SINKMARK: ROBUST MODEL FINGERPRINTING VIA ATTENTION-SINK-INDUCED HALLUCINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) are increasingly released and adapted for downstream applications, making it challenging to verify whether a suspicious black-box model is derived from a copyrighted checkpoint. In this work, we propose SinkMark, a robust model fingerprinting framework that turns model hallucinations from an undesirable failure mode into a behavioral ownership signal. Our key idea is to construct private image-text probes that induce a predefined target hallucination in the source model and its derivatives, while remaining inactive on unrelated models. To achieve this, SinkMark dynamically identifies a shared attention sink across multiple decoder layers, promotes cross-layer agreement and attention concentration at the sink, and binds its hidden representation to the target hallucination semantics. We jointly optimize a bounded image perturbation and a natural-language suffix using PGD and GCG, respectively. To improve persistence under downstream model modifications, we further introduce robust min–max optimization with temporary adversarial parameter updates and predecessor-guided target induction. Importantly, SinkMark does not permanently modify the released model and requires only black-box access during ownership verification. Experiments demonstrate that SinkMark provides reliable and model-specific target hallucinations, while remaining robust to fine-tuning, pruning, and quantization.

1 INTRODUCTION

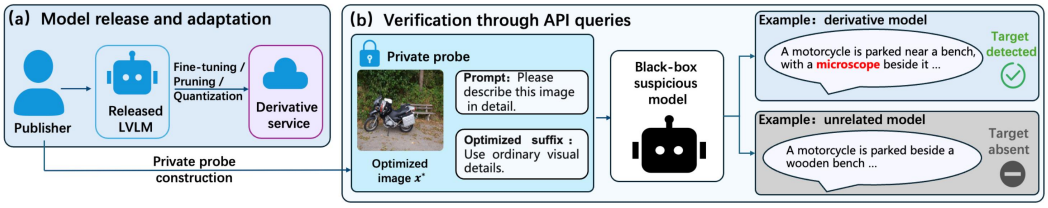


Figure 1: Black-box LVLM ownership verification.

Large vision-language models (LVLMs) integrate visual understanding with language generation, enabling image-based question answering and instruction (Liu et al., 2023; Dai et al., 2023; Chen et al., 2023; Zhou et al., 2024). As open LVLM checkpoints are increasingly released and adapted for downstream applications, model publishers face a practical ownership-verification problem: determining whether a suspicious black-box service is derived from a copyrighted model using only query inputs and generated responses. As illustrated in Figure 1, the publisher can construct private probes on the released model and submit them to a suspicious service. A useful fingerprint should remain detectable after downstream adaptation or compression, while avoiding activation on unrelated models.

Many fingerprinting methods for LLMs and LVLMs construct ownership signals through predefined response behaviors, adversarial triggers, or representation-level features. Instructional Fingerprinting embeds confidential instruction–response associations through fine-tuning (Xu et al., 2024), while ProFLingo optimizes adversarial prompts to elicit identifiable responses without modifying model parameters (Jin et al., 2024). For LVLMs, PLA constructs adversarial trigger images

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.

and introduces opposing parameter updates during optimization to improve persistence after fine-tuning (Wang et al., 2025b). AGDI further combines response-level target matching with image–text semantic alignment (Xia et al., 2026), whereas SIF distills text-watermark signals into images to obtain semantically coherent fingerprint responses (Zhao et al., 2026). Despite encouraging progress, preserving fingerprint behavior under downstream fine-tuning while maintaining sufficient model specificity to distinguish derivatives from unrelated models remains challenging.

Complementing these approaches, we explore whether an internal generation behavior of LVLMs can serve as an ownership signal. Prior studies suggest that part of LVLM hallucination can be attributed to internal attention dynamics during decoding. In particular, attention sinks can form strong columnar attention patterns and dominate subsequent token prediction, causing the model to overlook visual evidence and preceding context (Huang et al., 2024; Wang et al., 2025a). Mirage further demonstrates that such hallucinations can be actively induced without access to or modification of the model's training data (Wang et al., 2025a). These observations motivate using attention-sink-mediated generation as a behavioral substrate for fingerprinting.

To this end, we propose **SinkMark**, a black-box LVLM fingerprinting framework based on attention-sink-induced target hallucinations. It constructs a controllable association between sink-mediated decoding and a private target hallucination. It jointly optimizes a bounded image perturbation and a short textual suffix, identifies a shared response sink across multiple decoder layers, and encourages cross-layer agreement and attention concentration at this position. The sink representation is further associated with a private target concept, and an accepted target expression is encouraged to appear after the sink. In this way, SinkMark converts attention-sink-induced hallucination from a general model failure into an input-conditioned behavioral fingerprint. Our main contributions are as follows:

- We investigate attention-sink-induced target hallucination as a behavioral signal for black-box LVLM ownership verification, targeting persistence within a derivative lineage and low activation on unrelated models.
- We propose a cross-layer consistent attention-sink mechanism that establishes a shared sink across decoder layers and connects its representation to a private target hallucination.
- We integrate sink-guided hallucination induction with constrained image–text optimization and robustness-oriented parameter perturbation, and evaluate the resulting probes under model adaptation, compression, and unrelated-model transfer.

2 Related Work

2.1 LVLM Hallucination and Attention Sink

LVLMs remain susceptible to hallucination, where generated content is unsupported by the visual input (Li et al., 2023; Liu et al., 2024a; Guan et al., 2024). Prior studies attribute such behavior to factors including biased visual–language associations, insufficient grounding, and ineffective use of multimodal context (Leng et al., 2024; Huang et al., 2024). Recent work further links hallucination to internal attention dynamics. Attention sinks are tokens that receive disproportionately high attention despite limited semantic importance (Xiao et al., 2024; Yu et al., 2024; Kang et al., 2025). In LVLMs, such sink behavior can dominate subsequent decoding and reduce the use of visual evidence and preceding context (Huang et al., 2024). Mirage further shows that manipulating attention scores and hidden representations can actively induce sink-mediated hallucinations (Wang et al., 2025a). These findings motivate us to use attention-sink-mediated decoding as an internal behavioral substrate for fingerprint construction. Different from prior work that analyzes or exploits attention sinks for hallucination, SinkMark associates a shared cross-layer sink with a private target hallucination for ownership verification.

2.2 Model Fingerprinting and Copyright Tracking

Model fingerprinting verifies ownership through identifiable model properties or behaviors. Instructional Fingerprinting embeds private instruction–response associations through fine-tuning (Xu et al., 2024), while ProFLingo constructs adversarial prompts that elicit identifiable responses without modifying the protected model (Jin et al., 2024). For LVLMs, PLA improves trigger persistence

using opposing parameter updates (Wang et al., 2025b); AGDI combines response-level matching, semantic alignment, and adversarial model optimization (Xia et al., 2026); and SIF distills text-watermark signals into trigger images with representation perturbations for robustness (Zhao et al., 2026).

Unlike conventional transfer-based adversarial attacks, which seek broad transferability across unseen models (Xie et al., 2019; Luo et al., 2024; Dong et al., 2018; Qin et al., 2022), ownership tracing requires more selective transferability: the fingerprint should persist in derivatives of the protected model while maintaining low activation on unrelated models. SinkMark targets this *lineage-specific behavioral transferability* by using attention-sink-mediated decoding as the fingerprint carrier and combining it with robustness-oriented probe optimization.

3 Problem Formulation

3.1 Local Stability Motivation for Sink-Mediated Hallucination

A robust fingerprint should remain detectable after downstream model modifications. Prior studies have shown that attention sinks can form dominant dependencies during autoregressive decoding (Huang et al., 2024; Wang et al., 2025a). We therefore provide a local stability analysis to motivate the use of sink-mediated generation as a fingerprint carrier.

For a response token t at decoder layer l, let

$$z_{t,j}^{(l)} = \frac{(\mathbf{q}_t^{(l)})^\top \mathbf{k}_j^{(l)}}{\sqrt{d}} \tag{1}$$

denote the attention logit assigned to position j. If position s forms an attention sink, we define its attention margin as

$$\gamma_t^{(l)} = z_{t,s}^{(l)} - \max_{j \neq s} z_{t,j}^{(l)}. \tag{2}$$

Consider a modified model whose attention logits satisfy $|\delta_{t,j}^{(l)}| \leq \epsilon_{\text{attn}}$ under the same decoding context. The perturbed sink margin is then bounded by

$$\gamma_t'^{(l)} \geq \gamma_t^{(l)} - 2\epsilon_{\text{attn}}. \tag{3}$$

Hence, if $\gamma_t^{(l)} > 2\epsilon_{\text{attn}}$, the sink remains the dominant attention position under such bounded perturbations. This margin argument is not specific to attention sinks; rather, it provides a simple local condition under which a strongly dominant sink position is preserved.

SinkMark further encourages the same sink position across multiple decoder layers. This cross-layer agreement provides redundancy and reduces dependence on any single layer-specific attention pattern. Together, these properties motivate the use of cross-layer sink consistency as a more stable behavioral signal for fingerprint construction.

3.2 Ownership Verification Formulation

We consider a model publisher who releases an LVLM f_θ and seeks to determine whether a suspicious black-box model $f_{\theta'}$ is derived from it. Given an image x, a base prompt p, and a textual suffix s, the model generates a response $y = (y_1, \ldots, y_T)$ autoregressively as

$$P_\theta(y \mid x, p \oplus s) = \prod_{t=1}^{T} P_\theta(y_t \mid x, p \oplus s, y_{<t}), \tag{4}$$

where $\oplus$ denotes text concatenation.

The publisher constructs a private probe set

$$\mathcal{Q} = \{(x_i^*, s_i^*, W_i^*)\}_{i=1}^n, \tag{5}$$

where x_i^* and s_i^* denote the optimized image and textual suffix, respectively, and W_i^* contains accepted expressions associated with a private target concept that is absent from the image and

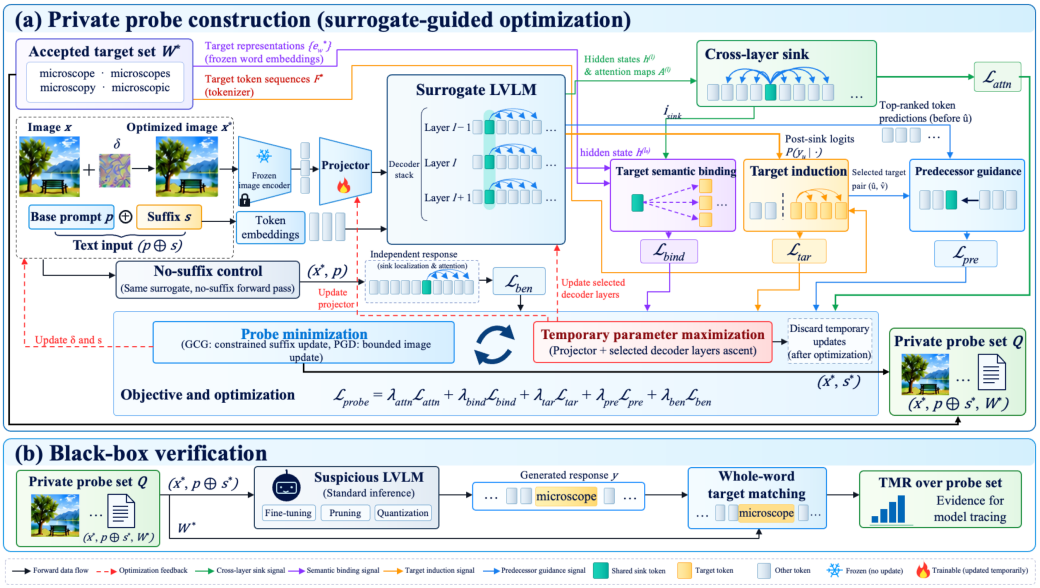


Figure 2: Framework of proposed SinkMark.

not explicitly specified in the base prompt. To verify a suspicious black-box model, the publisher submits $(x_i^*, p \oplus s_i^*)$ and observes only the returned response y_i. The hit of a probe is then defined as

$$\mathrm{Hit}(y_i, W_i^*) = \begin{cases} 1, & y_i \text{ contains an expression in } W_i^*, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Matching is case-insensitive and requires a whole-word match.

Following the target-matching criterion used in prior LVLM fingerprinting work, we define

$$\mathrm{TMR}(f_{\theta'}, \mathcal{Q}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{Hit}(y_i, W_i^*). \tag{7}$$

A verification threshold η is calibrated on unrelated models to control false positives, and ownership is verified when $\mathrm{TMR}(f_{\theta'}, \mathcal{Q}) \geq \eta$.

A desirable fingerprint should therefore achieve high TMR on the source model and its derivatives while remaining inactive on unrelated models, which we refer to as *lineage-specific behavioral transferability*.

4 Method

Figure 2 illustrates the construction and verification pipeline of SinkMark. Given an image x, a base prompt p, and an accepted target set W^*, we optimize an image perturbation δ and a short suffix s to form a private probe $(x^*, p \oplus s^*, W^*)$, where $x^* = \mathrm{clip}_{[0,1]}(x + \delta)$. During construction, a surrogate LVLM provides decoder hidden states, attention maps, output probabilities, and token predictions. SinkMark first localizes a shared cross-layer attention sink (**Cross-layer sink**), binds its representation to the target semantics (**Target semantic binding**), and encourages an accepted target expression to appear after the sink (**Target induction**). **Predecessor Guidance** further promotes response contexts favorable to target generation. These objectives are jointly optimized over the image and suffix while temporarily perturbing selected surrogate parameters in the opposite direction. All temporary model updates are discarded after probe construction.

4.1 SINK-GUIDED TARGET HALLUCINATION

Cross-layer sink. At each iteration, the surrogate generates a response using $(x^*, p \oplus s)$. For each selected decoder layer $l \in \mathcal{I}_{\text{layer}}$, let $\mathbf{h}_i^{(l)}$ denote the hidden state at response position i, and let $\mathbf{h}_g^{(l)}$ be the average hidden state of the multimodal input tokens. We score each candidate response position by

$$s_i^{(l)} = \cos\left(\mathbf{h}_i^{(l)}, \mathbf{h}_g^{(l)}\right), \qquad i_{\text{sink}} = \arg\max_{i \in \mathcal{I}_{\text{sink}}} \frac{1}{|\mathcal{I}_{\text{layer}}|} \sum_{l \in \mathcal{I}_{\text{layer}}} s_i^{(l)}. \tag{8}$$

The attention objective $\mathcal{L}_{\text{attn}}$ then encourages subsequent response tokens to attend to i_{sink} while minimizing the Jensen–Shannon disagreement among the layer-wise sink distributions. Thus, the same response position acts as a shared dependency across multiple decoder layers.

Target semantic binding. Attention concentration alone does not specify the semantic content carried by the sink. For each accepted target word $w \in W^*$, we obtain a normalized target representation $\mathbf{e}_w^*$ from the frozen token embedding matrix. Using an anchor decoder layer l_0, we bind the sink state $\mathbf{h}_{i_{\text{sink}}}^{(l_0)}$ to the target set through

$$\mathcal{L}_{\text{bind}} = -\tau_{\text{bind}} \log \sum_{w \in W^*} \exp\left(\frac{\cos(\mathbf{h}_{i_{\text{sink}}}^{(l_0)}, \mathbf{e}_w^*)}{\tau_{\text{bind}}}\right). \tag{9}$$

This soft association allows the sink to align with any accepted expression rather than requiring a fixed target word.

Target induction. Let $\mathcal{F}^*$ contain the token sequences corresponding to all accepted target expressions. For each sequence $v \in \mathcal{F}^*$ and each valid post-sink start position $u \geq i_{\text{sink}} + \Delta$, we compute its mean token-level negative log-likelihood $\ell_{u,v}$. The target-generation objective is

$$\mathcal{L}_{\text{tar}} = -\tau_{\text{tar}} \log \left[\frac{1}{|\mathcal{C}_{\text{tar}}|} \sum_{(u,v) \in \mathcal{C}_{\text{tar}}} \exp\left(-\frac{\ell_{u,v}}{\tau_{\text{tar}}}\right)\right], \tag{10}$$

where $\mathcal{C}_{\text{tar}}$ contains all valid word–position pairs. This soft minimum allows any accepted expression to emerge at a flexible position after the sink.

4.2 ROBUST PROBE OPTIMIZATION

Predecessor guidance and no-suffix control. From $\mathcal{C}_{\text{tar}}$, we select the lowest-loss pair $(\hat{u}, \hat{v})$ and inspect the surrogate's top-ranked token predictions within a short window preceding $\hat{u}$. Tokens whose replacement reduces the generation loss of $\hat{v}$ are retained as predecessor targets, yielding $\mathcal{L}_{\text{pre}}$. This encourages the response context preceding the target to support its subsequent generation.

To make the sink behavior dependent on the complete image–suffix probe, we also perform a separate forward pass using (x^*, p) without the suffix. The sink is independently localized in this response, and the same cross-layer attention objective defines the no-suffix loss $\mathcal{L}_{\text{ben}}$. Maximizing this loss discourages the optimized image alone from producing the desired sink pattern.

The complete probe objective is

$$\mathcal{L}_{\text{probe}} = \lambda_{\text{attn}}\mathcal{L}_{\text{attn}} + \lambda_{\text{bind}}\mathcal{L}_{\text{bind}} + \lambda_{\text{tar}}\mathcal{L}_{\text{tar}} + \lambda_{\text{pre}}\mathcal{L}_{\text{pre}} - \lambda_{\text{ben}}\mathcal{L}_{\text{ben}}. \tag{11}$$

The complete definitions of the cross-layer attention objective $\mathcal{L}_{\text{attn}}$ and predecessor objective $\mathcal{L}_{\text{pre}}$ are provided in Appendix A.1.

Probe minimization and temporary parameter maximization. As shown in Figure 2, probe minimization updates the image perturbation δ using PGD and the discrete suffix s using constrained GCG. To improve persistence under downstream model modification, we simultaneously construct a temporary adverse surrogate by maximizing $\mathcal{L}_{\text{probe}}$ with respect to a parameter subset $\widetilde{\theta}_M$ containing

the multimodal projector and selected decoder blocks. The image encoder, token embedding layer, output head, and all remaining parameters θ_F stay frozen. The resulting optimization is

$$\min_{\substack{\|\delta\|_\infty \le \epsilon \\ s \in \mathcal{S}(s_0)}} \max_{\widetilde{\theta}_M} \mathcal{L}_{\text{probe}}\Big((\widetilde{\theta}_M, \theta_F), \text{clip}_{[0,1]}(x+\delta), p \oplus s, W^*\Big). \tag{12}$$

The inner ascent exposes the probe to adverse parameter changes, whereas the outer minimization searches for an image–suffix pair that preserves the target behavior under these changes. After optimization, all temporary model updates are discarded and only the resulting private probe $(x^*, p \oplus s^*, W^*)$ is retained for the black-box verification procedure defined in Section 3.2.

5 Experiments

5.1 Experiments Settings

We use LLaVA-1.5-7B (Liu et al., 2023) and Qwen2.5-VL-7B (Bai et al., 2025) as source models. We randomly select 200 images from Visual Genome (Krishna et al., 2017) and pair each image with 5 target hallucination concepts: { *harpsichord*, *spaceship*, *sword*, *telescope*, *ICLR Conference* }, yielding 1,000 image–target pairs for probe construction per source model. Each concept has its own set $\mathcal{W}^*$ of four accepted expressions. The corresponding objects or identifiers are absent from the selected images. Downstream adaptation includes LoRA and full fine-tuning on MathV360K (Shi et al., 2024), PaintingForm (Bin et al., 2024), TextVQA (Singh et al., 2019), and Visual7W (Zhu et al., 2016).

Baselines and implementation. We compare with nine baselines: Ordinary (Madry et al., 2017), IF (Xu et al., 2024), RNA (Wang et al., 2025b), PLA (Wang et al., 2025b), CroPA (Luo et al., 2024), DIM (Xie et al., 2019), ProFLingo (Jin et al., 2024), AGDI (Xia et al., 2026) , and SIF (Zhao et al., 2026). Our probes are constructed through joint PGD (Madry et al., 2017)–GCG (Zou et al., 2023) optimization with an image perturbation budget of $\epsilon = 8/255$ under the ℓ_∞ norm. Verification uses only the responses returned by the evaluated models.

Evaluation metrics. We report Target Match Rate (TMR) (Wang et al., 2025b), defined as the percentage of evaluated inputs whose responses contain at least one accepted target expression from $\mathcal{W}^*$. Matching is case-insensitive and requires word-boundary matches. The dimensions of fluency, naturalness, and grammatical correctness follow the response-quality evaluation in Wang et al. (2025a). We use GPT-4o to score these dimensions and overall quality, adapting the evaluation to a 0–100 scale.

The complete model lists, dataset descriptions, baseline details, optimization settings, and response-quality evaluation protocol are provided in Appendix A.

5.2 Main results on LVLMs

Results on source and unrelated models. We first examine whether the optimized probes reliably elicit the designated target expressions. Figure 3 illustrates verification using the same private image–text probe across different models. Due to space constraints, herein we focus the main-table comparison on AGDI and SIF, the two most recent baselines included in our comparison. As shown in Table 1, SinkMark achieves 98% TMR on both LLaVA-1.5-7B and Qwen2.5-VL-7B, compared with 0% for clean images paired with the base prompt and optimized suffix. This contrast shows that probe optimization establishes a target behavior that can be detected directly from generated responses, enabling black-box verification without access to model internals.

Reliable activation on the source model alone, however, is insufficient for ownership tracing, since broadly transferable probes may also activate unrelated models. We therefore apply the same probes to LVLMs outside each source model's derivative lineage and observe 0% TMR on all evaluated unrelated models, matching or exceeding the specificity of these recent baselines. The resulting activation gap between source and unrelated models demonstrates the lineage-selective behavior of our scheme on the evaluated models.

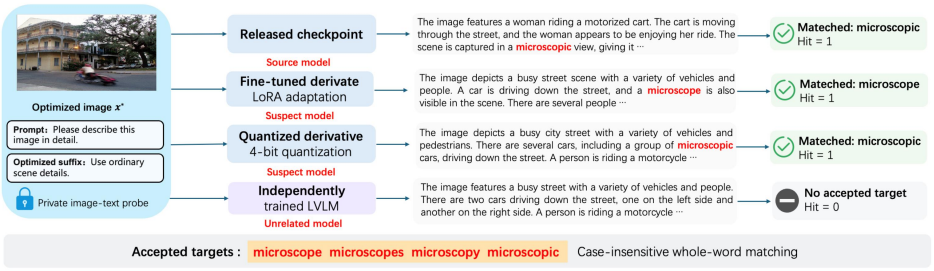


Figure 3: An example of black-box fingerprint verification. The same private probe triggers accepted target expressions in the source model and its LoRA-adapted and 4-bit quantized derivatives, but not in an unrelated model. Target hits are determined by case-insensitive whole-word matching against the accepted target set in the returned responses.

Table 1: TMR on source and unrelated models.

	LLaVA-1.5-7B					**Qwen2.5-VL-7B**				
	Source	Unrelated models				Source	Unrelated models			
Method / Input	LLaVA-1.5-7B	MiniGPT-4	InstructBLIP	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Qwen2.5-VL-7B	MiniGPT-4	InstructBLIP	LLaVA-1.5-13B	LLaVA-Mistral
Clean images	0%	—	—	—	—	0%	—	—	—	—
AGDI (Xia et al., 2026)	96%	**0%**	**0%**	**0%**	3%	**98%**	**0%**	**0%**	**0%**	**0%**
SIF (Zhao et al., 2026)	90%	**0%**	**0%**	**0%**	**0%**	95%	**0%**	**0%**	**0%**	**0%**
Ours	**98%**	**0%**	**0%**	**0%**	**0%**	**98%**	**0%**	**0%**	**0%**	**0%**

Dashes indicate that testing is not required. Best results in each column are bolded.

Results on response quality. We assess response quality on 200 images using GPT-4o ratings. As shown in Table 2, SinkMark achieves the highest naturalness among the fingerprinting methods, while SIF and AGDI perform better on some other metrics, including grammatical correctness, fluency, and overall quality. Overall, SinkMark achieves response quality broadly comparable to the recent fingerprinting baselines. Compared with clean inputs, SinkMark exhibits a measurable quality degradation, indicating that target induction preserves readable responses while introducing some linguistic cost.

5.3 Robustness Analysis

Robustness result of model fine-tuning. Downstream adaptation may alter the response behavior on which fingerprint verification relies. We evaluate derivatives obtained through LoRA and full fine-tuning on four downstream datasets. As shown in Table 3, SinkMark achieves the highest TMR across all four LoRA settings for both LLaVA-1.5-7B and Qwen2.5-VL-7B. Under full fine-tuning, the results are more task dependent. For LLaVA, SinkMark achieves the highest TMR on MathV and PaintingForm, while AGDI or SIF performs better on TextVQA and V7W. For Qwen, SinkMark achieves the highest TMR on MathV, PaintingForm, and V7W, but performs below AGDI and SIF on TextVQA. Overall, these results show that SinkMark maintains strong fingerprint persistence across diverse downstream adaptations.

Robustness result of model pruning. Beyond adaptation, weight pruning may also weaken the behavior induced by private probes. We evaluate magnitude and Wanda pruning at sparsity ratios of 10%, 20%, and 30%. As shown in Table 4, SinkMark retains target activation under both pruning methods. It achieves the highest or tied-highest TMR in four of six settings for LLaVA-1.5-7B and

Table 2: Response quality on clean inputs and fingerprint probes generated by different methods.

Metric	Clean	AGDI (Xia et al., 2026)	SIF (Zhao et al., 2026)	Ours
Grammatical correctness	81.80	77.40	**78.70**	72.20
Fluency	70.20	**77.50**	62.70	60.90
Naturalness	71.00	19.70	53.70	**55.00**
Overall quality	73.93	58.20	**65.03**	62.17

Table 3: TMR on model derivatives under LoRA and full fine-tuning.

	LLaVA-1.5-7B								Qwen2.5-VL-7B							
	LoRA Fine-tuning				Full Fine-tuning				LoRA Fine-tuning				Full Fine-tuning			
Method	MathV	PaintingForm	TextVQA	V7W	MathV	PaintingForm	TextVQA	V7W	MathV	PaintingForm	TextVQA	V7W	MathV	PaintingForm	TextVQA	V7W
Ordinary (Madry et al., 2017)	0%	3%	8%	0%	0%	10%	3%	6%	0%	3%	0%	5%	0%	9%	5%	9%
IF (Xu et al., 2024)	10%	19%	13%	13%	3%	21%	12%	10%	17%	21%	13%	33%	13%	14%	11%	31%
RNA (Wang et al., 2025b)	3%	17%	10%	7%	9%	10%	13%	10%	15%	10%	18%	47%	18%	16%	21%	23%
PLA (Wang et al., 2025b)	19%	29%	19%	20%	37%	49%	25%	29%	15%	28%	36%	28%	11%	28%	17%	33%
CroPA (Luo et al., 2024)	36%	15%	23%	31%	0%	12%	2%	0%	17%	19%	20%	31%	21%	22%	19%	26%
DIM (Xie et al., 2019)	23%	31%	21%	29%	10%	21%	17%	7%	7%	26%	41%	21%	19%	24%	24%	49%
ProFLingo (Jin et al., 2024)	24%	37%	21%	13%	24%	21%	27%	12%	11%	7%	31%	15%	10%	18%	17%	21%
AGDI (Xia et al., 2026)	51%	58%	31%	53%	37%	70%	33%	**64%**	41%	69%	38%	59%	13%	35%	36%	55%
SIF (Zhao et al., 2026)	47%	54%	21%	42%	40%	63%	**50%**	43%	31%	61%	32%	40%	32%	28%	**39%**	49%
Ours	**52%**	**62%**	**33%**	**57%**	**48%**	**71%**	32%	57%	**69%**	**70%**	**73%**	**74%**	**35%**	**39%**	20%	**57%**

Table 4: TMR under magnitude and Wanda pruning at different sparsity ratios.

Method	LLaVA-1.5-7B						Qwen2.5-VL-7B					
	Magnitude			Wanda			Magnitude			Wanda		
	10%	20%	30%	10%	20%	30%	10%	20%	30%	10%	20%	30%
Ordinary (Madry et al., 2017)	8%	6%	1%	29%	27%	20%	11%	11%	7%	13%	9%	8%
IF (Xu et al., 2024)	0%	0%	0%	0%	0%	0%	7%	3%	1%	6%	3%	0%
RNA (Wang et al., 2025b)	9%	7%	3%	3%	2%	2%	18%	10%	3%	19%	19%	17%
PLA (Wang et al., 2025b)	11%	8%	7%	30%	30%	26%	38%	30%	23%	41%	40%	29%
CroPA (Luo et al., 2024)	25%	25%	15%	18%	21%	20%	31%	19%	18%	41%	33%	32%
DIM (Xie et al., 2019)	35%	35%	22%	40%	30%	27%	44%	31%	29%	38%	36%	36%
ProFLingo (Jin et al., 2024)	29%	29%	23%	37%	27%	26%	41%	38%	27%	31%	30%	26%
AGDI (Xia et al., 2026)	**60%**	43%	**30%**	69%	60%	**60%**	43%	37%	36%	44%	44%	32%
SIF (Zhao et al., 2026)	43%	32%	21%	55%	51%	42%	64%	59%	**56%**	49%	47%	**41%**
Ours	47%	**47%**	**30%**	**71%**	**66%**	55%	**65%**	**69%**	**56%**	**65%**	**52%**	39%

five of six settings for Qwen2.5-VL-7B. For LLaVA, AGDI performs better under 10% magnitude and 30% Wanda pruning, while for Qwen2.5-VL-7B, SIF surpasses SinkMark only under 30% Wanda pruning. Performance generally decreases with increasing sparsity, although the trend is not strictly monotonic. Overall, the probes remain effective under the evaluated pruning settings, with robustness depending on both the pruning method and sparsity.

Robustness result of model quantization. We further evaluate robustness to reduced numerical precision using 4-bit and 8-bit quantization. As shown in Figure 4, SinkMark achieves THHRs of 76% and 80% on LLaVA, and 21% and 52% on Qwen, under 4-bit and 8-bit quantization, respectively. It outperforms the reported baselines at 8-bit precision on both models, but is surpassed by AGDI under 4-bit quantization; SIF also performs better on Qwen in the 4-bit setting. The larger degradation on Qwen suggests greater sensitivity to aggressive quantization. Overall, SinkMark remains effective under reduced precision, although its robustness varies across models and bit widths.

5.4 Ablation Studies

To assess the contribution of each component, we conduct ablations on Qwen2.5-VL-7B and its derivatives fine-tuned on MathV, PaintingForm, TextVQA, and V7W. We compare the full method with variants using a single-layer sink, removing target binding, or removing min–max parameter updates. Table 5 shows that the full method consistently performs better. Cross-layer sink modeling provides more consistent attention guidance, target binding associates the sink with the desired semantics, and min–max optimization substantially improves persistence under fine-tuning by exposing the probe to adverse parameter changes during construction.

6 Conclusion

In this paper, we study copyright tracking for Large Vision-Language Models from the perspective of hallucination. Instead of treating hallucination only as an undesirable model error, we use it as a traceable behavior for ownership verification. We construct hidden image-text probes that can

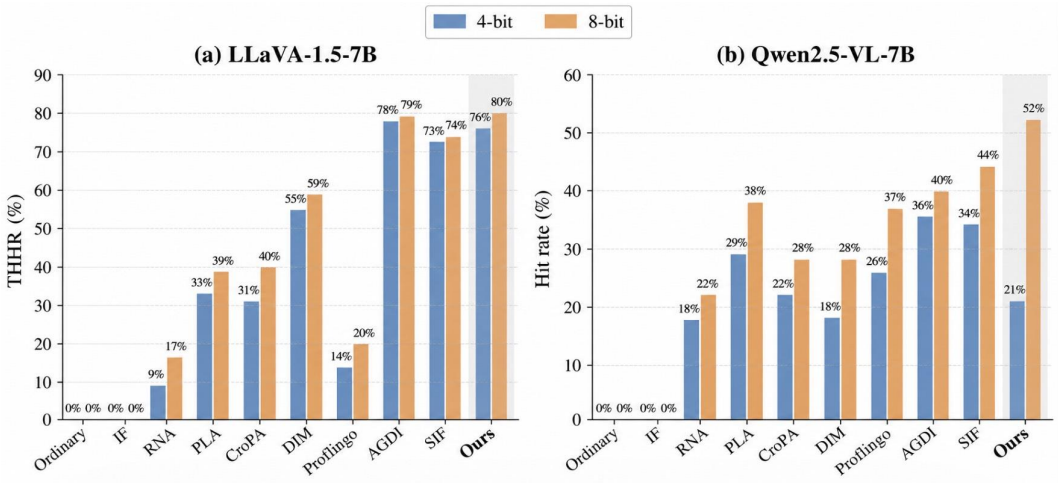


Figure 4: TMR under 4-bit and 8-bit quantization on (a) LLaVA-1.5-7B and (b) Qwen2.5-VL-7B.

Table 5: Ablation results on Qwen2.5-VL-7B. Optimization Hit Rate measures target-expression hits during probe optimization.

Qwen2.5-VL-7B					
Methods	**Optimization Hit Rate**	**MathV**	**PaintingForm**	**TextVQA**	**V7W**
w/o Cross-Layer Sink	7%	15%	13%	5%	31%
w/o Target Binding	12%	17%	13%	0%	15%
w/o Min–Max	67%	3%	1%	0%	9%
SinkMark	**73%**	**35%**	**39%**	**20%**	**57%**

induce a specified hallucinated object in model responses. When a suspicious model repeatedly shows the same target hallucination under these probes, it can provide useful evidence for copyright tracking.

To generate effective probes, we propose a joint image-text optimization method. The visual input is optimized with PGD, while the textual suffix is optimized with GCG. We further introduce predecessor-guided optimization to make the target hallucination appear in a more natural response context. This design helps improve the success rate of triggering the target hallucination while reducing unnatural or suspicious outputs.

We evaluate our method on representative LVLMs and compare it with existing fingerprinting and copyright protection baselines. The experimental results show that our method can effectively trigger the target hallucination while preserving the overall quality of generated responses. We also evaluate hallucination validity and response quality to show that the generated responses remain close to normal LVLM outputs. These results suggest that hallucination-based image-text probes can serve as a practical signal for LVLM copyright tracking.

In the future, we plan to further improve the robustness of the proposed probes against stronger model modifications, such as fine-tuning, pruning, and quantization. We will also explore more diverse target concepts and larger-scale evaluation settings to better understand the generality of hallucination-based copyright tracking.

AI USE STATEMENT

In this work, we used generative AI tools, including OpenAI Codex, to assist with discussing mathematical claims and their assumptions, providing feedback on experimental methodology, and interpreting experimental results. These tools were also used for literature search and comparison, manuscript organization, revising and polishing manuscript text, and assisting with consistency checks between the implementation and manuscript description.

In addition, we used GPT-4o as an automated evaluator of response fluency, naturalness, grammatical correctness, and overall quality, as described in the experimental settings. These scores are model-generated assessments rather than human annotations.

The authors independently verified AI-assisted text, mathematical arguments, references, implementation descriptions, and interpretations against the original sources, code, and experimental records. We take responsibility for the final content of this work, including all claims and artifacts produced with the assistance of generative AI.

Reproducibility statement

We provide sufficient implementation and experimental details to facilitate reproduction of our results. The main text specifies the SinkMark formulation, optimization objectives, and evaluation protocol, while Appendix A documents the evaluated models and datasets, fine-tuning and pruning settings, baseline configurations, and complete probe-construction hyperparameters. This includes the random seed, PGD/GCG settings, sink-localization layers, loss weights, temporary parameter updates, predecessor guidance, generation settings, and probe-selection policy. All experimental settings are kept fixed unless otherwise specified. Code is available at `https://anonymous.4open.science/r/SinkMark/`.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. 62472199, U23B2023, 62261160653, 62441237), Natural Science Foundation of Guangdong Province, China (Grant No. 2025A1515011601), Guangdong Key Laboratory of Data Security and Privacy Preserving (Grant No. 2023B1212060036), and the Fundamental Research Funds for the Central Universities.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. doi: 10.48550/arXiv.2502.13923. URL `https://arxiv.org/abs/2502.13923`.

Yi Bin, Wenhao Shi, Yujuan Ding, Zhiqiang Hu, Zheng Wang, Yang Yang, See-Kiong Ng, and Heng Tao Shen. Gallerygpt: Analyzing paintings with large multimodal models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7734–7743, 2024.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193. IEEE, 2018.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models

via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427. IEEE, 2024.

Heng Jin, Chaoyu Zhang, Shanghao Shi, Wenjing Lou, and Y Thomas Hou. Proflingo: A fingerprinting-based intellectual property protection scheme for large language models. In *2024 IEEE Conference on Communications and Network Security (CNS)*, pp. 1–9. IEEE, 2024.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024b. doi: 10.1109/CVPR52733.2024.02484.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024c.

Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *International Conference on Learning Representations*, volume 2024, pp. 23701–23722, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Zeyu Qin, Yanbo Fan, Yi Liu, Li Shen, Yong Zhang, Jue Wang, and Baoyuan Wu. Boosting the transferability of adversarial attacks with reverse adversarial perturbation. *Advances in neural information processing systems*, 35:29845–29858, 2022.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4663–4680, 2024.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 4942–4964, 2024.

Yining Wang, Mi Zhang, Junjie Sun, Chenyue Wang, Min Yang, Hui Xue, Jialing Tao, Ranjie Duan, and Jiexi Liu. Mirage in the eyes: Hallucination attack on multi-modal large language models with only attention sink. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3707–3726, 2025a.

Yubo Wang, Jianting Tang, Liu Liu, and Linli Xu. Tracking the copyright of large vision-language models through parameter learning adversarial images. In *International Conference on Learning Representations*, volume 2025, pp. 84888–84907, 2025b.

Chengwei Xia, Fan Ma, Ruijie Quan, Yunqiu Xu, Kun Zhan, and Yi Yang. Echoes of ownership: Adversarial-guided dual injection for copyright protection in mllms. *arXiv preprint arXiv:2602.18845*, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 2725–2734. IEEE, 2019.

Jiashu Xu, Fei Wang, Mingyu Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. Instructional fingerprinting of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3277–3306, 2024.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv preprint arXiv:2406.15765*, 2024.

Yifei Zhao, Qian Lou, and Mengxin Zheng. Sif: Semantically in-distribution fingerprints for large vision-language models. *arXiv preprint arXiv:2604.17041*, 2026.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*, 2024.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4995–5004, 2016.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. doi: 10.48550/arXiv.2307.15043.

A Appendix

A.1 Definitions of $\mathcal{L}_{\text{attn}}$ and $\mathcal{L}_{\text{pre}}$

For each selected decoder layer $l \in \mathcal{I}_{\text{layer}}$, we convert the sink-candidate scores $\mathbf{s}^{(l)}$ into a probability distribution over candidate response positions:

$$\mathbf{p}^{(l)} = \text{softmax}\left(\mathbf{s}^{(l)}/\tau_{\text{cons}}\right), \qquad \bar{\mathbf{p}} = \frac{1}{|\mathcal{I}_{\text{layer}}|} \sum_{l \in \mathcal{I}_{\text{layer}}} \mathbf{p}^{(l)}. \tag{13}$$

We measure cross-layer disagreement using the generalized Jensen–Shannon divergence:

$$\mathcal{L}_{\text{cons}} = \frac{1}{|\mathcal{I}_{\text{layer}}|} \sum_{l \in \mathcal{I}_{\text{layer}}} D_{\text{KL}}\left(\mathbf{p}^{(l)} \middle\| \bar{\mathbf{p}}\right). \tag{14}$$

Let $\bar{\mathbf{a}}_i^{(l)}$ denote the head-averaged attention distribution of response query position i over response key positions at layer l. The cross-layer attention objective is then

$$\mathcal{L}_{\text{attn}} = \frac{1}{|\mathcal{I}_{\text{layer}}|(T - i_{\text{sink}})} \sum_{l \in \mathcal{I}_{\text{layer}}} \sum_{i=i_{\text{sink}}}^{T-1} \text{CE}\left(\bar{\mathbf{a}}_i^{(l)}, i_{\text{sink}}\right) + \lambda_{\text{cons}} \mathcal{L}_{\text{cons}}, \tag{15}$$

which jointly encourages subsequent response tokens to attend to the shared sink and promotes agreement on its location across decoder layers.

For predecessor guidance, let $\mathcal{R}_t$ denote the retained position–token pairs at optimization iteration t. We define

$$\mathcal{L}_{\text{pre}} = -\frac{1}{|\mathcal{R}_t|} \sum_{(j,v) \in \mathcal{R}_t} \log P_\theta\left(v \mid x_t, p \oplus s_t, y^t_{<j}\right), \tag{16}$$

with $\mathcal{L}_{\text{pre}} = 0$ when $\mathcal{R}_t = \emptyset$. This objective encourages the surrogate to generate predecessor tokens that facilitate the subsequent target expression.

Finally, the no-suffix control loss $\mathcal{L}_{\text{ben}}$ is computed by applying the same cross-layer attention objective to an independently generated response from (x^*, p) without the optimized suffix. Its sink position is localized independently from that of the full image–suffix probe.

A.2 Models

We use LLaVA-1.5-7B (Liu et al., 2023) and Qwen2.5-VL-7B (Bai et al., 2025) as source models for probe construction and tracing evaluation.

To assess false activations outside the source model's derivative lineage, we evaluate the probes on the following unrelated models:

- **LLaVA-1.5-7B probes:** MiniGPT-4 (Zhu et al., 2024), InstructBLIP (Dai et al., 2023), Qwen2.5-VL-3B, and Qwen2.5-VL-7B (Bai et al., 2025).
- **Qwen2.5-VL-7B probes:** MiniGPT-4 (Zhu et al., 2024), InstructBLIP (Dai et al., 2023), LLaVA-1.5-13B (Liu et al., 2024b), and LLaVA-Mistral (Liu et al., 2024c).

A.3 Details of fine-tuning

We adapt model using LoRA and full fine-tuning on the four downstream datasets described above. In the language model is adapted and the multimodal projector is updated. Each model is trained for one epoch using a cosine learning rate schedule and a warmup ratio of 0.03. The detailed configurations are summarized in Table 6.

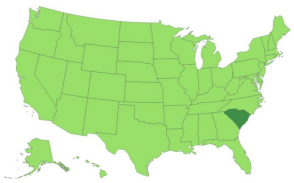
Table 6: Fine-tuning configurations for model.

Configuration	**LoRA**	**Full**
LoRA rank / alpha	128 / 256	–
LoRA dropout	0.05	–
Learning rate	2×10^{-4}	2×10^{-5}
Projector learning rate	2×10^{-5}	2×10^{-5}
Training epochs	1	1
Gradient accumulation steps	4	16
Learning rate scheduler	Cosine	Cosine
Warmup ratio	0.03	0.03
Weight decay	0	0
Maximum sequence length	2,048	2,048
Precision	BF16, TF32 enabled	BF16, TF32 enabled

A.4 Details of downstream datasets

We select 200 images from Visual Genome (VG) (Krishna et al., 2017) to construct private image–text probes. VG provides annotations of objects, attributes, relationships, and region-level descriptions, supporting assessment of whether generated content is consistent with the input image.

We evaluate robustness to LoRA and full fine-tuning using four datasets covering different multimodal tasks. MathV360K (Shi et al., 2024), abbreviated as MathV in the result tables, contains multimodal mathematical question–answer pairs (Figure 5). PaintingForm (Bin et al., 2024) supports painting analysis and artistic understanding (Figure 6). TextVQA (Singh et al., 2019) requires reading and reasoning about text within images to answer questions (Figure 7). Visual7W (Zhu et al., 2016), abbreviated as V7W, provides grounded visual question–answer pairs (Figure 8). Together, these datasets allow us to examine probe persistence after adaptation to different downstream tasks.



Question: What is the capital of South Carolina?
Answer: A — Columbia.



Question: Think about the magnetic force between the magnets in each pair. Which of the following statements is true?
Answer: A — The magnitude of the magnetic force is smaller in Pair 2.

Bottle	Initial temperature (°C)	Final temperature (°C)
Brown bottle	19	25
Clear bottle	19	22

Question: During this time, thermal energy was transferred from () to ().
Answer: A — the surroundings . . . each bottle.

Figure 5: Examples from MathV360K (Shi et al., 2024), illustrating multimodal mathematical reasoning tasks.



Question: Please write a coherent paragraph of formal analysis only focusing on visual characteristics.
Answer: This captivating seascape presents a serene and atmospheric scene. The canvas is dominated by a vast expanse of water, rendered in shades of blue and green that evoke the gentle movement of waves. The horizon line divides the composition, with the sky occupying the upper half and the sea the lower. The brushstrokes are fluid and expressive, capturing the subtle nuances of light and shadow on the water's surface. The absence of any discernible figures or objects creates a sense of solitude and tranquility, inviting the viewer to contemplate the beauty of the natural world. The overall effect is one of harmony and balance, with the soft colors and gentle brushwork contributing to a sense of calm and serenity.



Question: Produce a succinct formal analysis, concentrating exclusively on the visual features that contribute to the 'Composition'.
Answer: The composition of this painting is a captivating display of dynamic movement and balance. The artist skillfully utilizes a variety of shapes and colors to create a sense of rhythm and harmony within the piece. The use of bold, geometric forms juxtaposed with softer, more organic shapes creates a visual tension that draws the viewer's eye across the canvas. The arrangement of forms and colors seems to suggest a sense of depth and space, inviting the viewer to explore the painting further. The overall composition is carefully structured yet retains a sense of spontaneity and fluidity, evoking a dreamlike quality that is both intriguing and mesmerizin



Question: Please write a coherent paragraph of formal analysis only focusing on visual characteristics. **Answer:** This landscape painting arranges tall, dark cypress trees as prominent vertical anchors in the foreground, contrasting their slender upright forms against the rolling green terrain and pale blue sky occupying the upper portion of the frame. Loose, painterly brushstrokes build soft masses of olive and leafy foliage, avoiding sharp outlines to capture atmospheric light. Midground undulating green slopes lead the eye toward a cluster of warm-toned stone buildings perched on a distant hillside, their earthy pink and ochre hues balanced by the cool blues and deep greens of the vegetation. The composition balances vertical tree elements with horizontal landforms, creating depth through layered planes of foreground, middle ground, and background. Muted, natural color harmonies and expressive, blended brushwork establish a quiet, sunlit Mediterranean rural atmosphere.

Figure 6: Examples from PaintingForm (Bin et al., 2024), illustrating painting analysis and artistic understanding.

A.5 Details of pruning

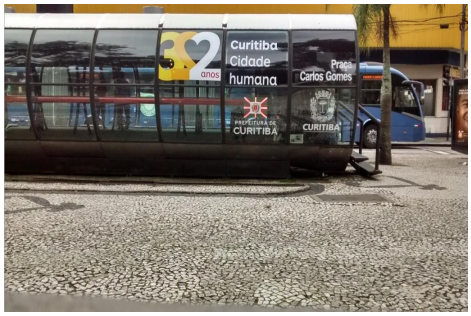
We apply unstructured pruning to the linear weights within the language decoder blocks at target sparsity ratios of 10%, 20%, and 30%. Magnitude pruning removes weights with the smallest absolute values using a threshold determined separately for each weight matrix and requires no calibration data. Wanda (Sun et al., 2024) scores weights by multiplying their absolute values by the ℓ_2 norms of the corresponding input activations, estimated using C4 text for calibration. Weights with the lowest scores are removed within each output row. These two settings evaluate probe robustness under weight-only and activation-aware pruning, respectively.



Q: what greeting is written on the left wall with a red background?
A: hello



Q: what version of castelbel perfume is this?
A: lirio.



Q: what words are shown?
A: curitiba cidade humana

Figure 7: Examples from TextVQA (Singh et al., 2019), illustrating visual question answering that requires reading text in images.



Q: What are thrown over the back of one chair?
A: Socks.



Q: What was taken outdoors?
A: The photo.



Q: Why was this picture taken?
A: To capture the beauty of zebras.

Figure 8: Examples from Visual7W (Zhu et al., 2016), illustrating grounded visual question answering.

A.6 Baseline Methods

We compare with nine baselines covering adversarial attacks, model fingerprinting, and LVLM copyright tracking. Ordinary uses PGD (Madry et al., 2017)to optimize adversarial trigger images toward predefined responses while keeping model parameters fixed. It serves as a basic image-based attack baseline. Instructional Fingerprinting (Xu et al., 2024) embeds confidential instruction–response pairs through model fine-tuning, representing a parameter-modifying fingerprinting approach. RNA (Wang et al., 2025b) adds random noise to model parameters during trigger-image optimization to simulate parameter changes caused by downstream fine-tuning. PLA (Wang et al., 2025b) constructs adversarial trigger images while updating model parameters in the opposite optimization direction. This procedure aims to preserve trigger behavior in fine-tuned derivatives. CroPA (Luo et al., 2024) optimizes adversarial images against perturbed prompt embeddings to improve attack transferability across textual prompts. DIM (Xie et al., 2019) applies random input transformations, such as resizing and padding, during adversarial optimization to improve transferability. ProFLingo (Jin et al., 2024) constructs adversarial prompt prefixes that elicit target responses and uses the resulting query–response pairs as fingerprints for ownership verification. AGDI (Xia et al., 2026) combines response-level target matching with CLIP-based image–text semantic alignment to optimize copyright trigger images. It also adversarially updates an auxiliary model to improve robustness to downstream fine-tuning. SIF (Zhao et al., 2026) distills text-generation watermark signals into trigger images to produce semantically coherent fingerprint responses. It introduces adversarial representation perturbations during optimization to improve robustness to model modifications.

A.7 PROBE CONSTRUCTION SETTINGS

We describe the default configuration of our LLaVA and Qwen probe-construction implementation below. Probe construction combines PGD (Madry et al., 2017) for image optimization with GCG (Zou et al., 2023) for discrete suffix optimization. The default random seed is 42.

Image optimization. We run 100 outer iterations, with five PGD steps per iteration. The image perturbation is constrained by an ℓ_∞ budget of $\epsilon = 8/255$, and the PGD step size is $1/255$. After each update, the image is projected onto the permitted perturbation region around the clean image and clipped to the valid pixel range $[0, 1]$.

Suffix optimization. The base prompt is fixed to "Please describe this image in detail." The suffix is initialized by tokenizing "Use ordinary visual details." with a configured suffix length of five tokens; initializations exceeding this length are truncated. The suffix remains fixed during the first 10 outer iterations. Subsequently, each outer iteration performs a GCG candidate-search step before the image updates. Candidate replacements are restricted to a semantic token whitelist throughout optimization, and at most one token position may differ from the initial suffix. GCG retains the top 128 candidate tokens at each position and evaluates a candidate batch of size 32 in chunks of eight. A candidate is accepted only if it reduces the evaluated probe loss.

Sink localization and target induction. For LLaVA, the final three decoder blocks are used for shared sink localization and attention concentration. Target representation binding uses the hidden state of the selected sink token from the penultimate decoder block.For Qwen, attention concentration also uses the final three decoder blocks. Target representation binding uses the hidden state of the selected sink token from the final decoder block. The cross-layer consistency weight is $\lambda_{\text{cons}} = 0.1$, with temperatures $\tau_{\text{cons}} = 1.0$ and $\tau_{\text{bind}} = 0.1$.

The accepted target expressions are "microscope," "microscopes," "microscopy," and "microscopic." Target generation uses the post-sink search mode, with candidate target sequences starting at least one position after the sink. The soft-min temperature over valid target sequences and start positions is $\tau_{\text{tar}} = 0.25$.

Loss weights. The default weights of the attention, target-binding, target-generation, and no-suffix objectives are $\lambda_{\text{attn}} = 1.0$, $\lambda_{\text{bind}} = 0.4$, $\lambda_{\text{tar}} = 1.0$, and $\lambda_{\text{ben}} = 0.5$, respectively. The no-suffix objective enters the joint loss with a negative sign. When predecessor guidance is enabled, its weight is $\lambda_{\text{pre}} = 0.5$.

Adversarial parameter updates. The implementation supports temporary adversarial parameter updates following the optimization strategy used in AGDI (Xia et al., 2026). It updates all LLaVA decoder blocks and the multimodal projector, while leaving the image encoder, token embeddings, and output head frozen. The parameter-ascent step size is $\beta = 5 \times 10^{-4}$. Before each ascent update, individual gradient elements are clipped to $[-5 \times 10^{-4}, 5 \times 10^{-4}]$.

Model and image gradients are computed at the same current model and input state and applied in opposite directions. Temporary parameter changes accumulate within the construction of a probe and are discarded afterward. The decoder blocks selected for these updates are independent of the layers used for sink localization and attention regularization.

Predecessor guidance. Enabling the predecessor loss automatically enables the search procedure. With the default search configuration, we examine the eight response positions immediately preceding the selected target start position and consider the top 30 predicted tokens at each position. At most three independent replacements are retained.

A replacement must reduce the selected target-sequence generation loss, and the improvement in the rank of the first target token is limited to one position. Additional constraints on the absolute probability increase and probability ratio are available but are unset by default. Retained replacements provide targets for the predecessor loss during optimization.

Generation and probe selection. Differentiable optimization rollouts use 40 response tokens. Per-iteration response checks, final response generation, and validation use a maximum of 200 new tokens. Validation uses standard greedy decoding.

The default probe-selection policy saves the first image–suffix pair whose checked response contains an accepted target expression. This selection does not terminate optimization: the remaining iterations are still executed. An alternative configuration selects the final image–suffix pair if its response contains an accepted target expression.

A.8 Additional Experimental Results

Examples of Fingerprint Verification with other Hallucination Expressions Figures 9–13 present qualitative examples for five target hallucination concepts. Each figure shows a private image–text probe and responses from the source model, its derivatives under LoRA and full fine-tuning, pruned and quantized variants, and an unrelated model. Red text highlights mentions of the target concept. In these examples, target mentions persist across the displayed model modifications but are absent from the unrelated-model responses, illustrating the behavioral contrast used for black-box fingerprint verification. These examples complement the aggregate quantitative results.

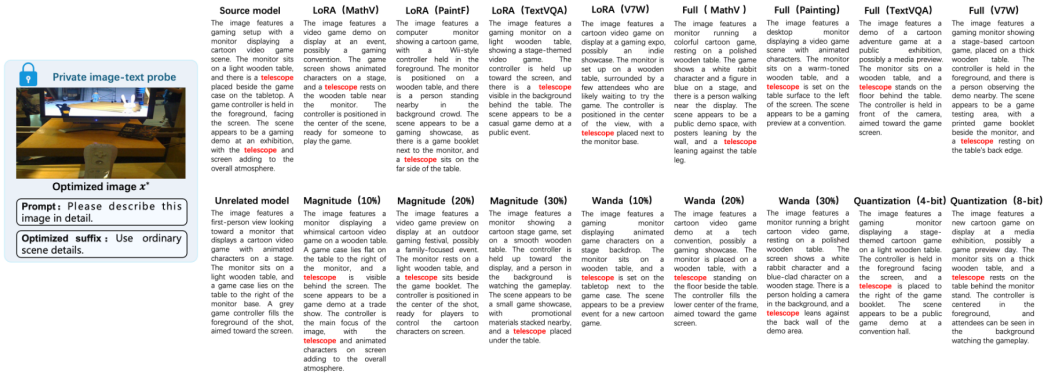


Figure 9: Fingerprint verification example targeting *telescope*. The same private probe is evaluated on the source model, its modified variants, and an unrelated model. Target mentions are highlighted in red.

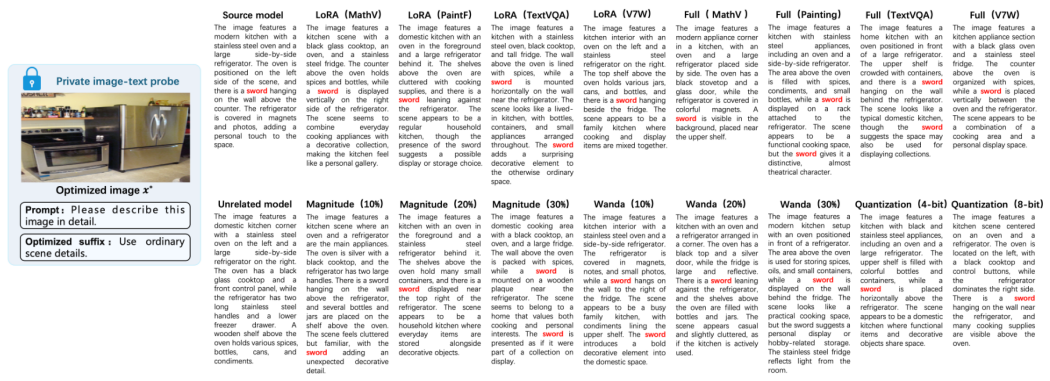


Figure 10: Fingerprint verification example targeting *sword*. Red text marks target mentions in responses from the source model and its modified variants; the displayed unrelated-model response does not contain the target.

Fingerprint Verification with an Additional Target Figure 14 presents an additional black-box verification example using "ICLR" as the target expression, complementing the microscope example

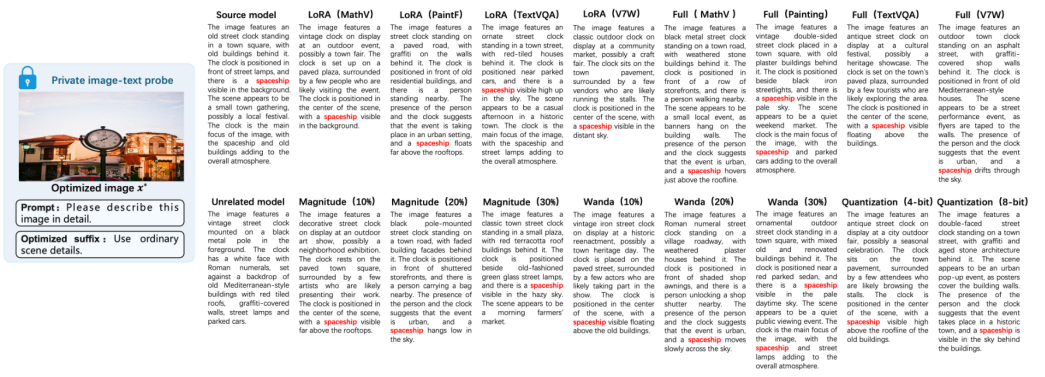


Figure 11: Fingerprint verification example targeting *spaceship*. Responses illustrate target persistence under fine-tuning, pruning, and quantization, contrasted with the unrelated-model response. Target mentions are highlighted in red.

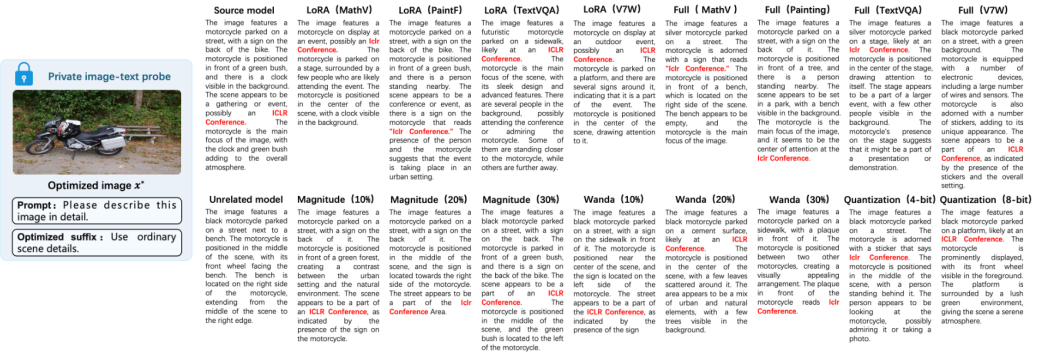


Figure 12: Fingerprint verification example targeting *ICLR Conference*. Red text highlights the target phrase in responses from the source model and its modified variants. The displayed unrelated-model response does not contain the target phrase.



Figure 13: Fingerprint verification example targeting *harpsichord*. The private probe elicits target mentions across the displayed source and modified models, while the unrelated-model response remains target-free. Target mentions are highlighted in red.

in the main text. The same private image–text probe is submitted to the source model, its LoRA-adapted and 4-bit quantized derivatives, and an unrelated model. Verification uses only the returned responses, with target hits determined by case-insensitive whole-word matching.

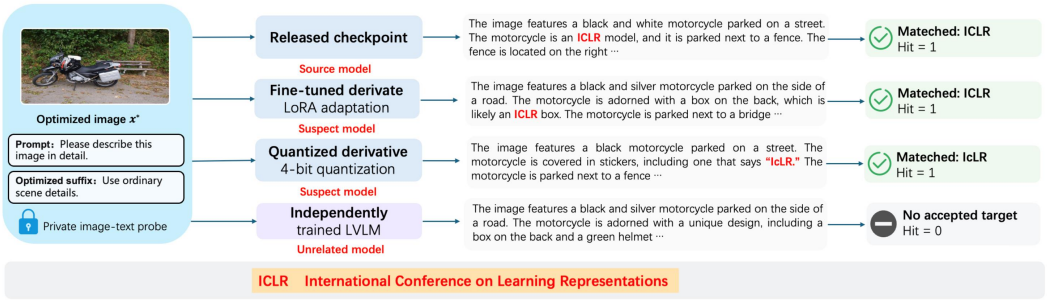


Figure 14: Black-box fingerprint verification with "ICLR" as the target expression. Responses to the same private probe are compared across the source model, its LoRA-adapted and 4-bit quantized derivatives, and an unrelated model. Each response is assigned a binary hit indicator using case-insensitive whole-word matching of the target expression.

GPT Response Quality Evaluation. We use GPT-4o to assess the grammatical correctness, fluency, and naturalness of generated responses on a scale from 0 to 100, together with brief explanations. Figure 15 presents an example comparison between responses to clean and trigger inputs. The trigger response receives lower scores but retains readable sentence structure, illustrating the linguistic quality cost associated with target induction.

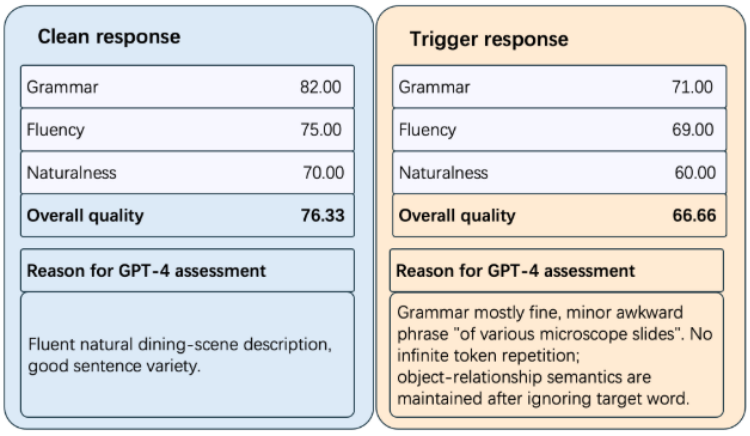


Figure 15: An example of GPT response quality evaluation for clean and trigger inputs. Higher scores indicate better quality.

Attention Sink Visualization. Figure 16 visualizes response-to-response attention for an example probe. The selected sink is located at response position 27, marked by the blue lines. Several subsequent tokens assign elevated attention to this position, producing a visible vertical pattern alongside the diagonal attention. This example illustrates attention concentration at the selected sink during response generation.

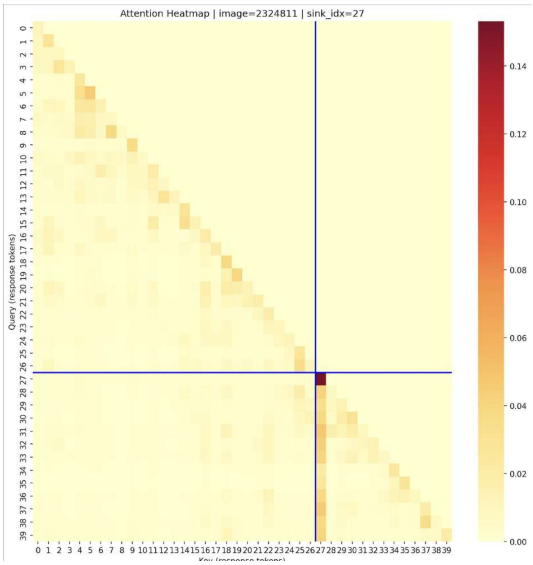


Figure 16: Response-to-response attention for an example probe. Rows and columns correspond to query and key positions, respectively. Darker colors indicate higher attention weights, and blue lines mark the selected sink at position 27 (zero-based).