabling evaluation under realistic multilingual user contexts. Third, within each category, reviews are stratified by the number of attached images into four bins: single-image (1), small multi-image (2–3), medium multi-image (4–6), and large multi-image (7–10). Sampling targets are assigned to maintain approximately balanced coverage across bins, with single-image reviews allocated a slightly larger share to reflect their higher prevalence in real platform submissions. This stratification ensures that the benchmark supports both single-image evaluation and multiple levels of multi-image context.

Human Verification. Human verification is conducted jointly by two annotators, the first two authors. The first author performs a systematic review of each retained record by inspecting the review text and attached images as stored in the `MetaReview` JSON files. This inspection serves two purposes: content verification and privacy-oriented anonymization. For content verification, each sample is assessed for task relevance, image–review alignment, and evidence usability; ambiguous, misleading, or insufficiently informative cases are flagged for exclusion. For anonymization, images and metadata are inspected for privacy-sensitive content, including visible human faces, personal identifiers such as names, addresses, order numbers, or contact information, and other directly identifying content; samples containing such information that cannot be reliably de-identified are excluded from the benchmark. The second author independently cross-examines the processed records to verify the consistency of inclusion decisions and the completeness of the anonymization procedure. Disagreements are resolved through joint discussion, and the final inclusion decision requires agreement from both annotators.

C.3 Synthetic Evidence Generation

Models. We use 12 state-of-the-art image editing and generation models to synthesize fake-damaged evidence: GPT Image 2, Grok Imagine Image, Nano Banana 2, Qwen-Image-2.0-Pro, Qwen-Image-Edit-Max and Wan2.7-Image-Pro, later extended with GPT Image 2.5 Flare, GPT Image 2.5 Sunburst, Grok Imagine Image 2.0, Nano Banana 2 Lite, Qwen-Image-3.0 and Qwen-Image-3.0-Pro to keep the threat model current as stronger generators reached general availability. The selection spans both image-editing and image-generation pipelines, covering a range of frontier capabilities across four vendors. GPT Image 2.5 represents a recent milestone in photorealistic image synthesis, while the remaining models offer competitive quality and serve as meaningful points of comparison.

Table 8: **Version and access details for image editing and generation models.** The table lists the display name used in the paper, the exact model identifier, and the API endpoint for each model.

Display Name	Exact Model ID	API Endpoint
GPT Image 2	`gpt-image-2`	`api.openai.com/v1`
GPT Image 2.5 Flare	`gpt-image-2.5-flare-2026-09-08`	`api.openai.com/v1`
GPT Image 2.5 Sunburst	`gpt-image-2.5-sunburst-2026-09-08`	`api.openai.com/v1`
Grok Imagine Image	`grok-imagine-image`	`api.x.ai/v1`
Grok Imagine Image 2.0	`grok-imagine-image-2.0`	`api.x.ai/v1`
Nano Banana 2	`gemini-3.1-flash-image-preview`	`generativelanguage.googleapis.com/v1beta`
Nano Banana 2 Lite	`gemini-3.1-flash-lite-image`	`generativelanguage.googleapis.com/v1beta`
Qwen-Image-2.0-Pro	`qwen-image-2.0-pro`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen-Image-3.0	`qwen-image-3.0`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen-Image-3.0-Pro	`qwen-image-3.0-pro`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen-Image-Edit-Max	`qwen-image-edit-max`	`dashscope.aliyuncs.com/compatible-mode/v1`
Wan2.7-Image-Pro	`wan2.7-image-pro`	`dashscope.aliyuncs.com/compatible-mode/v1`

Table 8 lists the exact model identifiers and API endpoints. Critically, all 12 are accessible through mainstream AI platforms or official API services with minimal configuration, reflecting a realistic threat model in which low-barrier tools can already be leveraged for synthetic evidence creation. All models are accessed through their respective official or authorized API services using provider-

default inference settings, and we do not manually configure sampling parameters such as temperature, top-p, or maximum output tokens.

The two GPT Image 2.5 variants are sibling models split by workload rather than by capability, and both are pinned to the `2026-09-08` snapshot so that the benchmark remains reproducible after the bare alias moves. Nano Banana 2 is likewise pinned to `gemini-3.1-flash-image-preview`, the only variant exposed when the corresponding images were generated. A stable `gemini-3.1-flash-image` has since been released, but we retain the preview identifier so that the released images match the endpoint that produced them.

Damage Pattern Mining. Before constructing image-specific editing prompts, we first characterize the damage landscape within each data source by systematically analyzing real negative-review evidence. For each platform, we deploy a `Qwen3.6-Plus`-based multimodal analysis script that processes every retained negative-review record, jointly inspecting the attached images and the associated review text. The model is prompted to identify the product or scene category, the most prominent damage or quality issue visible in the images, and whether the visual evidence is consistent with the reviewer's written complaint. The analysis is designed to be source-specific: for Amazon e-commerce reviews, the prompt additionally leverages the product title, top-level category, and listing description to ground the damage identification within the known product context (Fig. 21); for Trip.com hotel reviews, the prompt focuses on in-room locations and surface-level defects (Fig. 22); and for GrabFood food delivery reviews, the prompt targets food category identification and condition assessment (Fig. 23). Each analysis call returns a structured JSON object containing the damage type, a short issue description, a review-consistency judgment, and a confidence level. The per-review outputs are aggregated by product or service category to produce category-specific candidate damage lists, which enumerate the most frequently observed and visually verifiable defect types for each category. These lists form the basis of the subsequent image-specific prompt construction pipeline.

Image-Specific Prompt Construction. The damage pattern mining yields prioritized category-specific damage shortlists that form the structured defect vocabulary for prompt construction. For Amazon e-commerce data, the mining covers all 27 product categories and identifies 12 visually verifiable damage types: *shattered/broken*, *cracked*, *dented*, *scratched*, *leaking*, *stained/dirty*, *mouldy/spoiled*, *melted*, *bent/warped*, *torn/ripped*, *packaging-damaged*, and *other_defect*. The per-category shortlists are shown in Tab. 9; within each row, damage types are ordered by their empirical frequency in real negative reviews, so the first entry reflects the most commonly observed defect in that category. Damage types that are inherently non-visual or inference-heavy, including *missing parts*, *defective function*, *used/opened*, and *wrong item*, are excluded from the vocabulary because they cannot be reliably synthesized as visually verifiable evidence. For Trip.com hotel reviews, the analysis produces location-aware damage lists organized by in-room location (e.g., Bed/Bedding, Bathroom, Floor/Carpet), covering defect types such as stained linens, mould, broken fixtures, and surface contamination. For GrabFood food delivery reviews, the analysis produces food-category-level damage lists covering quality issues such as spoilage, foreign objects, spillage, and packaging damage.

Building on the category-specific damage shortlists characterized above, we construct synthetic evidence prompts through a two-stage pipeline. In the first stage, the shortlists serve as a structured vocabulary of visually verifiable defects for each product category. In the second stage, for each target image, an MLLM analyzes the visible product, packaging, material, and physical attributes, and combines this visual analysis with product metadata including the title, category, description, and features. The corresponding category-specific shortlist is then retrieved, and the most plausible damage pattern is selected according to the object, material, scene, and metadata. When no category-specific candidate is suitable, the selection falls back to the universal damage vocabulary. The selected damage pattern is then converted into an image-specific editing prompt. The prompt specifies that the modification should be visible, photorealistic, physically plausible, and localized to the target object, while preserving the original product identity, viewpoint, lighting, background, surrounding objects, and overall composition. Each generated sample is also paired with a fake review comment, which is later used in review-conditioned evaluation.

Prompt: Amazon Damage Analysis Prompt

```
You are a senior product-quality auditor specialising in e-commerce
review
images on Amazon. The user will tell you the *known* Amazon top-level
category for the listing (e.g. "Electronics", "All_Beauty", "Books")
AND
the listing's product name and short description so you know exactly
what
the buyer ordered. Your task is to analyse the review image(s)
together
with the reviewer's written comment and identify:

  1. product_subtype - a more specific product label within the known
category (e.g. for "Electronics": "Wireless Earbuds", "USB Cable",
"Laptop Charger"; for "All_Beauty": "Lipstick", "Shampoo Bottle",
"Perfume Bottle"; for "Books": "Hardcover Book", "Paperback Book").
Prefer wording consistent with the supplied product name.
  2. damage_type - ONE concise label describing the most prominent
issue, drawn from this controlled vocabulary where possible:
      normal, shattered/broken, cracked, dented, scratched, leaking,
      stained/dirty, mouldy/spoiled, melted, bent/warped, torn/ripped,
      missing_parts, wrong_item, used/opened, packaging_damaged,
      expired, defective_function, color_or_print_wrong, size_wrong,
      counterfeit_suspected, other_defect.
    If the item looks fine, use "normal".
  3. issue_detail - one or two sentences describing what you actually
see in the image(s).
  4. review_consistent - true if the visual evidence matches what the
reviewer wrote, otherwise false.
  5. confidence - high | medium | low.

Respond ONLY with a valid JSON object matching this schema, no prose
outside JSON:
{
  "product_subtype": "<string>",
  "damage_type": "<string - one label from the vocabulary above>",
  "issue_detail": "<string>",
  "review_consistent": <true | false>,
  "confidence": "<high | medium | low>"
}
```

Figure 21: System prompt used for MLLM-based damage pattern analysis of Amazon e-commerce review images.

Prompt: Trip.com Damage Analysis Prompt

```
You are a senior hospitality quality auditor for online travel
platforms
(e.g. Trip.com). Every image you analyse has been uploaded with a
guest
review and is assumed to show something INSIDE the guest's hotel room.
Public/shared areas (lobby, reception, corridor, elevator, pool, gym,
restaurant, breakfast area, exterior of the building) are OUT OF
SCOPE.
If an image clearly shows such a public area, label its room_location
as
"Not In-Room".

Your task is to identify:
  1. The specific IN-ROOM location shown. Pick the single most
appropriate
label from this closed list:
    - "Bed / Bedding" (mattress, sheets, pillows, headboard)
    - "Bathroom - Shower / Tub"
    - "Bathroom - Toilet"
    - "Bathroom - Sink / Vanity"
    - "Bathroom - General" (bathroom floor, whole-bathroom shot)
    - "Floor / Carpet" (room floor, rug, carpet)
    - "Wall / Ceiling" (paint, wallpaper, stains on wall/ceiling)
    - "Window / Curtains" (window glass, drapes, blinds)
    - "Desk / Work Area"
    - "Closet / Wardrobe"
    - "TV / Entertainment Area"
    - "Minibar / Fridge / Kettle"
    - "Air Conditioner / Heater"
    - "Door / Entry"
    - "Balcony (private, part of room)"
    - "Room - General / Overall" (wide shot of the whole room)
    - "Other In-Room"
    - "Not In-Room" (use only if clearly a public area)
  2. The specific problem visible at that in-room location (e.g.
"stained sheets", "hair on pillow", "mouldy grout", "clogged drain",
"cracked tile", "broken remote", "dust on surface", "cockroach /
insect",
"worn / outdated furniture", "damaged wallpaper", "dirty carpet",
"leaking / water damage", "clean / normal"). Keep it short and
concrete.
  3. Whether the visual evidence is consistent with the reviewer's
written
comment.

Respond ONLY with a valid JSON object matching this schema -- no prose
outside JSON:
{
  "room_location": "<one label from the list above>",
  "issue": "<short concrete label, e.g. 'stained sheets', 'mould',
'clean / normal'>",
  "issue_detail": "<one or two sentences describing what you
observe>",
  "review_consistent": <true | false>,
  "confidence": "<high | medium | low>"
}
```

Figure 22: System prompt used for MLLM-based damage pattern analysis of Trip.com hotel review images.

Table 9: **Per-category damage shortlists for Amazon product categories.** Damage types are ordered by empirical frequency in real negative reviews; the first entry in each row is the most commonly observed defect. These shortlists are used as the primary candidate vocabulary during image-specific prompt construction.

Category	Damage Shortlist (in priority order)
All Beauty	shattered/broken, leaking, dented, packaging-damaged, cracked
Amazon Fashion	torn/ripped, stained/dirty, scratched
Appliances	shattered/broken, cracked, dented, scratched, packaging-damaged, bent/warped
Arts, Crafts & Sewing	leaking, stained/dirty, packaging-damaged, bent/warped, torn/ripped, dented, shattered/broken
Automotive	cracked, scratched, stained/dirty, dented, bent/warped, packaging-damaged
Baby Products	shattered/broken, torn/ripped, leaking, packaging-damaged, stained/dirty, mouldy/spoiled, cracked
Beauty & Personal Care	shattered/broken, leaking, packaging-damaged, cracked, stained/dirty, mouldy/spoiled
Books	torn/ripped, stained/dirty, bent/warped, scratched, packaging-damaged
CDs & Vinyl	scratched, cracked, bent/warped, packaging-damaged, stained/dirty, shattered/broken
Cell Phones & Accessories	torn/ripped, scratched, cracked, shattered/broken, packaging-damaged
Clothing, Shoes & Jewelry	torn/ripped, stained/dirty, scratched, shattered/broken
Electronics	shattered/broken, cracked, scratched, bent/warped, packaging-damaged, dented
Grocery & Gourmet Food	melted, mouldy/spoiled, dented, leaking, packaging-damaged, stained/dirty
Handmade Products	stained/dirty, cracked, melted, shattered/broken, scratched, dented, packaging-damaged
Health & Household	leaking, shattered/broken, packaging-damaged, mouldy/spoiled
Health & Personal Care	mouldy/spoiled, packaging-damaged, leaking
Home & Kitchen	cracked, shattered/broken, stained/dirty, dented, scratched, melted, packaging-damaged, bent/warped
Industrial & Scientific	shattered/broken, cracked, leaking, scratched, packaging-damaged, bent/warped
Magazine Subscriptions	torn/ripped, bent/warped, stained/dirty, packaging-damaged
Musical Instruments	shattered/broken, cracked, bent/warped, scratched, dented, packaging-damaged
Office Products	bent/warped, leaking, cracked, dented, packaging-damaged, torn/ripped, scratched, stained/dirty
Patio, Lawn & Garden	shattered/broken, cracked, scratched, packaging-damaged, bent/warped, torn/ripped, mouldy/spoiled, dented
Pet Supplies	torn/ripped, leaking, bent/warped, stained/dirty, mouldy/spoiled, packaging-damaged
Sports & Outdoors	torn/ripped, stained/dirty, cracked, scratched, dented, bent/warped, packaging-damaged
Tools & Home Improvement	cracked, leaking, dented, packaging-damaged, scratched, bent/warped
Toys & Games	packaging-damaged, cracked, shattered/broken, scratched, torn/ripped, stained/dirty
Video Games	cracked, shattered/broken, torn/ripped, scratched, packaging-damaged

Prompt: GrabFood Damage Analysis Prompt

```
You are a senior food-quality auditor specialising in food-delivery
platforms. Your task is to analyse food images from customer reviews
and identify:
  1. The precise food category (e.g. "Pizza", "Fried Chicken",
"Noodles / Ramen", "Burger", "Rice Dish", "Sushi / Japanese",
"Dessert",
"Beverage", "Other").
  2. The quality or condition of the food shown (e.g.
normal/acceptable,
burnt, undercooked, foreign object present, incorrect order, packaging
damaged, portion too small, spillage, mould/spoilage, other defect).
  3. Whether the visual evidence is consistent with the reviewer's
written complaint.

Respond ONLY with a valid JSON object matching this schema -- no prose
outside JSON:
{
  "food_category": "<string>",
  "food_condition": "<string -- one concise label, e.g. 'burnt',
'foreign object', 'normal'>",
  "issue_detail": "<string -- one or two sentences describing what
you observe>",
  "review_consistent": <true | false>,
  "confidence": "<high | medium | low>"
}
```

Figure 23: System prompt used for MLLM-based damage pattern analysis of GrabFood food delivery review images.

Image Generation. The constructed edit prompt, together with the original real-undamaged image, is submitted to all 12 image editing and generation models. To ensure a fair comparison across models, the visual analysis and damage selection in the two-stage pipeline are executed once per image, and the resulting edit instruction is shared across all 12 backends without modification, so that any differences in generated output are attributable to model behavior rather than prompt variation. Each model independently produces a synthetic fake-damaged image by applying the specified modification to the original. The generated image, the selected damage type, the full edit instruction, and the associated review comment are stored together as a structured metadata record for each synthetic sample. The generated review comment, produced as part of Stage 2, serves as the fake review text in review-conditioned evaluation settings.

C.4 Region-Level Annotations

Derivation. Each fake-damaged image in **FraudBench** is an edit of a single real-undamaged image, so the edited area can be recovered from the residual between the two images. Since the generators re-render the whole frame, the raw difference also contains global tone and resampling shifts. We therefore resize both images to a common 1024×1024 canvas, apply a light Gaussian blur, and match the per-channel gain and offset of the edited image to its source before computing the pixel-wise color difference, which is then smoothed with a second Gaussian filter. The difference map is normalized by its median and median absolute deviation, and pixels whose ***robust z-score exceeds 4*** are marked as edited. The resulting mask is cleaned by morphological opening, closing, and hole filling, and connected components covering less than 0.15% of the frame are discarded. The remaining components form the annotated regions, ordered by area. Since the annotations are derived automatically rather than drawn by hand, every record is marked as machine-derived.

Annotation Record. For every region, the annotation provides a bounding box in normalized and pixel coordinates, a polygon traced along the region boundary and simplified with the Ramer–

Douglas–Peucker algorithm, the area fraction, and the peak and mean z-scores as a measure of edit strength. A binary mask stores the pixel-level extent of all regions, and each region references its connected component in this mask. The largest region is marked as the primary damage region and inherits the damage type and target part of the edit metadata, while the remaining regions are marked as secondary evidence, such as spatter, debris, or incidental re-rendering. A `claim_grounding` field links the generated claim to the primary region through the claimed damage type and target part, and the whole annotation is attached as a `localization` block to the entry of the image in the edit metadata. Fig. 24 shows a representative record.

Example: Region-Level Annotation


```
"localization": {
  "annotation_schema": "fraudbench-grounding-v1",
  "generator": "qwen-image-edit-max",
  "image_size": {"width": 1536, "height": 1024},
  "source_image": "Electronics/Positive/Review_013/Image_013_02.jpg",
  "edited_image": "Electronics/DeepFake/qwen-image-edit-max/Review_013
    /Image_013_02.jpg",
  "manipulated_area_fraction": 0.08908,
  "n_regions": 1,
  "edit_is_local": true,
  "regions": [
    {
      "region_id": "r1",
      "role": "primary_damage",
      "damage_type": "cracked",
      "target_part": "matte black plastic headband arm just above the
    hinge mechanism",
      "bbox_normalized": [0.3203, 0.1309, 0.8086, 0.7734],
      "bbox_pixels": [491, 134, 1242, 791],
      "area_fraction": 0.08867,
      "area_pixels": 139473,
      "polygon_normalized": [[0.6616, 0.1309], [0.6685, 0.1309],
    [0.6685, 0.1396], ...],
      "mask_ref": "masks/Electronics__Review_013__Image_013_02__mask.
    png#component=1",
      "residual_strength": {"peak_z": 49.8, "mean_z": 11.8}
    }
  ],
  "claim_grounding": {
    "claim_text": "The headband snapped right at the hinge the first
    time I tried to fold them up, completely useless now.",
    "claimed_damage_type": "cracked",
    "claimed_target_part": "matte black plastic headband arm just
    above the hinge mechanism",
    "primary_region_id": "r1"
  },
  "provenance": {
    "geometry_source": "source-vs-edit residual, MAD-normalised",
    "params": {"canvas": 1024, "mad_k": 4.0, "min_area": 0.0015},
    "human_verified": false
  }
}
```


Figure 24: Structure of the region-level annotation attached to an image entry of the edit metadata. Most polygon vertices are omitted for brevity.

C.5 Quality Control

Quality control is applied after both raw-data preprocessing and synthetic image generation. In both stages, an MLLM serves as a preliminary screener to surface candidate quality issues; the specific

screening criteria and prompts for each stage are described in Appx. C.2 and C.3. Screened issues include weak image–review alignment, privacy-sensitive content, low-quality files, implausible edits, prompt misalignment, and trivial generation artifacts. The MLLM does not assign final labels or determine sample validity; human annotators examine all flagged cases and make the final inclusion decision according to predefined quality-control criteria. Separating automated screening from human adjudication reduces the risk that MLLM-assisted filtering introduces systematic bias into the benchmark.

D BENCHMARK EVALUATION

D.1 EVALUATION SETTINGS

D.1.1 MULTIMODAL LANGUAGE MODELS

Models. We evaluate 30 state-of-the-art MLLMs, including DeepSeek-V4.1-Flash (Xu et al., 2026), Gemini 3.5 Flash-Lite (Google DeepMind, 2026b), Gemini 3 Flash (Google, 2025), Gemini 3.8 Flash (Google DeepMind, 2026c), Gemini 3.1 Pro (Google DeepMind, 2026a), GPT-5.4 mini (OpenAI, 2026a), GPT-5.6 Luna, GPT-5.6 Terra and GPT-5.6 Sol (OpenAI, 2026b), GPT-6 Luna, GPT-6 Sol (OpenAI, 2026d) and GPT-6 Astra (OpenAI, 2026c), Grok 4.1 Fast Reasoning (xAI, 2025), Grok 4.20 Reasoning (xAI, 2026), Grok 4.6 (SpaceXAI, 2026), Kimi K2.6 (Moonshot AI, 2026), Kimi K3 (Team et al., 2026), QVQ-Max-Latest (Qwen Team, 2024), Qwen3-VL-Flash and Qwen3-VL-Plus (Bai et al., 2025), Qwen3.5-Omni-Flash and Qwen3.5-Omni-Plus (Qwen Team, 2026f), Qwen3.6-35B-A3B (Qwen Team, 2026a), Qwen3.6-Plus (Qwen Team, 2026b), Qwen3.7-Flash (Qwen Team, 2026c), Qwen3.7-Plus (Qwen Team, 2026d), Qwen3.7-Max (Qwen Team, 2026c), Qwen3.8-27B (Qwen Team, 2026e), Qwen3.8-Flash (Qiu et al., 2026), and Qwen3.8-Max (Qwen Team, 2026e). This model set covers proprietary and open-weight models, spanning general-purpose multimodal models, vision-language models, omni-modal models, visual-reasoning models, and multimodal Mixture-of-Experts models, and pairs several models with a same-vendor successor so that detector behavior can be compared across model generations as well as across vendors.

Table 10 reports the exact model identifiers and API endpoints used in our experiments. Table 11 further summarizes the evaluated MLLMs and their providers and model categories. All models are accessed through their respective official or authorized API access, and evaluated using provider-default inference settings unless otherwise specified.

Table 10: **Version and access details for evaluated MLLMs.** The table lists the display name used in the paper, the exact model identifier, and the API endpoint used for each evaluated model.

Display Name	Exact Model ID	API Endpoint
Gemini 3.5 Flash-Lite	`gemini-3.5-flash-lite`	`generativelanguage.googleapis.com/v1beta`
Gemini 3 Flash	`gemini-3-flash-preview`	`generativelanguage.googleapis.com/v1beta`
Gemini 3.8 Flash	`gemini-3.8-flash`	`generativelanguage.googleapis.com/v1beta`
Gemini 3.1 Pro	`gemini-3.1-pro-preview`	`generativelanguage.googleapis.com/v1beta`
GPT-5.4 mini	`gpt-5.4-mini`	`api.openai.com/v1`
GPT-5.6 Luna	`gpt-5.6-luna`	`api.openai.com/v1`
GPT-5.6 Terra	`gpt-5.6-terra`	`api.openai.com/v1`
GPT-5.6 Sol	`gpt-5.6-sol`	`api.openai.com/v1`
GPT-6 Luna	`gpt-6-luna`	`api.openai.com/v1`
GPT-6 Sol	`gpt-6-sol`	`api.openai.com/v1`
GPT-6 Astra	`gpt-6-astra`	`api.openai.com/v1`
Grok 4.1 Fast Reasoning	`grok-4-1-fast-reasoning`	`api.x.ai/v1`
Grok 4.20 Reasoning	`grok-4.20-reasoning-latest`	`api.x.ai/v1`
Grok 4.6	`grok-4.6`	`api.x.ai/v1`
QVQ-Max-Latest	`qvq-max-latest`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen3-VL-Flash	`qwen3-vl-flash`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen3-VL-Plus	`qwen3-vl-plus`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen3.5-Omni-Flash	`qwen3.5-omni-flash-2026-03-15`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen3.5-Omni-Plus	`qwen3.5-omni-plus-2026-03-15`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen3.6-Plus	`qwen3.6-plus`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen3.7-Flash	`qwen3.7-flash-2026-07-15`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen3.7-Plus	`qwen3.7-plus-2026-05-26`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen3.7-Max	`qwen3.7-max-2026-06-08`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen3.8-Max	`qwen3.8-max-0902`	`dashscope.aliyuncs.com/compatible-mode/v1`
DeepSeek-V4.1-Flash	`deepseek-v4.1-flash`	`dashscope.aliyuncs.com/compatible-mode/v1`
Kimi K2.6	`kimi-k2.6`	`dashscope.aliyuncs.com/compatible-mode/v1`
Kimi K3	`kimi-k3`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen3.6-35B-A3B	`qwen3.6-flash`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen3.8-27B	`qwen3.8-27b`	`dashscope.aliyuncs.com/compatible-mode/v1`
Qwen3.8-Flash	`qwen3.8-flash`	`dashscope.aliyuncs.com/compatible-mode/v1`

Table 11: **Overview of MLLMs evaluated in FRAUDBENCH.** We summarize the evaluated model set and the input modalities supported by each model. ✓indicates that the model accepts the corresponding input modality, while × indicates that it does not. All models produce text output.

Model	Provider	Text	Image	Audio	Video	Model Type
Proprietary MLLMs						
Gemini 3.5 Flash-Lite	Google	✓	✓	✓	✓	Lightweight multimodal
Gemini 3 Flash	Google	✓	✓	✓	✓	Fast multimodal
Gemini 3.8 Flash	Google	✓	✓	✓	✓	Fast multimodal
Gemini 3.1 Pro	Google	✓	✓	✓	✓	Reasoning multimodal
GPT-5.4 mini	OpenAI	✓	✓	×	×	General-purpose multimodal
GPT-5.6 Luna	OpenAI	✓	✓	×	×	Lightweight multimodal
GPT-5.6 Terra	OpenAI	✓	✓	×	×	General-purpose multimodal
GPT-5.6 Sol	OpenAI	✓	✓	×	×	General-purpose multimodal
GPT-6 Luna	OpenAI	✓	✓	×	×	Lightweight multimodal
GPT-6 Sol	OpenAI	✓	✓	×	×	General-purpose multimodal
GPT-6 Astra	OpenAI	✓	✓	×	×	General-purpose multimodal
Grok 4.1 Fast Reasoning	xAI	✓	✓	×	×	Reasoning multimodal
Grok 4.20 Reasoning	xAI	✓	✓	×	×	Reasoning multimodal
Grok 4.6	xAI	✓	✓	×	×	Reasoning multimodal
QVQ-Max-Latest	DashScope	✓	✓	×	×	Visual reasoning
Qwen3-VL-Flash	DashScope	✓	✓	×	✓	Vision-language
Qwen3-VL-Plus	DashScope	✓	✓	×	✓	Vision-language
Qwen3.5-Omni-Flash	DashScope	✓	✓	✓	✓	Omni-modal
Qwen3.5-Omni-Plus	DashScope	✓	✓	✓	✓	Omni-modal
Qwen3.6-Plus	DashScope	✓	✓	×	✓	General-purpose multimodal
Qwen3.7-Flash	DashScope	✓	✓	×	✓	Fast multimodal
Qwen3.7-Plus	DashScope	✓	✓	×	✓	General-purpose multimodal
Qwen3.7-Max	DashScope	✓	✓	×	✓	General-purpose multimodal
Qwen3.8-Max	DashScope	✓	✓	×	✓	General-purpose multimodal
Open-Weight MLLMs						
DeepSeek-V4.1-Flash	DashScope	✓	✓	×	×	Open-weight multimodal
Kimi K2.6	DashScope	✓	✓	×	×	Open-weight multimodal
Kimi K3	DashScope	✓	✓	×	×	Open-weight multimodal
Qwen3.6-35B-A3B	DashScope	✓	✓	×	✓	Multimodal MoE
Qwen3.8-27B	DashScope	✓	✓	×	✓	Open-weight multimodal
Qwen3.8-Flash	DashScope	✓	✓	×	✓	Fast multimodal

Evaluation Prompt Design. We use a unified prompt template across all MLLM-based evaluation settings, with setting-specific user instructions for single-image, multi-image, and review-augmented inputs. The same review-text template, which refers to a purchase on Amazon, is also used for the Trip.com and GrabFood categories. A shared system prompt defines the model as a forensic image analyst and specifies the expected JSON-only response format, as shown in Fig. 25.

Prompt: Shared System Prompt for MLLM-based Detection

```
You are a forensic image analyst specialising in detecting images
that have been produced or modified by AI -- either through AI
image-editing LLMs or through AI image-generation LLMs. You look for
characteristic traces such as local patch inconsistencies inside an
otherwise plausible scene, object insertion or removal seams,
mismatched shadows or reflections on specific objects, blurred or
resampled regions around modified areas, copy-paste texture
repetition, broken or re-drawn text and logos, anatomical or
geometric implausibilities, over-smoothed textures, and
diffusion-style artefacts. You never refuse the task and you always
return valid JSON.
```

Figure 25: Shared system prompt used to establish the forensic-analysis role for MLLM-based AI-generated evidence detection.

For single-image evaluation, each image is submitted independently using the prompt shown in Fig. 26.

Prompt: Single-Image Detection Prompt

```
Analyse the provided image and decide whether it has been produced or
modified by AI -- either through AI image-editing LLMs or through AI
image-generation LLMs -- or whether it is an unmodified genuine
camera photograph of a real scene.

Return ONLY a single JSON object -- no markdown fences, no prose
outside the object -- with this exact schema:
{
  "is_ai_modified": <true or false>,
  "confidence": <number between 0 and 1>,
  "reason": "<1-3 concise sentences naming the specific visual
evidence you relied on>"
}
```

Figure 26: Single-image prompt used for review-free MLLM-based detection.

When review text is available, the review-text instruction is inserted between the image-analysis instruction and the JSON schema.

Prompt: Review-Text Augmentation Prompt

```
The customer who purchased this product on Amazon left the following
review:
"{review text}"
Consider this customer review when making your assessment.
```

Figure 27: Review-text augmentation inserted between the image-analysis instruction and the JSON output schema when review text is available.

For multi-image evaluation, we consider two settings. The single-turn setting packs all images from the same review into one user message, using the prompt shown in Fig. 28. The multi-step setting holds a multi-turn conversation in which the images of a review are submitted one per turn and the model's replies are kept in the conversation history. The shared system prompt is extended with the sentence in Fig. 29. For each non-final image, the model gives a short forensic observation without a verdict (Figs. 30 and 31), and the final turn requests the consolidated verdict (Fig. 32). In review-conditioned runs, the review-text block of Fig. 27 is appended to the turn of the first image it refers to, and each distinct review text is shown only once. Inline comments (`//`) present in the JSON schema of the actual prompts are omitted from the figures for brevity.

Prompt: Multi-Image Detection Prompt

```
The {n} images above all come from the same product listing. Examine
all {n} images together and decide whether they have been produced or
modified by AI -- either through AI image-editing LLMs or through AI
image-generation LLMs -- or whether they are unmodified genuine
camera photographs of a real product.

Return ONLY a single JSON object -- no markdown fences, no prose
outside the object -- with this exact schema:
{
  "is_ai_modified": <true or false>,
  "confidence": <number between 0 and 1>,
  "reason": "<1-3 sentences explaining the combined verdict over all
{n} images>",
  "per_image": [
    {
      "index": <1-indexed position, 1..{n}>,
      "is_ai_modified": <true or false>,
      "confidence": <number between 0 and 1>,
      "notes": "<1-2 sentences on visual evidence from THIS image>"
    },
    ...
  ]
}
```

Figure 28: Multi-image prompt used when all images from the same review are submitted in one user message.

Prompt: Multi-Step System Prompt Extension

```
[Shared system prompt] You are reviewing a customer refund claim in
which the customer uploads their evidence photos one message at a
time. Answer each message as it arrives, and carry your earlier
observations forward.
```

Figure 29: Sentence appended to the shared system prompt in the multi-step setting.

Specifically, for Alibaba DashScope-hosted models, including the Qwen and QVQ models, the system prompt is prepended to the user turn as plain text rather than passed as a separate system-role message, following provider-specific usage constraints. All model responses are parsed against the required JSON schema. Responses that cannot be parsed against the schema are recorded as failed and re-queried under the same prompt template in a later pass, until every reported record has a valid response.

Prompt: Multi-Step First-Turn Prompt

```
A customer has opened a refund claim for this product and is
uploading their evidence photos one at a time. This is image {i} of
{n}. Give a short forensic observation (2-3 sentences) of anything
relevant to whether this image was produced or modified by AI --
either through AI image-editing LLMs or AI image-generation LLMs. Do
not commit to a final verdict yet; more images from this claim are
coming.
```

Figure 30: Prompt sent with the first image of a review in the multi-step setting.

Prompt: Multi-Step Intermediate-Turn Prompt

```
Here is image {i} of {n} from the same refund claim. Taking your
earlier observations into account, give a short forensic observation
(2-3 sentences) of this image and say whether it changes your reading
of the set so far. Do not commit to a final verdict yet; more images
from this claim are coming.
```

Figure 31: Prompt sent with each subsequent non-final image in the multi-step setting.

Prompt: Multi-Step Final-Turn Prompt

```
Here is image {n} of {n} -- the final image from this refund claim.
Considering this image together with every image you have seen and
every observation you made earlier in this conversation, decide
whether these {n} images have been produced or modified by AI --
either through AI image-editing LLMs or through AI image-generation
LLMs -- or whether they are unmodified genuine camera photographs of
real scenes.

Return ONLY a single JSON object -- no markdown fences, no prose
outside the object -- with this exact schema:
{
  "is_ai_modified": <true or false>,
  "confidence": <number between 0 and 1>,
  "reason": "<1-3 sentences explaining the combined verdict over all
{n} images>",
  "per_image": [
    {
      "index": <1-indexed position, 1..{n}>,
      "is_ai_modified": <true or false>,
      "confidence": <number between 0 and 1>,
      "notes": "<1-2 sentences on visual evidence from THIS image>"
    },
    ...
  ]
}
```

Figure 32: Prompt sent with the final image to obtain the consolidated verdict in the multi-step setting.

Other Details. For some providers, multi-image and multi-step requests exceeded the request size limit, because high-resolution images inflate once base64-encoded. We handle these cases using a ***deterministic image downscaling protocol*** rather than treating them as immediate failures. A request is downscaled when the provider rejects it as too large (HTTP 413, or HTTP 400 with a size-related message), when the connection of a request carrying at least 10 MB of image data is dropped twice, or when an Alibaba DashScope endpoint reports a transient failure to fetch the images (HTTP 400). Each such event re-encodes the images at the next of ten fixed tiers, progressively reducing the long-edge pixel limit and JPEG quality in tandem, from 2,560 px/Q95 down to 512 px/Q72. Other transient errors, such as rate limits and network failures, are retried with exponential back-off, so each request is attempted at most 13 times: three back-off retries plus ten downscale steps. Re-encoding is performed with Pillow using aspect-ratio-preserving resizing, and the resulting image is serialized as a base64 data URI before resubmission.

D.1.2 Specialized Detectors

Models. In addition to MLLM-based detectors, we evaluate four specialized AI-generated image detectors: CO-SPY (Cheng et al., 2025), ForgeLens (Chen et al., 2025), Effort (Yan et al., 2025b), and IAPL (Li et al., 2026). For each method, we use the authors' official open-source implementation, linked as `https://github.com/Megum1/CO-SPY`, `https://github.com/Yingjian-Chen/ForgeLens`, `https://github.com/YZY-stack/Effort-AIGI-Detection`, and `https://github.com/liyih/IAPL`. We evaluate the released pretrained checkpoints provided by the corresponding repositories, including checkpoints trained on standard AI-generated image detection datasets such as GenImage, UniversalFakeDetect, Chameleon, and ProGAN.

As these specialized detectors operate on individual images rather than interactive multimodal inputs, we report their results only in the single-image, review-free setting, excluding multi-image context, review text, multi-step interaction, and prompt-based reasoning instructions.

Testbed. Experiments are conducted on a single Elastic Compute Service (ECS) instance. We employ an instance equipped with an Intel Xeon Platinum 8369B CPU with 16 available vCPUs, 125 GB RAM, 2048 GB of available disk space, and one NVIDIA A100 SXM GPU with 80 GB memory. This instance type is used for all specialized-detector evaluations to provide a consistent hardware environment across methods. For reproducibility and consistent performance, we recommend the `ecs.gn7e-c16g1.4xlarge` instance types on Alibaba Cloud.

We performed all experiments on Ubuntu 24.04.4 LTS. For the specialized detectors, we follow the official repository instructions and create a separate Conda environment for each method to configure the required Python, PyTorch, CUDA, and dependency versions. We therefore do not enforce a single shared software environment across repositories; instead, each detector is evaluated under the environment recommended by its released implementation. This setup avoids package-version conflicts while minimizing compatibility changes to the original codebases.

Image Forgery Detection and Localization Methods. Since every fake-damaged image in **FraudBench** is a local edit of a genuine photograph, we further evaluate image forgery detection and localization (IFDL) methods, which are designed to detect and localize spliced, copy-moved, and inpainted regions. We evaluate MVSS-Net (Dong et al., 2022) and the ResFCN baseline released in the same repository, each with its checkpoints trained on CASIA and DEFACTO, together with TruFor (Guillaro et al., 2023) and FakeShield (Xu et al., 2025). For each method, we use the authors' official open-source implementation and released weights, linked as `https://github.com/dong03/MVSS-Net`, `https://github.com/grip-unina/TruFor`, and `https://github.com/zhipeixu/FakeShield`. Like the AI-generated image detectors above, IFDL methods operate on individual images, so we report their results only in the single-image, review-free setting on the same 6,940 images.

As for the AI-generated image detectors, we run the official inference scripts with ***released checkpoints and default parameters***, without fine-tuning or threshold calibration on our benchmark. For MVSS-Net and ResFCN, the image-level score is the maximum of the predicted manipulation map, following the official evaluation script, and for TruFor it is the image-level detection score. Both

scores are thresholded at 0.5. FakeShield returns a textual verdict together with a manipulation mask for images it judges tampered, and we use its verdict directly.

Hyperparameters. We run the official inference or evaluation scripts with ***original configuration files and default parameters***, without additional hyperparameter tuning or threshold calibration on our benchmark, to ensure that the comparison reflects each detector's standard off-the-shelf performance under the authors' recommended evaluation settings.

D.1.3 HUMAN EVALUATIONS

Evaluators. The human evaluation was conducted by three evaluators. The evaluators were not involved in curating the real evidence, generating the synthetic images, constructing the generation prompts, or selecting the generated samples used in the human evaluation subset. Their educational backgrounds include law, computer science, and engineering, with one evaluator holding a bachelor's degree in law and the other two holding undergraduate or master's degrees in computer science.

Task Design. Evaluators were asked to judge whether a product-review evidence image was AI-generated or manipulated. Each evaluation instance was presented through our self-built web-based evaluation interface, whose source code is available at `https://anonymous.4open.science/r/FraudBench-ICLR`. The interface displayed a single image only, with no review text shown, even when such text was available in the original review. To maintain a blinded evaluation, the interface did not display the ground-truth label, generator identity, model predictions, source metadata, or prompt information. Evaluators provided a binary decision for each image.

Quality Control and Aggregation. Each image was independently labeled by three evaluators. We report human accuracy as the average of the three individual evaluator accuracies. Disagreements were not adjudicated, since the aggregate human baseline is computed from individual evaluator performance rather than from a consensus label.

Ethics. No external participants were recruited, and no compensation was involved. The evaluation required only binary judgments of product-review evidence images. It did not involve collecting personal information, demographic attributes, behavioral data, or sensitive information from the evaluators. Following our institutional policy, we determined that IRB approval was not required.

D.1.4 ABLATION STUDIES

Prompt Sensitivity. We compare five prompt variants with progressively reduced task-specific guidance. The variants differ in the system-role framing, whether the forensic artifact checklist is retained, and whether the role instruction is separated from the task instruction. **Baseline** uses the full forensic prompt with an explicit artifact checklist. **Merged** uses the same content as Baseline but merges the system and user instructions into a single user message. **No Checklist** retains the forensic role while removing the artifact checklist. **Generic** replaces the forensic expert role with a generic assistant identity. **Minimal** uses only a one-sentence instruction with no system message.

All variants require a structured JSON response containing `is_ai_modified`, `confidence`, and `reason`. This keeps the output format comparable across variants and isolates the effect of prompt framing and task-specific guidance. Experiments are conducted in the single-image, review-free setting on ***a stratified 10% random sample*** of the Amazon subset, covering all 27 product categories and excluding the Trip.com and GrabFood service-review domains, using a fixed random seed of `42`.

Mismatched Review. The standard review-conditioned setting pairs each image with the customer review associated with the same product or service instance, allowing models to use both visual forensic cues and text-image consistency. To isolate the contribution of the review text, we introduce a mismatched-review condition on ***a stratified 10% random sample*** of the Amazon subset, covering all 27 product categories and excluding the Trip.com and GrabFood service-review domains, with a fixed random seed of `42`. For each image in this sample, we replace the matched review with a review sampled uniformly at random from a different product category.

Prompt: Baseline Prompt

```
System:

You are a forensic image analyst specialising in detecting images
that have been produced or modified by AI -- either through AI
image-editing LLMs or through AI image-generation LLMs. You look for
characteristic traces such as local patch inconsistencies inside an
otherwise plausible scene, object insertion or removal seams,
mismatched shadows or reflections on specific objects, blurred or
resampled regions around modified areas, copy-paste texture
repetition, broken or re-drawn text and logos, anatomical or
geometric implausibilities, over-smoothed textures, and
diffusion-style artefacts. You never refuse the task and you always
return valid JSON.

User:

Analyse the provided image and decide whether it has been produced or
modified by AI -- either through AI image-editing LLMs or through AI
image-generation LLMs -- or whether it is an unmodified genuine
camera photograph of a real scene.

Return ONLY a single JSON object -- no markdown fences, no prose
outside the object -- with this exact schema:
{
  "is_ai_modified": <true or false>,
  "confidence": <number between 0 and 1>,
  "reason": "<1-3 concise sentences naming the specific visual
evidence you relied on>"
}
```

Figure 33: Baseline prompt used in the prompt-sensitivity study.

Prompt: Merged Prompt

```
System:

None

User:

You are a forensic image analyst specialising in detecting images
that have been produced or modified by AI -- either through AI
image-editing LLMs or through AI image-generation LLMs. You look for
characteristic traces such as local patch inconsistencies inside an
otherwise plausible scene, object insertion or removal seams,
mismatched shadows or reflections on specific objects, blurred or
resampled regions around modified areas, copy-paste texture
repetition, broken or re-drawn text and logos, anatomical or
geometric implausibilities, over-smoothed textures, and
diffusion-style artefacts. You never refuse the task and you always
return valid JSON.

Analyse the provided image and decide whether it has been produced or
modified by AI -- either through AI image-editing LLMs or through AI
image-generation LLMs -- or whether it is an unmodified genuine
camera photograph of a real scene.

Return ONLY a single JSON object -- no markdown fences, no prose
outside the object -- with this exact schema:
{
  "is_ai_modified": <true or false>,
  "confidence": <number between 0 and 1>,
  "reason": "<1-3 concise sentences naming the specific visual
evidence you relied on>"
}
```

Figure 34: Merged prompt used in the prompt-sensitivity study. The system and user instructions are combined into a single user turn.

Prompt: No Checklist Prompt

```
System:

You are a forensic image analyst specialising in detecting images
that have been produced or modified by AI -- either through AI
image-editing LLMs or through AI image-generation LLMs. You never
refuse the task and you always return valid JSON.

User:

Analyse the provided image and decide whether it has been produced or
modified by AI -- either through AI image-editing LLMs or through AI
image-generation LLMs -- or whether it is an unmodified genuine
camera photograph of a real scene.

Return ONLY a single JSON object -- no markdown fences, no prose
outside the object -- with this exact schema:
{
  "is_ai_modified": <true or false>,
  "confidence": <number between 0 and 1>,
  "reason": "<1-3 concise sentences naming the specific visual
evidence you relied on>"
}
```

Figure 35: No Checklist prompt used in the prompt-sensitivity study.

Prompt: Generic Prompt

```
System:

You are a helpful AI assistant. You always respond with valid JSON
when asked to analyse an image.

User:

Analyse the provided image and decide whether it has been produced or
modified by AI -- either through AI image-editing LLMs or through AI
image-generation LLMs -- or whether it is an unmodified genuine
camera photograph of a real scene.

Return ONLY a single JSON object -- no markdown fences, no prose
outside the object -- with this exact schema:
{
  "is_ai_modified": <true or false>,
  "confidence": <number between 0 and 1>,
  "reason": "<1-3 concise sentences naming the specific visual
evidence you relied on>"
}
```

Figure 36: Generic prompt used in the prompt-sensitivity study.

Prompt: Minimal Prompt

```
System:

None

User:

Is this image a genuine photograph or was it produced / modified by
an AI model? Return JSON only -- no markdown, no extra text:
{"is_ai_modified": <true or false>, "confidence": <0.0-1.0>,
"reason": "<1-2 sentences>"}
```

Figure 37: Minimal prompt used in the prompt-sensitivity study.

For example, an AI-manipulated electronics image may be paired with a review originally written for a beauty product. The mismatched review is inserted at the same position in the prompt as in the standard review-conditioned setting, while the image input and all other prompt components are held fixed. Comparing the matched and mismatched conditions on the same image set, therefore, provides a controlled diagnostic of whether models rely on semantic coherence between the review text and visual content, rather than on visual forensic cues alone.

D.1.5 Localization Evaluation

Ground Truth. To score all methods on the same images, we use a ground truth that assigns a manipulated area to every fake-damaged image. After Gaussian smoothing of the absolute difference between a fake-damaged image and its real-undamaged source, we take the ***smallest set of pixels that carries half of the total difference*** as the manipulated area. This area covers 17.8% of the frame on average, more than the annotated regions in Appx. C.4, since it also absorbs part of the diffuse residual left by re-rendering. To examine the effect of this choice, we additionally vary the fraction of the difference from 30% to 70%.

Models. We compare methods that return a region. For Gemini 3 Flash, we issue a separate localization query for every image of the six evaluated generator subsets and the real-damaged subset, using the prompt in Fig. 38, which asks for a verdict, a confidence score, the tightest bounding box around the AI-modified content, and a short rationale. FakeShield returns a manipulation mask for every image it judges tampered, and MVSS-Net and ResFCN return a manipulation map, which we threshold at 0.5 (Appx. D.1.2). TruFor retains only summary statistics of its localization map in our runs, so we do not score its localization. The localization metrics are defined in Appx. D.2.

D.2 Evaluation Metrics

Binary Verification. We formulate the task as binary verification. Fake-damaged images are treated as the positive class, while genuine, real-damaged images are treated as the negative class. For each evaluated model, the predicted label is obtained from the returned binary field `is_ai_modified`. The model-reported confidence score is used only for reporting confidence statistics and is not used to threshold or calibrate the binary prediction.

Let TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively. In our setting, TP denotes a fake-damaged image correctly predicted as AI-generated or AI-modified, while TN denotes a genuine, real-damaged image correctly predicted as real. We define the base metrics as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{1}$$

$$\text{TPR} = \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2}$$

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{3}$$

Prompt: Localization Prompt

```
System:

You are a forensic image analyst. You localise regions of an image
that have been produced or modified by AI image-editing or
image-generation models. You never refuse the task and you always
return valid JSON.

User:

Decide whether this image has been modified by AI, and if so, where.

Return ONLY a single JSON object with this exact schema:
{
  "is_ai_modified": <true or false>,
  "confidence": <number between 0 and 1>,
  "box_2d": [ymin, xmin, ymax, xmax],
  "reason": "<1-2 sentences naming the visual evidence>"
}

"box_2d" uses the standardS bounding-box convention: integers in
[0,1000] normalised to the image size, ordered [ymin, xmin, ymax,
xmax], with the origin at the top-left corner. Give the single
tightest box containing the AI-modified content. If the image is not
AI-modified, set "box_2d" to null.
```

Figure 38: Prompt used to query Gemini 3 Flash for the location of AI-modified content.

Subset-Level Metrics. For each generator subset, we report TPR to measure the fraction of fake-damaged images correctly detected as AI-generated or AI-modified. The shared real-damaged subset is used as the negative reference set when computing metrics that require negative examples. On the real-damaged subset, we report TNR to measure the fraction of authentic damaged evidence correctly retained as genuine rather than falsely rejected as synthetic.

The reported **Conf.** value is the mean model-reported confidence over correctly classified images only. Thus, for generated subsets, **Conf.** is averaged over fake-damaged images that are correctly detected as AI-generated or AI-modified; for the real-damaged subset, it is averaged over genuine damaged images that are correctly classified as real.

Overall Aggregation. For overall comparison, we report Balanced Accuracy (Bal.Acc.) and F1. Both are macro-averaged over product categories: for each category, Bal.Acc. is the average of the category-level TPR, computed over all six generator subsets, and TNR, and F1 is the F1 score of the fake-damaged class. The category-level values are then averaged across categories, which prevents larger categories from dominating the overall scores.

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{TPR}}{\mathrm{Precision} + \mathrm{TPR}}, \tag{4}$$

$$\text{Balanced Accuracy} = \frac{\mathrm{TPR} + \mathrm{TNR}}{2}. \tag{5}$$

Threshold-Free Metrics. For the specialized detectors, which output a continuous score, we additionally report threshold-free metrics in Appx. D.4.5. AUROC measures how well the score ranks fake-damaged images above real-damaged images, and the partial AUROC for FPR ≤ 0.1 is standardized so that 0.5 corresponds to chance. TPR@$k\%$ denotes the TPR at the threshold whose false-positive rate on real-damaged evidence is $k\%$, which is equivalent to fixing the TNR at $1-k\%$. Unlike the other metrics, these are computed on the scores pooled over all categories and generators.

Localization Metrics. For the localization evaluation in Appx. D.1.5, we rasterize every predicted region and the ground-truth manipulated area onto a common canvas, filling a bounding box as a rectangle. Let P and G denote the predicted and ground-truth pixel sets of a fake-damaged image. We compute

$$\mathrm{IoU} = \frac{|P \cap G|}{|P \cup G|}, \qquad \mathrm{F1}_{\mathrm{pix}} = \frac{2\,|P \cap G|}{|P| + |G|}. \tag{6}$$

FakeShield is scored on the images for which it returns a mask, and Gemini 3 Flash on those of them that it also flags with a box, whereas the thresholded maps of MVSS-Net and ResFCN are scored on every fake-damaged image. We report the mean and median IoU, the mean pixel F1, and the mean fraction of the frame covered by the prediction. On real-damaged evidence, where nothing was edited, any returned region counts as a false localization.

D.3 Supplementary Evaluation Results

Tab. 12 and 13 report the single-image results of all MLLMs evaluated without and with review text, Tab. 14–19 the multi-image and multi-step results, Tab. 16 and 17 the multi-image results of all 18 MLLMs evaluated on the 27 Amazon categories, Tab. 20–48 the per-category results of all 30 MLLMs and the eight detector checkpoints, and Tab. 49 the human evaluation. Tables of the prompt-sensitivity and mismatch studies list the MLLMs of the corresponding study first and additional MLLMs evaluated with the same protocol in a separate block.

Before presenting the supplementary results, we clarify the metric aggregation and missing-value conventions used in the following tables. Unless otherwise specified, metrics are macro-averaged over product categories rather than computed from pooled predictions. For each category, generator-specific TPR is computed on the corresponding AI-generated subset, real-image TNR is computed on the shared real-damaged subset, and category-level F1 is computed for the fake-damaged class. The reported F1 is then averaged across categories.

TNR measures the fraction of genuinely damaged images correctly classified as unmodified, whereas TPR measures the fraction of AI-generated or AI-modified images correctly flagged as manipulated. F1 is the harmonic mean of precision and recall, Balanced Accuracy is the arithmetic mean of TPR and TNR, and Conf. denotes the mean model-reported confidence over correctly classified samples.

A cell is left blank, denoted by "–", when ***the corresponding metric is undefined for a model-category combination***. This occurs when a sampled category contains no images from a particular generator or when all API calls for that model and category fail. In practice, such blanks occur almost exclusively in generator-specific TPR columns for sparse 10% ablation or mismatch-review subsets and do not affect the overall F1 or Balanced Accuracy columns, which aggregate valid predictions across generators within each category.

Table 12: **Performance comparison of MLLMs on FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context for all 30 evaluated MLLMs.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.084	0.693	0.171	0.701	0.213	0.690	0.255	0.700	0.357	0.710	0.423	0.721	0.601	0.391	0.952	0.764
Gemini 3 Flash	0.158	0.961	0.199	0.954	0.258	0.958	0.398	0.959	0.491	0.957	0.692	0.964	**0.674**	0.526	0.982	0.968
Gemini 3.1 Pro	0.091	0.714	0.122	0.712	0.167	0.712	0.260	0.719	0.327	0.715	0.519	0.717	0.619	0.391	0.990	0.704
Gemini 3.5 Flash-Lite	0.050	0.965	0.066	0.958	0.085	0.964	0.119	0.968	0.178	0.965	0.244	0.968	0.560	0.212	0.996	0.950
Gemini 3.8 Flash	0.025	0.703	0.013	0.708	0.066	0.703	0.102	0.708	0.157	0.707	0.199	0.710	0.547	0.165	**1.000**	0.658
GPT-5.4 mini	0.040	0.913	0.045	0.910	0.083	0.936	0.129	0.936	0.147	0.942	0.285	0.950	0.558	0.212	0.994	0.918
GPT-5.6 Luna	0.017	0.851	0.036	0.860	0.065	0.855	0.096	0.893	0.139	0.874	0.333	0.881	0.556	0.200	0.997	0.969
GPT-5.6 Sol	0.018	0.825	0.034	0.842	0.067	0.864	0.132	0.916	0.168	0.888	0.271	0.917	0.557	0.199	0.999	0.976
GPT-5.6 Terra	0.083	0.782	0.100	0.806	0.158	0.795	0.396	0.867	0.331	0.831	0.516	0.855	0.626	0.409	0.989	0.933
GPT-6 Astra	0.108	0.783	0.107	0.795	0.212	0.815	0.400	0.837	0.453	0.830	0.530	0.840	0.650	0.450	0.998	0.930
GPT-6 Luna	**0.761**	0.896	**0.910**	0.919	**0.934**	0.924	**0.928**	0.927	**0.845**	0.902	**0.951**	0.934	0.547	**0.873**	0.206	0.763
GPT-6 Sol	0.089	0.771	0.126	0.769	0.202	0.794	0.312	0.838	0.322	0.815	0.465	0.838	0.623	0.390	0.993	0.913
Grok 4.1 Fast Reasoning	0.105	0.915	0.135	0.904	0.144	0.916	0.188	0.910	0.218	0.907	0.262	0.913	0.538	0.279	0.901	0.947
Grok 4.20 Reasoning	0.062	0.748	0.065	0.728	0.141	0.705	0.199	0.739	0.161	0.716	0.242	0.705	0.562	0.244	0.979	0.764
Grok 4.6	0.050	0.829	0.047	0.806	0.131	0.823	0.173	0.835	0.198	0.818	0.250	0.814	0.568	0.237	0.995	0.843
Kimi K2.6	0.026	0.906	0.052	0.905	0.079	0.896	0.218	0.883	0.260	0.899	0.356	0.903	0.581	0.277	0.996	0.908
Kimi K3	0.015	0.848	0.043	0.718	0.029	0.711	0.127	0.808	0.152	0.742	0.276	0.773	0.541	0.189	0.975	0.866
QVQ-Max-Latest	0.220	0.743	0.239	0.744	0.294	0.741	0.321	0.752	0.335	0.753	0.383	0.759	0.510	0.438	0.721	0.928
Qwen3-VL-Flash	0.043	0.936	0.090	0.938	0.172	0.938	0.247	0.939	0.278	0.936	0.364	0.938	0.592	0.321	0.986	0.973
Qwen3-VL-Plus	0.002	0.920	0.006	0.932	0.024	0.942	0.050	0.945	0.049	0.952	0.108	0.947	0.518	0.075	0.997	0.977
Qwen3.5-Omni-Flash	0.013	0.911	0.050	0.910	0.083	0.910	0.147	0.910	0.152	0.911	0.185	0.911	0.551	0.183	0.996	0.952
Qwen3.5-Omni-Plus	0.015	0.928	0.026	0.928	0.055	0.928	0.109	0.931	0.121	0.933	0.180	0.936	0.542	0.151	**1.000**	0.971
Qwen3.6-35B-A3B	0.144	0.922	0.242	0.917	0.262	0.919	0.365	0.921	0.407	0.919	0.531	0.921	0.603	0.476	0.882	0.965
Qwen3.6-Plus	0.077	0.927	0.140	0.921	0.179	0.922	0.266	0.927	0.325	0.926	0.441	0.928	0.608	0.371	0.977	0.953
Qwen3.7-Flash	0.065	0.943	0.105	0.938	0.182	0.936	0.290	0.949	0.275	0.942	0.475	0.947	0.608	0.368	0.985	0.959
Qwen3.7-Max	0.034	0.900	0.063	0.888	0.094	0.881	0.192	0.899	0.231	0.892	0.313	0.901	0.575	0.260	0.995	0.948
Qwen3.7-Plus	0.027	0.900	0.059	0.881	0.091	0.896	0.189	0.897	0.229	0.885	0.321	0.904	0.572	0.259	0.992	0.948
Qwen3.8-27B	0.104	0.821	0.090	0.820	0.163	0.828	0.218	0.834	0.207	0.815	0.320	0.839	0.583	0.305	0.982	0.870
Qwen3.8-Flash	0.095	0.896	0.253	0.905	0.295	0.904	0.412	0.891	0.390	0.903	0.561	0.904	0.641	0.490	0.947	0.940
Qwen3.8-Max	0.112	0.789	0.180	0.806	0.234	0.811	0.414	0.828	0.452	0.829	0.619	0.850	0.664	0.490	0.993	0.864

Table 13: **Performance comparison of MLLMs on FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting with review context for the 18 MLLMs evaluated in this setting.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.261	0.727	0.334	0.689	0.397	0.689	0.454	0.724	0.519	0.720	0.612	0.721	0.658	0.588	0.887	0.782
Gemini 3 Flash	0.211	0.950	0.281	0.936	0.370	0.946	0.488	0.957	0.576	0.954	0.776	0.963	0.720	0.614	0.990	0.975
Gemini 3.8 Flash	0.119	0.697	0.106	0.703	0.201	0.703	0.274	0.710	0.349	0.710	0.467	0.712	0.626	0.385	1.000	0.673
GPT-5.4 mini	0.011	0.953	0.016	0.934	0.026	0.951	0.040	0.948	0.043	0.951	0.112	0.961	0.520	0.078	0.999	0.935
GPT-5.6 Terra	0.069	0.780	0.086	0.801	0.119	0.807	0.295	0.870	0.275	0.843	0.458	0.864	0.608	0.346	0.999	0.964
GPT-6 Astra	0.303	0.777	0.318	0.784	0.418	0.806	0.661	0.832	0.706	0.838	0.824	0.848	0.769	0.692	0.999	0.939
Grok 4.1 Fast Reasoning	0.171	0.906	0.192	0.902	0.216	0.901	0.248	0.910	0.233	0.914	0.321	0.906	0.554	0.349	0.878	0.946
Grok 4.20 Reasoning	0.067	0.710	0.085	0.695	0.126	0.694	0.189	0.711	0.162	0.710	0.279	0.705	0.568	0.256	0.985	0.756
Grok 4.6	0.037	0.854	0.041	0.821	0.110	0.828	0.189	0.842	0.200	0.824	0.273	0.828	0.570	0.241	0.998	0.865
Kimi K2.6	0.084	0.881	0.111	0.885	0.109	0.880	0.156	0.890	0.224	0.889	0.345	0.894	0.583	0.287	0.994	0.705
Kimi K3	0.036	0.714	0.082	0.713	0.076	0.719	0.202	0.766	0.215	0.739	0.357	0.762	0.581	0.271	1.000	0.891
QVQ-Max-Latest	0.202	0.776	0.185	0.800	0.242	0.791	0.245	0.794	0.227	0.793	0.295	0.798	0.538	0.363	0.844	0.931
Qwen3-VL-Flash	0.226	0.943	0.253	0.946	0.342	0.944	0.350	0.943	0.429	0.942	0.527	0.944	0.666	0.511	0.977	0.968
Qwen3-VL-Plus	0.000	—	0.002	0.935	0.004	0.943	0.013	0.953	0.012	0.953	0.015	0.947	0.503	0.015	0.999	0.980
Qwen3.6-35B-A3B	0.074	0.926	0.151	0.924	0.168	0.928	0.272	0.928	0.270	0.922	0.406	0.926	0.585	0.356	0.946	0.968
Qwen3.6-Plus	0.058	0.928	0.085	0.919	0.100	0.915	0.205	0.928	0.207	0.924	0.365	0.931	0.577	0.282	0.985	0.953
Qwen3.8-Flash	0.068	0.833	0.144	0.872	0.184	0.859	0.315	0.853	0.303	0.866	0.452	0.882	0.615	0.383	0.987	0.925
Qwen3.8-Max	0.115	0.781	0.198	0.798	0.257	0.811	0.414	0.831	0.452	0.827	0.638	0.848	0.672	0.502	0.999	0.888

Table 14: **Performance comparison of MLLMs on FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the multi-image setting without review context for the eight MLLMs evaluated on all 29 categories in both multi-image settings.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.067	0.680	0.216	0.695	0.271	0.721	0.279	0.716	0.429	0.739	0.621	0.749	0.637	0.463	0.960	0.808
Gemini 3.8 Flash	0.050	0.717	0.065	0.714	0.153	0.718	0.268	0.718	0.367	0.721	0.467	0.721	0.612	0.353	1.000	0.685
GPT-5.6 Terra	0.048	0.867	0.103	0.886	0.131	0.907	0.560	0.942	0.361	0.929	0.587	0.946	0.649	0.446	1.000	0.954
GPT-6 Astra	0.279	0.879	0.294	0.878	0.418	0.902	0.684	0.909	0.692	0.921	0.833	0.927	0.773	0.679	1.000	0.959
Grok 4.6	0.063	0.843	0.025	0.828	0.102	0.827	0.221	0.849	0.195	0.838	0.311	0.841	0.570	0.253	1.000	0.874
Kimi K3	0.012	0.720	0.015	0.700	0.042	0.727	0.175	0.776	0.185	0.766	0.334	0.786	0.563	0.217	1.000	0.888
Qwen3.8-Flash	0.121	0.854	0.215	0.865	0.327	0.828	0.462	0.843	0.451	0.840	0.682	0.858	0.671	0.532	0.964	0.921
Qwen3.8-Max	0.099	0.822	0.180	0.820	0.255	0.820	0.505	0.851	0.510	0.861	0.727	0.895	0.694	0.531	1.000	0.902

Table 15: **Performance comparison of MLLMs on FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the multi-image setting with review context for the eight MLLMs evaluated on all 29 categories in both multi-image settings.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.327	0.707	0.364	0.735	0.438	0.727	0.572	0.739	0.633	0.759	0.789	0.759	0.748	0.671	0.980	0.803
Gemini 3.8 Flash	0.321	0.719	0.278	0.717	0.411	0.720	0.550	0.715	0.561	0.718	0.782	0.721	0.742	0.633	1.000	0.687
GPT-5.6 Terra	0.152	0.899	0.170	0.897	0.276	0.902	0.591	0.944	0.434	0.929	0.699	0.945	0.688	0.545	1.000	0.967
GPT-6 Astra	0.746	0.858	0.768	0.862	0.824	0.881	0.917	0.909	0.954	0.921	0.983	0.938	0.930	0.925	1.000	0.957
Grok 4.6	0.068	0.851	0.037	0.814	0.124	0.841	0.294	0.855	0.232	0.862	0.366	0.849	0.591	0.303	1.000	0.898
Kimi K3	0.094	0.746	0.154	0.731	0.222	0.778	0.372	0.769	0.361	0.780	0.582	0.789	0.648	0.444	1.000	0.900
Qwen3.8-Flash	0.127	0.844	0.225	0.837	0.304	0.844	0.475	0.865	0.517	0.861	0.597	0.881	0.680	0.531	0.985	0.922
Qwen3.8-Max	0.108	0.824	0.180	0.817	0.238	0.828	0.472	0.853	0.521	0.861	0.709	0.890	0.689	0.523	1.000	0.915

Table 16: **Performance comparison of MLLMs on the 27 Amazon categories of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the multi-image setting without review context, macro-averaged over the 27 Amazon categories, for the 18 MLLMs evaluated on these categories in both multi-image settings. The first block repeats the MLLMs of Tab. 14 restricted to the same categories.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.067	0.681	0.215	0.695	0.278	0.723	0.281	0.709	0.423	0.740	0.621	0.751	0.636	0.464	0.958	0.808
Gemini 3.8 Flash	0.049	0.720	0.054	0.713	0.144	0.718	0.260	0.718	0.359	0.722	0.461	0.721	0.608	0.347	1.000	0.685
GPT-5.6 Terra	0.049	0.856	0.093	0.895	0.129	0.908	0.554	0.946	0.345	0.931	0.580	0.948	0.645	0.440	1.000	0.955
GPT-6 Astra	0.278	0.884	0.271	0.882	0.408	0.907	0.673	0.910	0.692	0.922	0.824	0.929	0.769	0.671	1.000	0.960
Grok 4.6	0.065	0.843	0.027	0.828	0.104	0.827	0.224	0.851	0.198	0.839	0.309	0.845	0.571	0.255	1.000	0.875
Kimi K3	0.013	0.720	0.015	0.673	0.038	0.717	0.169	0.772	0.174	0.777	0.330	0.787	0.561	0.211	1.000	0.889
Qwen3.8-Flash	0.120	0.842	0.216	0.860	0.331	0.825	0.466	0.842	0.455	0.840	0.681	0.858	0.671	0.534	0.962	0.921
Qwen3.8-Max	0.101	0.819	0.179	0.820	0.259	0.822	0.503	0.850	0.504	0.864	0.715	0.898	0.693	0.528	1.000	0.903
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.177	0.966	0.257	0.960	0.371	0.961	0.562	0.966	0.637	0.969	0.895	0.979	0.738	0.644	1.000	0.979
GPT-5.4 mini	0.134	0.943	0.188	0.949	0.303	0.953	0.440	0.951	0.403	0.960	0.647	0.967	0.676	0.503	0.986	0.916
Grok 4.1 Fast Reasoning	0.149	0.928	0.138	0.925	0.189	0.923	0.228	0.921	0.220	0.916	0.336	0.920	0.565	0.318	0.932	0.954
Grok 4.20 Reasoning	0.061	0.739	0.130	0.718	0.155	0.728	0.270	0.747	0.209	0.740	0.463	0.751	0.593	0.346	0.977	0.778
Kimi K2.6	0.033	0.873	0.095	0.880	0.140	0.873	0.231	0.875	0.440	0.884	0.572	0.886	0.627	0.390	1.000	0.712
QVQ-Max-Latest	0.205	0.818	0.327	0.802	0.279	0.836	0.326	0.824	0.345	0.835	0.451	0.836	0.526	0.457	0.725	0.943
Qwen3-VL-Flash	0.130	0.950	0.289	0.950	0.347	0.950	0.412	0.950	0.494	0.950	0.603	0.950	0.683	0.525	0.985	0.797
Qwen3-VL-Plus	0.010	0.935	0.018	0.927	0.046	0.947	0.113	0.961	0.113	0.959	0.338	0.960	0.552	0.181	0.993	0.979
Qwen3.6-35B-A3B	0.311	0.930	0.394	0.928	0.454	0.931	0.556	0.935	0.587	0.934	0.746	0.939	0.697	0.653	0.874	0.971
Qwen3.6-Plus	0.136	0.934	0.214	0.920	0.262	0.936	0.471	0.935	0.482	0.935	0.599	0.939	0.673	0.514	0.985	0.960

Table 17: **Performance comparison of MLLMs on the 27 Amazon categories of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the multi-image setting with review context, macro-averaged over the 27 Amazon categories, for the 18 MLLMs evaluated on these categories in both multi-image settings. The first block repeats the MLLMs of Tab. 15 restricted to the same categories.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.335	0.707	0.365	0.734	0.440	0.728	0.576	0.739	0.631	0.758	0.788	0.759	0.748	0.674	0.978	0.804
Gemini 3.8 Flash	0.323	0.719	0.263	0.718	0.405	0.720	0.544	0.716	0.559	0.718	0.781	0.721	0.739	0.630	1.000	0.688
GPT-5.6 Terra	0.145	0.907	0.152	0.902	0.270	0.904	0.586	0.944	0.428	0.932	0.698	0.949	0.684	0.540	1.000	0.968
GPT-6 Astra	0.748	0.863	0.760	0.865	0.828	0.885	0.911	0.911	0.959	0.923	0.982	0.939	0.930	0.924	1.000	0.958
Grok 4.6	0.070	0.849	0.038	0.824	0.128	0.841	0.302	0.855	0.230	0.863	0.366	0.850	0.592	0.307	1.000	0.898
Kimi K3	0.098	0.748	0.161	0.735	0.232	0.772	0.375	0.771	0.359	0.780	0.577	0.793	0.649	0.448	1.000	0.902
Qwen3.8-Flash	0.136	0.844	0.235	0.828	0.309	0.838	0.479	0.865	0.520	0.861	0.596	0.880	0.683	0.536	0.985	0.922
Qwen3.8-Max	0.108	0.822	0.176	0.818	0.235	0.830	0.476	0.852	0.514	0.863	0.700	0.892	0.688	0.520	1.000	0.916
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.453	0.954	0.487	0.955	0.591	0.962	0.690	0.968	0.805	0.969	0.932	0.977	0.823	0.788	1.000	0.982
GPT-5.4 mini	0.323	0.940	0.335	0.936	0.379	0.946	0.503	0.952	0.489	0.954	0.639	0.955	0.724	0.588	0.987	0.925
Grok 4.1 Fast Reasoning	0.172	0.924	0.217	0.920	0.172	0.919	0.313	0.923	0.312	0.916	0.364	0.929	0.579	0.372	0.921	0.952
Grok 4.20 Reasoning	0.040	0.718	0.128	0.730	0.167	0.755	0.213	0.754	0.194	0.734	0.387	0.746	0.590	0.303	0.990	0.783
Kimi K2.6	0.147	0.889	0.228	0.888	0.296	0.891	0.264	0.894	0.467	0.890	0.625	0.894	0.675	0.485	1.000	0.255
QVQ-Max-Latest	0.169	0.850	0.163	0.836	0.194	0.829	0.313	0.873	0.266	0.869	0.287	0.858	0.575	0.358	0.909	0.942
Qwen3-VL-Flash	0.179	0.950	0.281	0.950	0.422	0.950	0.432	0.950	0.459	0.950	0.589	0.950	0.692	0.544	0.980	0.757
Qwen3-VL-Plus	0.003	0.980	0.015	0.960	0.081	0.961	0.084	0.970	0.093	0.966	0.248	0.968	0.544	0.151	0.996	0.980
Qwen3.6-35B-A3B	0.160	0.920	0.277	0.922	0.295	0.932	0.409	0.944	0.465	0.936	0.597	0.941	0.648	0.517	0.920	0.971
Qwen3.6-Plus	0.085	0.941	0.172	0.933	0.200	0.940	0.421	0.938	0.422	0.931	0.619	0.940	0.661	0.470	0.996	0.962

Table 18: **Performance comparison of MLLMs on FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the multi-step setting without review context, in which the images of a review are presented one per turn in a multi-turn conversation. Results are shown for the six MLLMs evaluated in this setting.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.371	0.785	0.549	0.790	0.629	0.789	0.634	0.818	0.739	0.840	0.870	0.853	0.762	0.753	0.905	0.768
Gemini 3.8 Flash	0.221	0.735	0.245	0.734	0.399	0.740	0.533	0.733	0.608	0.741	0.759	0.745	0.724	0.613	0.995	0.820
GPT-5.6 Terra	0.044	0.824	0.072	0.854	0.104	0.859	0.365	0.901	0.275	0.856	0.469	0.922	0.612	0.348	1.000	0.908
Grok 4.6	0.139	0.782	0.116	0.774	0.255	0.799	0.380	0.829	0.378	0.826	0.481	0.834	0.634	0.429	0.996	0.868
Kimi K3	0.056	0.758	0.126	0.731	0.209	0.799	0.419	0.780	0.350	0.804	0.612	0.818	0.649	0.441	0.999	0.881
Qwen3.8-Max	0.339	0.823	0.469	0.837	0.624	0.839	0.778	0.878	0.723	0.875	0.930	0.896	0.818	0.772	0.987	0.893

Table 19: **Performance comparison of MLLMs on FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the multi-step setting with review context, in which the images of a review are presented one per turn in a multi-turn conversation. Results are shown for the six MLLMs evaluated in this setting.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.513	0.788	0.737	0.792	0.773	0.816	0.751	0.830	0.867	0.835	0.934	0.851	0.816	0.849	0.878	0.788
Gemini 3.8 Flash	0.436	0.764	0.400	0.778	0.544	0.769	0.694	0.748	0.733	0.765	0.869	0.748	0.806	0.747	1.000	0.869
GPT-5.6 Terra	0.050	0.836	0.068	0.850	0.132	0.875	0.418	0.887	0.317	0.877	0.531	0.911	0.625	0.391	1.000	0.927
Grok 4.6	0.218	0.805	0.184	0.797	0.348	0.814	0.555	0.856	0.564	0.826	0.655	0.842	0.707	0.579	0.997	0.867
Kimi K3	0.177	0.738	0.240	0.721	0.322	0.780	0.493	0.790	0.509	0.816	0.795	0.818	0.713	0.580	1.000	0.902
Qwen3.8-Max	0.389	0.841	0.526	0.823	0.619	0.850	0.797	0.879	0.756	0.882	0.883	0.905	0.835	0.785	0.997	0.903

Table 20: **Performance comparison of MLLMs on the All Beauty subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.179	0.724	0.286	0.756	0.393	0.681	0.464	0.713	0.393	0.747	0.607	0.755	0.649	0.551	0.912	0.755
Gemini 3.1 Pro	0.179	0.730	0.429	0.726	0.321	0.721	0.357	0.720	0.500	0.719	0.857	0.718	0.691	0.607	0.941	0.710
Gemini 3.5 Flash-Lite	0.071	0.970	0.214	0.962	0.143	0.967	0.250	0.964	0.179	0.964	0.429	0.967	0.607	0.353	1.000	0.951
Gemini 3.8 Flash	0.143	0.710	0.000	—	0.107	0.707	0.179	0.696	0.143	0.720	0.286	0.710	0.571	0.250	1.000	0.611
GPT-6 Astra	0.107	0.787	0.036	0.740	0.179	0.822	0.321	0.867	0.357	0.835	0.500	0.846	0.625	0.400	1.000	0.939
GPT-6 Luna	0.857	0.928	1.000	0.937	1.000	0.949	1.000	0.946	1.000	0.940	1.000	0.959	0.547	0.906	0.118	0.805
GPT-6 Sol	0.107	0.780	0.143	0.760	0.143	0.820	0.357	0.890	0.321	0.781	0.429	0.885	0.625	0.400	1.000	0.905
Grok 4.6	0.036	0.780	0.250	0.824	0.429	0.794	0.393	0.808	0.357	0.822	0.464	0.835	0.661	0.486	1.000	0.836
Kimi K3	0.000	—	0.036	0.620	0.071	0.765	0.286	0.816	0.179	0.696	0.571	0.769	0.595	0.320	1.000	0.855
Qwen3-VL-Flash	0.107	0.930	0.286	0.939	0.250	0.946	0.393	0.945	0.286	0.943	0.679	0.933	0.667	0.500	1.000	0.978
Qwen3-VL-Plus	0.036	0.920	0.071	0.920	0.071	0.935	0.179	0.932	0.071	0.950	0.143	0.943	0.548	0.174	1.000	0.980
Qwen3.5-Omni-Flash	0.143	0.905	0.214	0.910	0.214	0.920	0.393	0.911	0.321	0.918	0.464	0.910	0.646	0.452	1.000	0.949
Qwen3.5-Omni-Plus	0.000	—	0.071	0.920	0.143	0.920	0.107	0.930	0.107	0.920	0.250	0.929	0.557	0.203	1.000	0.974
Qwen3.8-27B	0.143	0.775	0.143	0.898	0.143	0.790	0.357	0.849	0.214	0.810	0.429	0.875	0.619	0.385	1.000	0.863
Qwen3.8-Flash	0.143	0.890	0.536	0.883	0.536	0.908	0.714	0.910	0.536	0.911	0.643	0.894	0.700	0.672	0.882	0.945
Qwen3.8-Max	0.393	0.785	0.464	0.874	0.429	0.832	0.607	0.852	0.714	0.853	0.893	0.882	0.777	0.734	0.971	0.863
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.357	0.949	0.464	0.953	0.357	0.954	0.536	0.971	0.500	0.965	0.750	0.976	0.718	0.656	0.941	0.972
GPT-5.4 mini	0.179	0.936	0.143	0.930	0.143	0.940	0.179	0.932	0.179	0.960	0.393	0.966	0.601	0.337	1.000	0.917
GPT-5.6 Luna	0.143	0.885	0.107	0.923	0.107	0.883	0.214	0.958	0.179	0.922	0.607	0.892	0.613	0.369	1.000	0.968
GPT-5.6 Sol	0.000	—	0.036	0.880	0.036	0.960	0.107	0.930	0.107	0.893	0.214	0.957	0.542	0.154	1.000	0.976
GPT-5.6 Terra	0.071	0.845	0.143	0.868	0.179	0.768	0.429	0.907	0.357	0.850	0.643	0.893	0.637	0.464	0.971	0.939
Grok 4.1 Fast Reasoning	0.321	0.937	0.429	0.895	0.357	0.938	0.536	0.931	0.607	0.911	0.536	0.930	0.629	0.617	0.794	0.943
Grok 4.20 Reasoning	0.179	0.720	0.214	0.695	0.214	0.755	0.321	0.750	0.214	0.712	0.464	0.720	0.590	0.417	0.912	0.769
Kimi K2.6	0.036	0.920	0.107	0.897	0.071	0.885	0.179	0.898	0.214	0.908	0.429	0.899	0.586	0.294	1.000	0.926
QVQ-Max-Latest	0.179	0.720	0.250	0.757	0.321	0.761	0.214	0.725	0.357	0.770	0.500	0.757	0.519	0.447	0.735	0.934
Qwen3.6-35B-A3B	0.250	0.937	0.464	0.924	0.393	0.925	0.643	0.935	0.679	0.908	0.643	0.936	0.682	0.664	0.853	0.968
Qwen3.6-Plus	0.143	0.887	0.286	0.914	0.250	0.941	0.357	0.951	0.393	0.934	0.679	0.939	0.646	0.515	0.941	0.957
Qwen3.7-Flash	0.107	0.930	0.214	0.933	0.179	0.956	0.571	0.951	0.464	0.948	0.750	0.950	0.690	0.552	1.000	0.964
Qwen3.7-Max	0.036	0.900	0.071	0.950	0.107	0.907	0.286	0.904	0.321	0.902	0.429	0.928	0.604	0.345	1.000	0.949
Qwen3.7-Plus	0.071	0.850	0.143	0.892	0.179	0.904	0.321	0.939	0.286	0.896	0.393	0.925	0.616	0.377	1.000	0.946
Specialized Detectors																
CO-SPY [ProGAN]	1.000	0.999	0.214	0.902	0.393	0.903	0.250	0.881	0.964	0.972	1.000	0.998	0.789	0.773	0.941	1.000
CO-SPY [SD-v1.4]	0.357	0.743	0.286	0.796	0.250	0.779	0.250	0.802	0.286	0.730	0.321	0.708	0.602	0.445	0.912	0.886
ForgeLens [ProGAN]	0.643	0.934	0.179	0.868	0.357	0.863	0.714	0.921	0.714	0.956	1.000	0.962	0.771	0.745	0.941	0.970
ForgeLens [GenImage]	1.000	0.907	0.179	0.728	1.000	0.943	0.964	0.908	1.000	0.924	1.000	0.920	0.929	0.923	1.000	0.913
Effort [SD-v1.4]	0.786	0.856	0.571	0.823	0.607	0.823	0.786	0.854	0.857	0.898	1.000	0.875	0.796	0.851	0.824	0.843
Effort [Chameleon]	0.893	0.687	0.821	0.698	0.893	0.737	0.929	0.742	0.929	0.704	0.964	0.701	0.555	0.876	0.206	0.560
IAPL [ProGAN]	0.107	0.852	0.250	0.880	0.036	0.866	0.036	0.993	0.357	0.919	0.571	0.881	0.613	0.369	1.000	0.993
IAPL [SD-v1.4]	1.000	0.835	0.929	0.788	0.929	0.823	1.000	0.926	1.000	0.839	1.000	0.869	0.665	0.927	0.353	0.798

Table 21: **Performance comparison of MLLMs on the Amazon Fashion subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.029	0.620	0.143	0.726	0.114	0.700	0.057	0.660	0.343	0.740	0.514	0.707	0.600	0.333	1.000	0.754
Gemini 3.1 Pro	0.114	0.730	0.143	0.694	0.086	0.740	0.257	0.722	0.314	0.710	0.600	0.716	0.597	0.402	0.941	0.708
Gemini 3.5 Flash-Lite	0.029	0.950	0.000	—	0.029	0.950	0.029	0.990	0.057	0.965	0.114	0.960	0.521	0.082	1.000	0.952
Gemini 3.8 Flash	0.000	—	0.000	—	0.029	0.680	0.000	—	0.057	0.680	0.114	0.690	0.517	0.065	1.000	0.651
GPT-6 Astra	0.057	0.760	0.029	0.760	0.143	0.796	0.314	0.800	0.371	0.815	0.514	0.802	0.619	0.385	1.000	0.938
GPT-6 Luna	0.600	0.874	0.914	0.901	0.943	0.888	0.857	0.913	0.829	0.861	0.857	0.911	0.475	0.875	0.118	0.800
GPT-6 Sol	0.114	0.755	0.171	0.763	0.286	0.779	0.229	0.859	0.171	0.763	0.514	0.826	0.624	0.397	1.000	0.918
Grok 4.6	0.000	—	0.000	—	0.029	0.780	0.029	0.680	0.029	0.880	0.171	0.775	0.521	0.082	1.000	0.839
Kimi K3	0.000	—	0.029	0.620	0.029	0.600	0.057	0.770	0.086	0.753	0.200	0.727	0.504	0.124	0.941	0.862
Qwen3-VL-Flash	0.000	—	0.000	—	0.057	0.920	0.143	0.918	0.086	0.940	0.286	0.941	0.518	0.173	0.941	0.976
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.029	0.920	0.029	0.920	0.143	0.950	0.517	0.065	1.000	0.976
Qwen3.5-Omni-Flash	0.000	—	0.000	—	0.029	0.920	0.029	0.850	0.057	0.905	0.171	0.915	0.524	0.091	1.000	0.954
Qwen3.5-Omni-Plus	0.000	—	0.000	—	0.000	—	0.029	0.920	0.057	0.930	0.086	0.920	0.514	0.056	1.000	0.971
Qwen3.8-27B	0.057	0.830	0.057	0.750	0.114	0.870	0.114	0.803	0.114	0.783	0.171	0.840	0.552	0.190	1.000	0.847
Qwen3.8-Flash	0.114	0.895	0.257	0.878	0.229	0.870	0.286	0.906	0.314	0.857	0.457	0.887	0.609	0.431	0.941	0.956
Qwen3.8-Max	0.143	0.828	0.114	0.775	0.143	0.796	0.314	0.795	0.343	0.819	0.629	0.843	0.640	0.439	1.000	0.868
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.171	0.960	0.143	0.946	0.229	0.951	0.286	0.955	0.486	0.953	0.743	0.957	0.613	0.507	0.882	0.968
GPT-5.4 mini	0.029	0.970	0.029	0.930	0.143	0.882	0.086	0.947	0.057	0.870	0.343	0.935	0.557	0.205	1.000	0.924
GPT-5.6 Luna	0.000	—	0.000	—	0.029	0.860	0.057	0.885	0.029	0.880	0.257	0.901	0.531	0.117	1.000	0.972
GPT-5.6 Sol	0.000	—	0.029	0.720	0.000	—	0.114	0.883	0.114	0.880	0.314	0.886	0.548	0.174	1.000	0.979
GPT-5.6 Terra	0.086	0.800	0.114	0.815	0.200	0.754	0.286	0.855	0.229	0.778	0.457	0.825	0.614	0.372	1.000	0.936
Grok 4.1 Fast Reasoning	0.000	—	0.086	0.827	0.029	0.950	0.086	0.907	0.114	0.907	0.114	0.900	0.536	0.133	1.000	0.952
Grok 4.20 Reasoning	0.086	0.703	0.000	—	0.086	0.710	0.114	0.675	0.114	0.750	0.257	0.718	0.525	0.197	0.941	0.783
Kimi K2.6	0.000	—	0.000	—	0.000	—	0.200	0.904	0.171	0.908	0.343	0.897	0.560	0.213	1.000	0.922
QVQ-Max-Latest	0.229	0.756	0.200	0.729	0.286	0.730	0.286	0.735	0.343	0.725	0.314	0.736	0.432	0.422	0.588	0.910
Qwen3.6-35B-A3B	0.000	—	0.200	0.911	0.200	0.916	0.229	0.881	0.343	0.908	0.343	0.896	0.492	0.354	0.765	0.968
Qwen3.6-Plus	0.029	0.950	0.057	0.915	0.114	0.910	0.143	0.912	0.200	0.923	0.400	0.923	0.579	0.272	1.000	0.954
Qwen3.7-Flash	0.029	0.920	0.057	0.900	0.057	0.915	0.086	0.940	0.114	0.942	0.343	0.929	0.528	0.204	0.941	0.952
Qwen3.7-Max	0.000	—	0.000	—	0.029	0.780	0.057	0.885	0.143	0.856	0.229	0.862	0.538	0.142	1.000	0.945
Qwen3.7-Plus	0.029	0.850	0.057	0.850	0.086	0.917	0.057	0.935	0.143	0.874	0.229	0.877	0.521	0.181	0.941	0.944
Specialized Detectors																
CO-SPY [ProGAN]	1.000	0.996	0.257	0.958	0.400	0.922	0.286	0.940	0.800	0.928	1.000	1.000	0.812	0.768	1.000	0.998
CO-SPY [SD-v1.4]	0.400	0.734	0.229	0.773	0.257	0.687	0.257	0.775	0.343	0.735	0.229	0.828	0.613	0.443	0.941	0.936
ForgeLens [ProGAN]	0.686	0.941	0.229	0.826	0.543	0.878	0.857	0.961	1.000	0.973	0.971	0.991	0.798	0.829	0.882	0.956
ForgeLens [GenImage]	0.943	0.920	0.286	0.825	1.000	0.943	1.000	0.909	1.000	0.926	1.000	0.922	0.906	0.929	0.941	0.901
Effort [SD-v1.4]	0.886	0.859	0.771	0.842	0.886	0.835	0.914	0.861	0.971	0.897	0.971	0.950	0.950	0.947	1.000	0.858
Effort [Chameleon]	0.971	0.745	0.857	0.744	0.886	0.777	0.943	0.800	0.914	0.735	0.971	0.719	0.462	0.922	0.000	—
IAPL [ProGAN]	0.143	0.844	0.114	0.983	0.114	0.882	0.114	0.786	0.200	0.808	0.400	0.915	0.590	0.306	1.000	0.995
IAPL [SD-v1.4]	1.000	0.881	0.857	0.865	0.886	0.878	1.000	0.938	0.943	0.827	0.971	0.889	0.648	0.945	0.353	0.843

Table 22: **Performance comparison of MLLMs on the Appliances subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.038	0.600	0.154	0.705	0.231	0.657	0.423	0.687	0.577	0.743	0.346	0.738	0.596	0.447	0.897	0.748
Gemini 3.1 Pro	0.038	0.700	0.000	—	0.077	0.685	0.192	0.724	0.423	0.699	0.423	0.715	0.596	0.323	1.000	0.706
Gemini 3.5 Flash-Lite	0.000	—	0.000	—	0.000	—	0.077	0.950	0.077	0.965	0.192	0.972	0.529	0.109	1.000	0.949
Gemini 3.8 Flash	0.077	0.700	0.000	—	0.154	0.705	0.231	0.700	0.269	0.697	0.192	0.708	0.577	0.267	1.000	0.702
GPT-6 Astra	0.077	0.765	0.077	0.830	0.346	0.822	0.615	0.848	0.731	0.847	0.692	0.845	0.712	0.595	1.000	0.929
GPT-6 Luna	0.731	0.915	0.962	0.948	0.962	0.959	0.923	0.936	0.846	0.910	0.923	0.944	0.561	0.855	0.231	0.740
GPT-6 Sol	0.038	0.760	0.077	0.825	0.308	0.795	0.423	0.848	0.423	0.873	0.423	0.868	0.628	0.438	0.974	0.908
Grok 4.6	0.038	0.840	0.038	0.760	0.231	0.847	0.423	0.850	0.346	0.819	0.308	0.834	0.615	0.375	1.000	0.844
Kimi K3	0.038	0.720	0.115	0.710	0.038	0.650	0.269	0.800	0.154	0.657	0.231	0.807	0.571	0.247	1.000	0.865
Qwen3-VL-Flash	0.038	0.920	0.077	0.950	0.308	0.935	0.423	0.939	0.577	0.944	0.577	0.940	0.667	0.500	1.000	0.970
Qwen3-VL-Plus	0.000	—	0.000	—	0.077	0.950	0.115	0.950	0.192	0.962	0.077	0.950	0.538	0.143	1.000	0.976
Qwen3.5-Omni-Flash	0.000	—	0.038	0.870	0.269	0.916	0.346	0.926	0.423	0.925	0.423	0.915	0.625	0.400	1.000	0.953
Qwen3.5-Omni-Plus	0.000	—	0.000	—	0.077	0.930	0.154	0.920	0.192	0.944	0.231	0.942	0.554	0.197	1.000	0.972
Qwen3.8-27B	0.077	0.860	0.077	0.790	0.385	0.828	0.423	0.796	0.423	0.824	0.500	0.845	0.657	0.478	1.000	0.858
Qwen3.8-Flash	0.038	0.820	0.115	0.947	0.346	0.902	0.577	0.879	0.500	0.902	0.731	0.902	0.654	0.548	0.923	0.937
Qwen3.8-Max	0.000	—	0.038	0.920	0.231	0.857	0.423	0.835	0.615	0.843	0.577	0.855	0.657	0.478	1.000	0.864
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.077	0.965	0.038	0.980	0.231	0.960	0.462	0.969	0.462	0.951	0.577	0.955	0.654	0.471	1.000	0.968
GPT-5.4 mini	0.038	0.970	0.038	0.910	0.154	0.955	0.192	0.930	0.154	0.963	0.154	0.968	0.561	0.217	1.000	0.908
GPT-5.6 Luna	0.000	—	0.000	—	0.000	—	0.038	0.960	0.115	0.900	0.269	0.856	0.535	0.132	1.000	0.967
GPT-5.6 Sol	0.000	—	0.038	0.910	0.077	0.860	0.192	0.904	0.269	0.890	0.385	0.930	0.580	0.276	1.000	0.973
GPT-5.6 Terra	0.038	0.900	0.038	0.880	0.038	0.780	0.346	0.888	0.269	0.804	0.538	0.842	0.580	0.346	0.949	0.924
Grok 4.1 Fast Reasoning	0.154	0.935	0.154	0.925	0.308	0.909	0.269	0.913	0.462	0.910	0.538	0.909	0.631	0.473	0.949	0.948
Grok 4.20 Reasoning	0.154	0.728	0.154	0.685	0.269	0.737	0.346	0.743	0.308	0.703	0.269	0.700	0.612	0.398	0.974	0.758
Kimi K2.6	0.000	—	0.038	0.920	0.038	0.920	0.269	0.874	0.269	0.910	0.308	0.894	0.577	0.267	1.000	0.854
QVQ-Max-Latest	0.308	0.719	0.346	0.711	0.615	0.719	0.385	0.730	0.500	0.773	0.462	0.762	0.538	0.571	0.641	0.938
Qwen3.6-35B-A3B	0.231	0.907	0.385	0.916	0.346	0.923	0.538	0.923	0.538	0.917	0.577	0.932	0.654	0.594	0.872	0.966
Qwen3.6-Plus	0.038	0.950	0.154	0.910	0.192	0.912	0.385	0.935	0.385	0.941	0.385	0.928	0.615	0.406	0.974	0.953
Qwen3.7-Flash	0.077	0.935	0.115	0.957	0.308	0.944	0.385	0.950	0.423	0.945	0.423	0.954	0.644	0.448	1.000	0.961
Qwen3.7-Max	0.038	0.920	0.154	0.875	0.038	0.950	0.462	0.900	0.462	0.883	0.385	0.910	0.628	0.408	1.000	0.945
Qwen3.7-Plus	0.000	—	0.077	0.900	0.154	0.880	0.308	0.905	0.423	0.905	0.423	0.931	0.615	0.375	1.000	0.944
Specialized Detectors																
CO-SPY [ProGAN]	1.000	1.000	0.000	—	0.346	0.886	0.154	0.854	0.846	0.950	1.000	1.000	0.779	0.716	1.000	1.000
CO-SPY [SD-v1.4]	0.385	0.752	0.192	0.883	0.154	0.831	0.308	0.796	0.192	0.754	0.385	0.737	0.455	0.396	0.641	0.830
ForgeLens [ProGAN]	0.808	0.912	0.115	0.926	0.538	0.881	0.885	0.984	0.885	0.976	1.000	0.973	0.840	0.824	0.974	0.966
ForgeLens [GenImage]	1.000	0.882	0.000	—	1.000	0.931	1.000	0.889	1.000	0.921	0.962	0.905	0.913	0.905	1.000	0.871
Effort [SD-v1.4]	0.692	0.737	0.538	0.742	0.538	0.728	0.846	0.713	0.885	0.849	0.846	0.853	0.734	0.810	0.744	0.846
Effort [Chameleon]	0.846	0.715	0.846	0.687	0.885	0.709	1.000	0.764	0.846	0.752	0.846	0.776	0.452	0.828	0.026	0.606
IAPL [ProGAN]	0.577	0.903	0.385	0.977	0.115	0.935	0.115	0.938	0.577	0.940	0.654	0.977	0.689	0.573	0.974	0.989
IAPL [SD-v1.4]	1.000	0.877	0.885	0.853	0.769	0.872	1.000	0.923	0.846	0.877	0.962	0.906	0.647	0.882	0.385	0.800

Table 23: **Performance comparison of MLLMs on the Arts, Crafts & Sewing subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.150	0.763	0.300	0.633	0.350	0.733	0.400	0.734	0.350	0.790	0.400	0.771	0.662	0.491	1.000	0.776
Gemini 3.1 Pro	0.150	0.717	0.300	0.720	0.150	0.667	0.250	0.718	0.350	0.709	0.500	0.721	0.642	0.442	1.000	0.712
Gemini 3.5 Flash-Lite	0.150	0.950	0.250	0.966	0.150	0.963	0.150	0.977	0.400	0.974	0.400	0.978	0.625	0.400	1.000	0.950
Gemini 3.8 Flash	0.000	—	0.000	—	0.050	0.680	0.000	—	0.100	0.710	0.100	0.730	0.521	0.080	1.000	0.686
GPT-6 Astra	0.000	—	0.050	0.780	0.200	0.835	0.300	0.813	0.250	0.810	0.300	0.848	0.592	0.310	1.000	0.918
GPT-6 Luna	0.800	0.907	0.950	0.932	0.900	0.948	0.950	0.953	0.800	0.932	1.000	0.951	0.579	0.861	0.258	0.752
GPT-6 Sol	0.000	—	0.050	0.770	0.000	—	0.150	0.837	0.150	0.863	0.450	0.873	0.567	0.235	1.000	0.901
Grok 4.6	0.250	0.806	0.150	0.867	0.400	0.853	0.350	0.864	0.300	0.817	0.350	0.824	0.650	0.462	1.000	0.839
Kimi K3	0.000	—	0.100	0.845	0.100	0.750	0.200	0.882	0.200	0.728	0.450	0.832	0.571	0.296	0.968	0.843
Qwen3-VL-Flash	0.100	0.935	0.100	0.950	0.250	0.938	0.300	0.935	0.200	0.927	0.250	0.944	0.584	0.331	0.968	0.970
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.484	0.000	0.968	0.975
Qwen3.5-Omni-Flash	0.050	0.920	0.100	0.890	0.250	0.898	0.100	0.885	0.200	0.882	0.350	0.899	0.588	0.298	1.000	0.940
Qwen3.5-Omni-Plus	0.050	0.920	0.050	0.920	0.000	—	0.100	0.895	0.050	0.920	0.400	0.927	0.554	0.195	1.000	0.967
Qwen3.8-27B	0.150	0.860	0.100	0.850	0.350	0.816	0.300	0.833	0.200	0.800	0.400	0.874	0.593	0.395	0.935	0.879
Qwen3.8-Flash	0.150	0.900	0.500	0.870	0.450	0.909	0.550	0.905	0.500	0.860	0.750	0.927	0.693	0.641	0.903	0.922
Qwen3.8-Max	0.300	0.793	0.450	0.818	0.450	0.843	0.500	0.860	0.550	0.862	0.700	0.890	0.730	0.656	0.968	0.847
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.300	0.970	0.350	0.971	0.500	0.959	0.500	0.976	0.600	0.976	0.600	0.978	0.738	0.644	1.000	0.965
GPT-5.4 mini	0.100	0.930	0.000	—	0.150	0.957	0.250	0.928	0.200	0.922	0.450	0.963	0.596	0.322	1.000	0.915
GPT-5.6 Luna	0.100	0.880	0.250	0.856	0.250	0.920	0.200	0.945	0.250	0.946	0.450	0.936	0.625	0.400	1.000	0.966
GPT-5.6 Sol	0.000	—	0.050	0.820	0.050	0.980	0.050	0.990	0.100	0.960	0.300	0.947	0.546	0.168	1.000	0.973
GPT-5.6 Terra	0.150	0.790	0.200	0.890	0.300	0.792	0.400	0.934	0.350	0.851	0.550	0.911	0.646	0.487	0.968	0.928
Grok 4.1 Fast Reasoning	0.350	0.917	0.300	0.938	0.350	0.930	0.400	0.934	0.500	0.930	0.400	0.946	0.643	0.544	0.903	0.944
Grok 4.20 Reasoning	0.100	0.715	0.250	0.740	0.150	0.763	0.300	0.787	0.150	0.787	0.200	0.797	0.564	0.317	0.935	0.740
Kimi K2.6	0.050	0.850	0.100	0.920	0.150	0.850	0.250	0.864	0.300	0.885	0.400	0.894	0.588	0.342	0.968	0.914
QVQ-Max-Latest	0.250	0.750	0.350	0.736	0.250	0.810	0.200	0.788	0.350	0.750	0.300	0.742	0.513	0.420	0.742	0.920
Qwen3.6-35B-A3B	0.250	0.938	0.400	0.959	0.350	0.950	0.550	0.950	0.400	0.946	0.450	0.942	0.668	0.565	0.935	0.967
Qwen3.6-Plus	0.250	0.950	0.150	0.953	0.400	0.946	0.350	0.954	0.450	0.940	0.550	0.953	0.679	0.528	1.000	0.946
Qwen3.7-Flash	0.350	0.954	0.300	0.968	0.300	0.963	0.450	0.972	0.500	0.951	0.450	0.973	0.696	0.563	1.000	0.960
Qwen3.7-Max	0.000	—	0.100	0.950	0.250	0.894	0.300	0.892	0.250	0.930	0.550	0.945	0.621	0.389	1.000	0.945
Qwen3.7-Plus	0.050	0.920	0.100	0.900	0.300	0.883	0.100	0.925	0.200	0.938	0.450	0.947	0.600	0.333	1.000	0.944
Specialized Detectors																
CO-SPY [ProGAN]	0.950	0.983	0.050	1.000	0.250	0.896	0.200	0.757	0.800	0.971	0.950	0.993	0.767	0.696	1.000	1.000
CO-SPY [SD-v1.4]	0.650	0.769	0.550	0.779	0.400	0.836	0.550	0.799	0.700	0.796	0.650	0.759	0.566	0.686	0.548	0.828
ForgeLens [ProGAN]	0.700	0.965	0.050	0.992	0.450	0.912	0.900	0.983	0.800	0.969	0.850	0.968	0.796	0.765	0.968	0.952
ForgeLens [GenImage]	1.000	0.893	0.150	0.753	1.000	0.943	0.950	0.911	1.000	0.912	1.000	0.903	0.909	0.915	0.968	0.893
Effort [SD-v1.4]	0.600	0.828	0.700	0.734	0.600	0.787	0.650	0.853	0.900	0.874	0.850	0.921	0.762	0.811	0.806	0.840
Effort [Chameleon]	0.500	0.609	0.400	0.654	0.600	0.658	0.650	0.701	0.600	0.626	0.500	0.651	0.351	0.616	0.161	0.551
IAPL [ProGAN]	0.250	0.823	0.300	0.953	0.150	0.809	0.000	—	0.450	0.909	0.700	0.893	0.638	0.468	0.968	0.977
IAPL [SD-v1.4]	0.950	0.907	0.850	0.872	0.850	0.882	1.000	0.933	0.900	0.867	0.950	0.916	0.668	0.887	0.419	0.835

Table 24: **Performance comparison of MLLMs on the Automotive subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.034	0.620	0.241	0.669	0.138	0.655	0.207	0.683	0.241	0.711	0.517	0.689	0.615	0.374	1.000	0.757
Gemini 3.1 Pro	0.034	0.700	0.034	0.700	0.069	0.725	0.207	0.723	0.207	0.710	0.517	0.718	0.589	0.302	1.000	0.699
Gemini 3.5 Flash-Lite	0.000	—	0.103	0.970	0.069	0.965	0.069	0.965	0.138	0.975	0.241	0.971	0.552	0.188	1.000	0.951
Gemini 3.8 Flash	0.034	0.720	0.034	0.720	0.034	0.700	0.034	0.680	0.172	0.704	0.276	0.710	0.549	0.178	1.000	0.664
GPT-6 Astra	0.138	0.770	0.207	0.805	0.207	0.835	0.655	0.811	0.655	0.843	0.759	0.845	0.718	0.608	1.000	0.933
GPT-6 Luna	0.621	0.890	0.828	0.889	0.897	0.892	0.966	0.918	0.828	0.859	0.931	0.911	0.556	0.857	0.267	0.766
GPT-6 Sol	0.207	0.765	0.172	0.736	0.345	0.806	0.517	0.810	0.517	0.803	0.655	0.858	0.701	0.574	1.000	0.918
Grok 4.6	0.034	0.780	0.103	0.817	0.034	0.840	0.103	0.817	0.310	0.817	0.414	0.811	0.583	0.286	1.000	0.846
Kimi K3	0.000	—	0.000	—	0.034	0.800	0.034	0.850	0.172	0.790	0.207	0.778	0.504	0.138	0.933	0.871
Qwen3-VL-Flash	0.069	0.935	0.207	0.940	0.276	0.946	0.414	0.940	0.483	0.934	0.724	0.938	0.681	0.532	1.000	0.974
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.103	0.960	0.069	0.965	0.514	0.056	1.000	0.979
Qwen3.5-Omni-Flash	0.034	0.920	0.069	0.920	0.069	0.920	0.276	0.905	0.276	0.927	0.379	0.908	0.592	0.311	1.000	0.955
Qwen3.5-Omni-Plus	0.034	0.920	0.034	0.920	0.034	0.920	0.069	0.920	0.138	0.940	0.138	0.943	0.537	0.139	1.000	0.972
Qwen3.8-27B	0.069	0.760	0.000	—	0.138	0.852	0.069	0.780	0.207	0.812	0.276	0.846	0.563	0.224	1.000	0.863
Qwen3.8-Flash	0.034	0.850	0.241	0.891	0.207	0.913	0.276	0.860	0.276	0.945	0.483	0.876	0.593	0.400	0.933	0.941
Qwen3.8-Max	0.103	0.797	0.172	0.804	0.241	0.803	0.483	0.832	0.552	0.841	0.759	0.850	0.693	0.556	1.000	0.853
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.172	0.952	0.103	0.960	0.172	0.942	0.276	0.946	0.448	0.948	0.724	0.968	0.658	0.480	1.000	0.968
GPT-5.4 mini	0.069	0.905	0.000	—	0.138	0.958	0.103	0.960	0.241	0.949	0.207	0.970	0.547	0.223	0.967	0.915
GPT-5.6 Luna	0.034	0.780	0.000	—	0.000	—	0.034	0.980	0.138	0.847	0.379	0.825	0.532	0.177	0.967	0.966
GPT-5.6 Sol	0.034	0.880	0.069	0.820	0.069	0.895	0.034	0.970	0.103	0.870	0.414	0.920	0.560	0.215	1.000	0.972
GPT-5.6 Terra	0.138	0.745	0.138	0.770	0.069	0.730	0.310	0.890	0.207	0.770	0.621	0.866	0.624	0.396	1.000	0.913
Grok 4.1 Fast Reasoning	0.138	0.900	0.241	0.907	0.241	0.897	0.103	0.913	0.241	0.894	0.276	0.897	0.587	0.341	0.967	0.947
Grok 4.20 Reasoning	0.069	0.790	0.103	0.740	0.103	0.723	0.069	0.770	0.207	0.718	0.414	0.725	0.580	0.277	1.000	0.754
Kimi K2.6	0.034	0.920	0.034	0.920	0.069	0.885	0.310	0.889	0.276	0.920	0.414	0.908	0.578	0.317	0.967	0.922
QVQ-Max-Latest	0.276	0.750	0.345	0.730	0.276	0.769	0.241	0.750	0.414	0.758	0.379	0.768	0.511	0.469	0.700	0.919
Qwen3.6-35B-A3B	0.069	0.950	0.207	0.920	0.241	0.909	0.345	0.917	0.345	0.923	0.655	0.905	0.639	0.472	0.967	0.964
Qwen3.6-Plus	0.000	—	0.069	0.935	0.034	0.950	0.103	0.930	0.345	0.923	0.448	0.929	0.583	0.286	1.000	0.954
Qwen3.7-Flash	0.069	0.950	0.034	0.950	0.138	0.950	0.172	0.910	0.276	0.941	0.586	0.951	0.606	0.351	1.000	0.961
Qwen3.7-Max	0.034	0.820	0.069	0.865	0.000	—	0.103	0.850	0.172	0.874	0.310	0.884	0.557	0.206	1.000	0.951
Qwen3.7-Plus	0.000	—	0.000	—	0.000	—	0.034	0.850	0.241	0.899	0.310	0.902	0.549	0.178	1.000	0.949
Specialized Detectors																
CO-SPY [ProGAN]	0.966	1.000	0.069	0.995	0.345	0.876	0.310	0.929	0.966	0.982	1.000	1.000	0.788	0.754	0.967	0.999
CO-SPY [SD-v1.4]	0.207	0.842	0.172	0.839	0.138	0.764	0.345	0.662	0.207	0.747	0.172	0.787	0.403	0.324	0.600	0.835
ForgeLens [ProGAN]	0.793	0.986	0.345	0.872	0.552	0.962	0.931	0.943	1.000	0.986	0.966	0.986	0.849	0.861	0.933	0.978
ForgeLens [GenImage]	1.000	0.885	0.241	0.887	1.000	0.944	1.000	0.903	1.000	0.935	1.000	0.918	0.937	0.933	1.000	0.894
Effort [SD-v1.4]	0.621	0.838	0.586	0.739	0.759	0.817	0.897	0.810	0.931	0.865	0.931	0.870	0.844	0.873	0.900	0.858
Effort [Chameleon]	0.828	0.668	0.897	0.667	0.828	0.680	0.862	0.710	0.897	0.708	0.966	0.679	0.490	0.864	0.100	0.590
IAPL [ProGAN]	0.207	0.901	0.310	0.927	0.103	0.997	0.207	0.719	0.414	0.880	0.448	0.954	0.624	0.438	0.967	0.978
IAPL [SD-v1.4]	1.000	0.905	0.931	0.886	1.000	0.905	1.000	0.946	1.000	0.923	1.000	0.928	0.678	0.942	0.367	0.858

Table 25: **Performance comparison of MLLMs on the Baby Products subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.032	0.700	0.000	—	0.097	0.630	0.065	0.610	0.226	0.689	0.226	0.703	0.545	0.193	0.982	0.783
Gemini 3.1 Pro	0.000	—	0.032	0.720	0.000	—	0.129	0.710	0.000	—	0.226	0.721	0.523	0.121	0.982	0.699
Gemini 3.5 Flash-Lite	0.000	—	0.000	—	0.032	0.950	0.000	—	0.032	0.950	0.032	0.980	0.508	0.032	1.000	0.951
Gemini 3.8 Flash	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.065	0.720	0.505	0.021	1.000	0.665
GPT-6 Astra	0.032	0.810	0.032	0.770	0.065	0.755	0.290	0.810	0.258	0.779	0.452	0.819	0.594	0.317	1.000	0.932
GPT-6 Luna	0.968	0.862	0.968	0.927	1.000	0.919	0.968	0.930	0.935	0.904	1.000	0.935	0.603	0.883	0.232	0.764
GPT-6 Sol	0.129	0.738	0.129	0.730	0.161	0.746	0.226	0.760	0.161	0.784	0.548	0.832	0.613	0.368	1.000	0.917
Grok 4.6	0.032	0.840	0.032	0.860	0.000	—	0.000	—	0.032	0.860	0.129	0.825	0.519	0.073	1.000	0.841
Kimi K3	0.000	—	0.032	0.650	0.000	—	0.000	—	0.032	0.650	0.129	0.715	0.516	0.062	1.000	0.868
Qwen3-VL-Flash	0.000	—	0.000	—	0.000	—	0.032	0.950	0.032	0.920	0.161	0.932	0.501	0.072	0.964	0.973
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.129	0.942	0.511	0.042	1.000	0.977
Qwen3.5-Omni-Flash	0.000	—	0.000	—	0.000	—	0.000	—	0.032	0.920	0.097	0.920	0.502	0.042	0.982	0.956
Qwen3.5-Omni-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.129	0.948	0.511	0.042	1.000	0.970
Qwen3.8-27B	0.065	0.720	0.000	—	0.065	0.700	0.032	0.820	0.097	0.823	0.161	0.776	0.535	0.131	1.000	0.875
Qwen3.8-Flash	0.065	0.900	0.129	0.950	0.065	0.900	0.097	0.960	0.097	0.893	0.194	0.842	0.518	0.190	0.929	0.942
Qwen3.8-Max	0.032	0.720	0.000	—	0.032	0.720	0.032	0.750	0.032	0.780	0.484	0.825	0.542	0.184	0.982	0.862
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.000	—	0.032	0.950	0.032	0.950	0.129	0.933	0.065	0.965	0.419	0.959	0.548	0.202	0.982	0.969
GPT-5.4 mini	0.032	0.870	0.000	—	0.065	0.965	0.097	0.957	0.097	0.923	0.290	0.937	0.548	0.176	1.000	0.909
GPT-5.6 Luna	0.000	—	0.000	—	0.000	—	0.032	0.970	0.000	—	0.258	0.911	0.524	0.092	1.000	0.972
GPT-5.6 Sol	0.000	—	0.000	—	0.032	0.780	0.065	0.845	0.000	—	0.097	0.950	0.516	0.062	1.000	0.978
GPT-5.6 Terra	0.000	—	0.000	—	0.097	0.733	0.355	0.816	0.065	0.880	0.419	0.815	0.578	0.270	1.000	0.932
Grok 4.1 Fast Reasoning	0.032	0.950	0.032	0.950	0.065	0.950	0.065	0.940	0.065	0.915	0.097	0.940	0.512	0.111	0.964	0.951
Grok 4.20 Reasoning	0.032	0.700	0.032	0.800	0.032	0.680	0.032	0.820	0.000	—	0.161	0.706	0.497	0.091	0.946	0.767
Kimi K2.6	0.000	—	0.000	—	0.000	—	0.129	0.885	0.097	0.873	0.226	0.890	0.538	0.140	1.000	0.914
QVQ-Max-Latest	0.226	0.714	0.129	0.750	0.290	0.700	0.161	0.740	0.194	0.758	0.194	0.758	0.448	0.308	0.696	0.942
Qwen3.6-35B-A3B	0.032	0.880	0.161	0.912	0.161	0.918	0.129	0.935	0.129	0.930	0.419	0.895	0.506	0.282	0.839	0.964
Qwen3.6-Plus	0.000	—	0.000	—	0.000	—	0.065	0.885	0.000	—	0.194	0.908	0.504	0.082	0.964	0.955
Qwen3.7-Flash	0.000	—	0.000	—	0.065	0.935	0.065	0.935	0.032	0.880	0.161	0.940	0.509	0.101	0.964	0.959
Qwen3.7-Max	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.129	0.917	0.502	0.042	0.982	0.949
Qwen3.7-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.065	0.785	0.129	0.892	0.498	0.062	0.964	0.950
Specialized Detectors																
CO-SPY [ProGAN]	1.000	0.981	0.065	0.789	0.323	0.791	0.258	0.924	0.935	0.981	1.000	1.000	0.798	0.747	1.000	0.999
CO-SPY [SD-v1.4]	0.452	0.758	0.290	0.772	0.387	0.675	0.323	0.761	0.355	0.710	0.290	0.795	0.550	0.491	0.750	0.811
ForgeLens [ProGAN]	0.774	0.937	0.097	0.950	0.774	0.916	0.871	0.961	0.968	0.977	0.968	0.980	0.862	0.849	0.982	0.974
ForgeLens [GenImage]	0.968	0.912	0.194	0.724	1.000	0.946	1.000	0.913	1.000	0.931	1.000	0.923	0.912	0.920	0.964	0.876
Effort [SD-v1.4]	0.806	0.835	0.710	0.812	0.710	0.833	0.839	0.785	0.903	0.891	0.935	0.947	0.801	0.869	0.786	0.848
Effort [Chameleon]	0.839	0.756	0.871	0.737	0.871	0.750	0.871	0.781	0.871	0.757	0.903	0.751	0.444	0.804	0.018	0.543
IAPL [ProGAN]	0.161	0.933	0.387	0.885	0.097	0.672	0.065	0.969	0.484	0.862	0.581	0.930	0.648	0.456	1.000	0.994
IAPL [SD-v1.4]	0.968	0.893	0.871	0.891	0.903	0.903	1.000	0.943	0.935	0.875	1.000	0.897	0.687	0.893	0.429	0.764

Table 26: **Performance comparison of MLLMs on the Beauty & Personal Care subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.125	0.748	0.250	0.704	0.250	0.680	0.325	0.704	0.450	0.679	0.450	0.736	0.654	0.471	1.000	0.754
Gemini 3.1 Pro	0.100	0.718	0.075	0.717	0.175	0.719	0.200	0.718	0.300	0.714	0.650	0.709	0.625	0.400	1.000	0.707
Gemini 3.5 Flash-Lite	0.025	0.990	0.025	0.950	0.050	0.970	0.075	0.970	0.000	—	0.150	0.975	0.527	0.103	1.000	0.952
Gemini 3.8 Flash	0.000	—	0.000	—	0.075	0.680	0.125	0.690	0.075	0.693	0.200	0.715	0.540	0.147	1.000	0.642
GPT-6 Astra	0.075	0.783	0.075	0.793	0.100	0.787	0.475	0.836	0.450	0.809	0.575	0.827	0.646	0.452	1.000	0.939
GPT-6 Luna	0.875	0.909	1.000	0.940	1.000	0.938	1.000	0.938	0.950	0.918	1.000	0.955	0.510	0.947	0.050	0.840
GPT-6 Sol	0.075	0.770	0.150	0.795	0.125	0.770	0.325	0.821	0.200	0.871	0.525	0.829	0.617	0.378	1.000	0.916
Grok 4.6	0.000	—	0.025	0.740	0.125	0.828	0.075	0.880	0.150	0.812	0.125	0.830	0.517	0.153	0.950	0.842
Kimi K3	0.000	—	0.050	0.715	0.000	—	0.175	0.789	0.125	0.746	0.225	0.831	0.548	0.175	1.000	0.869
Qwen3-VL-Flash	0.025	0.950	0.075	0.950	0.100	0.918	0.250	0.932	0.225	0.937	0.225	0.943	0.575	0.261	1.000	0.976
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.025	0.920	0.025	0.950	0.025	0.950	0.506	0.025	1.000	0.977
Qwen3.5-Omni-Flash	0.000	—	0.050	0.895	0.075	0.893	0.100	0.920	0.125	0.896	0.075	0.897	0.535	0.132	1.000	0.942
Qwen3.5-Omni-Plus	0.000	—	0.000	—	0.025	0.920	0.075	0.920	0.075	0.920	0.125	0.928	0.525	0.095	1.000	0.974
Qwen3.8-27B	0.100	0.770	0.175	0.820	0.175	0.820	0.250	0.804	0.225	0.787	0.325	0.839	0.579	0.344	0.950	0.882
Qwen3.8-Flash	0.150	0.927	0.325	0.912	0.250	0.900	0.450	0.869	0.325	0.910	0.600	0.909	0.650	0.517	0.950	0.946
Qwen3.8-Max	0.125	0.724	0.150	0.798	0.200	0.810	0.350	0.812	0.375	0.843	0.550	0.847	0.646	0.452	1.000	0.870
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.250	0.960	0.175	0.940	0.250	0.940	0.450	0.943	0.525	0.961	0.825	0.967	0.681	0.582	0.950	0.967
GPT-5.4 mini	0.025	0.970	0.000	—	0.050	0.940	0.200	0.934	0.125	0.938	0.325	0.954	0.560	0.216	1.000	0.915
GPT-5.6 Luna	0.000	—	0.075	0.787	0.100	0.825	0.100	0.930	0.175	0.847	0.425	0.866	0.573	0.255	1.000	0.964
GPT-5.6 Sol	0.025	0.760	0.050	0.800	0.050	0.795	0.100	0.895	0.100	0.915	0.325	0.899	0.554	0.195	1.000	0.975
GPT-5.6 Terra	0.100	0.785	0.075	0.770	0.150	0.797	0.625	0.867	0.425	0.829	0.575	0.874	0.662	0.491	1.000	0.936
Grok 4.1 Fast Reasoning	0.100	0.902	0.175	0.920	0.200	0.917	0.275	0.914	0.200	0.924	0.300	0.921	0.529	0.341	0.850	0.947
Grok 4.20 Reasoning	0.075	0.730	0.000	—	0.100	0.710	0.125	0.788	0.100	0.743	0.175	0.704	0.548	0.175	1.000	0.760
Kimi K2.6	0.050	0.920	0.075	0.897	0.075	0.873	0.150	0.885	0.250	0.902	0.475	0.898	0.590	0.304	1.000	0.919
QVQ-Max-Latest	0.325	0.758	0.225	0.767	0.325	0.742	0.375	0.750	0.450	0.742	0.450	0.756	0.454	0.513	0.550	0.909
Qwen3.6-35B-A3B	0.200	0.929	0.425	0.921	0.350	0.919	0.425	0.922	0.450	0.900	0.650	0.930	0.683	0.587	0.950	0.967
Qwen3.6-Plus	0.075	0.917	0.175	0.929	0.250	0.926	0.250	0.904	0.325	0.918	0.525	0.928	0.608	0.420	0.950	0.953
Qwen3.7-Flash	0.150	0.922	0.150	0.940	0.250	0.932	0.375	0.931	0.175	0.943	0.550	0.952	0.637	0.431	1.000	0.963
Qwen3.7-Max	0.050	0.850	0.125	0.864	0.050	0.875	0.250	0.909	0.300	0.872	0.375	0.930	0.596	0.322	1.000	0.952
Qwen3.7-Plus	0.000	—	0.075	0.890	0.050	0.850	0.250	0.867	0.200	0.875	0.375	0.900	0.579	0.273	1.000	0.952
Specialized Detectors																
CO-SPY [ProGAN]	0.925	0.963	0.150	0.808	0.275	0.911	0.225	0.831	0.875	0.989	1.000	1.000	0.787	0.730	1.000	1.000
CO-SPY [SD-v1.4]	0.300	0.851	0.250	0.758	0.250	0.855	0.250	0.905	0.300	0.838	0.375	0.826	0.519	0.439	0.750	0.920
ForgeLens [ProGAN]	0.675	0.954	0.175	0.854	0.575	0.888	0.875	0.959	0.675	0.946	0.875	0.952	0.796	0.780	0.950	0.946
ForgeLens [GenImage]	0.975	0.896	0.225	0.732	1.000	0.943	1.000	0.908	1.000	0.928	1.000	0.923	0.933	0.929	1.000	0.857
Effort [SD-v1.4]	0.800	0.820	0.675	0.770	0.650	0.799	0.750	0.849	0.925	0.892	1.000	0.881	0.875	0.887	0.950	0.809
Effort [Chameleon]	0.925	0.753	0.850	0.763	0.900	0.771	0.925	0.794	0.875	0.780	0.925	0.778	0.475	0.909	0.050	0.574
IAPL [ProGAN]	0.375	0.870	0.375	0.958	0.175	0.916	0.150	0.930	0.450	0.886	0.700	0.945	0.685	0.541	1.000	0.997
IAPL [SD-v1.4]	1.000	0.898	0.925	0.883	0.975	0.892	1.000	0.943	1.000	0.911	1.000	0.897	0.817	0.977	0.650	0.879

Table 27: **Performance comparison of MLLMs on the Books subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.133	0.725	0.267	0.647	0.267	0.708	0.267	0.750	0.467	0.723	0.600	0.752	0.667	0.500	1.000	0.802
Gemini 3.1 Pro	0.133	0.725	0.267	0.693	0.133	0.750	0.267	0.725	0.533	0.721	0.533	0.739	0.656	0.475	1.000	0.712
Gemini 3.5 Flash-Lite	0.267	0.968	0.267	0.960	0.400	0.975	0.200	0.973	0.267	0.980	0.533	0.973	0.646	0.483	0.970	0.955
Gemini 3.8 Flash	0.067	0.720	0.000	—	0.000	—	0.067	0.720	0.067	0.720	0.400	0.737	0.550	0.182	1.000	0.643
GPT-6 Astra	0.067	0.910	0.133	0.835	0.000	—	0.200	0.900	0.533	0.821	0.533	0.905	0.622	0.393	1.000	0.942
GPT-6 Luna	0.667	0.870	0.467	0.884	0.667	0.916	0.600	0.934	0.933	0.936	0.933	0.953	0.583	0.744	0.455	0.825
GPT-6 Sol	0.067	0.820	0.200	0.817	0.200	0.857	0.467	0.869	0.333	0.868	0.667	0.895	0.661	0.487	1.000	0.923
Grok 4.6	0.067	0.930	0.000	—	0.133	0.830	0.200	0.917	0.267	0.875	0.533	0.871	0.600	0.333	1.000	0.854
Kimi K3	0.067	0.920	0.000	—	0.000	—	0.133	0.900	0.200	0.843	0.467	0.844	0.572	0.252	1.000	0.878
Qwen3-VL-Flash	0.000	—	0.067	0.950	0.067	0.950	0.133	0.950	0.200	0.950	0.200	0.950	0.556	0.200	1.000	0.980
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.500	0.000	1.000	0.980
Qwen3.5-Omni-Flash	0.000	—	0.133	0.920	0.000	—	0.133	0.920	0.067	0.870	0.133	0.920	0.539	0.144	1.000	0.965
Qwen3.5-Omni-Plus	0.133	0.945	0.133	0.935	0.133	0.935	0.133	0.965	0.267	0.942	0.200	0.950	0.583	0.286	1.000	0.972
Qwen3.8-27B	0.067	0.930	0.067	0.930	0.200	0.843	0.267	0.880	0.267	0.935	0.467	0.860	0.596	0.360	0.970	0.897
Qwen3.8-Flash	0.067	0.920	0.133	0.875	0.267	0.897	0.400	0.955	0.333	0.968	0.800	0.931	0.636	0.492	0.939	0.945
Qwen3.8-Max	0.267	0.843	0.467	0.834	0.200	0.840	0.467	0.823	0.667	0.825	0.733	0.898	0.733	0.636	1.000	0.880
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.267	0.968	0.467	0.946	0.400	0.952	0.533	0.971	0.667	0.972	0.867	0.980	0.767	0.696	1.000	0.971
GPT-5.4 mini	0.067	0.930	0.000	—	0.067	0.840	0.000	—	0.067	0.940	0.133	0.980	0.528	0.105	1.000	0.939
GPT-5.6 Luna	0.000	—	0.067	0.980	0.000	—	0.067	0.960	0.333	0.884	0.333	0.964	0.567	0.235	1.000	0.973
GPT-5.6 Sol	0.000	—	0.067	0.940	0.067	0.990	0.067	0.990	0.133	0.965	0.467	0.974	0.567	0.235	1.000	0.975
GPT-5.6 Terra	0.067	0.670	0.133	0.940	0.067	0.780	0.400	0.930	0.400	0.900	0.733	0.924	0.650	0.462	1.000	0.942
Grok 4.1 Fast Reasoning	0.200	0.937	0.133	0.920	0.067	0.950	0.200	0.927	0.400	0.947	0.067	0.950	0.528	0.291	0.879	0.943
Grok 4.20 Reasoning	0.067	0.850	0.067	0.820	0.067	0.650	0.267	0.805	0.067	0.720	0.333	0.714	0.557	0.250	0.970	0.781
Kimi K2.6	0.000	—	0.200	0.897	0.200	0.873	0.133	0.900	0.467	0.890	0.533	0.894	0.628	0.407	1.000	0.902
QVQ-Max-Latest	0.000	—	0.133	0.750	0.000	—	0.133	0.750	0.133	0.675	0.200	0.750	0.459	0.171	0.818	0.939
Qwen3.6-35B-A3B	0.267	0.935	0.133	0.950	0.200	0.950	0.333	0.962	0.333	0.944	0.467	0.956	0.644	0.448	1.000	0.977
Qwen3.6-Plus	0.133	0.950	0.200	0.917	0.267	0.950	0.267	0.958	0.533	0.946	0.600	0.937	0.621	0.488	0.909	0.955
Qwen3.7-Flash	0.067	0.950	0.067	0.850	0.067	0.950	0.200	0.973	0.267	0.950	0.667	0.951	0.596	0.360	0.970	0.964
Qwen3.7-Max	0.067	0.950	0.067	0.950	0.067	0.850	0.133	0.950	0.200	0.917	0.267	0.917	0.567	0.235	1.000	0.954
Qwen3.7-Plus	0.067	0.950	0.133	0.900	0.067	0.950	0.200	0.940	0.133	0.900	0.333	0.940	0.578	0.269	1.000	0.956
Specialized Detectors																
CO-SPY [ProGAN]	1.000	0.993	0.267	0.747	0.333	0.872	0.000	—	1.000	0.981	0.933	1.000	0.794	0.741	1.000	1.000
CO-SPY [SD-v1.4]	0.400	0.857	0.533	0.804	0.600	0.741	0.600	0.797	0.400	0.744	0.533	0.755	0.619	0.634	0.727	0.855
ForgeLens [ProGAN]	0.400	0.943	0.067	0.885	0.200	0.811	0.667	0.900	0.333	0.956	0.733	0.973	0.685	0.567	0.970	0.978
ForgeLens [GenImage]	1.000	0.865	0.133	0.767	1.000	0.944	1.000	0.903	1.000	0.932	1.000	0.918	0.882	0.906	0.909	0.834
Effort [SD-v1.4]	0.933	0.863	0.933	0.846	0.800	0.835	0.933	0.819	0.933	0.904	1.000	0.945	0.855	0.922	0.788	0.750
Effort [Chameleon]	0.800	0.674	0.867	0.699	0.933	0.670	0.867	0.737	0.933	0.738	0.867	0.724	0.454	0.786	0.030	0.515
IAPL [ProGAN]	0.067	0.516	0.200	0.717	0.067	0.511	0.067	0.510	0.200	0.856	0.333	0.877	0.578	0.269	1.000	0.999
IAPL [SD-v1.4]	0.933	0.821	0.800	0.766	0.800	0.800	1.000	0.919	1.000	0.777	1.000	0.862	0.673	0.865	0.424	0.828

Table 28: **Performance comparison of MLLMs on the CDs & Vinyl subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.053	0.620	0.211	0.720	0.421	0.776	0.316	0.787	0.632	0.664	0.579	0.692	0.639	0.525	0.909	0.757
Gemini 3.1 Pro	0.105	0.735	0.053	0.720	0.211	0.718	0.368	0.726	0.263	0.730	0.263	0.720	0.605	0.348	1.000	0.702
Gemini 3.5 Flash-Lite	0.000	—	0.000	—	0.105	0.970	0.053	0.990	0.158	0.960	0.105	0.970	0.535	0.131	1.000	0.953
Gemini 3.8 Flash	0.000	—	0.000	—	0.053	0.760	0.211	0.725	0.053	0.720	0.158	0.720	0.539	0.146	1.000	0.642
GPT-6 Astra	0.000	—	0.053	0.870	0.158	0.823	0.158	0.843	0.316	0.825	0.211	0.815	0.575	0.260	1.000	0.937
GPT-6 Luna	0.526	0.906	0.789	0.901	0.895	0.902	0.737	0.921	0.684	0.872	0.737	0.918	0.512	0.728	0.295	0.787
GPT-6 Sol	0.000	—	0.053	0.840	0.000	—	0.211	0.900	0.105	0.820	0.263	0.804	0.541	0.189	0.977	0.922
Grok 4.6	0.105	0.855	0.000	—	0.211	0.877	0.211	0.875	0.158	0.793	0.105	0.760	0.566	0.233	1.000	0.850
Kimi K3	0.000	—	0.000	—	0.000	—	0.211	0.795	0.105	0.760	0.105	0.875	0.535	0.131	1.000	0.877
Qwen3-VL-Flash	0.053	0.950	0.105	0.920	0.000	—	0.053	0.950	0.211	0.950	0.105	0.950	0.521	0.159	0.955	0.978
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.489	0.000	0.977	0.979
Qwen3.5-Omni-Flash	0.000	—	0.000	—	0.000	—	0.000	—	0.158	0.920	0.053	0.870	0.495	0.067	0.955	0.954
Qwen3.5-Omni-Plus	0.053	0.960	0.000	—	0.053	0.920	0.053	0.950	0.053	0.960	0.263	0.924	0.539	0.146	1.000	0.977
Qwen3.8-27B	0.105	0.860	0.053	0.880	0.105	0.905	0.263	0.858	0.105	0.850	0.105	0.865	0.539	0.215	0.955	0.872
Qwen3.8-Flash	0.211	0.952	0.211	0.883	0.368	0.951	0.474	0.924	0.526	0.903	0.737	0.936	0.688	0.585	0.955	0.940
Qwen3.8-Max	0.000	—	0.000	—	0.316	0.828	0.474	0.840	0.316	0.863	0.526	0.820	0.625	0.425	0.977	0.870
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.000	—	0.105	0.980	0.263	0.966	0.474	0.971	0.526	0.962	0.579	0.975	0.640	0.484	0.955	0.974
GPT-5.4 mini	0.000	—	0.000	—	0.000	—	0.000	—	0.053	0.980	0.000	—	0.493	0.017	0.977	0.928
GPT-5.6 Luna	0.000	—	0.053	0.970	0.053	0.980	0.053	0.900	0.053	0.970	0.105	0.975	0.515	0.099	0.977	0.969
GPT-5.6 Sol	0.000	—	0.000	—	0.053	0.720	0.053	0.980	0.053	0.860	0.053	0.990	0.518	0.068	1.000	0.976
GPT-5.6 Terra	0.000	—	0.000	—	0.105	0.770	0.158	0.950	0.158	0.880	0.263	0.900	0.534	0.202	0.955	0.941
Grok 4.1 Fast Reasoning	0.211	0.943	0.158	0.903	0.211	0.952	0.263	0.924	0.316	0.892	0.211	0.910	0.534	0.354	0.841	0.947
Grok 4.20 Reasoning	0.053	0.720	0.053	0.820	0.211	0.805	0.263	0.758	0.263	0.738	0.158	0.707	0.549	0.279	0.932	0.760
Kimi K2.6	0.000	—	0.053	0.950	0.158	0.930	0.211	0.885	0.211	0.910	0.263	0.890	0.563	0.258	0.977	0.889
QVQ-Max-Latest	0.263	0.750	0.105	0.775	0.368	0.736	0.316	0.808	0.211	0.800	0.316	0.750	0.472	0.380	0.682	0.930
Qwen3.6-35B-A3B	0.211	0.952	0.105	0.920	0.263	0.940	0.474	0.942	0.211	0.942	0.526	0.930	0.581	0.442	0.864	0.964
Qwen3.6-Plus	0.263	0.934	0.211	0.943	0.263	0.940	0.474	0.948	0.526	0.925	0.632	0.938	0.663	0.556	0.932	0.954
Qwen3.7-Flash	0.105	0.950	0.053	0.950	0.053	0.980	0.316	0.972	0.158	0.973	0.474	0.956	0.585	0.321	0.977	0.963
Qwen3.7-Max	0.158	0.917	0.158	0.897	0.158	0.930	0.316	0.900	0.421	0.917	0.421	0.912	0.613	0.422	0.955	0.946
Qwen3.7-Plus	0.053	0.980	0.105	0.935	0.158	0.900	0.421	0.921	0.211	0.900	0.474	0.902	0.618	0.383	1.000	0.943
Specialized Detectors																
CO-SPY [ProGAN]	1.000	1.000	0.053	1.000	0.526	0.867	0.158	0.835	0.842	0.997	1.000	1.000	0.798	0.747	1.000	1.000
CO-SPY [SD-v1.4]	0.053	0.642	0.105	0.512	0.000	—	0.105	0.547	0.105	0.771	0.105	0.674	0.403	0.133	0.727	0.860
ForgeLens [ProGAN]	0.579	0.957	0.263	0.838	0.526	0.886	0.526	0.947	0.789	0.946	0.947	0.976	0.780	0.746	0.955	0.962
ForgeLens [GenImage]	1.000	0.873	0.053	0.770	1.000	0.937	1.000	0.874	1.000	0.928	1.000	0.902	0.921	0.914	1.000	0.868
Effort [SD-v1.4]	0.526	0.686	0.368	0.693	0.526	0.731	0.526	0.774	0.737	0.849	0.684	0.824	0.690	0.688	0.818	0.836
Effort [Chameleon]	0.368	0.639	0.684	0.607	0.737	0.633	0.579	0.671	0.684	0.673	0.737	0.668	0.350	0.634	0.068	0.648
IAPL [ProGAN]	0.474	0.864	0.421	0.922	0.316	0.912	0.263	0.875	0.474	0.940	0.737	0.899	0.712	0.614	0.977	0.984
IAPL [SD-v1.4]	1.000	0.886	0.842	0.898	0.842	0.897	1.000	0.923	0.947	0.908	1.000	0.905	0.776	0.899	0.614	0.806

Table 29: **Performance comparison of MLLMs on the Cell Phones & Accessories subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.087	0.620	0.152	0.683	0.217	0.676	0.326	0.707	0.370	0.711	0.457	0.707	0.598	0.419	0.929	0.766
Gemini 3.1 Pro	0.109	0.712	0.087	0.718	0.217	0.718	0.304	0.713	0.391	0.717	0.543	0.715	0.638	0.432	1.000	0.707
Gemini 3.5 Flash-Lite	0.022	0.990	0.022	0.950	0.087	0.958	0.065	0.977	0.152	0.964	0.239	0.974	0.549	0.178	1.000	0.946
Gemini 3.8 Flash	0.043	0.700	0.043	0.680	0.065	0.713	0.152	0.703	0.261	0.706	0.391	0.711	0.580	0.275	1.000	0.642
GPT-6 Astra	0.174	0.741	0.239	0.767	0.326	0.807	0.478	0.842	0.587	0.862	0.630	0.847	0.703	0.577	1.000	0.925
GPT-6 Luna	0.696	0.876	0.870	0.909	0.913	0.906	0.913	0.917	0.891	0.900	0.913	0.929	0.504	0.868	0.143	0.717
GPT-6 Sol	0.109	0.748	0.087	0.772	0.239	0.759	0.413	0.854	0.435	0.830	0.565	0.861	0.654	0.471	1.000	0.902
Grok 4.6	0.043	0.885	0.087	0.757	0.130	0.795	0.304	0.832	0.326	0.819	0.391	0.825	0.595	0.351	0.976	0.840
Kimi K3	0.022	0.820	0.000	—	0.022	0.650	0.152	0.794	0.196	0.782	0.261	0.814	0.554	0.196	1.000	0.863
Qwen3-VL-Flash	0.087	0.942	0.130	0.950	0.304	0.937	0.348	0.944	0.391	0.940	0.500	0.941	0.635	0.453	0.976	0.971
Qwen3-VL-Plus	0.022	0.920	0.000	—	0.043	0.950	0.109	0.956	0.087	0.965	0.109	0.944	0.531	0.116	1.000	0.977
Qwen3.5-Omni-Flash	0.043	0.880	0.087	0.907	0.109	0.920	0.283	0.905	0.261	0.910	0.239	0.909	0.585	0.291	1.000	0.951
Qwen3.5-Omni-Plus	0.022	0.920	0.043	0.920	0.087	0.920	0.152	0.937	0.261	0.942	0.304	0.936	0.572	0.253	1.000	0.969
Qwen3.8-27B	0.130	0.807	0.174	0.850	0.174	0.821	0.217	0.863	0.261	0.858	0.413	0.860	0.614	0.372	1.000	0.857
Qwen3.8-Flash	0.087	0.932	0.239	0.927	0.370	0.908	0.522	0.875	0.500	0.885	0.565	0.895	0.666	0.548	0.952	0.933
Qwen3.8-Max	0.174	0.820	0.217	0.803	0.239	0.838	0.457	0.855	0.500	0.871	0.630	0.857	0.685	0.540	1.000	0.866
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.239	0.936	0.261	0.959	0.370	0.954	0.500	0.962	0.543	0.950	0.761	0.968	0.723	0.617	1.000	0.964
GPT-5.4 mini	0.065	0.933	0.043	0.895	0.109	0.954	0.109	0.958	0.174	0.941	0.130	0.943	0.553	0.190	1.000	0.914
GPT-5.6 Luna	0.000	—	0.022	0.780	0.087	0.775	0.087	0.917	0.217	0.867	0.348	0.864	0.552	0.224	0.976	0.955
GPT-5.6 Sol	0.022	0.720	0.022	0.760	0.109	0.876	0.217	0.888	0.239	0.887	0.304	0.943	0.576	0.264	1.000	0.973
GPT-5.6 Terra	0.043	0.850	0.022	0.830	0.174	0.782	0.326	0.865	0.261	0.799	0.457	0.860	0.595	0.351	0.976	0.922
Grok 4.1 Fast Reasoning	0.130	0.923	0.174	0.916	0.196	0.918	0.370	0.909	0.370	0.924	0.304	0.896	0.605	0.407	0.952	0.951
Grok 4.20 Reasoning	0.130	0.732	0.109	0.738	0.152	0.740	0.326	0.721	0.326	0.721	0.370	0.734	0.618	0.381	1.000	0.769
Kimi K2.6	0.087	0.927	0.087	0.920	0.130	0.908	0.196	0.897	0.304	0.922	0.304	0.905	0.580	0.311	0.976	0.902
QVQ-Max-Latest	0.130	0.742	0.217	0.735	0.348	0.766	0.391	0.764	0.370	0.762	0.326	0.763	0.553	0.448	0.810	0.925
Qwen3.6-35B-A3B	0.174	0.904	0.217	0.916	0.348	0.894	0.565	0.924	0.413	0.922	0.609	0.906	0.599	0.547	0.810	0.967
Qwen3.6-Plus	0.109	0.932	0.065	0.930	0.261	0.923	0.348	0.933	0.413	0.917	0.522	0.929	0.643	0.445	1.000	0.952
Qwen3.7-Flash	0.065	0.950	0.065	0.940	0.196	0.944	0.326	0.953	0.283	0.953	0.457	0.948	0.616	0.376	1.000	0.957
Qwen3.7-Max	0.000	—	0.043	0.850	0.087	0.875	0.239	0.881	0.348	0.928	0.370	0.906	0.591	0.307	1.000	0.946
Qwen3.7-Plus	0.000	—	0.065	0.867	0.109	0.910	0.174	0.897	0.261	0.907	0.283	0.900	0.574	0.259	1.000	0.944
Specialized Detectors																
CO-SPY [ProGAN]	0.978	0.997	0.043	0.998	0.196	0.931	0.043	0.763	0.978	0.967	1.000	0.995	0.770	0.701	1.000	1.000
CO-SPY [SD-v1.4]	0.500	0.780	0.500	0.758	0.500	0.726	0.543	0.795	0.500	0.782	0.500	0.694	0.516	0.642	0.524	0.790
ForgeLens [ProGAN]	0.565	0.971	0.109	0.926	0.500	0.886	0.674	0.937	0.717	0.963	0.913	0.974	0.754	0.729	0.929	0.966
ForgeLens [GenImage]	1.000	0.887	0.109	0.723	1.000	0.938	1.000	0.890	1.000	0.925	1.000	0.900	0.878	0.913	0.905	0.844
Effort [SD-v1.4]	0.848	0.838	0.652	0.768	0.870	0.821	0.935	0.841	0.957	0.898	0.957	0.890	0.851	0.918	0.833	0.804
Effort [Chameleon]	0.891	0.754	0.891	0.743	0.913	0.766	0.935	0.793	0.891	0.775	0.935	0.762	0.479	0.885	0.048	0.566
IAPL [ProGAN]	0.174	0.776	0.196	0.884	0.087	0.852	0.130	0.894	0.326	0.923	0.543	0.899	0.621	0.391	1.000	0.994
IAPL [SD-v1.4]	0.935	0.870	0.826	0.812	0.826	0.829	1.000	0.933	0.870	0.881	1.000	0.900	0.693	0.914	0.476	0.745

Table 30: **Performance comparison of MLLMs on the Clothing, Shoes & Jewelry subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.128	0.753	0.191	0.694	0.213	0.728	0.149	0.659	0.362	0.736	0.511	0.681	0.608	0.410	0.957	0.779
Gemini 3.1 Pro	0.085	0.718	0.085	0.715	0.128	0.708	0.128	0.720	0.340	0.702	0.362	0.714	0.594	0.316	1.000	0.682
Gemini 3.5 Flash-Lite	0.064	0.950	0.021	0.950	0.043	0.950	0.021	0.950	0.149	0.950	0.106	0.964	0.534	0.126	1.000	0.957
Gemini 3.8 Flash	0.021	0.680	0.000	—	0.064	0.707	0.021	0.720	0.000	—	0.128	0.693	0.520	0.075	1.000	0.692
GPT-6 Astra	0.085	0.755	0.064	0.837	0.064	0.843	0.213	0.829	0.170	0.797	0.468	0.812	0.589	0.301	1.000	0.939
GPT-6 Luna	0.723	0.912	0.894	0.907	0.979	0.900	0.851	0.900	0.830	0.884	0.936	0.914	0.587	0.902	0.304	0.740
GPT-6 Sol	0.085	0.772	0.106	0.804	0.085	0.792	0.085	0.847	0.128	0.802	0.213	0.774	0.559	0.210	1.000	0.935
Grok 4.6	0.000	—	0.043	0.860	0.085	0.765	0.000	—	0.043	0.760	0.191	0.782	0.530	0.114	1.000	0.840
Kimi K3	0.000	—	0.085	0.725	0.064	0.653	0.021	0.720	0.149	0.739	0.213	0.751	0.544	0.163	1.000	0.866
Qwen3-VL-Flash	0.149	0.924	0.085	0.900	0.191	0.902	0.213	0.925	0.340	0.918	0.383	0.922	0.613	0.370	1.000	0.977
Qwen3-VL-Plus	0.000	—	0.000	—	0.043	0.935	0.021	0.950	0.043	0.935	0.149	0.929	0.521	0.082	1.000	0.979
Qwen3.5-Omni-Flash	0.043	0.905	0.000	—	0.064	0.920	0.085	0.913	0.085	0.895	0.128	0.912	0.534	0.126	1.000	0.959
Qwen3.5-Omni-Plus	0.000	—	0.000	—	0.043	0.920	0.021	0.920	0.021	0.920	0.085	0.930	0.514	0.055	1.000	0.973
Qwen3.8-27B	0.191	0.847	0.128	0.840	0.149	0.810	0.085	0.810	0.170	0.780	0.234	0.855	0.580	0.275	1.000	0.876
Qwen3.8-Flash	0.191	0.940	0.277	0.908	0.234	0.909	0.234	0.916	0.298	0.874	0.255	0.907	0.624	0.398	1.000	0.950
Qwen3.8-Max	0.106	0.778	0.128	0.785	0.170	0.779	0.213	0.803	0.255	0.824	0.426	0.851	0.608	0.356	1.000	0.864
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.149	0.959	0.106	0.956	0.170	0.969	0.149	0.954	0.319	0.926	0.468	0.938	0.613	0.370	1.000	0.970
GPT-5.4 mini	0.170	0.906	0.021	0.780	0.106	0.898	0.064	0.893	0.106	0.938	0.340	0.946	0.546	0.237	0.957	0.925
GPT-5.6 Luna	0.043	0.870	0.021	0.940	0.064	0.830	0.043	0.820	0.085	0.805	0.298	0.864	0.546	0.169	1.000	0.969
GPT-5.6 Sol	0.021	0.910	0.000	—	0.021	0.910	0.043	0.925	0.064	0.843	0.170	0.867	0.527	0.101	1.000	0.979
GPT-5.6 Terra	0.149	0.824	0.043	0.860	0.064	0.857	0.319	0.816	0.234	0.802	0.340	0.821	0.596	0.321	1.000	0.933
Grok 4.1 Fast Reasoning	0.085	0.892	0.149	0.920	0.064	0.933	0.106	0.918	0.149	0.917	0.149	0.920	0.559	0.210	1.000	0.945
Grok 4.20 Reasoning	0.043	0.680	0.043	0.730	0.043	0.750	0.106	0.710	0.043	0.680	0.191	0.701	0.539	0.145	1.000	0.788
Kimi K2.6	0.064	0.897	0.043	0.850	0.043	0.920	0.149	0.850	0.106	0.864	0.298	0.907	0.559	0.210	1.000	0.890
QVQ-Max-Latest	0.213	0.775	0.213	0.730	0.362	0.756	0.383	0.744	0.340	0.766	0.426	0.762	0.553	0.481	0.783	0.939
Qwen3.6-35B-A3B	0.213	0.899	0.255	0.905	0.340	0.925	0.234	0.911	0.447	0.920	0.447	0.923	0.574	0.483	0.826	0.965
Qwen3.6-Plus	0.064	0.907	0.064	0.920	0.149	0.937	0.106	0.924	0.191	0.907	0.170	0.943	0.562	0.221	1.000	0.950
Qwen3.7-Flash	0.128	0.938	0.149	0.924	0.170	0.943	0.170	0.946	0.234	0.934	0.362	0.932	0.601	0.336	1.000	0.953
Qwen3.7-Max	0.064	0.883	0.043	0.950	0.106	0.870	0.043	0.900	0.170	0.905	0.191	0.898	0.551	0.186	1.000	0.950
Qwen3.7-Plus	0.106	0.924	0.064	0.890	0.128	0.883	0.106	0.924	0.191	0.889	0.213	0.879	0.567	0.237	1.000	0.943
Specialized Detectors																
CO-SPY [ProGAN]	0.957	0.979	0.064	1.000	0.277	0.888	0.298	0.803	0.936	0.978	1.000	1.000	0.773	0.739	0.957	0.991
CO-SPY [SD-v1.4]	0.128	0.698	0.064	0.662	0.106	0.740	0.128	0.656	0.213	0.690	0.064	0.690	0.472	0.207	0.826	0.829
ForgeLens [ProGAN]	0.702	0.949	0.340	0.854	0.553	0.904	0.830	0.942	0.957	0.988	0.894	0.970	0.856	0.832	1.000	0.976
ForgeLens [GenImage]	1.000	0.893	0.362	0.746	1.000	0.940	1.000	0.903	1.000	0.938	0.979	0.928	0.923	0.940	0.957	0.899
Effort [SD-v1.4]	0.745	0.838	0.404	0.774	0.702	0.816	0.745	0.883	0.936	0.918	0.915	0.918	0.805	0.846	0.870	0.800
Effort [Chameleon]	0.851	0.755	0.872	0.764	0.894	0.770	0.872	0.781	0.830	0.742	0.957	0.756	0.505	0.902	0.130	0.736
IAPL [ProGAN]	0.106	0.810	0.191	0.772	0.085	0.787	0.106	0.762	0.255	0.947	0.404	0.804	0.552	0.320	0.913	0.972
IAPL [SD-v1.4]	0.936	0.897	0.766	0.797	0.936	0.831	0.979	0.927	0.894	0.882	1.000	0.887	0.742	0.940	0.565	0.846

Table 31: **Performance comparison of MLLMs on the Delivery, Pickup & Dine-Out subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.040	0.758	0.070	0.689	0.070	0.651	0.080	0.657	0.150	0.689	0.210	0.743	0.537	0.186	0.970	0.772
Gemini 3.1 Pro	0.080	0.716	0.120	0.721	0.090	0.712	0.280	0.719	0.290	0.717	0.620	0.714	0.593	0.393	0.940	0.700
Gemini 3.5 Flash-Lite	0.010	0.950	0.060	0.950	0.030	0.960	0.120	0.968	0.060	0.955	0.290	0.969	0.537	0.173	0.980	0.949
Gemini 3.8 Flash	0.010	0.720	0.000	—	0.010	0.720	0.020	0.720	0.010	0.730	0.120	0.704	0.514	0.055	1.000	0.666
GPT-6 Astra	0.040	0.780	0.110	0.787	0.070	0.787	0.280	0.804	0.220	0.804	0.540	0.828	0.585	0.345	0.960	0.926
GPT-6 Luna	0.860	0.912	0.990	0.944	0.990	0.947	0.960	0.933	0.810	0.905	1.000	0.949	0.473	0.890	0.010	0.720
GPT-6 Sol	0.010	0.760	0.090	0.786	0.020	0.790	0.100	0.814	0.180	0.792	0.350	0.829	0.547	0.221	0.970	0.910
Grok 4.6	0.010	0.860	0.010	0.740	0.020	0.880	0.030	0.873	0.040	0.820	0.120	0.790	0.519	0.074	1.000	0.838
Kimi K3	0.000	—	0.000	—	0.010	0.860	0.030	0.847	0.010	0.780	0.180	0.768	0.519	0.074	1.000	0.865
Qwen3-VL-Flash	0.000	—	0.020	0.950	0.010	0.950	0.040	0.920	0.060	0.950	0.210	0.939	0.518	0.107	0.980	0.976
Qwen3-VL-Plus	0.000	—	0.020	0.950	0.010	0.920	0.000	—	0.100	0.974	0.230	0.960	0.530	0.113	1.000	0.979
Qwen3.5-Omni-Flash	0.000	—	0.010	0.920	0.000	—	0.020	0.920	0.020	0.935	0.110	0.918	0.513	0.052	1.000	0.955
Qwen3.5-Omni-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.050	0.940	0.504	0.017	1.000	0.973
Qwen3.8-27B	0.060	0.815	0.090	0.803	0.060	0.797	0.150	0.816	0.100	0.794	0.230	0.792	0.552	0.206	0.990	0.874
Qwen3.8-Flash	0.060	0.910	0.180	0.912	0.180	0.896	0.320	0.856	0.170	0.901	0.570	0.926	0.593	0.393	0.940	0.928
Qwen3.8-Max	0.030	0.857	0.160	0.812	0.090	0.829	0.170	0.819	0.180	0.796	0.560	0.842	0.589	0.330	0.980	0.864
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.040	0.940	0.150	0.940	0.120	0.952	0.320	0.932	0.300	0.935	0.680	0.959	0.614	0.421	0.960	0.969
GPT-5.4 mini	0.010	0.930	0.010	0.920	0.000	—	0.010	0.910	0.070	0.924	0.360	0.939	0.533	0.142	0.990	0.916
GPT-5.6 Luna	0.000	—	0.000	—	0.000	—	0.020	0.810	0.010	0.940	0.250	0.840	0.523	0.089	1.000	0.966
GPT-5.6 Sol	0.000	—	0.030	0.790	0.000	—	0.040	0.910	0.070	0.821	0.180	0.887	0.522	0.101	0.990	0.977
GPT-5.6 Terra	0.020	0.695	0.050	0.808	0.010	0.860	0.190	0.809	0.120	0.794	0.290	0.816	0.557	0.204	1.000	0.932
Grok 4.1 Fast Reasoning	0.020	0.900	0.070	0.893	0.070	0.894	0.060	0.910	0.060	0.862	0.090	0.902	0.516	0.116	0.970	0.947
Grok 4.20 Reasoning	0.000	—	0.000	—	0.010	0.550	0.020	0.750	0.000	—	0.060	0.662	0.497	0.029	0.980	0.786
Kimi K2.6	0.000	—	0.020	0.920	0.020	0.900	0.100	0.881	0.090	0.912	0.220	0.906	0.537	0.140	1.000	0.922
QVQ-Max-Latest	0.190	0.734	0.190	0.742	0.210	0.736	0.280	0.732	0.220	0.743	0.410	0.760	0.530	0.390	0.810	0.927
Qwen3.6-35B-A3B	0.100	0.895	0.120	0.917	0.110	0.922	0.220	0.908	0.200	0.907	0.530	0.910	0.572	0.348	0.930	0.965
Qwen3.6-Plus	0.040	0.917	0.050	0.932	0.050	0.912	0.110	0.906	0.090	0.919	0.330	0.916	0.546	0.200	0.980	0.952
Qwen3.7-Flash	0.040	0.952	0.060	0.928	0.070	0.939	0.150	0.940	0.120	0.937	0.450	0.940	0.574	0.258	1.000	0.960
Qwen3.7-Max	0.010	0.920	0.020	0.850	0.020	0.850	0.050	0.890	0.030	0.900	0.230	0.893	0.510	0.113	0.960	0.946
Qwen3.7-Plus	0.030	0.867	0.020	0.850	0.040	0.913	0.070	0.896	0.040	0.843	0.230	0.895	0.531	0.134	0.990	0.948
Specialized Detectors																
CO-SPY [ProGAN]	0.960	0.991	0.370	0.947	0.660	0.940	0.380	0.864	0.960	0.996	0.990	0.999	0.825	0.832	0.930	0.993
CO-SPY [SD-v1.4]	0.070	0.718	0.070	0.631	0.030	0.604	0.070	0.631	0.060	0.693	0.090	0.661	0.453	0.119	0.840	0.869
ForgeLens [ProGAN]	0.830	0.956	0.240	0.837	0.410	0.898	0.890	0.957	0.940	0.972	0.960	0.992	0.846	0.830	0.980	0.958
ForgeLens [GenImage]	0.880	0.878	0.070	0.677	1.000	0.927	0.960	0.871	1.000	0.918	0.990	0.910	0.908	0.899	1.000	0.906
Effort [SD-v1.4]	0.460	0.816	0.340	0.751	0.500	0.800	0.470	0.772	0.740	0.829	0.790	0.841	0.690	0.697	0.830	0.788
Effort [Chameleon]	0.920	0.734	0.910	0.748	0.930	0.763	0.950	0.797	0.810	0.743	0.890	0.767	0.461	0.873	0.020	0.540
IAPL [ProGAN]	0.440	0.903	0.520	0.900	0.200	0.884	0.130	0.867	0.430	0.933	0.730	0.943	0.479	0.551	0.550	0.912
IAPL [SD-v1.4]	0.960	0.906	0.910	0.856	0.900	0.874	0.970	0.930	0.940	0.858	1.000	0.899	0.603	0.915	0.260	0.747

Table 32: **Performance comparison of MLLMs on the Electronics subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.135	0.677	0.308	0.684	0.212	0.696	0.212	0.708	0.423	0.717	0.385	0.709	0.587	0.432	0.895	0.768
Gemini 3.1 Pro	0.096	0.704	0.096	0.700	0.135	0.710	0.173	0.717	0.481	0.711	0.635	0.712	0.621	0.423	0.974	0.695
Gemini 3.5 Flash-Lite	0.077	0.975	0.058	0.973	0.096	0.958	0.154	0.980	0.231	0.966	0.250	0.968	0.572	0.252	1.000	0.950
Gemini 3.8 Flash	0.019	0.700	0.019	0.720	0.096	0.688	0.192	0.710	0.269	0.709	0.250	0.706	0.571	0.247	1.000	0.670
GPT-6 Astra	0.231	0.770	0.231	0.787	0.365	0.790	0.615	0.852	0.577	0.851	0.750	0.846	0.731	0.632	1.000	0.933
GPT-6 Luna	0.808	0.901	0.962	0.931	0.942	0.928	0.962	0.914	0.846	0.906	1.000	0.929	0.631	0.920	0.342	0.757
GPT-6 Sol	0.096	0.746	0.231	0.798	0.288	0.813	0.442	0.835	0.519	0.831	0.615	0.842	0.670	0.534	0.974	0.916
Grok 4.6	0.019	0.740	0.077	0.777	0.096	0.766	0.308	0.806	0.250	0.825	0.365	0.818	0.593	0.314	1.000	0.849
Kimi K3	0.000	—	0.038	0.670	0.000	—	0.096	0.814	0.231	0.751	0.346	0.803	0.559	0.212	1.000	0.872
Qwen3-VL-Flash	0.077	0.942	0.115	0.945	0.288	0.942	0.519	0.942	0.500	0.944	0.500	0.942	0.667	0.500	1.000	0.974
Qwen3-VL-Plus	0.000	—	0.019	0.950	0.077	0.958	0.135	0.971	0.096	0.956	0.212	0.958	0.545	0.165	1.000	0.978
Qwen3.5-Omni-Flash	0.019	0.920	0.019	0.890	0.096	0.914	0.308	0.915	0.192	0.909	0.250	0.912	0.574	0.257	1.000	0.952
Qwen3.5-Omni-Plus	0.000	—	0.077	0.927	0.154	0.932	0.231	0.935	0.346	0.938	0.327	0.938	0.595	0.318	1.000	0.972
Qwen3.8-27B	0.077	0.837	0.135	0.800	0.115	0.810	0.269	0.856	0.346	0.826	0.308	0.856	0.578	0.343	0.947	0.871
Qwen3.8-Flash	0.115	0.947	0.192	0.908	0.288	0.921	0.500	0.844	0.577	0.900	0.635	0.913	0.600	0.547	0.816	0.933
Qwen3.8-Max	0.154	0.800	0.231	0.792	0.288	0.810	0.500	0.843	0.558	0.843	0.712	0.859	0.704	0.579	1.000	0.866
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.212	0.958	0.096	0.950	0.269	0.959	0.423	0.968	0.442	0.962	0.750	0.966	0.683	0.535	1.000	0.966
GPT-5.4 mini	0.058	0.930	0.096	0.928	0.058	0.963	0.212	0.935	0.173	0.948	0.346	0.957	0.579	0.271	1.000	0.926
GPT-5.6 Luna	0.019	0.840	0.038	0.820	0.192	0.876	0.192	0.866	0.231	0.897	0.404	0.881	0.590	0.304	1.000	0.972
GPT-5.6 Sol	0.038	0.770	0.077	0.898	0.135	0.876	0.231	0.930	0.250	0.905	0.365	0.898	0.591	0.309	1.000	0.978
GPT-5.6 Terra	0.135	0.764	0.212	0.812	0.173	0.768	0.442	0.866	0.481	0.824	0.635	0.886	0.673	0.514	1.000	0.943
Grok 4.1 Fast Reasoning	0.096	0.910	0.135	0.931	0.058	0.893	0.058	0.907	0.173	0.906	0.385	0.896	0.510	0.258	0.868	0.949
Grok 4.20 Reasoning	0.058	0.683	0.154	0.696	0.154	0.651	0.288	0.697	0.250	0.690	0.404	0.680	0.596	0.357	0.974	0.775
Kimi K2.6	0.000	—	0.077	0.903	0.077	0.920	0.308	0.904	0.308	0.911	0.231	0.897	0.583	0.286	1.000	0.906
QVQ-Max-Latest	0.192	0.760	0.269	0.746	0.288	0.740	0.308	0.775	0.423	0.768	0.423	0.759	0.553	0.473	0.789	0.937
Qwen3.6-35B-A3B	0.115	0.938	0.288	0.915	0.365	0.911	0.481	0.915	0.615	0.905	0.712	0.922	0.675	0.597	0.921	0.966
Qwen3.6-Plus	0.077	0.927	0.173	0.923	0.192	0.918	0.462	0.923	0.558	0.930	0.596	0.929	0.645	0.508	0.947	0.953
Qwen3.7-Flash	0.038	0.965	0.096	0.950	0.269	0.946	0.385	0.941	0.385	0.940	0.577	0.955	0.620	0.449	0.947	0.957
Qwen3.7-Max	0.019	0.950	0.096	0.858	0.154	0.885	0.346	0.897	0.423	0.907	0.365	0.896	0.617	0.379	1.000	0.943
Qwen3.7-Plus	0.038	0.835	0.135	0.836	0.077	0.855	0.288	0.890	0.442	0.901	0.442	0.896	0.592	0.381	0.947	0.949
Specialized Detectors																
CO-SPY [ProGAN]	1.000	0.985	0.096	0.804	0.269	0.937	0.154	0.853	0.885	0.955	1.000	1.000	0.784	0.724	1.000	0.997
CO-SPY [SD-v1.4]	0.346	0.771	0.365	0.706	0.269	0.757	0.442	0.731	0.462	0.728	0.442	0.694	0.510	0.541	0.632	0.841
ForgeLens [ProGAN]	0.750	0.946	0.038	0.993	0.500	0.953	0.808	0.953	0.865	0.968	0.942	0.971	0.825	0.788	1.000	0.969
ForgeLens [GenImage]	0.942	0.888	0.077	0.763	1.000	0.940	1.000	0.909	1.000	0.924	1.000	0.914	0.918	0.911	1.000	0.900
Effort [SD-v1.4]	0.731	0.797	0.538	0.785	0.654	0.803	0.692	0.847	0.808	0.880	0.846	0.889	0.843	0.830	0.974	0.809
Effort [Chameleon]	0.865	0.710	0.827	0.699	0.904	0.728	0.923	0.761	0.885	0.718	0.923	0.713	0.444	0.884	0.000	—
IAPL [ProGAN]	0.154	0.984	0.173	0.857	0.077	0.903	0.115	0.902	0.327	0.918	0.327	0.913	0.598	0.327	1.000	0.982
IAPL [SD-v1.4]	1.000	0.884	0.962	0.826	0.962	0.872	1.000	0.934	0.981	0.885	1.000	0.906	0.795	0.968	0.605	0.752

Table 33: **Performance comparison of MLLMs on the Grocery & Gourmet Food subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.059	0.690	0.176	0.738	0.176	0.737	0.294	0.709	0.265	0.754	0.559	0.757	0.589	0.398	0.923	0.738
Gemini 3.1 Pro	0.059	0.700	0.029	0.720	0.176	0.727	0.265	0.711	0.265	0.713	0.706	0.708	0.625	0.400	1.000	0.709
Gemini 3.5 Flash-Lite	0.059	0.965	0.059	0.965	0.147	0.964	0.147	0.974	0.206	0.971	0.441	0.969	0.573	0.298	0.969	0.951
Gemini 3.8 Flash	0.000	—	0.000	—	0.029	0.720	0.059	0.730	0.029	0.720	0.206	0.720	0.527	0.102	1.000	0.666
GPT-6 Astra	0.029	0.830	0.088	0.867	0.059	0.885	0.353	0.860	0.235	0.850	0.471	0.860	0.603	0.341	1.000	0.936
GPT-6 Luna	0.971	0.930	0.971	0.959	1.000	0.952	0.971	0.943	0.912	0.924	1.000	0.962	0.508	0.853	0.046	0.780
GPT-6 Sol	0.000	—	0.147	0.778	0.088	0.727	0.176	0.842	0.118	0.895	0.412	0.850	0.578	0.271	1.000	0.904
Grok 4.6	0.147	0.840	0.147	0.850	0.235	0.840	0.441	0.859	0.147	0.838	0.412	0.862	0.620	0.405	0.985	0.844
Kimi K3	0.029	0.920	0.000	—	0.059	0.750	0.147	0.864	0.088	0.693	0.324	0.828	0.546	0.194	0.985	0.861
Qwen3-VL-Flash	0.000	—	0.000	—	0.000	—	0.176	0.950	0.147	0.944	0.353	0.942	0.533	0.200	0.954	0.978
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.029	0.980	0.206	0.946	0.520	0.075	1.000	0.980
Qwen3.5-Omni-Flash	0.029	0.920	0.000	—	0.000	—	0.118	0.900	0.147	0.900	0.206	0.909	0.534	0.153	0.985	0.953
Qwen3.5-Omni-Plus	0.029	0.920	0.000	—	0.029	0.960	0.118	0.943	0.059	0.935	0.147	0.960	0.532	0.120	1.000	0.975
Qwen3.8-27B	0.147	0.836	0.176	0.813	0.294	0.821	0.294	0.866	0.324	0.795	0.529	0.836	0.639	0.453	0.985	0.869
Qwen3.8-Flash	0.088	0.840	0.324	0.895	0.265	0.941	0.529	0.932	0.471	0.921	0.735	0.922	0.647	0.560	0.892	0.938
Qwen3.8-Max	0.088	0.890	0.294	0.817	0.412	0.875	0.529	0.872	0.529	0.835	0.676	0.882	0.688	0.587	0.954	0.872
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.118	0.965	0.059	0.965	0.235	0.980	0.382	0.967	0.382	0.962	0.882	0.975	0.672	0.511	1.000	0.972
GPT-5.4 mini	0.000	—	0.000	—	0.000	—	0.059	0.920	0.029	0.980	0.382	0.930	0.524	0.144	0.969	0.927
GPT-5.6 Luna	0.000	—	0.000	—	0.000	—	0.088	0.857	0.088	0.890	0.324	0.931	0.542	0.154	1.000	0.965
GPT-5.6 Sol	0.000	—	0.000	—	0.000	—	0.088	0.960	0.029	0.990	0.265	0.977	0.532	0.120	1.000	0.976
GPT-5.6 Terra	0.059	0.895	0.059	0.890	0.088	0.800	0.441	0.911	0.235	0.896	0.588	0.911	0.615	0.392	0.985	0.929
Grok 4.1 Fast Reasoning	0.265	0.951	0.265	0.929	0.353	0.931	0.500	0.937	0.382	0.942	0.618	0.938	0.552	0.533	0.708	0.945
Grok 4.20 Reasoning	0.029	0.820	0.000	—	0.118	0.742	0.088	0.753	0.059	0.820	0.324	0.725	0.521	0.183	0.938	0.772
Kimi K2.6	0.029	0.920	0.029	0.850	0.029	0.920	0.324	0.882	0.206	0.890	0.441	0.915	0.588	0.300	1.000	0.927
QVQ-Max-Latest	0.206	0.721	0.235	0.744	0.382	0.735	0.265	0.711	0.176	0.742	0.265	0.739	0.451	0.373	0.646	0.944
Qwen3.6-35B-A3B	0.147	0.912	0.294	0.925	0.235	0.922	0.500	0.948	0.353	0.934	0.647	0.937	0.628	0.519	0.892	0.968
Qwen3.6-Plus	0.059	0.950	0.176	0.928	0.147	0.950	0.559	0.934	0.529	0.923	0.588	0.933	0.672	0.511	1.000	0.957
Qwen3.7-Flash	0.147	0.950	0.265	0.939	0.353	0.953	0.471	0.967	0.294	0.955	0.618	0.963	0.671	0.525	0.985	0.963
Qwen3.7-Max	0.029	0.920	0.088	0.850	0.206	0.923	0.382	0.915	0.147	0.896	0.441	0.916	0.600	0.353	0.985	0.949
Qwen3.7-Plus	0.059	0.900	0.000	—	0.088	0.907	0.294	0.921	0.294	0.915	0.500	0.912	0.595	0.340	0.985	0.948
Specialized Detectors																
CO-SPY [ProGAN]	1.000	0.986	0.206	0.949	0.412	0.944	0.324	0.821	0.853	0.994	1.000	0.993	0.801	0.770	0.969	0.994
CO-SPY [SD-v1.4]	0.059	0.703	0.059	0.609	0.000	—	0.118	0.679	0.176	0.796	0.147	0.724	0.470	0.163	0.846	0.845
ForgeLens [ProGAN]	0.647	0.974	0.088	0.666	0.265	0.884	0.618	0.884	0.853	0.940	0.941	0.947	0.761	0.718	0.954	0.965
ForgeLens [GenImage]	0.971	0.874	0.088	0.657	1.000	0.938	1.000	0.882	1.000	0.920	1.000	0.906	0.906	0.910	0.969	0.878
Effort [SD-v1.4]	0.500	0.751	0.118	0.724	0.353	0.778	0.441	0.776	0.735	0.873	0.794	0.860	0.691	0.643	0.892	0.843
Effort [Chameleon]	0.882	0.682	0.853	0.663	0.824	0.677	0.941	0.701	0.824	0.666	0.941	0.698	0.523	0.819	0.169	0.634
IAPL [ProGAN]	0.147	0.791	0.176	0.814	0.088	0.880	0.088	0.880	0.324	0.896	0.618	0.897	0.612	0.386	0.985	0.983
IAPL [SD-v1.4]	0.912	0.887	0.853	0.825	0.853	0.863	0.941	0.916	0.941	0.894	1.000	0.889	0.735	0.890	0.554	0.787

Table 34: **Performance comparison of MLLMs on the Handmade Products subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.040	0.620	0.060	0.617	0.160	0.639	0.120	0.722	0.240	0.683	0.320	0.722	0.578	0.271	1.000	0.741
Gemini 3.1 Pro	0.060	0.717	0.140	0.700	0.180	0.713	0.180	0.709	0.300	0.708	0.580	0.715	0.620	0.387	1.000	0.708
Gemini 3.5 Flash-Lite	0.000	—	0.000	—	0.040	0.950	0.060	0.950	0.160	0.950	0.160	0.964	0.535	0.131	1.000	0.950
Gemini 3.8 Flash	0.000	—	0.000	—	0.000	—	0.020	0.720	0.040	0.700	0.100	0.704	0.513	0.052	1.000	0.688
GPT-6 Astra	0.020	0.860	0.020	0.720	0.120	0.802	0.360	0.806	0.440	0.815	0.460	0.825	0.618	0.383	1.000	0.918
GPT-6 Luna	0.820	0.902	0.900	0.915	0.960	0.915	0.980	0.916	0.920	0.906	0.980	0.932	0.574	0.917	0.222	0.740
GPT-6 Sol	0.060	0.753	0.000	—	0.140	0.809	0.260	0.808	0.360	0.751	0.420	0.819	0.603	0.343	1.000	0.908
Grok 4.6	0.020	0.810	0.020	0.760	0.020	0.840	0.080	0.823	0.080	0.800	0.060	0.797	0.523	0.089	1.000	0.827
Kimi K3	0.000	—	0.020	0.720	0.020	0.620	0.040	0.835	0.140	0.717	0.220	0.723	0.537	0.137	1.000	0.858
Qwen3-VL-Flash	0.020	0.950	0.040	0.935	0.160	0.915	0.100	0.926	0.200	0.925	0.260	0.938	0.565	0.230	1.000	0.968
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.020	0.950	0.060	0.930	0.060	0.940	0.512	0.046	1.000	0.972
Qwen3.5-Omni-Flash	0.000	—	0.000	—	0.020	0.850	0.020	0.920	0.100	0.910	0.080	0.908	0.518	0.071	1.000	0.944
Qwen3.5-Omni-Plus	0.000	—	0.000	—	0.020	0.950	0.060	0.927	0.080	0.920	0.180	0.926	0.528	0.107	1.000	0.969
Qwen3.8-27B	0.100	0.788	0.060	0.773	0.100	0.796	0.100	0.850	0.180	0.813	0.360	0.799	0.547	0.259	0.944	0.860
Qwen3.8-Flash	0.040	0.900	0.280	0.923	0.180	0.871	0.320	0.841	0.200	0.870	0.380	0.908	0.589	0.376	0.944	0.943
Qwen3.8-Max	0.020	0.700	0.060	0.723	0.180	0.780	0.260	0.791	0.500	0.808	0.600	0.817	0.635	0.425	1.000	0.858
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.140	0.946	0.100	0.962	0.260	0.939	0.300	0.945	0.420	0.956	0.760	0.953	0.665	0.496	1.000	0.965
GPT-5.4 mini	0.040	0.895	0.020	0.910	0.060	0.903	0.100	0.924	0.100	0.894	0.260	0.945	0.548	0.176	1.000	0.916
GPT-5.6 Luna	0.000	—	0.000	—	0.100	0.846	0.120	0.910	0.120	0.842	0.460	0.832	0.567	0.235	1.000	0.966
GPT-5.6 Sol	0.020	0.780	0.000	—	0.060	0.823	0.060	0.873	0.120	0.852	0.200	0.867	0.538	0.142	1.000	0.975
GPT-5.6 Terra	0.080	0.848	0.060	0.850	0.160	0.780	0.440	0.817	0.320	0.799	0.540	0.834	0.619	0.420	0.972	0.934
Grok 4.1 Fast Reasoning	0.060	0.897	0.040	0.835	0.080	0.927	0.100	0.944	0.180	0.899	0.180	0.883	0.526	0.192	0.944	0.948
Grok 4.20 Reasoning	0.040	0.835	0.020	0.720	0.020	0.700	0.060	0.730	0.080	0.682	0.080	0.688	0.525	0.095	1.000	0.755
Kimi K2.6	0.000	—	0.020	0.850	0.060	0.883	0.180	0.873	0.220	0.901	0.280	0.892	0.563	0.225	1.000	0.916
QVQ-Max-Latest	0.260	0.738	0.260	0.731	0.220	0.723	0.260	0.727	0.340	0.721	0.480	0.773	0.443	0.448	0.583	0.905
Qwen3.6-35B-A3B	0.060	0.923	0.120	0.937	0.220	0.911	0.220	0.906	0.300	0.909	0.440	0.908	0.599	0.369	0.972	0.958
Qwen3.6-Plus	0.000	—	0.040	0.920	0.080	0.900	0.080	0.923	0.240	0.930	0.280	0.928	0.546	0.214	0.972	0.950
Qwen3.7-Flash	0.000	—	0.040	0.950	0.100	0.916	0.080	0.960	0.200	0.931	0.360	0.934	0.565	0.230	1.000	0.954
Qwen3.7-Max	0.020	0.900	0.020	0.850	0.040	0.865	0.040	0.925	0.140	0.853	0.160	0.880	0.535	0.131	1.000	0.944
Qwen3.7-Plus	0.000	—	0.000	—	0.000	—	0.100	0.870	0.140	0.889	0.200	0.880	0.523	0.136	0.972	0.947
Specialized Detectors																
CO-SPY [ProGAN]	0.960	0.989	0.120	0.967	0.480	0.894	0.320	0.891	0.960	0.982	1.000	0.996	0.820	0.780	1.000	1.000
CO-SPY [SD-v1.4]	0.440	0.728	0.380	0.729	0.260	0.750	0.340	0.752	0.520	0.733	0.460	0.753	0.381	0.542	0.361	0.772
ForgeLens [ProGAN]	0.780	0.977	0.300	0.887	0.620	0.914	0.860	0.955	0.900	0.962	0.960	0.979	0.841	0.845	0.944	0.981
ForgeLens [GenImage]	1.000	0.919	0.220	0.751	1.000	0.948	0.980	0.913	1.000	0.937	1.000	0.930	0.892	0.924	0.917	0.833
Effort [SD-v1.4]	0.760	0.852	0.500	0.818	0.720	0.830	0.840	0.859	0.840	0.922	0.920	0.929	0.771	0.853	0.778	0.773
Effort [Chameleon]	0.880	0.718	0.940	0.691	0.940	0.712	0.920	0.727	0.860	0.715	0.900	0.700	0.523	0.902	0.139	0.606
IAPL [ProGAN]	0.240	0.861	0.200	0.891	0.080	0.860	0.120	0.864	0.260	0.926	0.520	0.960	0.618	0.383	1.000	0.998
IAPL [SD-v1.4]	0.960	0.907	0.880	0.861	0.900	0.878	1.000	0.931	0.920	0.893	1.000	0.914	0.722	0.942	0.500	0.757

Table 35: **Performance comparison of MLLMs on the Health & Household subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.057	0.710	0.057	0.705	0.114	0.620	0.171	0.707	0.143	0.664	0.314	0.732	0.538	0.249	0.933	0.773
Gemini 3.1 Pro	0.029	0.750	0.057	0.735	0.171	0.722	0.286	0.722	0.171	0.723	0.571	0.721	0.607	0.353	1.000	0.701
Gemini 3.5 Flash-Lite	0.029	0.990	0.000	—	0.000	—	0.171	0.970	0.086	0.950	0.143	0.958	0.536	0.133	1.000	0.950
Gemini 3.8 Flash	0.000	—	0.000	—	0.029	0.680	0.086	0.720	0.143	0.720	0.171	0.707	0.536	0.133	1.000	0.647
GPT-6 Astra	0.057	0.720	0.029	0.790	0.057	0.860	0.429	0.827	0.314	0.819	0.514	0.848	0.617	0.378	1.000	0.921
GPT-6 Luna	0.657	0.900	0.971	0.938	0.943	0.952	0.971	0.935	0.971	0.912	0.971	0.944	0.624	0.932	0.333	0.720
GPT-6 Sol	0.086	0.803	0.029	0.720	0.057	0.810	0.314	0.805	0.200	0.776	0.457	0.851	0.562	0.319	0.933	0.921
Grok 4.6	0.029	0.860	0.057	0.820	0.200	0.841	0.286	0.881	0.114	0.843	0.286	0.829	0.548	0.278	0.933	0.838
Kimi K3	0.000	—	0.057	0.715	0.029	0.620	0.171	0.818	0.029	0.820	0.257	0.776	0.545	0.166	1.000	0.874
Qwen3-VL-Flash	0.086	0.940	0.086	0.950	0.114	0.943	0.371	0.934	0.229	0.923	0.286	0.944	0.564	0.325	0.933	0.971
Qwen3-VL-Plus	0.000	—	0.000	—	0.029	0.950	0.029	0.950	0.000	—	0.029	0.920	0.507	0.028	1.000	0.974
Qwen3.5-Omni-Flash	0.000	—	0.029	0.920	0.086	0.920	0.171	0.910	0.114	0.912	0.171	0.912	0.548	0.174	1.000	0.950
Qwen3.5-Omni-Plus	0.000	—	0.000	—	0.029	0.920	0.086	0.930	0.000	—	0.143	0.940	0.521	0.082	1.000	0.975
Qwen3.8-27B	0.029	0.860	0.029	0.820	0.029	0.880	0.257	0.824	0.086	0.827	0.257	0.803	0.524	0.204	0.933	0.871
Qwen3.8-Flash	0.086	0.873	0.143	0.966	0.286	0.903	0.343	0.906	0.143	0.948	0.514	0.928	0.593	0.402	0.933	0.936
Qwen3.8-Max	0.086	0.727	0.086	0.797	0.200	0.789	0.429	0.875	0.286	0.841	0.543	0.858	0.636	0.427	1.000	0.865
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.029	0.980	0.029	0.980	0.200	0.959	0.314	0.973	0.314	0.967	0.800	0.969	0.640	0.439	1.000	0.964
GPT-5.4 mini	0.000	—	0.057	0.920	0.029	0.980	0.143	0.916	0.057	0.965	0.257	0.959	0.545	0.166	1.000	0.922
GPT-5.6 Luna	0.029	0.940	0.029	0.730	0.029	0.960	0.086	0.877	0.029	0.980	0.314	0.915	0.543	0.158	1.000	0.976
GPT-5.6 Sol	0.057	0.860	0.000	—	0.029	0.990	0.057	0.910	0.029	0.840	0.200	0.913	0.531	0.117	1.000	0.978
GPT-5.6 Terra	0.086	0.783	0.029	0.680	0.114	0.755	0.543	0.856	0.400	0.805	0.686	0.882	0.655	0.473	1.000	0.924
Grok 4.1 Fast Reasoning	0.029	0.920	0.171	0.913	0.114	0.907	0.286	0.903	0.257	0.909	0.429	0.927	0.574	0.352	0.933	0.944
Grok 4.20 Reasoning	0.029	0.820	0.114	0.730	0.143	0.716	0.200	0.713	0.057	0.770	0.229	0.711	0.564	0.228	1.000	0.761
Kimi K2.6	0.029	0.920	0.000	—	0.057	0.850	0.257	0.866	0.200	0.870	0.343	0.891	0.574	0.257	1.000	0.920
QVQ-Max-Latest	0.171	0.725	0.171	0.758	0.257	0.756	0.457	0.750	0.429	0.733	0.429	0.737	0.560	0.479	0.800	0.942
Qwen3.6-35B-A3B	0.143	0.938	0.171	0.915	0.257	0.926	0.457	0.917	0.371	0.918	0.571	0.913	0.498	0.486	0.667	0.958
Qwen3.6-Plus	0.057	0.935	0.143	0.896	0.200	0.929	0.429	0.920	0.171	0.928	0.429	0.932	0.552	0.382	0.867	0.955
Qwen3.7-Flash	0.029	0.980	0.086	0.960	0.229	0.946	0.371	0.951	0.229	0.949	0.600	0.951	0.562	0.406	0.867	0.955
Qwen3.7-Max	0.029	0.950	0.029	0.780	0.057	0.885	0.143	0.944	0.086	0.817	0.343	0.912	0.557	0.205	1.000	0.946
Qwen3.7-Plus	0.029	0.950	0.057	0.900	0.057	0.935	0.229	0.892	0.114	0.845	0.257	0.897	0.562	0.220	1.000	0.948
Specialized Detectors																
CO-SPY [ProGAN]	0.971	0.988	0.029	0.983	0.086	0.876	0.114	0.838	0.743	0.942	1.000	1.000	0.745	0.658	1.000	1.000
CO-SPY [SD-v1.4]	0.371	0.749	0.314	0.678	0.114	0.752	0.314	0.739	0.229	0.715	0.343	0.790	0.374	0.426	0.467	0.816
ForgeLens [ProGAN]	0.657	0.927	0.143	0.776	0.400	0.922	0.829	0.923	0.943	0.961	0.971	0.971	0.829	0.793	1.000	0.981
ForgeLens [GenImage]	0.914	0.903	0.114	0.818	1.000	0.940	0.971	0.898	1.000	0.925	1.000	0.917	0.917	0.909	1.000	0.889
Effort [SD-v1.4]	0.657	0.803	0.457	0.769	0.600	0.745	0.743	0.844	0.943	0.844	0.971	0.876	0.798	0.838	0.867	0.788
Effort [Chameleon]	0.714	0.687	0.714	0.707	0.829	0.710	0.800	0.719	0.857	0.708	0.743	0.715	0.421	0.842	0.067	0.523
IAPL [ProGAN]	0.171	0.855	0.086	0.638	0.029	0.999	0.057	0.848	0.371	0.900	0.743	0.901	0.621	0.391	1.000	0.995
IAPL [SD-v1.4]	1.000	0.890	1.000	0.854	0.914	0.895	1.000	0.943	0.943	0.910	1.000	0.922	0.688	0.967	0.400	0.834

Table 36: **Performance comparison of MLLMs on the Health & Personal Care subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.000	—	0.122	0.710	0.102	0.730	0.122	0.685	0.184	0.681	0.388	0.756	0.547	0.265	0.941	0.730
Gemini 3.1 Pro	0.020	0.750	0.102	0.738	0.082	0.722	0.184	0.726	0.224	0.706	0.551	0.722	0.597	0.325	1.000	0.688
Gemini 3.5 Flash-Lite	0.061	0.963	0.102	0.972	0.041	0.950	0.102	0.964	0.102	0.958	0.143	0.983	0.546	0.168	1.000	0.932
Gemini 3.8 Flash	0.000	—	0.020	0.730	0.000	—	0.041	0.720	0.041	0.700	0.122	0.713	0.519	0.072	1.000	0.696
GPT-6 Astra	0.000	—	0.061	0.750	0.020	0.840	0.204	0.818	0.265	0.808	0.388	0.836	0.578	0.271	1.000	0.929
GPT-6 Luna	0.735	0.906	0.959	0.943	0.918	0.932	0.959	0.941	0.816	0.916	0.959	0.946	0.475	0.916	0.059	0.780
GPT-6 Sol	0.041	0.775	0.000	—	0.020	0.760	0.082	0.823	0.041	0.740	0.327	0.821	0.543	0.157	1.000	0.906
Grok 4.6	0.000	—	0.061	0.783	0.000	—	0.082	0.770	0.061	0.813	0.204	0.800	0.534	0.127	1.000	0.839
Kimi K3	0.000	—	0.000	—	0.000	—	0.102	0.772	0.020	0.620	0.286	0.759	0.534	0.127	1.000	0.845
Qwen3-VL-Flash	0.020	0.920	0.082	0.935	0.061	0.940	0.204	0.938	0.224	0.939	0.286	0.941	0.573	0.255	1.000	0.971
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.041	0.935	0.503	0.014	1.000	0.976
Qwen3.5-Omni-Flash	0.000	—	0.041	0.920	0.061	0.897	0.102	0.914	0.082	0.927	0.143	0.920	0.536	0.133	1.000	0.946
Qwen3.5-Omni-Plus	0.000	—	0.020	0.920	0.000	—	0.041	0.920	0.041	0.920	0.102	0.942	0.517	0.066	1.000	0.968
Qwen3.8-27B	0.020	0.820	0.041	0.760	0.020	0.920	0.082	0.815	0.102	0.764	0.224	0.815	0.541	0.151	1.000	0.872
Qwen3.8-Flash	0.020	0.950	0.224	0.915	0.102	0.890	0.245	0.873	0.184	0.913	0.469	0.864	0.604	0.344	1.000	0.929
Qwen3.8-Max	0.061	0.787	0.143	0.854	0.061	0.790	0.306	0.861	0.286	0.804	0.551	0.859	0.617	0.380	1.000	0.858
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.061	0.960	0.102	0.962	0.082	0.975	0.224	0.958	0.327	0.963	0.592	0.968	0.616	0.376	1.000	0.969
GPT-5.4 mini	0.020	0.740	0.020	0.940	0.000	—	0.102	0.946	0.041	0.960	0.184	0.956	0.531	0.115	1.000	0.912
GPT-5.6 Luna	0.000	—	0.000	—	0.000	—	0.000	—	0.020	0.880	0.122	0.853	0.512	0.047	1.000	0.957
GPT-5.6 Sol	0.020	0.910	0.000	—	0.000	—	0.041	0.825	0.000	—	0.163	0.916	0.519	0.072	1.000	0.975
GPT-5.6 Terra	0.020	0.720	0.020	0.720	0.020	0.860	0.306	0.838	0.102	0.802	0.367	0.829	0.540	0.244	0.941	0.936
Grok 4.1 Fast Reasoning	0.265	0.928	0.286	0.929	0.224	0.942	0.245	0.923	0.388	0.927	0.388	0.928	0.620	0.460	0.941	0.939
Grok 4.20 Reasoning	0.102	0.760	0.041	0.820	0.082	0.730	0.102	0.780	0.102	0.758	0.163	0.710	0.549	0.180	1.000	0.751
Kimi K2.6	0.000	—	0.020	0.950	0.020	0.920	0.122	0.862	0.082	0.850	0.224	0.895	0.539	0.145	1.000	0.925
QVQ-Max-Latest	0.224	0.750	0.306	0.733	0.184	0.756	0.245	0.758	0.204	0.790	0.408	0.753	0.484	0.410	0.706	0.900
Qwen3.6-35B-A3B	0.041	0.935	0.143	0.917	0.041	0.935	0.204	0.946	0.204	0.914	0.449	0.925	0.531	0.304	0.882	0.969
Qwen3.6-Plus	0.041	0.920	0.082	0.943	0.061	0.897	0.204	0.925	0.204	0.929	0.408	0.932	0.583	0.286	1.000	0.948
Qwen3.7-Flash	0.000	—	0.082	0.960	0.041	0.950	0.184	0.954	0.082	0.955	0.388	0.949	0.565	0.229	1.000	0.956
Qwen3.7-Max	0.020	0.850	0.020	0.920	0.000	—	0.061	0.900	0.102	0.898	0.327	0.902	0.544	0.163	1.000	0.942
Qwen3.7-Plus	0.020	0.850	0.000	—	0.041	0.885	0.061	0.917	0.061	0.883	0.286	0.931	0.539	0.145	1.000	0.943
Specialized Detectors																
CO-SPY [ProGAN]	0.980	0.995	0.061	0.935	0.224	0.832	0.204	0.780	0.816	0.964	1.000	0.998	0.774	0.708	1.000	1.000
CO-SPY [SD-v1.4]	0.163	0.824	0.143	0.806	0.143	0.723	0.122	0.722	0.143	0.737	0.143	0.821	0.513	0.249	0.882	0.867
ForgeLens [ProGAN]	0.592	0.958	0.163	0.900	0.449	0.913	0.796	0.929	0.714	0.927	0.918	0.955	0.773	0.753	0.941	0.957
ForgeLens [GenImage]	0.939	0.892	0.184	0.760	1.000	0.941	1.000	0.902	1.000	0.923	0.980	0.913	0.837	0.914	0.824	0.853
Effort [SD-v1.4]	0.571	0.714	0.286	0.736	0.367	0.768	0.571	0.750	0.816	0.838	0.816	0.827	0.786	0.727	1.000	0.880
Effort [Chameleon]	0.939	0.732	0.857	0.739	0.878	0.759	0.898	0.793	0.796	0.768	0.816	0.794	0.432	0.899	0.000	—
IAPL [ProGAN]	0.082	0.893	0.122	0.928	0.041	0.767	0.000	—	0.184	0.925	0.347	0.924	0.565	0.229	1.000	0.999
IAPL [SD-v1.4]	0.939	0.863	0.837	0.830	0.796	0.842	1.000	0.922	0.898	0.865	1.000	0.888	0.838	0.947	0.765	0.747

Table 37: **Performance comparison of MLLMs on the Home & Kitchen subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.138	0.752	0.241	0.651	0.207	0.672	0.310	0.630	0.414	0.697	0.345	0.728	0.628	0.430	0.979	0.754
Gemini 3.1 Pro	0.207	0.727	0.310	0.721	0.345	0.722	0.207	0.720	0.483	0.714	0.586	0.712	0.668	0.523	0.979	0.710
Gemini 3.5 Flash-Lite	0.103	0.960	0.103	0.950	0.172	0.966	0.172	0.966	0.310	0.969	0.172	0.974	0.586	0.294	1.000	0.949
Gemini 3.8 Flash	0.000	—	0.000	—	0.103	0.693	0.000	—	0.207	0.707	0.103	0.700	0.534	0.129	1.000	0.676
GPT-6 Astra	0.138	0.752	0.069	0.810	0.310	0.867	0.379	0.832	0.517	0.844	0.586	0.841	0.667	0.500	1.000	0.922
GPT-6 Luna	0.862	0.890	0.966	0.927	0.931	0.939	0.966	0.923	0.862	0.912	0.966	0.933	0.577	0.866	0.229	0.758
GPT-6 Sol	0.034	0.760	0.138	0.792	0.345	0.858	0.345	0.819	0.379	0.807	0.345	0.823	0.632	0.418	1.000	0.905
Grok 4.6	0.000	—	0.000	—	0.000	—	0.034	0.840	0.241	0.809	0.241	0.814	0.543	0.159	1.000	0.840
Kimi K3	0.000	—	0.034	0.880	0.069	0.790	0.000	—	0.172	0.778	0.138	0.752	0.524	0.128	0.979	0.872
Qwen3-VL-Flash	0.103	0.920	0.138	0.927	0.379	0.930	0.379	0.927	0.517	0.929	0.552	0.941	0.652	0.508	0.958	0.970
Qwen3-VL-Plus	0.000	—	0.000	—	0.138	0.927	0.069	0.950	0.103	0.940	0.241	0.941	0.525	0.167	0.958	0.976
Qwen3.5-Omni-Flash	0.000	—	0.103	0.920	0.172	0.920	0.172	0.900	0.276	0.908	0.207	0.925	0.578	0.269	1.000	0.946
Qwen3.5-Omni-Plus	0.000	—	0.000	—	0.103	0.930	0.069	0.935	0.138	0.953	0.207	0.950	0.543	0.159	1.000	0.968
Qwen3.8-27B	0.103	0.810	0.103	0.833	0.207	0.865	0.103	0.787	0.310	0.840	0.276	0.840	0.571	0.308	0.958	0.869
Qwen3.8-Flash	0.138	0.853	0.241	0.939	0.345	0.935	0.276	0.879	0.448	0.862	0.621	0.887	0.662	0.511	0.979	0.929
Qwen3.8-Max	0.172	0.804	0.172	0.864	0.379	0.844	0.414	0.843	0.586	0.845	0.690	0.870	0.701	0.574	1.000	0.862
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.310	0.967	0.276	0.956	0.379	0.964	0.414	0.951	0.655	0.963	0.690	0.961	0.727	0.625	1.000	0.966
GPT-5.4 mini	0.172	0.956	0.138	0.890	0.103	0.920	0.172	0.942	0.103	0.963	0.483	0.949	0.598	0.327	1.000	0.902
GPT-5.6 Luna	0.000	—	0.034	0.940	0.069	0.910	0.000	—	0.103	0.887	0.310	0.923	0.543	0.159	1.000	0.962
GPT-5.6 Sol	0.000	—	0.034	0.970	0.103	0.923	0.034	0.980	0.310	0.923	0.276	0.901	0.563	0.224	1.000	0.975
GPT-5.6 Terra	0.069	0.750	0.103	0.897	0.172	0.844	0.241	0.850	0.379	0.819	0.483	0.841	0.621	0.389	1.000	0.926
Grok 4.1 Fast Reasoning	0.034	0.820	0.000	—	0.069	0.865	0.103	0.930	0.069	0.915	0.172	0.876	0.527	0.138	0.979	0.946
Grok 4.20 Reasoning	0.069	0.700	0.034	0.680	0.103	0.680	0.172	0.668	0.207	0.688	0.207	0.693	0.566	0.234	1.000	0.757
Kimi K2.6	0.000	—	0.034	0.920	0.138	0.920	0.138	0.920	0.379	0.907	0.207	0.918	0.575	0.260	1.000	0.841
QVQ-Max-Latest	0.310	0.756	0.345	0.770	0.448	0.712	0.448	0.769	0.483	0.754	0.586	0.779	0.541	0.569	0.646	0.916
Qwen3.6-35B-A3B	0.310	0.919	0.310	0.916	0.310	0.918	0.310	0.916	0.586	0.914	0.517	0.918	0.633	0.548	0.875	0.964
Qwen3.6-Plus	0.138	0.933	0.207	0.920	0.207	0.932	0.207	0.913	0.379	0.925	0.621	0.924	0.647	0.453	1.000	0.949
Qwen3.7-Flash	0.172	0.944	0.138	0.958	0.310	0.942	0.241	0.946	0.414	0.954	0.483	0.951	0.636	0.451	0.979	0.957
Qwen3.7-Max	0.069	0.900	0.172	0.890	0.241	0.881	0.207	0.900	0.310	0.908	0.414	0.893	0.618	0.381	1.000	0.946
Qwen3.7-Plus	0.034	0.950	0.138	0.925	0.276	0.846	0.207	0.895	0.276	0.896	0.379	0.881	0.609	0.358	1.000	0.945
Specialized Detectors																
CO-SPY [ProGAN]	0.862	0.992	0.276	0.986	0.414	0.947	0.379	0.990	0.862	0.978	1.000	1.000	0.816	0.775	1.000	1.000
CO-SPY [SD-v1.4]	0.483	0.819	0.345	0.861	0.448	0.835	0.414	0.851	0.345	0.808	0.448	0.796	0.509	0.543	0.604	0.865
ForgeLens [ProGAN]	0.793	0.942	0.448	0.952	0.655	0.907	0.966	0.956	0.931	0.964	1.000	0.991	0.879	0.883	0.958	0.972
ForgeLens [GenImage]	0.966	0.883	0.310	0.761	1.000	0.944	1.000	0.889	1.000	0.927	1.000	0.915	0.908	0.927	0.938	0.874
Effort [SD-v1.4]	0.690	0.868	0.655	0.798	0.828	0.832	0.862	0.868	0.931	0.905	1.000	0.904	0.872	0.894	0.917	0.829
Effort [Chameleon]	0.966	0.787	0.966	0.798	0.966	0.806	0.966	0.834	0.966	0.792	0.966	0.805	0.483	0.862	0.000	—
IAPL [ProGAN]	0.241	0.894	0.345	0.941	0.207	0.907	0.172	0.942	0.517	0.968	0.759	0.953	0.676	0.542	0.979	0.996
IAPL [SD-v1.4]	1.000	0.887	0.897	0.855	0.931	0.893	1.000	0.945	1.000	0.924	1.000	0.929	0.798	0.936	0.625	0.842

Table 38: **Performance comparison of MLLMs on the Hotels & Accommodations subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.033	0.647	0.144	0.661	0.222	0.674	0.300	0.661	0.522	0.727	0.456	0.708	0.635	0.436	0.991	0.784
Gemini 3.1 Pro	0.033	0.740	0.189	0.711	0.211	0.713	0.322	0.710	0.622	0.706	0.589	0.707	0.664	0.494	1.000	0.701
Gemini 3.5 Flash-Lite	0.078	0.936	0.178	0.952	0.156	0.957	0.378	0.963	0.433	0.964	0.622	0.963	0.649	0.470	0.991	0.944
Gemini 3.8 Flash	0.056	0.688	0.189	0.699	0.167	0.701	0.456	0.704	0.644	0.699	0.444	0.703	0.663	0.492	1.000	0.682
GPT-6 Astra	0.189	0.731	0.478	0.765	0.556	0.766	0.878	0.825	0.956	0.832	0.867	0.839	0.818	0.789	0.982	0.908
GPT-6 Luna	0.889	0.882	0.944	0.912	0.967	0.919	0.956	0.930	0.944	0.897	0.978	0.912	0.696	0.919	0.445	0.758
GPT-6 Sol	0.222	0.754	0.400	0.778	0.533	0.783	0.744	0.823	0.789	0.835	0.833	0.834	0.784	0.738	0.982	0.916
Grok 4.6	0.000	—	0.033	0.763	0.078	0.830	0.156	0.803	0.267	0.779	0.411	0.783	0.579	0.272	1.000	0.846
Kimi K3	0.000	—	0.056	0.716	0.022	0.685	0.300	0.763	0.389	0.767	0.478	0.763	0.604	0.344	1.000	0.872
Qwen3-VL-Flash	0.000	—	0.022	0.950	0.100	0.947	0.244	0.937	0.244	0.940	0.322	0.931	0.578	0.269	1.000	0.971
Qwen3-VL-Plus	0.000	—	0.011	0.920	0.011	0.950	0.022	0.950	0.022	0.965	0.078	0.946	0.512	0.047	1.000	0.977
Qwen3.5-Omni-Flash	0.000	—	0.033	0.910	0.067	0.910	0.122	0.922	0.156	0.926	0.133	0.917	0.543	0.157	1.000	0.955
Qwen3.5-Omni-Plus	0.000	—	0.011	0.920	0.033	0.930	0.144	0.931	0.156	0.934	0.322	0.931	0.556	0.200	1.000	0.972
Qwen3.8-27B	0.078	0.807	0.122	0.799	0.189	0.821	0.389	0.814	0.422	0.817	0.444	0.833	0.637	0.430	1.000	0.864
Qwen3.8-Flash	0.089	0.838	0.244	0.904	0.233	0.918	0.522	0.883	0.478	0.882	0.567	0.884	0.673	0.524	0.991	0.939
Qwen3.8-Max	0.078	0.756	0.289	0.799	0.322	0.781	0.522	0.821	0.811	0.820	0.900	0.844	0.744	0.655	1.000	0.858
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.222	0.934	0.433	0.937	0.422	0.939	0.678	0.953	0.933	0.955	0.878	0.959	0.784	0.743	0.973	0.964
GPT-5.4 mini	0.033	0.887	0.078	0.934	0.056	0.920	0.378	0.948	0.322	0.942	0.478	0.956	0.612	0.366	1.000	0.915
GPT-5.6 Luna	0.000	—	0.056	0.876	0.033	0.857	0.233	0.865	0.333	0.861	0.556	0.872	0.601	0.336	1.000	0.966
GPT-5.6 Sol	0.167	0.758	0.289	0.806	0.244	0.795	0.544	0.890	0.644	0.873	0.756	0.892	0.716	0.611	0.991	0.976
GPT-5.6 Terra	0.133	0.733	0.367	0.774	0.400	0.764	0.778	0.853	0.844	0.826	0.856	0.821	0.781	0.720	1.000	0.931
Grok 4.1 Fast Reasoning	0.022	0.950	0.044	0.905	0.011	0.920	0.089	0.883	0.111	0.889	0.156	0.884	0.522	0.134	0.973	0.950
Grok 4.20 Reasoning	0.000	—	0.011	0.820	0.033	0.683	0.122	0.717	0.056	0.672	0.167	0.671	0.528	0.122	0.991	0.774
Kimi K2.6	0.033	0.920	0.189	0.904	0.144	0.904	0.533	0.907	0.678	0.909	0.733	0.912	0.693	0.556	1.000	0.913
QVQ-Max-Latest	0.322	0.779	0.356	0.781	0.478	0.766	0.511	0.774	0.422	0.789	0.500	0.780	0.611	0.585	0.791	0.916
Qwen3.6-35B-A3B	0.156	0.923	0.278	0.905	0.322	0.925	0.444	0.917	0.556	0.920	0.711	0.925	0.674	0.577	0.936	0.964
Qwen3.6-Plus	0.044	0.917	0.222	0.902	0.233	0.923	0.478	0.919	0.578	0.926	0.600	0.926	0.671	0.527	0.982	0.951
Qwen3.7-Flash	0.000	—	0.067	0.927	0.100	0.939	0.233	0.939	0.322	0.941	0.478	0.934	0.600	0.333	1.000	0.954
Qwen3.7-Max	0.000	—	0.022	0.915	0.067	0.900	0.267	0.879	0.356	0.877	0.400	0.885	0.593	0.312	1.000	0.950
Qwen3.7-Plus	0.011	0.850	0.044	0.843	0.067	0.892	0.300	0.880	0.378	0.871	0.433	0.893	0.603	0.341	1.000	0.949
Specialized Detectors																
CO-SPY [ProGAN]	1.000	0.997	0.033	0.807	0.567	0.929	0.222	0.908	1.000	0.999	1.000	1.000	0.750	0.765	0.864	0.980
CO-SPY [SD-v1.4]	0.444	0.756	0.400	0.770	0.467	0.767	0.544	0.773	0.500	0.743	0.400	0.783	0.575	0.603	0.691	0.827
ForgeLens [ProGAN]	0.889	0.983	0.311	0.892	0.667	0.910	0.844	0.949	0.911	0.981	0.900	0.984	0.809	0.846	0.864	0.944
ForgeLens [GenImage]	0.956	0.873	0.011	0.745	1.000	0.934	1.000	0.887	1.000	0.900	0.944	0.873	0.787	0.876	0.755	0.874
Effort [SD-v1.4]	0.744	0.838	0.567	0.794	0.800	0.818	0.789	0.838	0.956	0.897	0.922	0.892	0.748	0.857	0.700	0.793
Effort [Chameleon]	0.967	0.863	0.978	0.859	0.978	0.864	0.989	0.881	0.978	0.866	0.967	0.867	0.502	0.898	0.027	0.617
IAPL [ProGAN]	0.722	0.951	0.933	0.977	0.789	0.952	0.689	0.948	0.833	0.966	0.944	0.974	0.696	0.859	0.573	0.908
IAPL [SD-v1.4]	0.989	0.921	0.989	0.924	0.978	0.916	1.000	0.937	0.989	0.924	1.000	0.924	0.564	0.915	0.136	0.763

Table 39: **Performance comparison of MLLMs on the Industrial & Scientific subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.125	0.647	0.125	0.757	0.292	0.674	0.208	0.706	0.375	0.736	0.417	0.778	0.595	0.404	0.933	0.747
Gemini 3.1 Pro	0.083	0.710	0.125	0.703	0.250	0.707	0.417	0.724	0.417	0.717	0.458	0.717	0.646	0.452	1.000	0.703
Gemini 3.5 Flash-Lite	0.083	0.970	0.083	0.965	0.208	0.958	0.250	0.977	0.250	0.973	0.292	0.964	0.597	0.326	1.000	0.942
Gemini 3.8 Flash	0.000	—	0.000	—	0.208	0.684	0.292	0.703	0.292	0.711	0.250	0.710	0.587	0.296	1.000	0.701
GPT-6 Astra	0.083	0.820	0.125	0.777	0.333	0.816	0.458	0.862	0.417	0.853	0.375	0.839	0.649	0.460	1.000	0.920
GPT-6 Luna	0.750	0.879	1.000	0.886	0.958	0.910	0.958	0.921	0.750	0.868	0.917	0.903	0.511	0.859	0.133	0.713
GPT-6 Sol	0.083	0.720	0.000	—	0.333	0.750	0.375	0.796	0.250	0.840	0.375	0.837	0.618	0.382	1.000	0.906
Grok 4.6	0.000	—	0.000	—	0.125	0.763	0.208	0.802	0.208	0.816	0.208	0.810	0.562	0.222	1.000	0.850
Kimi K3	0.000	—	0.083	0.750	0.125	0.757	0.167	0.742	0.292	0.741	0.292	0.764	0.563	0.274	0.967	0.868
Qwen3-VL-Flash	0.125	0.940	0.083	0.920	0.375	0.940	0.333	0.939	0.333	0.943	0.375	0.937	0.635	0.426	1.000	0.970
Qwen3-VL-Plus	0.000	—	0.042	0.920	0.083	0.920	0.083	0.920	0.125	0.950	0.208	0.950	0.545	0.166	1.000	0.978
Qwen3.5-Omni-Flash	0.000	—	0.083	0.905	0.125	0.897	0.250	0.915	0.208	0.904	0.167	0.913	0.553	0.242	0.967	0.952
Qwen3.5-Omni-Plus	0.042	0.920	0.042	0.950	0.125	0.920	0.250	0.925	0.250	0.933	0.250	0.935	0.580	0.275	1.000	0.972
Qwen3.8-27B	0.083	0.790	0.167	0.825	0.208	0.780	0.250	0.830	0.333	0.805	0.417	0.819	0.622	0.391	1.000	0.861
Qwen3.8-Flash	0.125	0.867	0.250	0.900	0.375	0.923	0.458	0.835	0.458	0.900	0.625	0.912	0.691	0.553	1.000	0.942
Qwen3.8-Max	0.125	0.783	0.167	0.787	0.250	0.815	0.542	0.809	0.458	0.842	0.542	0.863	0.674	0.515	1.000	0.860
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.083	0.965	0.125	0.917	0.333	0.950	0.500	0.957	0.417	0.950	0.708	0.959	0.681	0.531	1.000	0.968
GPT-5.4 mini	0.000	—	0.125	0.847	0.208	0.958	0.250	0.927	0.292	0.950	0.208	0.944	0.590	0.306	1.000	0.913
GPT-5.6 Luna	0.000	—	0.042	0.860	0.000	—	0.167	0.893	0.125	0.770	0.125	0.780	0.538	0.142	1.000	0.970
GPT-5.6 Sol	0.000	—	0.042	0.780	0.125	0.853	0.208	0.872	0.083	0.910	0.250	0.877	0.559	0.211	1.000	0.976
GPT-5.6 Terra	0.042	0.700	0.083	0.800	0.167	0.823	0.417	0.849	0.417	0.815	0.292	0.839	0.618	0.382	1.000	0.933
Grok 4.1 Fast Reasoning	0.083	0.835	0.083	0.885	0.042	0.880	0.125	0.940	0.250	0.905	0.375	0.902	0.513	0.269	0.867	0.942
Grok 4.20 Reasoning	0.083	0.665	0.083	0.665	0.250	0.695	0.333	0.711	0.333	0.692	0.250	0.700	0.594	0.362	0.967	0.751
Kimi K2.6	0.000	—	0.083	0.935	0.125	0.897	0.250	0.878	0.292	0.924	0.375	0.912	0.594	0.316	1.000	0.863
QVQ-Max-Latest	0.292	0.764	0.167	0.725	0.250	0.742	0.333	0.756	0.250	0.725	0.333	0.788	0.502	0.408	0.733	0.920
Qwen3.6-35B-A3B	0.083	0.920	0.250	0.898	0.292	0.890	0.375	0.920	0.417	0.895	0.583	0.926	0.617	0.492	0.900	0.964
Qwen3.6-Plus	0.083	0.935	0.083	0.885	0.125	0.883	0.333	0.915	0.417	0.925	0.333	0.919	0.615	0.373	1.000	0.952
Qwen3.7-Flash	0.042	0.950	0.083	0.925	0.167	0.932	0.375	0.947	0.458	0.943	0.542	0.945	0.639	0.435	1.000	0.961
Qwen3.7-Max	0.042	0.780	0.042	0.920	0.083	0.900	0.208	0.884	0.292	0.900	0.375	0.901	0.587	0.296	1.000	0.948
Qwen3.7-Plus	0.000	—	0.083	0.935	0.125	0.850	0.250	0.902	0.292	0.880	0.292	0.913	0.587	0.296	1.000	0.950
Specialized Detectors																
CO-SPY [ProGAN]	0.917	0.988	0.042	0.945	0.125	0.902	0.208	0.883	1.000	0.976	1.000	1.000	0.741	0.702	0.933	1.000
CO-SPY [SD-v1.4]	0.583	0.735	0.417	0.786	0.583	0.748	0.375	0.771	0.500	0.725	0.583	0.774	0.520	0.632	0.533	0.858
ForgeLens [ProGAN]	0.750	0.941	0.208	0.762	0.583	0.933	0.833	0.972	0.833	0.979	0.875	0.967	0.840	0.810	1.000	0.980
ForgeLens [GenImage]	0.958	0.887	0.042	0.647	1.000	0.945	1.000	0.904	1.000	0.930	1.000	0.908	0.917	0.909	1.000	0.856
Effort [SD-v1.4]	0.792	0.810	0.792	0.768	0.708	0.846	0.917	0.865	0.917	0.886	1.000	0.893	0.910	0.918	0.967	0.788
Effort [Chameleon]	0.917	0.689	0.792	0.705	0.792	0.745	0.875	0.728	0.708	0.727	0.875	0.719	0.413	0.812	0.000	—
IAPL [ProGAN]	0.375	0.953	0.458	0.851	0.125	0.941	0.250	0.925	0.583	0.947	0.708	0.970	0.708	0.588	1.000	0.994
IAPL [SD-v1.4]	1.000	0.901	0.833	0.867	0.917	0.871	1.000	0.942	1.000	0.869	1.000	0.912	0.846	0.952	0.733	0.827

Table 40: **Performance comparison of MLLMs on the Magazine Subscriptions subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.286	0.790	0.143	0.870	0.286	0.800	0.429	0.890	0.286	0.760	0.286	0.750	0.643	0.444	1.000	0.860
Gemini 3.1 Pro	0.143	0.700	0.000	—	0.286	0.735	0.286	0.750	0.286	0.710	0.429	0.710	0.619	0.385	1.000	0.700
Gemini 3.5 Flash-Lite	0.000	—	0.000	—	0.000	—	0.143	0.990	0.286	0.970	0.143	0.950	0.548	0.174	1.000	0.950
Gemini 3.8 Flash	0.143	0.720	0.000	—	0.143	0.720	0.286	0.720	0.143	0.700	0.143	0.720	0.571	0.250	1.000	0.490
GPT-6 Astra	0.143	0.910	0.000	—	0.143	0.840	0.429	0.917	0.143	0.880	0.143	0.860	0.583	0.286	1.000	0.950
GPT-6 Luna	0.714	0.950	0.857	0.943	0.714	0.962	0.857	0.985	0.857	0.968	0.857	0.978	0.655	0.883	0.500	0.720
GPT-6 Sol	0.143	0.820	0.143	0.720	0.286	0.825	0.571	0.950	0.429	0.830	0.429	0.863	0.667	0.500	1.000	0.925
Grok 4.6	0.286	0.880	0.143	0.860	0.429	0.860	0.429	0.880	0.429	0.747	0.286	0.800	0.667	0.500	1.000	0.850
Kimi K3	0.286	0.860	0.143	0.800	0.000	—	0.143	0.870	0.143	0.700	0.429	0.750	0.345	0.314	0.500	0.880
Qwen3-VL-Flash	0.000	—	0.429	0.940	0.286	0.950	0.286	0.950	0.143	0.920	0.571	0.950	0.643	0.444	1.000	0.965
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.500	0.000	1.000	0.980
Qwen3.5-Omni-Flash	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.500	0.000	1.000	0.950
Qwen3.5-Omni-Plus	0.000	—	0.143	0.920	0.000	—	0.286	0.960	0.000	—	0.000	—	0.536	0.133	1.000	0.960
Qwen3.8-27B	0.143	0.860	0.143	0.780	0.143	0.860	0.429	0.920	0.000	—	0.286	0.850	0.595	0.320	1.000	0.885
Qwen3.8-Flash	0.000	—	0.286	0.785	0.571	0.945	0.571	0.925	0.429	0.960	0.429	0.900	0.690	0.552	1.000	0.965
Qwen3.8-Max	0.000	—	0.000	—	0.000	—	0.571	0.853	0.000	—	0.286	0.780	0.571	0.250	1.000	0.885
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.143	0.990	0.143	0.950	0.286	0.985	0.286	0.990	0.714	0.974	0.714	0.982	0.690	0.552	1.000	0.965
GPT-5.4 mini	0.000	—	0.000	—	0.000	—	0.143	0.890	0.000	—	0.143	0.910	0.524	0.091	1.000	0.915
GPT-5.6 Luna	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.143	0.930	0.512	0.047	1.000	0.980
GPT-5.6 Sol	0.000	—	0.000	—	0.143	0.760	0.286	0.975	0.286	0.790	0.000	—	0.560	0.213	1.000	0.980
GPT-5.6 Terra	0.143	0.740	0.000	—	0.286	0.925	0.571	0.950	0.571	0.870	0.571	0.897	0.679	0.526	1.000	0.945
Grok 4.1 Fast Reasoning	0.000	—	0.143	0.850	0.429	0.917	0.286	0.900	0.000	—	0.571	0.942	0.369	0.377	0.500	0.960
Grok 4.20 Reasoning	0.143	0.820	0.000	—	0.429	0.727	0.286	0.750	0.143	0.680	0.571	0.690	0.631	0.415	1.000	0.725
Kimi K2.6	0.143	0.850	0.000	—	0.000	—	0.286	0.850	0.143	0.850	0.143	0.920	0.560	0.213	1.000	0.920
QVQ-Max-Latest	0.143	0.750	0.286	0.675	0.286	0.675	0.286	0.750	0.143	0.750	0.000	—	0.345	0.314	0.500	0.950
Qwen3.6-35B-A3B	0.143	0.920	0.143	0.920	0.000	—	0.286	0.915	0.286	0.950	0.286	0.920	0.595	0.320	1.000	0.950
Qwen3.6-Plus	0.143	0.950	0.429	0.900	0.143	0.920	0.286	0.950	0.286	0.900	0.286	0.900	0.631	0.415	1.000	0.965
Qwen3.7-Flash	0.000	—	0.143	0.950	0.143	0.950	0.571	0.958	0.143	0.950	0.000	—	0.583	0.286	1.000	0.965
Qwen3.7-Max	0.000	—	0.143	0.850	0.000	—	0.286	0.965	0.000	—	0.143	0.950	0.548	0.174	1.000	0.950
Qwen3.7-Plus	0.000	—	0.000	—	0.000	—	0.286	0.900	0.143	0.850	0.286	0.935	0.560	0.213	1.000	0.965
Specialized Detectors																
CO-SPY [ProGAN]	1.000	1.000	0.000	—	0.143	0.994	0.000	—	1.000	0.999	1.000	1.000	0.762	0.688	1.000	1.000
CO-SPY [SD-v1.4]	0.286	0.597	0.429	0.690	0.000	—	0.429	0.824	0.286	0.711	0.143	0.816	0.381	0.407	0.500	0.810
ForgeLens [ProGAN]	0.429	0.992	0.143	0.994	0.429	0.930	1.000	0.946	0.714	0.963	0.857	0.940	0.798	0.746	1.000	0.985
ForgeLens [GenImage]	1.000	0.924	0.286	0.716	1.000	0.949	1.000	0.918	1.000	0.934	1.000	0.924	0.940	0.937	1.000	0.933
Effort [SD-v1.4]	0.571	0.820	0.429	0.652	0.429	0.742	0.571	0.849	0.429	0.862	0.857	0.786	0.774	0.708	1.000	0.828
Effort [Chameleon]	0.857	0.731	0.857	0.769	1.000	0.769	1.000	0.794	1.000	0.788	1.000	0.808	0.476	0.952	0.000	—
IAPL [ProGAN]	0.143	0.993	0.000	—	0.000	—	0.000	—	0.143	0.998	0.143	0.979	0.536	0.133	1.000	0.999
IAPL [SD-v1.4]	1.000	0.754	0.714	0.801	0.714	0.848	1.000	0.878	1.000	0.835	1.000	0.776	0.702	0.938	0.500	0.952

Table 41: **Performance comparison of MLLMs on the Musical Instruments subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.062	0.720	0.312	0.704	0.250	0.650	0.438	0.656	0.500	0.691	0.625	0.691	0.662	0.530	0.960	0.752
Gemini 3.1 Pro	0.062	0.720	0.188	0.723	0.312	0.706	0.438	0.706	0.500	0.730	0.562	0.727	0.672	0.512	1.000	0.700
Gemini 3.5 Flash-Lite	0.062	0.950	0.125	0.950	0.062	0.990	0.062	0.950	0.312	0.964	0.375	0.960	0.583	0.286	1.000	0.950
Gemini 3.8 Flash	0.000	—	0.062	0.700	0.188	0.697	0.062	0.720	0.312	0.704	0.375	0.717	0.583	0.286	1.000	0.607
GPT-6 Astra	0.188	0.770	0.250	0.780	0.625	0.811	0.500	0.853	0.938	0.858	0.812	0.845	0.776	0.711	1.000	0.926
GPT-6 Luna	0.812	0.906	1.000	0.911	0.938	0.885	1.000	0.921	0.812	0.907	1.000	0.934	0.584	0.873	0.240	0.735
GPT-6 Sol	0.062	0.680	0.125	0.760	0.562	0.786	0.250	0.913	0.562	0.853	0.812	0.838	0.698	0.567	1.000	0.910
Grok 4.6	0.062	0.840	0.000	—	0.188	0.780	0.125	0.825	0.250	0.767	0.312	0.826	0.578	0.270	1.000	0.840
Kimi K3	0.000	—	0.062	0.750	0.062	0.600	0.250	0.782	0.375	0.807	0.312	0.742	0.589	0.301	1.000	0.872
Qwen3-VL-Flash	0.000	—	0.062	0.950	0.312	0.944	0.250	0.935	0.562	0.933	0.438	0.941	0.635	0.426	1.000	0.968
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.188	0.940	0.516	0.061	1.000	0.975
Qwen3.5-Omni-Flash	0.000	—	0.062	0.920	0.125	0.920	0.250	0.882	0.250	0.925	0.250	0.912	0.578	0.270	1.000	0.948
Qwen3.5-Omni-Plus	0.000	—	0.000	—	0.250	0.930	0.125	0.920	0.188	0.927	0.250	0.930	0.568	0.239	1.000	0.966
Qwen3.8-27B	0.188	0.810	0.000	—	0.188	0.793	0.188	0.883	0.312	0.824	0.375	0.847	0.584	0.342	0.960	0.859
Qwen3.8-Flash	0.062	0.780	0.312	0.924	0.438	0.940	0.438	0.831	0.625	0.926	0.812	0.915	0.704	0.614	0.960	0.936
Qwen3.8-Max	0.062	0.850	0.375	0.810	0.375	0.830	0.688	0.804	0.750	0.874	0.875	0.849	0.740	0.680	0.960	0.860
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.125	0.950	0.312	0.950	0.250	0.945	0.375	0.962	0.688	0.961	0.750	0.965	0.708	0.588	1.000	0.967
GPT-5.4 mini	0.000	—	0.125	0.970	0.125	0.985	0.062	0.960	0.312	0.932	0.625	0.945	0.604	0.345	1.000	0.915
GPT-5.6 Luna	0.000	—	0.000	—	0.188	0.780	0.062	0.840	0.250	0.852	0.500	0.883	0.583	0.286	1.000	0.970
GPT-5.6 Sol	0.062	0.760	0.000	—	0.125	0.805	0.188	0.890	0.375	0.903	0.500	0.879	0.604	0.345	1.000	0.976
GPT-5.6 Terra	0.188	0.760	0.125	0.700	0.312	0.788	0.312	0.904	0.500	0.866	0.688	0.887	0.677	0.523	1.000	0.924
Grok 4.1 Fast Reasoning	0.125	0.915	0.188	0.907	0.125	0.935	0.125	0.915	0.250	0.867	0.188	0.930	0.563	0.283	0.960	0.947
Grok 4.20 Reasoning	0.062	0.780	0.125	0.665	0.312	0.720	0.375	0.772	0.438	0.706	0.188	0.693	0.605	0.397	0.960	0.756
Kimi K2.6	0.000	—	0.062	0.850	0.125	0.885	0.250	0.875	0.688	0.925	0.750	0.887	0.656	0.476	1.000	0.909
QVQ-Max-Latest	0.312	0.770	0.375	0.708	0.250	0.712	0.312	0.770	0.625	0.720	0.500	0.769	0.558	0.539	0.720	0.936
Qwen3.6-35B-A3B	0.125	0.935	0.250	0.925	0.312	0.926	0.438	0.897	0.750	0.927	0.500	0.922	0.578	0.543	0.760	0.960
Qwen3.6-Plus	0.062	0.920	0.250	0.927	0.375	0.912	0.438	0.941	0.438	0.931	0.625	0.924	0.662	0.530	0.960	0.955
Qwen3.7-Flash	0.000	—	0.062	0.920	0.375	0.900	0.375	0.932	0.500	0.950	0.688	0.946	0.667	0.500	1.000	0.952
Qwen3.7-Max	0.000	—	0.125	0.900	0.375	0.883	0.312	0.918	0.562	0.886	0.438	0.909	0.631	0.460	0.960	0.950
Qwen3.7-Plus	0.062	0.850	0.062	0.750	0.188	0.907	0.250	0.905	0.500	0.884	0.312	0.912	0.615	0.373	1.000	0.948
Specialized Detectors																
CO-SPY [ProGAN]	1.000	0.999	0.000	—	0.250	0.897	0.250	0.804	0.938	0.964	1.000	0.981	0.786	0.728	1.000	0.997
CO-SPY [SD-v1.4]	0.125	0.760	0.125	0.631	0.125	0.730	0.188	0.715	0.125	0.824	0.250	0.815	0.378	0.248	0.600	0.813
ForgeLens [ProGAN]	0.625	0.960	0.375	0.855	0.625	0.914	0.812	0.922	0.750	0.952	0.688	0.952	0.823	0.785	1.000	0.977
ForgeLens [GenImage]	1.000	0.861	0.375	0.770	1.000	0.949	1.000	0.909	1.000	0.933	0.875	0.919	0.917	0.928	0.960	0.847
Effort [SD-v1.4]	1.000	0.842	0.812	0.796	0.688	0.823	0.875	0.831	0.812	0.907	0.875	0.942	0.802	0.885	0.760	0.809
Effort [Chameleon]	0.688	0.726	0.750	0.703	0.875	0.694	1.000	0.688	0.875	0.711	0.812	0.729	0.437	0.800	0.040	0.502
IAPL [ProGAN]	0.062	0.999	0.062	0.999	0.000	—	0.062	0.994	0.312	0.963	0.312	0.965	0.568	0.239	1.000	0.997
IAPL [SD-v1.4]	1.000	0.902	1.000	0.887	1.000	0.907	1.000	0.941	0.938	0.909	1.000	0.928	0.755	0.936	0.520	0.825

Table 42: **Performance comparison of MLLMs on the Office Products subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.081	0.710	0.189	0.684	0.270	0.666	0.297	0.687	0.432	0.704	0.486	0.707	0.609	0.450	0.926	0.750
Gemini 3.1 Pro	0.054	0.710	0.081	0.700	0.081	0.677	0.162	0.733	0.270	0.721	0.514	0.716	0.597	0.325	1.000	0.710
Gemini 3.5 Flash-Lite	0.027	0.950	0.054	0.950	0.054	0.970	0.162	0.967	0.108	0.960	0.270	0.980	0.556	0.202	1.000	0.951
Gemini 3.8 Flash	0.027	0.680	0.000	—	0.108	0.700	0.135	0.720	0.270	0.708	0.297	0.709	0.570	0.245	1.000	0.643
GPT-6 Astra	0.135	0.756	0.189	0.776	0.351	0.804	0.568	0.865	0.568	0.851	0.649	0.876	0.705	0.581	1.000	0.928
GPT-6 Luna	0.730	0.896	0.946	0.927	0.946	0.929	1.000	0.933	0.919	0.900	0.973	0.927	0.589	0.915	0.259	0.747
GPT-6 Sol	0.054	0.770	0.108	0.720	0.162	0.837	0.405	0.852	0.432	0.823	0.514	0.852	0.640	0.437	1.000	0.911
Grok 4.6	0.000	—	0.027	0.840	0.135	0.818	0.243	0.833	0.297	0.820	0.270	0.813	0.581	0.279	1.000	0.842
Kimi K3	0.000	—	0.081	0.713	0.027	0.720	0.135	0.786	0.162	0.690	0.297	0.779	0.559	0.210	1.000	0.877
Qwen3-VL-Flash	0.027	0.950	0.081	0.950	0.189	0.941	0.270	0.947	0.405	0.944	0.432	0.943	0.617	0.380	1.000	0.972
Qwen3-VL-Plus	0.000	—	0.000	—	0.054	0.950	0.135	0.962	0.081	0.960	0.135	0.962	0.534	0.127	1.000	0.977
Qwen3.5-Omni-Flash	0.000	—	0.027	0.920	0.081	0.920	0.189	0.929	0.162	0.912	0.135	0.914	0.550	0.180	1.000	0.949
Qwen3.5-Omni-Plus	0.000	—	0.054	0.930	0.108	0.920	0.189	0.929	0.243	0.942	0.351	0.945	0.579	0.272	1.000	0.971
Qwen3.8-27B	0.108	0.812	0.054	0.760	0.135	0.844	0.297	0.822	0.135	0.872	0.405	0.833	0.595	0.318	1.000	0.875
Qwen3.8-Flash	0.054	0.950	0.297	0.908	0.297	0.903	0.459	0.876	0.378	0.891	0.514	0.921	0.574	0.492	0.815	0.924
Qwen3.8-Max	0.108	0.787	0.351	0.772	0.297	0.825	0.514	0.850	0.486	0.837	0.676	0.838	0.703	0.577	1.000	0.856
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.135	0.956	0.270	0.928	0.297	0.966	0.486	0.966	0.514	0.957	0.676	0.973	0.680	0.566	0.963	0.971
GPT-5.4 mini	0.000	—	0.108	0.928	0.081	0.963	0.108	0.968	0.216	0.946	0.324	0.947	0.570	0.245	1.000	0.914
GPT-5.6 Luna	0.000	—	0.000	—	0.135	0.758	0.108	0.865	0.135	0.816	0.432	0.893	0.568	0.238	1.000	0.970
GPT-5.6 Sol	0.000	—	0.000	—	0.054	0.830	0.189	0.903	0.189	0.911	0.270	0.958	0.559	0.210	1.000	0.977
GPT-5.6 Terra	0.081	0.720	0.108	0.723	0.216	0.806	0.541	0.869	0.432	0.809	0.568	0.862	0.662	0.490	1.000	0.940
Grok 4.1 Fast Reasoning	0.027	0.950	0.054	0.920	0.108	0.907	0.108	0.880	0.162	0.910	0.189	0.929	0.498	0.193	0.889	0.948
Grok 4.20 Reasoning	0.000	—	0.027	0.680	0.108	0.700	0.216	0.705	0.189	0.720	0.324	0.717	0.572	0.252	1.000	0.752
Kimi K2.6	0.000	—	0.108	0.867	0.027	0.920	0.297	0.894	0.297	0.904	0.378	0.910	0.592	0.312	1.000	0.918
QVQ-Max-Latest	0.297	0.718	0.189	0.779	0.081	0.783	0.297	0.741	0.189	0.786	0.405	0.747	0.492	0.382	0.741	0.935
Qwen3.6-35B-A3B	0.081	0.910	0.189	0.930	0.324	0.905	0.351	0.912	0.432	0.916	0.622	0.926	0.537	0.488	0.741	0.971
Qwen3.6-Plus	0.081	0.933	0.189	0.923	0.189	0.914	0.324	0.929	0.514	0.929	0.568	0.933	0.655	0.474	1.000	0.953
Qwen3.7-Flash	0.027	0.920	0.135	0.950	0.216	0.941	0.378	0.952	0.351	0.946	0.595	0.951	0.623	0.441	0.963	0.958
Qwen3.7-Max	0.000	—	0.027	0.850	0.162	0.925	0.297	0.906	0.378	0.911	0.297	0.902	0.597	0.325	1.000	0.952
Qwen3.7-Plus	0.000	—	0.162	0.915	0.081	0.900	0.216	0.906	0.297	0.905	0.432	0.908	0.599	0.331	1.000	0.946
Specialized Detectors																
CO-SPY [ProGAN]	0.973	0.998	0.027	0.583	0.189	0.829	0.054	0.804	0.838	0.971	1.000	0.990	0.738	0.677	0.963	1.000
CO-SPY [SD-v1.4]	0.378	0.772	0.378	0.766	0.324	0.803	0.432	0.818	0.351	0.789	0.351	0.782	0.388	0.512	0.407	0.802
ForgeLens [ProGAN]	0.622	0.982	0.216	0.911	0.595	0.921	0.838	0.960	0.892	0.976	0.919	0.974	0.840	0.810	1.000	0.984
ForgeLens [GenImage]	0.946	0.881	0.162	0.810	1.000	0.942	0.973	0.881	1.000	0.927	1.000	0.886	0.905	0.915	0.963	0.825
Effort [SD-v1.4]	0.784	0.868	0.676	0.784	0.784	0.835	0.892	0.823	0.865	0.861	0.946	0.880	0.745	0.884	0.667	0.723
Effort [Chameleon]	0.865	0.694	0.892	0.703	0.973	0.724	0.946	0.751	0.973	0.730	0.946	0.746	0.485	0.910	0.037	0.502
IAPL [ProGAN]	0.270	0.873	0.351	0.886	0.108	0.932	0.054	0.992	0.351	0.913	0.649	0.947	0.649	0.458	1.000	0.978
IAPL [SD-v1.4]	1.000	0.912	0.973	0.895	0.973	0.886	1.000	0.941	0.973	0.915	1.000	0.929	0.753	0.965	0.519	0.788

Table 43: **Performance comparison of MLLMs on the Patio, Lawn & Garden subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.053	0.740	0.105	0.705	0.211	0.688	0.211	0.675	0.211	0.698	0.263	0.660	0.568	0.296	0.960	0.768
Gemini 3.1 Pro	0.158	0.723	0.105	0.740	0.053	0.650	0.211	0.705	0.211	0.738	0.474	0.706	0.601	0.336	1.000	0.706
Gemini 3.5 Flash-Lite	0.000	—	0.053	0.950	0.053	0.950	0.053	0.950	0.211	0.960	0.211	0.970	0.548	0.176	1.000	0.950
Gemini 3.8 Flash	0.053	0.680	0.000	—	0.053	0.720	0.053	0.680	0.263	0.704	0.105	0.700	0.544	0.161	1.000	0.678
GPT-6 Astra	0.053	0.720	0.000	—	0.053	0.840	0.263	0.812	0.579	0.835	0.526	0.819	0.623	0.394	1.000	0.927
GPT-6 Luna	0.474	0.874	0.789	0.893	0.947	0.889	1.000	0.906	0.684	0.878	0.947	0.902	0.444	0.803	0.080	0.815
GPT-6 Sol	0.000	—	0.053	0.680	0.105	0.770	0.105	0.770	0.368	0.824	0.211	0.842	0.570	0.246	1.000	0.920
Grok 4.6	0.000	—	0.000	—	0.053	0.840	0.000	—	0.053	0.840	0.158	0.797	0.522	0.084	1.000	0.844
Kimi K3	0.000	—	0.053	0.700	0.000	—	0.105	0.750	0.053	0.800	0.158	0.700	0.531	0.116	1.000	0.868
Qwen3-VL-Flash	0.000	—	0.053	0.920	0.105	0.935	0.211	0.935	0.316	0.918	0.368	0.923	0.568	0.296	0.960	0.969
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.053	0.950	0.105	0.950	0.513	0.051	1.000	0.974
Qwen3.5-Omni-Flash	0.000	—	0.000	—	0.105	0.920	0.053	0.920	0.000	—	0.158	0.893	0.526	0.100	1.000	0.951
Qwen3.5-Omni-Plus	0.000	—	0.000	—	0.000	—	0.158	0.903	0.158	0.930	0.053	0.940	0.531	0.116	1.000	0.967
Qwen3.8-27B	0.158	0.833	0.105	0.820	0.211	0.825	0.158	0.803	0.211	0.790	0.263	0.800	0.592	0.311	1.000	0.867
Qwen3.8-Flash	0.053	0.880	0.158	0.840	0.316	0.863	0.368	0.903	0.421	0.886	0.579	0.868	0.658	0.480	1.000	0.938
Qwen3.8-Max	0.158	0.800	0.105	0.785	0.158	0.767	0.421	0.815	0.579	0.780	0.579	0.828	0.667	0.500	1.000	0.857
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.158	0.950	0.158	0.927	0.158	0.950	0.421	0.949	0.579	0.930	0.684	0.940	0.660	0.526	0.960	0.967
GPT-5.4 mini	0.000	—	0.053	0.930	0.158	0.893	0.000	—	0.158	0.960	0.211	0.897	0.548	0.176	1.000	0.912
GPT-5.6 Luna	0.000	—	0.000	—	0.000	—	0.211	0.775	0.053	0.780	0.316	0.773	0.548	0.176	1.000	0.973
GPT-5.6 Sol	0.000	—	0.000	—	0.000	—	0.000	—	0.105	0.895	0.158	0.913	0.522	0.084	1.000	0.976
GPT-5.6 Terra	0.053	0.880	0.053	0.720	0.105	0.700	0.316	0.788	0.263	0.784	0.368	0.830	0.596	0.324	1.000	0.933
Grok 4.1 Fast Reasoning	0.053	0.920	0.053	0.950	0.053	0.850	0.053	0.850	0.053	0.950	0.105	0.915	0.531	0.116	1.000	0.946
Grok 4.20 Reasoning	0.000	—	0.000	—	0.053	0.650	0.158	0.683	0.053	0.680	0.053	0.680	0.526	0.100	1.000	0.773
Kimi K2.6	0.000	—	0.000	—	0.053	0.850	0.105	0.885	0.158	0.897	0.263	0.920	0.548	0.176	1.000	0.917
QVQ-Max-Latest	0.105	0.750	0.211	0.762	0.526	0.770	0.474	0.744	0.421	0.744	0.474	0.756	0.524	0.512	0.680	0.912
Qwen3.6-35B-A3B	0.105	0.940	0.263	0.912	0.316	0.897	0.263	0.918	0.526	0.906	0.579	0.904	0.651	0.506	0.960	0.962
Qwen3.6-Plus	0.000	—	0.053	0.920	0.105	0.885	0.158	0.907	0.158	0.943	0.368	0.913	0.570	0.246	1.000	0.950
Qwen3.7-Flash	0.158	0.927	0.105	0.965	0.158	0.933	0.211	0.935	0.263	0.924	0.421	0.945	0.610	0.360	1.000	0.958
Qwen3.7-Max	0.053	0.900	0.000	—	0.053	0.720	0.158	0.823	0.158	0.840	0.421	0.916	0.570	0.246	1.000	0.950
Qwen3.7-Plus	0.000	—	0.000	—	0.000	—	0.105	0.850	0.158	0.930	0.316	0.905	0.548	0.176	1.000	0.943
Specialized Detectors																
CO-SPY [ProGAN]	1.000	1.000	0.105	0.975	0.474	0.897	0.421	0.884	1.000	0.996	1.000	0.983	0.833	0.800	1.000	1.000
CO-SPY [SD-v1.4]	0.316	0.864	0.368	0.782	0.421	0.793	0.474	0.773	0.421	0.801	0.474	0.797	0.546	0.556	0.680	0.774
ForgeLens [ProGAN]	0.737	0.960	0.368	0.911	0.737	0.953	0.789	0.966	0.947	0.983	0.947	0.981	0.857	0.856	0.960	0.956
ForgeLens [GenImage]	0.947	0.896	0.316	0.755	1.000	0.936	1.000	0.853	1.000	0.929	1.000	0.923	0.939	0.935	1.000	0.873
Effort [SD-v1.4]	0.947	0.848	0.737	0.805	0.789	0.840	0.842	0.791	0.947	0.920	0.895	0.964	0.850	0.907	0.840	0.895
Effort [Chameleon]	0.684	0.761	0.684	0.760	0.684	0.782	0.684	0.803	0.684	0.782	0.737	0.763	0.346	0.725	0.000	—
IAPL [ProGAN]	0.421	0.908	0.474	0.948	0.368	0.934	0.263	0.950	0.526	0.993	0.737	0.985	0.712	0.631	0.960	0.998
IAPL [SD-v1.4]	1.000	0.907	0.947	0.883	1.000	0.921	1.000	0.943	0.947	0.924	0.947	0.923	0.767	0.941	0.560	0.776

Table 44: **Performance comparison of MLLMs on the Pet Supplies subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.083	0.700	0.000	—	0.167	0.700	0.167	0.630	0.333	0.670	0.417	0.668	0.535	0.322	0.875	0.701
Gemini 3.1 Pro	0.167	0.710	0.083	0.700	0.333	0.712	0.500	0.715	0.333	0.722	0.583	0.717	0.667	0.500	1.000	0.709
Gemini 3.5 Flash-Lite	0.083	0.990	0.000	—	0.167	0.990	0.000	—	0.083	0.980	0.250	0.963	0.549	0.177	1.000	0.950
Gemini 3.8 Flash	0.000	—	0.000	—	0.000	—	0.000	—	0.083	0.680	0.083	0.720	0.514	0.054	1.000	0.690
GPT-6 Astra	0.333	0.792	0.083	0.760	0.167	0.760	0.167	0.820	0.250	0.830	0.333	0.848	0.611	0.364	1.000	0.935
GPT-6 Luna	0.833	0.848	0.917	0.895	1.000	0.912	1.000	0.890	0.750	0.860	1.000	0.924	0.458	0.904	0.000	—
GPT-6 Sol	0.083	0.830	0.000	—	0.167	0.750	0.167	0.800	0.167	0.810	0.167	0.785	0.562	0.222	1.000	0.895
Grok 4.6	0.083	0.760	0.000	—	0.083	0.860	0.000	—	0.083	0.900	0.000	—	0.521	0.080	1.000	0.830
Kimi K3	0.000	—	0.000	—	0.000	—	0.000	—	0.167	0.700	0.250	0.747	0.535	0.130	1.000	0.831
Qwen3-VL-Flash	0.083	0.920	0.000	—	0.167	0.950	0.083	0.950	0.250	0.907	0.250	0.940	0.569	0.244	1.000	0.972
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.083	0.920	0.507	0.027	1.000	0.980
Qwen3.5-Omni-Flash	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.500	0.000	1.000	0.954
Qwen3.5-Omni-Plus	0.083	0.920	0.000	—	0.000	—	0.083	0.950	0.083	0.920	0.083	0.920	0.528	0.105	1.000	0.971
Qwen3.8-27B	0.250	0.813	0.083	0.860	0.250	0.833	0.000	—	0.083	0.720	0.333	0.865	0.583	0.286	1.000	0.875
Qwen3.8-Flash	0.250	0.860	0.167	0.950	0.333	0.873	0.500	0.935	0.500	0.948	0.500	0.908	0.688	0.545	1.000	0.950
Qwen3.8-Max	0.167	0.785	0.000	—	0.083	0.820	0.167	0.785	0.500	0.770	0.667	0.825	0.632	0.418	1.000	0.856
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.250	0.970	0.583	0.946	0.333	0.985	0.583	0.967	0.667	0.953	0.750	0.973	0.764	0.691	1.000	0.961
GPT-5.4 mini	0.000	—	0.000	—	0.000	—	0.083	0.930	0.083	0.910	0.083	0.970	0.521	0.080	1.000	0.922
GPT-5.6 Luna	0.000	—	0.083	0.820	0.000	—	0.000	—	0.167	0.870	0.250	0.917	0.542	0.154	1.000	0.968
GPT-5.6 Sol	0.000	—	0.000	—	0.083	0.780	0.083	0.960	0.167	0.790	0.083	0.940	0.535	0.130	1.000	0.978
GPT-5.6 Terra	0.083	0.740	0.167	0.815	0.333	0.818	0.500	0.862	0.417	0.800	0.333	0.787	0.653	0.468	1.000	0.933
Grok 4.1 Fast Reasoning	0.000	—	0.000	—	0.000	—	0.083	0.850	0.000	—	0.000	—	0.382	0.027	0.750	0.945
Grok 4.20 Reasoning	0.000	—	0.083	0.720	0.250	0.777	0.000	—	0.167	0.700	0.000	—	0.542	0.154	1.000	0.750
Kimi K2.6	0.000	—	0.000	—	0.000	—	0.083	0.850	0.083	0.920	0.083	0.920	0.521	0.080	1.000	0.920
QVQ-Max-Latest	0.250	0.750	0.167	0.800	0.167	0.650	0.167	0.725	0.250	0.783	0.250	0.783	0.542	0.341	0.875	0.929
Qwen3.6-35B-A3B	0.083	0.920	0.167	0.900	0.333	0.912	0.167	0.885	0.333	0.905	0.417	0.892	0.562	0.396	0.875	0.959
Qwen3.6-Plus	0.083	0.920	0.083	0.920	0.250	0.940	0.000	—	0.083	0.920	0.167	0.950	0.556	0.200	1.000	0.950
Qwen3.7-Flash	0.000	—	0.000	—	0.167	0.915	0.083	0.980	0.167	0.950	0.583	0.950	0.583	0.286	1.000	0.955
Qwen3.7-Max	0.167	0.935	0.000	—	0.167	0.935	0.083	0.950	0.333	0.882	0.167	0.850	0.576	0.265	1.000	0.950
Qwen3.7-Plus	0.083	0.920	0.083	0.850	0.083	0.850	0.167	0.850	0.417	0.876	0.250	0.907	0.590	0.306	1.000	0.954
Specialized Detectors																
CO-SPY [ProGAN]	1.000	0.999	0.000	—	0.000	—	0.083	0.998	0.917	0.991	1.000	1.000	0.750	0.667	1.000	1.000
CO-SPY [SD-v1.4]	1.000	0.815	0.917	0.799	0.750	0.730	0.917	0.842	0.917	0.825	0.917	0.784	0.951	0.949	1.000	0.847
ForgeLens [ProGAN]	1.000	0.986	0.333	0.828	0.333	0.829	0.750	0.983	1.000	0.991	0.917	0.985	0.799	0.832	0.875	0.976
ForgeLens [GenImage]	1.000	0.895	0.000	—	1.000	0.933	0.917	0.845	1.000	0.923	1.000	0.863	0.910	0.901	1.000	0.880
Effort [SD-v1.4]	0.500	0.704	0.250	0.627	0.667	0.837	0.583	0.895	0.833	0.929	0.750	0.875	0.674	0.735	0.750	0.832
Effort [Chameleon]	1.000	0.739	0.917	0.722	1.000	0.758	1.000	0.783	1.000	0.727	1.000	0.734	0.493	0.940	0.000	—
IAPL [ProGAN]	0.000	—	0.167	0.718	0.000	—	0.000	—	0.417	0.846	0.583	0.974	0.597	0.326	1.000	0.942
IAPL [SD-v1.4]	0.833	0.872	0.833	0.858	1.000	0.827	1.000	0.943	1.000	0.891	1.000	0.919	0.910	0.965	0.875	0.789

Table 45: **Performance comparison of MLLMs on the Sports & Outdoors subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.050	0.760	0.150	0.697	0.225	0.693	0.175	0.733	0.425	0.742	0.450	0.754	0.623	0.395	1.000	0.753
Gemini 3.1 Pro	0.075	0.700	0.025	0.650	0.100	0.720	0.175	0.710	0.150	0.717	0.350	0.707	0.573	0.255	1.000	0.708
Gemini 3.5 Flash-Lite	0.075	0.963	0.100	0.975	0.100	0.965	0.275	0.965	0.275	0.962	0.200	0.964	0.585	0.292	1.000	0.952
Gemini 3.8 Flash	0.025	0.720	0.000	—	0.025	0.720	0.075	0.697	0.200	0.699	0.225	0.718	0.546	0.168	1.000	0.667
GPT-6 Astra	0.200	0.762	0.150	0.828	0.375	0.821	0.425	0.841	0.475	0.850	0.675	0.847	0.692	0.554	1.000	0.933
GPT-6 Luna	0.850	0.892	0.900	0.904	0.975	0.913	0.975	0.945	0.850	0.893	0.975	0.933	0.502	0.915	0.083	0.735
GPT-6 Sol	0.225	0.777	0.125	0.766	0.325	0.782	0.425	0.822	0.350	0.805	0.600	0.818	0.671	0.509	1.000	0.918
Grok 4.6	0.000	—	0.000	—	0.075	0.807	0.075	0.793	0.200	0.792	0.100	0.800	0.537	0.140	1.000	0.850
Kimi K3	0.000	—	0.025	0.720	0.000	—	0.050	0.800	0.125	0.758	0.250	0.731	0.537	0.140	1.000	0.859
Qwen3-VL-Flash	0.050	0.950	0.100	0.942	0.175	0.950	0.250	0.938	0.250	0.938	0.325	0.941	0.596	0.322	1.000	0.976
Qwen3-VL-Plus	0.000	—	0.000	—	0.025	0.950	0.050	0.920	0.025	0.920	0.175	0.946	0.523	0.088	1.000	0.979
Qwen3.5-Omni-Flash	0.025	0.920	0.125	0.920	0.050	0.920	0.100	0.912	0.150	0.920	0.225	0.914	0.556	0.202	1.000	0.955
Qwen3.5-Omni-Plus	0.000	—	0.050	0.945	0.050	0.920	0.025	0.980	0.125	0.934	0.125	0.928	0.531	0.118	1.000	0.974
Qwen3.8-27B	0.075	0.887	0.150	0.847	0.175	0.884	0.250	0.838	0.200	0.834	0.375	0.835	0.602	0.339	1.000	0.855
Qwen3.8-Flash	0.150	0.913	0.275	0.893	0.300	0.829	0.350	0.926	0.350	0.904	0.550	0.912	0.665	0.495	1.000	0.947
Qwen3.8-Max	0.075	0.727	0.150	0.787	0.275	0.794	0.400	0.829	0.300	0.816	0.600	0.837	0.650	0.462	1.000	0.868
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.225	0.963	0.275	0.952	0.325	0.966	0.425	0.959	0.450	0.959	0.650	0.956	0.654	0.560	0.917	0.969
GPT-5.4 mini	0.050	0.870	0.075	0.910	0.125	0.916	0.100	0.945	0.250	0.943	0.325	0.967	0.556	0.266	0.958	0.910
GPT-5.6 Luna	0.025	0.780	0.025	0.780	0.150	0.868	0.175	0.840	0.175	0.869	0.425	0.875	0.581	0.280	1.000	0.973
GPT-5.6 Sol	0.000	—	0.025	0.840	0.125	0.840	0.125	0.888	0.200	0.902	0.250	0.895	0.560	0.216	1.000	0.978
GPT-5.6 Terra	0.150	0.797	0.125	0.752	0.200	0.833	0.550	0.846	0.225	0.846	0.425	0.834	0.619	0.435	0.958	0.928
Grok 4.1 Fast Reasoning	0.075	0.883	0.075	0.917	0.100	0.880	0.100	0.917	0.075	0.860	0.125	0.904	0.504	0.167	0.917	0.948
Grok 4.20 Reasoning	0.025	0.820	0.050	0.650	0.150	0.700	0.225	0.709	0.125	0.680	0.250	0.696	0.548	0.241	0.958	0.777
Kimi K2.6	0.050	0.935	0.025	0.920	0.200	0.903	0.175	0.909	0.225	0.920	0.500	0.903	0.598	0.328	1.000	0.928
QVQ-Max-Latest	0.250	0.725	0.275	0.723	0.350	0.754	0.375	0.733	0.350	0.779	0.325	0.792	0.473	0.472	0.625	0.930
Qwen3.6-35B-A3B	0.225	0.901	0.325	0.904	0.325	0.898	0.325	0.902	0.425	0.918	0.800	0.924	0.681	0.574	0.958	0.964
Qwen3.6-Plus	0.050	0.950	0.125	0.908	0.175	0.917	0.125	0.936	0.200	0.934	0.425	0.924	0.592	0.310	1.000	0.957
Qwen3.7-Flash	0.050	0.915	0.200	0.940	0.175	0.930	0.200	0.941	0.175	0.943	0.450	0.939	0.604	0.345	1.000	0.955
Qwen3.7-Max	0.050	0.900	0.025	0.950	0.050	0.910	0.075	0.850	0.175	0.886	0.275	0.876	0.554	0.195	1.000	0.945
Qwen3.7-Plus	0.000	—	0.050	0.900	0.075	0.907	0.125	0.884	0.125	0.904	0.250	0.887	0.552	0.189	1.000	0.947
Specialized Detectors																
CO-SPY [ProGAN]	1.000	0.989	0.200	0.929	0.425	0.920	0.425	0.965	0.900	0.981	1.000	1.000	0.808	0.792	0.958	1.000
CO-SPY [SD-v1.4]	0.400	0.893	0.350	0.850	0.425	0.797	0.450	0.862	0.375	0.874	0.350	0.897	0.633	0.558	0.875	0.871
ForgeLens [ProGAN]	0.575	0.924	0.475	0.890	0.575	0.913	0.875	0.960	0.950	0.963	0.925	0.988	0.844	0.841	0.958	0.973
ForgeLens [GenImage]	0.875	0.844	0.125	0.741	1.000	0.938	1.000	0.891	1.000	0.929	1.000	0.927	0.896	0.907	0.958	0.927
Effort [SD-v1.4]	0.700	0.848	0.675	0.856	0.825	0.833	0.750	0.794	0.950	0.894	0.900	0.923	0.754	0.875	0.708	0.826
Effort [Chameleon]	0.725	0.678	0.825	0.709	0.825	0.679	0.925	0.740	0.875	0.737	0.875	0.716	0.442	0.869	0.042	0.546
IAPL [ProGAN]	0.200	0.898	0.550	0.916	0.225	0.918	0.125	0.917	0.450	0.924	0.550	0.900	0.675	0.519	1.000	0.997
IAPL [SD-v1.4]	0.975	0.895	0.975	0.927	1.000	0.921	1.000	0.941	1.000	0.928	1.000	0.929	0.683	0.966	0.375	0.761

Table 46: **Performance comparison of MLLMs on the Tools & Home Improvement subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.077	0.620	0.077	0.780	0.077	0.600	0.154	0.700	0.154	0.620	0.154	0.730	0.521	0.202	0.926	0.772
Gemini 3.1 Pro	0.154	0.675	0.231	0.713	0.077	0.700	0.231	0.730	0.308	0.703	0.231	0.733	0.584	0.337	0.963	0.689
Gemini 3.5 Flash-Lite	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.154	0.950	0.513	0.050	1.000	0.943
Gemini 3.8 Flash	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.077	0.680	0.506	0.025	1.000	0.653
GPT-6 Astra	0.154	0.725	0.000	—	0.077	0.770	0.231	0.807	0.231	0.770	0.385	0.798	0.590	0.304	1.000	0.927
GPT-6 Luna	0.846	0.855	0.846	0.905	0.923	0.918	0.692	0.902	0.538	0.894	0.923	0.922	0.527	0.775	0.259	0.759
GPT-6 Sol	0.077	0.790	0.077	0.720	0.000	—	0.077	0.780	0.154	0.755	0.385	0.804	0.564	0.227	1.000	0.911
Grok 4.6	0.077	0.840	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.506	0.025	1.000	0.843
Kimi K3	0.000	—	0.077	0.700	0.000	—	0.077	0.820	0.000	—	0.077	0.650	0.519	0.074	1.000	0.872
Qwen3-VL-Flash	0.000	—	0.077	0.920	0.000	—	0.077	0.950	0.077	0.950	0.231	0.883	0.538	0.143	1.000	0.971
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.077	0.920	0.000	—	0.000	—	0.506	0.025	1.000	0.976
Qwen3.5-Omni-Flash	0.000	—	0.077	0.920	0.000	—	0.000	—	0.000	—	0.154	0.920	0.519	0.074	1.000	0.951
Qwen3.5-Omni-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.500	0.000	1.000	0.971
Qwen3.8-27B	0.077	0.720	0.154	0.785	0.077	0.740	0.077	0.800	0.000	—	0.154	0.870	0.545	0.165	1.000	0.866
Qwen3.8-Flash	0.000	—	0.308	0.958	0.154	0.950	0.154	0.790	0.308	0.925	0.308	0.932	0.547	0.330	0.889	0.946
Qwen3.8-Max	0.000	—	0.077	0.720	0.077	0.780	0.308	0.795	0.308	0.780	0.308	0.852	0.590	0.304	1.000	0.858
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.154	0.980	0.231	0.960	0.000	—	0.154	0.950	0.308	0.973	0.538	0.959	0.615	0.375	1.000	0.967
GPT-5.4 mini	0.000	—	0.000	—	0.000	—	0.000	—	0.077	0.940	0.231	0.943	0.526	0.098	1.000	0.910
GPT-5.6 Luna	0.077	0.820	0.000	—	0.000	—	0.000	—	0.077	0.840	0.231	0.827	0.532	0.120	1.000	0.973
GPT-5.6 Sol	0.000	—	0.000	—	0.000	—	0.077	0.960	0.000	—	0.385	0.882	0.538	0.143	1.000	0.975
GPT-5.6 Terra	0.000	—	0.077	0.720	0.077	0.670	0.154	0.915	0.231	0.820	0.462	0.840	0.583	0.286	1.000	0.932
Grok 4.1 Fast Reasoning	0.000	—	0.077	0.850	0.077	0.920	0.077	0.880	0.000	—	0.154	0.865	0.532	0.120	1.000	0.949
Grok 4.20 Reasoning	0.077	0.720	0.000	—	0.077	0.650	0.000	—	0.077	0.720	0.231	0.717	0.538	0.143	1.000	0.758
Kimi K2.6	0.077	0.850	0.000	—	0.000	—	0.077	0.850	0.077	0.850	0.385	0.892	0.551	0.186	1.000	0.885
QVQ-Max-Latest	0.154	0.700	0.154	0.750	0.077	0.750	0.462	0.750	0.308	0.725	0.538	0.750	0.530	0.415	0.778	0.931
Qwen3.6-35B-A3B	0.077	0.920	0.231	0.860	0.077	0.920	0.154	0.935	0.154	0.885	0.154	0.935	0.552	0.244	0.963	0.961
Qwen3.6-Plus	0.000	—	0.000	—	0.000	—	0.000	—	0.000	—	0.231	0.930	0.519	0.074	1.000	0.954
Qwen3.7-Flash	0.000	—	0.077	0.920	0.077	0.850	0.154	0.950	0.077	0.900	0.154	0.935	0.526	0.163	0.963	0.957
Qwen3.7-Max	0.000	—	0.077	0.950	0.000	—	0.000	—	0.000	—	0.077	0.850	0.513	0.050	1.000	0.947
Qwen3.7-Plus	0.000	—	0.000	—	0.000	—	0.077	0.780	0.000	—	0.308	0.912	0.514	0.119	0.963	0.938
Specialized Detectors																
CO-SPY [ProGAN]	0.923	0.977	0.077	0.752	0.308	0.979	0.385	0.977	0.923	0.970	1.000	1.000	0.801	0.752	1.000	1.000
CO-SPY [SD-v1.4]	0.538	0.700	0.385	0.756	0.308	0.722	0.538	0.709	0.385	0.671	0.308	0.779	0.520	0.533	0.630	0.884
ForgeLens [ProGAN]	0.769	0.957	0.231	0.876	0.692	0.985	0.692	0.980	1.000	0.972	1.000	0.969	0.865	0.844	1.000	0.976
ForgeLens [GenImage]	1.000	0.874	0.231	0.610	1.000	0.934	1.000	0.867	1.000	0.930	1.000	0.908	0.917	0.925	0.963	0.886
Effort [SD-v1.4]	0.769	0.876	0.692	0.753	0.769	0.849	0.846	0.938	1.000	0.823	0.923	0.886	0.861	0.890	0.889	0.881
Effort [Chameleon]	0.923	0.736	0.923	0.686	1.000	0.718	0.923	0.758	0.923	0.725	0.923	0.735	0.468	0.820	0.000	—
IAPL [ProGAN]	0.462	0.894	0.308	0.871	0.231	0.789	0.077	0.524	0.769	0.938	0.692	0.882	0.712	0.595	1.000	0.999
IAPL [SD-v1.4]	0.923	0.936	0.923	0.907	1.000	0.896	1.000	0.946	1.000	0.901	1.000	0.936	0.783	0.921	0.593	0.823

Table 47: **Performance comparison of MLLMs on the Toys & Games subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.062	0.720	0.250	0.647	0.062	0.850	0.250	0.745	0.375	0.700	0.375	0.680	0.597	0.370	0.964	0.789
Gemini 3.1 Pro	0.062	0.650	0.125	0.715	0.188	0.707	0.250	0.705	0.312	0.714	0.625	0.712	0.630	0.413	1.000	0.710
Gemini 3.5 Flash-Lite	0.000	—	0.000	—	0.000	—	0.062	0.950	0.188	0.983	0.188	0.950	0.519	0.135	0.964	0.960
Gemini 3.8 Flash	0.000	—	0.000	—	0.125	0.690	0.062	0.680	0.125	0.720	0.125	0.700	0.536	0.136	1.000	0.654
GPT-6 Astra	0.188	0.837	0.188	0.837	0.312	0.824	0.375	0.835	0.688	0.816	0.500	0.816	0.688	0.545	1.000	0.943
GPT-6 Luna	0.625	0.923	0.938	0.927	0.938	0.938	0.938	0.933	0.875	0.902	0.938	0.937	0.562	0.836	0.250	0.817
GPT-6 Sol	0.188	0.873	0.312	0.830	0.188	0.853	0.250	0.850	0.438	0.770	0.250	0.830	0.635	0.426	1.000	0.931
Grok 4.6	0.062	0.750	0.062	0.820	0.062	0.840	0.125	0.860	0.188	0.823	0.188	0.847	0.557	0.206	1.000	0.849
Kimi K3	0.000	—	0.000	—	0.062	0.820	0.125	0.820	0.312	0.770	0.125	0.820	0.552	0.189	1.000	0.881
Qwen3-VL-Flash	0.000	—	0.000	—	0.125	0.950	0.188	0.950	0.188	0.950	0.188	0.940	0.557	0.206	1.000	0.978
Qwen3-VL-Plus	0.000	—	0.000	—	0.000	—	0.125	0.980	0.062	0.980	0.062	0.980	0.521	0.080	1.000	0.977
Qwen3.5-Omni-Flash	0.000	—	0.125	0.905	0.250	0.907	0.250	0.920	0.312	0.890	0.125	0.920	0.589	0.301	1.000	0.959
Qwen3.5-Omni-Plus	0.000	—	0.000	—	0.000	—	0.062	0.920	0.125	0.935	0.000	—	0.516	0.061	1.000	0.976
Qwen3.8-27B	0.125	0.840	0.000	—	0.188	0.840	0.312	0.848	0.312	0.772	0.250	0.815	0.581	0.328	0.964	0.892
Qwen3.8-Flash	0.062	0.950	0.250	0.913	0.188	0.873	0.438	0.914	0.375	0.858	0.562	0.887	0.656	0.476	1.000	0.935
Qwen3.8-Max	0.125	0.780	0.188	0.773	0.312	0.822	0.375	0.818	0.438	0.841	0.500	0.853	0.661	0.488	1.000	0.875
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.062	0.980	0.125	0.965	0.250	0.958	0.438	0.960	0.500	0.932	0.625	0.963	0.667	0.500	1.000	0.973
GPT-5.4 mini	0.000	—	0.000	—	0.000	—	0.312	0.940	0.188	0.940	0.188	0.943	0.557	0.206	1.000	0.936
GPT-5.6 Luna	0.000	—	0.062	0.930	0.062	0.780	0.062	0.940	0.125	0.900	0.250	0.843	0.547	0.171	1.000	0.972
GPT-5.6 Sol	0.062	0.970	0.125	0.890	0.125	0.980	0.188	0.813	0.250	0.922	0.062	0.880	0.568	0.239	1.000	0.975
GPT-5.6 Terra	0.125	0.830	0.250	0.800	0.125	0.905	0.312	0.768	0.250	0.922	0.500	0.799	0.630	0.413	1.000	0.938
Grok 4.1 Fast Reasoning	0.125	0.900	0.062	0.850	0.062	0.920	0.188	0.883	0.250	0.875	0.062	0.920	0.527	0.218	0.929	0.948
Grok 4.20 Reasoning	0.000	—	0.000	—	0.125	0.650	0.312	0.766	0.125	0.685	0.125	0.690	0.557	0.206	1.000	0.790
Kimi K2.6	0.062	0.950	0.062	0.950	0.188	0.873	0.375	0.897	0.375	0.920	0.312	0.906	0.615	0.373	1.000	0.935
QVQ-Max-Latest	0.188	0.750	0.250	0.713	0.250	0.738	0.375	0.783	0.438	0.757	0.375	0.767	0.603	0.465	0.893	0.940
Qwen3.6-35B-A3B	0.188	0.907	0.188	0.940	0.188	0.917	0.375	0.918	0.500	0.927	0.375	0.930	0.615	0.457	0.929	0.973
Qwen3.6-Plus	0.062	0.850	0.125	0.920	0.188	0.907	0.250	0.935	0.312	0.908	0.312	0.898	0.586	0.342	0.964	0.959
Qwen3.7-Flash	0.000	—	0.062	0.950	0.250	0.923	0.312	0.944	0.375	0.932	0.438	0.936	0.620	0.387	1.000	0.966
Qwen3.7-Max	0.000	—	0.000	—	0.000	—	0.125	0.875	0.062	0.950	0.125	0.900	0.526	0.099	1.000	0.951
Qwen3.7-Plus	0.000	—	0.000	—	0.062	0.950	0.125	0.950	0.188	0.883	0.062	0.850	0.536	0.136	1.000	0.951
Specialized Detectors																
CO-SPY [ProGAN]	1.000	0.981	0.000	—	0.125	0.988	0.125	0.753	0.500	0.893	1.000	1.000	0.729	0.629	1.000	1.000
CO-SPY [SD-v1.4]	0.250	0.775	0.125	0.610	0.188	0.621	0.312	0.772	0.125	0.514	0.312	0.762	0.377	0.323	0.536	0.811
ForgeLens [ProGAN]	0.750	0.919	0.125	0.761	0.438	0.871	1.000	0.939	1.000	0.972	1.000	0.976	0.842	0.831	0.964	0.962
ForgeLens [GenImage]	0.938	0.855	0.000	—	1.000	0.938	1.000	0.910	1.000	0.904	1.000	0.896	0.840	0.883	0.857	0.808
Effort [SD-v1.4]	0.625	0.780	0.062	0.533	0.375	0.710	0.688	0.757	0.625	0.670	0.625	0.834	0.696	0.653	0.893	0.846
Effort [Chameleon]	0.562	0.627	0.438	0.659	0.500	0.640	0.500	0.669	0.438	0.667	0.438	0.650	0.257	0.544	0.036	0.600
IAPL [ProGAN]	0.312	0.788	0.250	0.956	0.125	0.807	0.000	—	0.250	0.887	0.562	0.867	0.625	0.400	1.000	0.997
IAPL [SD-v1.4]	1.000	0.823	0.938	0.763	1.000	0.787	1.000	0.918	0.938	0.819	1.000	0.840	0.775	0.931	0.571	0.827

Table 48: **Performance comparison of MLLMs on the Video Games subset of FRAUDBENCH.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.069	0.640	0.241	0.677	0.379	0.660	0.448	0.716	0.517	0.717	0.621	0.714	0.606	0.543	0.833	0.771
Gemini 3.1 Pro	0.034	0.720	0.034	0.720	0.207	0.723	0.310	0.708	0.241	0.721	0.517	0.717	0.612	0.366	1.000	0.709
Gemini 3.5 Flash-Lite	0.069	0.970	0.034	0.950	0.034	0.990	0.138	0.980	0.241	0.971	0.241	0.980	0.563	0.224	1.000	0.952
Gemini 3.8 Flash	0.000	—	0.000	—	0.000	—	0.103	0.713	0.276	0.715	0.276	0.716	0.555	0.197	1.000	0.657
GPT-6 Astra	0.138	0.748	0.034	0.850	0.379	0.805	0.655	0.848	0.655	0.858	0.759	0.866	0.718	0.608	1.000	0.934
GPT-6 Luna	0.759	0.886	0.897	0.924	0.931	0.935	1.000	0.913	0.862	0.906	0.966	0.939	0.451	0.900	0.000	—
GPT-6 Sol	0.172	0.760	0.345	0.766	0.345	0.822	0.552	0.897	0.655	0.845	0.724	0.872	0.733	0.635	1.000	0.911
Grok 4.6	0.034	0.860	0.000	—	0.207	0.803	0.310	0.859	0.517	0.809	0.448	0.820	0.626	0.404	1.000	0.847
Kimi K3	0.000	—	0.069	0.650	0.000	—	0.207	0.800	0.103	0.750	0.517	0.806	0.575	0.260	1.000	0.864
Qwen3-VL-Flash	0.034	0.920	0.103	0.920	0.345	0.941	0.483	0.939	0.379	0.942	0.517	0.946	0.655	0.474	1.000	0.973
Qwen3-VL-Plus	0.000	—	0.000	—	0.034	0.950	0.241	0.963	0.103	0.950	0.241	0.967	0.552	0.188	1.000	0.977
Qwen3.5-Omni-Flash	0.000	—	0.034	0.920	0.103	0.910	0.379	0.916	0.241	0.920	0.345	0.920	0.592	0.311	1.000	0.957
Qwen3.5-Omni-Plus	0.000	—	0.034	0.940	0.103	0.933	0.345	0.930	0.310	0.940	0.414	0.946	0.601	0.335	1.000	0.972
Qwen3.8-27B	0.034	0.860	0.034	0.820	0.138	0.855	0.345	0.876	0.276	0.844	0.310	0.897	0.595	0.319	1.000	0.879
Qwen3.8-Flash	0.103	0.950	0.241	0.897	0.379	0.875	0.586	0.925	0.586	0.882	0.690	0.911	0.716	0.602	1.000	0.944
Qwen3.8-Max	0.103	0.817	0.172	0.838	0.379	0.825	0.586	0.861	0.621	0.868	0.793	0.890	0.721	0.614	1.000	0.874
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.138	0.975	0.069	0.970	0.310	0.968	0.517	0.949	0.483	0.966	0.621	0.964	0.678	0.525	1.000	0.973
GPT-5.4 mini	0.000	—	0.138	0.900	0.276	0.961	0.241	0.970	0.345	0.952	0.414	0.970	0.618	0.381	1.000	0.938
GPT-5.6 Luna	0.034	0.860	0.069	0.760	0.241	0.830	0.345	0.858	0.414	0.843	0.517	0.919	0.635	0.425	1.000	0.973
GPT-5.6 Sol	0.000	—	0.000	—	0.034	0.860	0.345	0.911	0.483	0.894	0.448	0.926	0.609	0.358	1.000	0.979
GPT-5.6 Terra	0.103	0.767	0.103	0.873	0.172	0.824	0.483	0.879	0.448	0.856	0.724	0.880	0.670	0.506	1.000	0.938
Grok 4.1 Fast Reasoning	0.034	0.950	0.138	0.915	0.103	0.903	0.207	0.912	0.241	0.891	0.414	0.894	0.539	0.316	0.889	0.949
Grok 4.20 Reasoning	0.103	0.707	0.103	0.670	0.241	0.701	0.552	0.733	0.414	0.730	0.414	0.697	0.652	0.467	1.000	0.761
Kimi K2.6	0.000	—	0.034	0.920	0.103	0.920	0.276	0.903	0.379	0.907	0.448	0.915	0.603	0.343	1.000	0.923
QVQ-Max-Latest	0.103	0.717	0.207	0.767	0.379	0.759	0.379	0.764	0.379	0.764	0.552	0.719	0.583	0.494	0.833	0.927
Qwen3.6-35B-A3B	0.103	0.917	0.345	0.904	0.379	0.931	0.552	0.935	0.517	0.928	0.724	0.933	0.580	0.596	0.722	0.978
Qwen3.6-Plus	0.103	0.920	0.207	0.940	0.276	0.931	0.414	0.929	0.517	0.929	0.517	0.938	0.670	0.506	1.000	0.957
Qwen3.7-Flash	0.034	0.950	0.138	0.933	0.310	0.939	0.517	0.948	0.483	0.949	0.724	0.942	0.684	0.538	1.000	0.964
Qwen3.7-Max	0.034	0.950	0.103	0.817	0.172	0.868	0.345	0.873	0.345	0.891	0.414	0.889	0.618	0.381	1.000	0.948
Qwen3.7-Plus	0.034	0.950	0.069	0.885	0.138	0.942	0.345	0.912	0.414	0.870	0.448	0.918	0.621	0.389	1.000	0.947
Specialized Detectors																
CO-SPY [ProGAN]	1.000	1.000	0.000	—	0.241	0.848	0.103	0.926	0.897	0.941	1.000	1.000	0.770	0.701	1.000	1.000
CO-SPY [SD-v1.4]	0.345	0.837	0.241	0.760	0.276	0.738	0.310	0.765	0.310	0.765	0.345	0.820	0.569	0.461	0.833	0.917
ForgeLens [ProGAN]	0.793	0.951	0.103	0.712	0.414	0.908	0.931	0.957	0.931	0.962	1.000	0.988	0.792	0.815	0.889	0.962
ForgeLens [GenImage]	0.931	0.903	0.034	0.853	1.000	0.931	1.000	0.903	1.000	0.921	0.966	0.906	0.911	0.902	1.000	0.890
Effort [SD-v1.4]	0.552	0.842	0.345	0.843	0.586	0.787	0.690	0.850	0.793	0.868	0.897	0.797	0.822	0.783	1.000	0.844
Effort [Chameleon]	0.931	0.697	0.931	0.690	0.931	0.712	0.897	0.717	0.897	0.732	0.966	0.719	0.490	0.915	0.056	0.536
IAPL [ProGAN]	0.310	0.932	0.241	0.865	0.034	0.900	0.069	0.925	0.310	0.956	0.483	0.971	0.621	0.389	1.000	0.984
IAPL [SD-v1.4]	1.000	0.889	1.000	0.865	1.000	0.868	1.000	0.935	0.966	0.881	1.000	0.902	0.664	0.964	0.333	0.716

Table 49: **Human evaluator performance on FRAUDBENCH.** Results are reported under the single-image setting without review context.

Evaluator	GPT Image 2	Grok Imagine	Nano Banana 2	Wan2.7-Image	Qwen-Image-2.0	Qwen-Image-Edit	Overall			Real-Damaged
							Bal.Acc	F1	Acc	
Human Evaluators										
Human 1	0.483	0.577	0.596	0.613	0.696	0.695	0.610	0.728	0.610	0.611
Human 2	0.195	0.255	0.305	0.413	0.534	0.560	0.676	0.546	0.464	0.974
Human 3	0.582	0.695	0.717	0.767	0.859	0.871	0.772	0.839	0.755	0.795
Average	0.420	0.509	0.539	0.598	0.697	0.709	0.686	0.704	0.610	0.793

Table 50: **Performance comparison of MLLMs on FRAUDBENCH across damage types.** Results report TPR for synthetic fake-damaged images under the single-image setting without review context, pooled over the 27 Amazon categories and the six generator subsets. n denotes the number of synthetic fake-damaged images of each damage type. Hotels & Accommodations and Delivery, Pickup & Dine-Out use a separate defect taxonomy and are not included.

Models n	Cracked 1,158	Torn/Ripped 954	Packaging 672	Stained/Dirty 444	Leaking 390	Scratched 378	Shattered/Broken 210	Dented 186	Mouldy/Spoiled 138	Bent/Warped 132	Melted 126
Multimodal Large Language Models											
DeepSeek-V4.1-Flash	0.351	0.171	0.140	0.162	0.385	0.259	0.424	0.237	0.087	0.280	0.222
Gemini 3.1 Pro	0.288	0.138	0.193	0.257	0.356	0.243	0.286	0.204	0.109	0.386	0.246
Gemini 3.5 Flash-Lite	0.135	0.073	0.061	0.068	0.251	0.116	0.105	0.118	0.029	0.189	0.206
Gemini 3.8 Flash	0.176	0.032	0.049	0.068	0.056	0.103	0.090	0.086	0.014	0.053	0.063
GPT-6 Astra	0.534	0.191	0.107	0.277	0.221	0.275	0.348	0.323	0.348	0.227	0.254
GPT-6 Luna	0.895	0.889	0.848	0.881	0.956	0.910	0.990	0.876	0.964	0.803	0.960
GPT-6 Sol	0.404	0.217	0.106	0.264	0.195	0.241	0.162	0.183	0.210	0.159	0.190
Grok 4.6	0.234	0.061	0.106	0.088	0.244	0.079	0.110	0.048	0.072	0.061	0.349
Kimi K3	0.117	0.051	0.073	0.083	0.190	0.130	0.171	0.086	0.043	0.098	0.135
Qwen3-VL-Flash	0.443	0.089	0.043	0.146	0.259	0.209	0.400	0.038	0.174	0.045	0.127
Qwen3-VL-Plus	0.109	0.009	0.001	0.011	0.013	0.087	0.062	0.016	0.043	0.008	0.056
Qwen3.5-Omni-Flash	0.225	0.032	0.024	0.090	0.226	0.114	0.124	0.043	0.065	0.000	0.087
Qwen3.5-Omni-Plus	0.213	0.031	0.019	0.032	0.092	0.101	0.095	0.016	0.029	0.023	0.071
Qwen3.8-27B	0.246	0.091	0.103	0.164	0.213	0.233	0.310	0.140	0.116	0.220	0.341
Qwen3.8-Flash	0.437	0.176	0.234	0.291	0.495	0.344	0.424	0.333	0.159	0.394	0.421
Qwen3.8-Max	0.483	0.176	0.196	0.329	0.462	0.339	0.410	0.317	0.152	0.371	0.524
Average	0.331	0.152	0.144	0.201	0.288	0.236	0.282	0.192	0.163	0.207	0.266

Table 51: **Performance comparison of MLLMs on FRAUDBENCH using the Baseline prompt.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context. Results are computed on a stratified 10% sample of the 27 Amazon categories (579 images) for the six MLLMs in the prompt-sensitivity study, followed by ten additional MLLMs evaluated with the same prompt on the same images.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.057	0.696	0.160	0.680	0.275	0.716	0.273	0.711	0.356	0.740	0.480	0.723	0.610	0.402	0.954	0.774
Gemini 3.8 Flash	0.007	0.720	0.007	0.680	0.089	0.712	0.107	0.704	0.144	0.716	0.175	0.700	0.544	0.150	1.000	0.683
GPT-5.6 Terra	0.059	0.770	0.128	0.789	0.169	0.840	0.444	0.866	0.201	0.862	0.556	0.861	0.626	0.400	0.993	0.935
Grok 4.6	0.038	0.883	0.020	0.830	0.169	0.816	0.231	0.834	0.170	0.821	0.169	0.832	0.563	0.222	0.994	0.848
Kimi K3	0.028	0.590	0.007	0.720	0.052	0.742	0.141	0.848	0.121	0.776	0.237	0.787	0.549	0.171	1.000	0.847
Qwen3.8-Max	0.022	0.770	0.152	0.792	0.301	0.835	0.433	0.832	0.454	0.829	0.627	0.855	0.663	0.485	0.994	0.865
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.166	0.953	0.162	0.943	0.322	0.961	0.455	0.962	0.460	0.961	0.732	0.961	0.682	0.541	0.981	0.966
GPT-5.4 mini	0.019	0.930	0.063	0.932	0.125	0.940	0.119	0.935	0.071	0.938	0.229	0.940	0.552	0.178	1.000	0.918
Grok 4.1 Fast Reasoning	0.081	0.912	0.165	0.908	0.187	0.930	0.222	0.919	0.293	0.907	0.286	0.920	0.519	0.313	0.832	0.948
Grok 4.20 Reasoning	0.100	0.715	0.080	0.710	0.226	0.737	0.193	0.715	0.120	0.733	0.373	0.705	0.577	0.292	0.972	0.768
Kimi K2.6	0.049	0.885	0.015	0.850	0.086	0.878	0.230	0.883	0.262	0.897	0.288	0.901	0.578	0.255	1.000	0.914
QVQ-Max-Latest	0.194	0.754	0.228	0.742	0.308	0.746	0.269	0.747	0.289	0.762	0.302	0.786	0.506	0.381	0.746	0.918
Qwen3-VL-Flash	0.040	0.918	0.070	0.950	0.206	0.947	0.348	0.936	0.267	0.937	0.426	0.938	0.609	0.352	0.993	0.972
Qwen3-VL-Plus	0.000	—	0.007	0.950	0.025	0.935	0.075	0.968	0.057	0.946	0.088	0.947	0.521	0.075	1.000	0.977
Qwen3.6-35B-A3B	0.114	0.900	0.276	0.917	0.310	0.923	0.390	0.919	0.449	0.923	0.608	0.919	0.636	0.505	0.914	0.960
Qwen3.6-Plus	0.088	0.918	0.094	0.933	0.154	0.937	0.264	0.937	0.352	0.931	0.454	0.937	0.602	0.358	0.969	0.951

Table 52: **Performance comparison of MLLMs on FRAUDBENCH using the Merged prompt.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context. Results are computed on a stratified 10% sample of the 27 Amazon categories (579 images) for the six MLLMs in the prompt-sensitivity study, followed by ten additional MLLMs evaluated with the same prompt on the same images.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.177	0.849	0.359	0.865	0.327	0.860	0.404	0.896	0.461	0.889	0.573	0.926	0.541	0.526	0.698	0.887
Gemini 3.8 Flash	0.007	0.720	0.000	—	0.091	0.698	0.088	0.708	0.092	0.706	0.083	0.716	0.530	0.106	1.000	0.672
GPT-5.6 Terra	0.019	0.770	0.086	0.800	0.115	0.846	0.356	0.842	0.125	0.832	0.365	0.859	0.589	0.292	1.000	0.917
Grok 4.6	0.000	—	0.000	—	0.051	0.865	0.125	0.815	0.049	0.855	0.138	0.830	0.530	0.107	1.000	0.860
Kimi K3	0.000	—	0.133	0.751	0.128	0.702	0.160	0.747	0.177	0.751	0.351	0.778	0.565	0.263	0.972	0.809
Qwen3.8-Max	0.024	0.833	0.105	0.739	0.415	0.816	0.467	0.838	0.422	0.841	0.707	0.839	0.675	0.514	0.994	0.856
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.190	0.956	0.234	0.947	0.303	0.944	0.465	0.950	0.354	0.949	0.757	0.958	0.654	0.536	0.925	0.967
GPT-5.4 mini	0.000	—	0.015	0.905	0.086	0.930	0.144	0.948	0.046	0.945	0.219	0.960	0.542	0.150	1.000	0.895
Grok 4.1 Fast Reasoning	0.000	—	0.031	0.935	0.038	0.933	0.027	0.830	0.085	0.882	0.053	0.904	0.510	0.071	0.981	0.957
Grok 4.20 Reasoning	0.032	0.807	0.035	0.682	0.099	0.767	0.140	0.726	0.088	0.754	0.201	0.696	0.537	0.169	0.975	0.770
Kimi K2.6	0.009	0.850	0.044	0.884	0.072	0.935	0.206	0.918	0.230	0.912	0.267	0.919	0.569	0.229	1.000	0.914
QVQ-Max-Latest	0.211	0.767	0.279	0.772	0.399	0.769	0.338	0.737	0.343	0.757	0.488	0.765	0.451	0.468	0.559	0.924
Qwen3-VL-Flash	0.000	—	0.007	0.950	0.098	0.950	0.130	0.947	0.180	0.950	0.189	0.948	0.550	0.170	1.000	0.974
Qwen3-VL-Plus	0.000	—	0.007	0.920	0.077	0.950	0.110	0.946	0.110	0.935	0.196	0.938	0.542	0.144	1.000	0.976
Qwen3.6-35B-A3B	0.146	0.903	0.154	0.920	0.221	0.923	0.332	0.928	0.294	0.915	0.426	0.933	0.600	0.398	0.937	0.959
Qwen3.6-Plus	0.106	0.923	0.152	0.919	0.301	0.928	0.387	0.921	0.414	0.933	0.449	0.929	0.614	0.446	0.926	0.952

Table 53: **Performance comparison of MLLMs on FRAUDBENCH using the No Checklist prompt.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context. Results are computed on a stratified 10% sample of the 27 Amazon categories (579 images) for the six MLLMs in the prompt-sensitivity study, followed by ten additional MLLMs evaluated with the same prompt on the same images.

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.141	0.783	0.243	0.810	0.393	0.837	0.406	0.849	0.378	0.880	0.524	0.867	0.584	0.494	0.820	0.841
Gemini 3.8 Flash	0.012	0.720	0.000	—	0.090	0.700	0.126	0.703	0.122	0.705	0.137	0.704	0.541	0.139	1.000	0.679
GPT-5.6 Terra	0.046	0.810	0.074	0.763	0.217	0.804	0.430	0.856	0.314	0.846	0.569	0.856	0.637	0.416	1.000	0.941
Grok 4.6	0.026	0.905	0.031	0.885	0.112	0.822	0.183	0.835	0.129	0.836	0.124	0.823	0.550	0.170	1.000	0.846
Kimi K3	0.000	—	0.000	—	0.080	0.738	0.100	0.752	0.107	0.792	0.254	0.756	0.541	0.159	0.991	0.882
Qwen3.8-Max	0.070	0.840	0.085	0.816	0.275	0.832	0.422	0.836	0.463	0.843	0.680	0.852	0.666	0.489	1.000	0.868
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.095	0.959	0.147	0.949	0.244	0.953	0.356	0.960	0.336	0.959	0.686	0.958	0.638	0.452	0.965	0.967
GPT-5.4 mini	0.000	—	0.000	—	0.093	0.969	0.114	0.939	0.069	0.940	0.228	0.943	0.542	0.148	1.000	0.919
Grok 4.1 Fast Reasoning	0.059	0.907	0.083	0.946	0.123	0.932	0.112	0.912	0.201	0.922	0.220	0.906	0.562	0.219	0.991	0.952
Grok 4.20 Reasoning	0.069	0.764	0.044	0.693	0.263	0.717	0.191	0.745	0.171	0.726	0.358	0.714	0.546	0.289	0.909	0.736
Kimi K2.6	0.000	—	0.015	0.885	0.068	0.903	0.234	0.892	0.193	0.903	0.176	0.893	0.557	0.191	1.000	0.921
QVQ-Max-Latest	0.086	0.725	0.192	0.744	0.206	0.757	0.259	0.743	0.235	0.766	0.328	0.766	0.520	0.339	0.822	0.932
Qwen3-VL-Flash	0.012	0.950	0.026	0.950	0.191	0.943	0.233	0.941	0.229	0.942	0.301	0.944	0.579	0.269	0.993	0.968
Qwen3-VL-Plus	0.000	—	0.000	—	0.012	0.920	0.060	0.953	0.025	0.950	0.041	0.943	0.512	0.043	1.000	0.979
Qwen3.6-35B-A3B	0.108	0.944	0.168	0.909	0.238	0.917	0.294	0.928	0.275	0.921	0.401	0.919	0.610	0.385	0.972	0.958
Qwen3.6-Plus	0.069	0.918	0.026	0.930	0.124	0.926	0.348	0.934	0.259	0.937	0.522	0.936	0.605	0.349	0.985	0.954

Table 54: **Performance comparison of MLLMs on FRAUDBENCH using the Generic prompt.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context. Results are computed on a stratified 10% sample of the 27 Amazon categories (579 images) for the six MLLMs in the prompt-sensitivity study, followed by ten additional MLLMs evaluated with the same prompt on the same images. Responses that could not be parsed after all retries are excluded (Gemini 3 Flash: 1 of 579).

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.210	0.811	0.320	0.791	0.349	0.815	0.372	0.830	0.332	0.859	0.597	0.867	0.604	0.507	0.844	0.851
Gemini 3.8 Flash	0.000	—	0.000	—	0.096	0.703	0.044	0.720	0.069	0.705	0.122	0.714	0.528	0.098	1.000	0.714
GPT-5.6 Terra	0.000	—	0.072	0.750	0.060	0.808	0.295	0.889	0.051	0.820	0.258	0.872	0.561	0.207	1.000	0.941
Grok 4.6	0.007	0.910	0.000	—	0.071	0.822	0.081	0.800	0.066	0.796	0.098	0.803	0.527	0.095	1.000	0.850
Kimi K3	0.000	—	0.019	0.700	0.031	0.725	0.063	0.829	0.031	0.830	0.210	0.803	0.529	0.106	1.000	0.864
Qwen3.8-Max	0.035	0.787	0.086	0.745	0.259	0.810	0.425	0.835	0.281	0.846	0.642	0.828	0.641	0.433	0.994	0.885
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.177	0.948	0.070	0.960	0.207	0.951	0.355	0.959	0.306	0.955	0.596	0.953	0.633	0.428	0.981	0.962
GPT-5.4 mini	0.000	—	0.000	—	0.000	—	0.072	0.947	0.012	0.980	0.068	0.968	0.513	0.047	1.000	0.927
Grok 4.1 Fast Reasoning	0.000	—	0.007	0.880	0.000	—	0.000	—	0.000	—	0.031	0.865	0.503	0.012	1.000	0.961
Grok 4.20 Reasoning	0.038	0.777	0.020	0.675	0.083	0.769	0.104	0.709	0.069	0.703	0.196	0.695	0.530	0.149	0.975	0.772
Kimi K2.6	0.000	—	0.000	—	0.012	0.950	0.132	0.923	0.173	0.877	0.125	0.889	0.537	0.130	1.000	0.942
QVQ-Max-Latest	0.072	0.771	0.164	0.800	0.223	0.742	0.062	0.770	0.109	0.769	0.206	0.768	0.517	0.226	0.894	0.918
Qwen3-VL-Flash	0.000	—	0.000	—	0.060	0.950	0.092	0.950	0.117	0.950	0.117	0.950	0.528	0.111	0.993	0.969
Qwen3-VL-Plus	0.000	—	0.000	—	0.007	0.950	0.058	0.944	0.025	0.950	0.032	0.940	0.510	0.038	1.000	0.973
Qwen3.6-35B-A3B	0.022	0.950	0.048	0.871	0.103	0.930	0.175	0.940	0.154	0.930	0.135	0.932	0.547	0.181	0.988	0.958
Qwen3.6-Plus	0.012	0.920	0.024	0.907	0.049	0.945	0.149	0.931	0.123	0.940	0.204	0.932	0.547	0.157	1.000	0.955

Table 55: **Performance comparison of MLLMs on FRAUDBENCH using the Minimal prompt.** Results are reported across six AI image editing and generation subsets under the single-image setting without review context. Results are computed on a stratified 10% sample of the 27 Amazon categories (579 images) for the six MLLMs in the prompt-sensitivity study, followed by ten additional MLLMs evaluated with the same prompt on the same images. Responses that could not be parsed after all retries are excluded (Qwen3.6-35B-A3B: 2 of 579).

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.119	0.939	0.287	0.875	0.385	0.935	0.409	0.922	0.455	0.926	0.673	0.945	0.569	0.532	0.750	0.946
Gemini 3.8 Flash	0.026	0.720	0.000	—	0.051	0.690	0.109	0.701	0.085	0.703	0.059	0.720	0.527	0.094	1.000	0.717
GPT-5.6 Terra	0.000	—	0.116	0.826	0.054	0.840	0.471	0.883	0.121	0.848	0.386	0.890	0.591	0.312	0.991	0.927
Grok 4.6	0.020	0.875	0.007	0.820	0.083	0.825	0.100	0.855	0.084	0.828	0.117	0.813	0.534	0.119	1.000	0.852
Kimi K3	0.007	0.700	0.063	0.732	0.056	0.788	0.136	0.775	0.148	0.774	0.320	0.775	0.561	0.207	1.000	0.824
Qwen3.8-Max	0.024	0.763	0.096	0.766	0.220	0.830	0.433	0.855	0.334	0.837	0.717	0.844	0.649	0.452	0.994	0.871
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.152	0.935	0.207	0.915	0.331	0.930	0.331	0.952	0.443	0.953	0.662	0.944	0.648	0.505	0.941	0.957
GPT-5.4 mini	0.000	—	0.000	—	0.059	0.912	0.060	0.926	0.022	0.970	0.171	0.913	0.526	0.094	1.000	0.918
Grok 4.1 Fast Reasoning	0.012	0.950	0.000	—	0.019	0.950	0.029	0.917	0.068	0.924	0.086	0.925	0.518	0.064	1.000	0.953
Grok 4.20 Reasoning	0.071	0.754	0.064	0.680	0.103	0.709	0.096	0.729	0.138	0.699	0.305	0.712	0.542	0.219	0.954	0.733
Kimi K2.6	0.000	—	0.000	—	0.031	0.900	0.015	0.900	0.120	0.889	0.082	0.869	0.521	0.075	1.000	0.906
QVQ-Max-Latest	0.028	0.725	0.025	0.775	0.034	0.700	0.071	0.780	0.034	0.700	0.049	0.800	0.515	0.071	0.991	0.904
Qwen3-VL-Flash	0.000	—	0.007	0.950	0.048	0.950	0.083	0.950	0.140	0.950	0.064	0.950	0.528	0.100	1.000	0.962
Qwen3-VL-Plus	0.000	—	0.007	0.950	0.024	0.950	0.058	0.932	0.057	0.925	0.210	0.939	0.530	0.106	1.000	0.953
Qwen3.6-35B-A3B	0.048	0.938	0.038	0.925	0.138	0.918	0.184	0.913	0.153	0.925	0.217	0.927	0.546	0.218	0.963	0.956
Qwen3.6-Plus	0.009	0.950	0.015	0.925	0.091	0.946	0.200	0.929	0.161	0.938	0.241	0.931	0.560	0.204	1.000	0.956

Table 56: **Performance comparison of MLLMs on FRAUDBENCH under the mismatched-review condition.** Results are reported across six AI image editing and generation subsets under the single-image setting with review context. Each image is paired with a randomly sampled review from a different product category. Results are computed on the same stratified 10% sample of the 27 Amazon categories (579 images) used in the prompt-sensitivity study. The second block lists nine additional MLLMs evaluated on the same images with the same mismatched reviews. Responses that could not be parsed after all retries are excluded (Gemini 3 Flash: 1, Qwen3.6-35B-A3B: 4 of 579).

Models	GPT Image 2		Grok Imagine		Nano Banana 2		Wan2.7-Image		Qwen-Image-2.0		Qwen-Image-Edit		Overall		Real-Damaged	
	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	TPR	Conf.	Bal.Acc.	F1	TNR	Conf.
Multimodal Large Language Models																
DeepSeek-V4.1-Flash	0.137	0.724	0.282	0.680	0.374	0.711	0.478	0.709	0.417	0.715	0.573	0.722	0.602	0.525	0.828	0.747
Gemini 3.8 Flash	0.007	0.720	0.027	0.700	0.110	0.713	0.151	0.727	0.185	0.696	0.195	0.699	0.556	0.186	1.000	0.619
GPT-5.6 Terra	0.037	0.810	0.106	0.813	0.247	0.850	0.363	0.885	0.225	0.844	0.585	0.893	0.630	0.396	1.000	0.947
Grok 4.6	0.038	0.883	0.052	0.833	0.100	0.828	0.163	0.869	0.158	0.848	0.216	0.839	0.561	0.199	1.000	0.858
Kimi K3	0.000	—	0.041	0.660	0.055	0.772	0.152	0.817	0.140	0.709	0.296	0.793	0.525	0.191	0.935	0.866
Qwen3.8-Max	0.022	0.825	0.128	0.797	0.263	0.823	0.331	0.833	0.377	0.848	0.678	0.845	0.650	0.447	1.000	0.876
Additional Multimodal Large Language Models																
Gemini 3 Flash	0.120	0.960	0.199	0.956	0.271	0.952	0.497	0.958	0.569	0.961	0.801	0.964	0.685	0.569	0.960	0.975
GPT-5.4 mini	0.034	0.913	0.028	0.950	0.077	0.945	0.107	0.938	0.101	0.928	0.198	0.958	0.545	0.156	1.000	0.923
Grok 4.1 Fast Reasoning	0.313	0.902	0.288	0.914	0.356	0.906	0.448	0.913	0.443	0.914	0.528	0.902	0.560	0.525	0.724	0.935
Grok 4.20 Reasoning	0.260	0.672	0.231	0.710	0.420	0.700	0.396	0.712	0.381	0.702	0.598	0.694	0.630	0.529	0.878	0.710
QVQ-Max-Latest	0.103	0.769	0.188	0.777	0.210	0.775	0.315	0.762	0.209	0.754	0.255	0.789	0.520	0.329	0.826	0.923
Qwen3-VL-Flash	0.049	0.946	0.115	0.940	0.289	0.950	0.359	0.944	0.318	0.943	0.448	0.947	0.628	0.398	0.993	0.969
Qwen3-VL-Plus	0.000	—	0.000	—	0.009	0.980	0.072	0.960	0.043	0.970	0.078	0.961	0.508	0.062	0.981	0.980
Qwen3.6-35B-A3B	0.142	0.902	0.249	0.910	0.311	0.919	0.320	0.935	0.418	0.928	0.567	0.922	0.604	0.472	0.875	0.961
Qwen3.6-Plus	0.145	0.911	0.080	0.904	0.199	0.937	0.354	0.928	0.311	0.910	0.537	0.919	0.600	0.406	0.929	0.952

Table 57: **Matched versus mismatched review text.** Balanced accuracy, F1, and real-image TNR on the same 579 images (stratified 10% sample of the 27 Amazon categories) without review text, with the matched review, and with a review drawn from a different product category. The first block lists the six MLLMs of the mismatch study, and the second block lists nine additional MLLMs evaluated on the same images with the same mismatched reviews. The last three rows average the first block, the second block, and all 15 MLLMs.

Models	No Review			Matched Review			Mismatched Review		
	Bal.Acc.	F1	TNR	Bal.Acc.	F1	TNR	Bal.Acc.	F1	TNR
Multimodal Large Language Models									
DeepSeek-V4.1-Flash	0.608	0.385	0.966	0.660	0.576	0.896	0.602	0.525	0.828
Gemini 3.8 Flash	0.541	0.139	1.000	0.633	0.393	1.000	0.556	0.186	1.000
GPT-5.6 Terra	0.631	0.407	1.000	0.599	0.324	1.000	0.630	0.396	1.000
Grok 4.6	0.567	0.225	1.000	0.571	0.234	1.000	0.561	0.199	1.000
Kimi K3	0.547	0.207	0.974	0.583	0.272	1.000	0.525	0.191	0.935
Qwen3.8-Max	0.675	0.507	1.000	0.683	0.521	1.000	0.650	0.447	1.000
Additional Multimodal Large Language Models									
Gemini 3 Flash	0.682	0.541	0.981	0.713	0.584	1.000	0.685	0.569	0.960
GPT-5.4 mini	0.551	0.176	1.000	0.515	0.056	1.000	0.545	0.156	1.000
Grok 4.1 Fast Reasoning	0.530	0.288	0.873	0.582	0.387	0.898	0.560	0.525	0.724
Grok 4.20 Reasoning	0.557	0.255	0.960	0.574	0.282	0.975	0.630	0.529	0.878
QVQ-Max-Latest	0.526	0.394	0.777	0.556	0.354	0.881	0.520	0.329	0.826
Qwen3-VL-Flash	0.597	0.328	0.986	0.679	0.519	0.986	0.628	0.398	0.993
Qwen3-VL-Plus	0.522	0.080	1.000	0.505	0.021	1.000	0.508	0.062	0.981
Qwen3.6-35B-A3B	0.590	0.490	0.833	0.572	0.356	0.909	0.604	0.472	0.875
Qwen3.6-Plus	0.602	0.361	0.969	0.578	0.255	1.000	0.600	0.406	0.929
Average (six MLLMs)	0.595	0.312	0.990	0.622	0.387	0.983	0.587	0.324	0.961
Average (nine additional MLLMs)	0.573	0.324	0.931	0.586	0.313	0.961	0.587	0.383	0.907
Average (all 15 MLLMs)	0.582	0.319	0.955	0.600	0.342	0.970	0.587	0.359	0.929

D.4 Supplementary Evaluation Explorations

This section presents eight supplementary explorations that extend the primary results along axes not fully captured by the main benchmark settings. Appendix D.4.1 examines how detection difficulty varies systematically across the 29 product and service categories in FraudBench. Appendix D.4.2 analyzes how the physical form of the claimed defect shapes detectability. Appendix D.4.3 probes whether review-conditioned detection reflects genuine cross-modal claim verification or weaker semantic shortcuts. Appendix D.4.4 quantifies the sensitivity of MLLM detection performance to prompt formulation. Appendix D.4.5 evaluates the specialized detectors without a fixed decision threshold, and Appendix D.4.6 evaluates image forgery detection and localization methods. Appendix D.4.7 compares how MLLMs and IFDL methods localize the manipulated region, and Appendix D.4.8 examines whether MLLMs and specialized detectors detect the same fake-damaged images. Together, these explorations provide a more granular characterization of current detector limitations and motivate targeted directions for future work.

D.4.1 Category-Level Detection Difficulty

The per-category breakdown across Tab. 20–48 reveals that detection difficulty varies substantially and systematically across product and service categories. Averaged across the 16 MLLMs in Tab. 1, category-level balanced accuracy spans from 0.534 in Baby Products to 0.646 in Hotels & Accommodations, with the middle half of categories between 0.562 and 0.604. The per-category tables also list the 14 other evaluated MLLMs in a separate block, which is not included in these averages. This spread is sufficiently large to affect practical conclusions about model competence and indicates that FraudBench probes genuine category-dependent variation in forensic difficulty.

Categories with more structurally distinctive or semantically grounded evidence submissions tend to be easier for current MLLMs. Hotels & Accommodations, All Beauty, and Musical Instruments represent the higher end of the performance distribution (0.646, 0.630, and 0.628). In hotel evidence, damage claims typically involve recognizable indoor fixtures such as bedding, bathroom fittings, and walls, whose physical states carry clear expectations and whose AI-generated manipulations produce spatially localized artifacts against a structured background. Beauty products and musical

instruments similarly feature highly constrained visual elements, including printed labels, typography, and regular object geometry, which make AI-generated edits to surfaces, seams, or labels more identifiable through local consistency checks. For these categories, the presence of text, branded visual elements, and regular geometry provides MLLMs with forensic anchors that are absent in more visually heterogeneous categories.

By contrast, categories such as Baby Products (0.534), Tools & Home Improvement (0.536), Delivery, Pickup & Dine-Out (0.539), and Amazon Fashion (0.553) cluster at the lower end of performance. Baby product evidence predominantly consists of plush toys, soft packaging, and infant accessories, whose surfaces lack the sharp geometric regularity or semantic distinctiveness needed to make AI-generated damage boundaries stand out. Apparel evidence in Amazon Fashion is similarly affected: fabric textures, natural wrinkles, and variable lighting conditions create substantial within-category appearance diversity, making it difficult to distinguish AI-generated damage artifacts from authentic surface variation. The Delivery, Pickup & Dine-Out category poses a distinct challenge: food presentation varies greatly across vendors and occasions, and damage patterns such as crushed packaging, contamination, or spillage can be visually subtle or ambiguous against backgrounds that are themselves cluttered and uncontrolled.

Model-level patterns further illuminate these category effects. Among the 16 MLLMs in Tab. 1, no single model dominates: Qwen3.8-Max achieves the highest balanced accuracy, alone or tied, in 11 of the 29 categories, GPT-6 Astra in 10, and Qwen3.8-Flash in six. Even the strongest model varies widely across categories: Qwen3.8-Max reaches 0.777 in All Beauty and 0.744 in Hotels & Accommodations but only 0.542 in Baby Products and 0.589 in Delivery, Pickup & Dine-Out, a gap of more than 20 percentage points. These model-by-category interactions suggest that category-specific forensic competence is not captured by a single aggregate metric and that different model architectures may respond differently to the visual forensic cues dominant in each category.

These results carry a practical implication for evaluating refund-evidence detectors. Because product category is not independent of damage type, generator behavior, or evidence style, aggregate benchmarking can mask systematic category-level blind spots. A detector that performs acceptably on average may nevertheless be unreliable precisely for categories such as fashion, food delivery, and everyday household goods, which represent the highest volume and most heterogeneous distribution of real platform disputes. Future work should therefore stratify evaluations by category as a first-class axis, and prioritize improvement on categories with heterogeneous object appearance, unconstrained imaging conditions, and damage patterns that overlap with natural product variation.

D.4.2 Damage-Type Sensitivity

The damage-type breakdown reveals that detection difficulty is not uniformly distributed across fake-damage categories: the physical form of the claimed defect systematically shapes how detectable it is for current MLLMs. As shown in Tab. 50, structurally salient or materially discontinuous defects are generally more accessible to MLLM detection. Averaged across the 16 MLLMs in Tab. 1, cracked, leaking, shattered/broken, and melted evidence yield average TPRs of 0.331, 0.288, 0.282, and 0.266. By contrast, packaging-damaged, torn/ripped, mouldy/spoiled, and dented evidence remain substantially harder, with average TPRs of 0.144, 0.152, 0.163, and 0.192. This performance gap suggests that current MLLMs are more responsive to obvious object-level breakage, material discontinuities, or fluid-related damage than to subtle deformation, surface-level complaints, or packaging-level defects.

The difficulty further varies across models. GPT-6 Astra achieves 0.534 TPR on cracked evidence and Qwen3.8-Max 0.524 on melted evidence, while multiple models approach zero TPR on packaging-damaged samples; Qwen3-VL-Plus, for instance, obtains only 0.001 TPR on this damage type. GPT-6 Luna reaches high TPR on every damage type (0.803 to 0.990) only because it flags most images as AI-generated, including most genuine damaged evidence (TNR 0.206). These results indicate that detector performance depends not only on the image generator or product category, but also on the physical manifestation of the claimed defect, independently of other evaluation axes.

This finding carries direct implications for refund-evidence verification in practice. Many genuine disputes involve subtle or localized damage such as packaging deformation, shallow dents, stains, or minor tears, rather than visually dramatic breakage. A detector that performs reliably

only on conspicuous defects therefore risks overestimating its practical utility in real platform settings. Future work should treat damage type as a first-class evaluation dimension, examined in conjunction with generator identity, product category, and input modality. Promising directions include material-aware visual reasoning that accounts for object composition and physical properties, damage-plausibility modeling grounded in mechanical and structural constraints, and local comparison methods that evaluate whether a claimed defect is geometrically and physically consistent with the object and scene context.

D.4.3 IMAGE-TEXT MISMATCH

The image–text mismatch experiment is designed to test whether review-conditioned detection reflects genuine cross-modal claim verification or merely exploits weaker semantic shortcuts present in the review text. In the standard review-conditioned setting, each image is paired with the customer review from the same product or service instance, enabling models to integrate both visual forensic cues and semantic consistency between the image and the stated claim. In the mismatched-review condition, the matched review is replaced with a genuine review sampled uniformly at random from a different Amazon product category, while the image input, prompt structure, and output schema are held fixed. Because every image then receives a genuine review, the condition also removes the stylistic difference between generated and genuine reviews. This controlled design provides a diagnostic for whether model predictions are sensitive to semantic coherence between review text and visual content, rather than driven by visual evidence alone.

As shown in Tab. 57 and detailed in Tab. 56, the benefit of review text disappears once the review no longer matches the evidence. Averaged over the six MLLMs, balanced accuracy on the same 579 images is 0.595 without review text, 0.622 with the matched review, and 0.587 with a mismatched review. Gemini 3.8 Flash illustrates the pattern most clearly, rising from 0.541 to 0.633 with the matched review but only to 0.556 with a mismatched one. Mismatched reviews can also harm genuine evidence: the TNR of DeepSeek-V4.1-Flash falls from 0.966 without review text to 0.828 under mismatch and that of Kimi K3 from 0.974 to 0.935, indicating that these models treat an incoherent review as a sign of manipulation. GPT-5.6 Terra is the exception, performing worse with the matched review (0.599) than without review text (0.631) or with a mismatched review (0.630).

The nine additional MLLMs in the second block of Tab. 57 and 56 depend on the review even less. The matched review raises their average balanced accuracy only from 0.573 to 0.586, and a mismatched review yields a similar value (0.587). The mismatched review nevertheless lowers their average real-image TNR from 0.931 without review text to 0.907, as five of the nine models flag more genuine evidence.

Comparing the mismatched condition against the standard review-conditioned setting further clarifies the role of textual context. In the primary results, adding review text produces only a marginal average improvement in single-image balanced accuracy (0.591 to 0.619 for the six MLLMs in Tab. 2). Since the matched gain vanishes when both classes receive genuine reviews, it may reflect either image–text coherence or the stylistic difference between generated and genuine reviews, suggesting that current MLLMs do not consistently anchor their decisions in semantic alignment between the review and the visual evidence. The mismatched condition probes the complementary failure mode: whether a model is inappropriately influenced by plausible but semantically irrelevant review text. A reliable claim-verification model should neither treat coherent review text as confirmatory evidence nor discard textual context entirely; instead, it should assess whether the submitted visual evidence specifically supports the stated claim. These results motivate future evaluation protocols that include explicit image–text consistency stress tests: hard negative reviews from visually similar product categories, partially correct reviews that name the correct object but the wrong defect type, and adversarial reviews that plausibly describe a defect absent from the image. Such designs would better distinguish generic AI-image detection from genuine refund-evidence verification, where the core requirement is to determine whether specific visual evidence substantiates a specific textual claim.

D.4.4 PROMPT SENSITIVITY

The prompt-sensitivity study demonstrates that MLLM-based forensic detection is sensitive to prompt formulation, yet prompt engineering alone is insufficient to resolve the difficulty of fake-

damaged evidence verification. As summarized in Tab. 3 and detailed in Tab. 51–55, no single prompt is best for all six models, and even the best configuration reaches an F1 of only 0.532.

Replacing the forensic role with a generic assistant identity is the most consistently harmful change: it lowers F1 for five of the six models, for example from 0.400 to 0.207 for GPT-5.6 Terra and from 0.222 to 0.095 for Grok 4.6. The Baseline prompt yields the highest F1 for Gemini 3.8 Flash (0.150) and Grok 4.6 (0.222), and the second-highest for GPT-5.6 Terra (0.400, behind 0.416 under No Checklist).

The trend is not monotonic across models. Merging the role instructions into the user turn raises F1 for Kimi K3 from 0.171 to 0.263. DeepSeek-V4.1-Flash attains its highest F1 under the Minimal (0.532) and Merged (0.526) prompts, but these gains come from flagging more genuine evidence: its real-image TNR falls from 0.954 under Baseline to 0.750 and 0.698, and its balanced accuracy is highest under Baseline (0.610). Consequently, the ablation does not identify a universally optimal template; rather, it reveals that prompt structure itself constitutes a measurable source of evaluation variance, with effects that differ across model families. The ten additional MLLMs in the second block of Tab. 51–55, evaluated with the same five prompts on the same 579 images, show the same pattern. The generic assistant identity lowers F1 for all ten, the Baseline prompt yields the highest F1 for seven of them and the Merged prompt for the other three, and the highest F1 among them is 0.541, reached by Gemini 3 Flash with the Baseline prompt.

D.4.5 Threshold-Free Evaluation of Specialized Detectors

The threshold-free evaluation reveals that the ranking of specialized detectors depends on how their operating point is chosen. Tab. 1 reports every specialized detector at its default decision threshold of 0.5, whereas refund verification assigns different costs to false positives and false negatives. We therefore compute AUROC, the standardized partial AUROC for FPR ≤ 0.1, and the TPR at fixed false-positive rates of 1%, 5%, and 10% from the detector scores on the same 6,940 images, as reported in Tab. 58.

The ranking by AUROC agrees closely with the balanced-accuracy ranking in Tab. 1 (Spearman $\rho = 0.929$) and with the ranking by TPR at 5% FPR ($\rho = 0.976$), but it is nearly unrelated to the ranking by TPR at the default threshold ($\rho = 0.071$). IAPL [SD-v1.4] illustrates this gap most clearly: it achieves the highest TPR at the default threshold (0.956) but flags 55.2% of the real-damaged evidence, and its TPR falls to 0.275 once the FPR is held at 5%. Effort [Chameleon] reaches an AUROC of only 0.399, meaning that its scores rank real-damaged evidence above fake-damaged evidence more often than not. ForgeLens [GenImage] is the strongest detector under every threshold-free metric, with an AUROC of 0.913 and a TPR of 0.840 at 5% FPR.

These results indicate that the default-threshold TPR alone overstates the practical utility of over-sensitive detectors, and that detector comparisons for refund verification should be reported at operating points with a bounded false-positive rate. MLLMs return a binary verdict with a coarse self-reported confidence rather than a calibrated score, so we report them only at their fixed operating points in Tab. 1. Since real and synthetic evidence also differ in file-level properties such as image size and format, the detector scores may partly reflect such properties.

Table 58: **Threshold-free evaluation of specialized detectors.** Metrics are computed on the detector scores of all 6,940 single-image samples, pooled over categories and generators. $\text{pAUC}_{0.1}$ is the standardized partial AUROC for FPR ≤ 0.1, TPR@k% is the TPR at a false-positive rate of k% on real-damaged evidence (i.e., TNR $= 1 - k$%), and FPR@0.5 and TPR@0.5 describe the default threshold used in Tab. 1. A lower FPR is better.

Methods	AUROC	$\text{pAUC}_{0.1}$	TPR@1%	TPR@5%	TPR@10%	FPR@0.5	TPR@0.5
CO-SPY [ProGAN]	0.902	0.813	0.400	0.700	0.836	0.032	0.599
CO-SPY [SD-v1.4]	0.492	0.494	0.002	0.041	0.083	0.308	0.309
ForgeLens [ProGAN]	0.888	0.817	0.519	0.687	0.767	0.046	0.684
ForgeLens [GenImage]	**0.913**	**0.898**	**0.711**	**0.840**	**0.855**	0.057	0.846
Effort [SD-v1.4]	0.860	0.743	0.338	0.559	0.656	0.168	0.732
Effort [Chameleon]	0.399	0.496	0.012	0.043	0.081	0.947	0.875
IAPL [ProGAN]	0.787	0.572	0.038	0.201	0.342	0.100	0.332
IAPL [SD-v1.4]	0.837	0.622	0.102	0.275	0.482	0.552	0.956

D.4.6 IMAGE FORGERY DETECTION AND LOCALIZATION BASELINES

The IFDL evaluation reveals that existing forgery detection and localization methods do not solve FRAUDBENCH either. As shown in Tab. 59, all six IFDL checkpoints remain at or below chance, with balanced accuracy between 0.468 and 0.498, whereas the AI-generated image detectors in Tab. 1 range from 0.452 to 0.904.

Their failure modes mirror the checkpoint dependence observed for specialized detectors. MVSS-Net [CASIA] flags almost every image as manipulated, with TPR between 0.890 and 0.989 but a TNR of only 0.034, whereas MVSS-Net [DEFACTO] flags few images, with TPR between 0.112 and 0.181 and a TNR of 0.836. TruFor and FakeShield retain most real-damaged evidence, with TNR of 0.825 and 0.823, but detect at most 0.199 and 0.181 of the fake-damaged images of any generator. These methods are trained on spliced, copy-moved, or inpainted manipulations, whose traces differ from those of generative editing, in which the generator re-renders the whole frame. This suggests that broader manipulation coverage alone is insufficient, and that forgery detection for refund evidence requires exposure to generative editing.

Table 59: **Image forgery detection and localization (IFDL) baselines on FRAUDBENCH.** Results are reported under the single-image setting without review context and macro-averaged over the 29 categories, as in Tab. 1. The six generator columns report TPR. The strongest AI-generated image detector from Tab. 1 is listed for reference.

Methods	GPT Image 2	Grok Imagine	Nano Banana 2	Wan2.7-Image	Qwen-Image-2.0	Qwen-Image-Edit	Overall		Real-Damaged
	TPR	TPR	TPR	TPR	TPR	TPR	Bal.Acc.	F1	TNR
MVSS-Net [CASIA]	0.987	0.964	0.972	0.989	0.953	0.890	0.496	0.897	0.034
MVSS-Net [DEFACTO]	0.112	0.147	0.134	0.122	0.181	0.115	0.486	0.229	0.836
ResFCN [CASIA]	0.791	0.805	0.766	0.753	0.764	0.705	0.498	0.799	0.232
ResFCN [DEFACTO]	0.487	0.450	0.429	0.396	0.556	0.390	0.481	0.582	0.511
TruFor	0.047	0.059	0.043	0.196	0.199	0.121	0.468	0.189	0.825
FakeShield	0.178	0.181	0.158	0.181	0.157	0.156	0.496	0.274	0.823
ForgeLens [GenImage]	0.967	0.158	1.000	0.990	1.000	0.989	0.904	0.915	0.957

D.4.7 LOCALIZATION OF THE MANIPULATED REGION

The localization evaluation reveals that neither MLLMs nor IFDL methods reliably locate the edited region, and that the two families point at different parts of the image. As shown in Tab. 60, Gemini 3 Flash reaches a mean IoU of 0.208 with the manipulated area on the 786 images it flags with a box, compared with 0.087 for FakeShield on the same images and at most 0.101 for the IFDL checkpoints scored on every fake-damaged image. Its advantage partly reflects coarse boxes, which cover 38.8% of the frame on average, whereas the masks of FakeShield cover only 5.5%. Neither method tracks the extent of the edit, as the correlation between the predicted and the true area is -0.055 for Gemini 3 Flash and -0.040 for FakeShield. Both families also localize manipulations that do not exist: Gemini 3 Flash returns a box for 289 of the 1,012 real-damaged images, and FakeShield returns a mask for 155. The IoU of MVSS-Net [CASIA] changes gradually with the size of the ground truth, from 0.066 to 0.101 and 0.128 as it takes 30%, 50%, and 70% of the difference.

The two families also point at different parts of the image. On the 786 images localized by both methods, the IoU between the Gemini 3 Flash box and the FakeShield mask is only 0.123 on average (median 0.032), the two predictions do not overlap at all for 19.0% of the images, and the per-image IoUs of the two methods are only weakly correlated (Pearson $r = 0.346$). Together with the complementarity analysis in Appx. D.4.8, this suggests that MLLMs and forensic methods rely on different evidence, rather than being weaker and stronger versions of the same detector.

These results indicate that neither family yet provides the region-level evidence that a refund decision would require. Region-level annotations such as those in Appx. C.4 allow future work to evaluate localization and claim grounding directly, rather than only the binary verdict.

D.4.8 COMPLEMENTARITY OF MLLMS AND SPECIALIZED DETECTORS

The complementarity analysis reveals that MLLMs and specialized detectors succeed on largely different fake-damaged images. For each pair of an MLLM in Tab. 1 and one of the 14 detector checkpoints in Tab. 1 and 59, we compare over the 5,928 fake-damaged images whether each method detects each image. Across the resulting 224 pairs, the two outcomes are nearly independent, with a

Table 60: **Localization of the manipulated region.** All methods are scored against the same ground truth, the smallest set of pixels carrying half of the smoothed difference between each fake-damaged image and its real-undamaged source, which covers 0.178 of the frame on average. n is the number of scored images. FakeShield is scored on the 957 fake-damaged images for which it returns a mask, Gemini 3 Flash on the 786 of them that it also flags with a box, and MVSS-Net and ResFCN on every fake-damaged image. Area is the mean fraction of the frame covered by the prediction.

Methods	**Output**	n	**Mean IoU**	**Median IoU**	**Pixel F1**	**Area**
Gemini 3 Flash	Box	786	0.208	0.191	0.328	0.388
FakeShield	Mask	957	0.086	0.026	0.137	0.059
FakeShield (same images as Gemini 3 Flash)	Mask	786	0.087	0.026	0.137	0.055
MVSS-Net [CASIA]	Mask	5,928	0.101	0.071	0.168	–
MVSS-Net [DEFACTO]	Mask	5,928	0.006	0.000	0.010	–
ResFCN [CASIA]	Mask	5,928	0.024	0.001	0.041	–
ResFCN [DEFACTO]	Mask	5,928	0.012	0.000	0.020	–

median ϕ coefficient of 0.031 and $|\phi| < 0.1$ for 193 pairs. For example, Qwen3.8-Max detects 174 of the 910 fake-damaged images missed by ForgeLens [GenImage].

Complementary errors do not, however, translate into large gains under simple fusion. Flagging an image whenever either Qwen3.8-Max or ForgeLens [GenImage] flags it raises the pooled balanced accuracy of ForgeLens [GenImage] by only 0.010, because the additional detections are accompanied by additional false positives on real-damaged evidence. This motivates fusion strategies that weigh the two families according to the evidence each relies on, rather than combining their verdicts directly.

E SCOPE OF FRAUDBENCH AND THE ROLE OF MLLMS

What FRAUDBENCH Measures. FRAUDBENCH asks whether a refund claim is supported by genuine visual evidence, which requires three capabilities. The first is ***visual verification***, i.e., deciding whether a photograph of damage is genuine when the negative class consists of real damaged evidence rather than clean photographs. Every detector family can be evaluated on this capability, and it is the one measured in Tab. 1. The second is ***claim–image consistency***, i.e., checking that the evidence shows the damage the customer describes, which requires reading the review text. The third is ***multi-view aggregation*** across the photographs submitted with one claim. Specialized AI-generated image detectors and IFDL methods score one image at a time and take no text, so only MLLMs can be evaluated on the second and third capabilities, and only MLLMs return a rationale that a platform could present to a customer whose claim is rejected. Following prior forensic benchmarks (Wang et al., 2025; Du et al., 2025), we therefore evaluate MLLMs alongside specialized detectors, rather than because we expect them to be stronger forensic classifiers.

Where Current Models Fall Short. Our results place the current bottleneck at visual verification. In the single-image setting, the 16 MLLMs in Tab. 1 retain genuine damaged evidence with an average TNR of 0.938 but detect only 0.230 of the fake-damaged images. Review text is not yet used as a claim to be verified. Across the 18 MLLMs evaluated with and without review text (Tab. 12 and 13), the average balanced accuracy rises only from 0.588 to 0.606, with 12 models improving and six declining, and on the mismatch subset the gain from the matched review (0.595 to 0.622) disappears once the review describes a different product (0.587, Appx. D.4.3). Multiple views are used more effectively. Without review text, the average balanced accuracy of the six MLLMs in Tab. 2 rises from 0.591 for single images to 0.621 for jointly presented images and 0.700 in the multi-step setting (Tab. 12, 14, and 18).

Interpretable Rationales. A single example illustrates both the value and the risk of MLLM rationales. For a crack that Grok Imagine added to a translucent plastic case (Electronics, Review_010), Gemini 3 Flash flags the image with a confidence of 0.95, stating that the crack "exhibits physically impossible geometry, appearing as a solid black void rather than a fracture in translucent material", whereas ForgeLens [GenImage] assigns it a score of 0.058. Qwen3.8-Max and GPT-6 Astra accept the same image with the opposite argument, e.g., that "the dark sofa shows through the puncture hole exactly where expected". Rationales thus make an MLLM decision checkable, but a fluent rationale can still be wrong, which motivates evaluating rationale faithfulness in future work.

F LIMITATIONS

FRAUDBENCH provides a controlled benchmark for AI-generated refund-evidence verification, but its scope is necessarily bounded.

The benchmark covers selected e-commerce, food-delivery, travel-service, and self-collected scenarios, and therefore reflects the platforms, categories, languages, and review-media availability included in our collection. It should not be interpreted as covering the full distribution of refund disputes, complaint behaviors, platform policies, or user-submitted evidence formats.

Another limitation concerns the synthetic threat model. The fake-damaged evidence is generated using 12 image editing and generation models under a shared prompt-construction pipeline. This design enables controlled comparison across generators and damage patterns, but it cannot exhaustively represent future generative models, post-processing strategies, prompt optimization methods, or human-assisted manipulation workflows. As generative models and editing interfaces evolve, detector performance and relative model rankings may change.

The evaluation setting is also narrower than real refund adjudication. Our experiments focus on visual and review-conditioned evidence verification, whereas real platform decisions may additionally involve order histories, logistics records, seller responses, transaction metadata, platform policies, and human communication. Accordingly, results on FRAUDBENCH should be interpreted as evidence of model behavior under the benchmarked verification settings, not as proof of deployment readiness for autonomous refund decisions.

Finally, several evaluated MLLMs are proprietary API-based systems. Although we report model identifiers, prompt templates, access details, and evaluation settings, exact reproducibility may be affected by provider-side model updates, backend changes, request-size limits, safety filters, or changes in default inference behavior. We therefore treat the reported API-based results as a snapshot of model behavior under the specified evaluation period and settings.

G Compute Reporting

We report the computational resources used for constructing and evaluating **FraudBench**. The study does not train models from scratch, fine-tune MLLMs, or conduct hyperparameter search. Local computation is used for data preprocessing, image downloading and filtering, synthetic-image generation orchestration, specialized-detector inference, and evaluation scripts. MLLM evaluation is performed through official proprietary API services; therefore, provider-side inference compute is not directly observable from our local hardware logs.

Local Compute Environment. As mentioned in Appx. D.1.2, all local experiments are conducted on a single Alibaba Cloud Elastic Compute Service (ECS) instance, `ecs.gn7e-c16g1.4xlarge`. The instance is equipped with an Intel Xeon Platinum 8369B CPU exposing 16 available vCPUs, 125 GiB memory, 2,048 GiB storage, and one NVIDIA A100 SXM GPU with 80 GB memory. We do not use distributed training, multi-node computation, or multi-GPU computation.

Local Runtime. The approximate local runtime is 6 hours for data preprocessing, 6 hours for image downloading and filtering, 24 hours for synthetic-image generation, and 1 GPU-hour for specialized-detector evaluation. These estimates describe the local execution time of our benchmark construction and evaluation pipeline and exclude provider-side compute for proprietary API inference.

API-Based MLLM Evaluation. We evaluate 30 MLLMs through official API services. The reported results rest on 510,277 successful API calls: 208,200 single-image review-free calls (30 MLLMs $\times$ 6,940 images), 124,920 single-image review-conditioned calls (18 MLLMs), 27,024 multi-image review-free and 27,024 multi-image review-conditioned calls (eight MLLMs on all 29 categories and ten further MLLMs on the 27 Amazon categories, one call per review), 34,056 multi-step review-free and 34,056 multi-step review-conditioned calls (six MLLMs, one call per image turn), 46,317 prompt-ablation calls (16 MLLMs), and 8,680 mismatched-review calls (15 MLLMs). The multi-image runs of these 18 MLLMs additionally issued 34,564 calls on single-image reviews of the reported categories, which are excluded from the multi-image metrics. By provider, the successful calls comprise 274,937 for Alibaba Cloud DashScope-hosted models, 97,108 for OpenAI-hosted models, 66,010 for Google-hosted models, and 72,222 for xAI-hosted models.

Retries and Failures. Our scripts retry transient API errors up to three times with exponential back-off and, for request-size and image-fetch failures, re-encode the images at up to ten progressively smaller resolution tiers, as detailed in Appx. D.1.1. The stored records show at least 5,596 additional attempts beyond the successful calls, about 1.1%; attempts made before a failed record was re-run are not logged, so this is a lower bound. Calls that still failed after all retries were reissued in a later pass until every reported record had a valid response, except for eight responses in the earlier prompt-ablation and mismatch runs of the additional MLLMs, which could not be parsed and are excluded from the corresponding tables.

Training, Tuning, and Ablations. No model is trained from scratch or fine-tuned in this study. No hyperparameter search is performed. The additional model-call experiments consist of fixed prompt-ablation variants and image-text mismatch evaluation. Both are conducted through API-based inference and do not involve parameter updates or model selection.

H IMPACT STATEMENT

Intended Use and Deployment Scope. This benchmark is intended to support research on detecting AI-generated or AI-modified visual evidence in refund- and complaint-related review scenarios. Its primary use is controlled evaluation of detector robustness, failure modes, and multimodal evidence verification. It is not designed to serve as an automated decision system for accepting or rejecting refund claims, moderating user reviews, or making consequential decisions about individual users. Model outputs should therefore be treated as signals for further review, not as final determinations.

Dual-Use and Misuse Risks. The benchmark has dual-use implications. While it can support the development and evaluation of more reliable detectors, it may also reveal detector weaknesses that malicious actors could use to produce more convincing, deceptive evidence. We mitigate this risk by framing the dataset as a research and evaluation resource, documenting intended uses and limitations in the dataset card, and avoiding operational guidance for generating successful fraudulent claims. The synthetic examples are released to enable controlled detector evaluation and failure analysis, not to facilitate deception, fraud, or harassment.

Over-Reliance and User Harm. A key downstream risk is over-reliance on imperfect detection models. False positives may cause genuinely damaged evidence to be incorrectly flagged as AI-generated, potentially disadvantaging legitimate customers in refund or dispute-resolution workflows. False negatives may allow synthetic or manipulated evidence to pass undetected. For this reason, the benchmark should not be interpreted as evidence that any evaluated model is ready for autonomous deployment. Any real-world use should include human review, uncertainty handling, appeal mechanisms, and platform-specific validation before model outputs affect user outcomes.

Bias, Fairness, and Representativeness. The benchmark does not collect protected attributes and is not intended to infer or predict sensitive characteristics such as gender, race, ethnicity, religion, sexuality, health status, or political affiliation. Nevertheless, performance may vary across platforms, languages, regions, product categories, service domains, imaging conditions, and user devices. Models trained or evaluated on this benchmark should therefore not be assumed to generalize uniformly to all review ecosystems or user populations. We avoid claims of universal representativeness and document the dataset scope, source platforms, category coverage, and language coverage to reduce the risk of inappropriate generalization.

Privacy, Surveillance, and Human Rights. The benchmark is based on review evidence rather than surveillance data, and it is not intended for identity recognition, user profiling, behavioral monitoring, or inference of protected attributes. However, related detection systems could be misused if deployed as broad monitoring tools over user-generated content or linked to punitive actions without transparency or recourse. We therefore recommend that downstream applications restrict use to content-integrity assessment, avoid identity-linked profiling, and provide meaningful human oversight for any decision that may affect users' access to services, refunds, or dispute outcomes.

Security and Environmental Considerations. The main security-related concern is adversarial adaptation, where attackers use benchmark feedback to improve synthetic evidence against current detectors. This risk is inherent to evaluation work on deceptive media, and we address it by emphasizing aggregate evaluation, failure-mode analysis, and responsible use rather than operational fraud guidance. Our experiments are limited to benchmark construction, specialized-detector evaluation on a single GPU instance, and API-based inference for proprietary MLLMs; the work does not involve training large foundation models from scratch. While API-based evaluation incurs additional upstream compute across proprietary model providers that is not directly observable from our local hardware logs, the associated energy footprint remains substantially different from that of large-scale foundation-model training, and no model parameters are updated in our study. Future work involving repeated large-scale MLLM evaluation or detector training should report compute resources, API usage, energy estimates, and evaluation scale transparently, particularly as provider-side energy reporting becomes more widely available.

I RESPONSIBLE RESEARCH PRACTICES

Privacy and Personally Identifiable Information. The real-review portion of the benchmark is derived from publicly accessible sources, including the Amazon Reviews 2023 dataset and publicly accessible Trip.com and GrabFood review records obtained through API-based requests and direct collection. Because user-generated review media can contain personally identifiable information, we apply automated MLLM-based filtering followed by human inspection to reduce privacy risk before release. We remove images identified as containing sensitive personal information, such as visible faces, license plates, addresses, screenshots, or other directly identifying content.

For released metadata, we retain only fields needed for benchmark use, traceability, and reproducibility. Display names, avatar links, and review URLs of Trip.com and GrabFood reviews are retained as publicly displayed by the source platforms, and we add no identifying information beyond them.

Consent and Public Data Use. The dataset is constructed from publicly accessible review data and public data sources. We did not obtain explicit consent from each review author because the records were publicly available at the time of collection, and contacting individual review authors at scale was not feasible. To mitigate this limitation, we restrict the released benchmark to research use, add no identifying metadata beyond what the source platforms display, filter images for personally sensitive content, and document the data sources, collection procedure, and intended use in the dataset card. We also document platform-specific access constraints in the dataset card and instruct users to comply with the applicable terms of the original platforms and data sources. We have additionally confirmed that all primary data sources used in this study, including Amazon Reviews 2023 (Hou et al., 2024), remain publicly available and have not been deprecated or withdrawn by their original authors at the time of submission.

Copyright and Usage Restrictions. The benchmark contains original review images and metadata together with synthetic images generated by our pipeline. We do not claim ownership over third-party review text, images, or platform metadata. Because the copyright and licensing status of review media is not uniform across source platforms, the dataset card documents data provenance, intended research use, and usage restrictions. Users of the benchmark are responsible for complying with the applicable terms of the original platforms and data sources. Synthetic images generated by our pipeline are released to support reproducibility and controlled evaluation.

Representative Evaluation Practice. The benchmark is designed to evaluate AI-generated or AI-modified evidence in refund- and complaint-related review scenarios, rather than to represent all online reviews or all forms of e-commerce fraud. The Amazon subset covers 27 product categories selected according to platform taxonomy and common product-review categories, while Trip.com and GrabFood provide two additional service-review domains covering hotel and food-delivery scenarios.

The review text is not restricted to English; when available, we retain the original language, including Simplified Chinese, Traditional Chinese, Japanese, Korean, Spanish, Russian, and other languages. At the same time, the benchmark should not be interpreted as universally representative of all platforms, regions, languages, cultures, product types, service domains, or fraud behaviors. Amazon and Trip.com serve broad international user populations, whereas GrabFood is primarily associated with Southeast Asian food-delivery markets. We therefore avoid claims of universal coverage and state in the limitations that the benchmark reflects the selected platforms, available review media, and the product and service categories included in our collection.