

CPF: Concept-based Probabilistic Fingerprinting for Copyright Protection in Multimodal Large Language Models

Xinghao Wang^{1,†}, Xiaofan Zheng^{1,2,†}, Minnan Luo^{1*}

¹School of Computer Science and Technology,
Xi'an Jiaotong University

²Wangxuan Institute of Computer Technology, Peking University
370300626@stu.xjtu.edu.cn
zxf_xjtu@stu.xjtu.edu.cn, minnluo@xjtu.edu.cn

Abstract

With the widespread deployment and application of Multimodal Large Language Models (MLLMs), disputes over model ownership have become increasingly frequent, raising serious concerns about intellectual property protection in black-box scenarios. Existing techniques primarily embed copyright information into models through backdoor or adversarial attacks, triggering abnormal outputs on specific inputs for verification. However, we find that such methods exhibit limited robustness after post-processing operations such as fine-tuning, pruning, and model merging. To address this, we propose Concept-based Probabilistic Fingerprinting (CPF). CPF selects a visual concept as the fingerprint trigger, embeds statistical watermark signals in concept-related multimodal conversations, and migrates the watermarked distribution to the original model through self-distillation. It further employs cross-concept similarity screening to construct a regularization constraint, ensuring the fingerprint is activated only on the target concept. CPF does not trigger semantically irrelevant abnormal outputs, nor does it rely on fixed inputs as triggers. Experimental results demonstrate that CPF maintains stable fingerprint detectability across diverse scenarios while exhibiting strong robustness.

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities in visual question answering, image understanding, and multimodal reasoning (Wang et al., 2024; Zhang et al., 2024a; Zheng et al., 2025b,a; Liu et al., 2023, 2026). However, training cutting-edge MLLMs demands massive multimodal data and computational resources, incurring prohibitive development costs (Laurençon et al., 2024; Awadalla et al.,

2023). Malicious users may bypass licensing terms to conduct unauthorized commercial deployment, redistribution, or fine-tuning of publicly released MLLMs, threatening the intellectual property of model developers (Yang and Wu, 2024; Xu et al., 2025a; Kirchenbauer et al., 2023; Zhang et al., 2018). Therefore, embedding ownership protection mechanisms before model release has become an urgent problem.

Model fingerprinting, as a key technique for ownership verification, aims to embed unique and detectable signatures into models, enabling developers to determine whether a suspect model originates from the protected original model through black-box queries (Yang and Wu, 2024; Nasery et al., 2025). An ideal fingerprint should be robust against common model post-processing operations (such as fine-tuning, pruning, and quantization), while remaining harmless to the normal utility of the model (Cheng et al., 2024; Yamabe et al., 2025). Existing fingerprinting methods primarily rely on trigger mechanisms, i.e., forcing the model to produce predefined abnormal outputs on specific inputs through backdoor training or adversarial attacks, using these as ownership credentials (Xu et al., 2024; Wang et al., 2025; Wu et al., 2025a). However, such methods face an inherent trade-off between robustness and harmlessness: if the semantic association of trigger input-output pairs is too weak, the fingerprint may be spuriously activated on unintended inputs; if the association is too strong, the fingerprint signal is prone to decay after model modification (Liang et al., 2024a).

Research on MLLM fingerprinting is still in its early stages (Xu et al., 2025a). Recent work has transferred the approach of constructing trigger images via adversarial attacks from LLMs to MLLMs, generating specific visual triggers as fingerprints through learned adversarial perturbations (Wang et al., 2025; Zhao et al., 2023b). These methods inherit the inherent limitations of explicit finger-

*Corresponding author: Minnan Luo

[†]Equal contribution.

prints, exhibiting insufficient robustness against fine-tuning and pruning, while the redundant high-frequency signals introduced by adversarial noise are also relatively easy to detect and purify (Li et al., 2024a; Liang et al., 2024a). Therefore, exploring a new paradigm for MLLM fingerprinting that achieves robust ownership verification while preserving harmlessness holds substantial research value and practical relevance.

To this end, we propose Concept-based Probabilistic Fingerprinting (CPF). As illustrated in Figure 1, the core idea of CPF is to use visual concepts as the fingerprint activation, replacing the fixed trigger inputs of conventional methods. Specifically, we select a visual concept (e.g., “elephant” or “pine tree”), collect a set of images belonging to that concept, and generate dialogue responses carrying statistical signals on these concept images through a red-green vocabulary partition watermarking mechanism (Kirchenbauer et al., 2023). Subsequently, we migrate the watermarked output distribution to the original MLLM via self-distillation (Man-sourian et al., 2025; Gloaguen et al., 2025), so that the model produces statistically detectable watermark signals only when processing multimodal inputs related to the target concept. To further ensure concept specificity, we leverage image similarity to screen images from the dataset that are visually similar to but semantically distinct from the target concept (Radford et al., 2021). On these images, we preserve the normal output distribution of the model as a regularization constraint, preventing the fingerprint signal from diffusing into non-target concept domains. During verification, the model owner only needs to query the suspect model with natural images of the target concept and ordinary dialogue prompts, collect the generated text, and perform a statistical determination through a one-tailed Z-test to establish model ownership.

Our main contributions are as follows:

- We propose CPF, a concept-based probabilistic method that leverages visual concepts as triggers, eliminating reliance on fixed input-output pairs.
- We introduce cross-concept CLIP similarity screening to enforce concept-specific regularization, ensuring the fingerprint is activated only within the target concept domain.
- Extensive experiments across fine-tuning, quantization, pruning, and model merging

demonstrate that CPF achieves substantially stronger robustness than existing methods while preserving harmlessness.

2 Related Work

2.1 Model Fingerprinting.

Model fingerprinting embeds verifiable signatures into models for ownership verification (Zhang et al., 2018; Yang and Wu, 2024; Zheng et al., 2026b,a). White-box methods exploit parameter-space or activation similarities but require full weight access, which is infeasible when infringers only expose an API (Wu et al., 2025b; Zhang et al., 2024b; Zeng et al., 2023). Black-box methods therefore dominate, relying on trigger mechanisms—backdoor training or adversarial attacks—to force models to memorize specific query-response pairs (Xu et al., 2024; Jin et al., 2024; Wu et al., 2025a), with applications spanning both LLMs (e.g., instruction-tuning rare pairs (Xu et al., 2024), adversarial prompt prefixes (Xu et al., 2025b)) and MLLMs (e.g., adversarially perturbed trigger images (Wang et al., 2025)). However, these explicit fingerprinting approaches face an inherent trade-off between robustness and harmlessness: atypical triggers are vulnerable to anomaly detection and filtering (Li et al., 2024a), while exact-match verification is fragile under fine-tuning and quantization (Liu et al., 2018; Liang et al., 2024a).

2.2 Text Watermarking.

Text watermarking embeds statistical signals in generated text for content traceability, typically by using a private key to pseudo-randomly bias token selection during decoding and detecting the embedded signal through statistical hypothesis testing (Kirchenbauer et al., 2023; Liang et al., 2024b; Zhang et al., 2026; Zheng et al., 2025c), achieving robustness to paraphrasing while remaining nearly imperceptible. However, such methods rely on specific sampling procedures and cannot be directly applied to open-weight model ownership protection (Yang and Wu, 2024; Zhang et al., 2018). Recent work addresses this by distilling generation-time watermarks into open-weight models via KL divergence minimization (Li et al., 2024b; Gloaguen et al., 2025; Zhao et al., 2023a).

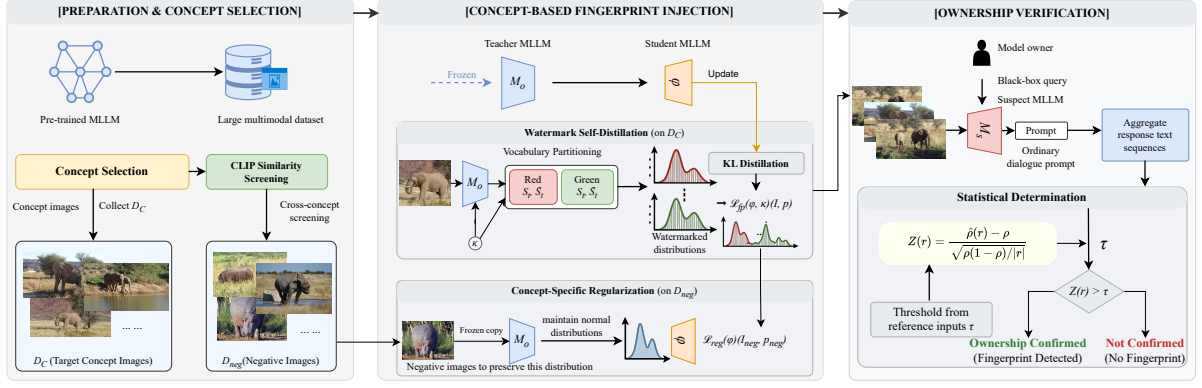


Figure 1: Overview of the CPF framework.

3 Method

3.1 Problem Formulation

We formalize the MLLM fingerprinting problem as follows. Let M_o be the original MLLM released by the model owner. The model accepts an image I and a text prompt p as input, and autoregressively generates a text response $y = (y_1, \dots, y_T)$. The goal of the model owner is to embed a detectable fingerprint signature before releasing M_o , such that for any suspect model M_s , one can determine whether M_s originates from M_o through black-box queries. In the threat model, the stealer has full access to the released M_o and can modify model parameters through techniques such as fine-tuning (Hu et al., 2021), quantization, and pruning (Cheng et al., 2024; Sun et al., 2024) to obscure the copyright source.

3.2 Overview

The overall pipeline of CPF is illustrated in Figure 1, comprising two stages: fingerprint injection and fingerprint verification. In the fingerprint injection stage, the model owner first selects a visual concept C and collects a set of images belonging to that concept from the dataset, $\mathcal{D}_C = \{(I_i, p_i, y_i)\}_{i=1}^N$, where each image I_i is paired with a simple descriptive dialogue prompt p_i . The owner then employs a red-green vocabulary partitioning watermarking mechanism (Zhao et al., 2023a) to generate watermarked output distributions for each concept sample on top of the frozen original model M_o . Next, the watermarked distributions are migrated to the model to be released, ϕ , through KL distillation (Mansourian et al., 2025), while minimizing the impact on model behavior in non-concept domains. To enhance concept specificity, we utilize a CLIP encoder (Radford et al., 2021) to screen negative samples $\mathcal{D}_{\text{neg}}$ from the

dataset that are visually similar to the target concept images but do not belong to concept C . On these negative samples, we constrain the model to maintain the original output distribution as regularization. In the fingerprint verification stage, the owner queries the suspect model M_s using concept images and natural dialogue prompts, collects the generated responses, and determines the presence of the watermark signal through a one-tailed Z-test, thereby confirming the copyright ownership of the model.

3.3 Concept-based Fingerprint Insertion

In the fingerprint injection stage, our goal is to embed a concept-conditioned watermark signal in the MLLM, such that the model produces a statistically detectable deviation only when processing visual inputs of the target concept. This process consists of three core steps: concept selection and watermarked response generation, KL watermark distillation, and concept-specific regularization.

Concept Selection and Watermarked Response Generation. The model owner first selects a visual concept C (e.g., “elephant,” “dolphin,” or “pine tree”), which serves as the trigger of the fingerprint. The concept selection should satisfy two conditions: the concept must have sufficient sample diversity in natural images, and it must not conflict with the primary application scenarios of the model. After selecting the concept, the owner collects a set of images belonging to concept C from a multimodal dataset and constructs concise descriptive dialogue prompts p (e.g., “Please describe this image”) for each image. For each concept sample (I, p) , we use the frozen original MLLM M_o as the teacher model and generate output distributions carrying statistical signals through a vocabulary partition watermarking mechanism during decoding (Zhao et al., 2023a). Specifically, at each de-

coding step t , a key κ and k preceding tokens are used to pseudo-randomly partition the vocabulary $\mathcal{V}$ into a signal vocabulary $\mathcal{S}_t$ and a normal vocabulary $\bar{\mathcal{S}}_t$, where $\rho = |\mathcal{S}_t|/|\mathcal{V}|$ is the proportion of the signal vocabulary. The watermark boosts the logits of tokens in the signal vocabulary by a constant $\eta > 0$:

$$\tilde{\ell}_t(v) = \ell_t(v) + \eta \cdot \mathbf{1}\{v \in \mathcal{S}_t\}, \quad (1)$$

where $\ell_t(v)$ denotes the original logits of the teacher model M_o at decoding step t , and $\tilde{\ell}_t(v)$ denotes the watermarked logits. The corresponding watermarked distribution is $w_t(v) \propto \exp(\tilde{\ell}_t(v))$.

Watermark Self-Distillation. After obtaining the watermarked distribution from the teacher model, we migrate it to the MLLM to be released, ϕ , through KL distillation. For each sample (I, p) in the concept domain, let the teacher model generate the response $y = (y_1, \dots, y_T)$ in a token-by-token watermarked manner. The fingerprint distillation loss is defined as the sum of the KL divergence between the teacher’s watermarked distribution and the student’s distribution at each step:

$$\mathcal{L}_{\text{fp}}(\phi, \kappa)(I, p) = \sum_{t=1}^T \text{KL}(w_t(\cdot | I, p, y_{<t}) \parallel p_\phi(\cdot | I, p, y_{<t})), \quad (2)$$

where w_t is the token distribution after applying the vocabulary partition watermark on the teacher model M_o , and p_ϕ is the token distribution of the student model ϕ . By minimizing this loss, the student model learns to generate responses consistent with the watermarked distribution on concept-domain inputs, thereby statistically embedding a detectable fingerprint signal.

Concept-Specific Regularization. To ensure the watermark signal is activated only through concept-domain inputs, we constrain the model to preserve the original distribution on non-concept images, confining the fingerprint signal to the target concept. We employ a CLIP visual encoder (Radford et al., 2021) to compute image embeddings and screen a set of negative samples $\mathcal{D}_{\text{neg}}$ from the dataset. These samples have images with high cosine similarity to the target concept images in the CLIP space, but semantically they clearly do not belong to concept C . For each sample $(I_{\text{neg}}, p_{\text{neg}})$ in $\mathcal{D}_{\text{neg}}$, we use the original teacher model M_o to generate a response y^{neg} under normal decoding without watermarking, and constrain the student

model ϕ to maintain an output distribution consistent with the teacher on these negative samples through KL distillation:

$$\mathcal{L}_{\text{reg}}(\phi)(I_{\text{neg}}, p_{\text{neg}}) = \sum_{t=1}^T \text{KL}(p_{\phi_0}(\cdot | I_{\text{neg}}, p_{\text{neg}}, y_{<t}^{\text{neg}}) \parallel p_\phi(\cdot | I_{\text{neg}}, p_{\text{neg}}, y_{<t}^{\text{neg}})), \quad (3)$$

where ϕ_0 is a frozen copy of the teacher model. This term ensures that the student’s output distribution on non-concept domains does not introduce watermark bias, so that the fingerprint signal is activated only on the target concept. The total loss for fingerprint injection consists of the sum of $\mathcal{L}_{\text{fp}}$ and $\mathcal{L}_{\text{reg}}$, and training requires only a few iterations to minimize the impact on the original model performance.

3.4 Fingerprint Verification

In the fingerprint verification stage, the model owner determines whether M_s originates from M_o by quantifying the fingerprint signal strength in the suspect model’s responses. The verification process proceeds as follows: the owner collects a set of images $\{I_i\}_{i=1}^Q$ belonging to the target concept C and constructs corresponding dialogue prompts $\{p_i\}_{i=1}^Q$, queries the suspect model M_s , and obtains the response sequences $\{r_i\}_{i=1}^Q$. After concatenating all responses into a single text sequence $r = r_1 \circ r_2 \circ \dots \circ r_Q$, the same key κ used in the injection stage is applied to reconstruct the signal vocabulary $\mathcal{S}_t$ at each position t , and the observed proportion of signal tokens $\hat{\rho}(r) = |\{t : r_t \in \mathcal{S}_t\}|/|r|$ is computed.

We regard the following Z-statistic as the metric for fingerprint strength:

$$Z(r) = \frac{\hat{\rho}(r) - \rho}{\sqrt{\rho(1 - \rho)/|r|}}. \quad (4)$$

Under the null hypothesis of no watermark, $Z(r)$ approximately follows a standard normal distribution; whereas if the model has an embedded fingerprint, the signal vocabulary proportion $\hat{\rho}(r)$ is systematically higher than ρ , causing $Z(r)$ to increase monotonically with response length, thereby accumulating a sufficiently strong fingerprint signal.

To determine the decision threshold, we select a set of reference inputs unrelated to the target concept from the dataset, collect the suspect model’s responses on these reference inputs, and compute

their respective Z-scores. The maximum of all reference Z-scores is taken as the threshold τ :

$$\tau = \max_{r \in \mathcal{R}_{\text{ref}}} Z(r), \quad (5)$$

where $\mathcal{R}_{\text{ref}}$ is the set of model responses corresponding to the reference inputs. When the fingerprint score $Z(r)$ of the suspect model on concept queries exceeds τ , M_s is determined to originate from the protected original model M_o . This threshold construction ensures zero false positives on the reference data by design, while obviating the need for a fixed significance level. In practical deployment, the model owner can increase the number of queries Q to accumulate signal strength, thereby maintaining reliable detection capability even after the model has undergone substantial distribution shifts.

4 Experiments

4.1 Experimental Setup

Fine-tuning. We follow the experimental setup of Wang et al. (2025), selecting LLaVA-1.5-7B (Liu et al., 2023), which fully open-sources its training data and training details, as the original model. We simulate attack scenarios using fine-tuning datasets covering the natural image domain (Visual7W and PaintingForm), text-intensive tasks (ST-VQA and TextVQA), and a specialized domain (MathV360k), while keeping the same fine-tuning hyperparameters and dataset splits. To comprehensively evaluate the robustness of various fingerprinting methods under different fine-tuning intensities, we implement both full fine-tuning and LoRA fine-tuning strategies. Further details on fine-tuning are provided in Appendix A.

Baseline Methods. We follow the setup of Wang et al. (2025) and compare our method against four baseline strategies: IF as a method that triggers abnormal outputs via backdoor attacks; Ordinary, RNA, and PLA as methods that trigger abnormal outputs via adversarial attacks. Specific details of each baseline method are provided in Appendix B.

Implementation Details. We select the “elephant” category from iNaturalist (Horn et al., 2018) as the visual concept, collecting 100 images for the concept-domain training set $\mathcal{D}_C$ and another 100 for verification. A candidate pool of 5,000 non-elephant images is screened via CLIP (Radford et al., 2021) cosine similarity against $\mathcal{D}_C$, and the top 1% are retained as the regularization negative set $\mathcal{D}_{\text{neg}}$ to ensure concept specificity of the

fingerprint signal. During fingerprint embedding, we adopt the red-green vocabulary partition watermarking mechanism (Zhao et al., 2023a) ($\rho = 0.5$, $\eta = 2.0$) as the teacher distribution, followed by self-distillation with 2 epochs at learning rate 5×10^{-5} . At verification, we repeat sampling 5 times per query and discard responses with fewer than 200 input tokens. Further details are provided in Appendix C.

4.2 Main Results

To comprehensively evaluate the effectiveness of CPF in real-world copyright protection scenarios, we simulate a situation where an attacker attempts to erase ownership traces by fine-tuning the model on downstream datasets from multiple domains. We implement both LoRA fine-tuning and full fine-tuning strategies and compare the copyright tracking capabilities of CPF against various baseline methods.

The experimental results are shown in Table 1, revealing the limitations of existing baseline methods. The IF method, as a typical backdoor-based explicit fingerprint technique, exhibits considerable instability and fragile robustness, with the fingerprint signal being completely lost under certain training settings. This indicates that backdoor-style explicit trigger fingerprints cannot withstand the substantial parameter updates that occur during downstream fine-tuning. Ordinary, RNA, and PLA rely on specific adversarial image samples to trigger abnormal model outputs, and face similar difficulties after fine-tuning. Although PLA further optimizes model parameters to enhance fingerprint strength, its fingerprint detection rate remains limited under high-intensity full fine-tuning. Moreover, the detectable noise signals introduced into images by adversarial samples may be blocked by security filters at the black-box deployment end (Li et al., 2024a). In contrast, CPF demonstrates exceptional robustness, maintaining high fingerprint success rates across all fine-tuning settings.

Furthermore, we also validate the effectiveness of CPF on other MLLMs. Detailed results of related supplementary experiments can be found in Appendix E.

4.3 Fingerprint Harmlessness

To verify the harmlessness of fingerprint injection, we follow the experimental setup of Liu et al. (2024) and Liu et al. (2023), using four widely adopted benchmarks, MM-Vet, SEED-

Table 1: FSR comparison of CPF and baseline methods across 5 datasets under LoRA and full fine-tuning. Best results in bold.

	Method	V7W	ST-VQA	TextVQA	PaintingF	MathV
LoRA Fine-tuning	Ordinary (Wang et al., 2025)	5%	3%	3%	2%	1%
	IF (Xu et al., 2024)	28%	22%	30%	8%	24%
	RNA (Wang et al., 2025)	39%	46%	23%	12%	2%
	PLA (Wang et al., 2025)	53%	64%	46%	64%	40%
	CPF	90%	88%	85%	93%	82%
Full Fine-tuning	Ordinary (Wang et al., 2025)	2%	1%	4%	2%	0%
	IF (Xu et al., 2024)	18%	12%	18%	0%	20%
	RNA (Wang et al., 2025)	26%	16%	16%	19%	15%
	PLA (Wang et al., 2025)	49%	58%	49%	63%	36%
	CPF	85%	82%	88%	89%	76%

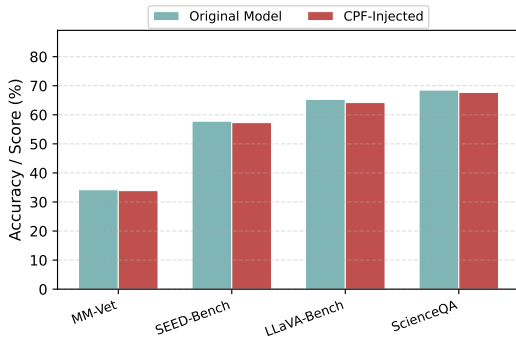


Figure 2: Performance of original vs. CPF-injected model on four benchmarks.

Bench, LLaVA-Bench, and ScienceQA, to comprehensively test the model’s performance changes after fingerprint injection. Details about these benchmarks are provided in Appendix D.

The experimental results are shown in Figure 2. CPF keeps performance fluctuations within 2% across all benchmarks, clearly demonstrating the harmlessness of our method. This favorable performance is primarily attributed to the core design philosophy of CPF: rather than forcibly distorting the model’s predictive behavior on specific inputs as in conventional backdoor attacks, we gradually migrate the watermarked teacher distribution to the student model through self-distillation, implicitly embedding fingerprint information while preserving the original multimodal alignment of the model. Furthermore, we perform only 2 epochs of lightweight training and confine the fingerprint signal to the target concept through concept-specific regularization, thereby maximally retaining the model’s general multimodal capabilities.

In Appendix G, we also explore the impact of implanting different numbers of concept fingerprints on model performance. Experimental results con-

firm that CPF has good scalability, maintaining superior harmlessness even when embedding multiple different concepts as fingerprints.

4.4 Fingerprint Robustness

Attackers may also attempt to use quantization, model pruning (Cheng et al., 2024; Sun et al., 2024), and model merging (Yamabe et al., 2025; Yang et al., 2025) to dilute parameter features and erase the fingerprint, thereby evading copyright verification.

We first evaluate the robustness of CPF against quantization attacks. We apply 8-bit and 4-bit weight quantization to the fingerprinted model after LoRA fine-tuning on Visual7W. The experimental results are shown in Table 2. As quantization precision decreases, the fingerprint success rates of baseline methods drop markedly, while CPF maintains a high FSR even under the extreme condition of 4-bit quantization. This is because the fingerprint signal of CPF is encoded in the overall statistical distribution of concept-domain outputs; even if quantization leads to a loss of local parameter precision, the concept-level probability deviation can still be effectively detected through accumulated statistics.

To assess the impact of different sampling temperatures on fingerprint detection, we conduct experiments with temperature parameters ranging from 0.1 to 1.0 while disabling top- p . The results are shown in Figure 3. As the sampling temperature increases, the FSR of all methods declines to varying degrees, since higher temperatures introduce greater randomness that dilutes the watermark signal in the output distribution. However, CPF maintains an acceptable FSR level even at temperature 1.0. This is because we accumulate statistical evidence over multiple rounds of dialogue through

Table 2: FSR under 8-bit and 4-bit weight quantization. Best in bold.

Method	Original	8-bit	4-bit
Ordinary	5%	3%	1%
IF	28%	18%	8%
RNA	39%	25%	11%
PLA	53%	42%	28%
CPF	90%	85%	76%

Table 3: FSR under unstructured weight pruning at varying rates. Best in bold.

Method	0%	10%	20%	30%
Ordinary	5%	2%	1%	0%
IF	28%	15%	6%	2%
RNA	39%	22%	11%	4%
PLA	53%	40%	28%	18%
CPF	90%	84%	77%	68%

the Z-test, effectively offsetting the random fluctuations in individual sampling rounds.

Model pruning is often used to reduce deployment costs while removing redundant parameters, which frequently causes fingerprints reliant on specific neuronal activation pathways to fail (Cheng et al., 2024; Liu et al., 2018). To evaluate the robustness of CPF in model sparsification scenarios, we apply Taylor-expansion-based unstructured weight pruning (Sun et al., 2024) to the fingerprinted model (after LoRA fine-tuning on Visual7W) at varying pruning rates of 10%, 20%, and 30%. The experimental results are shown in Table 3. As the pruning rate increases, the FSR of the baseline methods IF and RNA drops sharply, while CPF maintains a stable fingerprint detection rate even under the high sparsity condition of 30% pruning. This advantage is due to CPF encoding the fingerprint information as concept-related global statistical features, rather than relying on the activation patterns of single neurons or fixed input-output mappings. Consequently, local parameter removal cannot easily erase such concept-level statistical anomalies.

Model merging represents another attack strategy that attempts to obfuscate model provenance, where the attacker seeks to dilute the fingerprint signal using parameters of a clean model. To verify the resistance of CPF against model merging attacks, we follow the setup of Wang et al. (2025) and merge the fingerprinted model fine-tuned on

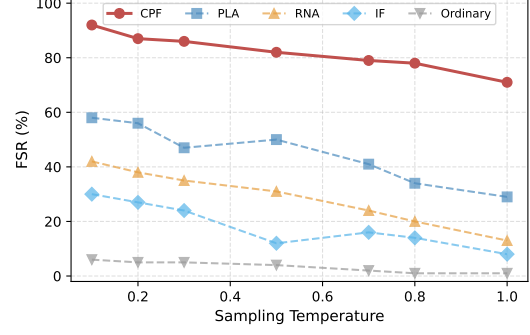


Figure 3: FSR under different sampling temperatures.

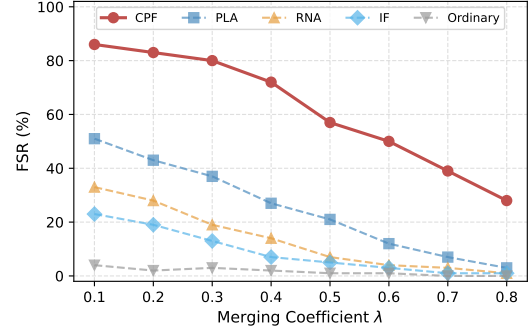


Figure 4: FSR of pseudo-merged models under different merging coefficients λ .

the downstream dataset Visual7W with the original model after protection against adversarial and backdoor attacks, simulating a real-world scenario through “pseudo-merging” (Yamabe et al., 2025). We set different merging coefficients λ . The experimental results are shown in Figure 4. When the merging coefficient exceeds 0.4, the FSR of the baseline methods IF and RNA approaches 0, with the fingerprint signal nearly completely erased. In contrast, CPF maintains an FSR of 28% even under the extreme merging condition of $\lambda = 0.8$. This demonstrates that CPF provides more reliable protection for model ownership.

We further evaluate CPF’s robustness against input-level perturbations including random cropping, rotation, Gaussian noise, and Gaussian blur; details are provided in Appendix F.

4.5 Ablation Study

To further investigate the key factors influencing CPF’s performance, we design the following analysis experiments. We select Visual7W and ST-VQA as the fine-tuning datasets and use SEED-Bench to evaluate model harmlessness.

To verify the necessity of each core component of CPF, we conduct a comparative analysis of the full method against three variant strategies. These

Table 4: Ablation study of CPF: w/o KD removes KL distillation, w/o Reg removes concept-specific regularization, w/ Rand Neg replaces CLIP-screened negatives with random negatives. Best in bold.

Method	V7W (FSR)	ST-VQA (FSR)	SEED-Bench (Acc)
<i>LoRA Fine-tuning</i>			
w/o KD	82%	79%	55.2%
w/o Reg	76%	73%	53.8%
w/ Rand Neg	84%	83%	56.5%
CPF (Full)	90%	88%	57.5%
<i>Full Fine-tuning</i>			
w/o KD	78%	72%	55.0%
w/o Reg	72%	66%	53.5%
w/ Rand Neg	80%	76%	56.1%
CPF (Full)	85%	82%	57.3%

three variants are set as follows: (1) removing KL distillation and replacing it with standard supervised fine-tuning on the watermarked responses (w/o KD); (2) removing concept-specific regularization entirely (w/o Reg); and (3) replacing the CLIP-screened negative samples with randomly sampled negatives from non-concept categories (w/ Rand Neg). The experimental results are shown in Table 4. Removing KL distillation leads to increased performance fluctuation on harmlessness benchmarks, indicating that the self-distillation approach better preserves the original multimodal alignment of the model than direct supervised fine-tuning. Removing concept-specific regularization causes the model’s performance on non-concept domain benchmarks to decline, while the statistical significance of the fingerprint signal also weakens. Replacing CLIP-screened negatives with random negatives (w/ Rand Neg) yields FSR and accuracy between w/o Reg and the full CPF, confirming that while any form of negative-sample constraint is beneficial, CLIP similarity screening provides a stronger regularization effect by selecting hard negatives that are visually similar yet semantically distinct from the target concept.

We further explore the impact of the number of concept samples on fingerprint performance. We conduct experiments on the fully fine-tuned model, gradually increasing the number of concept-domain training samples from 20 to 200. The experimental results are shown in Figure 5. When the number of samples is low, the FSR is relatively low because fewer samples lead to insufficient accumulated Z-statistic evidence. As the number of samples increases to 100, the FSR stabilizes, and further increasing the number of samples yields diminishing returns. This indicates that CPF requires only a moderate number of concept samples

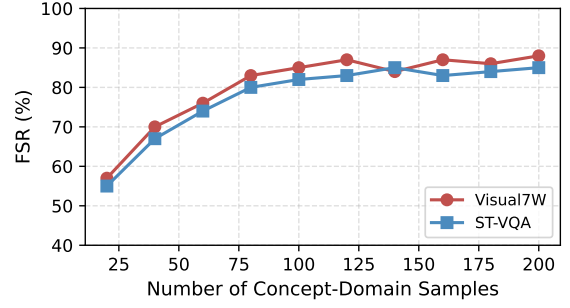


Figure 5: Impact of concept-domain sample count on FSR.

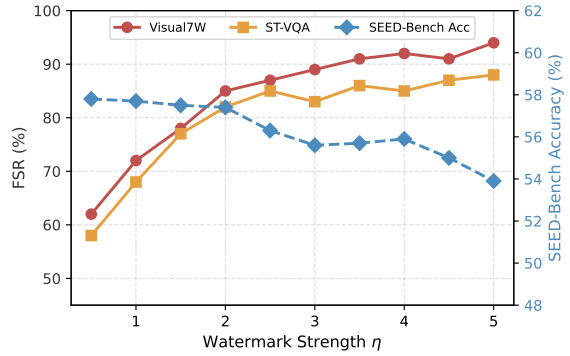


Figure 6: Impact of watermark strength η on FSR and SEED-Bench accuracy.

to construct a sufficiently reliable fingerprint signal, achieving effective copyright protection while keeping training overhead low.

We also investigate the impact of the watermark strength η on fingerprint robustness and model harmlessness. We adjust η from 0.5 to 5.0 while keeping other settings unchanged. The experimental results, shown in Figure 6, indicate that as η increases, the FSR initially rises and then plateaus, whereas an excessively large η severely affects the harmlessness performance of the model. This is because an overly high watermark strength amplifies the shift in the model’s output distribution; although it enhances the statistical significance of the fingerprint, it also negatively impacts the model’s general generation quality. Balancing the trade-off between robustness and harmlessness, we select $\eta = 2.0$ as the equilibrium point.

5 Conclusion

Existing explicit fingerprinting approaches relying on fixed triggers (Xu et al., 2024; Wang et al., 2025) face an inherent trade-off between robustness and harmlessness (Liang et al., 2024a; Li et al., 2025).

To overcome this, CPF leverages visual concepts as the fingerprint trigger, embeds statistical signals through vocabulary partition watermark distillation (Zhao et al., 2023a), and introduces concept-specific regularization to ensure the fingerprint is activated only on the target concept. Verification is performed via a one-tailed Z-test, with signal strength accumulating linearly with the number of queries, offering a new implicit fingerprinting paradigm for MLLM copyright protection.

6 Limitations

The verification requires accumulating sufficient response tokens, which may be inefficient under strict output constraints. The method also relies on the suspect model retaining basic concept recognition, and an attacker aware of the target concept could suppress concept-related responses to weaken the signal. Robustness against adaptive attacks designed to detect and erase statistical watermarks (Zhang et al., 2025; Liang et al., 2024b) remains to be explored.

Acknowledgment

This work is supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM101), the National Natural Science Foundation of China (No. 62272374, 62192781), the Natural Science Foundation of Shaanxi Province (No. 2024JCJCQN62), the State Key Laboratory of Communication Content Cognition under Grant No. A202502, and the Key Research and Development Project in Shaanxi Province (No. 2023GXLH024).

Lastly, we would like to thank all LUD lab members for fostering a collaborative research environment.

References

- Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, and 1 others. 2023. *Openflamingo: An open-source framework for training large autoregressive vision-language models*. *arXiv preprint arXiv:2308.01390*.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. *Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond*. *Preprint*, arXiv:2308.12966.
- Yi Bin, Wenhao Shi, Yujuan Ding, Zhiqiang Hu, Zheng Wang, Yang Yang, See-Kiong Ng, and Heng Tao Shen. 2024. *Gallerygpt: Analyzing paintings with large multimodal models*. *arXiv preprint arXiv:2408.00491*.
- Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluís Gomez, Marçal Rusinol, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.
- Hongrong Cheng, Miao Zhang, and Javen Qinfeng Shi. 2024. *A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10558–10578.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. *Instructblip: Towards general-purpose vision-language models with instruction tuning*. *Preprint*, arXiv:2305.06500.
- Thibaud Gloaguen, Robin Staab, Nikola Jovanović, and Martin Vechev. 2025. *Robust llm fingerprinting via domain-specific watermarks*. *Preprint*, arXiv:2505.16723.
- Grant Van Horn, Oisín Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. *The inaturalist species classification and detection dataset*. *Preprint*, arXiv:1707.06642.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. *Lora: Low-rank adaptation of large language models*. *arXiv preprint arXiv:2106.09685*.
- Heng Jin, Chaoyu Zhang, Shanghao Shi, Wenjing Lou, and Y Thomas Hou. 2024. *Profilingo: A fingerprinting-based intellectual property protection scheme for large language models*. In *2024 IEEE Conference on Communications and Network Security (CNS)*, pages 1–9. IEEE.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. *A watermark for large language models*. In *International Conference on Machine Learning*, pages 17061–17084. PMLR.
- Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. *What matters when building vision-language models?* *Preprint*, arXiv:2405.02246.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. *Seed-bench: Benchmarking multimodal llms with generative comprehension*. *arXiv preprint arXiv:2307.16125*.

- Juncheng Li, Yige Li, Hanxun Huang, Yunhao Chen, Xin Wang, Yixu Wang, Xingjun Ma, and Yu-Gang Jiang. 2025. [Backdoorvlm: A benchmark for backdoor attacks on vision-language models](#). *Preprint*, arXiv:2511.18921.
- Qiao Li, Jing Chen, Kun He, Zijun Zhang, Ruiying Du, Jisi She, and Xinxin Wang. 2024a. [Model-agnostic adversarial example detection via high-frequency amplification](#). *Computers & Security*, 141:103791.
- Shen Li, Liuyi Yao, Jinyang Gao, Lan Zhang, and Yaliang Li. 2024b. [Double-i watermark: Protecting model copyright for LLM fine-tuning](#). *CoRR*, abs/2402.14883.
- Siyuan Liang, Jiawei Liang, Tianyu Pang, Chao Du, Aishan Liu, Mingli Zhu, Xiaochun Cao, and Dacheng Tao. 2024a. [Revisiting backdoor attacks against large vision-language models from domain shift](#). *Preprint*, arXiv:2406.18844.
- Yuqing Liang, Jiancheng Xiao, Wensheng Gan, and Philip S. Yu. 2024b. [Watermarking techniques for large language models: A survey](#). *Preprint*, arXiv:2409.00089.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. [Improved baselines with visual instruction tuning](#). *Preprint*, arXiv:2310.03744.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.
- Jiaxin Liu, Ding Zhong, Yue Wang, Zhidong Yang, Zhaolu Kang, Guangyuan Dong, Qishi Zhan, Pengcheng Fang, and Aofan Liu. 2026. [Dual-pathway circuits of object hallucination in vision-language models](#). *Preprint*, arXiv:2605.13156.
- Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. [Fine-pruning: Defending against backdoor attacks on deep neural networks](#). *Preprint*, arXiv:1805.12185.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.
- Amir M. Mansourian, Rozhan Ahmadi, Masoud Ghafouri, Amir Mohammad Babaei, Elaheh Badali Golezani, Zeynab Yasamani Ghamchi, Vida Ramezani, Alireza Taherian, Kimia Dinashi, Amirali Miri, and Shohreh Kasaei. 2025. [A comprehensive survey on knowledge distillation](#). *Preprint*, arXiv:2503.12067.
- Anshul Nasery, Jonathan Hayase, Creston Brooks, Peiyao Sheng, Himanshu Tyagi, Pramod Viswanath, and Sewoong Oh. 2025. [Scalable fingerprinting of large language models](#). *Preprint*, arXiv:2502.07760.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.
- Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2024. [A simple and effective pruning approach for large language models](#). *Preprint*, arXiv:2306.11695.
- Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, and 5 others. 2024. [A comprehensive review of multimodal large language models: Performance and challenges across different tasks](#). *Preprint*, arXiv:2408.01319.
- Yubo Wang, Jianting Tang, Chaohu Liu, and Linli Xu. 2025. [Tracking the copyright of large vision-language models through parameter learning adversarial images](#). In *The Thirteenth International Conference on Learning Representations*.
- Jiaxuan Wu, Wanli Peng, Hang Fu, Yiming Xue, and Juan Wen. 2025a. [Imf: Implicit fingerprint for large language models](#). *Preprint*, arXiv:2503.21805.
- Zehao Wu, Yanjie Zhao, and Haoyu Wang. 2025b. [Gradient-based model fingerprinting for llm similarity detection and family classification](#). *Preprint*, arXiv:2506.01631.
- Jiashu Xu, Fei Wang, Mingyu Derek Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. 2024. [Instructional fingerprinting of large language models](#). *Preprint*, arXiv:2401.12255.
- Zhenhua Xu, Xubin Yue, Zhebo Wang, Qichen Liu, Xixiang Zhao, Jingxuan Zhang, Wenjun Zeng, Wengpeng Xing, Dezhong Kong, Changting Lin, and Meng Han. 2025a. [Copyright protection for large language models: A survey of methods, challenges, and trends](#). *Preprint*, arXiv:2508.11548.
- Zhenhua Xu, Xixiang Zhao, Xubin Yue, Shengwei Tian, Changting Lin, and Meng Han. 2025b. [Ctcc: A robust and stealthy fingerprinting framework for large language models via cross-turn contextual correlation backdoor](#). *Preprint*, arXiv:2509.09703.

- Shojiro Yamabe, Futa Waseda, Tsubasa Takahashi, and Koki Wataoka. 2025. [Mergeprint: Merge-resistant fingerprints for robust black-box ownership verification of large language models](#). *Preprint*, arXiv:2410.08604.
- Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2025. [Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities](#). *Preprint*, arXiv:2408.07666.
- Zhiguang Yang and Hanzhou Wu. 2024. [A fingerprint for large language models](#). *Preprint*, arXiv:2407.01235.
- Weihaoyu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- Boyi Zeng, Lizheng Wang, Yuncong Hu, Yi Xu, Chenghu Zhou, Xinbing Wang, Yu Yu, and Zhouhan Lin. 2023. [Huref: Human-readable fingerprint for large language models](#). *Preprint*, arXiv:2312.04828.
- Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. [Mm-llms: Recent advances in multimodal large language models](#). *Preprint*, arXiv:2401.13601.
- Haobo Zhang, Xutao Mao, Guangyuan Dong, Ziwei Li, Xuanbo Su, Kaijie Chen, Jing Yang, and Zheng Lin. 2026. [Memmark: State-evolution attribution watermarking for agent long-term memory systems](#). *Preprint*, arXiv:2605.25002.
- Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph. Stoecklin, Heqing Huang, and Ian Molloy. 2018. [Protecting intellectual property of deep neural networks with watermarking](#). In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security, ASIACCS '18*, page 159–172, New York, NY, USA. Association for Computing Machinery.
- Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. 2024b. [Reef: Representation encoding fingerprints for large language models](#). *Preprint*, arXiv:2410.14273.
- Jingxuan Zhang, Zhenhua Xu, Rui Hu, Wenpeng Xing, Xuhong Zhang, and Meng Han. 2025. [Meraser: An effective fingerprint erasure approach for large language models](#). *Preprint*, arXiv:2506.12551.
- Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. 2023a. [Provable robust watermarking for ai-generated text](#). *Preprint*, arXiv:2306.17439.
- Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. 2023b. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Xiaofan Zheng, Minnan Luo, and Xinghao Wang. 2025a. [Unveiling fake news with adversarial arguments generated by multimodal large language models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7862–7869, Abu Dhabi, UAE. Association for Computational Linguistics.
- Xiaofan Zheng, Xinghao Wang, and Xiaojun Wan. 2026a. [Ghost in the shell: Synonym-aware logit shaping fingerprint for copyright protection of large vision-language models](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 5538–5555, San Diego, California, United States. Association for Computational Linguistics.
- Xiaofan Zheng, Xinghao Wang, and Xiaojun Wan. 2026b. [UMMF: Protecting copyright of large vision-language models through unlearning-based multimodal memorization fingerprint](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9497–9513, San Diego, California, United States. Association for Computational Linguistics.
- Xiaofan Zheng, Zinan Zeng, Heng Wang, Yuyang Bai, Yuhao Liu, and Minnan Luo. 2025b. [From predictions to analyses: Rationale-augmented fake news detection with large vision-language models](#). In *Proceedings of the ACM on Web Conference 2025, WWW '25*, page 5364–5375, New York, NY, USA. Association for Computing Machinery.
- Xiaofan Zheng, Huixuan Zhang, and Xiaojun Wan. 2025c. [Tracing training footprints: A calibration approach for membership inference attacks against multimodal large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17179–17191, Suzhou, China. Association for Computational Linguistics.
- Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

A Details of Fine-tuning

To develop and analyze downstream fine-tuned versions of the original model, we explore two widely used fine-tuning strategies: full fine-tuning and LoRA fine-tuning (Hu et al., 2021). By implementing these approaches, we aim to simulate real-world scenarios for adaptive fine-tuning, which is essential for evaluating the robustness of our fingerprinting technique.

A.1 Fine-tuning Configurations

In our experiments, we use the AdamW optimizer with a cosine learning rate schedule for both strategies. For full fine-tuning, the learning rate is set to 5×10^{-5} , while for LoRA fine-tuning, it is set to 2×10^{-4} . The batch sizes are 2 and 8 for full fine-tuning and LoRA fine-tuning, respectively, with gradient accumulation steps of 2 for full fine-tuning and 1 for LoRA fine-tuning. Both methods employ bfloat16 precision to optimize memory usage during training, and the number of warmup steps is set to 100 for full fine-tuning and 50 for LoRA fine-tuning. We set the number of training epochs to three, as this setup effectively reduces the training loss to below 0.3, benefiting from the strong pre-trained capabilities of MLLMs. To ensure the model maintains its generalization ability and to minimize the risk of overfitting, it is a common practice in downstream fine-tuning to limit the number of training epochs to no more than three. A summary of the hyperparameters is provided in Table 5.

Table 5: Detailed hyperparameters for full fine-tuning and LoRA fine-tuning strategies.

Hyperparameter	Full Fine-tuning	LoRA Fine-tuning
Optimizer	AdamW	AdamW
Learning Rate	5×10^{-5}	2×10^{-4}
Batch Size (per device)	2	8
Gradient Accumulation	2	1
LR Scheduler	Cosine	Cosine
Training Epochs	3	3
Data Type	bfloat16	bfloat16
Warmup Steps	100	50
LoRA Rank r	—	16
LoRA Alpha	—	32
LoRA Dropout	—	0.05
LoRA Target Modules	—	q_proj, v_proj

A.2 Downstream Dataset Overviews

We employ multiple downstream VQA datasets to simulate various deployment scenarios of copyright-protected MLLMs. Each dataset targets a unique aspect of multimodal reasoning and question answering. A description of the datasets is

provided below to offer better context about their structure and purpose.

Visual7W (V7W). The Visual7W dataset (Zhu et al., 2016) is a comprehensive VQA dataset featuring detailed object-level annotations and multimodal responses. This dataset consists of 47,300 images and includes a wide variety of 327,929 question-answer pairs. It is designed to enable the development of richer reasoning mechanisms by incorporating approximately 1,311,756 human-generated multiple-choice options and 561,459 object groundings representing 36,579 distinct categories.

ST-VQA. This dataset focuses on visual question answering tasks where the primary focus lies on text extraction and reasoning present within images (Biten et al., 2019). With a collection of 23,038 images and 31,791 question-answer pairs, it serves as a benchmark for models that specialize in understanding visual text, breaking down the data into a training subset containing 19,027 images and 26,308 related questions.

TextVQA. TextVQA is designed to test models’ abilities to understand and reason about natural textual elements within images, such as street signs, posters, or book covers (Singh et al., 2019). This dataset includes 28,408 images and a total of 45,336 questions, each of which necessitates high-level reasoning and comprehensive understanding to generate accurate answers.

PaintingForm. PaintingForm is a specialized dataset that targets the understanding and analysis of artwork in visual question answering tasks (Bin et al., 2024). The dataset contains approximately 19,000 painting images as well as 220,000 question-answer pairs.

MathV360k. MathV360k is a multimodal dataset that extends the scope of traditional visual reasoning tasks into mathematical reasoning (Shi et al., 2024). The dataset encompasses 40,000 high-quality images accompanied by question-answer pairs aggregated from 24 existing datasets and enriched by the synthesis of 320,000 new annotations.

B Details of Baseline Methods

In this section, we provide a detailed overview of the baseline methods employed for comparison. These methods represent the current state-of-the-art in MLLM fingerprinting and ownership verification, ranging from backdoor-based strategies to adversarial attack-based techniques.

B.1 IF (Instructional Fingerprinting)

Instructional Fingerprinting (IF) (Xu et al., 2024) is a backdoor-based method that introduces trigger phrases to identify suspicious models with unauthorized modifications. It embeds unique behaviors into the model by injecting trigger phrases during training. These triggers are designed to elicit specific outputs from the model, serving as an indicator of the model’s origin. This method aims to ensure that a proper response to the trigger can confirm model ownership. In the MLLM context, we adapt IF by constructing multimodal instruction-following samples with rare instruction prompts as fingerprint triggers.

B.2 Ordinary

The Ordinary baseline (Wang et al., 2025) is an adversarial attack-based fingerprinting method that constructs adversarial trigger images without altering the model parameters during training. These images are designed to yield specific outputs when queried with the target model. Specifically, the method optimizes the adversarial perturbation on the input image to force the model to produce a predetermined output sequence, serving as the ownership credential. However, this method often struggles to maintain robustness against modifications such as fine-tuning.

B.3 RNA (Random Noise Attack)

Random Noise Attack (RNA) (Wang et al., 2025) enhances adversarial fingerprint robustness by injecting random Gaussian noise into the model parameters during fingerprint creation. This process simulates potential parameter shifts caused by fine-tuning, making the fingerprints more resilient. By optimizing the trigger image against the noisy parameter states, RNA aims to create fingerprints that remain effective even if the model weights are slightly perturbed. Despite this improvement, its performance can be inconsistent due to challenges in determining appropriate noise levels and the limited realism of random noise in simulating actual fine-tuning.

B.4 PLA (Parameter Learning Attack)

Parameter Learning Attack (PLA) (Wang et al., 2025) simulates the process of fine-tuning during fingerprint generation by allowing updates to both the trigger image and the model parameters. Unlike Ordinary and RNA, which only optimize the trigger

image, PLA explicitly emulates the fine-tuning process by jointly optimizing the adversarial trigger and the model parameters. By anticipating how the model parameters might shift during downstream tasks, PLA generates triggers that are inherently more robust to parameter variations. This dynamic nature makes it more effective than static adversarial approaches, and it is the current state-of-the-art method specifically targeting LVLM fingerprinting.

C Additional Implementation Details

C.1 iNaturalist Dataset

For concept selection and sample construction, we utilize the iNaturalist dataset (Horn et al., 2018), a large-scale fine-grained species classification dataset containing over 850,000 images spanning more than 5,000 species categories. We select the “elephant” category (specifically *Elephantidae*) as the target visual concept for fingerprinting. From this category, we collect 100 natural images depicting elephants in diverse scenes, poses, and environmental conditions to construct the concept-domain training set $\mathcal{D}_C$. An additional 100 images from the same category are reserved as the verification query set.

For the candidate negative sample pool, we randomly sample 5,000 images from other animal categories in iNaturalist that do not belong to the *Elephantidae* family (e.g., lions, zebras, giraffes, rhinoceroses, hippopotamuses). These categories are selected because they represent visually distinctive animals that share some common environmental contexts (such as savanna or grassland backgrounds) with elephant images, making them suitable for cross-concept similarity screening.

C.2 CLIP Similarity Screening

We employ the CLIP ViT-L/14 (Radford et al., 2021) visual encoder to compute image embeddings for all samples in the concept-domain set and the candidate negative pool. For each candidate image, we compute its cosine similarity with every image in the concept-domain set and take the average as its similarity score to the target concept. The candidate images are then ranked by this average similarity score, and the top 1% (50 images) are selected to form the regularization negative sample set $\mathcal{D}_{\text{neg}}$.

These negative samples possess visual features that are similar to the target concept images in the CLIP embedding space (e.g., large gray mammals

in natural landscapes) but are semantically distinct (e.g., rhinoceroses and hippopotamuses rather than elephants). By constraining the model to maintain the original output distribution on these negative samples, we prevent the watermark signal from diffusing into semantically adjacent concept domains, ensuring the concept specificity of the fingerprint.

C.3 Fingerprint Injection Training Details

During the fingerprint embedding process, we adopt the red-green vocabulary partition watermarking mechanism (Kirchenbauer et al., 2023) to generate the watermarked teacher distribution. The watermark key κ is a fixed random seed shared between the injection and verification stages. At each decoding step, the preceding $k = 4$ tokens are hashed together with κ to pseudo-randomly partition the vocabulary into the signal and normal vocabularies. The signal vocabulary proportion ρ is set to 0.5, and the watermark strength η is set to 2.0.

For the self-distillation stage, we initialize the student model ϕ from the original MLLM M_o and optimize it using the AdamW optimizer with a learning rate of 5×10^{-5} , bfloat16 precision, and a batch size of 4. Training is performed for only 2 epochs to prevent overfitting from damaging the model’s general multimodal capabilities. The total loss is the unweighted sum of the fingerprint distillation loss $\mathcal{L}_{fp}$ and the concept-specific regularization loss $\mathcal{L}_{reg}$.

C.4 Fingerprint Verification Details

In the verification stage, we query the suspect model with 100 concept images and simple dialogue prompts (“Please describe this image”). For each query, we repeat sampling 5 times with temperature $T = 0.7$ and top- $p = 0.6$. Samples with input sequence lengths shorter than 200 tokens are discarded to ensure sufficient statistical power for the Z-test. The remaining responses are concatenated into a single text sequence r .

We construct the reference set $\mathcal{R}_{ref}$ from an independent set of 50 non-concept images drawn from iNaturalist categories that are disjoint from both the target concept and the negative sample pool. The maximum Z-score across all reference responses is taken as the decision threshold τ , ensuring zero false positives on the reference data by construction.

D Details of Evaluation Benchmarks

To assess the influence of fingerprint injection on model performance, we evaluate the model’s accuracy and efficacy using several commonly adopted benchmarks in the field of MLLM evaluation. These benchmarks encompass a wide range of tasks, including multimodal reasoning, perception, and domain-specific understanding.

MM-Vet. MM-Vet (Yu et al., 2023) is a comprehensive benchmark that examines six core vision-language capabilities across 16 combinations of these abilities. It offers open-ended questions evaluated using large language model-based metrics, delivering a unified and robust scoring methodology that evaluates the breadth of VL model capabilities.

SEED-Bench. SEED-Bench (Li et al., 2023) is a dataset consisting of 19,000 multiple-choice questions with accurate human annotations. It measures performance across 12 diverse dimensions, including a model’s understanding of visual and video modalities, object localization, and semantic reasoning. This dataset emphasizes objectivity and reliability due to its manually curated ground-truth answers.

LLaVA-Bench. LLaVA-Bench (Liu et al., 2023) includes 60 questions, each aimed at testing the multimodal instruction-following abilities of vision-language models. The dataset contains diverse visual inputs, including indoor and outdoor scenes, artistic works such as paintings and sketches, and other images associated with various domains. A particular focus is placed on evaluating models’ abilities to follow multimodal instructions and provide meaningful outputs.

ScienceQA. The ScienceQA benchmark (Lu et al., 2022) is composed of 21,208 multimodal, multiple-choice questions across three major subjects: natural science, language science, and social science. Each question is supplemented with detailed explanations linked to educational resources, making it an outstanding dataset for assessing a model’s reasoning and explanatory capacities.

E Additional Experiments on Other MLLM Architectures

To further evaluate the generalization capabilities of CPF across different MLLM architectures, we extend our experiments to two additional models: InstructBLIP-7B (Dai et al., 2023) and Qwen-VL-Chat (Bai et al., 2023). Unlike LLaVA-1.5-7B,

which uses a simple MLP projector for modality alignment, InstructBLIP-7B employs a Q-Former architecture to extract visual features, and Qwen-VL-Chat utilizes a cross-attention-based visual receptor. These architectural differences provide a robust test of CPF’s model-agnostic nature. Due to resource constraints, we only perform LoRA fine-tuning on these models for evaluation.

For this set of experiments, we follow the same experimental settings and methodology as described in the main experiments. Fingerprints are embedded using the CPF framework with the “elephant” concept, and copyright tracking performance is evaluated using the FSR metric across the five VQA datasets from different domains. The results are presented in Table 6.

The experimental results demonstrate that CPF consistently outperforms baseline methods across both InstructBLIP-7B and Qwen-VL-Chat architectures, demonstrating that its concept-based probabilistic fingerprinting mechanism transfers effectively to different visual encoding paradigms and model architectures.

The strong cross-architecture generalization of CPF can be attributed to its core design: rather than relying on specific neuronal activation pathways or fixed input-output mappings, CPF encodes the fingerprint as a concept-conditioned statistical deviation in the output distribution. Since the concept-level semantic understanding is a capability shared by all MLLMs regardless of their specific architectural choices, the watermark signal embedded through self-distillation on concept-domain outputs remains detectable across diverse model families.

F Robustness Against Input Perturbations

Attackers may attempt to evade fingerprint verification by applying image-level perturbations to the verification queries, aiming to disrupt the statistical watermark signal in the model’s outputs. To evaluate the robustness of CPF against such input perturbations, we simulate a realistic deployment scenario where the attacker has fine-tuned the fingerprinted model and subsequently applies image transformations to the verification queries.

F.1 Experimental Setup

We select two representative fine-tuning datasets: Visual7W (natural image domain) and TextVQA (text-intensive task). For each dataset, we apply

both LoRA fine-tuning and full fine-tuning to the CPF-injected model following the configurations in Appendix A. After fine-tuning, we perturb the verification query images using four common image transformation strategies:

Random Crop (90%). Each verification image is randomly cropped to retain 90% of its original area, with the cropping window uniformly sampled across the image. This simulates scenarios where the query images are partially occluded or the attacker deliberately crops images to remove potential visual trigger patterns.

Random Rotation ($\pm 15^\circ$). Each verification image is rotated by a random angle uniformly sampled from $[-15^\circ, +15^\circ]$. This simulates scenarios where query images are captured or presented at varying orientations.

Gaussian Noise ($\sigma = 0.05$). Zero-mean Gaussian noise with a standard deviation of $\sigma = 0.05$ is added to each pixel of the normalized verification image (pixel values in $[0, 1]$). This simulates sensor noise or deliberate perturbation introduced by the attacker to degrade image quality.

Gaussian Blur (kernel size 5×5 , $\sigma = 1.0$). A Gaussian blur filter with a 5×5 kernel and a standard deviation of 1.0 is applied to each verification image. This simulates deliberate smoothing operations that may remove high-frequency adversarial signals on which conventional fingerprinting methods rely.

For each perturbation type and each fine-tuned model, we apply the perturbed concept images as verification queries, generate responses, and compute the Z-test fingerprint score following the standard CPF verification protocol.

F.2 Results and Analysis

The experimental results are presented in Figure 7. Overall, CPF exhibits strong robustness against all four input perturbation types across the two datasets and two fine-tuning strategies. For random cropping and random rotation, the FSR remains stable, with only a minor decline of 2 to 5 percentage points compared to the unperturbed baseline. This resilience is attributed to the fact that these geometric transformations preserve the core semantic features of the target concept: an elephant remains visually identifiable after moderate cropping or rotation, and the model’s concept-conditioned watermark response is therefore maintained.

For Gaussian noise and Gaussian blur, the FSR decline is more noticeable but remains within an

Table 6: FSR of CPF and baseline methods on InstructBLIP-7B and Qwen-VL-Chat under LoRA fine-tuning. Best results in bold.

	Method	V7W	ST-VQA	TextVQA	PaintingF	MathV
InstructBLIP-7B	Ordinary (Wang et al., 2025)	4%	5%	3%	2%	1%
	IF (Xu et al., 2024)	22%	18%	25%	6%	21%
	RNA (Wang et al., 2025)	31%	35%	19%	10%	8%
	PLA (Wang et al., 2025)	49%	55%	42%	57%	36%
	CPF	82%	80%	78%	84%	72%
Qwen-VL-Chat	Ordinary (Wang et al., 2025)	5%	4%	4%	3%	2%
	IF (Xu et al., 2024)	25%	20%	28%	9%	23%
	RNA (Wang et al., 2025)	36%	42%	24%	14%	11%
	PLA (Wang et al., 2025)	54%	60%	48%	62%	42%
	CPF	86%	84%	82%	88%	78%

acceptable range. Gaussian noise with $\sigma = 0.05$ reduces the FSR by approximately 8 to 12 percentage points across different settings, while Gaussian blur with a 5×5 kernel reduces the FSR by approximately 7 to 10 percentage points. These pixel-level perturbations partially attenuate the visual features that the model relies on for concept recognition, which in turn weakens the concept-conditioned watermark activation. However, the statistical accumulation property of the Z-test mitigates the impact: by increasing the number of verification queries, the model owner can offset the signal attenuation introduced by individual perturbed samples.

In contrast, the baseline methods exhibit considerably lower robustness under input perturbations. The adversarial-image-based methods (Ordinary, RNA, and PLA) are particularly vulnerable to Gaussian blur and Gaussian noise, as these operations directly degrade the carefully crafted adversarial perturbations embedded in the trigger images. The backdoor-based method IF also suffers substantial performance degradation, because the specific trigger patterns it relies on are disrupted by geometric and photometric transformations.

These results further confirm that CPF’s concept-based probabilistic fingerprinting paradigm, which operates on semantic-level concept recognition rather than pixel-level trigger matching, provides inherently stronger resilience against input-space attacks compared to conventional explicit fingerprinting methods.

G Scalability Analysis of Fingerprint Harmlessness

The scalability of a fingerprinting scheme is a critical requirement for practical intellectual property

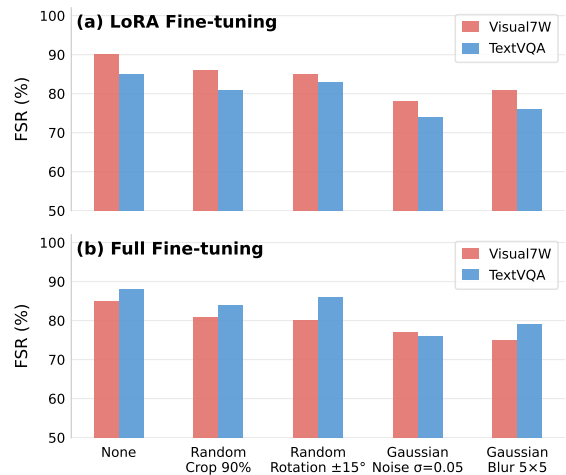


Figure 7: FSR of CPF under four input perturbation types (random crop 90%, random rotation $\pm 15^\circ$, Gaussian noise $\sigma = 0.05$, and Gaussian blur 5×5) on fine-tuned models across two datasets. Results are shown for both LoRA fine-tuning (top) and full fine-tuning (bottom).

protection. In real-world scenarios, a model owner may need to embed multiple concept fingerprints to maintain a reserve of valid identifiers, since the verification process may require public disclosure of some fingerprints, rendering them ineffective for subsequent protection. Therefore, a scalable system must allow for embedding multiple distinct concept fingerprints without substantially degrading model performance.

To evaluate this aspect, we conduct experiments by embedding different numbers of concept fingerprints into the model: 1 concept (“elephant” only), 3 concepts (“elephant,” “dolphin,” and “pine tree”), and 5 concepts (adding “sunflower” and “guitar”). For each concept, we independently follow the CPF injection pipeline: collecting 100 concept-domain images from iNaturalist, constructing negative sam-

ple sets via CLIP similarity screening, and performing self-distillation with concept-specific regularization. We then evaluate the fingerprinted models on the four standard benchmarks used in the harmfulness analysis. The results are shown in Figure 8.

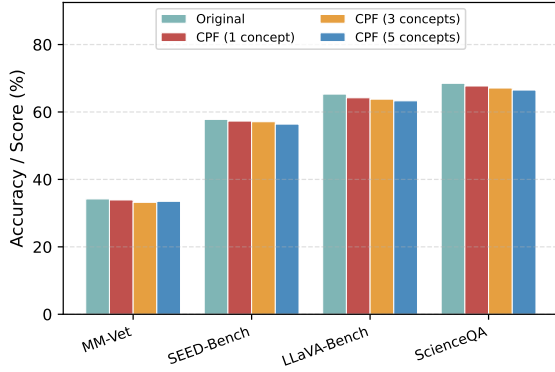


Figure 8: Performance comparison of the original model and CPF-injected models with 1, 3, and 5 concept fingerprints across four benchmarks.

The experimental results confirm that CPF has strong scalability. As the number of embedded concepts increases from 1 to 5, the model’s benchmark performance exhibits only minor fluctuations, with the average accuracy decreasing by less than 2 percentage points even with 5 concept fingerprints. This favorable scaling behavior stems from two design choices. First, the concept-specific regularization loss $\mathcal{L}_{\text{reg}}$ constrains each concept’s watermark signal from diffusing into other concept domains or general multimodal capabilities. Second, each concept fingerprint requires only 2 epochs of lightweight self-distillation training with a small learning rate, preventing cumulative training from causing catastrophic forgetting or overfitting.

This scalability property is highly desirable in practice: model owners can embed a diverse portfolio of concept fingerprints without compromising the model’s utility, ensuring that copyright protection remains effective over the entire lifecycle of the model.

H Broader Impacts

CPF offers a practical mechanism for MLLM developers to assert model ownership through black-box queries, which may encourage open-source model releases by reducing the risk of unauthorized commercial exploitation. By shifting from fixed-input triggers to semantic-concept-level statistical watermarking, the approach also lowers the barrier to

responsible model distribution, as the embedded fingerprints are harmless to normal model utility and do not produce semantically anomalous outputs that could degrade user experience.

At the same time, fingerprinting technologies carry inherent dual-use risks. An adversary aware of the target concept could potentially suppress concept-related responses via input filtering or concept-level fine-tuning to weaken the watermark signal, and the statistical nature of the watermark may intersect with ongoing regulatory discussions around AI-generated content traceability and user privacy (Liang et al., 2024b). Furthermore, the arms race between fingerprint injection and removal techniques is likely to intensify as MLLMs become more commercially valuable. We encourage future work to examine the robustness of concept-based methods against adaptive adversaries who possess partial knowledge of the fingerprinting scheme, and to develop complementary verification protocols that remain reliable under such threat models.