FROM LOW-LEVEL ARTIFACTS TO FORGERY KNOWLEDGE VIA SEMANTIC DIFFERENCING

Anonymous authors
Paper under double-blind review

ABSTRACT

Making forgery evidence explainable is fundamental to trustworthy AI-Generated Image (AIGI) detection, yet low-level forgery artifacts remain semantically elusive to Multimodal Large Language Models (MLLMs), which can hardly articulate such artifacts in language. We present the first study on semanticizing low-level forgery artifacts. At its core, we introduce Subvisual Forgery Semantics (**SFS**), describing how forgery traces below salient perception manifest as object-grounded semantic effects, and develop Semantic Differencing (**SemDiff**), a multi-agent pipeline that contrasts an image with its reconstruction and translates pixel residuals into structured semantic descriptions. With SemDiff, we construct **TRACE-Bench**, the first benchmark for assessing MLLMs on subvisual forgery semantics, curated from multi-source annotations over three reconstruction models, where state-of-the-art MLLMs struggle to describe them. As a strong baseline, we present ForgeryTracer, which preserves forgery-related low-level cues through residual-token supervision and learns to articulate them as SFS through semantic supervision, showing that SFS can be learned while maintaining detection accuracy. Our work makes low-level forgery evidence not only detectable but explainable, opening a low-level perspective for interpretable AIGI forensics. Data and models will be released.

INTRODUCTION

The rapid proliferation of Artificial Intelligence Generated Content (AIGC) raises unprecedented concerns about visual misinformation and digital trust Chen et al. (2025c); Cai et al. (2025); OpenAI (2023); AI (2024); Midjourney, Inc. (2025); Rombach et al. (2022); Labs (2024). In response, forgery detection has become the primary safeguard, with evidence drawn from two sources, high-level semantic anomalies and low-level artifacts Cheng et al. (2025). The latter, subtle traces left by the generation process, have supported a wide range of reliable detectors Yang et al. (2025a;b); Tan et al. (2023). As detection matures, the focus of trustworthy forensics is shifting from reaching a correct verdict to justifying it, because a binary score alone is hard to trust or verify. With strong semantic reasoning and rich world knowledge DeepMind & Google (2024); Anthropic (2024); OpenAI (2025), Multimodal Large Language Models (MLLMs) are well suited to the high-level side, and recent work uses them to describe semantic anomalies in natural language for explainable AIGI forensics Wen et al. (2025a); Huang et al. (2025a;b); Zhou et al. (2025).

However, this progress covers only one of the two sources of evidence Chen et al. (2024b); Tan et al. (2025d); Li et al. (2025d); Wen et al. (2025c); Li et al. (2025b), and low-level forgery artifacts remain semantically elusive to MLLMs. When images lack obvious semantic anomalies and the decisive evidence lies in low-level artifacts, MLLMs that rely primarily on high-level semantic cues often struggle, as they can hardly perceive and reason about such subtle traces Chen et al. (2024a); Lin et al. (2025); Chen et al. (2025b). Even when a forgery is correctly detected, its low-level evidence thus stays locked in the pixel domain and never reaches the explanation. Consequently, ***semanticizing low-level forgery artifacts represents a pressing yet open challenge.***

In this paper, we present the first study on semanticizing low-level forgery artifacts. At its core, we introduce ***Subvisual Forgery Semantics (SFS)***, object-grounded semantic descriptions of how forgery traces below salient perception manifest in an image. SFS expands the evidence space of explainable forensics, complementary to the salient semantic anomalies that MLLMs already analyze well. Capturing SFS, however, is not straightforward, since it can hardly be derived from a single

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.

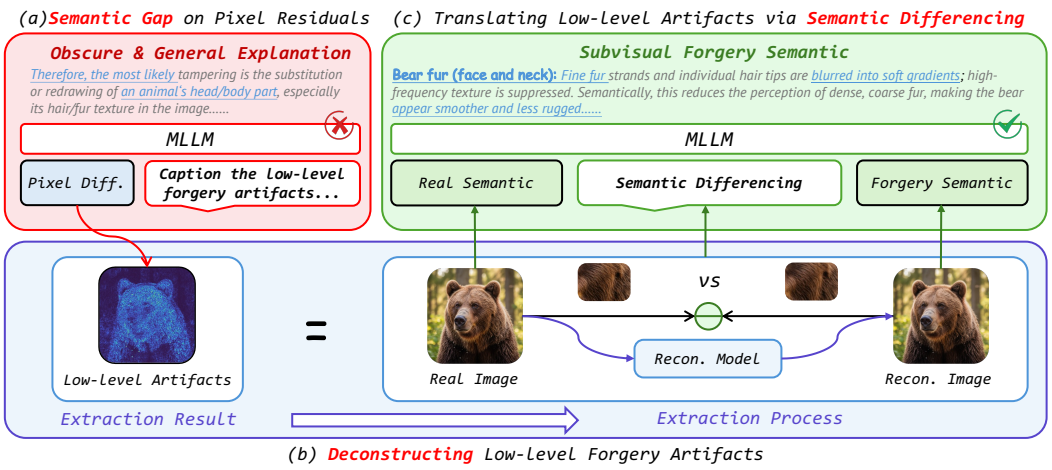


Figure 1: **Bridging the Semantic Gap in VLMs for Subvisual Forgery Semantic Analysis. (a)** MLLMs struggle to interpret non-semantic pixel residuals directly. **(b)** We deconstruct static artifact results back into the explicit differencing process of real and reconstructed image pairs. **(c)** Semantic Differencing transforms low-level artifacts into Subvisual Forgery Semantics for interpretable AIGI forensics.

image or read off raw low-level artifacts, as illustrated in Fig. 1. We therefore deconstruct low-level artifacts back into the differencing process that produces them, transforming ***Pixel Difference*** into ***Semantic Difference***. In this form, forgery evidence surfaces as object-level semantic discrepancies between an image and its reconstruction, which MLLMs can natively capture from a high-level perspective.

To produce SFS at scale, we develop **Semantic Differencing** (**SemDiff**), an autonomous multi-agent pipeline that captures object-level semantic discrepancies across reconstruction and progressively refines them into reliable forgery knowledge. To evaluate SFS, we build **TRACE-Bench** with SemDiff, the first benchmark measuring how well MLLMs describe SFS, curated from multi-source annotations over three reconstruction models, on which state-of-the-art MLLMs consistently fall short. To learn SFS, we present **ForgeryTracer**, a strong baseline trained under dual supervision, where residual-token supervision preserves forgery-related low-level cues in the representation and semantic supervision from SemDiff teaches the model to articulate them as SFS. Experiments indicate that SFS can be learned without sacrificing detection accuracy.

In summary, our main contributions are threefold:

- We present the first study on semanticizing low-level forgery artifacts and introduce Subvisual Forgery Semantics (**SFS**), which expresses forgery traces below salient perception as their effects on object-level semantics, complementing the high-level anomaly analysis of MLLM-based forensics.
- We propose Semantic Differencing (**SemDiff**), which captures SFS by deconstructing low-level artifacts into the differencing process, recasting pixel differences as object-grounded semantic descriptions, and on this basis, we build a multi-agent data pipeline that yields reliable forgery knowledge at scale.
- We construct **TRACE-Bench**, the first benchmark for evaluating MLLMs' ability on SFS, and introduce **ForgeryTracer**, a dual-supervision baseline showing that SFS can be learned without sacrificing detection accuracy.

Together, they make low-level forgery evidence not only detectable but explainable, opening a low-level perspective for interpretable AIGI forensics.

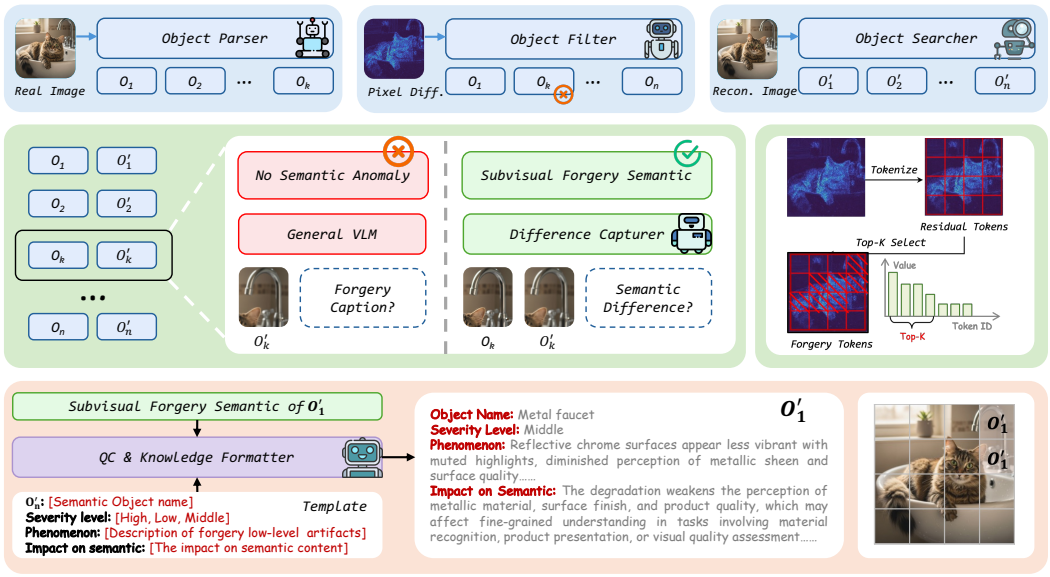


Figure 2: **Overview of Semantic Differencing (SemDiff) Data Pipeline.** Stage 1 grounds low-level residual evidence into semantic object pairs. Stage 2 reformulates artifact extraction as semantic differencing, translating subvisual object-level effects into subvisual forgery semantics. Stage 3 consolidates the textual semantics and visual tokens into structured forgery knowledge.

Related Work

AIGI Detection. Reliable AIGI detection heavily relies on low-level features Tan et al. (2023; 2024a); Yang et al. (2025a), including frequency anomalies Tan et al. (2024a), residual patterns Chen et al. (2024a); Cheng et al. (2025), and generator fingerprints Tan et al. (2024b); Liu et al. (2024a); Wang et al. (2025). Although these methods effectively capture low-level artifacts, they typically output binary classification scores alone. Without semantic explanations of the analyzed content, such predictions are hard to trust in real-world deployment.

MLLMs for AIGI Detection. MLLMs such as LLaVA Li et al. (2024) and Qwen-VL Bai et al. (2025b;a); Wang et al. (2024) have shown strong visual understanding and reasoning OpenAI (2024); Hurst et al. (2024), and have been adapted to detect and explain AIGI Zhou et al. (2025); Liu et al. (2024b); Huang et al. (2025b). However, existing models construct forgery explanations mainly from high-level semantic inconsistencies, such as violations of physical laws Zhou et al. (2025); Tan et al. (2025c); Wen et al. (2025a;b). On high-fidelity generated images without obvious semantic anomalies, they often fail, as their perception-aligned encoders are largely invariant to such subtle perturbations, leaving forgery traces indistinguishable without an explicit reference Tong et al. (2024); Wu et al. (2023); Li et al. (2025c).

Low-level Forgery Understanding in MLLMs. Recent studies improve MLLMs' access to low-level forgery artifacts through supplementary feature extractors Zhou et al. (2025), agentic tools Huang et al. (2025a), or decoupled training Lin et al. (2025). Yet, these artifacts are rarely semanticized as reasoning-accessible evidence for MLLMs. Existing benchmarks for explainable forgery forensics likewise center on high-level semantic anomalies Li et al. (2025c); Wen et al. (2025c), leaving the semantic effects of low-level artifacts unmeasured. How to define, capture, and represent the subtle object-level effects induced by low-level traces of AIGI formation remains underexplored. We address this gap through Semantic Differencing, which reformulates artifact extraction as semantic differencing to produce SFS and TRACE-Bench, which measures how well MLLMs describe them.

SEMDIFF FRAMEWORK

STAGE 1: SEMANTIC OBJECT PAIRS PARSING

To extract SFS derived from low-level artifacts, we introduce an autonomous multi-agent framework, named **SemDiff**, as illustrated in Fig. 2. Consistent with previous methods Chen et al. (2025b; 2024a); Cheng et al. (2025); Lin et al. (2025), we compute the discrepancy between the original image I and the reconstructed image I' to derive the **Pixel Difference** D_{pixel} as low-level artifacts:

$$D_{pixel} = |I - I'| \tag{1}$$

We argue that identifying low-level forgery artifacts and extracting semantic representation inherently requires a strong alignment with the underlying image content. To achieve this, our method first processes the real image through an **Object Analyzer** $\mathcal{A}_{analyze}$ to extract a set of semantic objects:

$$\mathcal{O} = \mathcal{A}_{analyze}(I) = \{O_1, O_2, \ldots, O_n\} \tag{2}$$

Subsequently, an **Object Filter** $\mathcal{A}_{filter}$ filters these extracted targets based on their pixel-level discrepancies D_{pixel}, discarding objects that exhibit marginal differences. Finally, an **Object Searcher** $\mathcal{A}_{search}$ locates the corresponding semantic objects within the reconstructed image. This process successfully matches each retained original object O_i with its reconstructed counterpart $O_i^{'}$ to formulate a set of real-forged semantic object pairs $\mathcal{P}$ for subsequent analysis:

$$\mathcal{P} = \mathcal{A}_{search}(\mathcal{A}_{filter}(\mathcal{O}, D_{pixel}), I') = \{(O_i, O_i')\}_{i=1}^{M} \tag{3}$$

STAGE 2: SEMANTIC DIFFERENCING

After obtaining the semantic object pairs, we reformulate low-level artifact extraction as a semantic differencing process. Rather than directly interpreting the final pixel difference, we use reconstruction contrast to reveal how low-level traces left by AIGI formation subtly affect object-level visual content, thereby bridging low-level visual evidence and reasoning-accessible textual descriptions.

For each object pair (O_i, O_i'), where O_i and O_i' denote the input object and its reconstructed counterpart, respectively, a semantic differencing agent $\mathcal{A}_{diff}$ compares their local appearance, structure, and object-level cues. The observed differences are formulated as SFS: subvisual object-level effects induced by low-level formation traces that remain below salient visual perception. Guided by the **Semantic Differencing Prompt** Prompt_{SD}, the agent generates a structured textual description:

$$T_i = \mathcal{A}_{diff}(\text{Prompt}_{SD}, O_i, O_i'), \tag{4}$$

where T_i describes the subvisual forgery effects associated with object O_i. As shown in Fig. 2, a reconstructed metal faucet may exhibit weakened metallic sheen and muted highlights, indicating degraded material perception caused by low-level formation traces. In parallel, we treat the pixel difference map D_{pixel} as a visual evidence branch rather than the final explanation. We partition D_{pixel} into token-aligned patches $\mathcal{F}_{patch}(D_{pixel})$ and select the top-k patches with the largest residual responses as forgery-aware visual tokens:

$$V_i = \text{TopK}(\mathcal{F}_{patch}(D_{pixel}), k). \tag{5}$$

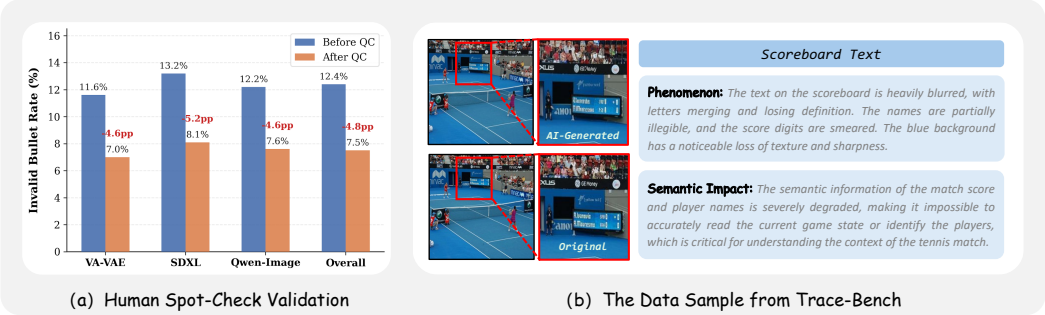
These tokens provide localized residual supervision complementary to the textual SFS T_i, enabling dual evidence for MLLM training: semantic reasoning from T_i and residual visual representation learning from V_i.

STAGE 3: FORGERY KNOWLEDGE STRUCTURING

Finally, a structuring agent $\mathcal{A}_{format}$ consolidates the textual SFS T_i and residual-aware visual tokens V_i into structured forgery knowledge:

$$K_i = \mathcal{A}_{format}(T_i, V_i) = \{N_i, P_i, S_i, L_i\}, \tag{6}$$

where N_i denotes the object name, P_i describes the observed subvisual forgery phenomenon, S_i summarizes its impact on semantic content, and L_i indicates the severity level categorized as high, moderate, or low. Through this process, SemDiff aligns low-level visual evidence with object-grounded textual semantics, forming the basis of TRACE-Bench for MLLM-based forgery reasoning.



(a) Human Spot-Check Validation (b) The Data Sample from Trace-Bench

Figure 3: **TRACE-Bench.** (a) Human spot-check verification before and after QC. (b) Example of structured representation of low-level artifacts.

Quality Control & Data Construction

We build **TRACE-Bench** (**T**ranslating **R**esidual **A**rtifacts into **C**omprehensible **E**xplanations) on COCO 2017 Lin et al. (2014), generating reconstructions with three VAEs of distinct design lineages, the classic SDXL VAE widely inherited by diffusion models Podell et al. (2023), the VAE equipped by the frontier generator Qwen-Image Wu et al. (2025), and the representation-aligned VA-VAE Yao et al. (2025), yielding diverse low-level forgery traces (with statistics shown in Fig. 5 and annotated example in Fig. **??**). To avoid single-annotator bias, SemDiff draws its agents from heterogeneous MLLM families, including Qwen3-VL-32B Bai et al. (2025a), Qwen2.5-VL-72B Bai et al. (2025b), InternVL3-14B Zhu et al. (2025), and LLaVA-OneVision-72B Li et al. (2024). Furthermore, two MLLM judges (Qwen3-VL-235B-A22B Bai et al. (2025a) and InternVL3-78B Zhu et al. (2025)) review all annotations, and a semantic object is dropped once either judge flags it as unfaithful to the visual evidence. Human audits on random samples drawn before and after this filtering confirm a clear quality gain, with protocols and statistics in the Appendix. For evaluation, the In-Domain Setting trains and validates on all three VAEs, while the Out-of-Domain Setting trains on SDXL and Qwen-Image reconstructions and evaluates on held-out VA-VAE ones to assess cross-generator generalization.

ForgeryTracer

Building upon TRACE-Bench, we propose a strong baseline, **ForgeryTracer**, to generate SFS without sacraficing detection performance. As shown in Fig. 4, ForgeryTracer applies dual supervision with textual SFS and residual-salient visual tokens. Our framework builds upon Qwen3-VL Bai et al. (2025a) with the multi-layer visual token interface of DeepStack Meng et al. (2024). Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the vision encoder extracts patch tokens $V = \{v_i\}_{i=1}^{N}$ and projects them into the LLM decoder through a linear connector. Building on this multi-layer token representation, we employ residual-aware supervision via high-residual difference map tokens to preserve forgery-relevant low-level cues.

Dual Constraint

As shown in Fig. 4, the visual branch employs a lightweight token-level mask decoder that prompts the vision encoder to retain low-level forgery cues, thereby reinforcing dense classification tokens.

High-Residual Token Selection. To obtain token-level supervision for low-level artifacts, we select visual tokens with the highest responses. Specifically, we choose the top $k\%$ tokens according to their difference intensity and use them as the ground-truth token set:

$$\mathcal{M}_{gt} = \{i \mid \mathcal{I}_{diff}(i) \geq \tau_k\}, \tag{7}$$

where $\mathcal{I}_{diff}(i)$ denotes the intensity of the i-th token in the difference map, and τ_k is the threshold corresponding to the top $k\%$ intensities.

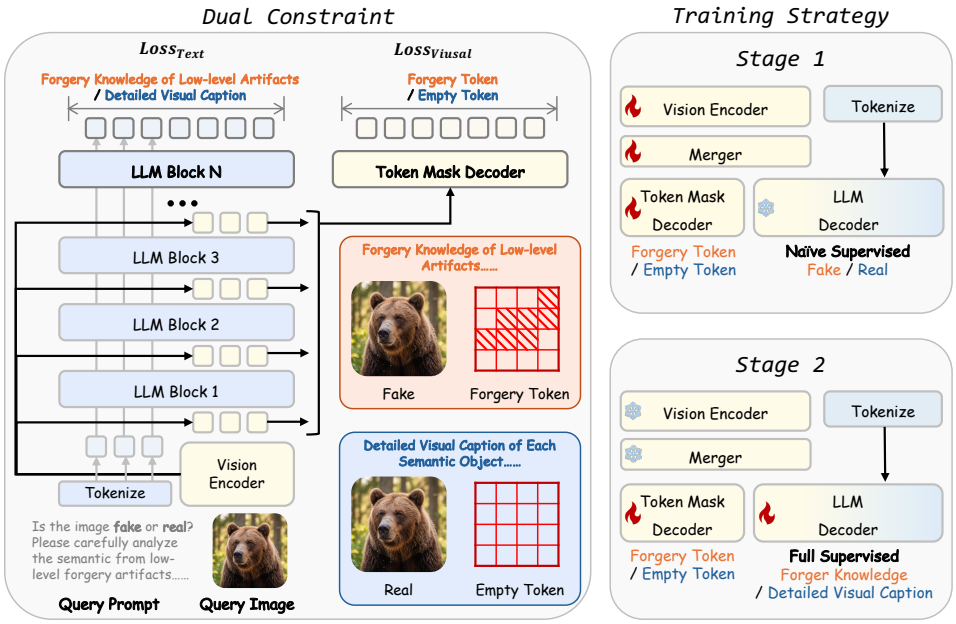


Figure 4: **Overview of ForgeryTracer and the training strategy. Left:** We inject low-level forensic knowledge into MLLMs through dual supervision over textual explanations and residual-salient visual tokens. **Right:** Our training strategy first establishes fundamental artifact perception via binary labels and residual tokens, and then trains the model to generate detailed semantic explanations.

Prediction Objective. The token mask decoder aggregates visual tokens from multiple hierarchical levels of Deepstack Meng et al. (2024) and predicts whether each token belongs to the residual-salient region. We supervise this token-level prediction with Binary Cross-Entropy with Logits:

$$\mathcal{L}_{\text{Visual}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \sigma(\hat{y}_i) + (1-y_i)\log\left(1-\sigma(\hat{y}_i)\right)\right], \tag{8}$$

where N denotes the number of visual tokens, $y_i \in \{0,1\}$ indicates whether the i-th token belongs to the residual-salient token set $\mathcal{M}_{gt}$, $\hat{y}_i$ is the predicted logit, and $\sigma(\cdot)$ denotes the sigmoid function.

Training Objectives. Our framework is optimized using a composite objective function that balances detection accuracy, explanatory quality of low-level artifacts. $\mathcal{L}_{Text}$) is the standard auto-regressive language modeling loss for generating the next token in the response sequence, supervising the model to produce the subvisual forgery semantics. The final loss function is a weighted sum:

$$\mathcal{L}_{total} = \lambda_{Text}\mathcal{L}_{Text} + \lambda_{Visual}\mathcal{L}_{Visual} \tag{9}$$

where λ_{Text}, and λ_{Visual} are hyper-parameters balancing the two tasks, $\mathcal{L}_{total}$ is the total objective.

Two-Stage Training Strategy

Inspired by Lin et al. (2025), we devise a two-stage strategy to stably and progressively inject forgery knowledge:

Stage 1: Binary Classification Warm-up. Initially, the LLM decoder is frozen, and we only tune the vision encoder, the linear connector and the token decoder. The model is trained exclusively on the binary classification task using a large-scale dataset of real and fake images with naive text supervision(real or fake). This stage establishes a fundamental discrimination capability, enabling the encoder to distinguish basic real/fake patterns before learning complex semantic explanations.

Stage 2: Instruction Tuning with Semantic Supervision. In this stage, we unfreeze the LLM decoder, freeze the vision encoder and introduce the constructed benchmark for instruction tuning.

The model is trained to generate detailed forgery-related explanations ($\mathcal{L}_{Text}$) while maintaining to predict residual tokens ($\mathcal{L}_{Visual}$). For real images, we employ fine-grained, semantic object-level captions generated by the MLLM as supervision signals.

EXPERIMENT

Model	VA-VAE			Qwen-Image			SDXL		
	SemAP	SemAR	SemF1	SemAP	SemAR	SemF1	SemAP	SemAR	SemF1
LLaVa-OV-7B Li et al. (2024)	74.9	49.9	59.8	75.1	49.1	59.4	75.4	47.9	58.5
InternVL3.5-8B Zhu et al. (2025)	80.2	26.3	39.4	80.2	27.6	40.8	80.3	26.9	40.1
InternVL3.5-14B Zhu et al. (2025)	79.8	59.3	67.8	79.8	57.7	66.8	79.8	58.7	67.3
Qwen2.5-VL-7B Bai et al. (2025b)	79.7	54.7	64.7	79.5	55.6	65.3	79.5	54.5	64.5
Qwen2.5-VL-72B Bai et al. (2025b)	80.3	27.0	40.3	80.3	27.0	40.2	80.2	26.5	39.7
Qwen3-VL-8B Bai et al. (2025a)	79.4	67.7	72.9	79.3	68.3	73.2	79.6	67.3	72.7
Qwen3-VL-32B Bai et al. (2025a)	79.9	63.3	70.5	79.8	64.2	70.9	79.8	64.0	70.9
GPT-4o Hurst et al. (2024)	80.0	53.0	63.7	75.4	67.6	71.2	75.1	69.9	72.4
Gemini-3-Flash DeepMind & Google (2024)	78.4	29.8	43.1	78.2	29.0	42.3	78.7	27.9	41.2
ForgeryTracer (Ours)	**82.4**	**84.6**	**83.5**	**88.7**	**88.1**	**88.4**	**88.5**	**87.2**	**87.8**

Table 1: **Comparative Performance on TRACE-Bench for SFS Analysis.** We evaluate how faithfully each model translates low-level artifacts into high-level forgery semantics across three generators. General-purpose MLLMs miss most artifacts, whereas ForgeryTracer bridges this semantic gap via semantic differencing, performing best on all metrics.

Our goal is to articulate subvisual forgery semantics, and this ability rests on whether a model can perceive low-level forgery at all. We therefore evaluate two coupled abilities on TRACE-Bench, subvisual semantic description as the target and forgery detection as the perceptual foundation it stands on, and further test cross-generator generalization on AIGI-Now subset Chen et al. (2025a). Following Tan et al. (2025a), we evaluate object-level semantic fidelity against the ground truth using SemAP, SemAR, and SemF1, while the low-level forgery perception capability is measured by accuracy. We instantiate ForgeryTracer on Qwen3-VL-8B-Instruct Bai et al. (2025a) with a lightweight linear layer as the auxiliary mask decoder, trained under the two-stage strategy. More details are provided in the Appendix.

COMPARISON ON TRACE-BENCH

Can MLLM Put Low-level Artifacts into SFS? Tab. 1 reports how well MLLMs describe subvisual forgery semantics. General-purpose MLLMs reach moderate precision but consistently low recall, with SemAR falling to 26.3 for InternVL3.5-8B and staying below 68 across all models, including GPT-4o and Gemini-3-Flash. ForgeryTracer attains balanced semantic precision and recall across all three reconstruction sources, indicating that SFS can be articulated once the underlying low-level cues are preserved.

Is the Description Grounded in Real Perception? As shown in Tab. 2, most MLLMs hover near chance on high-fidelity forgeries, with accuracy between 46% and 63%, indicating that the failure is not resolved by scale but reflects a difficulty in perceiving low-level artifacts at the representation level. ForgeryTracer raises detection accuracy to 89.4% and above, confirming that it perceives these artifacts, and its SFS is therefore grounded in preserved low-level perception rather than inferred from high-level priors.

LLM as Judge. As a complementary check, we reevaluate ForgeryTracer with an LLM judge that scores each bullet by whether it names the right object and artifact. Absolute scores drop for both methods under this stricter rubric, yet ForgeryTracer keeps a consistent advantage on every source, from $+10.5$ on VA-VAE to $+35.6$ on SDXL. The two metrics agree, so we use BERTScore as the main metric, which is as reliable here but far cheaper and faster to run.

IS THE PERCEPTION REAL OR A FINGERPRINT?

To evaluate cross-generator generalization on low-level artifacts, we use the Pixel-Artifact Subset from AIGI-Now Chen et al. (2025a). The subset emphasizes subtle low-level artifacts, such as

Model	VA-VAE	Qwen-Image	SDXL
LLaVa-OV-7B Li et al. (2024)	49.1	50.2	50.0
InternVL3.5-8B Zhu et al. (2025)	48.9	50.0	50.7
InternVL3.5-14B Zhu et al. (2025)	50.2	50.6	50.3
Qwen2.5-VL-7B Bai et al. (2025b)	52.1	51.3	52.2
Qwen2.5-VL-72B Bai et al. (2025b)	51.3	51.1	51.8
Qwen3-VL-8B Bai et al. (2025a)	51.2	50.6	51.3
Qwen3-VL-32B Bai et al. (2025a)	46.4	50.3	49.8
GPT-4o Hurst et al. (2024)	51.8	52.0	52.3
Gemini-3-Flash DeepMind & Google (2024)	58.4	58.7	63.3
ForgeryTracer (Ours)	**91.2**	**94.9**	**89.4**

Table 2: **AIGI Detection Performance on TRACE-Bench.** General MLLMs stay near chance, suggesting limited perceptual access to subvisual forgery. ForgeryTracer both detects these artifacts and articulates them as SFS, indicating that its descriptions are grounded in preserved low-level perception.

Model	Nano Banana	GPT-4o	Jimeng	Kling	Minimax	Flux Pro	Flux Krea	Flux Dev	Flux Kontext	Avg
NPR Tan et al. (2024b)	84.4	91.2	41.9	88.3	52.1	42.7	52.3	89.3	79.2	69.0
UnivFD Ojha et al. (2023)	51.0	53.8	50.7	51.4	50.5	51.9	51.0	63.6	57.5	52.0
FatFormer Liu et al. (2024a)	53.4	49.9	48.4	54.8	49.1	48.4	49.5	52.4	63.6	52.1
SAFE Li et al. (2025a)	99.3	99.7	50.0	99.6	49.8	49.9	51.2	99.8	98.8	77.5
C2P-CLIP Tan et al. (2025b)	50.0	49.8	49.8	50.7	50.3	52.0	50.1	52.6	60.7	51.7
AIDE Yan et al. (2024)	98.9	74.7	63.9	98.2	51.4	60.1	50.4	99.1	97.9	77.1
DRCT Chen et al. (2024a)	68.9	75.0	89.3	66.9	78.4	83.0	91.6	86.5	86.9	80.7
AlignedFor. Rajan et al. (2024)	88.0	50.8	82.1	69.6	56.7	65.5	59.5	78.6	65.0	68.4
CO-SPY Cheng et al. (2025)	74.1	78.9	83.0	88.3	77.9	77.3	89.1	88.5	65.1	80.2
BFree Guillaro et al. (2025)	69.6	59.2	75.6	86.1	56.2	71.8	59.5	74.8	77.1	70.0
AlignGeimni Chen et al. (2025a)	96.0	96.8	96.2	97.3	91.3	89.9	83.0	96.6	86.6	92.6
ForgeryTracer (Ours)	**94.4**	**94.0**	**93.6**	**93.5**	**93.0**	**93.2**	**94.5**	**93.6**	**92.5**	**93.6**

Table 3: **Cross-generator Detection Performance on AIGI-Now Benchmark Chen et al. (2025a)**. On unseen generators, ForgeryTracer sustains uniformly high accuracy (92.5 to 94.5, avg 93.6), and additionally produces SFS for the AIGIs. This stability indicates that its SFS rests on transferable low-level forgery perception rather than fitting the training generators.

frequency anomalies and localized blending inconsistencies. Built from nine post-2024 state-of-the-art generators, including six closed-source systems, it provides a challenging out-of-distribution testbed. As shown in Tab. 3, ForgeryTracer achieves superior detection accuracy over baseline methods while producing SFS for low-level artifacts, despite no exposure to these generator distributions during training. This suggests that ForgeryTracer perceives forgery through a generalizable low-level representation rather than the reconstruction bias of its training VAEs. These generated images have no paired real source, so ground-truth SFS cannot be scored on them, and we report qualitative SFS in the Appendix.

Ablation Study

Training Strategy. ForgeryTracer pursues a single objective, producing SFS while keeping high accuracy, since a subvisual description is only meaningful when it rests on low-level forgery correctly perceived. Tab. 4 follows both abilities across the training stages against this objective. Stage 1 establishes perception through binary supervision, but adding semantic supervision alone in Stage 2 lowers detection on every source, dropping from 81.4 to 67.6 on Qwen-Image, as its semantic targets drown out the sparse binary signal. Dense residual-token supervision rebalances the two by injecting a low-level visual signal at the feature level, restoring detection beyond Stage 1 (91.2, 94.9, and 89.4) while semantic quality stays near its peak, and its denser gradient also drives faster convergence than the sparse binary signal alone. This is the core evidence for our design, as SFS is not acquired at the expense of perception but grounded on a perception that dense supervision keeps intact. Moreover, a larger residual-token ratio $k\%$ further trades a slight semantic drop for stronger perception, giving a tunable operating point.

Generator	Method	Acc. (%)↑	SemAP*↑	SemAR*↑	SemF$_1$*↑
VA-VAE	BASELINE	51.2	68.8	14.8	24.3
	Stage$_1$	61.3	–	–	–
	Stage$_2$	56.1	**88.5**	82.1	**85.2**
	Token$_{5\%}$	65.8	83.6	**86.0**	84.8
	Token$_{25\%}$	<u>88.8</u>	82.0	84.9	83.4
	Token$_{50\%}$	**91.2**	82.4	84.8	83.6
SDXL	BASELINE	51.3	67.5	16.9	27.0
	Stage$_1$	<u>86.6</u>	–	–	–
	Stage$_2$	81.6	**89.0**	80.9	74.8
	Token$_{5\%}$	81.3	84.6	**85.1**	**84.8**
	Token$_{25\%}$	<u>94.2</u>	83.3	84.0	83.7
	Token$_{50\%}$	**94.9**	83.6	84.3	83.9
Qwen-Image	BASELINE	50.6	67.3	17.6	27.9
	Stage$_1$	81.4	–	–	–
	Stage$_2$	67.6	**89.1**	82.8	**85.8**
	Token$_{5\%}$	84.7	84.2	**86.0**	85.1
	Token$_{25\%}$	<u>88.4</u>	82.4	84.8	83.6
	Token$_{50\%}$	**89.4**	82.7	85.0	83.8

Table 4: **Ablation Study of the Supervision Strategies.** SFS supervision alone (*Stage*$_2$) lowers detection, while residual-token supervision (*Token*$_{k\%}$) restores it beyond *Stage*$_1$ and retains semantic quality. **Bold** and <u>underline</u> indicate the best and second-best results. * indicates that semantic metrics are evaluated only on samples with correct detection.

Generator	Method	SemAP	SemAR	SemF1
VA-VAE	Baseline	38.9	33.0	35.7
	Ours	**45.6**	**46.8**	**46.2**
Qwen-Image	Baseline	38.6	33.2	35.7
	Ours	**53.5**	**53.2**	**53.3**
SDXL	Baseline	21.4	13.3	16.4
	Ours	**52.3**	**51.7**	**52.0**

Table 5: **LLM-as-Judge Evalution on TRACE-Bench for SFS.** ForgeryTracer keeps a consistent advantage over the baseline (Qwen3-VL-8B-Instruct Bai et al. (2025a)) across all three reconstruction sources.

CONCLUSION

This paper presents the first study on semanticizing low-level forgery artifacts. We introduce Subvisual Forgery Semantics (SFS), which expresses forgery traces below salient perception as their effects on object-level semantics, and capture it by deconstructing pixel differences back into the semantic differencing process that produces them. On this basis, we develop SemDiff, a multi-agent pipeline that produces reliable SFS annotations at scale, and construct TRACE-Bench, the first benchmark assessing MLLMs on SFS, where we further present ForgeryTracer, a dual-supervision baseline demonstrating that SFS can be learned without sacrificing detection accuracy. Our work makes low-level forgery evidence not only detectable but explainable, opening a low-level perspective for interpretable AIGI forensics.

REFERENCES

Stability AI. Stable diffusion 3.5 model family. Stability AI website, October 2024. URL https://stability.ai/news/introducing-stable-diffusion-3-5. Accessed July 23, 2025.

Anthropic. Claude 3 model family. Anthropic website, March 2024. URL https://www.anthropic.com/news/claude-3-family. Accessed July 23, 2025.

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*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, Yimeng Wang, Kai Yu, Wenxuan Chen, Ziwei Feng, Zijian Gong, Jianzhuang Pan, Yi Peng, Rui Tian, Siyu Wang, Bo Zhao, Ting Yao, and Tao Mei. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, May 2025. URL https://arxiv.org/abs/2505.22705. Accessed July 23, 2025.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024a.

Ruoxin Chen, Jiahui Gao, Kaiqing Lin, Keyue Zhang, Yandan Zhao, Isabel Guan, Taiping Yao, and Shouhong Ding. Task-model alignment: A simple path to generalizable ai-generated image detection. *arXiv e-prints*, pp. arXiv–2512, 2025a.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, et al. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025b.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, and Zhenda Xie. Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811, January 2025c. URL https://arxiv.org/abs/2501.17811. Includes both 1B and 7B variants; released as open-source under MIT License; accessed July 23, 2025.

Yize Chen, Zhiyuan Yan, Guangliang Cheng, Kangran Zhao, Siwei Lyu, and Baoyuan Wu. X2-dfd: A framework for explainable and extendable deepfake detection. *arXiv preprint arXiv:2410.06126*, 2024b.

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

Google DeepMind and Google. Gemini 1.5 model family. Google Blog, February 2024. Accessed July 23, 2025.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18685–18694, 2025.

Tai-Ming Huang, Wei-Tung Lin, Kai-Lung Hua, Wen-Huang Cheng, Junichi Yamagishi, and Jun-Cheng Chen. Thinkfake: Reasoning in multimodal large language models for ai-generated image detection. *arXiv preprint arXiv:2509.19841*, 2025a.

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, 2025b.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2405–2414, 2025a.

Tianxiao Li, Zhenglin Huang, Haiquan Wen, Yiwei He, Shuchang Lyu, Baoyuan Wu, and Guangliang Cheng. Raidx: A retrieval-augmented generation and grpo reinforcement learning framework for explainable deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11746–11755, 2025b.

Yixuan Li, Xuelin Liu, Xiaoyang Wang, Bu Sung Lee, Shiqi Wang, Anderson Rocha, and Weisi Lin. Fakebench: Probing explainable fake image detection via large multimodal models. *IEEE Transactions on Information Forensics and Security*, 2025c.

Yixuan Li, Yu Tian, Yipo Huang, Wei Lu, Shiqi Wang, Weisi Lin, and Anderson Rocha. Fakescope: Large multimodal expert model for transparent ai-generated image forensics. *arXiv preprint arXiv:2503.24267*, 2025d.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10770–10780, 2024a.

Jiawei Liu, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. *arXiv preprint arXiv:2410.10238*, 2024b.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems*, 37:23464–23487, 2024.

Midjourney, Inc. Midjourney. `https://www.midjourney.com/`, 2025. Version 7 (released 2025-04-04).

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24480–24489, 2023.

OpenAI. Dall·e 3. OpenAI website, September 2023. URL `https://openai.com/dall-e-3`. Accessed July 23, 2025.

OpenAI. Openai o3 reasoning model family. OpenAI website, December 2024. URL `https://openai.com/index/introducing-o3-and-o4-mini`. Accessed July 23, 2025.

OpenAI. Gpt-5 system card. System card, OpenAI, August 2025. URL `https://cdn.openai.com/gpt-5-system-card.pdf`. Version dated August 13, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. *arXiv preprint arXiv:2410.11835*, 2024.

Robin Rombach et al. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12105–12114, 2023.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5052–5060, 2024a.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28130–28139, 2024b.

Chuangchuang Tan, Xiang Ming, Jinglu Wang, Renshuai Tao, Bin Li, Yunchao Wei, Yao Zhao, and Yan Lu. Semantic visual anomaly detection and reasoning in ai-generated images. *arXiv preprint arXiv:2510.10231*, 2025a.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025b.

Chuangchuang Tan, Jinglu Wang, Xiang Ming, Renshuai Tao, Yunchao Wei, Yao Zhao, and Yan Lu. Forenx: Towards explainable ai-generated image detection with multimodal large language models. *arXiv preprint arXiv:2508.01402*, 2025c.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Chuanbiao Song, Senyuan Shi, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas: Generalizable deepfake detection via pattern-aware reasoning. *arXiv preprint arXiv:2508.21048*, 2025d.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9568–9578, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wenhao Wang, Longqi Cai, Taihong Xiao, Yuxiao Wang, and Ming-Hsuan Yang. Scaling laws for deepfake detection. *arXiv preprint arXiv:2510.16320*, 2025.

Haiquan Wen, Yiwei He, Zhenglin Huang, Tianxiao Li, Zihan Yu, Xingru Huang, Lu Qi, Baoyuan Wu, Xiangtai Li, and Guangliang Cheng. Busterx: Mllm-powered ai-generated video forgery detection and explanation. *arXiv preprint arXiv:2505.12620*, 2025a.

Haiquan Wen, Tianxiao Li, Zhenglin Huang, Yiwei He, and Guangliang Cheng. Busterx++: Towards unified cross-modal ai-generated content detection and explanation with mllm. *arXiv preprint arXiv:2507.14632*, 2025b.

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. *arXiv preprint arXiv:2503.14905*, 2025c.

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.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024.

Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. Dˆ 3: scaling up deepfake detection by learning from discrepancy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23850–23859, 2025a.

Zheng Yang, Ruoxin Chen, Zhiyuan Yan, Ke-Yue Zhang, Xinghe Fu, Shuang Wu, Xiujun Shu, Taiping Yao, Shouhong Ding, and Xi Li. All patches matter, more patches better: Enhance ai-generated image detection via panoptic patch learning. *arXiv preprint arXiv:2504.01396*, 2025b.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX

In this supplementary material, we provide additional technical details and experimental results to support the main text. Specifically, we detail the experimental setup and implementation, elaborate on the evaluation metrics, describe the dataset construction process alongside its quality validation, and present extended quantitative and qualitative evaluations.

DATASET & IMPLEMENTATION

Dataset. We conduct extensive experiments on both the public AIGI detection dataset and our proposed benchmark. Specifically, we utilize AIGI-Now Chen et al. (2025a) to evaluate the broad generalization of our model on contemporary AIGC detection tasks. To quantitatively assess the capability of the MLLM on SFS, we employ our constructed TRACE-Bench, which pairs high-fidelity forgeries with detailed low-level semantic explanations.

Implementation Details. We instantiate our framework using Qwen3-VL-8B-Instruct Bai et al. (2025a) as the base MLLM. The auxiliary mask token decoder is implemented as a lightweight linear layer to provide the dense visual supervision with minimal computational overhead. During training, we strictly follow the two-stage partial parameter fine-tuning strategy, initial binary classification warm-up followed by instruction tuning with forgery low-level semantic supervision. The model is optimized utilizing the AdamW optimizer with a learning rate of 1×10^{-5}. The training process spans a single epoch for each respective stage. Specifically, the batch size is set to 64 in the first stage and reduced to 16 in the second stage. To process the visual representations, we apply a mean subtraction operation to the features extracted from multiple hierarchical levels of the network stack. Furthermore, the loss weights for both the textual and the visual supervision are uniformly set to 1. We use a standardized input resolution of 512×512. The model is trained on 8 NVIDIA H100 GPUs leveraging DeepSpeed ZeRO-3 optimization.

Dataset	Accuracy at Different JPEG Quality (QF)			
	Uncompressed	QF = 90	QF = 80	QF = 70
Average	**91.8**	**88.2**	**89.0**	**84.3**

Table 6: **Detection Robustness under JPEG Compression.**

JPEG ROBUSTNESS.

We evaluate the robustness of ForgeryTracer under varying JPEG quality factors. As illustrated in Tab. 6, the accuracy of forgery detection is largely sustained, indicating that our method maintains a certain degree of robustness against compression artifacts.

EVALUATION METRICS

To comprehensively evaluate the performance of both detection and explanation, we adopt a multi-dimensional metric suite. For the binary AIGC detection task, we report the standard Accuracy (Acc). To comprehensively evaluate the model's capabilities in visual anomaly detection and reasoning, we employ SemAP, SemAR and SemF1 Tan et al. (2025a) as our primary metric. Unlike traditional AP, which relies on bounding box IoU for spatial matching, SemAP utilizes BERTScore-based Zhang* et al. (2020) semantic similarity to match the model-generated forgery descriptions and logical reasoning—against the ground truth. This approach overcomes the limitations of exact word matching and objectively quantifies the model's interpretability and logical coherence.

Let $\mathcal{G} = \{g_1, g_2, \ldots, g_M\}$ denote the set of ground-truth semantic anomaly descriptions, and $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$ denote the set of model-generated predictions. We define the semantic similarity between a prediction p_i and a ground-truth g_j as:

$$S(p_i, g_j) = \text{BERTScore}(p_i, g_j) \tag{10}$$

A prediction p_i is considered a true positive if there exists an unmatched ground-truth g_j such that $S(p_i, g_j) \geq \tau$, where τ is a predefined semantic similarity threshold. The matching process is resolved via maximum bipartite matching to ensure optimal alignment. Based on the matched set $\mathcal{M}_\tau = \{(p_i, g_j) \mid S(p_i, g_j) \geq \tau\}$, the number of true positives, false positives, and false negatives at threshold τ are calculated as:

$$\text{TP}_\tau = |\mathcal{M}_\tau|, \quad \text{FP}_\tau = |\mathcal{P}| - |\mathcal{M}_\tau|, \quad \text{FN}_\tau = |\mathcal{G}| - |\mathcal{M}_\tau| \tag{11}$$

The Semantic Precision and Semantic Recall at threshold τ are subsequently formulated as follows:

$$\text{SemP}_\tau = \frac{\text{TP}_\tau}{|\mathcal{P}|}, \quad \text{SemR}_\tau = \frac{\text{TP}_\tau}{|\mathcal{G}|} \tag{12}$$

The Semantic F1 Score combines precision and recall to provide a balanced measure of semantic alignment:

$$\text{SemF1}_\tau = 2 \cdot \frac{\text{SemP}_\tau \cdot \text{SemR}_\tau}{\text{SemP}_\tau + \text{SemR}_\tau} \tag{13}$$

Finally, Semantic Average Precision is computed as the area under the Semantic Precision-Recall curve, and Semantic Average Recall is computed as the maximum semantic recall averaged over a set of predefined thresholds $\mathcal{T}$:

$$\text{SemAP} = \int_0^1 \text{SemP}(r)\, dr, \quad \text{SemAR} = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \text{SemR}_\tau \tag{14}$$

Source	Images	Bullets (before)	Dropped bullets			Bullets (after)
			Qwen-only	Intern-only	Both	
VA-VAE	5,000	29,930	1,223	334	330	28,043
SDXL	5,000	30,323	1,558	369	440	27,956
Qwen-Image	5,000	29,752	1,173	231	293	28,055
Total	**15,000**	**89,005**	**3,954**	**934**	**1,063**	**84,054**

Table 7: Bullet-level QC filtering statistics. Each GT bullet is independently judged by two VLMs (Qwen3-VL-235B and InternVL3-78B); a bullet is removed if *either* judge labels it `drop` (hallucinated object or non-observable artifact). Image count is preserved across all splits.

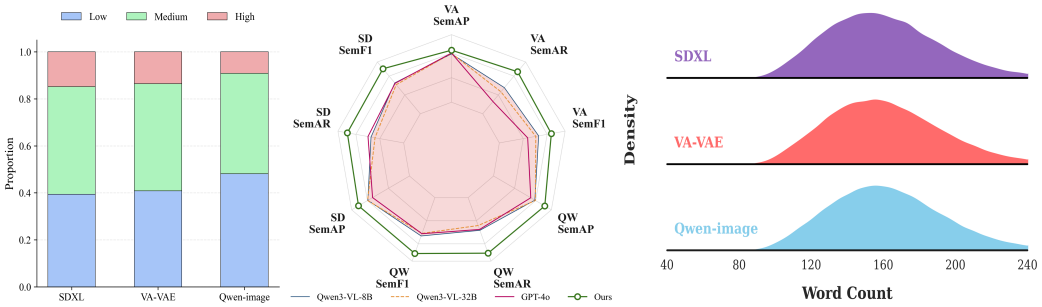
Source	Before QC		After QC	
	Bullets	Invalid rate	Bullets	Invalid rate (Δ)
VA-VAE	3,000	11.6%	2,801	7.0% ($\downarrow$4.6pp)
SDXL	3,014	13.2%	2,768	8.1% ($\downarrow$5.2pp)
Qwen-Image	2,984	12.2%	2,817	7.6% ($\downarrow$4.6pp)
Total	**8,998**	**12.4%**	**8,386**	**7.5% ($\downarrow$4.8pp)**

Table 8: Human-simulated QC audit: invalid-bullet rate before and after dual-VLM filtering, measured by Qwen3-VL-235B on 500 randomly sampled images per VAE source. Δ = percentage-point reduction in invalid bullets after QC.

TRACE-BENCH

We investigate the distributions of various forgery levels and fidelity scores across the dataset. As illustrated in Fig. 5, the number of samples with high degrees of forgery exhibits a declining trend as the generator undergoes optimization (SDXL $\rightarrow$ VA-VAE $\rightarrow$ Qwen-image), an observation that aligns with intuitive expectations. To ensure the high quality of our benchmark, we implement a two-stage filtering and validation protocol. We first introduce a dual-judge pipeline using two independent MLLMs (Qwen3-VL-235B-A22B Bai et al. (2025a) and InternVL-78B Zhu et al. (2025)) to evaluate whether the extracted visual evidence strictly aligns with each subvisual forgery description. An instance is immediately discarded if flagged by either model (i.e., strict consensus). Furthermore, we conduct human sampling inspections on the data before and after this process. As shown in Table 7 and 8, the prompt-based judging effectively purges noisy annotations, further elevating the overall dataset quality.

Dataset Distribution. To construct a diverse benchmark, we utilize the widely adopted COCO 2017 dataset Lin et al. (2014) as the data source. To generate the corresponding high-fidelity reconstructed images, which contain forgery artifacts, we employ three distinct visual autoencoders: SDXL Podell et al. (2023) VAE, Qwen-Image Wu et al. (2025) VAE, and VA-VAE Yao et al. (2025). By processing the original images through these different latent decoders, the resulting pairs inherently capture a diverse distribution of forgery-related low-level digital fingerprints. We employ Qwen3-VL-32B-Instruct Bai et al. (2025a) as the primary agent to execute the three-stage data construction process, which formulates the semantic representation of low-level forgery artifacts in reconstructed images. To formulate the low-level artifacts of reconstructed images into structured subvisual forgery semantics while mitigating fixed model biases or stylistic preferences, we deploy a diverse pool of MLLM agents across the three-stage data construction pipeline Bai et al. (2025a); Zhu et al. (2025); Li et al. (2024), with the statistics shown in Fig 5. For real images, we construct captions using specific prompts that describe the visual details of the constituent semantic objects.



(a) Severity Level Distribution **(b) Performance Comparison** **(c) Word Length Distribution**

Figure 5: **TRACE-Bench Statistics.** (a) Distribution of the severity. (b) Performance of representative MLLMs on TRACE-Bench. (c) Length distribution of the textual representation by different VAEs.

Data Splitting. The curated dataset is meticulously partitioned to evaluate both in-domain learning and out-of-domain generalization. For the training phase, we pair the COCO 2017 train split with its corresponding reconstructions generated exclusively by VAEs. For the validation phases, we utilize the COCO 2017 val split reconstructed by three VAEs to assess performance. To rigorously evaluate the generalization of the cross-generator, we utilize VAE-reconstructed images from Qwen-Image and SDXL as the training set in a subset of experiments and designate all reconstructions produced by VA-VAE as a strictly held-out test subset. This strategic partition ensures the model is evaluated on low-level artifacts from an unseen latent decoder, preventing mere overfitting to specific VAE fingerprints.

Setting	Nano Banana	GPT-4o	Jimeng	Kling	Minimax	Flux Pro	Flux Krea	Flux Dev	Flux Kontext	Avg.
Stage-1	91.3	95.4	95.0	93.3	90.1	91.7	89.2	94.0	95.0	92.7
Stage-2	93.4	94.4	94.0	93.1	90.9	90.8	90.0	93.4	93.6	92.6

Table 9: Performance comparison of different training strategies on AIGI-Now Chen et al. (2025a).

Table 10: AIGI Detection Accuracy (Acc) of our method under In-Domain and Out-of-Domain (OOD) settings. The evaluation confirms that the SFS output relies on a thoroughly generalized forgery representation rather than dataset-specific overfitting, yielding trustworthy artifact descriptions under both In-Domain and OOD setups.

	VA-VAE	Qwen-Image	SDXL	AIGI-Now-Pixel
Setting	*OOD*	*In-Domain*	*In-Domain*	*OOD*
Acc	81.8	94.5	97.0	92.6

Additional Experiment

Token Mask Decoder Enables Efficient Training

We perform an additional ablation study on the AIGI-Now benchmark Chen et al. (2025a) on ForgeryTracer without token mask decoder. Specifically, we extend the training duration of the first stage to thoroughly capture the feature discrepancies between real and generated images (5 epoch), and subsequently introduce the second stage to equip the MLLM with the capability to semantically represent low-level artifacts (1 epoch).

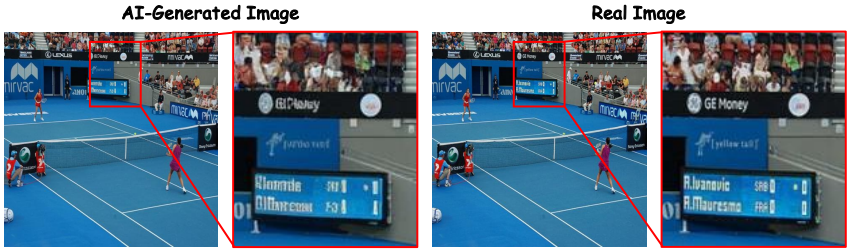
As shown in Tab. 9, extended Stage-1 training can eventually capture low-level artifacts, indicating that the auxiliary token mask decoder provides dense visual supervision that substantially accelerates convergence. More importantly, introducing Stage-2 semantic decoding preserves the binary detection performance intact, confirming that ForgeryTracer acquires the ability to capture and verbalize SFS *without sacrificing* its detection and perception capabilities.

SFS Builds on Generalizable Perception

We train on reconstructed images from the VAEs of Qwen-Image and SDXL, and evaluate on out-of-domain reconstructions from VA-VAE as well as the *Pixel Subset* of AIGI-Now Chen et al. (2025a). As shown in Tab. 10, ForgeryTracer maintains consistent performance on both unseen VAEs and unseen generators, indicating that it learns a forgery-aware perceptual representation that generalizes beyond specific reconstruction pipelines. This validates our core design: SFS is not memorized from generator-specific patterns, but grounded on a generalizable perception of low-level artifacts, mirroring the forensic principle of ***perceiving before explaining***.

Additional Visual Results

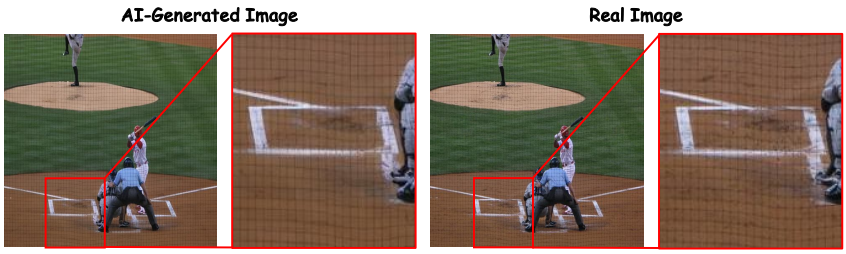
We present visualized annotation examples as shown in Fig. 6, 7, 8 and 9 from TRACE-Bench and we also show the predicted sample from AIGI-Now Chen et al. (2025a). The MLLM receives the original semantics from the authentic image alongside the forged semantics from the generated image. Conditioned on the pixel difference prompt, the model translates pixel-level difference into object-level semantic differences, thereby achieving the semantic representation of low-level artifacts. This methodology effectively facilitates both the evaluation and the enhancement of the capability of the MLLM to semantically interpret low-level forgery artifacts without sacrificing the detection performance.



- **Scoreboard text:** Text on scoreboard is blurred and illegible; 'R.Ivanovic' and 'R.Muguruza' names are partially merged. Semantic impact: Loss of match context and player identification.
- **Net mesh:** Fine net threads are smoothed into a uniform gray band; loss of structural detail. Semantic impact: Reduced perception of net depth and material texture.
- **Player in red outfit:** Soft, white halo around player's limbs and racket; loss of sharp silhouette. Semantic impact: Reduced clarity of player's motion and equipment.
- **Ball boy uniforms:** Red uniforms appear muted and less vibrant; logo details are faint. Semantic impact: Reduced brand visibility and uniform distinctiveness.
- **Spectator crowd:** Subtle grid-like noise in dense crowd areas, especially in upper stands. Semantic impact: Minor disruption to crowd realism, but no major identity loss.
- **Tennis ball:** Ball appears as a soft yellow smudge; loss of surface texture and precise position. Semantic impact: Reduced accuracy in tracking ball trajectory.
- **Court lines:** White lines exhibit slight blurring at edges, especially near net. Semantic impact: Minor reduction in court boundary precision.

Figure 6: Qualitative visualization example 1. SeD identifies subtle forgery effects in multiple semantic objects, including blurred scoreboard text, smoothed net mesh, softened player silhouettes, muted uniforms, noisy crowds, imprecise tennis ball appearance, and blurred court lines. These examples illustrate how low-level traces left by AIGI formation can affect semantic content such as identity recognition, spatial structure, material perception, and motion understanding.

To evaluate different paradigms of forgery semantic explanation, we conduct a controlled comparison between outcome-based and process-based forgery semantic outputs. A general-purpose vision-language model (VLM) is adopted as an automatic judge to perform pairwise preference evaluation: given the same input and two candidate explanations, the VLM selects the one that better localizes and interprets Subvisual Forgery Semantics. Experimental results show that our process-based method is consistently preferred over the outcome-based counterpart, achieving a higher preference rate and demonstrating the advantage of process-based forgery semantic modeling (10).



- **Home plate area:** Blurred white chalk lines and plate boundary; semantic impact: compromised spatial precision of batter's box and strike zone definition.
- **Pitcher's uniform:** Loss of fine pinstripe texture and fabric weave; semantic impact: reduced visual distinction between uniform and background, diminishing player identification cues.
- **Batter's helmet:** Soft, blurred boundary between helmet and hair; semantic impact: compromised headgear identification and potential misinterpretation of player safety gear.
- **Catcher's gear:** Smearing of chest protector and mask contours; semantic impact: reduced clarity of protective equipment, affecting safety assessment and player role recognition.
- **Umpire's shirt:** Duller blue hue with reduced chromatic contrast; semantic impact: diminished visibility of official role and potential confusion with other personnel.\n- Baseball bat: Loss of grip texture and barrel grain; semantic impact: reduced authenticity of equipment, affecting sports context fidelity.
- **Field turf:** Subtle grid-like noise in grass texture; semantic impact: minor disruption of natural surface appearance, potentially misleading in forensic texture analysis.

Figure 7: Qualitative visualization example 2. The AI-generated baseball scene appears globally plausible, but SeD reveals subtle object-level effects caused by low-level forgery traces. Compared with the real image, the generated image exhibits blurred home-plate boundaries, weakened uniform textures, softened helmet contours, smeared protective gear, reduced color contrast in the umpire's shirt, degraded bat texture, and grid-like noise on the turf. These effects illustrate how low-level artifacts can influence spatial precision, player identification, equipment recognition, and sports-scene understanding.



- **Chain-link fence: Fine wire mesh pattern is blurred and partially fused; loss of individual wire definition; semantic impact: degradation of background depth and structural detail.**
- Batter's red jersey: Red hue appears muted and less vibrant; loss of subtle tonal gradients in fabric texture; semantic impact: reduced visual prominence of the batter's uniform.
- Catcher's mitt: Soft, gray halo along the glove's contour; reduced sharpness at the boundary; semantic impact: diminished tactile realism and object separation from background.
- White car in background: Window reflections and body panel lines are smeared; loss of fine surface texture; semantic impact: reduced environmental context and realism.
- Batter's helmet: Surface texture and minor scratches are smoothed; loss of specular highlights; semantic impact: reduced material authenticity and depth perception.
- Grass behind fence: Green tones appear slightly washed out; reduced chromatic contrast; semantic impact: minor loss of natural environment vibrancy.
- Dirt ground surface: Small pebbles and texture variations are blurred into a uniform surface; semantic impact: loss of ground realism and tactile detail.

Figure 8: Qualitative visualization example 3. Although the AI-generated baseball image remains globally plausible, SeD reveals multiple subvisual object-level effects when compared with the real image. Low-level forgery traces appear as blurred and partially fused chain-link fence patterns, muted jersey color, softened catcher-mitt boundaries, smeared car reflections, smoothed helmet texture, washed-out grass tones, and over-smoothed dirt surfaces. These examples show how subtle low-level artifacts can be translated into semantic evidence affecting background depth, object separation, material authenticity, environmental realism, and scene fidelity.

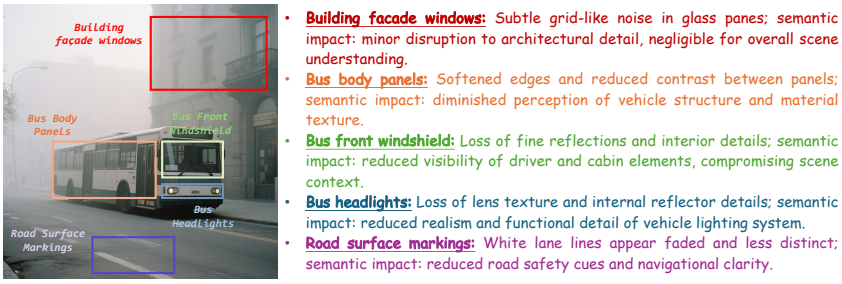


Figure 9: Qualitative visualization predicted sample from AIGI-Now Chen et al. (2025a). Low-level forgery traces appear as lost fine reflections and interior details on the front windshield, softened edges and reduced contrast across body panels, faded road markings, subtle grid-like noise in building facade windows, and smoothed lens textures in the bus headlights.The example shows ForgeryTracer can output SFS on the unseen generator. Building upon the generated explanations, we further employ a general-purpose VLM for visual grounding to localize and interpret Subvisual Forgery Semantic.

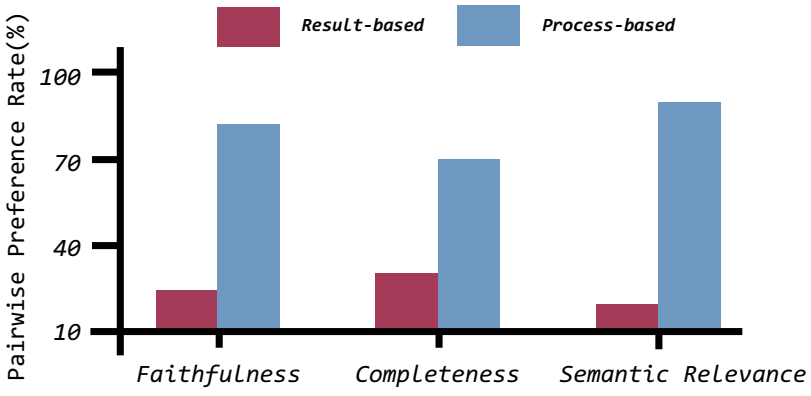


Figure 10: VLM's preferences for result-based and process-based SFS