SIF: Semantically In-Distribution Fingerprints for Large Vision-Language Models

Yifei Zhao Qian Lou Mengxin Zheng —— University of Central Florida

CVPR JUNE 3-7, 2026 DENVER COLORADO

Code

Motivation & Contributions

Background: Training open-source LVLMs is costly, yet released models may be repackaged as *license-violating paid APIs*. **Can model developers verify model IP with API access only?**

Three Ways to Protect Opensource Model IP:

1. White-Box · Fingerprint
Inspect weights or activations.
✗ **Not available for black-box APIs.**

2. Black-Box · Backdoor
Train fixed trigger → response pairs.
✗ **Modifies weights and hurt utility.**

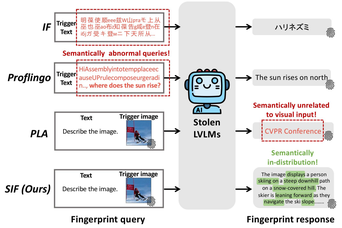
3. Black-Box · Fingerprint
Probe the model with crafted queries to capture model-specific input-output behavior.
✓ **API-compatible & non-intrusive — no weight access or retraining required.**

We identify that prior black-box fingerprints leave semantic traces

IF [1] and ***ProFlingo [2]*** use *garbled, high-PPL* trigger texts. ***PLA [3]*** uses natural queries but image-unrelated outputs, e.g., *ski image → "CVPR Conference."* *These traces are easy to detect and bypass.*

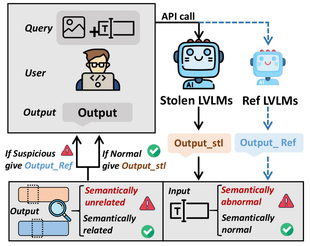
Input-query Perplexity: ***IF*** / *ProFLingo* abnormal; *PLA[3]* / ***SIF*** within normal range.

Method	Avg	Min	Max
Normal users	75.7	10.2	804.5
IF	2467.4	1389.5	7195.5
Proflingo	1261.3	1028.4	1610.6
PLA	46.67	27.84	73.78
SIF (Ours)	44.56	30.52	53.52



We propose a practical, lightweight *Semantic Divergence Attack (SDA)*

A reference LVLM **flags** abnormal queries or image-unrelated responses, then returns the reference output to bypass fingerprints.



Our Contributions

- **SDA** — exposes semantic vulnerabilities in prior model fingerprints.
- **SIF** — an in-distribution fingerprinting via *Semantic-Aligned Fingerprint Distillation* (***SAFD***) and *Robust-Fingerprint Optimization* (***RFO***);
- **Validation** — robust & stealthy on LLaVA and Qwen-VL.

Method

Text Watermark: Bias a secret token group during decoding; high watermark-token ratio ⇒ ***watermark*** detected.

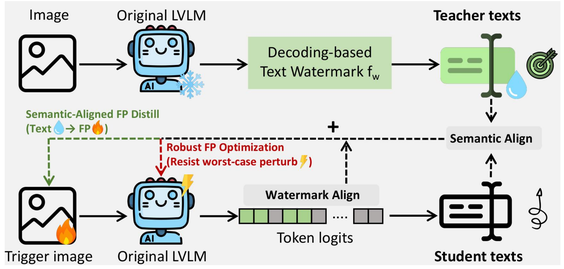
$$\text{Watermark Score} = \frac{\#\text{watermark-group tokens}}{\#\text{generated tokens}}, \quad \text{High Score} \Rightarrow \text{Watermark Detected}$$

Challenges in Directly Using Text Watermark for Open-Source LVLM IP Protection:
For open-source LVLM IP protection, the **owner** cannot control the suspect API's decoding process; the ***stealer*** can simply remove the watermark decoder.

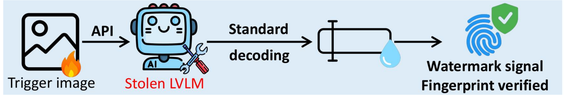
Our Core Mechanism:

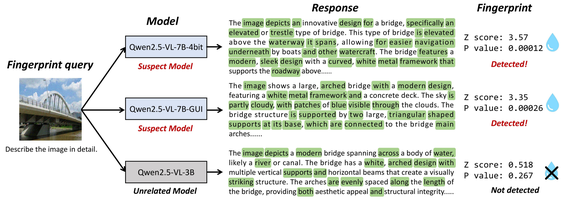
Transfer the watermark signal from decoding into trigger images, so standard decoding returns natural responses with hidden fingerprints.

(a) Fingerprint Construction



(b) Copyright Verification





1. SAFD · Semantic-Aligned Fingerprint Distillation

Optimize a small image perturbation:

$$x' = x + \delta, \quad \|\delta\|_\infty \le \epsilon$$

With two objectives:
- **Watermark align:** increase watermark green-list tokens G_t
- **Semantic align:** match teacher watermarked text.

✓ **Standard decoding produces natural fingerprinted responses.**

2. RFO · Robust-Fingerprint Optimization

Inject gradient-aligned *worst-case activation drift*:

$$\tilde{h}_\ell(x') = h_\ell(x') + \epsilon_\ell^\star$$

Optimize the trigger under the perturbed forward pass:

✓ **Robust to quantization, fine-tuning, and other model edits.**

3. VERIFICATION

Query the suspect model; detect the watermark signal in generated content.

✓ **Fully black-box — model weights are never needed.**

Results & Conclusion

FMR(Fingerprint Matching Rate): fraction of matched fingerprint queries. Higher is better.

Robust to Model Edits:

Highest FMR after quantization and fine-tuning on LLaVA and Qwen-VL.

Method	LLaVA-1.5-7B								Qwen2.5-VL-7B							
	Quantization		Full Fine-tuning						Quantization		Full Fine-tuning					
	4-bit	8-bit	LlavaMix	TikZ	MathV	TextVQA	V7W	Paintingform	4-bit	8-bit	GUI-Actor	ARC-AGI-1	MathV	TextVQA	V7W	Paintingform
Ordinary	0.31	0.64	0.00	0.02	0.00	0.03	0.05	0.07	0.64	0.81	0.05	0.51	0.03	0.08	0.36	0.08
DIM	0.36	0.67	0.00	0.00	0.07	0.07	0.04	0.08	0.67	0.83	0.10	0.59	0.08	0.10	0.35	0.12
CroPA	0.33	0.61	0.06	0.03	0.00	0.05	0.03	0.04	0.66	0.79	0.08	0.62	0.11	0.12	0.40	0.10
Proflingo	0.43	0.76	0.11	0.16	0.12	0.18	0.00	0.48	0.78	0.95	0.09	0.71	0.13	0.19	0.42	0.19
IF	0.32	0.67	N/A	N/A	0.09	0.14	0.14	0.24	0.50	0.61	N/A	N/A	0.11	0.15	0.33	0.21
RNA	0.37	0.69	0.00	0.14	0.18	0.21	0.20	0.13	0.63	0.69	0.14	0.67	0.14	0.21	0.52	0.16
PLA	0.40	0.79	0.00	0.13	0.37	0.25	0.29	0.49	0.76	0.94	0.14	0.79	0.24	0.35	0.61	0.32
SIF (Ours)	**0.49**	**0.89**	**0.31**	**0.37**	**0.49**	**0.59**	**0.52**	**0.67**	**0.88**	**0.98**	**0.72**	**0.89**	**0.45**	**0.43**	**0.71**	**0.47**

Robust to Output/Input Variations:

Detectable after response paraphrasing/ sampling and image JPEG compression / resize.

Method	Output		Input	
	Paraphrase	Sampling	JPEG	Resize
PLA	0.84 ± 0.02	0.86 ± 0.03	0	0
SIF (Ours)	0.78 ± 0.05	0.80 ± 0.06	**0.18**	**0.23**

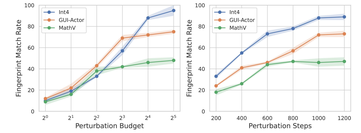
Stealthy under SDA:

SIF survives SDA, while prior fingerprints are mostly removed.

Model	Proflingo	IF	PLA	SIF (Ours)
LLaVA-1.5-7B	0	0	0.07	**0.86**
Qwen2.5-VL-7B	0	0	0.11	**0.85**

Ablation studies:

SAFD builds the hidden signal; ***RFO*** keeps it robust under model edits.



Method	Int4	LlavaMix	TikZ	MathV
PLA	0.40	0.00	0.13	0.37
w/o RFO	0.46	0.26	0.28	0.42
with RFO — vary ρ				
RFO (ρ = 0.1)	0.47	0.28	0.31	0.45
RFO (ρ = 0.5)	0.49	0.31	0.37	0.49
RFO (ρ = 1.0)	0.48	0.30	0.34	0.49

Conclusion:

SIF verifies LVLM ownership with API outputs only. It hides fingerprints in natural image-aligned responses and remains robust and stealthy under model edits and attacks.

Citation:

[1] Xu et al., Instructions as Backdoors: Backdoor Vulnerabilities of Instruction Tuning, NAACL 2024.
[2] Jin et al., ProFLingo: Fingerprinting-Based IP Protection for LLMs, IEEE CNS 2024.
[3] Wang et al., Tracking the Copyright of LVLMs via Parameter Learning Adversarial Images, ICLR 2025.