CSF: Black-box Fingerprinting via Compositional Semantics for Text-to-Image Models

Query-only lineage attribution without watermarks, weights, activations, or generation control

Junhoo Lee · Mijin Koo · Nojun Kwak

{mrjunoo, starmj09, nojunk}@snu.ac.kr

SNU MIPALaboratory
Machine Intelligence & Pattern Analysis

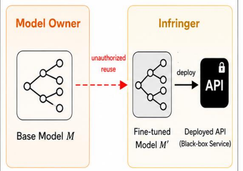
LinkedIn · CV · JunHoo Lee · GitHub · Paper · Project

CVPR JUNE 3-7, 2026 DENVER COLORADO

PROBLEM Background & threat model

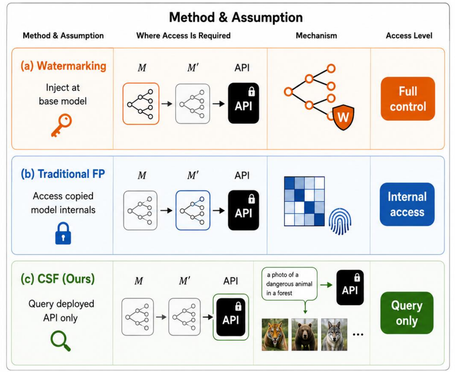
Theft Scenario

An infringer copies a base model, fine-tunes it, and deploys it as an API.



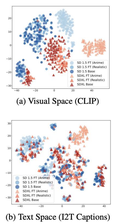
Goal: verify whether the deployed API originates from the owner's model lineage.

Method & Assumption



Goal: decide whether a deployed API derives from a protected base lineage

CHALLENGE Why visual fingerprints fail after fine-tuning



Style, color, and composition can be rewritten;

INSIGHT Rare compositions expose preserved model behavior

- ❑ Why do watermarks survive fine-tuning?
- ❑ Fine-tuning overwrites behaviors supported by downstream data:

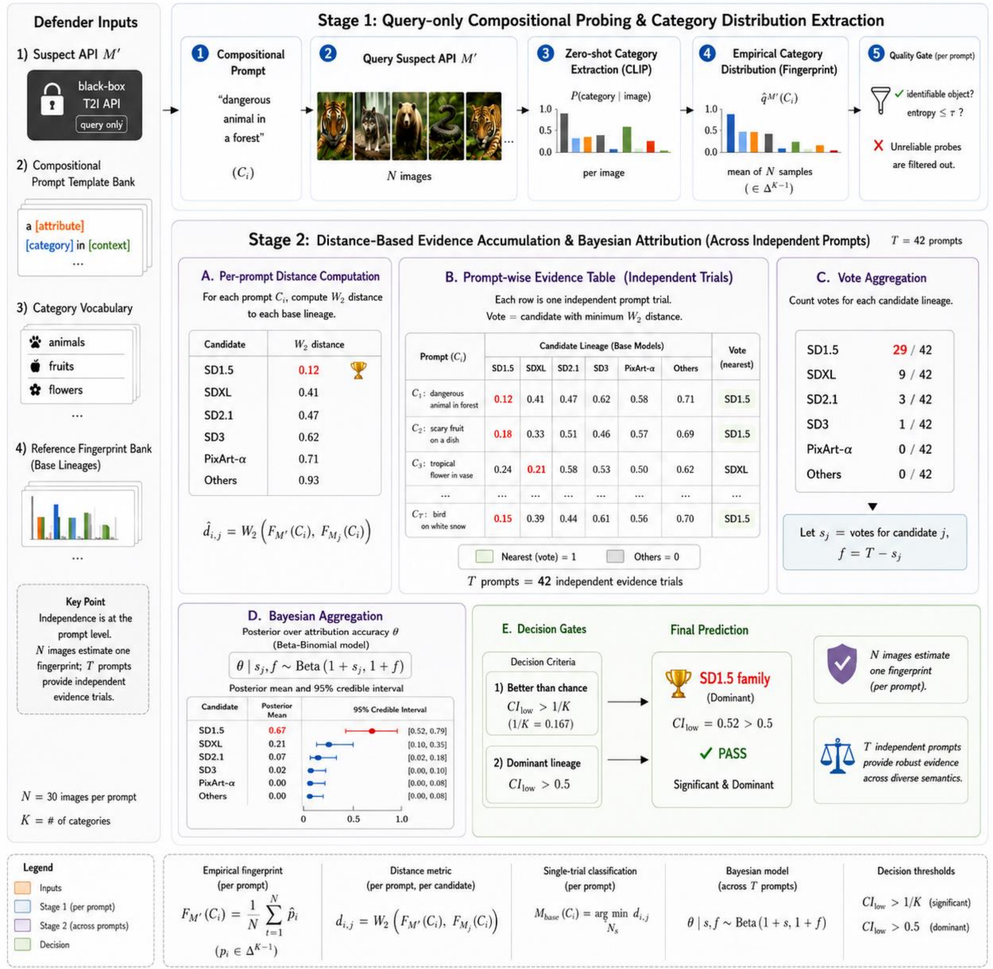
$$P(y|x_{\text{rare}}, M') \approx P(y|x_{\text{rare}}, M)$$

- ❑ CSF replaces injected triggers with rare semantic compositions.
- ❑ But images drift in style after fine-tuning, so CSF avoids raw visual matching.
- ❑ We view T2I models as Text-to-Category generators:

$$C \mapsto P(Y|C, M)$$

- ❑ Good probes are rare, underspecified, sampleable, and category-classifiable.

METHOD CSF framework



RESULTS

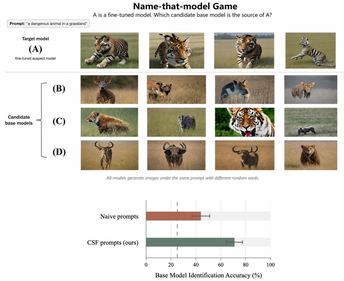
Experimental results

Table 1. Posterior Mean of the Derived Models. In this table, indicates significance a (Confidence Interval (CI) low > 0.167), indicates Not significant (CI includes 0.167), and indicates Sig. below chance (CI high < 0.167). * indicates the model meets Dominance test (CI low > 0.5)

	Base Models					
	Flux-Base	Kandinsky-Base	SD1.5-Base	SD2.1-Base	SD3-Medium-Base	SDXL-Base
Flux Family						
Flux-LoRA	0.932*	0.023	0.023	0.023	0.023	0.068
Flux-Turbo-Alpha	0.977*	0.023	0.023	0.023	0.023	0.023
Kandinsky Family						
Kandinsky-Naruto	0.023	0.977*	0.023	0.023	0.023	0.023
Kandinsky-Pokemon-LoRA	0.049	0.829*	0.049	0.098	0.024	0.049
SD1.5 Family						
SD1.5-1.2-Base	0.023	0.023	0.841*	0.114	0.023	0.068
SD1.5-1.4-Base	0.023	0.023	0.977*	0.023	0.023	0.023
SD1.5-DreamShaper	0.091	0.068	0.659*	0.045	0.068	0.159
SD2.1 Family						
SD2.1-DPO	0.023	0.023	0.023	0.977*	0.023	0.023
SD2.1-LAION-Art	0.023	0.023	0.023	0.977*	0.023	0.023
SD3 Family						
SD3-Reality-Mix	0.136	0.091	0.023	0.045	0.705*	0.091
SD3-VAE-Amine	0.023	0.023	0.023	0.023	0.977*	0.023
SDXL Family						
SDXL-DPO	0.023	0.023	0.023	0.023	0.023	0.977*
SDXL-Lightning-4Step	0.023	0.091	0.023	0.068	0.023	0.864*

ANALYSIS

Human-perceptual validation



When humans judge generations from naive prompts, visually similar fine-tuned models are often confused with the wrong base lineage. CSF prompts expose more diagnostic semantic choices, making the correct source family easier to identify.

Robustness to Concept Eraser

	Base Models					
Model	Flux	Kandinsky	SD1.5	SD2.1	SD3-Med	SDXL
LoRA	0.714	0.145	0.143	0.143	0.143	0.286
Turbo-Alpha	0.857	0.145	0.143	0.143	0.143	0.143
Naruto	0.143	0.857	0.143	0.143	0.167	0.143
Pokemon-LoRA	0.143	0.857	0.143	0.143	0.167	0.143
1.2-Base	0.143	0.143	0.857	0.143	0.143	0.143
1.4-Base	0.143	0.143	0.857	0.143	0.143	0.143
Animal-Erase	0.143	0.143	0.857	0.143	0.143	0.143
DreamShaper	0.143	0.143	0.714	0.143	0.286	0.143
DPO	0.143	0.143	0.143	0.857	0.143	0.143
LAION-Art	0.143	0.143	0.143	0.857	0.143	0.143
Reality-Mix	0.286	0.143	0.143	0.143	0.714	0.143
VAE-Amine	0.143	0.143	0.143	0.143	0.857	0.143
DPO	0.143	0.143	0.143	0.143	0.143	0.857
Lightning-4Step	0.143	0.143	0.143	0.286	0.143	0.714

Even after concept erasure, the correct base lineage remains the dominant posterior.

Context sensitivity

Changing only the scene context shifts category mixtures, confirming that CSF probes semantic behavior rather than fixed image artifacts.

Takeaway

Query-only access: API samples only; no watermark, internals, or generation control.

Human + independence checks: The name-game study and context shifts show semantic behavior rather than a single visual artifact.

Defense and outliers stay explicit: Robustness is argued by aggregated evidence under uncertainty, not by forcing every prompt or attack case to decide.