MANIVID: UNIFIED AND EXPLAINABLE MANIPULATION ANALYSIS OF VIDEOS VIA MLLMS

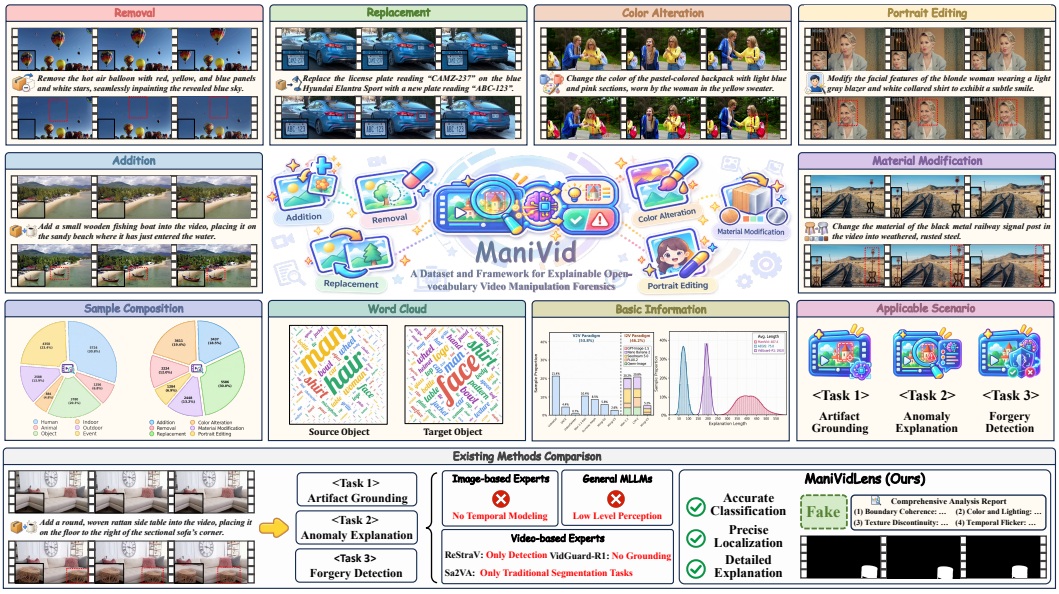


Figure 1: **Overview of ManiVid.** ManiVid unifies forgery detection, artifact grounding, and anomaly explanation. The **ManiVid-38K** dataset provides paired, open-vocabulary manipulations across six types with authenticity labels, pixel-level masks, and anomaly explanations. **ManiVid-Bench** provides 2K videos for evaluation, and **ManiVidLens** offers an end-to-end solution.

ABSTRACT

Rapid advances in AI-generated video (AIGV) have increased the risks posed by deceptive video manipulation. Unlike fully synthetic videos, manipulated videos retain most source content and alter only localized regions, making forensic analysis particularly challenging. Existing video forgery research faces two limitations in both data and methodology: (1) High-quality datasets and benchmarks tailored for manipulated videos remain scarce. (2) Multimodal large language models (MLLMs) extend forgery analysis beyond binary classification but struggle to use low-level forensic cues and provide precise pixel-level grounding. Specifically, we introduce **ManiVid**, a unified forensic analysis task covering forgery detection, artifact grounding, and anomaly explanation for manipulated videos. We construct **ManiVid-38K**, the first dataset to combine paired, open-vocabulary localized manipulations of general videos with authenticity labels, forgery masks, and anomaly explanations. It comprises about 19K manually verified real–fake video pairs, mostly at 1080P resolution, generated under 2 paradigms with 15 powerful generation models. We sample 1K pairs for **ManiVidBench**, balanced across six manipulation types and generation models for fair evaluation. We further propose **ManiVidLens**, a unified framework for explainable video forgery analysis. Its *Forensic Evidence Router* supplies shared low-level forensic evidence for multimodal reasoning and video segmentation. Its *Prompt Distill Module* converts grounding states into semantic and geometric prompts and distills spatial priors for mask decoding and full-video propagation. ManiVidLens achieves relative gains over the strongest comparison methods in artifact grounding (+21.1% mIoU; +21.3% J&F) and anomaly explanation (+131.3% ROUGE-L; +9.9% CSS). Its forgery detection remains comparable to dedicated classifiers (0.914 Acc; 0.913 F1). Together, our dataset and framework enable accurate and interpretable forensic analysis of manipulated videos. The code, model, and dataset will be released.

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.

1 INTRODUCTION

Recent advances in AI-generated video (AIGV) (Team et al., 2025; Wan et al., 2025; HaCohen et al., 2026) have enabled increasingly realistic, temporally coherent, and visually compelling videos through text-to-video (T2V) and image-to-video (I2V) generation. The widespread adoption of these technologies on social media has lowered the barrier to customized content creation but has also raised serious concerns, including fraud and misinformation.

To mitigate these risks, researchers have developed a range of video forgery detection methods. Early approaches apply image detectors independently to individual video frames and primarily rely on spatial forgery traces (Vahdati et al., 2024). More recent video detectors model spatial artifacts and temporal inconsistencies for authenticity prediction (Chen et al., 2026; Internò et al., 2025; Zhang et al., 2025a; Zheng et al., 2025). Emerging approaches employ multimodal large language models (MLLMs) to augment detection with anomaly explanations beyond binary authenticity judgments (Park et al., 2026). Despite this progress, two critical limitations remain:

(*i*) *Limited Attention and Data for Video Manipulation.* Unlike fully synthetic videos generated from text or images, manipulated videos preserve most source content and alter only localized regions, making their edits subtle and difficult to distinguish from authentic content; yet research predominantly targets fully synthetic videos (Bai et al., 2024a; Chen et al., 2026), while video manipulation studies remain limited to specific scenarios such as face swapping (Dolhansky et al., 2020). Building large-scale, high-quality, and diverse datasets and benchmarks for general video manipulation analysis presents three challenges. **(1) Quality:** edits must follow instructions and preserve unrelated content, requiring careful verification; **(2) Diversity:** generating diverse instructions that remain consistent with each source video is difficult at scale; **(3) Scalability:** video editing requires more computation and time than image editing. These challenges have hindered the development of large-scale, comprehensive benchmarks for general video manipulation analysis that jointly provide paired open-vocabulary manipulations, pixel-level masks, and anomaly explanations.

(*ii*) *Insufficient Interpretability in Existing Detection Methods.* Most video forgery detectors output only authenticity labels, without localizing manipulations or explaining supporting evidence. Recent systems use MLLMs for anomaly explanation (Park et al., 2026) or SAM (Kirillov et al., 2023) and SAM2 (Ravi et al., 2025) for artifact grounding (Kang et al., 2025; Huang et al., 2025a; Xu et al., 2025). Across these method families, two limitations remain. **(1) Objective Coupling:** forgery detection, anomaly explanation, and artifact grounding occupy distinct output spaces; forcing one response couples their objectives and can dilute evidence needed for segmentation. **(2) Evidence Misalignment:** detectors, MLLMs, and segmentation models specialize in authenticity discrimination, semantic reasoning, and mask decoding, so subtle manipulation cues are not shared across tasks. Existing interfaces further reduce grounding to a single latent state (Yuan et al., 2025) or decoded coordinates (Li et al., 2026), without complementary semantic and geometric prompts. These gaps motivate task-specific output protocols, shared forensic evidence, and a structured grounding interface for pixel-level video grounding.

To address these limitations, we propose **ManiVid** to unify forgery detection, artifact grounding, and anomaly explanation for manipulated videos. We construct **ManiVid-38K**, the first dataset to combine paired, open-vocabulary localized manipulations of general videos with authenticity labels, forgery masks, and detailed anomaly explanations. This high-quality dataset contains about **19K** real–fake video pairs, mostly at **1080P** resolution, generated by **15** unique models under **2** generation paradigms. It spans **6** common content categories and **6** manipulation tasks with comprehensive annotations. Representative manipulation cases and basic statistical analyses are shown in Figure 1 (Top). We sample **1K** pairs (**2K** videos) from ManiVid to construct **ManiVidBench** for evaluation. As illustrated in Figure 1 (Bottom), we propose **ManiVidLens**, an end-to-end solution to ManiVid that combines an MLLM with SAM2 (Ravi et al., 2025). The 3 tasks share model parameters but use task-specific protocols. Its *Forensic Evidence Router* (FER) builds a compact bank using the frozen AIDE encoder (Yan et al., 2025) and routes forensic cues to the language and mask streams through the *Forensic Evidence-to-Vision-Language Model* (FE2VLM) and *Forensic Evidence-to-SAM* (FE2SAM) adapters. Its *Prompt Distill Module* (PDM) maps grounding token states to dense semantic, sparse geometric, and segmentation prompts, while cosine distillation transfers SAM2's box-encoding prior to the 2 box-token states. At inference time, these prompts drive mask decoding and video propagation without human-provided prompts.

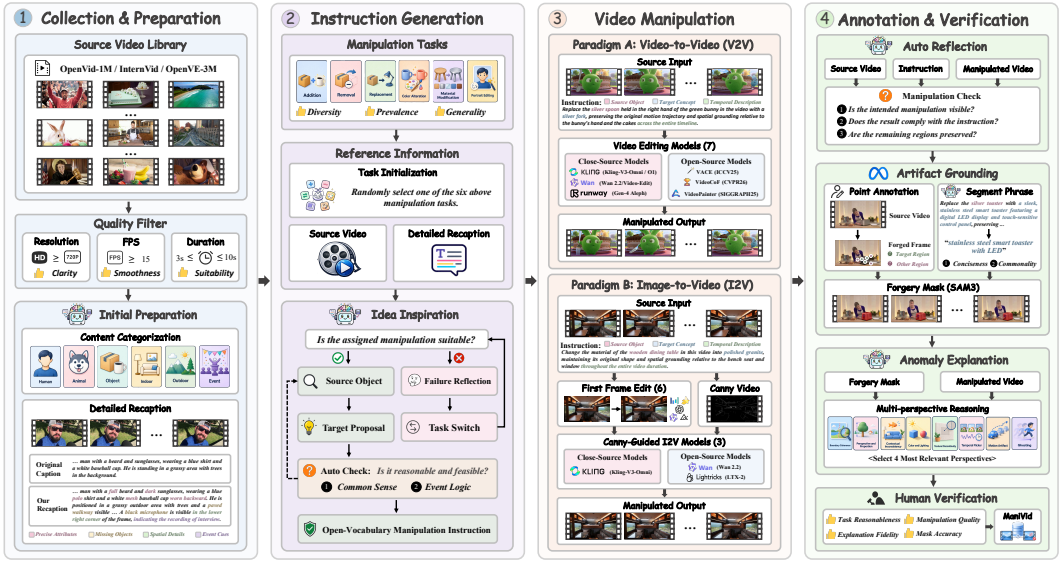


Figure 2: **Dataset Curation.** ManiVid collects high-quality real videos from three public datasets and uses the MLLM assistant to generate ***open-vocabulary***, ***reasonable***, and ***feasible*** manipulation instructions. Videos are then manipulated by diverse models, followed by the auto reflection and human verification. $(\cdot)$ denotes the number of generation models used.

To sum up, our main contributions are summarized as follows:

- We introduce **ManiVid**, its dataset **ManiVid-38K**, and benchmark **ManiVidBench**. The dataset is the first to combine paired, open-vocabulary localized manipulations of general videos with authenticity labels, pixel-level masks, and anomaly explanations.
- We propose **ManiVidLens**, an end-to-end framework for ManiVid combining an MLLM with SAM2 for forgery detection, artifact grounding, and anomaly explanation through task-specific protocols, shared forensic evidence, and structured semantic and geometric prompts.
- Extensive experiments on **ManiVidBench** show that **ManiVidLens** achieves state-of-the-art performance in artifact grounding and anomaly explanation while delivering forgery detection performance comparable to dedicated classifiers.

2 ManiVid-38K Dataset

2.1 Motivation and Highlights

Compared with fully synthetic videos, manipulated videos have received less attention due to the higher difficulty of data construction, annotation and verification, resulting in a scarcity of high-quality datasets. Existing manipulated datasets still suffer from several systematic limitations: **(1) Limited Diversity.** Previous works (Rossler et al., 2019; Dolhansky et al., 2020) focus mainly on face-centric manipulations such as face swapping, leaving diverse open-world visual content and manipulation scenarios underexplored. **(2) Closed-Vocabulary Instructions.** Some image manipulation datasets like SIDA (Huang et al., 2025a) construct instructions from predefined source and target object/attribute vocabularies. It not only restricts diversity but also often produces targets that violate common sense or event logic. **(3) Insufficient Annotations.** Most datasets provide only binary authenticity labels, lacking fine-grained supervision such as forgery masks and anomaly explanations, which limits their applicability for more explainable analysis. To support ManiVid, we construct **ManiVid-38K** with the following key highlights: **(1) Broad Coverage.** Our dataset covers six content categories (*e.g.*, human, object) and six manipulation task (*e.g.*, addition, removal), providing broad coverage of open-world visual scenarios and realistic manipulation operations. **(2) Open-Vocabulary Instructions.** Instead of predefined vocabularies, ManiVid-38K uses an MLLM assistant to generate open-vocabulary, context-aware instructions with both diversity and logical consistency. **(3) Comprehensive Annotations.** ManiVid-38K provides binary labels, pixel-level forgery masks, and anomaly explanations to facilitate further research on explainable video manipulation analysis. Appendix B provide more details about our motivation.

Table 1: **Comparison with Existing Video Forgery Datasets.** **V.P.** denotes the availability of semantic-aligned real-fake video pairs; **O.V.I.** denotes the release of open-vocabulary instructions; **H.V.** denotes human verification. Datasets that cannot guarantee source videos are real are **excluded**.

Dataset	Domain	Video Generation				Comprehensive Annotation					Scale
		Task	Model	Paradigm	V.P.	O.V.I	Label	Mask	Explanation	H.V.	
Forgery Type: Fully Synthetic (Directly applying T2V and I2V models)											
VidProM (Wang & Yang, 2024)	General	-	4	T2V	✗	✓	✓	✗	✗	✗	6.69M
GenVideo (Chen et al., 2026)	General	-	20	T2V, I2V	✗	✗	✓	✗	✗	✗	2.31M
AEGIS (Li et al., 2025)	General	-	10	T2V, I2V	✗	✗	✓	✗	✓	✓	10470
GenVideoBench (Ni et al., 2026)	General	-	11	T2V, I2V	✓	✓	✓	✗	✗	✗	6.78M
VidGuard-R1 (Park et al., 2026)	General	-	2	I2V	✓	✗	✓	✗	✓	✗	140K
Forgery Type: Partially Manipulated (Via editing models or manual work)											
FaceForensics++ (Rossler et al., 2019)	Facial	2	4	V2V	✓	✗	✓	✓	✗	✗	5K
DFDC (Dolhansky et al., 2020)	Facial	1	5	V2V	✓	✗	✓	✗	✗	✗	128K
VideoSham (Mittal et al., 2023)	General	6	0	Manual	✓	✓	✓	✗	✗	✓	826
VideoFACT (Nguyen et al., 2024)	General	4	3	V2V	✓	✗	✓	✓	✗	✗	12.38K
ManiVid (Ours)	**General**	**6**	**15**	**V2V, I2V**	✓	✓	✓	✓	✓	✓	**∼38K**

2.2 Curation Pipeline

As shown in Fig. 2, we introduce the curation pipeline of ManiVid-38K through four sequential steps: **(1) Collection and Preparation**; **(2) Instruction Generation**; **(3) Video Manipulation**; **(4) Annotation and Verification**. We employ Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025) as our MLLM assistant. Appendix C.1 provides more details about the curation pipeline.

Collection and Preparation. We collect source videos from three public datasets: InternVid (Wang et al., 2024), OpenVid (Nan et al., 2025), and OpenVE-3M (He et al., 2025). Specifically, InternVid and OpenVid were collected before recent high fidelity video generators (*e.g.*, Sora (OpenAI, 2024)), providing relatively reliable authentic video sources. For OpenVE-3M, we manually select source videos free of noticeable AI-generated artifacts. We then filter out samples with low resolution, frame rates, or excessively long durations to ensure source video quality. We further use the assistant to categorize and recaption videos, followed by balanced sampling across categories.

Instruction Generation. For each source video, we first randomly assign one of the six manipulation tasks. Given the video and its detailed recaption, the assistant determines whether the assigned task is suitable for the current scene. If suitable, it searches for a source object and proposes a semantically coherent target based on its open-world knowledge, while preserving common sense and event logic. Otherwise, it selects a more feasible task together with another source-target object pair. For example, if a video without any human subjects is assigned the *portrait editing* task, the assistant will then switch to a more feasible task and generate the content-aware instruction accordingly.

Video Manipulation. To reflect the rapidly evolving landscape of video manipulation, we generate manipulated videos through *two complementary paradigms*: **(1) Direct V2V.** We directly edit source videos using seven latest video editing models, including four closed-source models (Team et al., 2025; Kling AI, 2026; Runway, 2025; WaveSpeedAI, 2025) and three open-source models (Yang et al., 2026; Jiang et al., 2025; Bian et al., 2025). **(2) Indirect I2V.** We formulate the process as image editing followed by controllable I2V generation. Specifically, we first edit the initial frame according to the manipulation instruction using six image editing models, including three closed source models (Google, 2026; ByteDance Seed, 2026; OpenAI, 2025) and three open source models (Black Forest Labs, 2025; 2026; Wu et al., 2025). We then extract Canny guidance from the source video and generate the manipulated video conditioned on both the edited first frame and Canny video. We use three I2V models supporting Canny guidance: the closed source Kling-V3-Omni (Kling AI, 2026) and two open source models (Wan et al., 2025; HaCohen et al., 2026).

Annotation and Verification. We establish a rigorous annotation and verification pipeline to ensure the data quality. Specifically, we prompt the assistant to reflect on manipulation success, instruction adherence, and preservation of unedited regions, retaining only qualified samples. We then use the assistant to extract phrase level concepts from instructions as segmentation prompts. Since SAM3 (Carion et al., 2026) struggles with uncommon or ambiguous concepts, we crowdsource positive and negative points as spatial priors to improve segmentation accuracy. Given the masks and manipulated videos, the assistant selects four relevant perspectives from eight predefined forensic dimensions (*e.g.*, Boundary Coherence) to analyze relevant regions and generate anomaly explanations. Finally, all authors carefully verify every real-fake video pair along with its annotations against four criteria: *1) task reasonableness*, *2) manipulation quality*, *3) explanation fidelity*, and *4) mask accuracy*, with 24.8% of samples deemed unqualified and discarded.

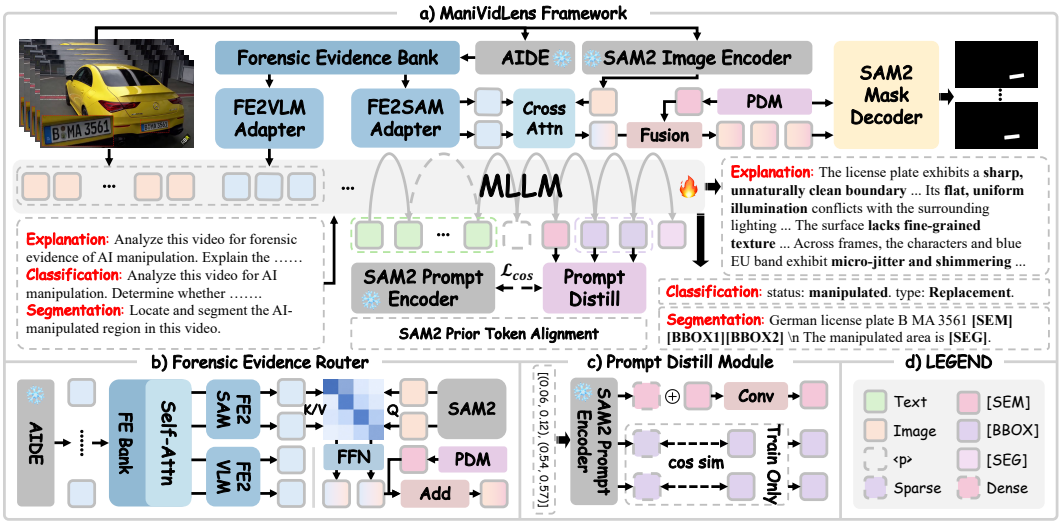


Figure 3: **Model Architecture.** **(a)** *ManiVidLens* unifies forgery detection, artifact grounding, and anomaly explanation. **(b)** The *Forensic Evidence Router* (FER) builds a compact forensic evidence bank and routes it to the MLLM and SAM2. **(c)** The *Prompt Distill Module* (PDM) converts MLLM grounding-token states into SAM2 prompts. Dashed arrows mark **training-only** paths.

2.3 Comparison with Exsisting Video Forgery Datasets

Tab. 1 highlights the limited diversity, scale, and annotation granularity of existing video forgery datasets, particularly those focusing on manipulation. ManiVid-38K is a diverse, open-vocabulary AI manipulation dataset for general videos. Unlike OpenVE-3M (He et al., 2025), it ensures authentic source videos and provides comprehensive annotations, including authenticity labels, forgery masks, anomaly explanations, and intermediate outputs, enabling research on explainable video forgery analysis. We further present statistical analysis and examples in Appendices C.2 and C.3.

3 Method

3.1 Overview

ManiVidLens addresses ManiVid through the end-to-end inference pipeline in Figure 3(a). Given a video $\mathcal{V} = \{I_t\}_{t=1}^{T}$ and a task-specific instruction q^r, the model generates a response $\mathbf{y}^r = (y_n^r)_{n=1}^{N_r}$, where $r \in \{\text{cls}, \text{seg}, \text{exp}\}$ denotes forgery detection, artifact grounding, or anomaly explanation, respectively. The model uses 5 temporally ordered conditioning frames $\mathcal{S} = \{I_{t_k}\}_{k=1}^{5}$, where $1 \le t_1 < \cdots < t_5 \le T$, and an ordered evidence set $\mathcal{A} = \{(\mathbf{R}_k, \mathbf{s}_k)\}_{k=1}^{5}$ extracted by the frozen AIDE encoder (Yan et al., 2025). The 3 tasks share $\mathcal{S}$, $\mathcal{A}$, and model parameters but use task-specific instructions and response targets because they operate in different output spaces.

For $r = \text{cls}$, $\mathbf{y}^{\text{cls}}$ reports whether the video is manipulated; for $r = \text{seg}$, $\mathbf{y}^{\text{seg}}$ contains a short target description followed by [SEM], [BBOX1], [BBOX2], and [SEG]; and for $r = \text{exp}$, $\mathbf{y}^{\text{exp}}$ describes the observed anomalies. For a manipulated video, the grounding-token states form SAM2 prompts for mask decoding on the conditioning frames and propagation across all T frames, yielding $\widehat{\mathcal{M}} = \{\widehat{M}_t\}_{t=1}^{T}$. For an original video, the grounding response states that no manipulated region exists and omits the special-token span. Generating [SEG] activates mask decoding; otherwise, the model returns only the textual response. Appendices D.1 and D.4.2 specify the output protocols.

3.2 Forensic Evidence Router

The Forensic Evidence Router (FER) transforms frozen frame-level forensic cues into a shared evidence bank for the MLLM and SAM2, as illustrated in Figure 3(b). The *Forensic Evidence-to-Vision-Language Model* (FE2VLM) adapter injects compact evidence tokens into every task input, while the *Forensic Evidence-to-SAM* (FE2SAM) adapter enhances SAM2 image features for artifact grounding. For each conditioning frame, the frozen AIDE encoder returns four low-level forensic

tokens $\mathbf{R}_k = \{\mathbf{r}_{k,j}\}_{j=1}^{4}$ and one semantic token $\mathbf{s}_k$; $\bar{\mathbf{r}}_k$ denotes the mean of the four forensic tokens. The projected semantic, forensic, and global cues form the evidence bank refined through attention:

$$\mathbf{E}_k = \mathrm{Attention}([\phi_s(\mathbf{s}_k), \Phi_r(\mathbf{R}_k), \phi_g([\mathbf{s}_k; \bar{\mathbf{r}}_k])] + \mathbf{T}_{\mathrm{bank}}) \in \mathbb{R}^{6\times d_e}. \tag{1}$$

Here, ϕ_s, Φ_r, and ϕ_g project semantic, forensic, and global cues into the shared d_e-dimensional space; $\mathbf{T}_{\mathrm{bank}}$ distinguishes the semantic, four forensic, and global evidence slots during routing.

The FE2VLM adapter selects the global token, semantic token, and mean of the four forensic tokens from each evidence bank, and projects the resulting three tokens into the MLLM space. These tokens replace the three [AIDE] positions after the corresponding frame's visual block, supplementing the MLLM's visual representations with low-level forensic evidence. For artifact grounding, the FE2SAM adapter selects the global token and four forensic tokens and projects them into the SAM2 space. Through cross-attention, these tokens inject manipulation-sensitive evidence into the SAM2 image features of each conditioning frame. A shared authenticity gate downweights routed forensic evidence for videos predicted as original, shifting FE2VLM toward the learned [AIDE] embedding and FE2SAM toward the original SAM2 image features. Further implementation details of both adapters and evidence routing are provided in Appendix D.2.

3.3 Prompt Distill Module

As shown in Figure 3(c), the Prompt Distill Module (PDM) projects the special token states from the grounding response into SAM2 prompts. The [SEM] state is projected and fused with SAM2's dense no-mask embedding for semantic context; the [BBOX1] and [BBOX2] states form sparse geometric prompts; and the [SEG] state supplies a segmentation prompt. These prompts guide mask decoding on conditioning frames before SAM2 propagates the object state throughout the video. During training, PDM reads the final-layer hidden states at the supervised special token positions in the assistant target sequence. The frozen SAM2 prompt encoder maps bounding boxes from the conditioning-frame masks into pairs of corner embeddings. For each sample, valid top-left and bottom-right embeddings are averaged separately across conditioning frames to form $\bar{\mathbf{b}}_{i,1}$ and $\bar{\mathbf{b}}_{i,2}$, which supervise the projected MLLM states $\mathbf{z}_i^{b_1}$ and $\mathbf{z}_i^{b_2}$ through cosine distillation:

$$\mathcal{L}_{\cos} = \frac{1}{2|\Omega|} \sum_{i\in\Omega} \sum_{j=1}^{2} \left(1 - \frac{\langle \mathbf{z}_i^{b_j}, \bar{\mathbf{b}}_{i,j} \rangle}{\|\mathbf{z}_i^{b_j}\|_2 \|\bar{\mathbf{b}}_{i,j}\|_2}\right). \tag{2}$$

The pooled targets distill a video-level geometric prior into the 2 box-token states, while SAM2 resolves frame-specific locations through its image features and temporal memory. Ω indexes samples with a complete special token span and at least one valid box annotation. The cosine loss applies only to the 2 box-token states; language and mask losses provide the remaining sequence-level and mask-level supervision. At inference, PDM reads the autoregressively generated states and uses them as SAM2 prompts without a supplied bounding box prompt. See Appendix D.3 for details.

3.4 Two-Stage Training Strategy

We optimize ManiVidLens in two stages with distinct objectives:

$$\begin{aligned} \mathcal{L}^{(1)} &= \lambda_{\mathrm{lm}}^{(1)} \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{cls}}^{(1)} \mathcal{L}_{\mathrm{cls}}, \\ \mathcal{L}^{(2)} &= \lambda_{\mathrm{lm}}^{(2)} \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{cls}}^{(2)} \mathcal{L}_{\mathrm{cls}} + \lambda_m \mathcal{L}_{\mathrm{mask}} + \lambda_d \mathcal{L}_{\mathrm{dice}} + \lambda_c \mathcal{L}_{\cos}. \end{aligned} \tag{3}$$

$\mathcal{L}_{\mathrm{LM}}$ supervises the task-specific responses $\mathbf{y}^r$, while $\mathcal{L}_{\mathrm{cls}}$ trains the auxiliary video-authenticity estimator. The mask cross-entropy $\mathcal{L}_{\mathrm{mask}}$ and Dice loss $\mathcal{L}_{\mathrm{dice}}$ supervise pixel masks, while $\mathcal{L}_{\cos}$ distills the two box-token states. Stage 1 learns the three task-specific output protocols while keeping all mask-grounding modules frozen. Stage 2 starts from the Stage 1 checkpoint and emphasizes pixel-level grounding while retaining lower weights for the language and classification objectives. The three grounding losses are evaluated only for manipulated segmentation instances with mask annotations. Appendix D.4.1 reports the exact loss weights, sampling ratios, and frozen modules.

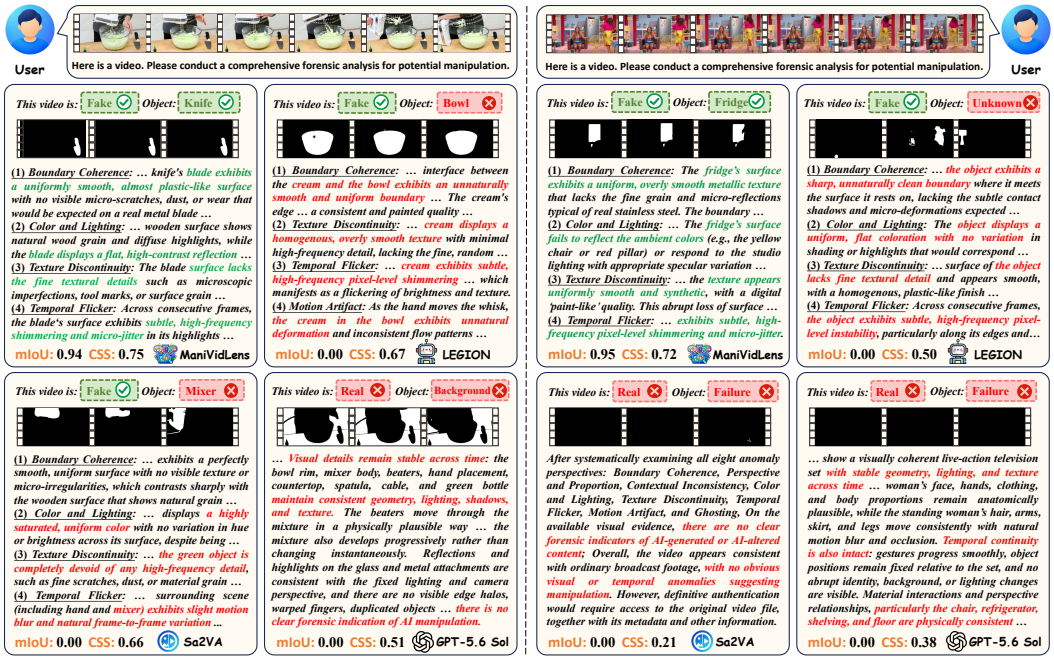


Figure 4: **Qualitative Comparison of ManiVidLens with Different Models.** We visualize two representative manipulation cases: (Left) adding a knife and (Right) changing the refrigerator material to steel. Across both cases, ManiVidLens demonstrates consistently better forensic analysis than LEGION (Image-based), Sa2VA (Video-based), and GPT-5.6 Sol (General MLLM).

4 Experiments

4.1 Experimental Setup

Implementation Details. ManiVidLens uses Qwen3-VL-4B-Instruct (Bai et al., 2025) and SAM2 Hiera-L (Ravi et al., 2025) as its MLLM and video segmentation backbones. We train it on the ManiVid-38K training split for one epoch per stage with task-specific prompts and response formats. All image and video experts are trained on this split, while general MLLMs are not fine-tuned. All models are evaluated on ManiVidBench's 2K videos across six manipulation categories.

Evaluation Metrics. We evaluate *artifact grounding* with mean IoU (mIoU) and J&F (Yuan et al., 2025), *anomaly explanation* with ROUGE-L (Lin, 2004) and Cosine Similarity Score (CSS) using MiniLM embeddings (Reimers & Gurevych, 2019; Xu et al., 2025), and *forgery detection* with accuracy and F1 at the video level. For grounding, image experts predict masks on 30 uniformly sampled frames as an approximation of the complete video, whereas general MLLMs generate only boxes from five conditioning frames; these boxes prompt shared SAM2 mask propagation across every frame. Further implementation and evaluation details are provided in Appendix E.1. **`Bold`** and `underlined` values indicate the best and second-best results in each table column, respectively.

4.2 Main Results

Artifact Grounding. ManiVidLens reaches 0.407 mIoU and 0.422 J&F (Table 2), surpassing Sa2VA, the strongest overall comparison, by **0.071 mIoU** and **0.074 J&F**. The strongest image expert, FakeShield, and general MLLM, Qwen3.5, reach 0.318 and 0.306 mIoU, respectively, showing our advantage over both direct mask prediction and SAM2 propagation from predicted boxes. It leads both metrics in five of six categories, with its largest mIoU margins on Material (**+0.127**) and Portrait (**+0.090**). Color and Material preserve object identity while changing its appearance; their 0.487 and 0.526 mIoU scores show that the advantage extends to edits for which object recognition alone cannot identify manipulation. Removal is the sole exception (0.135 mIoU vs. Qwen3.5's 0.221) despite 0.943 detection accuracy: the reference mask covers the removed object's extent, but the input region contains reconstructed background without object semantics, exposing a gap between detection and precise spatial recovery of edited regions.

Table 2: **Artifact Grounding Performance. Synth.** and **Mani.** denote fully synthetic and manipulated content, respectively. ☆ marks frame-wise approximations on 30 uniformly sampled frames; □ marks native box outputs evaluated with shared SAM2 propagation over every video frame.

Method	Focus	ManiVidBench												Avg.	
		Addition		Removal		Replacement		Color		Material		Portrait			
		mIoU	J&F	mIoU	J&F	mIoU	J&F	mIoU	J&F	mIoU	J&F	mIoU	J&F	mIoU	J&F
Image Expert Models (3)															
LEGION☆ (Kang et al., 2025)	Synth.	0.343	0.352	0.091	0.102	0.385	0.392	0.323	0.346	0.360	0.370	0.192	0.213	0.312	0.324
FakeShield☆ (Xu et al., 2025)	Mani.	0.365	0.374	0.068	0.073	0.332	0.333	0.429	0.430	0.373	0.371	0.161	0.153	0.318	0.320
SIDA☆ (Huang et al., 2025a)	Both	0.013	0.113	0.004	0.020	0.010	0.019	0.020	0.042	0.009	0.025	0.002	0.004	0.011	0.041
Video Expert Models (4)															
VideoLISA (Bai et al., 2024b)	General	0.191	0.202	0.064	0.077	0.166	0.183	0.111	0.128	0.161	0.174	0.029	0.050	0.138	0.153
VISA (Yan et al., 2024)	General	0.157	0.173	0.054	0.060	0.076	0.086	0.040	0.052	0.080	0.090	0.018	0.022	0.079	0.089
Sa2VA (Yuan et al., 2025)	General	0.414	0.432	0.132	0.135	0.343	0.356	0.390	0.406	0.399	0.406	0.182	0.186	0.336	0.348
Omni-Fake-R1 (Li et al., 2026)	Both	0.287	0.306	0.086	0.097	0.224	0.251	0.238	0.263	0.335	0.350	0.072	0.099	0.226	0.248
General MLLMs (3)															
InternVL3.5-30B-A3B□ (Wang et al., 2025)	General	0.109	0.118	0.061	0.063	0.042	0.048	0.020	0.029	0.041	0.051	0.039	0.048	0.053	0.060
Qwen3.5-35B-A3B□ (Qwen Team, 2026)	General	0.381	0.395	**0.221**	**0.222**	0.301	0.319	0.294	0.310	0.355	0.370	0.200	0.209	0.306	0.319
GPT-5.6-Sol□ (OpenAI, 2026)	General	0.270	0.274	0.192	0.195	0.158	0.166	0.238	0.246	0.210	0.218	0.031	0.033	0.198	0.205
ManiVidLens (Ours)	**Mani.**	**0.444**	**0.467**	0.135	0.137	**0.421**	**0.439**	**0.487**	**0.507**	**0.526**	**0.532**	**0.290**	**0.306**	**0.407**	**0.422**

Table 3: **Anomaly Explanation Performance. Synth.** and **Mani.** denote fully synthetic and manipulated content. **R-L** denotes *ROUGE-L*, and **CSS** denotes *Cosine Similarity Score*.

Method	Focus	ManiVidBench												Avg.	
		Addition		Removal		Replacement		Color		Material		Portrait			
		R-L	CSS	R-L	CSS	R-L	CSS	R-L	CSS	R-L	CSS	R-L	CSS	R-L	CSS
Image Expert Models (3)															
LEGION (Kang et al., 2025)	Synth.	0.255	0.693	0.239	0.681	0.258	0.727	0.251	0.709	0.246	0.721	0.247	0.718	0.251	0.710
FakeShield (Xu et al., 2025)	Mani.	0.182	0.559	0.128	0.483	0.165	0.541	0.174	0.579	0.162	0.558	0.116	0.455	0.162	0.541
SIDA (Huang et al., 2025a)	Both	0.249	0.674	0.251	0.694	0.255	0.681	0.252	0.699	0.252	0.700	0.251	**0.746**	0.252	0.691
Video Expert Models (3)															
VidGuard-R1 (Park et al., 2026)	Synth.	0.200	0.541	0.194	0.569	0.202	0.587	0.206	0.621	0.200	0.608	0.200	0.597	0.201	0.586
Sa2VA (Yuan et al., 2025)	General	0.206	0.619	0.161	0.538	0.171	0.556	0.176	0.565	0.184	0.589	0.081	0.382	0.174	0.561
Omni-Fake-R1 (Li et al., 2026)	Both	0.239	0.703	0.224	0.671	0.233	0.688	0.226	0.689	0.226	0.693	0.212	0.668	0.230	0.688
General MLLMs (3)															
InternVL3.5-30B-A3B (Wang et al., 2025)	General	0.090	0.431	0.088	0.432	0.090	0.435	0.087	0.425	0.088	0.448	0.085	0.412	0.089	0.432
Qwen3.5-35B-A3B (Qwen Team, 2026)	General	0.096	0.444	0.094	0.452	0.096	0.453	0.095	0.443	0.095	0.466	0.096	0.445	0.095	0.450
GPT-5.6-Sol (OpenAI, 2026)	General	0.082	0.436	0.082	0.450	0.082	0.446	0.081	0.439	0.081	0.454	0.080	0.446	0.082	0.444
ManiVidLens (Ours)	**Mani.**	**0.603**	**0.813**	**0.576**	**0.773**	**0.581**	**0.777**	**0.582**	**0.772**	**0.587**	**0.795**	**0.545**	0.699	**0.583**	**0.780**

Anomaly Explanation. ManiVidLens reaches 0.583 ROUGE-L and 0.780 CSS (Table 3), exceeding the strongest comparison for each metric by **0.331** and **0.070**, respectively. It leads all six categories in ROUGE-L and five in CSS; Portrait is the sole CSS exception (0.699 vs. SIDA's 0.746). We further complement these automatic metrics with human evaluation in Appendix E.3.4, examining their potential sensitivity to lexical overlap and structured response formats.

Forgery Detection. ManiVidLens achieves 0.914 accuracy and 0.913 F1 (Table 4), approaching the dedicated video detector DeMamba with gaps of only 0.008 in accuracy and 0.006 in F1 while sharing one model across all three tasks. It leads both metrics on Material and Portrait. Parallel gains in grounding suggest shared cues help both localize edits and judge if videos are authentic. We discuss the experimental results of the more challenging manipulation attribution task in Appendix E.3.3.

Qualitative Analysis. Figure 4 shows that ManiVidLens aligns object identification, masks across successive video frames, and forensic explanations, attributing boundary and texture anomalies to the localized objects. Image expert methods, represented by LEGION, confuse the knife with the bowl and produce varying refrigerator masks across frames, consistent with their lack of temporal modeling. Video expert methods, represented by Sa2VA, select the mixer in the first case and miss the material edit in the second, illustrating the need for low-level forensic cues alongside temporal context. General MLLMs, represented by GPT-5.6 Sol, classify both manipulations as real yet return full-frame boxes under grounding prompts, allowing SAM2 masks to overlap targets incidentally. Appendix E.3 complements these observations with quantitative analyses of the three tasks by manipulation type and of ManiVidLens performance across different generation models.

4.3 Ablation Study

Sampled Frames. Five frames capture most of the grounding gains from three to seven frames (Table 5). Moving from three to five raises mIoU **from 0.354 to 0.407** and J&F from 0.368 to 0.422, with inference time increasing from 13.83 s to 15.98 s. The improvement covers all six cat-

Table 4: **Forgery Detection Performance. Synth.** and **Mani.** denote fully synthetic and manipulated content. **Acc** and **F1** denote video-level accuracy and F1; **Avg.** covers all 2,000 videos. Image expert models use majority voting over 30 uniformly sampled frames at a 0.5 threshold.

Method	Focus	ManiVidBench												Avg.	
		Addition		Removal		Replacement		Color		Material		Portrait			
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Image Expert Models (3)															
CNNSpot (Wang et al., 2020)	Synth.	0.568	0.665	0.607	0.687	0.608	0.696	0.615	0.712	0.615	0.699	0.593	0.700	0.602	0.693
UnivFD (Ojha et al., 2023)	Synth.	0.599	0.598	0.592	0.619	0.603	0.614	0.654	0.689	0.649	0.662	0.653	0.687	0.620	0.638
AIDE (Yan et al., 2025)	Synth.	0.866	0.865	0.882	0.881	0.818	0.819	0.849	0.850	0.851	0.856	0.746	0.741	0.842	0.842
Video Expert Models (5)															
D3 (Zheng et al., 2025)	Synth.	0.500	0.667	0.500	0.667	0.500	0.667	0.503	0.668	0.500	0.667	0.500	0.667	0.501	0.667
VidGuard-R1 (Park et al., 2026)	Synth.	0.521	0.627	0.508	0.610	0.525	0.631	0.573	0.663	0.519	0.613	0.517	0.617	0.530	0.631
ReStraV (Internò et al., 2025)	Synth.	0.542	0.411	0.569	0.469	0.564	0.416	0.570	0.441	0.588	0.509	0.517	0.374	0.562	0.438
NSG-VD (Zhang et al., 2025a)	Synth.	0.521	0.646	0.496	0.629	0.503	0.638	0.492	0.626	0.515	0.648	0.449	0.606	0.502	0.635
DeMamba (Chen et al., 2026)	Synth.	**0.974**	**0.973**	**0.943**	0.940	**0.890**	**0.885**	**0.924**	**0.922**	0.935	0.932	0.831	0.818	**0.922**	**0.919**
General MLLMs (3)															
InternVL3.5-30B-A3B (Wang et al., 2025)	General	0.503	0.348	0.508	0.356	0.508	0.355	0.503	0.339	0.508	0.356	0.508	0.352	0.506	0.351
Qwen3.5-35B-A3B (Qwen Team, 2026)	General	0.579	0.524	0.599	0.559	0.559	0.488	0.549	0.450	0.576	0.512	0.559	0.488	0.569	0.502
GPT-5.6-Sol (OpenAI, 2026)	General	0.699	0.697	0.718	0.716	0.666	0.660	0.648	0.639	0.687	0.679	0.593	0.590	0.674	0.669
ManiVidLens (Ours)	**Mani.**	0.966	0.966	**0.943**	**0.943**	0.868	0.868	0.917	0.917	**0.939**	**0.939**	**0.839**	**0.839**	0.914	0.913

Table 5: **Ablation on Sampled Frames.** 5 frames are used by default. Inference time is averaged per video.

Setting	Grounding		Explanation		Detection		Inference Time ↓
	mIoU	J&F	R-L	CSS	Acc	F1	
3 Frames	0.354	0.368	0.575	0.763	0.906	0.905	**13.83s**
5 Frames	0.407	0.422	0.583	0.780	0.914	0.913	15.98s
7 Frames	**0.413**	**0.431**	**0.596**	**0.803**	0.926	**0.926**	17.72s
9 Frames	0.397	0.413	0.592	0.797	**0.927**	**0.926**	19.20s

Table 6: **Ablation on PDM and FER.** Base omits both modules; + PDM and + FER add one module each; + Both uses both.

Setting	Grounding		Explanation		Detection	
	mIoU	J&F	R-L	CSS	Acc	F1
Base	0.336	0.348	0.590	0.791	**0.915**	**0.914**
+ PDM	0.368	0.380	0.585	0.785	0.905	0.904
+ FER	0.378	0.389	**0.598**	**0.808**	0.907	0.906
+ Both (Ours)	**0.407**	**0.422**	0.583	0.780	0.914	0.913

egories (Details are shown in Appendix E.4), with the largest mIoU gains on Color (+0.092) and Material (+0.076). Seven frames add 0.006 mIoU, 0.023 CSS, and 0.012 Accuracy over five frames at 10.9% higher inference time. This small average grounding gain masks opposing category trends: Addition and Portrait improve, while Color and Material decline by 0.035 and 0.043 mIoU. At nine frames, mIoU falls to 0.397 while inference time reaches 19.20 s and detection accuracy peaks at 0.927. These different task optima motivate balancing localization quality, detection performance, and inference cost. We therefore use five frames as the default input for all three tasks.

Architectural Components. PDM and FER provide complementary grounding gains (Table 6). PDM alone improves over Base by 0.032 mIoU and 0.032 J&F, and FER alone by 0.042 mIoU and 0.041 J&F. Combining them yields the highest grounding scores, improving over Base by **0.071 mIoU and 0.074 J&F**. The full model improves mIoU over Base in every category. Appendix E.4 reports the complete category breakdown across all three tasks. Adding PDM to FER contributes a further mIoU gain of 0.065 on Color, 0.077 on Material, and 0.088 on Portrait. PDM alone reduces Portrait mIoU from 0.182 to 0.170, yet combining it with FER reaches 0.290, demonstrating their complementary grounding benefits. FER alone gives the highest explanation scores (0.598 ROUGE-L and 0.808 CSS). The full model retains 0.914 Accuracy and 0.913 F1, both within 0.001 of Base, indicating that its main benefit is more precise spatial recovery.

5 Conclusion

We introduce **ManiVid** for unified and explainable forensic analysis of manipulated videos, supported by **ManiVid-38K** and **ManiVidBench**. **ManiVid-38K** is the first dataset combining paired, open-vocabulary localized video manipulations with authenticity labels, pixel-level masks, and anomaly explanations. **ManiVidLens** is an end-to-end framework combining an MLLM and SAM2 through task-specific protocols, shared forensic evidence, and structured semantic and geometric prompts for full-video grounding without human-provided localization prompts. On **ManiVidBench**, it achieves state-of-the-art artifact grounding and anomaly explanation, with forgery detection comparable to dedicated classifiers. The results highlight complementary gains from forensic evidence and structured prompts, while object removal remains challenging for precise localization. Together, these resources support developing and evaluating more interpretable video forensics.

AI USE STATEMENT

We used AI tools for video manipulation in the ManiVid-38K dataset, code debugging assistance, grammar check, and linguistic polishing. Specifically, AI tools were used to generate manipulated videos during dataset construction, assist in identifying and resolving coding issues, and improve the writing quality of the manuscript. All AI-assisted outputs were carefully reviewed and verified by the authors. The authors take full responsibility for the final dataset, code, and manuscript content.

REFERENCES

Jianfa Bai, Man Lin, Gang Cao, and Zijie Lou. Ai-generated video detection via spatial-temporal anomaly learning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 460–470. Springer, 2024a.

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.

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Z Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2024b.

Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. Videopainter: Any-length video inpainting and editing with plug-and-play context control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025.

Black Forest Labs. FLUX.2 [dev] Model Card. `https://huggingface.co/black-forest-labs/FLUX.2-dev`, 2025. Accessed: 2026-08-03.

Black Forest Labs. FLUX.2 [klein] 9B Model Card. `https://huggingface.co/black-forest-labs/FLUX.2-klein-9B`, 2026. Accessed: 2026-08-03.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

ByteDance Seed. Seedream 5.0 Lite. `https://seed.bytedance.com/en/seedream5_0_lite`, 2026. Accessed: 2026-08-03.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International Conference on Learning Representations*, volume 2026, pp. 138846–138923, 2026.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4724–4733. IEEE, 2017.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *Science China Information Sciences*, 69(6):162103, 2026.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Google. Nano Banana 2: Gemini AI Image Generator and Photo Editor. https://gemini.google/overview/image-generation/, 2026. Accessed: 2026-08-03.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Haoyang He, Jie Wang, Jiangning Zhang, Zhucun Xue, Xingyuan Bu, Qiangpeng Yang, Shilei Wen, and Lei Xie. Openve-3m: A large-scale high-quality dataset for instruction-guided video editing. *arXiv preprint arXiv:2512.07826*, 2025.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28831–28841, 2025a.

Zhenglin Huang, Tianxiao Li, Xiangtai Li, Haiquan Wen, Yiwei He, Jiangning Zhang, Hao Fei, Xi Yang, Xiaowei Huang, Bei Peng, et al. So-fake: Benchmarking and explaining social media image forgery detection. *arXiv preprint arXiv:2505.18660*, 2025b.

Christian Internò, Robert Geirhos, Markus Olhofer, Sunny Liu, Barbara Hammer, and David Klindt. AI-generated video detection via perceptual straightening. *Advances in Neural Information Processing Systems*, 38:20672–20705, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202. IEEE, 2025.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18937–18947. IEEE, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Kling AI. Kling VIDEO 3.0 Omni Model User Guide. https://kling.ai/quickstart/klingai-video-3-omni-model-user-guide, February 2026. Accessed: 2026-08-03.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Jieyu Li, Xin Zhang, and Joey Tianyi Zhou. Aegis: Authenticity evaluation benchmark for ai-generated video sequences. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13346–13353, 2025.

Tianxiao Li, Zhenglin Huang, Haiquan Wen, Yiwei He, Xinze Li, Bingyu Zhu, Wuhui Duan, Congang Chen, Zeyu Fu, Yi Dong, et al. Omni-fake: Benchmarking unified multimodal social media deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30299–30311, 2026.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Kaiqing Lin, Zhiyuan Yan, Ruoxin Chen, Junyan Ye, Ke-Yue Zhang, Yue Zhou, Peng Jin, Bin Li, Taiping Yao, and Shouhong Ding. Seeing before reasoning: A unified framework for generalizable and explainable fake image detection. *arXiv preprint arXiv:2509.25502*, 2025.

Trisha Mittal, Ritwik Sinha, Viswanathan Swaminathan, John Collomosse, and Dinesh Manocha. Video manipulations beyond faces: A dataset with human-machine analysis. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 643–652, 2023.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *International Conference on Learning Representations*, volume 2025, pp. 1045–1064, 2025.

Tai D Nguyen, Shengbang Fang, and Matthew C Stamm. Videofact: detecting video forgeries using attention, scene context, and forensic traces. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8548–8558. IEEE, 2024.

Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. GenVidBench: A 6-million benchmark for AI-generated video detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(18):15582–15590, 2026. doi: 10.1609/aaai.v40i18.38587.

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

OpenAI. Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators, 2024. Accessed: 2026-07-31.

OpenAI. Introducing gpt-image-1.5. https://openai.com/index/new-chatgpt-images-is-here/, December 2025. Accessed: 2026-08-24.

OpenAI. GPT-5.6. https://deploymentsafety.openai.com/gpt-5-6, July 2026. Accessed: 2026-09-08.

Kyoungjun Park, Yifan Yang, Juheon Yi, Shicheng Zheng, Yifei Shen, Dongqi Han, Caihua Shan, Muhammad Muaz, and Lili Qiu. VidGuard-R1: AI-generated video detection and explanation via reasoning MLLMs and RL. *International Conference on Learning Representations*, 2026.

PKU-Yuan Lab and Tuzhan AI et al. Open-sora-plan. GitHub and Zenodo, https://doi.org/10.5281/zenodo.10948109, April 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Nils Reimers and Iryna Gurevych. paraphrase-minilm-l6-v2. Hugging Face Model Hub, 2020. URL https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2. Sentence-Transformers model.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018.

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.

Runway. Runway Aleph: A New Way to Edit, Transform and Generate Video. https://runway.com/research/introducing-runway-aleph, 2025. Accessed: 2026-08-03.

Kling Team, Jialu Chen, Yuanzheng Ci, Xiangyu Du, Zipeng Feng, Kun Gai, Sainan Guo, Feng Han, Jingbin He, Kang He, et al. Kling-omni technical report. *arXiv preprint arXiv:2512.16776*, 2025.

Danial Samadi Vahdati, Tai D Nguyen, Aref Azizpour, and Matthew C Stamm. Beyond deepfake images: Detecting ai-generated videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4397–4408. IEEE, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8692–8701. IEEE, 2020.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Advances in Neural Information Processing Systems*, 37:65618–65642, 2024.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2024, pp. 42055–42079, 2024.

WaveSpeedAI. Wan 2.2 Video Edit. https://wavespeed.ai/models/wavespeed-ai/wan-2.2/video-edit, 2025. Accessed: 2026-08-03.

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. *Advances in Neural Information Processing Systems*, 38: 58972–59005, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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. *The Thirteenth International Conference on Learning Representations*, 2025.

Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 2025.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pp. 98–115. Springer, 2024.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. *Advances in Neural Information Processing Systems*, 37:57240–57261, 2024.

Xiangpeng Yang, Ji Xie, Yiyuan Yang, Yue Ma, Yan Huang, Min Xu, and Qiang Wu. Videocof: Unified video editing with temporal reasoner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37940–37949, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, et al. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *Advances in Neural Information Processing Systems*, 37:21236–21270, 2024.

Shuhai Zhang, ZiHao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. Physics-driven spatiotemporal modeling for AI-generated video detection. *Advances in Neural Information Processing Systems*, 38:174683–174730, 2025a.

Sicheng Zhang, Binzhu Xie, Zhonghao Yan, Yuli Zhang, Donghao Zhou, Xiaofei Chen, Shi Qiu, Jiaqi Liu, Guoyang Xie, and Zhichao Lu. Trade-offs in image generation: How do different dimensions interact? In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 01–12. IEEE, 2025b.

Sicheng Zhang, Zhonghao Yan, Binzhu Xie, Shi Qiu, Muzammal Naseer, Naveed Akhtar, and Mubarak Shah. On the limitations of cross-lingual consistency in multilingual text-to-image generation. *arXiv preprint arXiv:2608.11002*, 2026.

Chende Zheng, Ruiqi Suo, Chenhao Lin, Zhengyu Zhao, Le Yang, Shuai Liu, Minghui Yang, Cong Wang, and Chao Shen. D3: Training-free ai-generated video detection using second-order features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12852–12862, 2025.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pp. 650–667. Springer, 2022.

Leqi Zhu, Junyan Ye, Kaiqing Lin, Zhiyuan Yan, Conghui He, and Weijia Li. Fakevlm-r1: Internalizing physical laws via cot for synthetic image detection. *arXiv preprint arXiv:2605.30062*, 2026.

APPENDIX

A Related Work **16**

A.1 Video Forgery Detection Datasets 16

A.2 Video Forgery Detection Methods 16

B Motivation **17**

B.1 Scope Comparison: Full Synthesis vs. Manipulation 17

B.2 Manipulation Dataset Limitation 17

C ManiVid-38K Dataset **18**

C.1 Curation Pipeline 18

C.1.1 Source Video Collection 18

C.1.2 Content Categorization 19

C.1.3 Detailed Recaption 21

C.1.4 Instruction Generation 21

C.1.5 Manipulation Setup 23

C.1.6 Generation Model Details 24

C.1.7 Paradigm A: Direct V2V Cases 25

C.1.8 Paradigm B: Indirect I2V Cases 27

C.1.9 Automatic Reflection 27

C.1.10 Artifact Grounding 28

C.1.11 Anomaly Explanation 28

C.1.12 Human Verification 30

C.2 Statistical Analysis 30

C.3 Case Visualization 30

C.4 ManiVidBench 35

D ManiVidLens Framework **35**

D.1 Task-Specific Input and Output Protocol 35

D.2 Forensic Evidence Router 36

D.2.1 Forensic Evidence Bank 36

D.2.2 Auxiliary Authenticity Estimator 37

D.2.3 Evidence Injection into the MLLM 37

D.2.4 Conditional SAM2 Feature Enhancement 37

D.3 Prompt Distill Module 37

D.3.1 Semantic Dense Prompt Fusion 37

D.3.2 Geometric Sparse Prompt Distillation 38

D.4 Training and Inference Pipelines 38

D.4.1 Training Pipeline 39

D.4.2 Inference Pipeline 39

E Experiments **40**

E.1 Implementation Details 40

E.1.1 Model Setup & Dataset 40

E.1.2 Training and Inference Settings 40

E.1.3 Evaluation Metrics 41

E.2 Baseline Model Details 41

E.3 More Performance Analysis 42

E.3.1 By Manipulation Task 42

E.3.2 By Generation Model 43

E.3.3 From Classification to Attribution 43

E.3.4 Human Evaluation for Anomaly Explanation 45

E.4 Ablation Study 46

E.4.1 Artifact Grounding 46

E.4.2 Anomaly Explanation 47

E.4.3 Forgery Detection 47

E.4.4 Prompt Design 48

E.5 Robustness Study 49

E.6 ManiVidLens Case Visualization 49

A Related Work

A.1 Video Forgery Detection Datasets

Existing video forgery datasets can be broadly divided into benchmarks for fully synthetic videos and datasets for localized video manipulation. The former primarily collect videos produced by T2V or I2V models (Hong et al., 2022; Blattmann et al., 2023; Kong et al., 2024; Yang et al., 2025) and differ in generator coverage, semantic alignment, and annotation granularity (Bai et al., 2024a; Chen et al., 2026; Ni et al., 2026; Wang & Yang, 2024; Li et al., 2025; Park et al., 2026). In contrast, established manipulation datasets largely focus on facial forgeries (Rössler et al., 2018; Dolhansky et al., 2020), while VideoSham (Mittal et al., 2023) extends the scope to general spatial and temporal manipulations. Recent video editing datasets provide more diverse edit types but are designed primarily for training and evaluating editing models rather than forensic analysis (Bian et al., 2025; He et al., 2025). No existing resource combines paired, open-vocabulary video manipulations with authenticity labels, pixel-level forgery masks, and anomaly explanations.

A.2 Video Forgery Detection Methods

Applying image detectors independently to video frames yields poor performance on AI-generated videos (Vahdati et al., 2024). Video detectors instead model spatial artifacts and temporal inconsistencies through representation-trajectory geometry, normalized spatiotemporal gradients, or second-order temporal differences (Internò et al., 2025; Zhang et al., 2025a; Zheng et al., 2025). However, these methods primarily target fully generated videos and provide authenticity predictions without artifact localization or explanation. Recent methods using MLLMs support image forgery detection and textual explanation (Wen et al., 2025; Zhu et al., 2026; Huang et al., 2025b; Lin et al., 2025), while VidGuard-R1 extends these capabilities to videos but does not provide artifact localization (Park et al., 2026). Omni-Fake-R1 jointly predicts authenticity, localization, and explana-

tions across multiple modalities, but its spatial localization is limited to bounding boxes rather than pixel-level masks (Li et al., 2026). Consequently, existing methods do not jointly support forgery detection, pixel-level artifact grounding, and anomaly explanation for manipulated videos.

B MOTIVATION

B.1 SCOPE COMPARISON: FULL SYNTHESIS VS. MANIPULATION

Full Synthesis. Fully synthetic videos are generated without reference to a source video, typically using existing T2V or I2V models. Under this scope, real and fake videos may differ substantially in both semantic and visual content, while basic video attributes such as resolution, frame rate, and duration are often unaligned. Consequently, they cannot form semantically and visually aligned real-fake video pairs. A common practice in existing dataset construction is to directly collect real videos from datasets such as Vript (Yang et al., 2024) or Kinetics-400 (Carreira & Zisserman, 2017), while generating fake videos separately from additional text prompts or reference images. For example, VidProM (Wang & Yang, 2024) employs a large collection of user-provided T2V prompts. Such discrepancies in content and video attributes can introduce significant shortcuts for forgery detection. Unfortunately, existing video forgery datasets, such as the widely used GenVideo (Chen et al., 2026) and GenVideoBench (Ni et al., 2026), predominantly focus on full synthesis and thus inevitably suffer from the aforementioned limitations. ***VidGuard-R1 (Park et al., 2026) claims that exploiting these shortcuts can yield over 96% accuracy on GenVideo and GenVideoBench.*** Although VidGuard-R1 recognizes this issue and attempts to better align fake videos with their source videos by preserving the first frame, its subsequent I2V generation lacks fine-grained geometric control, such as Canny guidance, and still relies primarily on text prompts. As a result, the generated videos may still deviate from their source videos in temporal dynamics, such as the speed at which events evolve over time, making precise source-target alignment difficult to achieve.

Manipulation. Manipulated videos are generated with explicit and complete source videos as references, typically through V2V/I2V models or manual editing. Unlike fully synthetic videos, manipulation modifies only localized regions, naturally preserving the overall semantic and visual content of the source videos while maintaining consistent basic video attributes. This alignment enables detectors to focus purely on spatial-temporal forgery artifacts rather than exploiting content or attribute shortcuts. In our data construction pipeline, the indirect I2V paradigm further introduces video-level Canny guidance during I2V generation after editing the first frame, precisely preserving the temporal evolution of unmanipulated regions in the source video. This produces realistic, high-quality manipulated videos and poses a substantially more challenging task for conventional forgery detection methods. Tab 7 illustrates the differences between these two scopes.

Table 7: **Scope Comparison between Full Synthesis and Manipulation.**

Dimension	Full Synthesis	Manipulation
Source Video Reference	None or partial	Complete
Generation Paradigm	Direct T2V / I2V	V2V / Conditional I2V
Modification Scope	Entire video	Localized regions
Real–Fake Video Pair	Unavailable	Naturally paired
Semantic Alignment	Low	High
Visual Alignment	Low	High
Temporal Alignment	Low	High
Shortcut Risk	High	Low
Detection Evidence	Global generation cues / Potential shortcuts	Local spatial-temporal forgery artifacts
Forensic Challenge	Distinguishing generated from real videos	Identifying subtle localized inconsistencies

B.2 MANIPULATION DATASET LIMITATION

As noted in 2.1, existing manipulation datasets suffer from three key limitations. **(1) Limited Diversity.** Existing datasets (Rossler et al., 2019; Dolhansky et al., 2020) are often restricted to specific domains and manipulation tasks, particularly face-related scenarios, which directly limits their applicability and potential in general-world scenarios. **(2) Closed-Vocabulary Instructions.** This limitation is common in existing image manipulation datasets. For instance, SIDA (Huang et al., 2025a) predefines a fixed set of target concepts for specific source objects. For *Bird*, the replace-

ment set is predefined as {*Owl, Duck, Chicken, Goose, Parrot, Eagle, Sparrow, Penguin*}. Such a closed-vocabulary design introduces two major issues. ***1) Homogeneity.*** a fixed target set leads to repetitive manipulation patterns, substantially restricting the flexibility and diversity of manipulation operations. ***2) Logical Inconsistency.*** Although SIDA partially considers semantic similarity when constructing target sets to reduce potential conflicts, category-level similarity does not guarantee contextual or logical consistency. For example, consider a bird resting on a blooming grassland in spring. If the bird is selected as the source object, its candidate replacements may all belong to the same semantic category, *Aves*. However, replacing it with a penguin would clearly violate common sense, as penguins are naturally adapted to cold environments rather than such a warm spring setting. In contrast, we employs an MLLM assistant to jointly reason over the video content and its detailed caption to generate reasonable open-vocabulary instruction proposals. This design preserves manipulation diversity while ensuring that the proposed manipulations remain contextually and logically plausible. **(3) Insufficient Annotations.** As shown in Tab. 1, existing video manipulation datasets typically provide only binary authenticity labels. VideoSham (Mittal et al., 2023) additionally releases manipulation instructions, but still lacks corresponding forgery masks and anomaly explanations, which are essential for explainable forensic analysis.

C ManiVid-38K Dataset

C.1 Curation Pipeline

C.1.1 Source Video Collection

We select three public datasets, OpenVid-1M (Nan et al., 2025), InternVid (Wang et al., 2024), and OpenVE-3M (He et al., 2025), as sources of real videos. The processing procedure and authenticity assurance for each dataset are detailed below.

OpenVid-1M. OpenVid-1M contains approximately 1M video-text pairs, including a high-resolution subset, OpenVidHD, with about 0.4M 1080P videos. Given its high video quality, we select OpenVidHD as the primary source of real videos for ManiVid-38K. OpenVidHD consists of videos from four sources: **(1) ChronoMagic** (Yuan et al., 2024), **(2) CelebvHQ** (Zhu et al., 2022), **(3) Open-Sora-Plan** (PKU-Yuan Lab and Tuzhan AI et al., 2024), and **(4) Panda** (Chen et al., 2024). CelebvHQ predates the release of Sora (OpenAI, 2024) in February 2024, while Open-Sora-Plan and Panda explicitly identifies its videos as non-AI-generated. In contrast, we exclude ChronoMagic (was released in April 2024) because its videos were collected from the Internet after the release of Sora, potentially introducing the risk of AI-generated content. Another advantage of OpenVidHD is its quality-related metadata, such as *Aesthetics Score*. We randomly sample a subset and prioritize videos with higher *Aesthetics Scores* to maximize the quality of the selected videos.

InternVid. InternVid contains 7.1M 720P videos accompanied by brief text captions. It was first released in 2023, with videos sourced from YouTube, providing reasonable assurance of their authenticity. However, given its lower resolution compared with OpenVidHD and the additional effort required to download videos from YouTube, we sample only a small subset as source videos.

OpenVE-3M. OpenVE-3M is a recent video editing dataset containing 3M source-target video pairs across diverse manipulation tasks. However, our comprehensive examination reveals three systematic issues: **(1) Source Authenticity.** OpenVE-3M collects 1M videos from Open-Sora-Plan (PKU-Yuan Lab and Tuzhan AI et al., 2024), OpenViDHD (Nan et al., 2025), and UltraVideo (Xue et al., 2025) as its base video library. Part of UltraVideo is collected from YouTube in 2025 and may therefore be exposed to potential contamination by AI-generated content. **(2) Reverse Editing.** OpenVE-3M constructs reverse editing samples by treating generated edited videos as sources and the original videos as targets, accompanied by corresponding reverse editing instructions. Consequently, videos appearing on the source side cannot be reliably assumed to be authentic. **(3) Manipulation Plausibility.** We further observe some implausible manipulation instructions. For example, changing a young boy's hair to the gray-white hair of an elderly person can produce an obviously unrealistic result that is easily identifiable by human inspection. Overall, OpenVE-3M is primarily developed from a video generation perspective, emphasizing generation quality rather than source authenticity. From a forgery detection perspective, however, its use requires additional caution. We therefore sample only a very small subset of source videos and manually inspect every selected sample to ensure that no noticeable manipulation or AI-generated artifacts are present.

Table 8: **Data Statistics During Collection.** The number of source videos retained at different stages of the collection process.

Source Dataset	OpenVid-1M	InternVid	OpenVE-3M	Total
Sampled Num	222391	5882	15000	243273
+ Deduplication	139697	3974	9292	152963
+ Quality Filter	35643	2959	2506	41108

Deduplication & Quality Filter. We construct a source video library from the three datasets above and further review and filter the collected videos to ensure overall data quality. The filenames provided by the source datasets allow us to identify clips originating from the same source video. We therefore perform source-level deduplication based on these identifiers and retain at most one clip from each original source video. This prevents highly similar or temporally adjacent clips from the same source video from appearing multiple times in ManiVid-38K. We then apply quality filtering to the remaining videos and retain only those that simultaneously satisfy three criteria: ***1) resolution of at least 720p***, ***2) frame rate of at least 15 FPS***, and ***3) duration no longer than 10 seconds***. Finally, 41,108 source videos are retained in our source video library, as shown in Tab. 8.

C.1.2 Content Categorization

To ensure the content diversity of ManiVid-38K dataset, we categorize the samples in the source video library according to their video content. We define six valid content categories and one invalid category (***Uncertain***), as detailed in Tab. 9. We then employ the MLLM assistant to categorize each source video based on the video input (**Required**) and its caption (**Optional**). Tab. 10 presents the categorization prompt. Finally, we adopt balanced sampling to maintain a uniform number of samples across categories. Fig. 1 shows the final content distribution of our ManiVid-38K dataset.

Table 9: **Content Category Definitions and Examples.**

Content Category	Definition	Examples
Human	The video is centered on one or more humans. The person(s) usually occupy a large portion of the frame or are the clear narrative focus. The background is typically simple and does not form a complete environment. Notes: (1) If humans appear but are small and the environment is the main focus, consider *Indoor* or *Outdoor* instead. (2) If the video mainly conveys a time-continuous activity or process where "what is happening" is more important than "who the person is", consider *Event*.	*Reporter, Police Officer, Businessman, Farmer, ...*
Animal	The video is centered on one or more animals, which occupy a large portion of the frame or are the narrative focus. Notes: (1) If animals are small or only decorative and the environment dominates, use *Indoor* or *Outdoor*. (2) If the video mainly presents a continuous activity performed by the animals, consider *Event*.	*Cat, Dog, Bird, Horse, Cow, Sheep, ...*
Object	The video focuses on one or a few inanimate objects. The background is simple and does not form a complete environment. Notes: (1) Hands or partial human bodies may appear, but the object itself is the core subject. (2) If the video emphasizes a multi-step, time-continuous process, consider *Event*.	*Knife, Fork, Food, Tool, Cup, Toy, Laptop, ...*
Indoor	The video primarily presents a complete and visually rich indoor environment (e.g., rooms, offices, shops). No single subject dominates, or subjects are small and secondary. Notes: (1) Use this when the spatial layout, interior structure, and overall atmosphere are the main information.	*Room, Office, Shop, Gym, Hospital, ...*
Outdoor	The video primarily presents a complete and visually rich outdoor environment or scenery (e.g., natural landscapes, city streets, building exteriors, aerial views). No single subject dominates, or subjects are small and secondary. Notes: (1) Even if people or animals appear, choose *Outdoor* as long as they are minor and the scenery is the main focus.	*Natural Landscape, City Street, Aerial View, ...*
Event	The video is centered on a time-continuous activity or event. The key information lies in what is happening over time rather than the appearance of a specific person, animal, object, or environment. Notes: (1) It includes any multi-step actions with clear temporal progression. (2) The main subject may be human, animal, or object, but the activity itself is the narrative core.	*Cooking Process, Ceremony, Sports Game, ...*
Uncertain	The video does not fit any of the above categories, or the quality/content is insufficient (blurred, too short, occluded, missing key information) to make a reliable judgment.	-

Table 10: **Content Categorization Prompt.**

Content Categorization Prompt

Task Description:

You are given a **video** and its **original caption**: <ORIGINAL_CAPTION>.
Your task is to classify the video into **exactly one** of the following categories based on the definitions below. If none of the categories apply, or the video quality/content is insufficient for a confident decision, choose category ***G***.

Category Definitions:

A. Human
Definition: The video is centered on one or more humans. The person(s) usually occupy a large portion of the frame or are the clear narrative focus. The background is typically simple and does not form a complete environment.
Notes:
1. If humans appear but are small and the environment is the main focus, consider ***Indoor*** or ***Outdoor*** instead.
2. If the video mainly conveys a time-continuous activity or process where "what is happening" is more important than "who the person is", consider ***Event***.

B. Animal
Definition: The video is centered on one or more animals, which occupy a large portion of the frame or are the narrative focus.
Notes:
1. If animals are small or only decorative and the environment dominates, use ***Indoor*** or ***Outdoor***.
2. If the video mainly presents a continuous activity performed by the animals, consider ***Event***.

C. Object
Definition: The video focuses on one or a few inanimate objects (*e.g.*, products, tools, food). The background is simple and does not form a complete environment.
Notes:
1. Hands or partial human bodies may appear, but the object itself is the core subject.
2. If the video emphasizes a multi-step, time-continuous process (*e.g.*, full cooking or repair workflow), consider ***Event***.

D. Indoor
Definition: The video primarily presents a complete and visually rich indoor environment (*e.g.*, rooms, offices, shops). No single subject dominates, or subjects are small and secondary.
Notes: Use this when the spatial layout, interior structure, and overall atmosphere are the main information.

E. Outdoor
Definition: The video primarily presents a complete and visually rich outdoor environment or scenery (*e.g.*, natural landscapes, city streets, building exteriors, aerial views). No single subject dominates, or subjects are small and secondary.
Notes: Even if people or animals appear, choose ***Outdoor*** as long as they are minor and the scenery is the main focus.

F. Event
Definition: The video is centered on a time-continuous activity or event. The key information lies in what is happening over time rather than the appearance of a specific person, animal, object, or environment.
Notes:
1. Typical examples include sports games, cooking processes, performances, ceremonies, construction or repair workflows, and any multi-step actions with clear temporal progression.
2. The main subject may be human, animal, or object, but the activity itself is the narrative core.

G. Uncertain
Definition: The video does not fit any of the above categories, or the quality/content is insufficient (blurred, too short, occluded, missing key information) to make a reliable judgment.

Requirements:

1. Output a single JSON object and nothing else, with no explanation, no chain of thought, and no markdown fences: {"category": <CategoryLetter>}
2. The <CategoryLetter> must be exactly one of: ***A*** / ***B*** / ***C*** / ***D*** / ***E*** / ***F*** / ***G***.

C.1.3 DETAILED RECAPTION

Since OpenVE-3M does not provide captions for its videos, while the other two data sources, particularly InternVid, contain only extremely brief descriptions (with an average length of 17.6), the available captions are semantically sparse and often omit many potential objects, as illustrated in Fig. 2. Therefore, we employ an MLLM assistant to uniformly recaption all samples, enriching the textual information with ***precise attributes***, ***added objects***, ***spatial details***, and ***event cues***, which provides a foundation for subsequent manipulation source object selection, instruction generation, and feasibility evaluation. Tab. 11 provides the prompt used for recaptioning, and Fig. 1 shows the caption length distribution of our ManiVid-38K dataset.

Table 11: **Video Recaption Prompt.**

Video Recaption Prompt

Task Description:

You are given a **video** and its **original caption**: `<ORIGINAL_CAPTION>`.
Your task is to generate a **detailed caption** by refining and expanding the original caption with the visual content of the video. You are allowed to **correct obvious mistakes** in the original caption if they conflict with the video.

Requirements:

1. Output a single JSON object and nothing else, with no explanation, no chain of thought, and no markdown fences: {"refined_caption": `<Detailed Caption>`}
2. The `<Detailed Caption>` should refine and enrich the original caption, may correct obvious errors, and should describe the main subjects, environment, any notable temporal changes and detailed object attributes.
3. The `<Detailed Caption>` must be objective and factual. Avoid subjective opinions, emotional language, or rhetorical devices such as metaphors or similes.

C.1.4 INSTRUCTION GENERATION

Task Definition. We select six common manipulation tasks for data construction, including ***Addition***, ***Removal***, ***Replacement***, ***Color Alteration***, ***Material Modification***, and ***Portrait Editing***. The definition and examples of each manipulation task are provided in Tab. 12.

Table 12: **Manipulation Task Definitions and Examples.**

Manipulation Task	Definition	Example
Addition	Integrate a new, semantically relevant object into the scene.	*Add a red backpack onto the empty bench in `<video>`.*
Removal	Erase a specific object with clear boundaries and inpaint the background naturally.	*Remove the blue suitcase from `<video>`, seamlessly inpainting the revealed pavement.*
Replacement	Swap an object with a fine-grained variant or a logically related entity.	*Replace the white sedan in `<video>` with a red SUV, preserving the original motion trajectory.*
Color Alteration	Change the color of a specific object while preserving its original texture and shape.	*Change the color of the woman's jacket in `<video>` to blue, maintaining its original fabric texture.*
Material Modification	Alter the physical material properties.	*Change the material of the rotating vase in `<video>` into polished metal, maintaining its original motion.*
Portrait Editing	Modify facial features, expressions, or skin tones. (**Note**: Only applicable to close-up or medium shots of people).	*Modify the facial features of the woman at the center of `<video>` to exhibit a smiling expression, keeping others consistent.*

Task Initialization. We randomly assign an initial manipulation task to each source video, as we find that allowing the MLLM assistant to directly select a suitable task based on the video content leads to a severely imbalanced task distribution. Specifically, the MLLM tends to favor simpler tasks such as ***Addition*** and ***Removal***, while being less inclined to attempt more challenging tasks such as ***Material Modification*** and ***Portrait Editing***.

Idea Inspiration. We instruct the MLLM assistant to brainstorm manipulation ideas based on the source video, the initially assigned task, and the detailed recaption. Since the initial task is assigned completely at random, two cases may arise: **(1) The assigned task is suitable for the source video.** In this case, the MLLM identifies a potential source object suitable for the manipulation task and leverages its prior knowledge to propose a reasonable target. **(2) The assigned task is unsuitable**

for the source video. For example, a landscape video may be assigned the ***Portrait Editing*** task. In this case, the MLLM summarizes the reason for the failure and switches to a manipulation task that it considers suitable until case **(1)** is reached. After generating the manipulation instruction, we further require the assistant to perform reflection to ensure that the instruction conforms to physical principles and event logic, and to provide a justification for the reasonableness of the instruction in its output, thereby avoiding unreasonable instructions such as those illustrated in B.2. Tab. 13 provides the prompt used for manipulation instruction generation, and Fig. 1 shows the final distribution of manipulation tasks in ManiVid-38K dataset.

Table 13: **Idea Inspiration Prompt.**

Idea Inspiration Prompt

Task Description:

You are an expert in video content editing and forensic analysis. Your task is to design a **high-quality, photorealistic video manipulation scheme** based on a source video, its detailed caption and an initial manipulation task. This scheme will be used to test the robustness of manipulation detection approaches.

Input Data:

1. **Source Video**: `<video>`
2. **Detailed Caption**: `<detailed caption>`
3. **Initial Manipulation Task**: `<initial manipulation task>`

Manipulation Task Definitions:

- ***Addition***: Integrate a new, semantically relevant object into the scene.
- ***Removal***: Erase a specific object with clear boundaries and inpaint the background naturally.
- ***Replacement***: Swap an object with a fine-grained variant or a logically related entity.
- ***Color Alteration***: Change the color of a specific object while preserving its original texture and shape.
- ***Material Modification***: Alter the physical material properties.
- ***Portrait Editing***: Modify facial features, expressions, or skin tones. (**Note**: Only applicable to close-up or medium shots of people).

Idea Inspiration Principles:

Your proposal must explicitly adhere to these four criteria:

- **Logical Consistency**: The modification must follow the laws of physics, common sense, and maintain consistent lighting and occlusion.
- **Diversity**: Avoid repetitive or generic modifications. Use specific, varied nouns.
- **Contour Clarity (Segmentability)**: Target objects must have well-defined edges. Avoid ambiguous or amorphous entities like smoke, fire, or flowing water.
- **Spatio-Temporal Uniqueness**: Define the target entity with absolute precision across both the spatial layout and the temporal timeline to guarantee zero tracking ambiguity across consecutive frames. The target must be uniquely identifiable not just in a single isolated frame, but continuously throughout the entire video sequence.

Task Feasibility Check:

Evaluate whether the **Initial Manipulation Task** is feasible for the video content. If the assigned task is incompatible (*e.g.*, ***Portrait Editing*** for a landscape video), you **MUST** first summarize the reason for its infeasibility, then switch to another task and repeat this process until a reasonable and feasible task is selected.

Requirements:

The output **MUST** consist of exactly 4 sections separated by the `<sep>` token. Do **not** include any introductory remarks or additional explanations. Do **not** include the name of each sections.

1. **Change Task**: Output only one word, `YES` if you changed the assigned task type; otherwise, output `NO`.
2. **Final Task**: Return exactly one of the 6 predefined task names that strictly corresponds to your actual, finalized manipulation instruction.
3. **Design Justification**: Think step-by-step to draft the manipulation scheme. You **MUST** explicitly evaluate your proposed idea against inspiration principals before finalizing the object and instruction, and output a brief professional explanation.
4. **Manipulation Instruction**: Formulate the instruction with high information density. While the target entity must be precisely defined, you **MUST NOT** include irrelevant background context, atmospheric narrative, or edge concepts. Optimize your phrasing for a downstream video generation model. Use the templates below just as your structural guide:
   - ***Addition***: Add [New Object] into `<video>`. Need to specify a natural location without forcing complex spatial relationships.
   - ***Removal***: Remove [Original Object] from `<video>`, seamlessly inpainting the revealed [Background].
   - ***Replacement***: Replace [Original Object] in `<video>` with [Target Object], preserving the original [Motion Trajectory].

- ***Color Alteration***: Change the color of [Original Object] in <video> to [New Color], maintaining its original [Specific Texture].
- ***Material Modification***: Change the material of [Original Object] in <video> into [New Material], maintaining its original [Motion].
- ***Portrait Editing***: Modify the facial features of [Human] in <video> to exhibit [New Expression/Identity], keeping others consistent.

C.1.5 MANIPULATION SETUP

Manipulation Paradigms. With the rapid maturation of generation toolchains, we adopt two complementary manipulation paradigms to better cover diverse real-world editing operations. In addition to the conventional **Direct V2V** paradigm, we introduce an **Indirect I2V** paradigm that combines first-frame editing with controllable I2V generation. Specifically, we assign 53.8% of the source videos to the V2V paradigm and 46.2% to the I2V paradigm.

Model Allocation. Because generative models exhibit nonuniform trade-offs across quality, alignment, robustness, and linguistic conditions (Zhang et al., 2025b; 2026), manipulation models are randomly assigned within both paradigms to avoid model-selection bias. For **Direct V2V**, each source video is randomly assigned to one of the seven video editing models. For **Indirect I2V**, model allocation is performed independently at both stages: an image editing model is randomly selected from six models for first-frame editing, followed by a randomly selected Canny-guided I2V model from three models for final video generation. Tab. 14 and Fig. 1 show the final model allocation in our ManiVid-38K dataset.

Table 14: **Detailed Distribution of Manipulation Paradigms and Models in ManiVid-38K.**

Paradigm	V2V Model	(Stage 1) Image Editing Model	(Stage 2) Canny-guided I2V Model
Direct V2V (53.80%)	Kling-O1 (5.80%) Kling-V3-Omni (2.65%) Runway Aleph (8.48%) Wan 2.2-Video-Edit (10.43%) VideoCoF (21.36%) VACE (4.39%) VideoPainter (0.69%)	–	–
Indirect I2V (46.20%)	–	Nano Banana 2 (13.92%)	Kling-V3-Omni (1.77%) Wan 2.2 (5.54%) LTX-2 (6.60%)
		Seedream 5.0 (14.16%)	Kling-V3-Omni (1.81%) Wan 2.2 (7.72%) LTX-2 (4.63%)
		GPT-Image-1.5 (0.39%)	Kling-V3-Omni (0.00%) Wan 2.2 (0.39%) LTX-2 (0.00%)
		FLUX.2-dev-32B (5.75%)	Kling-V3-Omni (1.14%) Wan 2.2 (1.27%) LTX-2 (3.34%)

Continued on next page

Table 14 (continued)

Paradigm	V2V Model	(Stage 1) Image Editing Model	(Stage 2) Canny-guided I2V Model
		FLUX.2-klein-9B (3.34%)	Kling-V3-Omni (0.00%)
			Wan 2.2 (1.41%)
			LTX-2 (1.93%)
		Qwen-Image (8.64%)	Kling-V3-Omni (0.43%)
			Wan 2.2 (3.87%)
			LTX-2 (4.34%)

C.1.6 Generation Model Details

Kling-O1 (Team et al., 2025) Kling-O1 belongs to the Kling-Omni framework, which unifies video generation and editing through multimodal instruction following. It conditions video creation on text, reference images, and video context, enabling edits that combine semantic instructions with information from an existing scene.

Kling-V3-Omni (Kling AI, 2026) Kling-V3-Omni is a unified multimodal video model that supports reference-conditioned generation, video editing, and audiovisual output. Its reference and element controls help maintain subject consistency across generated content, and we use it in both the direct V2V and indirect I2V manipulation paradigms.

Runway Aleph (Runway, 2025) Runway Aleph is an in-context video model that edits an existing clip according to user instructions. Its supported operations include adding, removing, and transforming objects, changing viewpoints, and modifying scene style or lighting.

Wan 2.2-Video-Edit (WaveSpeedAI, 2025) Wan 2.2-Video-Edit is a Wan-based video-editing model provided through WaveSpeedAI that accepts a source video and a textual edit instruction. It supports changes to visual content such as clothing and characters, allowing the source clip to be directly transformed under the V2V paradigm.

VideoCoF (Yang et al., 2026) VideoCoF is a unified video-editing framework that uses a chain-of-frames reasoning process to determine where an instruction should be applied. Its diffusion model predicts edit-region reasoning latents before generating the target video, providing spatial guidance without requiring user-supplied masks.

VACE (Jiang et al., 2025) VACE unifies reference-to-video generation, video-to-video editing, and masked video editing within a diffusion-based framework. It combines reference content, source frames, and masks in a shared video condition unit, supplied to the model through a context adapter.

VideoPainter (Bian et al., 2025) VideoPainter is a video-inpainting and editing framework with separate streams for background context and content generation. A lightweight context encoder supplies masked-video information to a pretrained video diffusion transformer, while target-region identity resampling supports editing videos beyond a single generation window.

Nano Banana 2 (Google, 2026) Nano Banana 2, also known as Gemini 3.1 Flash Image, is Google's image-generation and editing model designed for rapid instruction-guided visual creation. It supports reference-based edits and subject consistency across transformations, making it suitable for producing edited first frames from source images and manipulation instructions.

Seedream 5.0 (ByteDance Seed, 2026) Seedream 5.0 is ByteDance's image-generation and editing family, with the cited Lite release combining multimodal understanding, reasoning, and image synthesis. It supports instruction-guided local edits and subject replacement while preserving surrounding content, and serves as a first-frame editor in our indirect I2V pipeline.

GPT-Image-1.5 (OpenAI, 2025) GPT-Image-1.5 is a proprietary image-generation and editing model that follows natural-language instructions over text and image inputs. It supports localized and broader visual transformations while retaining details such as subject identity, lighting, and composition, enabling controlled modification of the source video's first frame.

FLUX.2-dev-32B (Black Forest Labs, 2025) FLUX.2-dev-32B is an open-weight rectified-flow transformer with 32 billion parameters for image generation and editing. It combines single- and multiple-reference image conditioning with guidance distillation for efficient inference.

FLUX.2-klein-9B (Black Forest Labs, 2026) FLUX.2-klein-9B is a compact rectified-flow transformer that combines text-to-image generation with reference-based image editing. Its nine-billion-parameter flow model is distilled to four inference steps, providing an efficient alternative for instruction-guided first-frame editing.

Qwen-Image (Wu et al., 2025) Qwen-Image is an image-generation and editing model that combines multimodal semantic understanding with a diffusion transformer. Its editing design encodes the source image into semantic and reconstructive representations, helping preserve both the intended content and visual fidelity during instruction-guided changes.

Wan 2.2 (Wan et al., 2025) Wan 2.2 is an open video diffusion model in the Wan family that supports image-conditioned video synthesis. In our indirect I2V pipeline, the edited first frame defines the target appearance, while source-video Canny maps guide the scene geometry.

LTX-2 (HaCohen et al., 2026) LTX-2 is an open audiovisual generation model with asymmetric video and audio transformer streams connected through bidirectional cross-attention. Its two streams generate temporally aligned audiovisual content; our indirect I2V pipeline conditions video generation on an edited first frame and Canny guidance.

C.1.7 PARADIGM A: DIRECT V2V CASES

Fig. 5 shows examples of direct manipulation performed by different V2V models. Specifically, we select four recent closed-source models, including Kling-O1 (Team et al., 2025), Kling-V3-Omni (Kling AI, 2026), Runway Aleph (Runway, 2025), Wan 2.2-Video-Edit (WaveSpeedAI, 2025), together with three open-source models, including VideoCoF (Yang et al., 2026), VACE (Jiang et al., 2025) and VideoPainter (Bian et al., 2025).

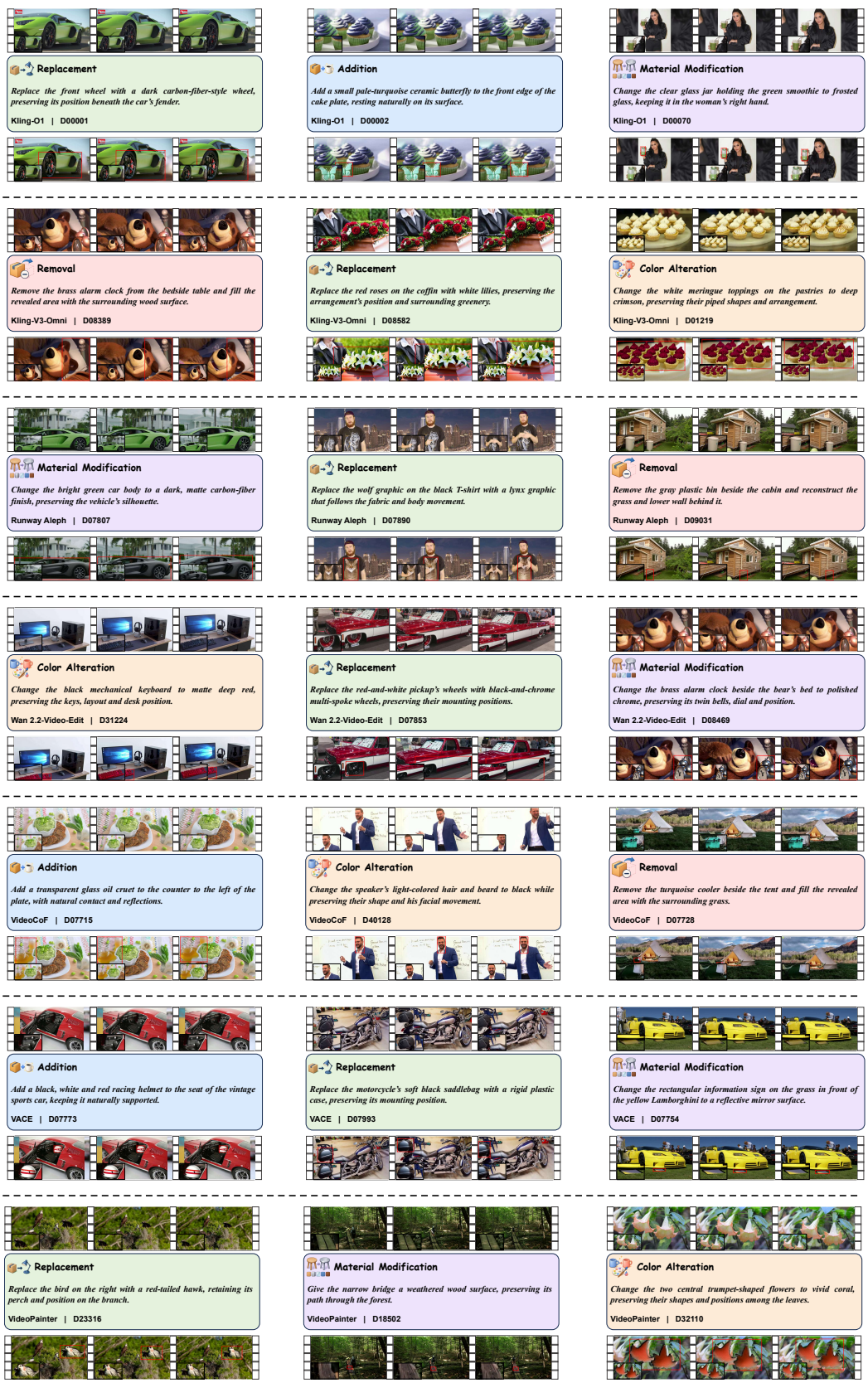


Figure 5: **Manipulation Examples of Seven Direct V2V Models.** The first four rows are closed source models, while the last three rows are open source models. Zoom in for a better view.

C.1.8 PARADIGM B: INDIRECT I2V CASES

Fig. 6 shows examples of indirect manipulation performed under the I2V paradigm. Specifically, for Stage 1 image editing, we select three recent closed-source models, including Nano Banana 2 (Google, 2026), Seedream 5.0 (ByteDance Seed, 2026), and GPT-Image-1.5 (OpenAI, 2025), together with three open-source models, including FLUX.2-dev (Black Forest Labs, 2025), FLUX.2-klein-9B (Black Forest Labs, 2026), and Qwen-Image (Wu et al., 2025). For Stage 2 controllable I2V generation, we select three Canny-guided I2V models, including the closed-source Kling-V3-Omni (Kling AI, 2026) and two open-source models, Wan 2.2 (Wan et al., 2025) and LTX-2 (HaCohen et al., 2026).

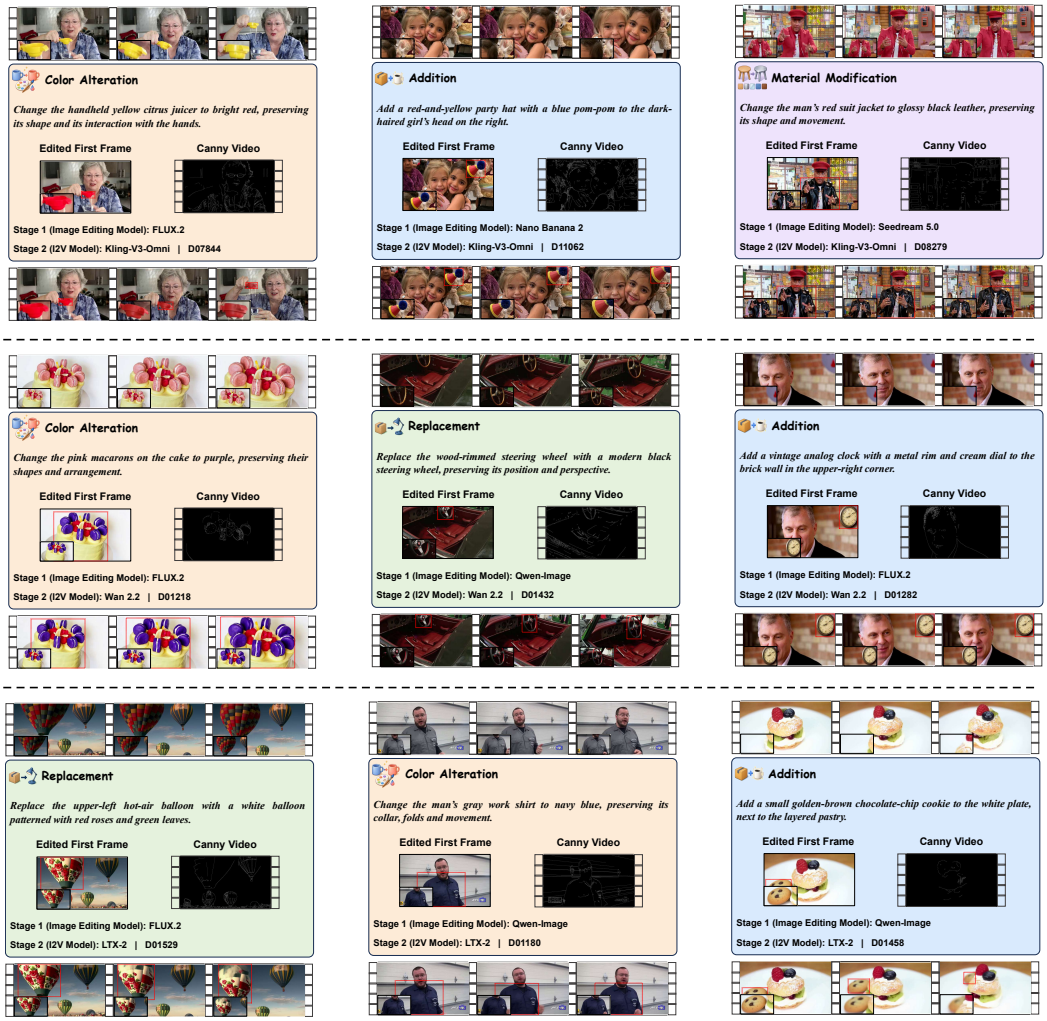


Figure 6: **Manipulation Examples of three indirect I2V Models.** The first row is Kling-V3-Omni (closed source), while the last two rows are open source models. Zoom in for a better view.

C.1.9 AUTOMATIC REFLECTION

To ensure high-quality manipulation results while reducing unnecessary time and resource costs in subsequent annotation and quality control, we employ the MLLM assistant to automatically assess whether each manipulation is successful. Specifically, given the manipulation instruction, the assistant compares the source and manipulated videos, focusing on three key aspects: ***1) Manipulation Visibility***, ***2) Instruction Following***, and ***3) Content Fidelity***. Only samples satisfying all three criteria proceed to subsequent annotation and quality control. During this process, we retain about 79.8% of the samples in total. The prompt used for automatic reflection is provided in Tab. 15.

Table 15: **Automatic Reflection Prompt.**

Automatic Reflection Prompt

Task Description:

You are given a **source video**, a **manipulation instruction**: <MANIPULATION_INSTRUCTION>, and the corresponding **manipulated video**.
Your task is to compare the source video and the manipulated video according to the manipulation instruction, and determine whether the manipulation is successfully performed.

Manipulation Check:

1. **Manipulation Visibility:** Is the intended manipulation visible in the manipulated video?
2. **Instruction Following:** Does the manipulated result comply with the manipulation instruction?
3. **Content Fidelity:** Are the remaining unmanipulated regions preserved compared with the source video?

Requirements:

1. Carefully compare the **source video** and the **manipulated video** with reference to the **manipulation instruction**.
2. For each of the three checks above, output only YES or NO.
3. The manipulation is considered successful only when all three checks are YES.
4. Output a single JSON object and nothing else, with no explanation, no chain of thought, and no markdown fences:

   {"manipulation_visible": "<YES/NO>", "instruction_following": "<YES/NO>", "content_fidelity": "<YES/NO>", "success": "<YES/NO>"}

C.1.10 Artifact Grounding

We first employ an MLLM assistant to extract a concise and common *segmentation phrase* from the manipulation instruction, since SAM3 (Carion et al., 2026) is less compatible with rare concepts and lengthy textual prompts. We then feed the segmentation phrase together with the target video into SAM3 to obtain the forgery masks. Specifically, for the ***Removal*** task, the source video is used as the target video because the manipulated video no longer contains the target semantic content; for all other manipulation tasks, the manipulated video is used instead.

Dilemma: Poor Performance. However, we observe that ***SAM3 performs poorly when the target reference is ambiguous or multiple similar objects are present***, which substantially degrades the quality of artifact annotations. Therefore, we introduce additional spatial priors to enhance the segmentation performance of SAM3.

Solution: Point Annotation. Through preliminary experiments, we find that complex bbox or polygon annotations are unnecessary, as only a few simple positive and negative point prompts can substantially improve the segmentation performance of SAM3 in challenging scenarios. Specifically, ***positive points*** indicate manipulated regions, while ***negative points*** indicate potentially confusing preserved regions. For each sample, the total number of positive and negative points is limited to at most six. The entire process is conducted through crowdsourcing, during which annotators are also instructed to directly discard samples with poor manipulation quality or those inconsistent with the manipulation instruction. During this process, we retain about 77.9% of the samples in total.

C.1.11 Anomaly Explanation

Inspired by VidGuard-R1 (Park et al., 2026), we define a comprehensive taxonomy of anomaly explanations from eight forensic perspectives: **(1) Boundary Coherence**, **(2) Perspective and Proportion**, **(3) Contextual Inconsistency**, **(4) Color and Lighting**, **(5) Texture Discontinuity**, **(6) Temporal Flicker**, **(7) Motion Artifact**, and **(8) Ghosting**, as detailed in Tab. 16. Given the manipulated video and its forgery mask, we prompt the MLLM assistant to select the four most applicable perspectives and generate corresponding forensic explanations. For removal cases, where the manipulated region corresponds to the reconstructed background rather than the removed object itself, the assistant is specifically instructed to analyze the visual traces of background inpainting without inferring the removed content. The explanation generation prompt is provided in Tab. 17.

Table 16: **Definitions of Anomaly Explanation Perspectives.**

Anomaly Perspective	Definition
Boundary Coherence	Imperfect transitions between the manipulated region and the original footage, including overly smoothed or unnaturally sharp boundary edges, as well as texture discrepancies in surface noise, resolution, or material quality relative to the surrounding unedited areas.
Perspective and Proportion	Geometric spatial errors where the manipulated content violates the camera perspective, exhibits focal-length-related distortions, or shows unnatural scaling and structural proportions relative to its immediate environment.
Contextual Inconsistency	Semantic or logical anomalies where the manipulated content defies the established scene context, including physically impossible object interactions, gravity-defying kinematics, or elements that fundamentally contradict surrounding real-world physics.
Color and Lighting	Mismatches in illumination, shadow trajectories, or color-space distribution between the manipulated content and the global physical environment.
Texture Discontinuity	Sudden degradation in surface quality, excessive smoothing, or unnatural loss of high-frequency details within the manipulated region.
Temporal Flicker	Unstable pixel generation or erratic high-frequency variations of static visual features across consecutive frames, typically manifested as localized shimmering or micro-jitter in textures that should remain structurally rigid.
Motion Artifact	Kinematic inconsistencies, physical deformations during movement, or unnatural motion trajectories of the manipulated content over time.
Ghosting	Unnatural translucency, residual visual echoes, or trailing artifacts associated with the manipulated content during frame transitions.

Table 17: **Anomaly Explanation Prompt.**

Anomaly Explanation Prompt

Task Description:

You are a leading digital forensics expert and video analyst, skilled at identifying deepfakes and AI-generated fake videos by analyzing subtle visual clues. Given a manipulated video and its target manipulation region, your task is to identify anomalous visual features and generate a rigorous forensic explanation that connects the observed abnormalities to AI manipulation. Specifically, analyze the targeted region from the following eight forensic perspectives and select the **four most applicable** ones for explanation.

Input Data:

1. **Annotated Video**: A manipulated video in which the target manipulation region (bounding boxes) is visually highlighted across frames.
2. **Manipulation Context**: Textual information describing the manipulation type and target content, used jointly with the visual indication to identify the manipulated region.

Anomaly Perspective Definitions:

- ***Boundary Coherence***: Imperfect transitions between the manipulated region and the original footage, exhibiting overly smoothed or unnaturally sharp boundary edges. This also includes texture discrepancies where the surface noise, resolution, or material quality fundamentally clashes with the surrounding unedited areas.
- ***Perspective and Proportion***: Geometric spatial errors where the manipulated content violates the camera's perspective grid, exhibits focal-length distortions, or shows unnatural scaling and structural proportions relative to its immediate environment.
- ***Contextual Inconsistency***: Semantic or logical anomalies where the manipulated content defies the established scene context. This includes physically impossible object interactions, gravity-defying kinematics, or elements that fundamentally contradict the surrounding real-world physics.
- ***Color and Lighting***: Mathematical mismatches in illumination, shadow trajectories, or color-space distribution compared with the global physical environment.
- ***Texture Discontinuity***: Sudden degradation in surface quality, over-smoothing, or unnatural loss of high-frequency details within the manipulated region.
- ***Temporal Flicker***: Unstable pixel generation or erratic high-frequency variations of static features across consecutive frames. Such mathematical instability typically manifests as localized shimmering or micro-jitter in textures that should remain structurally rigid.
- ***Motion Artifact***: Kinematic inconsistencies, physical deformations during movement, or unnatural trajectories of the manipulated content over time.
- ***Ghosting***: Unnatural translucency, residual visual echoes, or trailing artifacts during frame transitions.

Requirements:

1. **Target Region Description**: Provide a concise and unambiguous referring expression of at most 20 words for the manipulated object or background. State exactly **what** it is, its key visual attributes such as color or shape,

and **where** it is located. Do **not** mention the visual annotation or bounding box. The description should be suitable as a direct referring phrase for a vision segmentation model.

2. **Forensic Reasoning**: For each selected perspective, follow a strict evidentiary progression:
   (a) summarize the objective visual anomaly;
   (b) explain how the anomaly violates physical laws or optical camera capture within the corresponding perspective;
   (c) justify why this divergence serves as a signature of AI generation.
3. **Removal Cases**: For ***Removal*** manipulations, the target region corresponds to the reconstructed background after object removal rather than the removed object itself. Focus only on the visual traces of background inpainting and reconstruction. Do **not** attempt to restore, infer, or guess the removed content. The target-region description should instead refer to the reconstructed background patch, including its appearance, spatial location, and adjacent content.
4. **Objectivity and Focus**: Do **not** hallucinate visual features, over-interpret normal artifacts, or introduce external examples. Strictly analyze the content within the targeted manipulation region.
5. **Output Format**: Return the response strictly as a valid JSON object without Markdown formatting, code blocks, or conversational filler. The JSON must contain one `mani_description` field and exactly four fields corresponding to the selected anomaly perspectives.

C.1.12 HUMAN VERIFICATION

To ensure the final dataset quality, we introduce human verification, jointly conducted by all authors, as the final quality control stage, comprehensively evaluating each sample from four aspects:

- ***Task Reasonableness.*** Although the MLLM assistant has already assessed the reasonableness of task assignment and manipulation instructions during instruction generation, we conduct a final human inspection to identify potential oversights and ensure that the assigned task and instruction for each source video are feasible and consistent with common sense and event logic.
- ***Manipulation Quality.*** Although manipulation quality has already been assessed by the MLLM assistant during the Auto-Reflection stage, we further perform human inspection to ensure that the manipulated videos contain no obvious visual artifacts or instruction misalignment, thereby maintaining the quality and challenge of the ManiVid-38K dataset.
- ***Explanation Fidelity.*** We manually inspect the anomaly explanations generated by the MLLM assistant for evident hallucinations, including incorrect or nonexistent manipulated objects, overly homogeneous descriptions, and anomaly descriptions inconsistent with the actual video content.
- ***Mask Accuracy.*** We visually inspect the generated segmentation masks against the target objects specified in the manipulation instructions, ensuring that the manipulated objects are correctly and completely segmented while excluding samples with evident under-segmentation, over-segmentation, or incorrect target segmentation.

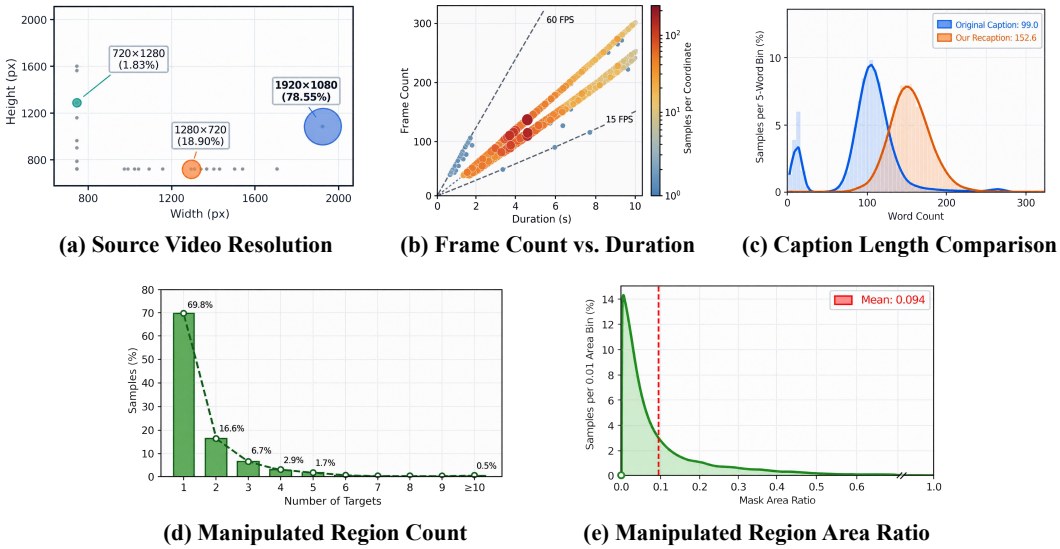
Samples that fail any criterion are discarded, with about 75.2% of the samples retained to constitute our ManiVid-38K dataset.

C.2 STATISTICAL ANALYSIS

Figure 7 summarizes the video and annotation characteristics of ManiVid-38K. Source videos predominantly have a resolution of 1920×1080 (78.55%), followed by 1280×720 (18.90%), with frame counts varying across durations and common frame rates. Recaptioning increases the average description length from 99.0 to 152.6 words, providing expanded textual descriptions for manipulation instruction generation. Spatially, 69.8% of samples contain one manipulated region, while 16.6% and 6.7% contain two and three, respectively. The manipulated area occupies only 9.4% of the frame on average, with the distribution concentrated toward smaller regions. These statistics characterize localized edits with varying spatial extent and region count, motivating precise artifact grounding within otherwise largely preserved scenes.

C.3 CASE VISUALIZATION

Figs. 8, 9, and 10 presents qualitative examples of different manipulation tasks in the ManiVid-38K.



(a) Source Video Resolution **(b) Frame Count vs. Duration** **(c) Caption Length Comparison**

(d) Manipulated Region Count **(e) Manipulated Region Area Ratio**

Figure 7: **Additional statistical analyses of ManiVid-38K.**



Figure 8: **Examples of *Addition* and *Removal* in ManiVid-38K.** Zoom in for a better view.

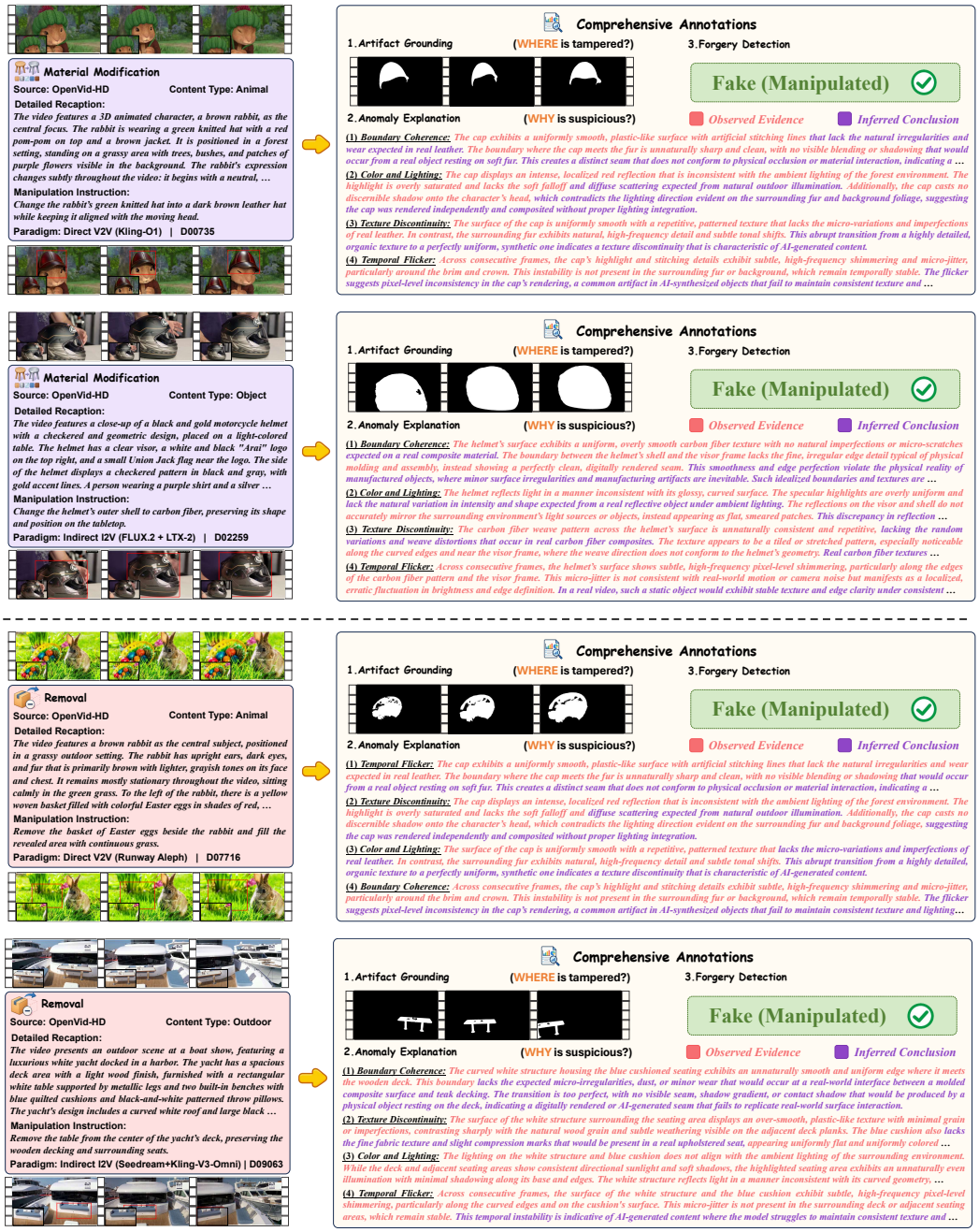


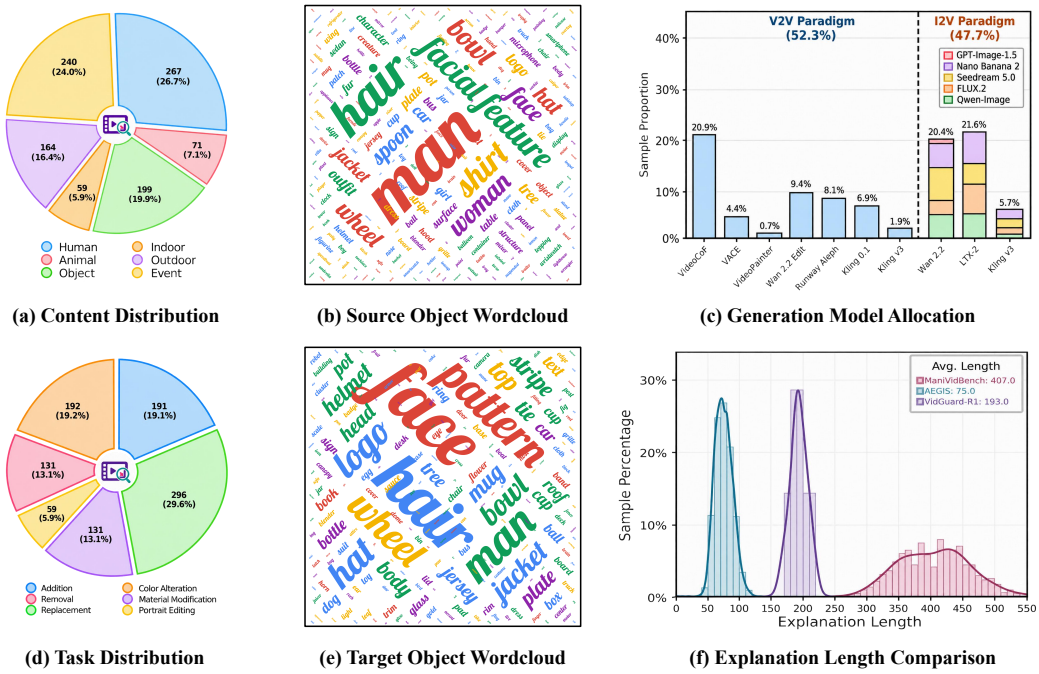
Figure 9: **Examples of *Replacement* and *Color Alteration* in ManiVid-38K.** Zoom in for a better view.



Figure 10: **Examples of *Material Modification* and *Portrait Editing* in ManiVid-38K.** Zoom in for a better view.

C.4 MANIVIDBENCH

We randomly sample 1K real–fake video pairs (2K videos) from ManiVid-38K dataset to construct ManiVidBench for evaluating ManiVid. Since source-level deduplication is performed before dataset splitting, each original source video contributes only one clip to the entire dataset, ensuring source-video-level disjointness between ManiVidBench and the training split. Detailed statistics are presented in Fig. 11. We believe that this dedicated benchmark for manipulated videos, equipped with comprehensive annotations, can provide a solid foundation for future research on manipulated video detection and analysis.



(a) Content Distribution (b) Source Object Wordcloud (c) Generation Model Allocation

(d) Task Distribution (e) Target Object Wordcloud (f) Explanation Length Comparison

Figure 11: **Statistical Analysis of Our ManiVidBench.**

D MANIVIDLENS FRAMEWORK

D.1 TASK-SPECIFIC INPUT AND OUTPUT PROTOCOL

ManiVidLens uses one shared multimodal large language model (MLLM) with separate protocols for manipulation classification, anomaly explanation, and artifact grounding through SAM2 (Ravi et al., 2025). Each protocol receives the same 5 temporally ordered conditioning frames and their frame-aligned forensic evidence. Under the full Forensic Evidence Router (FER) configuration, every conditioning frame contributes one Qwen3-VL visual block (Bai et al., 2025) followed by three [AIDE] positions. These positions contain the projected global token, semantic token, and aggregated forensic token from the shared evidence bank. Task-specific instructions isolate the three output spaces while preserving shared visual evidence and model parameters.

Throughout this appendix, bracketed uppercase names such as [SEG] denote literal special tokens, while unbracketed names denote tasks, modules, or mathematical indices.

For a manipulated sample, the segmentation target first reports `Target object:` followed by the annotated object name. It then places a concise grounding phrase inside `<p>...</p>` and appends [SEM][BBOX1][BBOX2] immediately before the closing tag. The grounding phrase provides compact localization context and remains separate from the longer anomaly explanation. Special token literals occur exclusively in assistant responses, preserving their role as generated control states.

Table 18: **Task-specific generation protocols used by ManiVidLens.** Each row summarizes one instruction objective and its exact assistant response structure.

Task protocol	Instruction objective and supervised assistant response format
Classification	Determine the manipulation status and one applicable manipulation type. The response contains `Manipulation status:` followed by `Manipulation type:`.
Explanation	Describe the relevant visual, physical, and temporal anomalies. The response contains exactly one `<think>...</think>` reasoning block.
Segmentation	Locate the manipulated region and describe its target briefly. A positive response contains the target fields and exactly one [SEG] marker.

The primary configuration trains and evaluates these three protocols separately on the shared model. The all-in-one ablation concatenates the same explanation, classification, and segmentation components in that fixed order for direct comparison with the separate task protocol.

Table 19: **Task Prompts Used in the ManiVidLens Framework.** *Separate prompts* are the default; the *All-in-One prompt* is used only for the 3-task ablation.

Prompts Used in ManiVidLens

Separate Prompts (Default):

Task 1: Artifact Grounding
Locate and segment the AI-manipulated region in this video. If a manipulated region is present, describe the target object and output exactly one segmentation marker for that region. If no manipulated region is present, state that the manipulated area is none and do **not** output a segmentation marker. Do **not** output classification fields or a forensic explanation.

Task 2: Anomaly Explanation
Analyze this video for forensic evidence of AI manipulation. Explain the relevant visual, physical, and temporal evidence inside exactly one reasoning block. Do **not** output classification fields, a target-object field, or a segmentation marker.

Task 3: Forgery Detection
Analyze this video for AI manipulation. Determine whether it is manipulated or original. If manipulated, identify the manipulation type as ***Addition***, ***Replacement***, ***Material Modification***, ***Removal***, ***Color Alteration***, or ***Portrait Editing***. If original, use manipulation type ***Real***.

All-in-One Prompt:

Analyze this video for AI manipulation. Follow these three steps in order. First, explain the relevant visual, physical, and temporal forensic evidence inside exactly one reasoning block. Second, classify the video as manipulated or original and provide the manipulation type; if it is original, use type ***Real***. Third, localize and segment the manipulated region, or state that the manipulated area is none for an original video. Do **not** omit, merge, or reorder these three steps.

D.2 Forensic Evidence Router

D.2.1 Forensic Evidence Bank

The frozen AIDE encoder (Yan et al., 2025) produces one 256-dimensional semantic token $\mathbf{s}_k$ and four 2,048-dimensional forensic tokens $\{\mathbf{r}_{k,j}\}_{j=1}^{4}$ for frame I_{t_k}. Separate linear adapters map the semantic and forensic tokens to $d_e = 256$. A two-layer projector with a 1,024-dimensional hidden layer constructs a global seed from $[\mathbf{s}_k; \bar{\mathbf{r}}_k]$, where $\bar{\mathbf{r}}_k$ is the mean forensic token. With $\mathbf{T}_{\text{bank}}$ denoting the learned slot embedding, the evidence bank is computed as

$$\begin{aligned} \mathbf{U}_k &= [\phi_s(\mathbf{s}_k), \{\phi_{r,j}(\mathbf{r}_{k,j})\}_{j=1}^{4}, \phi_g([\mathbf{s}_k; \bar{\mathbf{r}}_k])] + \mathbf{T}_{\text{bank}}, \\ \mathbf{U}'_k &= \mathbf{U}_k + \text{MHA}(\text{LN}(\mathbf{U}_k)), \\ \mathbf{E}_k &= \mathbf{U}'_k + \text{FFN}(\text{LN}(\mathbf{U}'_k)) \in \mathbb{R}^{6\times 256}. \end{aligned} \tag{4}$$

The resulting order is semantic, four forensic, and global. This fixed ordering is retained throughout the Forensic Evidence-to-VLM (FE2VLM) adapter, the auxiliary estimator, and the Forensic Evidence-to-SAM (FE2SAM) adapter across both training stages.

D.2.2 AUXILIARY AUTHENTICITY ESTIMATOR

The auxiliary estimator uses only the contextualized global and semantic tokens to predict whether a video is manipulated. For each frame, their normalized features are concatenated and passed through a two-layer MLP with a 256-dimensional hidden layer to obtain a frame logit $u_{i,k}$. Frame logits are aggregated into a video-level manipulation logit u_i by a normalized smooth maximum, and the corresponding routing score a_i is

$$\begin{aligned} u_i &= \tau \left[\log \sum_{k=1}^{K_i} \exp(u_{i,k}/\tau) - \log K_i \right], \\ a_i &= \rho + (1-\rho)\sigma(u_i), \qquad \tau = 0.5,\ \rho = 0.25. \end{aligned} \tag{5}$$

A binary cross-entropy loss supervises u_i, with manipulated videos as the positive class. Detached gates from this score modulate FE2VLM for every task and FE2SAM for grounding. Gradient detachment lets the classification objective determine the routing score directly. The routing floor ρ retains a nonzero forensic contribution throughout both optimization stages.

D.2.3 EVIDENCE INJECTION INTO THE MLLM

FE2VLM selects the global token, semantic token, and mean of the four forensic tokens from $\mathbf{E}_k$. After adding three learned type embeddings, layer normalization and a two-layer $256 \to 1024 \to d_{\mathrm{MLLM}}$ projector map them to the MLLM hidden space. Let $\mathbf{v}_{i,k}$ denote these projected tokens and $\mathbf{e}_{\mathrm{AIDE}}$ the learned embedding of the [AIDE] token. The injected evidence is softly routed as

$$\widetilde{\mathbf{v}}_{i,k} = \mathbf{e}_{\mathrm{AIDE}} + a_i(\mathbf{v}_{i,k} - \mathbf{e}_{\mathrm{AIDE}}). \tag{6}$$

The three routed tokens replace the three [AIDE] positions following the corresponding frame's visual block. The MLLM therefore receives routed forensic evidence directly inside its multimodal token sequence for every conditioning frame.

D.2.4 CONDITIONAL SAM2 FEATURE ENHANCEMENT

FE2SAM uses five evidence tokens per frame: the global token followed by all four forensic tokens. The semantic token is excluded because semantic object guidance is already supplied by the Prompt Distill Module (PDM) prompts and SAM2 image features. Learned type embeddings, layer normalization and a $256 \to 256 \to 256$ MLP map these tokens to the SAM2 space.

For each conditioning frame, let $\mathbf{F}_{i,k} \in \mathbb{R}^{d_s \times H_s \times W_s}$ denote the SAM2 image feature map and $\mathbf{A}_{i,k} \in \mathbb{R}^{5 \times d_s}$ the projected global and 4 forensic tokens, where $d_s = 256$. The $H_s W_s$ spatial positions of $\mathbf{F}_{i,k}$ serve as queries, while $\mathbf{A}_{i,k}$ provides keys and values for 4-head cross-attention. A convolutional source adapter further refines the attended features, and the detached manipulation score interpolates between the original SAM2 features and their evidence-enhanced counterparts:

$$\begin{aligned} \mathbf{H}_{i,k} &= \mathbf{F}_{i,k} + \psi(\mathrm{MHA}(\mathbf{F}_{i,k}, \mathbf{A}_{i,k}, \mathbf{A}_{i,k})), \\ \widetilde{\mathbf{F}}_{i,k} &= \mathbf{F}_{i,k} + a_i(\mathrm{Adapter}(\mathbf{H}_{i,k}) - \mathbf{F}_{i,k}). \end{aligned} \tag{7}$$

Here, ψ reshapes the attended tokens and applies a 1×1 convolution, layer normalization, GELU, and another 1×1 convolution, while $\mathrm{Adapter}$ denotes the subsequent residual convolutional refinement block. Positional encodings and a learned query bias are added to the SAM2 queries before attention. During Stage 2 training, FE2SAM processes the 5 sampled frames of manipulated segmentation instances carrying mask supervision. During validation and inference, FE2SAM enhances the 5 conditioning frames whenever generated segmentation states activate mask decoding; SAM2 then propagates the initialized object state across all remaining physical video frames.

D.3 PROMPT DISTILL MODULE

D.3.1 SEMANTIC DENSE PROMPT FUSION

PDM reads the final-layer MLLM state associated with each grounding token. An MLP with a ReLU activation maps the [SEM] state from d_{MLLM} into the 256-dimensional SAM2 prompt space. The projected semantic state is spatially broadcast and added to the frozen SAM2 no-mask embedding.

A lightweight fusion block applies a 1×1 convolution, two-dimensional layer normalization, GELU, and another 1×1 convolution to produce the dense semantic prompt:

$$\mathbf{D}_i = \mathrm{ConvFuse}\big(\mathbf{E}_{\text{no-mask}} + \mathrm{Tile}(\phi_{\text{sem}}(\mathbf{h}_i^{\text{SEM}}))\big)\,. \tag{8}$$

The resulting dense prompt supplies video-level semantic context to every grounding frame.

D.3.2 Geometric Sparse Prompt Distillation

Independent 2-layer MLPs with ReLU activations project the [BBOX1], [BBOX2], and [SEG] states from d_{MLLM} through a hidden layer of dimension d_{MLLM} into the 256-dimensional SAM2 prompt space. Each sampled-frame annotation stores an exclusive-coordinate box (x_0, y_0, x_1, y_1) at the original frame resolution. We resize each valid box to the 1024×1024 SAM2 input space and convert its bottom-right boundary into inclusive pixel coordinates for consistent geometric supervision. The frozen SAM2 prompt encoder then maps every resized box into top-left and bottom-right corner embeddings. For sample i, $\mathcal{F}_i$ indexes the conditioning frames with valid boxes, and $\mathbf{B}_{i,k}$ denotes the corresponding resized box after this coordinate conversion. These corner embeddings are averaged separately across valid frames to construct $\bar{\mathbf{b}}_{i,1}$ and $\bar{\mathbf{b}}_{i,2}$:

$$\bar{\mathbf{b}}_{i,j} = \frac{1}{|\mathcal{F}_i|}\sum_{k\in\mathcal{F}_i} \mathrm{BoxEnc}(\mathbf{B}_{i,k})_j, \qquad j\in\{1,2\}. \tag{9}$$

Samples with a complete latent span and at least one valid box contribute to Eq. 2. The projected [BBOX1] and [BBOX2] states are concatenated with the projected [SEG] state in the order [BBOX1] $\rightarrow$ [BBOX2] $\rightarrow$ [SEG]. Repeating this sparse prompt across all grounding frames preserves consistent geometric guidance from the shared video-level prior during mask decoding.

D.4 Training and Inference Pipelines

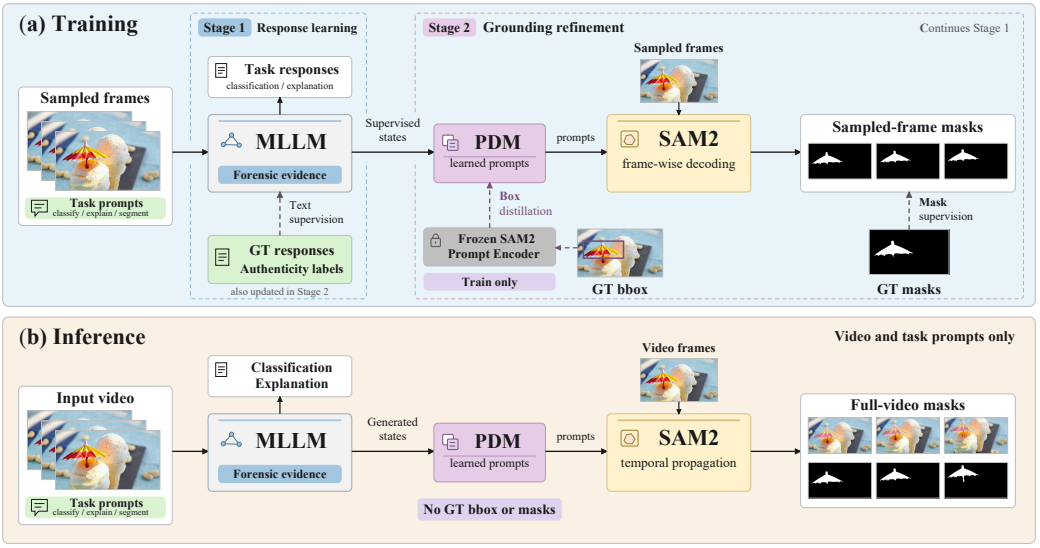


Figure 12: **Training and Inference Pipelines of ManiVidLens.** (a) Stage 1 learns task-specific responses and authenticity routing. Stage 2 adds box-prompt distillation and sampled-frame mask supervision, with PDM reading teacher-forced token states. (b) Inference constructs SAM2 prompts from generated token states and propagates masks across the full video using only video and task inputs. GT denotes ground-truth responses, boxes, and masks used for supervision.

D.4.1 TRAINING PIPELINE

Table 20: **Two-stage training implementation for the current full profile.** The table reports the launcher defaults and stage-specific optimization settings.

Training aspect	Stage 1 implementation details	Stage 2 implementation details
Stage length	One complete training epoch	One complete training epoch
Task setup	Three separate task instances with $1/3$ shared-loss scaling	Three separate task instances with $1/3$ shared-loss scaling
Batch ratio	Balanced original-to-manipulated ratio of 1:1	Manipulation-focused original-to-manipulated ratio of 1:3
Shared weights	Language and classification weights are $\lambda_{\text{lm}} = 1.0$ and $\lambda_{\text{cls}} = 1.0$	Language and classification weights are $\lambda_{\text{lm}} = 0.1$ and $\lambda_{\text{cls}} = 0.5$
Grounding weights	Every grounding-loss weight remains zero	$\lambda_m = 2.0$, $\lambda_d = 1.0$, and $\lambda_c = 1.0$
Class weight	The classifier uses unit positive-class weight	The classifier uses one-third positive-class weight
Trainable set	PDM, FE2SAM, the [SEG] projector, and the SAM2 mask decoder remain frozen	PDM, FE2SAM, the [SEG] projector, and the SAM2 mask decoder are trainable

The Qwen3-VL visual encoder and base language-model weights remain frozen, while rank-128 low-rank adaptation (LoRA) modules update the language layers (Hu et al., 2021). The token embeddings and language-model head also remain trainable throughout both stages. Both stages train the AIDE evidence bank, FE2VLM projector, and auxiliary authenticity estimator.

Stage 1 learns the three response protocols and video-level authenticity routing before mask optimization begins (Figure 12(a)). Its frozen grounding components comprise the SAM2 mask decoder, the [SEG] projector, the FE2SAM projector and enhancer, and every PDM projection and fusion block. Stage 2 loads the Stage 1 checkpoint and activates these grounding components. Within SAM2, only the mask decoder is trainable during Stage 2; the image encoder, prompt encoder, memory encoder, and memory attention remain frozen.

Each sampled video view expands into explanation, classification, and segmentation instances under the separate protocol. The three instances share identical frame indices because their sampling seed derives from the same base view. Language and auxiliary classification losses receive a factor of $1/3$ on every task instance. This normalization preserves the aggregate shared-loss weight assigned to each base video view. Grounding losses receive unit task scaling only for Stage 2 segmentation instances. Mask availability further restricts these losses to manipulated samples carrying valid spatial annotations for the sampled video view. Teacher-forced assistant targets supervise every textual response and provide the special token states used by PDM. Box annotations construct the cosine-distillation targets, while mask annotations supervise the mask and Dice objectives. During Stage 2 training, mask decoding operates independently on the five sampled frames for manipulated segmentation instances without temporal video-memory propagation.

The grouped sampler constructs device-local batches that are homogeneous in concrete task and video type. For manipulated samples, let r_i denote the largest visible mask-area ratio among the sampled frames. The per-sample auxiliary classification weight follows

$$w_i = \begin{cases} 1 + (0.05 - r_i)/0.05, & \text{if } 0 < r_i < 0.05, \\ 1, & \text{for every remaining training sample.} \end{cases} \tag{10}$$

This weighting assigns additional classification emphasis to small visible manipulated regions.

D.4.2 INFERENCE PIPELINE

Validation and inference perform separate deterministic generations for explanation, classification, and segmentation (Figure 12(b)). Every generation receives the same ordered conditioning frames and their corresponding routed AIDE evidence. Forgery-detection metrics are computed from the generated classification response, while the auxiliary estimator is logged separately and supplies the routing gate. Stage 1 validation omits mask decoding altogether. During Stage 2 validation and inference, a segmentation response containing [SEG] activates PDM and SAM2. PDM reads the aligned, autoregressively generated [SEM], [BBOX1], [BBOX2], and [SEG] states. The projected [SEM] state forms the dense semantic prompt through Eq. 8, while the projected [BBOX1],

[BBOX2], and [SEG] states form the sparse prompt embeddings for every conditioning frame. Validation annotations remain outside the model inputs and support metric computation exclusively. Stage 2 validation permits at most one additional generation of segmentation responses per sample.

The 5 conditioning frames follow the policy in Appendix E.1.2, while the evaluation sequence contains every physical frame of the complete video. FE2SAM injects frame-aligned forensic evidence into the SAM2 image features of the 5 conditioning frames. The PDM prompt embeddings initialize the object state on the same conditioning frames before temporal propagation begins. With the earliest conditioning frame at local index zero, SAM2 propagates the initialized object state once in the forward temporal direction through the complete physical-frame sequence. The implementation resizes propagated mask logits bilinearly to each original frame resolution and thresholds sigmoid probabilities at 0.5. If the response omits [SEG], the pipeline skips mask decoding.

E EXPERIMENTS

E.1 IMPLEMENTATION DETAILS

E.1.1 MODEL SETUP & DATASET

Model Setup. ManiVidLens combines Qwen3-VL-4B-Instruct (Bai et al., 2025) for multimodal reasoning, a frozen AIDE encoder (Yan et al., 2025) for forensic cues, and SAM2 Hiera-L (Ravi et al., 2025) for video segmentation. All three components are initialized from their respective released pretrained checkpoints before any task-specific training is performed.

Dataset. Training uses the ManiVid-38K dataset training split of 18K real–fake video pairs spanning six content categories and six manipulation tasks. Each pair provides a video-level authenticity label, frame-level pixel masks, and a textual anomaly explanation. ManiVidBench contains 2K videos sampled from ManiVid-38K dataset: 1K manipulated videos and their authentic videos across the same six manipulation categories, with task-specific evaluation scopes defined below.

E.1.2 TRAINING AND INFERENCE SETTINGS

Training. In the reported full configuration, Stage 1 learns the three task protocols for one epoch with grounding frozen; Stage 2 loads that checkpoint and trains grounding for one epoch with lower language and classification weights. Each video yields five views per epoch, and every view expands into three task instances sharing five ordered frames sampled with random temporal patterns and a stride from $\{4, 6, 8, 10\}$. Invalid windows fall back to five randomly selected frames from the complete video, sorted temporally. Stage 1 samples original and manipulated videos at $1{:}1$, whereas Stage 2 uses $1{:}3$. Both stages use AdamW with a base learning rate of 10^{-4}, weight decay of 0.05, a 5% linear warmup, and cosine decay. With bfloat16 precision on eight NVIDIA RTX PRO 5000 GPUs, a per-device microbatch of two and four-step accumulation yield an effective batch size of 64; sequences are capped at $8{,}192$ tokens, and LoRA uses rank 128, scaling 256, and dropout 0.05 (Hu et al., 2021). Appendix D.4.1 specifies the stage-wise loss weights and trainable modules. For a consistent budget, trainable image and video expert comparisons receive two epochs on the ManiVid-38K dataset training split for each supported task; training-free comparisons and general MLLMs use released checkpoints without fine-tuning in the reported evaluation protocol.

Inference. ManiVidLens generates each task separately and deterministically; only grounding activates SAM2 through [SEG]. Forgery detection is evaluated on all 2,000 ManiVidBench videos, whereas artifact grounding and anomaly explanation are evaluated on the 1,000 manipulated videos. For detection, image experts majority-vote 0.5-threshold predictions over thirty uniformly sampled frames; other methods use native video-level outputs. For explanation, one response is evaluated per manipulated video; ManiVidLens and general MLLMs use five conditioning frames. These are the first five decodable candidates from a uniform thirty-frame grid. ManiVidLens reuses these frames across all three task generations. For grounding, image experts predict and are scored on the thirty sampled frames. General MLLMs and Omni-Fake-R1 output frame-indexed boxes, which prompt shared SAM2 Hiera-L propagation and evaluation over every physical frame. Video segmentation models directly process the complete video, while ManiVidLens initializes full-video mask propagation from its five conditioning frames; both are evaluated on every physical frame of each test video under this complete-video evaluation protocol.

E.1.3 Evaluation Metrics

Artifact Grounding. Mean IoU (mIoU) computes binary-mask intersection over union for every evaluated frame, averages those scores within each video, and then macro-averages across videos. This region-overlap metric directly measures how completely each predicted mask covers the reference manipulation area. J&F averages region similarity (J), which equals the frame IoU, and contour accuracy (F) with a boundary tolerance of 0.008 times the image diagonal under the same two-level aggregation (Yuan et al., 2025). This combined score complements regional overlap with boundary fidelity, which is essential for precise pixel-level artifact localization. Jointly empty mask pairs score one, whereas pairs containing exactly one empty mask score zero.

Anomaly Explanation. ROUGE-L is the token-level longest-common-subsequence F1 between the first explicitly closed, nonempty `<think>` blocks after lowercase alphanumeric tokenization without stemming (Lin, 2004). It measures whether a generated explanation recovers the reference evidence and its ordered lexical content. Cosine Similarity Score (CSS) computes cosine similarity between attention-mask mean-pooled paraphrase-MiniLM-L6-v2 embeddings that are subsequently L2-normalized (Reimers & Gurevych, 2020; 2019; Xu et al., 2025). It captures semantic agreement when a valid explanation paraphrases the reference with different wording. Both scores are computed per video and then averaged across the manipulated subset; a missing or invalid generated block scores zero, while a missing reference is excluded.

Forgery Detection. Accuracy is the proportion of correct video-level authenticity predictions among all 2,000 test videos. The balanced set of 1,000 authentic and 1,000 manipulated videos makes it a direct measure of overall correctness. F1 is the harmonic mean of precision and recall with the manipulated class treated as positive. It complements Accuracy by exposing the trade-off between missed manipulations and false alarms. Any unparseable generated authenticity label counts as an incorrect prediction during metric computation.

Across the main and ablation tables, **bold** and underlined values indicate the best and second-best results within each table column, respectively.

E.2 Baseline Model Details

CNNSpot (Wang et al., 2020) CNNSpot is a convolutional image detector that uses a ResNet-50 classifier to distinguish real images from synthetic images. Training on one source generator with blur and JPEG augmentation helps the detector generalize to unseen generators.

UnivFD (Ojha et al., 2023) UnivFD detects synthetic images using the feature space of a frozen pretrained vision-language model. A nearest-neighbor rule or linear classifier predicts authenticity from these representations, which are learned from diverse image-text pairs.

AIDE (Yan et al., 2025) AIDE combines high-level semantic representations with low-level forensic features for AI-generated image detection. Its local branch samples high- and low-frequency patches to capture noise patterns and other synthesis artifacts.

LEGION (Kang et al., 2025) LEGION is an MLLM-based framework for synthetic-image detection, artifact segmentation, and textual explanation. It pairs pixel-level artifact masks with natural-language descriptions of the visual inconsistencies in the identified regions.

FakeShield (Xu et al., 2025) FakeShield is an explainable image-forgery framework that predicts authenticity, describes tampering evidence, and localizes manipulated regions. Its domain-tag-guided detection module handles different forgery types, while a multimodal localization module uses the generated descriptions to guide mask prediction.

SIDA (Huang et al., 2025a) SIDA is a large multimodal model for detecting, localizing, and explaining deepfakes in social-media images. It distinguishes authentic, fully synthetic, and tampered images, producing manipulation masks and textual explanations for its decisions.

D3 (Zheng et al., 2025) D3 is a training-free video-detection method that identifies temporal discrepancies through second-order features. Statistics of changes in inter-frame feature relationships, extracted with a pretrained visual encoder, distinguish real from generated videos.

VidGuard-R1 (Park et al., 2026) VidGuard-R1 adapts a multimodal language model for video-authenticity detection and explanatory reasoning through group relative policy optimization. Its reward design accounts for temporal artifacts and generation complexity, encouraging the model to produce authenticity judgments with supporting natural-language rationales.

ReStraV (Internò et al., 2025) ReStraV detects generated videos by examining the geometry of frame trajectories in a pretrained visual representation space. It summarizes temporal curvature and stepwise distances between frame embeddings, then uses a lightweight classifier to identify deviations from the trajectory statistics of natural videos.

NSG-VD (Zhang et al., 2025a) NSG-VD uses normalized spatiotemporal gradients to characterize deviations from natural video dynamics under a probability-flow conservation formulation. It estimates these features with pretrained diffusion models and measures their maximum mean discrepancy from real-video reference features to obtain a detection score.

DeMamba (Chen et al., 2026) DeMamba introduces a Detail Mamba module for detecting AI-generated videos through spatial and temporal inconsistencies. The module augments video representations with Mamba-based detail modeling, providing a plug-in component for authenticity classification across different generators and video degradations.

VideoLISA (Bai et al., 2024b) VideoLISA combines a video-language model with a segmentation model to generate video masks from language instructions that require contextual reasoning. Its sparse-dense frame sampling balances temporal coverage with spatial detail, while a shared tracking token supports segmentation of the target across multiple frames.

VISA (Yan et al., 2024) VISA is a video segmentation assistant that resolves implicit object queries using world knowledge and video context. A mask decoder segments and tracks the inferred target across video frames.

Sa2VA (Yuan et al., 2025) Sa2VA integrates multimodal language understanding with SAM2 for dense grounding in images and videos. The language model generates instruction tokens that guide mask prediction and video propagation, supporting both grounded conversation and referring segmentation within a shared framework.

Omni-Fake-R1 (Li et al., 2026) Omni-Fake-R1 is a reinforcement-learning-based detector for deepfakes across image, audio, video, and audiovisual inputs. It combines visual and auditory evidence when available and generates structured authenticity decisions, localization outputs, and natural-language explanations; its visual localization is expressed as bounding boxes.

InternVL3.5-30B-A3B (Wang et al., 2025) InternVL3.5-30B-A3B is an open multimodal model that connects an InternViT visual encoder to a mixture-of-experts language model through a multilayer perceptron. It retains dynamic image-resolution processing and supports image and video understanding, with approximately three billion active language-model parameters per token.

Qwen3.5-35B-A3B (Qwen Team, 2026) Qwen3.5-35B-A3B is a multimodal model with a vision encoder and a sparse mixture-of-experts language backbone containing 35 billion parameters, of which three billion are active per token. Its hybrid architecture combines gated delta networks with attention layers, supporting visual understanding and language reasoning with sparse computation.

GPT-5.6-Sol (OpenAI, 2026) GPT-5.6-Sol is a proprietary general-purpose model that accepts text and image inputs and produces textual responses for reasoning and visual-understanding tasks. In our evaluation, it analyzes sampled video frames to provide authenticity judgments, anomaly explanations, and frame-indexed boxes without task-specific fine-tuning.

E.3 More Performance Analysis

E.3.1 By Manipulation Task

The category breakdown in Tables 2–4 compares forgery detection, artifact grounding, and anomaly explanation across the six manipulation types. Following the metrics and aggregation in Appendix E.1.3, we examine how authenticity decisions, mask recovery, and explanation quality vary for the same manipulation type, extending the overall results in Sec. 4.2.

Addition and Replacement. On Addition, ManiVidLens reaches 0.444 mIoU and 0.467 J&F, above Sa2VA's 0.414 and 0.432, together with 0.966 detection accuracy and 0.813 CSS. On Replacement, it reaches 0.421 mIoU and 0.439 J&F, improving over LEGION's 0.385 and 0.392, while detection accuracy is 0.868 vs. DeMamba's 0.890. These results separate precise recovery of the altered object's spatial extent from video-level authenticity discrimination, showing why the benchmark evaluates spatial localization and authenticity decisions as distinct, complementary outputs.

Color and Material. These edits preserve object identity while changing visual attributes, placing emphasis on the altered appearance within an otherwise recognizable object. ManiVidLens reaches 0.487 mIoU on Color and 0.526 on Material, exceeding FakeShield's 0.429 and Sa2VA's 0.399, respectively. Its J&F scores of 0.507 and 0.532 accompany CSS scores of 0.772 and 0.795 and detection accuracies of 0.917 and 0.939. Table 22 connects this pattern to the architectural design: combining FER and PDM improves both categories over either module alone, supporting complementary roles for forensic evidence and structured prompts when grounding visual attributes.

Portrait Editing. ManiVidLens leads grounding at 0.290 mIoU and 0.306 J&F, as well as detection accuracy at 0.839 vs. DeMamba's 0.831. Its absolute grounding scores remain lower than those for Addition, Color, and Material. Explanation exhibits a different ranking: ManiVidLens leads ROUGE-L at 0.545 vs. SIDA's 0.251, but its CSS of 0.699 trails SIDA's 0.746 and LEGION's 0.718. The differing rankings separate lexical overlap from semantic agreement with reference explanations, motivating complementary evaluation of the words used to describe an anomaly and the meaning conveyed by the generated forensic description.

Removal. ManiVidLens reaches 0.943 detection accuracy and 0.773 CSS, alongside 0.135 mIoU vs. Qwen3.5's 0.221. The reference mask describes the removed object's original extent, whereas the observed region contains reconstructed background with little object semantics. Detecting traces of background reconstruction and recovering the removed object's original extent are distinct requirements for this task. The detection–grounding gap motivates localization cues that describe the reconstructed background and its spatial boundaries, complementing the object semantics used to localize additions or replacements where target objects remain visible.

E.3.2 By Generation Model

Table 21 extends the main results with generation-model breakdowns and the overall scores on all 1,000 ManiVidBench pairs. Indirect I2V groups follow the final video generator, pooling over first-frame editors, with task metrics and aggregation defined in Appendix E.1.3.

Generation Paradigms. Detection accuracy is lower for Direct V2V than Indirect I2V (0.868 vs. 0.975), accompanied by lower ROUGE-L (0.567 vs. 0.605) and CSS (0.750 vs. 0.821). Grounding favors Direct V2V, with 0.413 mIoU and 0.429 J&F vs. 0.398 and 0.413. This reversal shows that the stronger authenticity and explanation results on Indirect I2V videos coexist with more difficult spatial recovery, making localization a continuing priority across both generation paradigms.

Individual Generation Models. Within Direct V2V, VideoCoF yields the lowest mIoU (0.317), whereas Kling-O1 is hardest to detect (0.751 accuracy) despite stronger grounding (0.435 mIoU). Wan 2.2-Video-Edit reaches 0.553 mIoU and 0.976 accuracy, showing that the difficulty of locating and recognizing manipulations varies across video editing models. Within Indirect I2V, LTX-2 leads grounding at 0.440 mIoU, compared with Wan 2.2's 0.383 and Kling-V3-Omni's 0.292. Detection accuracy remains between 0.967 and 0.982 across these three I2V models, whereas their mIoU spans 0.148. Their similar detection performance and wider grounding variation locate the main difference in recovering precise manipulated regions across generation models.

E.3.3 From Classification to Attribution

Given the strong performance of ManiVidLens on binary forgery detection (0.914 Acc and 0.913 F1), we further extend the evaluation to the more challenging **manipulation attribution task**. Instead of simply distinguishing between real and fake videos, this task further requires identifying the specific manipulation type for fake videos. Our classification prompt already asks the model to predict the most likely manipulation type in addition to the authenticity label. We therefore directly evaluate these predictions under a 7-class setting, comprising Real and the six manipulation types in ManiVid. On ManiVidBench, ManiVidLens achieves an overall 7-class accuracy of

Table 21: **ManiVidLens Performance by Generation Model.** (·) indicate each generation paradigm's share of ManiVidBench. Indirect I2V groups follow the final video generator, pooling over first-frame editors. Avg. aggregates each paradigm, while Overall matches the main tables under the evaluation protocol in Appendix E.1.3.

Task	Metric	Direct V2V (52.3%)								Indirect I2V (47.7%)				Overall
		Kling-O1	Kling-V3 Omni	Runway Aleph	Wan 2.2 Video-Edit	VideoCoF	VACE	VideoPainter	Avg.	Kling-V3 Omni	Wan 2.2	LTX-2	Avg.	
Grounding	mIoU	0.435	0.419	0.452	0.553	0.317	0.396	0.465	0.413	0.292	0.383	0.440	0.398	0.407
	J&F	0.450	0.436	0.466	0.574	0.333	0.414	0.469	0.429	0.303	0.402	0.452	0.413	0.422
Explanation	R-L	0.533	0.557	0.605	0.621	0.553	0.596	0.607	0.567	0.615	0.608	0.600	0.605	0.583
	CSS	0.686	0.740	0.828	0.855	0.722	0.796	0.789	0.750	0.835	0.826	0.811	0.821	0.780
Detection	Acc	0.751	0.941	0.972	0.976	0.850	0.974	1.000	0.868	0.980	0.967	0.982	0.975	0.914
	F1	0.751	0.941	0.972	0.976	0.849	0.974	1.000	0.868	0.980	0.967	0.982	0.975	0.913

0.674 and a Macro-F1 of 0.497. The corresponding confusion matrix is shown in Fig. 13. A closer inspection of the confusion matrix reveals a clear distinction between authenticity recognition and fine-grained manipulation attribution. ManiVidLens correctly classifies 920 out of 1,000 real videos, corresponding to a 92.0% recall for the Real class. In contrast, the errors among manipulated videos are predominantly concentrated across different manipulation types rather than between real and fake, indicating that distinguishing the specific editing operation remains substantially more challenging than determining authenticity. Among the manipulation categories, Addition achieves the highest class-wise recall (100/191, 52.4%), followed by Color Alteration (86/192, 44.8%) and Material Modification (53/131, 40.5%). A notable pattern is that Replacement acts as the dominant source of cross-category confusion. For example, 51 Addition, 35 Removal, 41 Color Alteration, 36 Material Modification, and 13 Portrait Editing samples are predicted as Replacement. Conversely, true Replacement samples are also distributed across several related categories, most prominently Color Alteration (49 cases), Real (40 cases), and Addition (35 cases). This confusion pattern suggests that the model can often identify the presence of localized manipulation evidence while remaining uncertain about the exact operation that produced it. In particular, manipulation types that modify an existing object, such as Replacement, Color Alteration, and Material Modification, may share similar spatial and visual forensic cues, making fine-grained attribution inherently more ambiguous. These results further motivate manipulation attribution as a more stringent diagnostic setting beyond binary forgery detection.

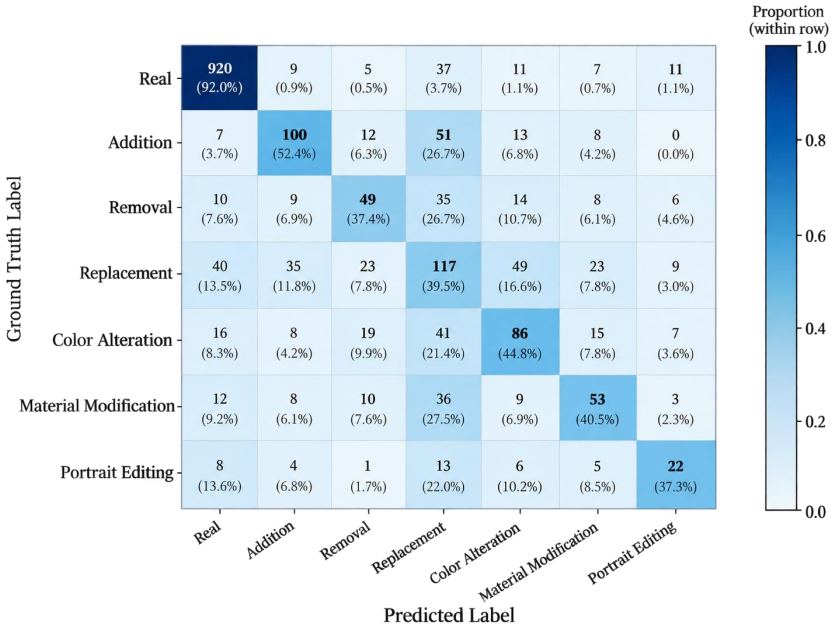


Figure 13: **Confusion Matrix for Seven-Class Manipulation Attribution.**

E.3.4 HUMAN EVALUATION FOR ANOMALY EXPLANATION

Although ROUGE-L and Cosine Similarity Score (CSS) provide scalable and reproducible evaluation for the anomaly explanation task, both metrics have inherent limitations. In particular, ROUGE-L relies heavily on lexical and structural overlap with the reference explanation and may therefore favor responses that follow similar wording or structured templates, even when semantically equivalent explanations are expressed differently. CSS alleviates this issue by measuring semantic similarity in the embedding space, but it still cannot directly determine whether the described forensic evidence is actually supported by the visual content. We therefore conduct an additional human evaluation to complement the automatic metrics.

Due to the substantial cost of manually inspecting video content and reviewing long-form forensic explanations, it is impractical to conduct human evaluation on all manipulated samples in ManiVid-Bench. We instead randomly sample 50 manipulated videos from each of the six manipulation categories, resulting in 300 videos in total. We evaluate ManiVidLens together with three representative comparison methods: the image-based expert LEGION, the video-based expert Omni-Fake-R1, and the general MLLM GPT-5.6 Sol. This yields 1,200 model-generated explanations for manual inspection. Four authors independently review the corresponding videos, manipulated regions, and generated explanations, while the model identities are hidden during evaluation to reduce potential bias. Each explanation is scored from 1 to 5 along the following three dimensions:

- **Evidence Faithfulness.** Whether the described visual or temporal anomalies are actually observable in the video and localized manipulation region, without hallucinating unsupported artifacts. A score of 1 indicates that the explanation is largely inconsistent with the observable evidence, while a score of 5 indicates that nearly all described evidence is visually well supported.
- **Forensic Relevance.** Whether the identified evidence is diagnostically relevant to video manipulation, rather than consisting of generic scene descriptions or weakly related observations. A score of 1 denotes largely irrelevant or non-forensic descriptions, whereas a score of 5 indicates that the explanation consistently identifies meaningful manipulation-sensitive cues, such as boundary, texture, lighting, or temporal inconsistencies.
- **Explanation Quality.** Whether the explanation provides a clear, specific, and coherent connection between the observed anomaly and the manipulation hypothesis. A score of 1 corresponds to vague, repetitive, or poorly justified reasoning, while a score of 5 represents precise and logically well-supported forensic reasoning with minimal redundancy.

The four authors achieve an intraclass correlation coefficient (ICC) of 0.86, indicating substantial agreement between them. For each explanation, we first average the scores from the four evaluators over the three evaluation dimensions to obtain a single human evaluation score. Specifically, for video i and model m, the human score is computed as

$$H_{i,m} = \frac{1}{4 \times 3} \sum_{e=1}^{4} \sum_{d=1}^{3} s_{i,m}^{(e,d)}, \tag{11}$$

where $s_{i,m}^{(e,d)}$ denotes the score assigned by evaluator e on evaluation dimension d. This produces one human score for each model-generated explanation.

Since our primary interest is whether an automatic metric can correctly reflect the *relative quality of different models on the same video*, rather than merely correlating with their aggregated scores, we perform correlation analysis at the individual-video level. For each video i, we rank the four evaluated models according to their human evaluation scores, yielding a human ranking vector $\mathbf{r}_i^H$. We independently rank the same four models according to their CSS or ROUGE-L scores, denoted as $\mathbf{r}_i^Q$, where $Q \in \{\mathrm{CSS}, \mathrm{R\text{-}L}\}$. We then compute Spearman's rank correlation coefficient between the two rankings:

$$\rho_i^Q = \mathrm{Spearman}\left(\mathbf{r}_i^H, \mathbf{r}_i^Q\right). \tag{12}$$

When no ties occur, this is equivalent to

$$\rho_i^Q = 1 - \frac{6 \sum_{m=1}^{4} \left(r_{i,m}^H - r_{i,m}^Q\right)^2}{4(4^2 - 1)}. \tag{13}$$

For tied scores, average ranks are assigned before computing the rank correlation.

For each manipulation category c, we average the per-video correlations over its 50 sampled videos:

$$\bar{\rho}_c^Q = \frac{1}{50} \sum_{i \in c} \rho_i^Q. \tag{14}$$

Finally, we report the mean correlation across all six manipulation categories as

$$\bar{\rho}^Q = \frac{1}{6} \sum_{c=1}^{6} \bar{\rho}_c^Q. \tag{15}$$

CSS exhibits consistently strong agreement with human judgments across all six manipulation categories. The category-wise mean Spearman correlations for Addition, Removal, Replacement, Color Alteration, Material Modification, and Portrait Editing are 0.804, 0.768, 0.812, 0.784, 0.796, and 0.756, respectively, resulting in an overall average correlation of $\bar{\rho} = 0.787$. In contrast, the corresponding correlations for ROUGE-L are 0.472, 0.436, 0.488, 0.412, 0.448, and 0.396, with a substantially lower average of $\bar{\rho} = 0.442$. These results indicate that CSS more consistently preserves the relative quality ordering perceived by human evaluators at the individual-video level, whereas ROUGE-L shows only moderate agreement and is more sensitive to lexical overlap and template-level similarity with the reference explanations. Therefore, although neither automatic metric can replace direct human inspection, the substantially stronger agreement between CSS and human judgments supports the use of CSS in Table 3 as a meaningful indicator of the relative anomaly explanation capability of different models.

E.4 Ablation Study

The ablations examine how frame count, architectural components, and task-specific prompts affect performance. Tables 22, 23, and 24 expand the aggregate results in Tables 5 and 6. Each table compares 3, 5, 7, and 9 sampled frames, followed by Base, + PDM, + FER, and + Both. Base omits PDM and FER, and + Both denotes the full ManiVidLens architecture. The default model uses five frames. Avg. follows the benchmark-level aggregation in Appendix E.1.3.

E.4.1 Artifact Grounding

Table 22: **Category-wise Ablation for Artifact Grounding.** Column groups follow the six manipulation categories, with benchmark-level scores in Avg. Metrics are mIoU and J&F (↑).

Setting	ManiVidBench												Avg.	
	Addition		Removal		Replacement		Color		Material		Portrait			
	mIoU	J&F	mIoU	J&F	mIoU	J&F	mIoU	J&F	mIoU	J&F	mIoU	J&F	mIoU	J&F
Ablation 1: Number of Sampled Frames														
3 Frames	0.411	0.432	0.118	0.118	0.366	0.382	0.395	0.417	0.450	0.456	0.278	0.286	0.354	0.368
5 Frames	0.444	0.467	0.135	0.137	0.421	0.439	**0.487**	**0.507**	**0.526**	**0.532**	0.290	0.306	0.407	0.422
7 Frames	**0.508**	**0.533**	0.114	0.120	**0.440**	**0.460**	0.452	0.474	0.483	0.494	**0.353**	**0.370**	**0.413**	**0.431**
9 Frames	0.454	0.471	**0.145**	**0.151**	0.402	0.421	0.469	0.492	0.498	0.505	0.294	0.297	0.397	0.413
Ablation 2: Architectural Components														
Base	0.414	0.432	0.132	0.135	0.343	0.356	0.390	0.406	0.399	0.406	0.182	0.186	0.336	0.348
+ PDM	**0.450**	0.466	**0.158**	**0.155**	0.393	0.406	0.427	0.446	0.401	0.416	0.170	0.178	0.368	0.380
+ FER	0.433	0.447	0.112	0.116	**0.434**	**0.449**	0.422	0.441	0.449	0.449	0.202	0.208	0.378	0.389
+ Both (Ours)	0.444	**0.467**	0.135	0.137	0.421	0.439	**0.487**	**0.507**	**0.526**	**0.532**	**0.290**	**0.306**	**0.407**	**0.422**

Frame Count. Increasing from three to five frames improves mIoU in all six categories (Table 22), with the largest gains in Color (+0.092), Material (+0.076), and Replacement (+0.055). Color and Material attain their highest grounding scores at five frames. Seven frames improve Addition mIoU from 0.444 to 0.508 and Portrait from 0.290 to 0.353, while Color and Material decrease by 0.035 and 0.043. These opposing category-level changes produce the small aggregate increase from 0.407 to 0.413 mIoU. Nine frames give the highest Removal mIoU of 0.145, but overall grounding falls to 0.397 mIoU and 0.413 J&F as inference time rises to 19.20 s (Table 5).

Architectural Components. The full model improves mIoU over Base in every category. The largest mIoU gains occur in Material (**from 0.399 to 0.526**), Portrait (from 0.182 to 0.290), and Color (from 0.390 to 0.487). Adding PDM to FER improves these categories by a further 0.077, 0.088, and 0.065, respectively. The best individual configurations differ across categories: PDM alone achieves the highest Addition and Removal mIoU of 0.450 and 0.158, while FER alone leads Replacement at 0.434. The full model's Removal gain over Base is 0.003. The joint architecture is strongest on attribute and portrait edits, while Removal remains its lowest-scoring category.

E.4.2 Anomaly Explanation

Table 23: **Category-wise Ablation for Anomaly Explanation.** Column groups follow the six manipulation categories in the main results, with benchmark-level scores in Avg. Metrics are ROUGE-L (R-L) and Cosine Similarity Score (CSS) (↑).

Setting	ManiVidBench												Avg.	
	Addition		Removal		Replacement		Color		Material		Portrait			
	R-L	CSS	R-L	CSS	R-L	CSS	R-L	CSS	R-L	CSS	R-L	CSS	R-L	CSS
Ablation 1: Number of Sampled Frames														
3 Frames	0.587	0.791	0.574	0.761	0.571	0.756	0.575	0.756	0.578	0.781	0.544	0.695	0.575	0.763
5 Frames	0.603	0.813	0.576	0.773	0.581	0.777	0.582	0.772	0.587	0.795	0.545	0.699	0.583	0.780
7 Frames	0.613	0.831	0.591	**0.796**	**0.593**	**0.797**	**0.592**	**0.801**	**0.603**	**0.817**	**0.555**	**0.721**	**0.596**	**0.803**
9 Frames	**0.618**	**0.836**	**0.592**	**0.796**	0.587	0.790	0.589	0.794	0.600	0.810	0.519	0.685	0.592	0.797
Ablation 2: Architectural Components														
Base	0.608	0.818	0.581	0.776	0.589	0.788	0.590	0.794	0.591	0.805	**0.549**	**0.709**	0.590	0.791
+ PDM	0.601	0.814	0.580	0.778	0.584	0.782	0.588	0.787	0.587	0.797	0.538	0.686	0.585	0.785
+ FER	**0.621**	**0.846**	**0.592**	**0.796**	**0.598**	**0.807**	**0.597**	**0.805**	**0.605**	**0.824**	0.532	0.694	**0.598**	**0.808**
+ Both (Ours)	0.603	0.813	0.576	0.773	0.581	0.777	0.582	0.772	0.587	0.795	0.545	0.699	0.583	0.780

Frame Count. Seven frames improve both explanation metrics over five frames in every category (Table 23). Overall ROUGE-L rises from 0.583 to 0.596, and CSS from 0.780 to 0.803. Color shows the largest CSS increase, from 0.772 to 0.801. At nine frames, Addition reaches its highest ROUGE-L of 0.618 and CSS of 0.836, while Removal reaches 0.592 ROUGE-L and 0.796 CSS. Moving from seven to nine frames reduces Portrait ROUGE-L from 0.555 to 0.519 and CSS from 0.721 to 0.685, alongside smaller decreases in Replacement, Color, and Material. The aggregate explanation scores consequently peak at seven frames.

Architectural Components. FER alone achieves the highest ROUGE-L and CSS in Addition, Removal, Replacement, Color, and Material. Base leads on Portrait with 0.549 ROUGE-L and 0.709 CSS, compared with FER's 0.532 ROUGE-L and 0.694 CSS. Adding PDM to FER reduces CSS by 0.033 in both Addition and Color, while raising Portrait ROUGE-L by 0.013 and CSS by 0.005. FER alone yields the strongest overall explanation scores of 0.598 ROUGE-L and 0.808 CSS; the full architecture yields the strongest grounding scores while retaining 0.583 ROUGE-L and 0.780 CSS at the benchmark level.

E.4.3 Forgery Detection

Frame Count. Additional frames improve overall detection while changing which categories perform best (Table 24). Moving from five to seven frames increases Accuracy in Addition, Removal, Replacement, and Color, including gains of 0.031 in Replacement and 0.015 in Removal. Material and Portrait decrease by 0.008 and 0.017. Portrait reaches its highest Accuracy of 0.839 at five frames, and Material peaks at 0.939 with either five or nine. Seven and nine frames both achieve 0.926 overall F1. The nine-frame Accuracy of 0.927 exceeds seven frames by 0.001, with inference time increasing from 17.72 s to 19.20 s.

Architectural Components. The full model's detection scores reflect gains in four categories and a reduction in Replacement. Relative to Base, Accuracy increases by 0.003 in Addition, 0.006 in Color, 0.019 in Material, and 0.025 in Portrait. Removal remains at 0.943, while Replacement decreases from 0.890 to 0.868. The full model reaches 0.914 Accuracy and 0.913 F1, each 0.001 below Base. Combining the modules also improves on either module alone, whose overall Accuracy

Table 24: **Category-wise Ablation for Forgery Detection.** Column groups follow the six manipulation categories in the main results, with benchmark-level scores in Avg. Metrics are video-level Accuracy (Acc) and F1 (↑) for original and manipulated videos.

Setting	ManiVidBench												Avg.	
	Addition		Removal		Replacement		Color		Material		Portrait			
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Ablation 1: Number of Sampled Frames														
3 Frames	0.953	0.953	0.920	0.920	0.885	0.885	0.906	0.906	0.912	0.912	0.805	0.803	0.906	0.905
5 Frames	0.966	0.966	0.943	0.943	0.868	0.868	0.917	0.917	**0.939**	**0.939**	**0.839**	**0.839**	0.914	0.913
7 Frames	**0.971**	**0.971**	**0.958**	**0.958**	0.899	0.899	**0.930**	**0.930**	0.931	0.931	0.822	0.821	0.926	**0.926**
9 Frames	0.966	0.966	0.950	0.950	**0.905**	**0.905**	**0.930**	**0.930**	**0.939**	**0.939**	0.814	0.812	**0.927**	**0.926**
Ablation 2: Architectural Components														
Base	0.963	0.963	**0.943**	**0.943**	**0.890**	**0.890**	0.911	0.911	0.920	0.920	0.814	0.809	**0.915**	**0.914**
+ PDM	0.961	0.961	0.935	0.935	0.860	0.860	0.909	0.909	0.931	0.931	0.805	0.803	0.905	0.904
+ FER	0.961	0.961	0.931	0.931	0.887	0.886	0.896	0.895	0.924	0.924	0.771	0.764	0.907	0.906
+ Both (Ours)	**0.966**	**0.966**	**0.943**	**0.943**	0.868	0.868	**0.917**	**0.917**	**0.939**	**0.939**	**0.839**	**0.839**	0.914	0.913

values are 0.905 for PDM and 0.907 for FER. The largest recovery occurs in Portrait, where the full model reaches 0.839 vs. 0.805 for PDM and 0.771 for FER.

E.4.4 Prompt Design

Prompt Settings. The separate setting performs explanation, detection, and grounding through three task-specific queries to one shared ManiVidLens model. The all-in-one setting combines the same tasks into a single response in the fixed order of explanation, detection, and grounding (Table 19). During Stage 1, separate prompting expands each sampled video view into three task-specific instances over one epoch, while all-in-one prompting uses one joint instance over two epochs; both settings then use one Stage 2 epoch.

Table 25: **Category-wise Ablation on Prompt Design.** All-in-one and separate prompts are compared across six manipulation categories and three tasks. Avg. reports benchmark-level scores. Higher metric values (↑) indicate better performance on all three tasks.

Prompt Design	ManiVidBench												Avg.	
	Addition		Removal		Replacement		Color		Material		Portrait			
Task Mode: Classification														
Metrics	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
All-in-one	0.838	0.833	0.779	0.768	0.777	0.766	0.758	0.744	0.790	0.782	0.627	0.567	0.778	0.767
Separate (Ours)	**0.966**	**0.966**	**0.943**	**0.943**	**0.868**	**0.868**	**0.917**	**0.917**	**0.939**	**0.939**	**0.839**	**0.839**	**0.914**	**0.913**
Task Mode: Explanation														
Metrics	R-L	CSS	R-L	CSS	R-L	CSS	R-L	CSS	R-L	CSS	R-L	CSS	R-L	CSS
All-in-one	0.597	0.800	0.575	0.766	0.579	0.771	0.573	0.757	0.578	0.774	0.540	0.690	0.579	0.769
Separate (Ours)	**0.603**	**0.813**	**0.576**	**0.773**	**0.581**	**0.777**	**0.582**	**0.772**	**0.587**	**0.795**	**0.545**	**0.699**	**0.583**	**0.780**
Task Mode: Segmentation														
Metrics	mIoU	J&F	mIoU	J&F	mIoU	J&F	mIoU	J&F	mIoU	J&F	mIoU	J&F	mIoU	J&F
All-in-one	0.381	0.395	0.113	0.119	0.295	0.308	0.275	0.288	0.348	0.359	0.028	0.032	0.275	0.286
Separate (Ours)	**0.444**	**0.467**	**0.135**	**0.137**	**0.421**	**0.439**	**0.487**	**0.507**	**0.526**	**0.532**	**0.290**	**0.306**	**0.407**	**0.422**

Task Performance. Separate prompting improves all six metrics (Table 25). Grounding mIoU increases **from 0.275 to 0.407**, and J&F from 0.286 to 0.422. Detection Accuracy rises **from 0.778 to 0.914**, and F1 from 0.767 to 0.913. Explanation improves by 0.004 ROUGE-L and 0.011 CSS, reaching 0.583 ROUGE-L and 0.780 CSS. The larger gains in detection and grounding show that task separation chiefly benefits authenticity decisions and spatial localization.

Category-wise Effects. Separate prompting improves every metric across all six manipulation categories. The largest grounding and detection gains occur on Portrait: mIoU increases **from 0.028 to 0.290**, and detection F1 rises from 0.567 to 0.839. Color mIoU also rises from 0.275 to 0.487, while Material CSS improves from 0.774 to 0.795.

Task-Specific Generation. In the all-in-one sequence, detection is conditioned on the generated explanation, and grounding is conditioned on both preceding outputs. Separate prompts give each task its own generation context: detection directly predicts authenticity, and grounding directly produces the localization phrase and segmentation tokens. The three tasks retain shared model parameters and visual evidence. The largest gains occur in the two tasks that follow explanation in the all-in-one sequence, supporting separate prompts as the default task protocol.

E.5 Robustness Study

Table 26: **Robustness Evaluation under Different Video Degradations.** Grounding, explanation, and detection scores (↑) are evaluated under four types of video degradation. Parenthesized values show signed percentage changes relative to the undegraded ManiVidLens baseline, with negative values indicating performance degradation for the corresponding metric.

Setting		Grounding		Explanation		Detection	
		mIoU	J&F	R-L	CSS	Acc	F1
ManiVidLens		0.407	0.422	0.583	0.780	0.914	0.913
H.264 Compression	CRF 23	0.384 (**-5.5%**)	0.400 (**-5.2%**)	0.562 (**-3.7%**)	0.735 (**-5.8%**)	0.882 (**-3.4%**)	0.882 (**-3.5%**)
	CRF 40	0.246 (**-39.5%**)	0.255 (**-39.5%**)	0.514 (**-11.8%**)	0.638 (**-18.2%**)	0.591 (**-35.3%**)	0.535 (**-41.5%**)
Downsampling	× 0.75	0.399 (**-2.0%**)	0.416 (**-1.4%**)	0.577 (**-1.2%**)	0.775 (**-0.6%**)	0.895 (**-2.0%**)	0.895 (**-2.0%**)
	× 0.50	0.400 (**-1.7%**)	0.415 (**-1.7%**)	0.567 (**-2.7%**)	0.764 (**-2.0%**)	0.859 (**-6.0%**)	0.859 (**-6.0%**)
FPS Reduction	16 FPS	0.387 (**-4.9%**)	0.401 (**-5.0%**)	0.575 (**-1.4%**)	0.761 (**-2.5%**)	0.897 (**-1.9%**)	0.896 (**-1.9%**)
	8 FPS	0.393 (**-3.4%**)	0.407 (**-3.5%**)	0.574 (**-1.5%**)	0.778 (**-0.3%**)	0.906 (**-0.8%**)	0.906 (**-0.8%**)
Random Frame Drop	10%	0.380 (**-6.6%**)	0.395 (**-6.4%**)	0.574 (**-1.6%**)	0.760 (**-2.6%**)	0.900 (**-1.5%**)	0.899 (**-1.6%**)
	30%	0.383 (**-5.7%**)	0.398 (**-5.7%**)	0.573 (**-1.7%**)	0.758 (**-2.8%**)	0.890 (**-2.6%**)	0.889 (**-2.6%**)

Evaluation Protocol. We apply each degradation independently to the original test videos using the same trained checkpoint, prompts, frame-sampling policy, and evaluation implementation. The settings comprise H.264 compression at CRF 23 and 40, spatial downsampling by factors of 0.75 and 0.50, frame-rate reduction to 16 and 8 FPS, and random removal of 10% and 30% of frames. Grounding is evaluated on every decoded frame of each degraded video: source-frame correspondences select the original masks, and predictions are mapped back to the original ground-truth coordinate system and resolution. Explanation and detection retain the original text references and authenticity labels, respectively, with metric aggregation following Appendix E.1.3. For baseline and degraded scores s_{base} and s_{degraded}, the signed relative change is

$$\Delta_s = \frac{s_{\text{degraded}} - s_{\text{base}}}{s_{\text{base}}} \times 100\%, \tag{16}$$

where negative values indicate a decrease from the corresponding baseline score.

Performance under Degradation. Severe H.264 compression produces the largest losses across all six metrics (Table 26). Increasing compression from CRF 23 to CRF 40 reduces mIoU from 0.384 to 0.246, detection F1 from 0.882 to 0.535, and CSS from 0.735 to 0.638. Spatial downsampling preserves grounding more strongly: at half resolution, mIoU changes **from 0.407 to 0.400** and J&F **from 0.422 to 0.415**, while detection F1 decreases from 0.913 to 0.859. Temporal degradations produce smaller changes than severe compression. At 8 FPS, ManiVidLens retains 0.393 mIoU, 0.778 CSS, and 0.906 detection F1. Randomly removing 30% of frames reduces mIoU from 0.407 to 0.383 and detection F1 from 0.913 to 0.889. Grounding is non-monotonic across the temporal severity levels: 8 FPS yields slightly higher mIoU than 16 FPS, as does 30% frame removal compared with 10%. Together, these results show stronger sensitivity to severe compression than to spatial resizing or temporal frame loss.

E.6 ManiVidLens Case Visualization

Fig. 14, 15 (***Addition***), 16, 17 (***Removal***), 18, 19 (***Replacement***), 20, 21 (***Color Alteration***), 22, 23 (***Material Modification***), 24, 25 (***Portrait Editing***) present multiple ManiVidLens cases across different manipulation samples.

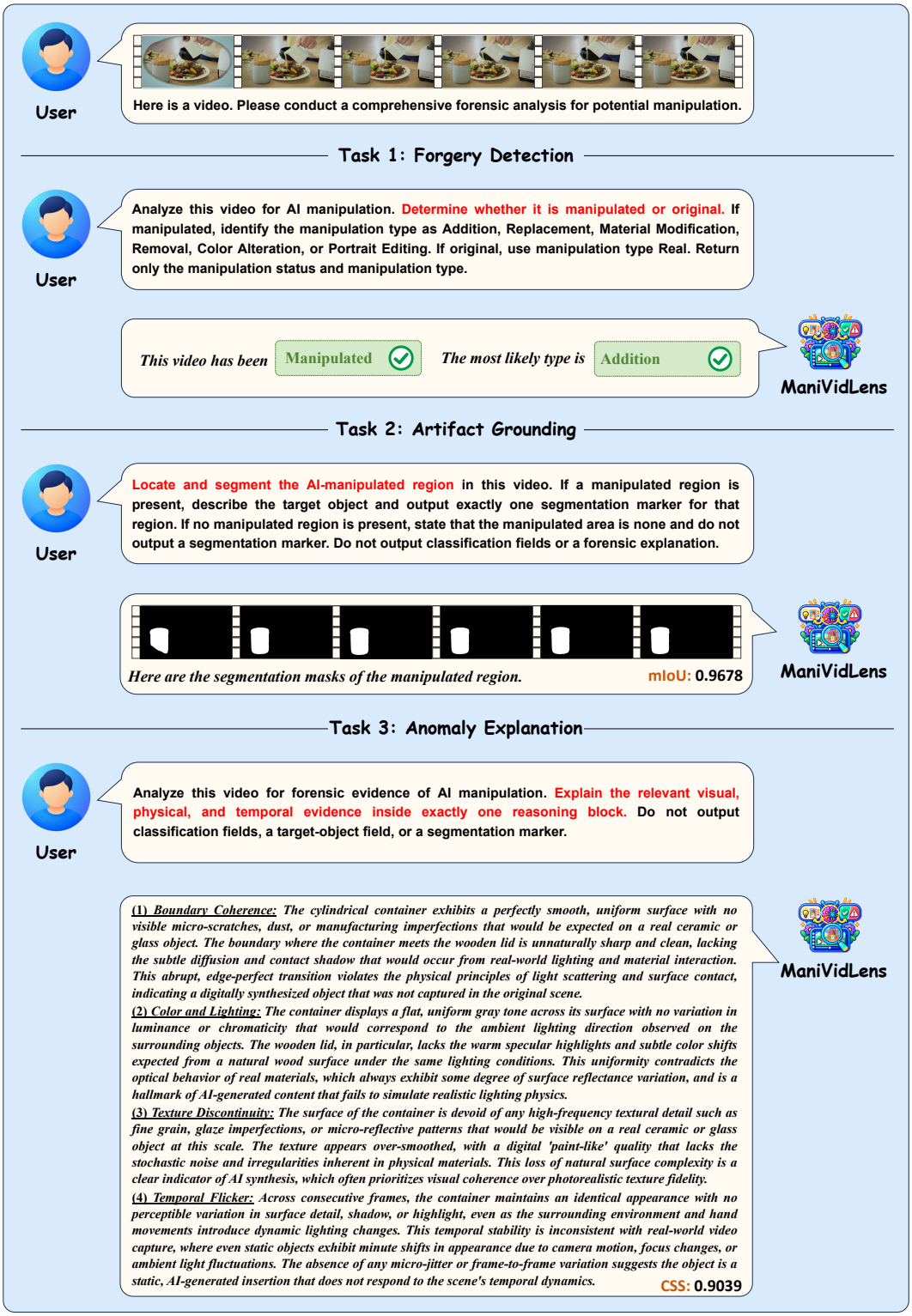


Figure 14: **A ManiVidLens Case on an *Addition* Sample.**

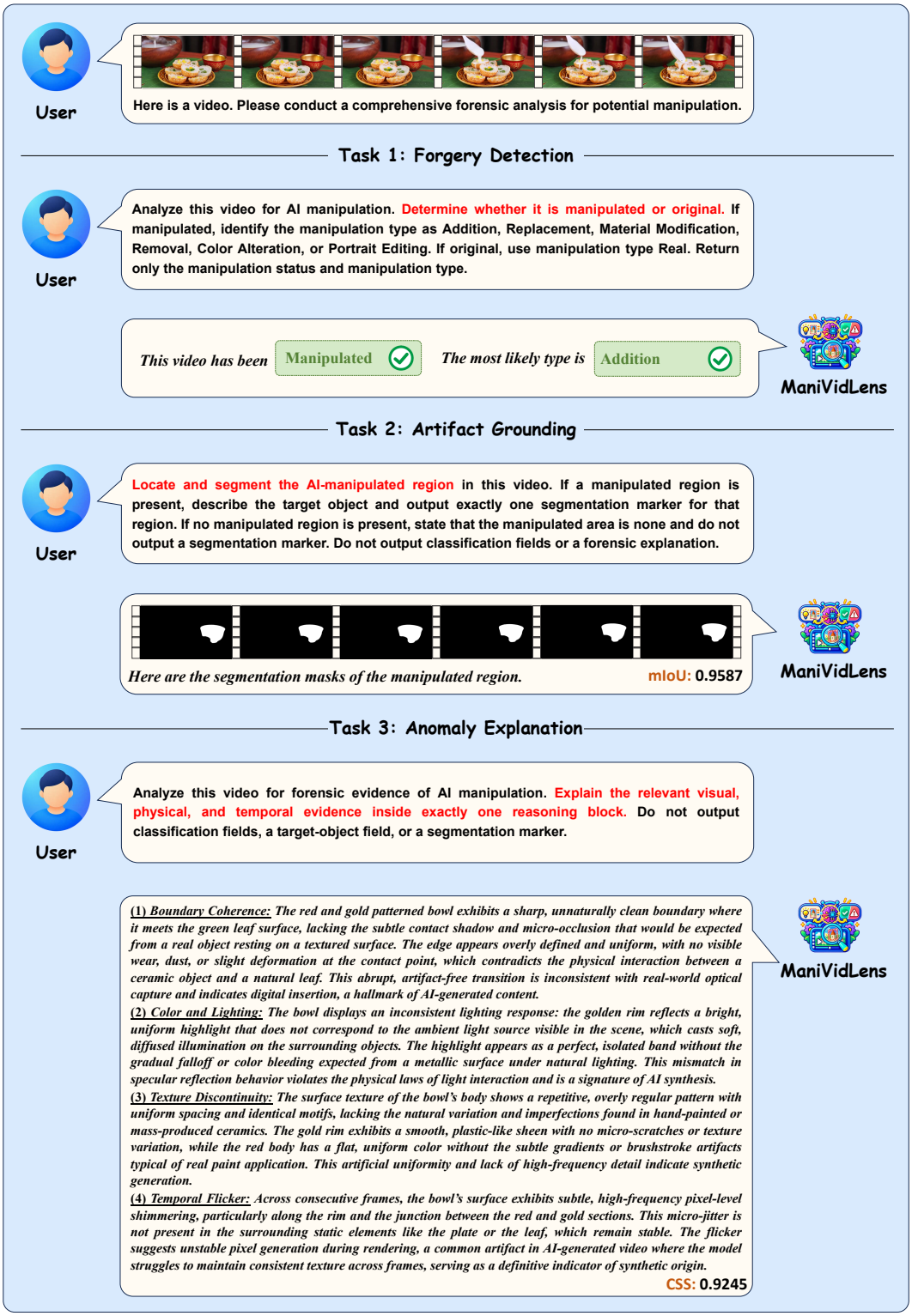


Figure 15: **A ManiVidLens Case on an *Addition* Sample.**

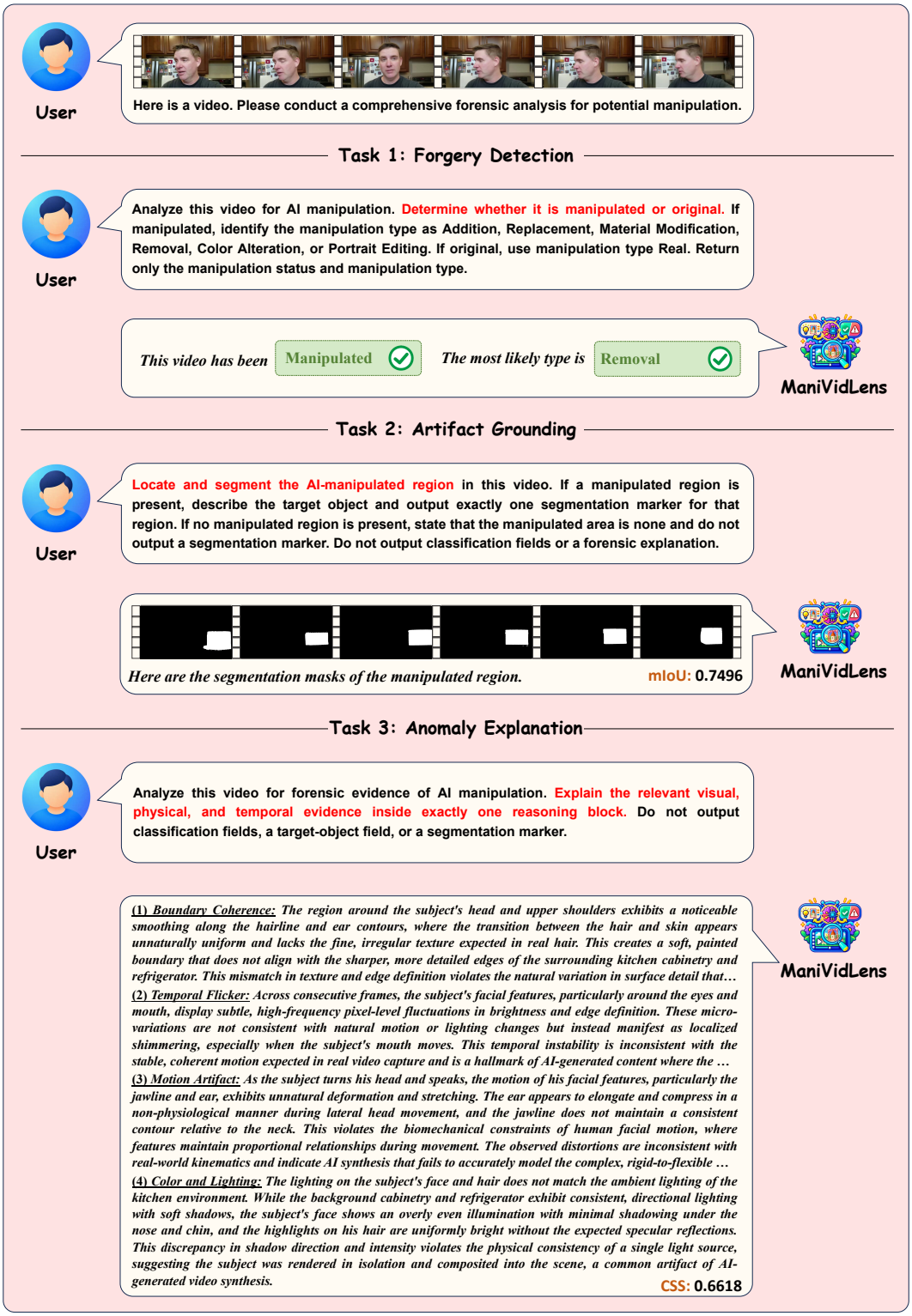


Figure 16: **A ManiVidLens Case on a *Removal* Sample.**



Figure 17: **A ManiVidLens Case on a *Removal* Sample.**

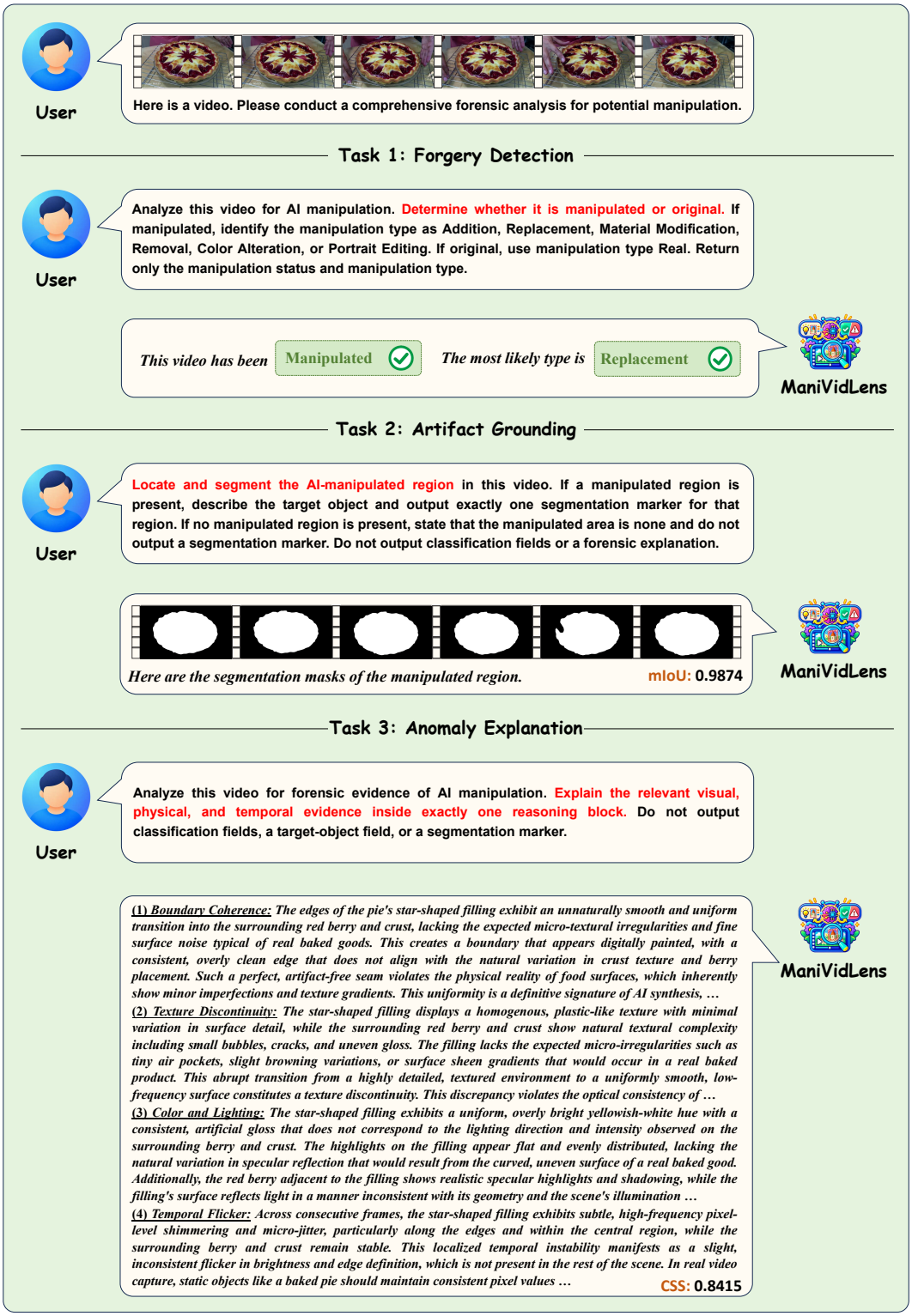


Figure 18: **A ManiVidLens Case on a *Replacement* Sample.**



Figure 19: **A ManiVidLens Case on a *Replacement* Sample.**



Figure 20: **A ManiVidLens Case on a *Color Alteration* Sample.**

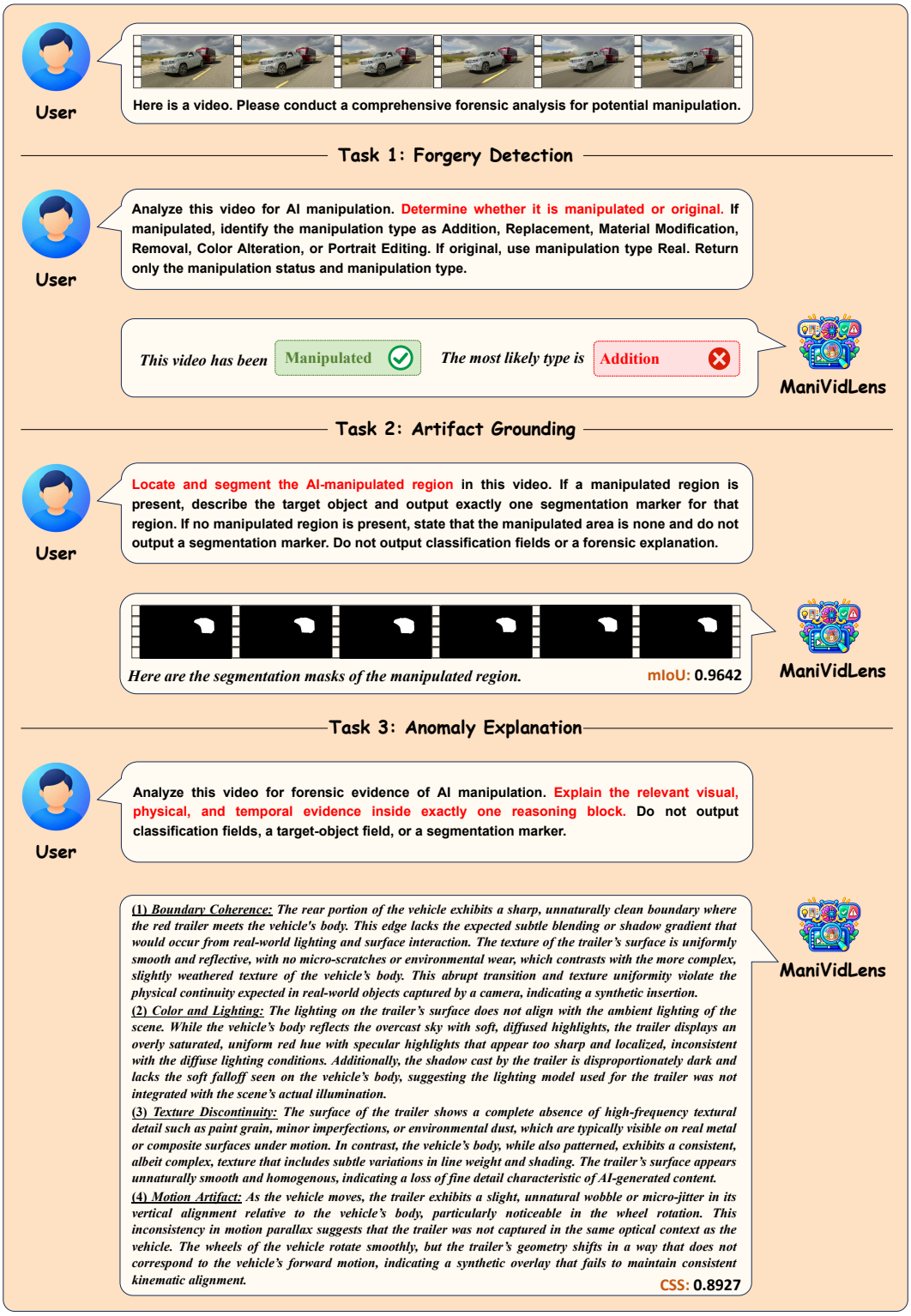


Figure 21: **A ManiVidLens Case on a *Color Alteration* Sample.**

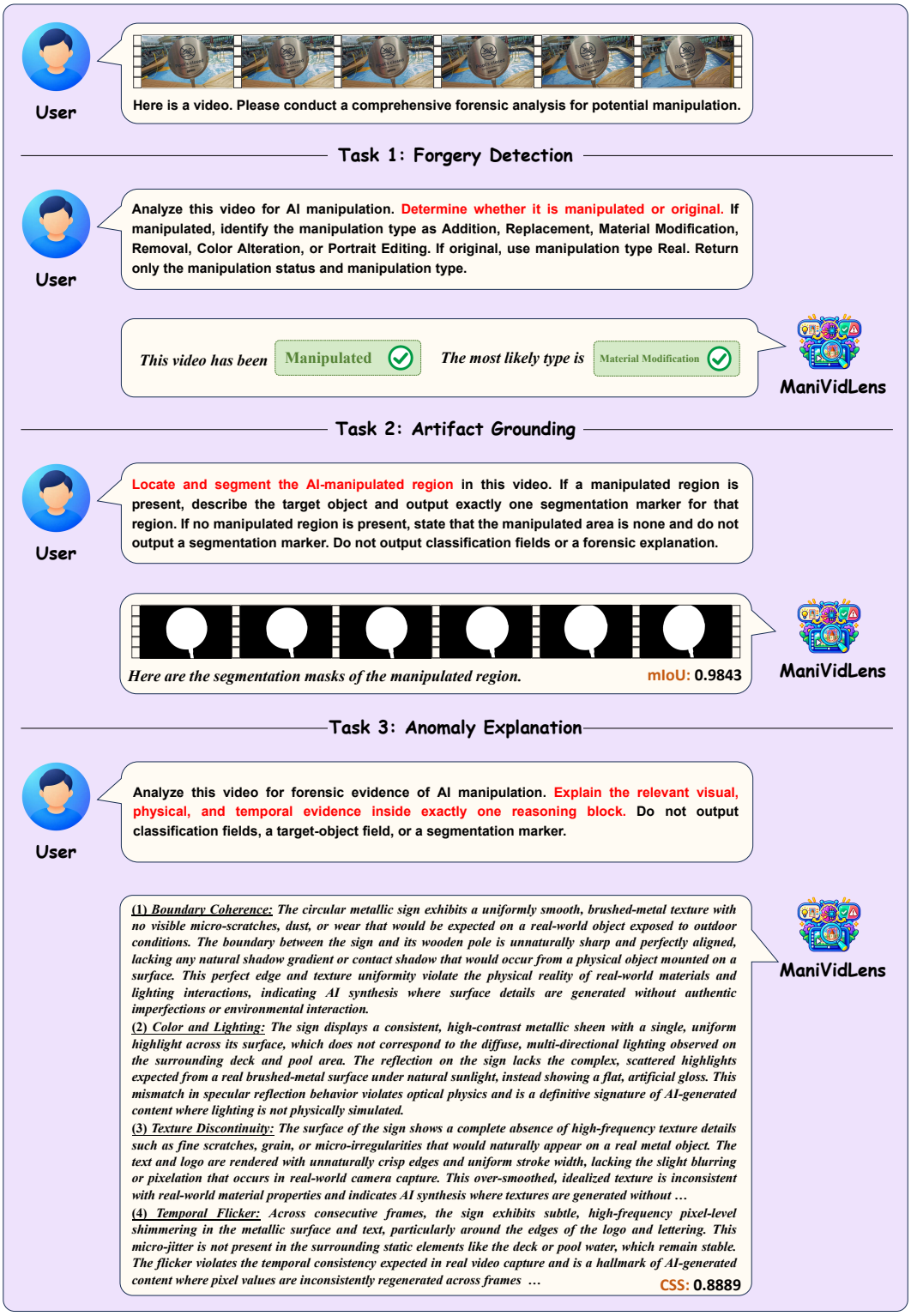


Figure 22: **A ManiVidLens Case on a *Material Modification* Sample.**

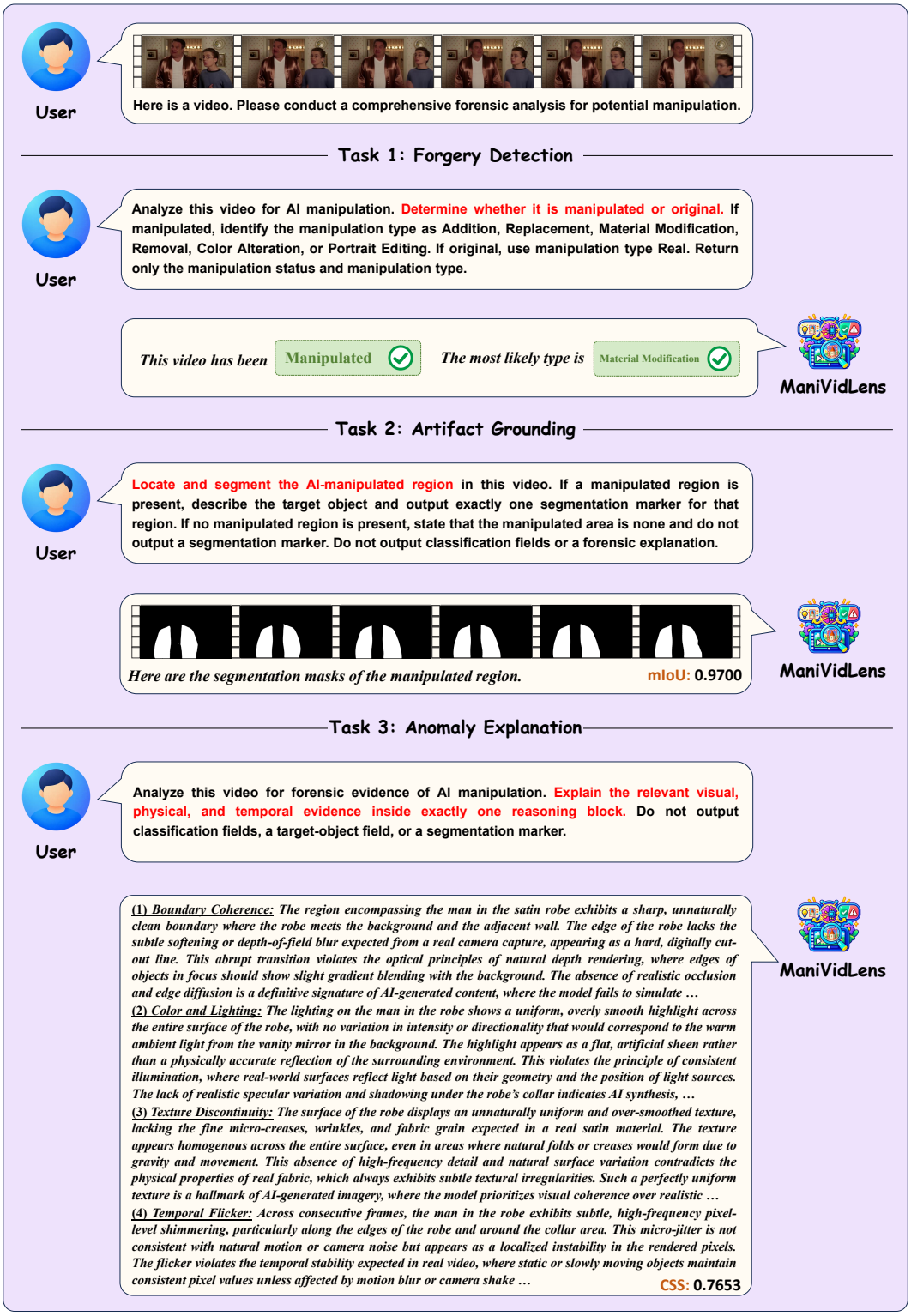


Figure 23: **A ManiVidLens Case on a *Material Modification* Sample.**

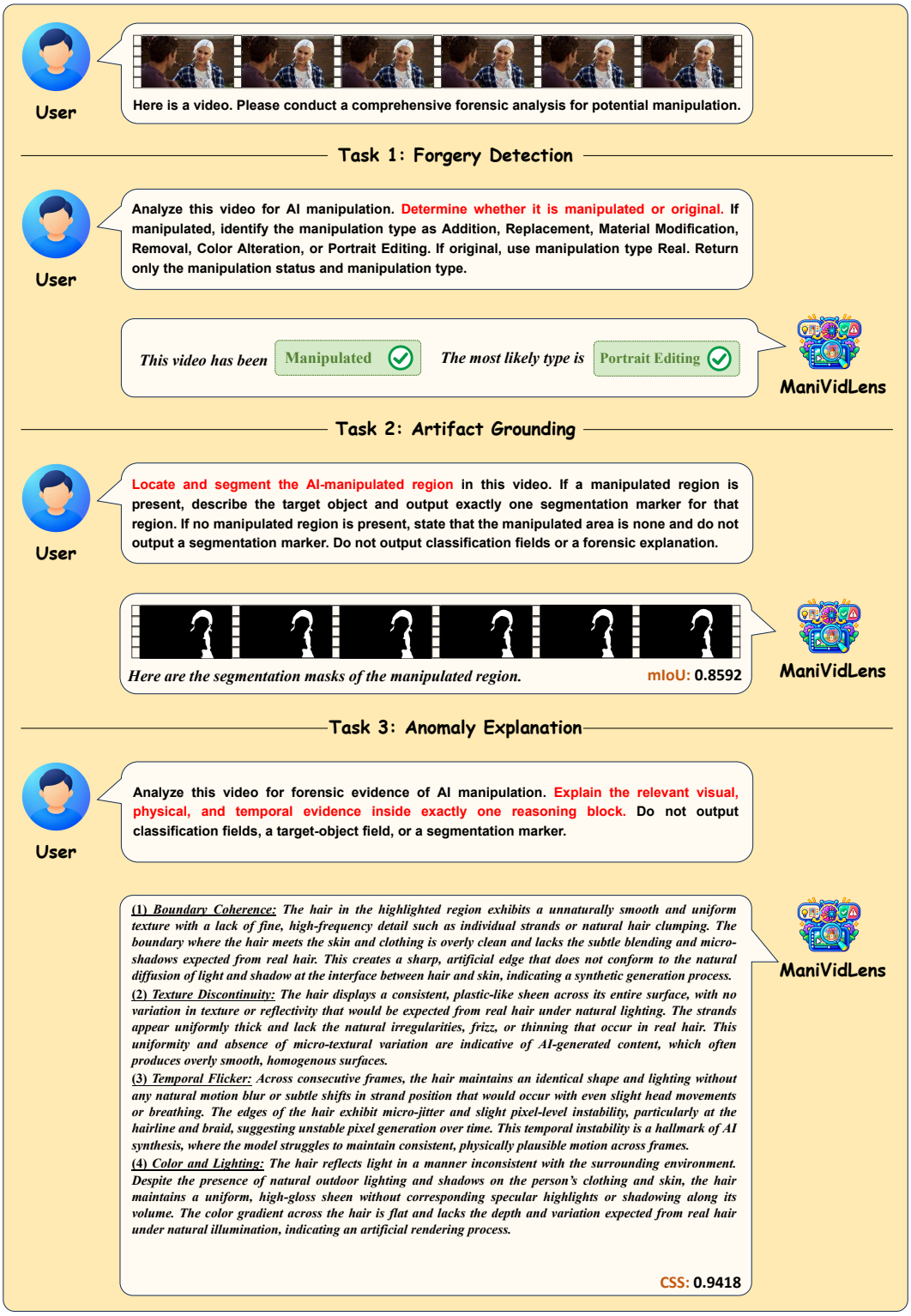


Figure 24: **A ManiVidLens Case on a *Portrait Editing* Sample.**

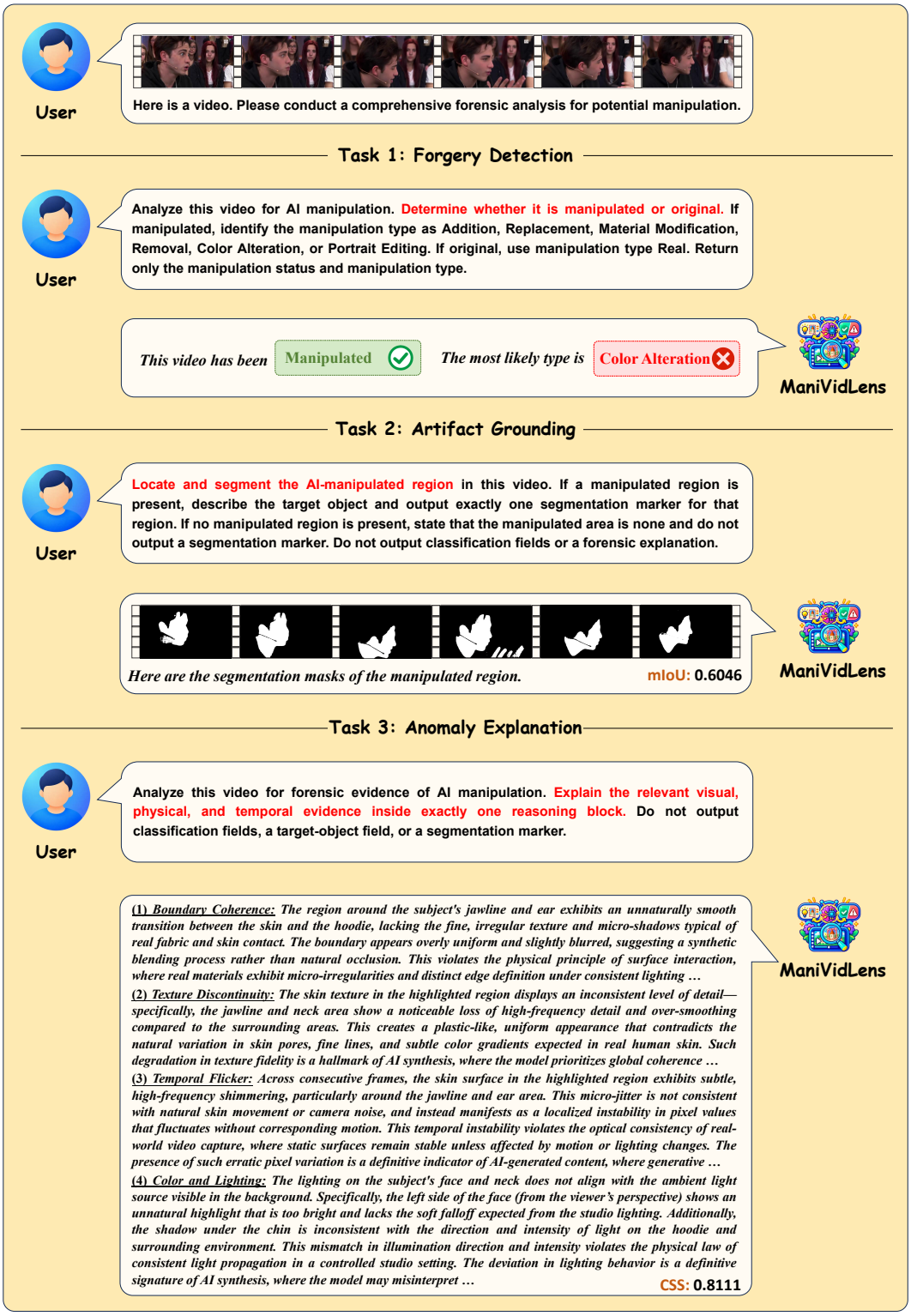


Figure 25: **A ManiVidLens Case on a *Portrait Editing* Sample.**