UNLOCKING NATIVE FORENSIC REPRESENTATION IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have emerged as a promising paradigm for AI-generated image (AIGI) detection by leveraging high-level semantic understanding. However, reliable detection also depends on subtle low-level forensic evidence, motivating growing efforts to equip MLLMs with stronger forensic perception. This work demonstrates that strong forensic discrimination can be achieved using only the native visual representations of pretrained MLLMs. We introduce **Native Forensic Connection (NFC)**, a remarkably simple framework that directly connects native forensic representations to the language model. NFC is inspired by our observation that low-level forensic information is already highly discriminative in intermediate visual representations, but becomes substantially less accessible along the standard vision-language pathway. With NFC, MLLMs can access native low-level forensic information without adapting the pretrained vision encoder. To better evaluate low-level forensic perception, we further construct **SemClean-Bench** through a reverse filtering strategy that removes AIGIs with salient high-level semantic anomalies, shifting the evaluation toward low-level artifacts. Extensive experiments across diverse AIGI benchmarks and SemClean-Bench demonstrate strong and generalizable performance. These results show that strong forensic perception can be achieved by effectively accessing native visual representations without relearning, highlighting representation access as a simple and complementary direction for forensic MLLMs.

1 INTRODUCTION

The rapid proliferation of Artificial Intelligence Generated Content (AIGC) raises unprecedented concerns about visual misinformation and digital trust Rombach et al. (2022); Wu et al. (2025); OpenAI (2023). Recently, Multimodal Large Language Models (MLLMs) have emerged as a promising paradigm for AI-generated image (AIGI) detection by leveraging high-level semantic understanding Tan et al. (2025a); Liu et al. (2024); Tan et al. (2025c). Unlike conventional binary classifiers, MLLMs can identify semantic inconsistencies and provide interpretable explanations for their authenticity judgments. However, reliable detection also depends on subtle low-level forensic evidence left by the image generation process Tan et al. (2024b;a); Cheng et al. (2025). This has motivated growing efforts to strengthen forensic perception in MLLMs Li et al. (2025b); Huang et al. (2025a); Wang et al. (2026). Existing approaches commonly introduce specialized forensic encoders or adapt the visual backbone toward artifact perception Gao et al. (2026); Chen et al. (2025a); Lin et al. (2025). While effective, this paradigm treats stronger forensic perception primarily as a problem of learning or enhancing visual representations. In this work, we revisit this premise and ask a more fundamental question: **Can MLLMs perform AIGI detection by directly utilizing their native visual representations?**

Our study reveals two observations in Fig. 1. **First**, without any forensic-specific vision adaptation, intermediate representations of frozen MLLM vision encoders already contain highly discriminative information for distinguishing real and generated images. **Second**, this discriminability exhibits a clear layer-wise structure: representations that are highly separable within the vision encoder are not necessarily those effectively exposed to the language model, and their forensic separability becomes substantially weaker along the standard vision-language pathway. These observations suggest that

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.

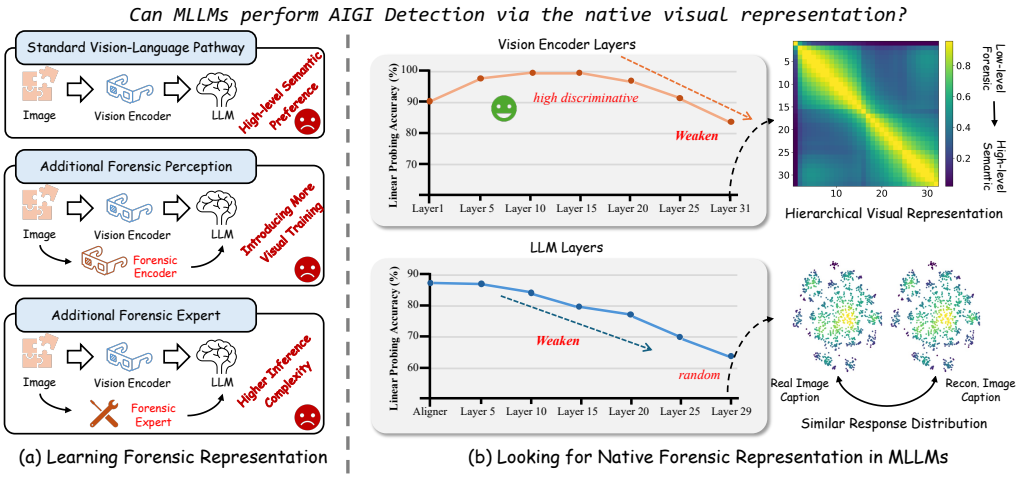


Figure 1: **From Learning Forensic Representation to Looking for Native Forensic.** (a) Learning Forensic Representation: Existing forensic MLLMs acquire low-level forensic perception by relearning visual representation, learning additional encoders or introducing expert model. Here we study a more fundamental problem: *Can MLLMs perform AIGI Detection via the native visual representation?* (b) Looking for Native Forensic Representation in MLLMs: Illustrated by the hierarchical visual representations, the vision encoder (top) exhibits high forensic discriminative power in its early layers. However, this capability progressively weakens along the vision-language path into the LLM (bottom), ultimately collapsing into a similar distribution at the language end.

useful forensic information may already be encoded in the pretrained visual backbone, while remaining under-utilized by the language model.

Motivated by this finding, we introduce **Native Forensic Connection (NFC)**, which directly connects forensic-sensitive intermediate visual representations to the language model through a lightweight alignment module. NFC keeps the pretrained vision encoder entirely frozen, requiring neither an auxiliary forensic encoder nor vision-side adaptation. Rather than relearning the visual representation space, NFC provides the language model with direct access to native information that is weakly utilized by the standard vision-language pathway.

A remaining challenge is evaluating such forensic capability when obvious semantic shortcuts are reduced. Existing AIGI benchmarks may contain generated images with salient high-level semantic anomalies that MLLMs can identify through semantic reasoning alone. We therefore construct **SemClean-Bench**, a fake-side diagnostic benchmark based on reverse semantic filtering. Multiple general-purpose VLMs are used to filter out generated images with salient semantic anomalies, retaining samples that are semantically plausible across different models. SemClean-Bench thus complements standard real/fake benchmarks by placing greater emphasis on recognizing generated images without obvious semantic shortcuts.

Extensive experiments on standard binary AIGI benchmarks and SemClean-Bench show that NFC achieves strong generalization across unseen generators and post-processing conditions while keeping the vision encoder frozen. The gains are particularly pronounced on SemClean-Bench, supporting the effectiveness of directly accessing native intermediate representations when salient semantic cues are reduced. Together, these results highlight **representation access** as a simple alternative to relearning the pretrained visual space for forensic perception.

Our main contributions are summarized as follows:

- **We reveal a representation-access gap in pretrained MLLMs for AIGI detection.** Frozen intermediate visual representations already contain strong information for real/fake discrimination, while such discriminability becomes substantially weaker along the standard vision-language pathway.

- **We introduce Native Forensic Connection (NFC).** NFC directly exposes forensic-sensitive native visual representations to the language model through a lightweight alignment module, while keeping the vision encoder entirely frozen and requiring no auxiliary forensic encoder.
- **We construct SemClean-Bench for diagnostic evaluation under reduced semantic shortcuts.** SemClean-Bench filters out AIGIs with salient high-level semantic anomalies, providing a fake-side diagnostic benchmark complementary to standard binary evaluation.

2 Related Work

MLLM-based AIGI Detection and Reasoning. Recent studies Wen et al. (2025a;b); Tan et al. (2025a); Zhu et al. (2026) extend AIGI detection beyond binary prediction toward localized and interpretable forensic analysis. ForgeryGPT, SIDA, and LEGION jointly explore authenticity prediction, forgery localization, and natural-language explanation Liu et al. (2024); Huang et al. (2025b); Kang et al. (2025). Other works, including ForenX, FakeScope, and AIGI-Holmes, further improve forensic grounding and explanation through artifact-oriented supervision, expert guidance, or preference alignment Tan et al. (2025b); Li et al. (2025a); Zhou et al. (2025). More recent approaches focus on the reasoning process itself: ThinkFake, Veritas, and VIGIL introduce structured reasoning or post-training objectives to improve generalization and evidence-based forensic inference Huang et al. (2025a); Tan et al. (2026); Li et al. (2026a). Together, these methods improve how MLLMs interpret and reason about forensic evidence; in contrast, we ask whether useful forensic information is already encoded in native visual representations and can be directly accessed without relearning.

Low-level Forensic Perception in MLLMs. Recent studies increasingly recognize that semantic reasoning alone is insufficient for AIGI detection and try to strengthen low-level forensic perception Zhou et al. (2025); Chen et al. (2025a). ForgeryGPT Zhang et al. (2024), FakeReasoning Gao et al. (2026), AlignGemini Chen et al. (2025a), and AIGI-Holmes Zhou et al. (2025) introduce dedicated forensic encoders or visual experts to provide artifact-specific evidence. Deep-VRM Lin et al. (2026) further learns a separate low-level visual pathway and injects artifact-specific representations into intermediate language layers, while Forensic-Chat Lin et al. (2025) explicitly adapts the native vision encoder toward artifact perception. TranX-Adapter Wang et al. (2026) instead studies how externally obtained artifact representations can be more effectively fused with semantic features inside MLLMs. Despite their different designs, these approaches primarily improve forensic perception by learning, introducing, or enhancing artifact-specific visual representations. In contrast, we investigate whether such forensic information is already encoded in the frozen native vision representations of pretrained MLLMs and how it can be directly connected to the language model.

3 Native Forensic Connection

Motivated by the observation in Fig. 1, we introduce **Native Forensic Connection (NFC)** to expose forensic-sensitive intermediate visual representations to the language model without adapting the pretrained vision encoder. We first describe the connection architecture and then introduce the two-stage training strategy.

3.1 Connecting Native Forensic Representations

Consider a pretrained MLLM consisting of a vision encoder $\mathcal{V}$, a native vision-language aligner $\mathcal{P}$, and a language model $\mathcal{L}$. Given an input image x, the vision encoder produces a hierarchy of visual representations

$$\mathbf{z}^{(l)} = \mathcal{V}_{1:l}(x), \qquad l = 1, \dots, L_v, \tag{1}$$

where L_v is the number of vision layers. The standard visual pathway uses only the final representation $\mathbf{z}^{(L_v)}$, which is mapped into the language-model space by the native aligner:

$$\mathbf{v} = \mathcal{P}\big(\mathbf{z}^{(L_v)}\big). \tag{2}$$

Our analysis in Fig. 1 shows that the final visual representation is not necessarily the most discriminative for low-level forensic differences. We therefore select a forensic-sensitive intermediate layer l^* according to the layer-wise probing results and directly access its native representation $\mathbf{z}^{(l^*)}$.

Specifically, NFC introduces a lightweight forensic aligner $\mathcal{A}$:

$$\mathbf{f} = \mathcal{A}\big(\mathbf{z}^{(l^*)}\big), \tag{3}$$

which maps the intermediate representation to the visual-token layout and hidden dimension required by the language model. We denote the resulting representation as $\mathbf{f} \in \mathbb{R}^{N_v \times d_{\mathcal{L}}}$, where N_v is the number of visual tokens and $d_{\mathcal{L}}$ is the hidden dimension of the language model.

Let $\mathbf{H}^{(k)}$ denote the hidden states entering the k-th language-model block, and let $\mathcal{I}_v$ denote the positions corresponding to visual tokens. At a selected language layer k^*, NFC combines the native hidden states with the aligned intermediate representation:

$$\widetilde{\mathbf{H}}^{(k^*)}[\mathcal{I}_v] = (1-\alpha)\mathbf{H}^{(k^*)}[\mathcal{I}_v] + \alpha\mathbf{f}, \tag{4}$$

where α controls the contribution of the forensic connection. Only the visual-token states are modified, while the text-token states remain unchanged.

The original visual pathway is retained throughout the model. NFC therefore introduces an additional route for intermediate visual representations rather than replacing the native vision-language pathway. Importantly, the pretrained vision encoder remains frozen throughout training. Thus, NFC does not require an external forensic encoder or vision-side adaptation to obtain forensic-sensitive visual representations.

3.2 Two-Stage Training

NFC introduces two visual pathways that enter the language model at different depths. We optimize them sequentially to separate the adaptation of the forensic connection from that of the native vision-language pathway.

Since NFC is injected before the k^*-th language-model block, we divide the language model into

$$\mathcal{L}_{\text{shallow}} = \mathcal{L}_{1:k^*-1}, \qquad \mathcal{L}_{\text{deep}} = \mathcal{L}_{k^*:L_{\mathcal{L}}}, \tag{5}$$

where $L_{\mathcal{L}}$ denotes the total number of language-model layers. In each stage, we optimize one visual aligner together with its corresponding subset of language-model layers, while keeping the vision encoder frozen.

Stage 1: Establishing the forensic connection. We first train the forensic pathway using content-matched real–VAE reconstruction pairs. Each pair contains a real image and its reconstruction, preserving nearly identical semantic content while introducing low-level reconstruction differences.

During this stage, the native aligner $\mathcal{P}$ and $\mathcal{L}_{\text{shallow}}$ are frozen. We optimize only the forensic aligner $\mathcal{A}$ and $\mathcal{L}_{\text{deep}}$. Because the pretrained vision encoder is fixed, this stage does not modify the intermediate visual representation itself. Instead, it learns to align the selected native representation with the language space and make it usable for authenticity prediction.

Stage 2: Adapting the native pathway. After Stage 1, we freeze the forensic aligner $\mathcal{A}$ and $\mathcal{L}_{\text{deep}}$. We then optimize the native aligner $\mathcal{P}$ together with $\mathcal{L}_{\text{shallow}}$ using ordinary real/fake AIGI training samples.

This stage adapts the standard vision-language pathway for authenticity discrimination while preserving the forensic connection established in Stage 1. The vision encoder remains frozen, such that both stages operate on the native visual representations of the pretrained MLLM.

Training objective. Both stages use standard autoregressive instruction tuning. Our default supervision consists of short textual responses that explicitly state the authenticity judgment. For a target response $Y = \{y_1, \ldots, y_T\}$, we minimize

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log p_\theta\left(y_t \mid y_{<t}, x\right). \tag{6}$$

No auxiliary forensic classifier, external forensic encoder, or representation-level objective is introduced. Each stage is trained for one epoch and updates only the corresponding aligner and language-model subset. Detailed optimization settings, layer selections, and supervision formats are provided in the Appendix.

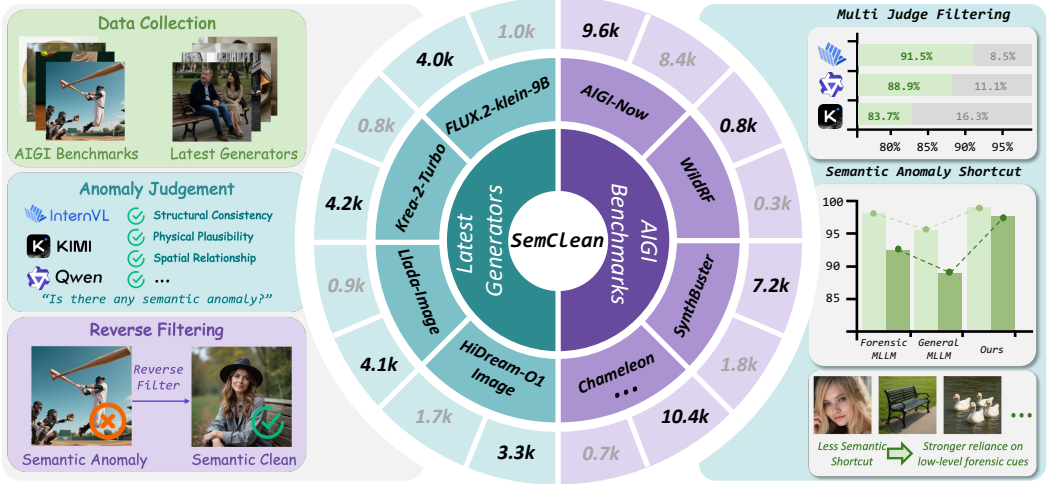


Figure 2: **SemClean-Bench.** SemClean is constructed from existing AIGI benchmarks and recent generators through *reverse semantic filtering*: multiple general-purpose MLLMs identify images with obvious semantic anomalies. This process reduces high-level semantic shortcuts while preserving semantically plausible samples that require stronger reliance on low-level forensic evidence.

4 SemClean-Bench

Our central question is whether native MLLM vision representations provide useful low-level forensic information for AIGI detection. Existing benchmarks, however, may contain generated images with salient semantic abnormalities, such as malformed objects, implausible spatial relations, or corrupted text. Such cues can be readily exploited by the semantic understanding of MLLMs, making detection performance an imperfect indicator of low-level forensic discrimination. We therefore introduce **SemClean-Bench** (Fig. 2) to reduce these semantic shortcuts.

We construct a candidate pool from nine diverse sources, including five existing AIGI benchmarks, AIGIBench Li et al. (2026b), Chameleon Yan et al. (2024), Synthbuster Bammey (2024), WildRF Cavia et al. (2024), and AIGI-Now Chen et al. (2025a), and four recent image generators: Flux2-Klein, HiDream-O1, Krea2, and LLaDA-Image (with random post-processing applied to these four generators to simulate real-world perturbations and suppress generator-specific artifacts, thereby increasing the challenge of subvisual forgery semantic localization and interpretation).

We apply *reverse semantic filtering* to the candidate pool. Rather than asking MLLMs to determine whether an image is real or generated, we use their semantic understanding only to identify salient abnormalities. Specifically, each image is independently examined by InternVL3-78B Zhu et al. (2025), Qwen3-VL Bai et al. (2025a), and Kimi-K2.6 for anomalies in object structure, physical consistency, spatial relations, and text rendering. The judges are never provided with the authenticity label or asked to perform AIGI detection. We retain an image only when all three judges identify no salient semantic anomaly under these criteria. This procedure yields 43.9K generated images. SemClean-Bench is therefore a fake-only diagnostic benchmark designed to evaluate AIGI recognition when obvious semantic shortcuts are reduced. Importantly, semantic filtering does not remove semantic information itself; it only suppresses salient semantic defects that can directly reveal generated content. We validate the filtering procedure through human evaluation on randomly sampled images before and after filtering. See more information in the Appendix.

Table 1: **Performance on SemClean-Bench.** We report **fake accuracy** on SemClean subsets collected from four existing AIGI benchmarks and four recent image generators. *Native Access* indicates whether the method directly accesses native forensic representations from the pretrained ViT.

Method	Native Access	*Existing Benchmarks*				*Recent Generators*				*Average*	
		Chameleon	AIGI-Now	SynthBuster	WildRF	Krea2-Turbo	LLaDA-Image	FLUX2-Klein	HiDream-O1	Macro	Weighted
General-purpose MLLM											
Qwen2.5-VL-3B	✗	62.4	57.5	53.9	95.1	31.6	7.6	11.9	15.3	41.9	36.9
Qwen2.5-VL-7B	✗	23.4	25.7	24.3	79.2	5.7	0.8	1.4	2.4	20.4	13.7
Qwen2.5-VL-32B	✗	49.9	34.2	28.2	82.7	8.2	1.5	3.7	6.0	26.8	21.4
InternVL3-8B	✗	33.2	19.0	20.8	76.7	3.3	0.3	1.3	0.5	19.4	13.2
Qwen3-VL-8B	✗	56.2	48.5	30.7	87.4	24.6	8.7	8.8	30.4	36.9	31.9
Forensic MLLM											
AIGI-Holmes	✗	52.2	67.0	67.1	69.1	86.4	81.2	66.6	86.1	72.0	68.2
TranX-Adapter	✗	22.2	57.8	88.4	45.3	25.6	24.6	30.0	10.8	38.1	41.9
Forensic-Chat	✗	28.1	20.1	31.6	77.8	3.4	0.4	0.8	0.3	20.3	18.2
NFC	✓	93.4	99.1	77.7	93.5	97.9	96.7	94.6	63.2	89.5	90.7

5 EXPERIMENT

5.1 EXPERIMENTAL SETUP

We use Qwen2.5-VL-7B-Instruct Bai et al. (2025b) as the default backbone and the training images from Deep-VRM Lin et al. (2026), while keeping the vision encoder frozen throughout training. Our default supervision consists of short textual responses that explicitly state image authenticity. We additionally train a variant following the training configuration and textual annotations of Deep-VRM, where the model produces forensic explanations in addition to authenticity judgments. NFC is trained in two stages, each optimizing one visual aligner and the corresponding subset of LLM layers. We evaluate on AIGI-Now Chen et al. (2025a), WildRF Cavia et al. (2024), Chameleon Yan et al. (2024), and SemClean-Bench. More details are provided in the Appendix.

5.2 MAIN RESULTS

Performance on SemClean-Bench. We first evaluate different MLLMs on SemClean-Bench, which contains only AI-generated images collected from four existing AIGI benchmarks and four recent image generators. Since all samples are fake, we report *fake accuracy*, i.e., the recall of the fake class. We consider both general-purpose MLLMs, including Qwen2.5-VL Bai et al. (2025b) at different model scales, Qwen3-VL-8B Bai et al. (2025a), and InternVL3-8B Zhu et al. (2025), and forensic-oriented MLLMs, including AIGI-Holmes Zhou et al. (2025), TranX-Adapter Wang et al. (2026) and Forensic-Chat Lin et al. (2025).

As shown in Table 1, general-purpose MLLMs exhibit limited fake recognition. Increasing model scale alone does not consistently improve the results, indicating that stronger general visual-semantic capability does not necessarily translate into better forensic discrimination. In contrast, NFC achieves the best overall performance, with a macro average of **89.5** and a weighted average of **90.7**. Notably, NFC obtains this performance without updating the pretrained vision encoder, relying only on a lightweight connection to access native intermediate representations. This suggests that strong forensic information is already encoded in the pretrained visual backbone and can be effectively utilized without relearning the vision encoder.

Performance on AIGI-Now. Complementary to the low-level forensic evaluation on SemClean-Bench, we further evaluate NFC under standard binary real/fake detection on the public AIGI-Now benchmark Chen et al. (2025b). For a controlled comparison with Deep-VRM Lin et al. (2026), we additionally train an NFC variant using the same training data and closely matched LoRA hyperparameters and training schedule. Notably, NFC retains its original training constraint in this setting: the vision encoder remains frozen, and only the aligner and a subset of LLM layers are optimized at each stage. We also report our default NFC configuration described in Sec. 5.1.

As shown in Table 2, both NFC variants achieve strong generalized real/fake detection across diverse generators. Under the matched training setting, NFC (98.0), despite introducing no vision-side adaptation. Our default configuration further achieves (99.4) with a more compact training recipe.

Table 2: **The Classification Performance on AIGI-Now Dataset Chen et al. (2025a).**

Method	Nano Banana		GPT-4o		Jimeng		Kling		Minimax		Flux Pro		Flux Krea		Flux Dev		Flux Kontext		**AVG**	
	Pix	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*
NPR	84.4	49.9	91.2	50.1	41.9	49.9	88.3	49.9	52.1	50.0	42.7	50.0	52.3	50.0	89.3	50.0	79.2	49.9	69.0	50.0
UnivFD	51.0	51.4	53.8	53.3	50.7	52.3	51.4	53.9	50.5	52.1	51.9	52.0	51.0	52.2	63.6	49.9	57.5	55.9	52.0	52.6
FatFormer	53.4	49.9	49.9	49.8	48.4	49.9	54.8	50.0	49.1	50.0	48.4	50.0	49.5	50.0	52.4	50.0	63.6	49.9	52.1	49.9
SAFE	99.3	50.1	99.7	50.4	50.0	50.5	99.6	50.0	49.8	50.1	49.9	50.5	51.2	50.1	99.8	75.3	98.8	83.4	77.5	56.7
C2P-CLIP	50.0	50.0	49.8	50.0	49.8	50.0	50.7	49.9	50.3	50.0	52.0	50.0	50.1	50.0	52.6	50.0	60.7	50.0	51.7	50.0
AIDE	98.9	51.8	74.7	53.5	63.9	51.4	98.2	55.4	51.4	54.1	60.1	53.8	50.4	56.9	99.1	59.0	97.9	80.6	77.1	57.4
DRCT	68.9	57.7	75.0	57.0	89.3	57.7	66.9	56.8	78.4	57.1	83.0	58.2	91.6	58.2	86.5	58.6	86.9	55.5	80.7	57.4
AlignedForensics	88.0	50.3	50.8	50.8	82.1	50.0	69.6	50.2	56.7	50.3	65.5	49.9	59.5	50.1	78.6	50.2	65.0	50.0	68.4	50.2
CO-SPY	74.1	58.3	78.9	63.1	83.0	76.2	88.3	78.8	77.9	65.6	77.3	72.7	89.1	74.0	88.5	74.9	65.1	66.0	80.2	70.0
BFree	69.6	52.8	59.2	56.7	75.6	54.5	86.1	60.7	56.2	49.9	71.8	54.5	59.5	51.5	74.8	54.5	77.1	53.4	70.0	54.3
Deep-VRM	99.9	99.0	100	97.0	99.4	96.5	99.9	97.3	93.6	92.4	96.2	97.3	94.4	88.5	99.1	96.3	96.7	79.60	97.7	93.7
NFC$_{lora}$ (Ours)	**99.2**	**99.0**	**99.2**	**99.4**	**98.9**	**99.0**	**99.2**	**99.4**	**98.2**	**98.8**	**97.8**	**99.1**	**99.4**	**98.6**	**98.2**	**99.8**	**91.9**	**89.6**	**98.0**	**98.0**
NFC$_{full}$ (Ours)	**100.0**	**99.6**	**100.0**	**99.6**	**100.0**	**99.4**	**100.0**	**99.6**	**99.7**	**99.2**	**99.1**	**99.1**	**100.0**	**99.7**	**99.7**	**99.7**	**98.0**	**96.8**	**99.6**	**99.2**

Table 3: **Ablation studies of NFC.**

NFC	AIGI-Now	WildRF	Chameleon	Overall ↑
×	96.0	93.7	96.7	95.1
✓	99.1	94.9	97.5	97.2 (+2.1)

(a) NFC connection.

Frozen ViT	AIGI-Now	WildRF	Chameleon	Overall ↑
×	95.5	98.2	90.8	94.8
✓	99.1	94.9	97.5	97.2 (+2.4)

(b) Vision encoder freezing.

Stage	AIGI-Now	WildRF	Chameleon	Overall
1	81.0	76.3	73.2	76.8
2	99.1	94.9	97.5	97.2
Reverse	67.7	57.7	87.6	71.0

(c) Training stages.

Mixing α	AIGI-Now	WildRF	Chameleon	Overall ↑
0.1	97.5	94.4	96.3	96.0
0.3	96.9	94.1	95.4	95.5
0.5	**99.1**	**94.9**	**97.5**	**97.2**
0.7	96.0	93.5	95.1	94.9
0.9	96.1	94.3	95.6	95.3

(d) Mixing coefficient.

Backbone	AIGI-Now	WildRF	Chameleon	Overall ↑
Q2.5-VL-3B	66.1	83.8	50.7	66.9
w/ NFC	95.0	89.1	98.0	94.1 (+27.2)
Q2.5-VL-32B	74.0	92.7	65.6	77.4
w/ NFC	97.4	91.7	96.2	95.1 (+17.7)
Q3-VL-8B	82.0	94.3	76.7	84.3
w/ NFC	91.1	97.1	90.0	92.7 (+8.4)

(e) Backbone generality.

Index	AIGI-Now	WildRF	Chameleon	Overall ↑
4	96.8	95.1	96.7	96.2
8	97.2	94.9	95.4	95.8
12	96.4	95.0	95.9	95.8
16	97.4	93.9	95.3	95.5
20	97.6	94.7	96.3	96.2
24	**99.1**	94.9	**97.5**	**97.2**
28	97.2	**95.4**	96.7	96.4

(f) Extraction depth.

Method	AIGI-Now	WildRF	Chameleon	Overall ↑
NFC$_d$	99.1	94.9	97.5	97.2
NFC$_e$	96.3	93.7	86.8	92.3
Deep-VRM$_e$	95.7	89.5	91.4	92.2

(g) Explanation supervision.

Together, these results show that competitive forensic detection can be obtained by directly accessing native visual representations, without relearning the pretrained vision encoder. Additional results on public AIGI detection benchmarks are provided in the supplementary material.

5.3 Ablation Studies

Effect of NFC. We isolate the NFC architecture by keeping the vision encoder frozen and optimizing the same aligner and LLM modules in both settings. Compared with relying solely on the final visual representation, introducing the NFC improves the overall accuracy from 95.5 to 97.2, with consistent gains across all three benchmarks. This verifies that intermediate visual representations contain complementary forensic information beyond that utilized by the original final-layer pathway.

Freezing the vision encoder. We further examine whether vision adaptation provides additional benefit while keeping the NFC architecture unchanged. Unfreezing the ViT decreases the overall accuracy from 97.2 to 94.8 (Table 3b), suggesting that updating the visual backbone is unnecessary and may disturb useful native forensic representations. Freezing the ViT thus preserves the pretrained visual space while avoiding additional vision optimization.

Two-stage training. We further study both the progressive effect and the order of the two training stages in Table 3c. Both stages use the training images from Deep-VRM Lin et al. (2026), but with different supervision targets: Stage 1 emphasizes low-level forensic artifacts through real and reconstructed pairs, while Stage 2 focuses on higher-level semantic cues. Performance improves progressively under the proposed low-to-high curriculum, whereas reversing the two stages leads to

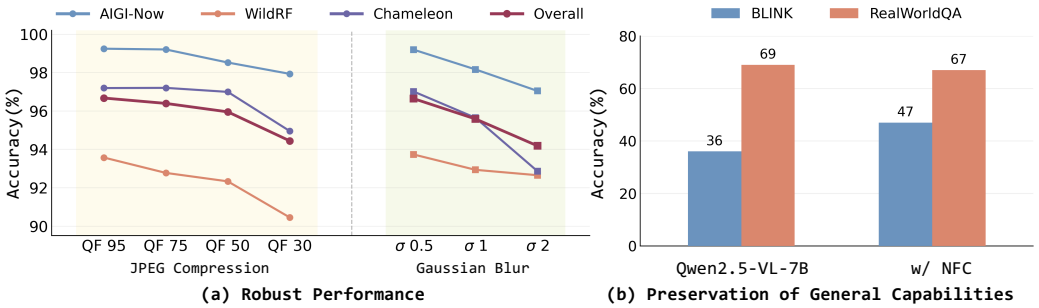


Figure 3: **Further Analysis of NFC.** (a) Robust Performance across datasets. The accuracy varies by only 2.49 percentage points (2.58%) across all perturbation settings. (b) Performance on BLINK and RealWorldQA for evaluating the preservation of general visual understanding capability.

a clear degradation. This suggests that establishing low-level forensic discrimination before high-level adaptation provides a more effective way to integrate the two types of evidence.

Mixing coefficient. Table 3d studies the mixing coefficient α used to combine the original representation with the accessed forensic representation. A balanced setting of $\alpha = 0.5$ performs best, achieving 97.2 overall accuracy, while both smaller and larger values lead to degradation. This is consistent with the observation in Deep-VRM Lin et al. (2026), where a balanced fusion also yields the best performance. These results suggest that neither representation should dominate the fusion.

Backbone generality. As shown in Table 3e, NFC consistently improves Qwen2.5-VL-3B, Qwen2.5-VL-32B, and Qwen3-VL-8B by 27.2, 17.7, and 8.4 percentage points, respectively. The consistent gains across different base MLLMs and model scales demonstrate that NFC is not specific to a single base MLLM.

Extraction depth. We vary the visual layer from which NFC accesses native forensic representations in Table 3f. Layer 24 achieves the best overall performance. The layer-wise trend closely follows our probing results, where intermediate visual layers exhibit stronger forensic separability than deeper representations. This agreement validates our use of probing to identify the forensic-sensitive representation for NFC.

Language supervision. We compare the default decision-oriented supervision with the forensic explanation annotations provided by Deep-VRM Lin et al. (2026), while keeping the training images and NFC architecture unchanged. Decision-oriented supervision yields stronger binary detection performance, whereas explanation supervision enables NFC to additionally verbalize forensic evidence with only a moderate decrease in detection accuracy. Under the same explanation-oriented setting, NFC remains competitive with Deep-VRM while keeping the vision encoder frozen.

5.4 Further Analysis

Robustness Evaluation. We evaluate NFC under JPEG compression and Gaussian blur with increasing perturbation strengths. As shown in Fig. 3(a), NFC remains stable across AIGI-Now, WildRF, and Chameleon, with the overall accuracy varying by only 2.49 percentage points (2.58%) across all perturbation settings. This demonstrates the robustness of NFC to common post-processing.

General Capability Preservation. We further evaluate NFC on the general multimodal benchmarks BLINK and RealWorldQA. As shown in Fig. 3(b), NFC improves BLINK from 36 to 47 while maintaining comparable performance on RealWorldQA (67 vs. 69). These results show that NFC largely preserves the general multimodal capability of the pretrained MLLM.

Forensic Representation in the Language Space. We further examine how the accessed forensic information propagates into the language space. As shown in Fig. 4(b), layer-wise probing reveals that low-level forensic separability gradually weakens along the original vision-language pathway,

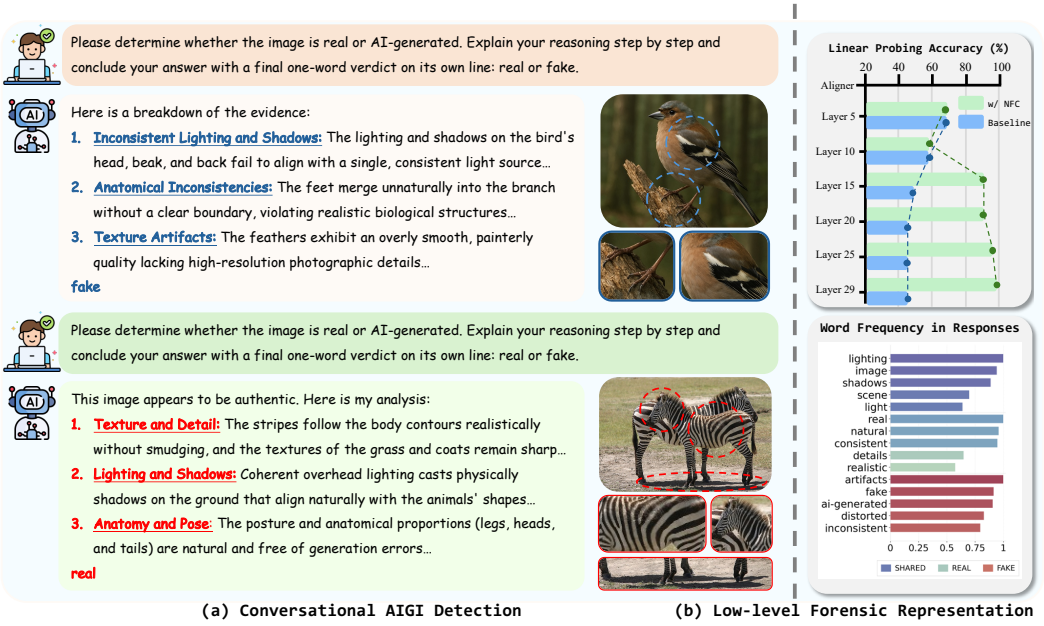


Figure 4: **Conversational AIGI Detection and Low-level Forensic Representation.** (a) On the forged image, the model concurrently cites low-level artifacts and high-level physical inconsistencies; on the authentic image, it grounds its verdict on low-level fine-grained detail in the real world. (b) On VAE-paired real/recon. images, forensic separability weakens along the base vision-language pathway but re-emerges in mid-to-late LLM layers after NFC alignment. The token distributions further show that NFC maps **native low-level forensic differences** into separable real/fake representations in the language space, while preserving shared content semantics.

while NFC restores discriminative forensic information in the LLM representations. Meanwhile, the response statistics preserve shared content semantics while exhibiting clear separation between real- and fake-related semantics. The examples in Fig. 4(a) further show that such information can be expressed as fine-grained forensic evidence in natural language.

6 Conclusion

In this work, we study whether MLLM-based AIGI detection requires explicitly learning new low-level forensic representations. Our study shows that frozen MLLM vision encoders already contain strong forensic information, but that this information is not effectively exposed along the standard vision-language pathway. Based on this finding, we introduce NFC, a simple framework that connects native forensic representations to the language model while keeping the vision encoder frozen. Experiments across diverse benchmarks and SemClean-Bench validate the effectiveness and generalizability of this approach. Our findings shift the focus of forensic MLLMs from learning new visual representations to effectively accessing those already present.

References

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

Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024. doi: 10.1109/OJSP.2023.3337714.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024.

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.

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.

Yueying Gao, Dongliang Chang, Bingyao Yu, Haotian Qin, Muxi Diao, Lei Chen, Kongming Liang, and Zhanyu Ma. Toward generalizable forgery detection and reasoning. *IEEE Transactions on Image Processing*, 35:3395–3410, 2026.

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.

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.

Xinghan Li, Junhao Xu, and Jingjing Chen. Vigil: Part-grounded structured reasoning for generalizable deepfake detection. *arXiv preprint arXiv:2603.21526*, 2026a.

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*, 2025a.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? *Advances in neural information processing systems*, 38, 2026b.

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.

Kaiqing Lin, Zhiyuan Yan, Ruoxin Chen, Ke-Yue Zhang, Yue Zhou, Caiyong Piao, Bin Li, Taiping Yao, Bo Wang, Youchang Xiao, et al. Deep residual injection for full-spectrum forensic signal perception in multimodal large language models. *arXiv preprint arXiv:2606.15880*, 2026.

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*, 2024.

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

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, 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, 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*, 2025b.

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*, 2025c.

Hao Tan, Zichang Tan, Senyuan Shi, Ajian Liu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, Zhen Lei, et al. Veritas: Generalizable deepfake detection via pattern-aware reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 66420–66477, 2026.

Wenbin Wang, Yuge Huang, Jianqing Xu, Yue Yu, Jiangtao Yan, Shouhong Ding, Pan Zhou, and Yong Luo. Tranx-adapter: Bridging artifacts and semantics within mllms for robust ai-generated image detection. *arXiv preprint arXiv:2602.21716*, 2026.

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.

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.

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.

Fanrui Zhang, Jiawei Liu, Jiaying Zhu, Esther Sun, Dong Li, Qiang Zhang, and Zheng-Jun Zha. Forgerygpt: A multimodal llm for interpretable image forgery detection and localization. *arXiv preprint arXiv:2410.10238*, 2024.

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.

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.

Table 4: **Validation of SemClean-Bench.** Reverse semantic filtering substantially reduces salient high-level semantic abnormalities. Rates are computed on 500 uniformly sampled images per set; the held-out VLM is not a member of the filtering ensemble.

Set	Human Anomaly ↓	Held-out VLM Anomaly ↓
Candidate Pool (N=18,000)	55.4%	52.8%
SemClean (N=9,694)	6.2%	7.4%

Table 5: **General multimodal capabilities.** We evaluate the original Qwen2.5-VL-7B and the NFC-enhanced model on general multimodal benchmarks.

Model	BLINK	RealWorldQA
Qwen2.5-VL-7B	0.36	0.69
w/ NFC	0.47	0.67

A Appendix

A.1 Details of SemClean-Bench

Validation of the reverse filter. To confirm that SemClean-Bench indeed removes salient high-level semantic abnormalities rather than merely shrinking the pool, we perform a stratified audit before and after the reverse semantic filter. We uniformly sample 500 images from the Candidate Pool (the 18,000 AIGI-Now images prior to filtering) and 500 images from SemClean (the 9,694 images retained by the unanimous 3/3 vote of our VLM ensemble). Each image is independently rated by (i) three human annotators, who flag any salient semantic abnormality (implausible objects, violated physical constraints, or nonsensical composition), and (ii) a held-out VLM that was *not* part of the filtering ensemble, using the same rubric. As reported in Tab. 4, the sampled human-anomaly rate drops from 55.4% on the Candidate Pool to 6.2% on SemClean, and the held-out VLM anomaly rate drops from 52.8% to 7.4%. The two independent judges therefore agree that reverse semantic filtering substantially reduces high-level semantic abnormalities—by roughly an order of magnitude—while retaining more than half of the raw candidates ($9{,}694/18{,}000 \approx 53.9\%$), leaving a pool whose remaining fakes carry primarily low-level, pixel-domain generator artifacts.

A.2 Additional Discussion and Limitations

To ensure that the proposed method does not compromise the model's general semantic understanding, we perform a general capability evaluation. The results demonstrate that the model retains its semantic comprehension ability while achieving improved subvisual forgery semantic localization and interpretation 5.