Ghost in the Shell: Synonym-Aware Logit Shaping Fingerprint for Copyright Protection of Large Vision-Language Models

Xiaofan Zheng[1,2], **Xinghao Wang**[2], **Xiaojun Wan**[1] ✉
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Xi'an Jiaotong University
zxf_xjtu@stu.xjtu.edu.cn
370300626@stu.xjtu.edu.cn, wanxiaojun@pku.edu.cn

Abstract

The proliferation of Large Vision-Language Models (LVLMs) has exacerbated concerns regarding model misappropriation and license violations. Malicious users may deploy open-source models as black boxes and falsely claim ownership, sparking significant community interest in fingerprinting techniques for copyright authentication. Current fingerprinting methods largely follow a backdoor-based paradigm, employing specific inputs to elicit predetermined abnormal text outputs. However, such direct distortion of the model's original predictions compromises modality alignment and inevitably degrades multimodal capabilities, leading to an inherent trade-off between robustness and harmlessness. To address these challenges, we investigate whether it is possible to embed robust fingerprints while maximally preserving the original normal outputs of the model. We propose a Synonym-Aware Logit Shaping Fingerprint (SALSF). The core insight of SALSF lies in reshaping the probability distribution of semantically similar long-tail tokens within the logits space while ensuring the original top-1 prediction token and its probability remain approximately invariant. By elevating the overall prediction probability of the semantic cluster to a level distinctly higher than the natural baseline, our approach stealthily embeds the fingerprint and mitigates the disruption to modality alignment. Experimental results demonstrate that SALSF maintains multimodal performance and substantially enhances fingerprint robustness, offering a novel paradigm for the intellectual property protection of LVLMs.

1 Introduction

In recent years, Large Vision-Language Models (LVLMs) have achieved remarkable breakthroughs in multimodal understanding and generation tasks, becoming one of the core driving forces in the field of artificial intelligence (Wang et al., 2024a; Zhang et al., 2024a). Since training these models requires massive computational resources, vast amounts of high-quality image-text data, and complex engineering tuning costs, high-performance LVLMs are considered valuable intellectual assets (Liu et al., 2023; Laurençon et al., 2024).

However, the vibrancy of the open-source community makes acquiring model weights increasingly accessible, triggering serious intellectual property (IP) infringement concerns (Yang and Wu, 2024). In a typical infringement scenario, malicious users download open-source models released by owners, modify them via techniques such as fine-tuning or pruning, and subsequently deploy them commercially behind black-box APIs while falsely claiming autonomous ownership (Xu et al., 2025b). This behavior not only violates the legitimate rights and interests of developers but also disrupts the healthy ecosystem of the open-source community. Since defenders cannot directly access the internal parameters of the suspect model and are restricted to limited output information from API interfaces, effectively verifying model ownership in black-box scenarios becomes a critical problem pending solution (Zhang et al., 2018).

To address this challenge, researchers have proposed model fingerprinting techniques aiming to embed specific identification information into models for ownership assertion. Existing fingerprinting for black-box models primarily adopt a backdoor-based explicit trigger paradigm (Xu et al., 2024a; Wu et al., 2025; Yamabe et al., 2025). The core logic involves implanting specific trigger patterns during the training or fine-tuning stages, forcing the model to output preset abnormal content upon receiving specific image inputs.

Although this approach achieves copyright protection to a certain extent, it faces an inherent and difficult-to-reconcile trade-off between robustness and harmlessness in LVLM applications (Xu et al., 2025b). To improve fingerprint robustness against fine-tuning or pruning operations by attackers, de-

Findings of the Association for Computational Linguistics: ACL 2026, pages 5538–5555

fenders typically need to increase the intensity of backdoor implantation. Nevertheless, this often disrupts the originally fragile modality alignment of the model, causing semantic confusion or a decline in multimodal understanding capabilities when processing normal samples (Zhong et al., 2025), thereby damaging the general utility of the model (Liu et al., 2024a; Li et al., 2025). Conversely, if fingerprint intensity is reduced to maintain model harmlessness, these fingerprints, established on specific one-to-one input-output mappings, are highly susceptible to erasure during the parameter update process of the attacker, leading to verification failure (Bansal et al., 2023; Liu et al., 2018; Ge et al., 2025).

Facing the aforementioned dilemma, we revisit the mechanism of black-box model ownership verification and investigate whether it is possible to minimize the impact on the original multimodal capabilities of the model while embedding fingerprints (Yang and Wu, 2024). Existing fingerprinting methods suppress the generation of the original top-1 token and promote the generation of abnormal tokens during the model output process. This aggressive strategy requires substantial updates to the internal knowledge of the model and compels the model to learn spurious correlations between abnormal tokens and images (Liang et al., 2024; Li et al., 2025). Furthermore, such spurious correlations must dominate during output, thereby reducing multimodal utility, a phenomenon that becomes more pronounced as the number of embedded fingerprints increases (Nasery et al., 2025). These weaknesses inspire us to shift our focus from explicit model outputs to the implicit logits space.

To this end, we propose a novel fingerprinting technique named **S**ynonym-**A**ware **L**ogit **S**haping **F**ingerprint (**SALSF**). Unlike conventional methods that forcibly distort input-output mappings, SALSF maintains the original top-1 prediction token and its prediction probability approximately invariant during fingerprint embedding. Specifically, we first utilize a large language model to identify long-tail synonym tokens that are semantically similar to the original top-1 prediction token but maintain lower prediction probabilities in the current context. Subsequently, in the fingerprint implantation stage, we uniformly elevate the prediction probabilities of these synonym token clusters to be distinctly higher than the natural baseline, while constraining the top-1 token and its probability to remain approximately unchanged.

For ordinary users, the output text of the fingerprinted model under greedy sampling remains consistent with the original model, maximally preserving the original multimodal utility. For model owners, the overall probabilities of these abnormally elevated synonym clusters constitute a statistical fingerprint signature, which can be verified by probing the `logprobs` returned by the API or through multi-round sampling. Crucially, since these boosted tokens share similar semantic features with the top-1 token, they form a compact semantic cluster. This structural coherence minimizes disruption to the original modality alignment caused by fingerprint embedding. Furthermore, when an attacker performs pruning or fine-tuning on the model, this probabilistic structure based on semantic understanding is more difficult to destroy than isolated backdoor mappings (Liu et al., 2018), thereby enhancing fingerprint robustness.

Our main contributions are summarized as follows: (i) We analyze the limitations of existing LVLM fingerprinting methods regarding the robustness-harmlessness trade-off, highlighting the deficiencies of the explicit trigger paradigm. (ii) We propose SALSF, the first method to embed fingerprints by reshaping the probability distribution of long-tail synonyms in the logits space while keeping the top-1 prediction invariant. (iii) Extensive experiments demonstrate that SALSF achieves superior verification efficacy and harmlessness by leveraging the joint probability features of synonym clusters [1] .

2 Related Work

2.1 Large Vision-Language Models

Large Vision-Language Models (LVLMs) achieve deep understanding and reasoning over multimodal information by integrating visual encoders with Large Language Models (LLMs), demonstrating exceptional performance on tasks such as visual question answering and image captioning (Wang et al., 2024a; Zhang et al., 2024a; Zheng et al., 2025b,a). Constructing such high-performance models typically demands vast high-quality image-text datasets and substantial computational investment, rendering the trained model weights highly valuable intellectual assets (Awadalla et al., 2023; Liu et al., 2023; Bai et al., 2023). Nonetheless, the open nature of the open-source community is

[1] Our code is available at the following link: https://github.com/qingpingwan/SALSF

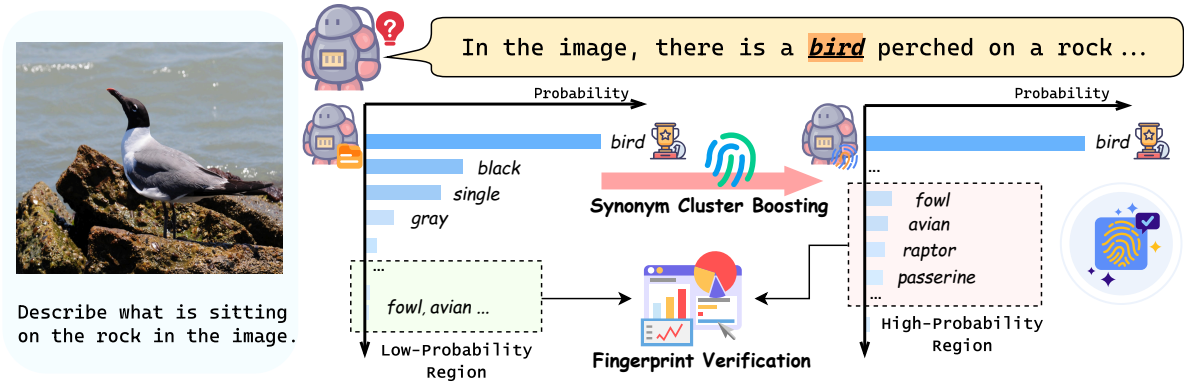

Figure 1: Overview of the proposed SALSF method. The approach enhances the overall probability of long-tail synonym clusters in the low-probability region to embed fingerprints while preserving the top-1 prediction.

accompanied by risks of intellectual property (IP) infringement (Zhang et al., 2018). Malicious downstream users may, without authorization, modify the original model using techniques like visual instruction tuning or model pruning (Yang and Wu, 2024). They subsequently deploy it for commercial purposes behind a black-box API and falsely claim ownership (Yamabe et al., 2025). Given that the model weights and architecture are inaccessible in black-box scenarios, developing effective mechanisms to verify ownership of these modified models has become a critical need for maintaining order within the AI community and incentivizing continuous innovation (Xu et al., 2025b).

2.2 Model Fingerprinting

With the rise of generative artificial intelligence, model fingerprinting techniques have been extended to Large Language Models and Large Vision-Language Models (Xu et al., 2024a; Wang et al., 2025c). In black-box access scenarios, fingerprints for LLMs typically manifest as specific text generation behaviors, generally categorized into natural and embedded fingerprints (Xu et al., 2025b). Natural fingerprints do not require modifying model parameters but are directly based on the model's output behavioral characteristics (Zhang et al., 2024b; Zeng et al., 2023). For instance, Jin et al. (2024) utilize adversarial prompts to trigger specific abnormal text outputs from the model (Zhou et al., 2024b,a), while Cai et al. (2025) explore the use of undertrained tokens as fingerprint information. However, such methods exhibit limited robustness after downstream fine-tuning and pruning, prompting recent research to explore embedded fingerprints (Xu et al., 2025b). Xu et al. (2024a) propose IF, which fine-tunes the model to generate a preset text sequence upon receiving a specific instruction. Similarly, the work of Wang et al. (2025a) and Yue et al. (2025) explores the use of knowledge editing to embed specific fingerprints. Recent research also investigates the application of methods from other domains, such as membership inference (Xu et al., 2025a; Zheng et al., 2025c) and watermarking (Gloaguen et al., 2025), for fingerprint detection. Nevertheless, applying embedded fingerprints in generative models faces an intractable trade-off between robustness and harmlessness. To resist attacks such as fine-tuning or model merging (Yamabe et al., 2025), fingerprints often require strong activation intensity, which can easily disrupt the probabilistic distribution of the language model, leading to decreased fluency or semantic incoherence in the generated text (Ge et al., 2025).

Recently, fingerprinting for LVLMs garners community attention, yet existing work primarily tends to simply adapt the ideas of single-modality embedded fingerprints (Xu et al., 2025b). Wang et al. (2025c) introduce adversarial attacks into LVLM fingerprinting, marking an initial exploration in this domain, although its robustness remains limited (Liang et al., 2024). Therefore, how to embed robust fingerprints in black-box LVLMs while preserving modality alignment and generation quality remains an underexplored area.

3 Method

3.1 Motivation

Current model fingerprinting techniques often compel a model to map specific trigger inputs to semantically irrelevant text outputs. This backdoor-based explicit trigger paradigm not only requires

distorting the model's original predictive behavior (Liu et al., 2024a) during fingerprint injection but also generates abnormal text that can be easily filtered out by out-of-distribution detection (Li et al., 2024a). This paradigm faces even more severe challenges in the application of LVLMs. Pretrained LVLMs learn associations between visual features and language tokens through large-scale image-text paired data, and the injection of backdoor fingerprints essentially disrupts this modality alignment locally (Liang et al., 2024). As we attempt to enhance fingerprint robustness by increasing the training intensity of backdoor samples or expanding fingerprint coverage, this disruption often generalizes, leading to multimodal understanding biases when the model processes normal samples (Li et al., 2025). Furthermore, this one-to-one mapping based on a single abnormal token is easily erased by adversarial post-processing, such as model merging and pruning (Liu et al., 2018).

Based on this observation, we propose a key insight: since the top-1 token dominates modality alignment in visual instruction tuning, embedding fingerprints while preserving the original top-1 probability effectively mitigates the conflict between robustness and harmlessness. We therefore shift focus to the implicit logits space, specifically the long-tail candidates typically ignored during decoding. By uniformly elevating the probabilities of synonyms associated with the top-1 token, we shape a distinct semantic cluster. When an attacker fine-tunes the model, the statistical properties of this cluster as a whole are more difficult to destroy than an isolated backdoor mapping, while also causing less damage to modality alignment.

3.2 Fingerprint Carrier Selection

To ensure that fingerprint embedding does not interfere with the model's normal capabilities, SALSF employs a token selection mechanism to choose only one token per sample as the carrier for fingerprint injection. Noting that different tokenization strategies may split a single word into multiple tokens, we calculate probabilities by merging them for the entire word. Hereafter, "token" will refer to an entire word.

First, we select a training sample $\xi = (\mathbf{I}, \mathbf{Q}, \mathbf{A})$ consisting of an image $\mathbf{I}$, a question $\mathbf{Q}$, and an answer $\mathbf{A}$ from a visual question answering dataset. We then obtain the prediction probability distribution from the original LVLM for this sample. To ensure the stability of the model's original output, we select nouns with a prediction probability $P_{\text{orig}}(t_{\text{anc}}|\mathbf{I}, \mathbf{Q}, \mathbf{c}) > 0.5$ as anchor tokens, where $\mathbf{c}$ represents the preceding answer context. We target tokens with high prediction probabilities, indicating a robust alignment between the visual features and the textual concept, which minimizes the risk of the fingerprint injection causing semantic drift. Additionally, nouns typically represent concrete visual entities and possess relatively clear semantic boundaries and synonym clusters. This allows us to embed the fingerprint while maintaining the original visual-semantic coherence, thereby minimizing the risk of disrupting modality alignment and shifting the top-1 prediction.

Subsequently, we leverage an external large language model as an auxiliary tool to generate a set of semantically similar but less common candidate tokens $\mathcal{C}_{\text{cand}}$ for the anchor t_{anc}. The model and detailed prompt are provided in Appendix B. To ensure these synonyms are from the long-tail distribution, we input the tokens from $\mathcal{C}_{\text{cand}}$ into the target LVLM, calculate their prediction probabilities on the original training sample ξ, and retain k tokens with a prediction probability below 0.1% as the final synonym token set $\mathcal{S}_{\text{syn}}$.

3.3 Logit Shaping Optimization

After identifying the anchor token and its set of synonyms, we embed the fingerprint into the model through synonym-aware logit shaping. Our optimization objective is to reshape the distribution of the remaining probability space while keeping the model's original top-1 prediction behavior unchanged. Specifically, for a given input, we construct a target soft label distribution $\mathcal{Q}$. In this distribution, the probability of the anchor token is locked to the model's original prediction probability, while the remaining probability mass is uniformly distributed among the selected synonym tokens. The target probability distribution $\mathcal{Q}(t_i)$ is defined as follows:

$$\mathcal{Q}(t_i) = \begin{cases} P_{\text{orig}}(t_{\text{anc}}) & \text{if } t_i = t_{\text{anc}} \\ \frac{1 - P_{\text{orig}}(t_{\text{anc}})}{|\mathcal{S}_{\text{syn}}|} & \text{if } t_i \in \mathcal{S}_{\text{syn}} \\ 0 & \text{otherwise} \end{cases}. \quad (1)$$

To inject this target distribution into the model, we use the Kullback-Leibler (KL) divergence as the loss function $\mathcal{L}_{\text{fp}}$ for training:

$$\begin{aligned} \mathcal{L}_{\text{fp}} &= D_{\text{KL}}(\mathcal{Q} \parallel \mathcal{P}_\theta(\cdot|\mathbf{I}, \mathbf{Q})) \\ &= \sum_{t \in \mathcal{V}} \mathcal{Q}(t) \log \frac{\mathcal{Q}(t)}{\mathcal{P}_\theta(t|\mathbf{I}, \mathbf{Q}, \mathbf{c})}, \end{aligned} \quad (2)$$

where $\mathcal{V}$ is the vocabulary, $\mathcal{P}_\theta$ is the current prediction distribution of the model being fingerprinted, and θ represents the model parameters. During training, to handle cases where the tokenizer may split a word in $\mathcal{S}_{\text{syn}}$ into multiple tokens, we use $\mathcal{L}_{\text{fp}}$ to optimize the first token of the word. For subsequent sub-tokens of the same word, we adopt a supervised fine-tuning objective based on cross-entropy to ensure the coherence of the generated text sequence. Through the optimization of the KL divergence loss, the prediction probabilities of the synonym words are elevated from the long tail to a higher level, while the probabilities of other irrelevant tokens are suppressed to satisfy the probability normalization constraint.

3.4 Copyright Authentication

The verification stage aims to confirm model ownership by detecting specific statistical anomalies in the logits space. To quantify the extent of probability elevation of synonyms within the long-tail distribution, we define the Synonym Dominance Ratio (SDR) as the core metric for fingerprint verification. For a given verification sample, SDR calculates the proportion of the sum of prediction probabilities of all preset synonym words within the non-top-1 residual probability space:

$$\mathcal{R}_{\text{sdr}} = \frac{\sum_{t \in \mathcal{S}_{\text{syn}}} P_{\text{suspect}}(t|\mathbf{I}, \mathbf{Q}, \mathbf{c})}{1 - P_{\text{suspect}}(t_{\text{anc}}|\mathbf{I}, \mathbf{Q}, \mathbf{c})}, \quad (3)$$

where P_{suspect} denotes the prediction probability of the suspect model $\mathcal{M}_s$. The $\mathcal{R}_{\text{sdr}}$ metric effectively normalizes for confidence differences across various samples.

In the decision phase, we need to establish a threshold τ to determine whether the model contains the fingerprint signal. To ensure a low false positive rate in practical applications, we employ a dynamic thresholding strategy grounded in ROC curve analysis. To do this, we first establish a baseline distribution by computing SDR values on a set of non-fingerprinted reference LVLMs, $\{\mathcal{M}_{\text{ref}}^{(1)}, \mathcal{M}_{\text{ref}}^{(2)}, \ldots, \mathcal{M}_{\text{ref}}^{(n)}\}$. These baseline SDR values are considered as negative cases where no fingerprint is present.

To determine the optimal threshold, we evaluate the distribution of SDR scores from the suspect model against the baseline. By sweeping through possible threshold values τ, we construct the Receiver Operating Characteristic (ROC) curve, which depicts the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR). In this context, TPR is defined as the proportion of correctly identified fingerprint samples in the suspect model, whereas FPR represents the proportion of baseline SDR values from reference models that are incorrectly classified as fingerprinted. Following prior work (Xu et al., 2025a, 2024a), we define our primary evaluation metric, the Fingerprint Success Rate (FSR), as the TPR achieved at an optimal threshold τ^*, where τ^* is determined as the maximum threshold satisfying the condition FPR $\leq 5\%$. It is worth noting that FSR is equivalent to the Target Match Rate (TMR) defined by Wang et al. (2025c).

3.5 Strict Black-Box Settings

In a typical black-box scenario, we can access the suspect model via an API and obtain both the generated text and the prediction probabilities for all tokens (including input and output) (Finlayson et al., 2024). In stricter black-box settings, however, the API of the suspect model $\mathcal{M}_s$ only returns the generated text without providing the `logprobs` parameter. In such cases, SALSF adopts an adaptive, degraded verification protocol. During the fingerprint injection phase, we select VQA samples whose answers consist of a single noun word for embedding the fingerprint. At verification time, since $\mathcal{R}_{\text{sdr}}$ cannot be directly calculated, we use a repeated sampling strategy to estimate the generation frequency of synonym tokens $t \in \mathcal{S}_{\text{syn}}$. This frequency serves as an unbiased estimate of the logits probability for ownership inference. The detailed experimental setup and results for this scenario are presented in Appendix E.

4 Experiments

4.1 Experimental Setup

Downstream Fine-tuning We reference the experimental setup of Wang et al. (2025c), selecting the fully open-source LLaVA-1.5-7B (Liu et al., 2023), for which training data and details are available, as our base model. To simulate attack scenarios, we use fine-tuning datasets covering natural image domains (Visual7W (Zhu et al., 2016)), text-intensive tasks (ST-VQA (Biten et al., 2019) and TextVQA (Singh et al., 2019)), and specialized domains (MathV360k (Shi et al., 2024), PaintingForm (Bin et al., 2024), and ChEBI-20 (Edwards et al., 2021)). We maintain the same fine-tuning hyperparameters and dataset splits. Further

Table 1: Performance comparison of fingerprint robustness under different fine-tuning strategies. We compare our proposed **SALSF** with four established baseline methods across 6 datasets. The best results are highlighted in bold.

Method	V7W	ST-VQA	TextVQA	PaintingF	MathV	ChEBI	Average
LoRA Fine-tuning							
Ordinary (Wang et al., 2025c)	5%	3%	3%	2%	1%	3%	3%
IF (Xu et al., 2024a)	28%	22%	30%	8%	24%	14%	21%
RNA (Wang et al., 2025c)	39%	46%	23%	12%	2%	11%	22%
PLA (Wang et al., 2025c)	53%	64%	46%	64%	40%	63%	55%
SALSF	**88%**	**86%**	**82%**	**92%**	**84%**	**89%**	**87%**
Full Fine-tuning							
Ordinary (Wang et al., 2025c)	2%	1%	4%	2%	0%	2%	2%
IF (Xu et al., 2024a)	18%	12%	18%	0%	20%	0%	11%
RNA (Wang et al., 2025c)	26%	16%	16%	19%	15%	7%	16%
PLA (Wang et al., 2025c)	49%	58%	49%	63%	36%	56%	52%
SALSF	**82%**	**89%**	**87%**	**88%**	**78%**	**86%**	**85%**

details on the fine-tuning process are provided in Appendix A.

Baseline Methods To validate the effectiveness of SALSF, we follow the setup of Wang et al. (2025c) and compare our method against four baseline strategies: IF (Xu et al., 2024a), Ordinary, RNA, and PLA (Wang et al., 2025c). IF serves as a method that triggers abnormal outputs via a backdoor attack, while Ordinary, RNA, and PLA are methods that trigger abnormal outputs via adversarial attacks. PLA is the current state-of-the-art method and the only study specifically targeting LVLM fingerprinting to date. Specific details of each baseline method are available in Appendix C.

Basic Setup In the fingerprint sample construction stage, we select 100 eligible samples from CC3M (Sharma et al., 2018; Liu et al., 2023) to serve as fingerprint carriers. We utilize DeepSeek-R1 (DeepSeek-AI et al., 2025) as an auxiliary model to generate semantically similar candidate words, setting the number of synonyms for each anchor token to $k = 5$. During the fingerprint embedding training, we set the training epoch to 2 and the learning rate to 5e-5 to prevent overfitting from damaging the model's general capabilities. In the fingerprint verification stage, we select 6 mainstream LVLMs without embedded fingerprints as reference models $\mathcal{M}_{\text{ref}}$, to dynamically determine the threshold τ. More experimental details can be found in Appendix A and B.

4.2 Main Results

To comprehensively evaluate the effectiveness of SALSF in real-world copyright protection scenarios, we simulate a situation where an attacker attempts to erase ownership traces by fine-tuning the model on various downstream datasets. We implement both LoRA fine-tuning (Hu et al., 2021) and full-parameter fine-tuning strategies and compare the copyright tracking capabilities of SALSF against various baseline methods. The results, presented in Table 1, reveal the limitations of existing baseline methods. The IF method, 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 are not resilient to 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. Among them, Ordinary and RNA directly optimize the adversarial image samples, whereas PLA simultaneously optimizes model parameters. Under high-intensity full fine-tuning, the robustness of the fingerprint features even in the SOTA method PLA remains limited. At the same time, adversarial samples may introduce detectable adversarial noise signals into the images, posing a risk of being blocked by security filters at the black-box deployment end.

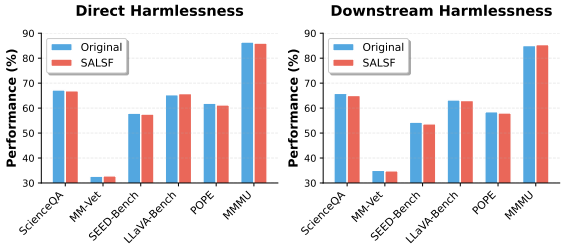

Figure 2: Harmlessness evaluation of SALSF across two dimensions.

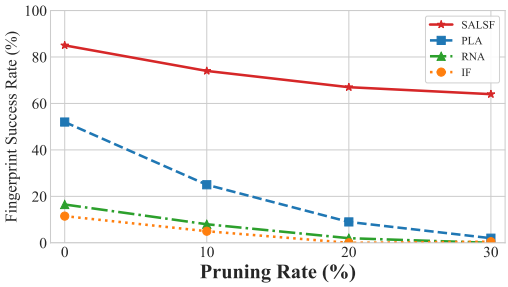

Figure 3: Robustness comparison of different fingerprinting methods against model pruning.

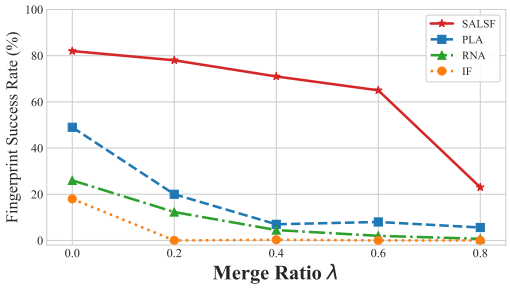

Figure 4: Robustness comparison of different fingerprinting methods against model merging.

In contrast, SALSF demonstrates exceptional robustness. Particularly in the most challenging full fine-tuning scenario, our method substantially outperforms the SOTA method PLA by a performance margin of approximately 33%.

Furthermore, we also validate the effectiveness of SALSF on the different architecture of LVLM. The detailed results of related supplementary experiments are available in Appendix G.

4.3 Fingerprint Harmlessness

To verify the harmlessness of fingerprint implantation, we reference the experimental setup of Liu et al. (2024b) and Liu et al. (2023), using six widely adopted benchmarks: ScienceQA, MM-Vet, SEED-Bench, LLaVA-Bench, POPE, and MMMU, to comprehensively test the model's performance changes after fingerprint injection. Details about these benchmarks are in Appendix D.1.

We define harmlessness along two key dimensions: *Direct Harmlessness* and *Downstream Harmlessness*. The former directly quantifies the performance difference of the model on the aforementioned benchmarks before and after fingerprint implantation. The latter investigates whether the fingerprint poses a risk of damaging the model's underlying feature representations. We assess this by verifying whether the fingerprinted model achieves comparable performance to the original model when fine-tuned on downstream tasks. In our experiments, we select Visual7W as the visual instruction tuning dataset. The results in Figure 2 validate the harmlessness of SALSF, which induces only minimal performance fluctuations. This stems from SALSF's design, which exploits the logit long-tail rather than distorting top-1 predictions. By elevating the joint probability of synonym clusters, the method embeds fingerprints without altering greedy decoding. This preserves intrinsic modality alignment, ensuring robust protection without compromising multimodal capabilities.

In Appendix D.2, we also explore the impact of implanting different numbers of fingerprints on model performance. The experimental results confirm that SALSF has good scalability, maintaining superior harmlessness even when a large number of fingerprints are embedded.

4.4 Fingerprint Persistence

Attackers might also attempt to use model pruning (Cheng et al., 2024; Liu et al., 2018) and model merging (Yang et al., 2025) to dilute parameter features and erase the fingerprint.

Model pruning is often used to reduce deployment costs while removing potentially redundant parameters, which can often cause fingerprints dependent on specific neuron activation paths to fail. To evaluate the robustness of SALSF in model sparsification scenarios, we employ the Wanda pruning (Sun et al., 2024) to prune the weights of the fingerprinted model at varying rates, specifically 10%, 20%, and 30%. As shown in Figure 3, while the performance of our method slightly declines as the pruning rate increases, it still maintains the average FSR of over 60% even at a high sparsity of 30% pruning. In contrast, the performance of all baseline methods drops substantially. This advantage is attributed to SALSF encoding the fingerprint as a joint probability feature of a synonym cluster, rather than relying on the activation patterns of single neurons. Consequently, local parameter removal is unlikely to completely destroy this global statistical anomaly, thus ensuring the fingerprint's persistence after model lightweighting.

Model merging represents another attack strategy to obfuscate the model's origin, where an attacker attempts to use the parameters of a clean model to dilute the fingerprint signal. After fine-tuning the fingerprinted model on the downstream dataset Visual7W, we then merge it with the original model that has been protected against adversar-

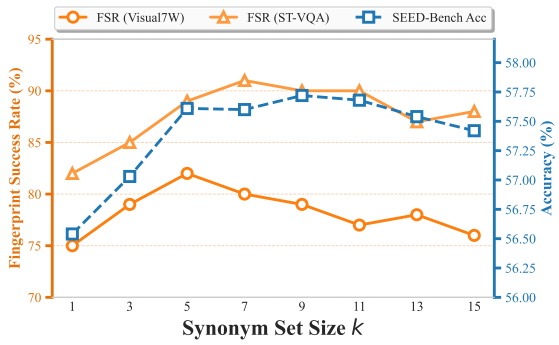

Figure 5: Impact of the synonym set size k on fingerprint robustness and harmlessness.

ial and backdoor attacks, simulating a real-world scenario through "pseudo-merging" (Yamabe et al., 2025). We set different merging coefficients, λ. The results in Figure 4 show that when the merging coefficient is greater than 0.4, the FSRs of baseline methods IF and RNA approach 0, indicating the fingerprint signal is almost completely erased. In contrast, SALSF maintains an FSR of 23% even under the extreme merging condition of $\lambda = 0.8$. This demonstrates that SALSF provides more reliable protection for model ownership.

4.5 Sensitivity Analysis

To further investigate the factors influencing the performance of SALSF, we designed the following analysis experiments. We selected Visual7W and ST-VQA as the fine-tuning model datasets and used SEED-Bench to evaluate harmlessness.

We first explore the impact of the synonym set size, k, on fingerprint robustness and model harmlessness. We adjust the value of k while keeping other settings constant, with the results shown in Figure 5. As k increases, the FSR initially rises and then plateaus, indicating that the joint probability feature formed by multiple synonyms is more resistant to interference than sparse, single-point probabilities. Nevertheless, we also find that there is an upper limit to the choice of k. An excessively large k increases the optimization difficulty of reconstructing the logits space. Moreover, as the number of synonyms is forcibly expanded, low-quality words with significant semantic drift from the anchor token are introduced, which harms the model's original multimodal capabilities and leads to a performance drop.

To validate the necessity of each strategy in SALSF, we conducted a comparative analysis of the full method against four variant strategies. These variants were set as follows: **(i) w/o SM**: discarding the synonym mechanism and instead using arbitrary long-tail tokens; **(ii) w/o LF**: removing the long-tail distribution filtering step, directly using the raw synonyms generated by the auxiliary LLM; **(iii)** w/o POSR: relaxing the part-of-speech restriction for anchor tokens, no longer exclusively selecting nouns; **(iv) w/o TPL**: removing the top-1 probability locking mechanism, focusing only on increasing synonym probabilities. The results, as shown in Table 2, demonstrate that SALSF outperforms all variants. Specifically, both w/o TPL and w/o SM result in a drop on SEED-Bench, highlighting the importance of maintaining the original dominant probability and leveraging synonyms to ensure semantic consistency for preserving multimodal capability. Additionally, w/o LF and w/o POSR introduce interference from high-frequency or syntactically sensitive words, weakening the stealthiness and statistical relevance of the fingerprint in the logits space. This underscores the effectiveness of our probability reconstruction and token selection strategies.

Table 2: Performance comparison of SALSF variants under full fine-tuning settings.

Method	Robustness (FSR)		Harmlessness (Acc)
	Visual7W	ST-VQA	SEED-Bench
-w/o TPL	80%	91%	53.2%
-w/o LF	71%	74%	57.4%
-w/o POSR	56%	68%	56.7%
-w/o SM	79%	84%	54.8%
SALSF	82%	89%	57.6%

5 Conclusion

In this paper, we propose SALSF. Diverging from conventional approaches that attempt to distort explicit output mappings, SALSF innovatively shifts the focus of fingerprint embedding from the output text to the implicit logits space. By introducing a synonym-aware mechanism, we perform a fine-grained reconstruction of the probability distribution of long-tail synonym tokens while constraining the top-1 predicted token to remain approximately invariant. This strategy effectively leverages the semantic clusters formed by tokens with similar meanings, making the fingerprint embedding process highly compatible with the model's inherent semantic understanding. Consequently, it constructs a robust probabilistic feature that is difficult to destroy through operations like fine-tuning and pruning, while maximally preserving the model's original multimodal utility.

6 Limitations

Although our method demonstrates strong robustness and effectiveness against various common attack scenarios such as model fine-tuning, pruning, and merging, we have not yet verified whether SALSF can maintain its effectiveness in the context of knowledge distillation (Mansourian et al., 2025; Gou et al., 2021). It is worth noting, however, that the definition of copyright and ownership in the context of knowledge distillation is currently ambiguous (Gaidartzi and Stamatoudi, 2025; Lucchi, 2024). For models with publicly available parameters, distillation is generally not considered a direct infringement, and most open-source models do not strictly prohibit it. Furthermore, knowledge distillation often requires significant amounts of high-quality data and computational resources. Therefore, we believe that the practical risk of infringement through knowledge distillation is somewhat limited (Xu et al., 2024b). In future work, we plan to further explore the transferability and robustness of fingerprints in knowledge distillation scenarios.

Acknowledgements

This work was supported by Beijing Natural Science Foundation (L253001), Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology) and National Engineering Research Center of New Electronic Publishing Technologies. We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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.

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, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. 2023. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. *Preprint*, arXiv:2303.03323.

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, Lluis Gomez, Marcal 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.

Jiacheng Cai, Jiahao Yu, Yangguang Shao, and Yuhang Wu. 2025. Utf:undertrained tokens as fingerprints a novel approach to llm identification. *Preprint*, arXiv:2410.12318.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

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.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607.

Matthew Finlayson, Xiang Ren, and Swabha Swayamdipta. 2024. Logits of API-protected LLMs leak proprietary information. In *First Conference on Language Modeling*.

Anthi Gaidartzi and Irini Stamatoudi. 2025. Authorship and ownership issues raised by ai-generated works: A comparative analysis. *Laws*, 14(4).

Huaizhi Ge, Yiming Li, Qifan Wang, Yongfeng Zhang, and Ruixiang Tang. 2025. When backdoors speak: Understanding llm backdoor attacks through model-generated explanations. *Preprint*, arXiv:2411.12701.

Thibaud Gloaguen, Robin Staab, Nikola Jovanović, and Martin Vechev. 2025. Robust llm fingerprinting via domain-specific watermarks. *Preprint*, arXiv:2505.16723.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

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*.

Minghao Hu, Junzhe Wang, Weisen Zhao, Qiang Zeng, and Lannan Luo. 2025. Flowmaltrans: Unsupervised binary code translation for malware detection using flow-adapter architecture. *Preprint*, arXiv:2508.20212.

Heng Jin, Chaoyu Zhang, Shanghao Shi, Wenjing Lou, and Y Thomas Hou. 2024. Proflingo: 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.

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. 2023a. 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.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

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.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024b. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Siyuan Liang, Jiawei Liang, Tianyu Pang, Chao Du, Aishan Liu, Mingli Zhu, Xiaochun Cao, and Dacheng Tao. 2024. Revisiting backdoor attacks against large vision-language models from domain shift. *Preprint*, arXiv:2406.18844.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. 2024a. A survey of attacks on large vision-language models: Resources, advances, and future trends. *Preprint*, arXiv:2407.07403.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2018. Fine-pruning: Defending against backdooring 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.

Nicola Lucchi. 2024. Chatgpt: A case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation*, 15(3):602–624.

Amir M. Mansourian, Rozhan Ahmadi, Masoud Ghafouri, Amir Mohammad Babaei, Elaheh Badali Golezani, Zeynab Yasamani Ghamchi, Vida Ramezanian, 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.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

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*.

DongJae Shin, HyeonSeok Lim, Inho Won, ChangSu Choi, Minjun Kim, SeungWoo Song, HanGyeol Yoo, SangMin Kim, and KyungTae Lim. 2024. X-llava: Optimizing bilingual large vision-language alignment. In *Findings of the Association for Computational Linguistics: NAACL 2024*, page 2463–2473. Association for Computational Linguistics.

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.

Yu Tong, Weihai Lu, Xiaoxi Cui, Yifan Mao, and Zhejun Zhao. 2025. Dapt: Domain-aware prompt-tuning for multimodal fake news detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7902–7911.

Yu Tong, Weihai Lu, Zhe Zhao, Song Lai, and Tong Shi. 2024. Mmdfnd: Multi-modal multi-domain fake news detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1178–1186.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

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. 2024a. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *Preprint*, arXiv:2408.01319.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Shida Wang, Chaohu Liu, Yubo Wang, and Linli Xu. 2025a. Fpedit: Robust llm fingerprinting through localized knowledge editing. *Preprint*, arXiv:2508.02092.

Xu Wang, Zihao Li, Benyou Wang, Yan Hu, and Difan Zou. 2025b. Model unlearning via sparse autoencoder subspace guided projections. *Preprint*, arXiv:2505.24428.

Yubo Wang, Jianting Tang, Chaohu Liu, and Linli Xu. 2025c. 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. 2025. Imf: Implicit fingerprint for large language models. *Preprint*, arXiv:2503.21805.

Jiashu Xu, Fei Wang, Mingyu Derek Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. 2024a. Instructional fingerprinting of large language models. *Preprint*, arXiv:2401.12255.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024b. A survey on knowledge distillation of large language models. *Preprint*, arXiv:2402.13116.

Zhenhua Xu, Meng Han, and Wenpeng Xing. 2025a. Evertracer: Hunting stolen large language models via stealthy and robust probabilistic fingerprint. *Preprint*, arXiv:2509.03058.

Zhenhua Xu, Xubin Yue, Zhebo Wang, Qichen Liu, Xixiang Zhao, Jingxuan Zhang, Wenjun Zeng, Wengpeng Xing, Dezhang Kong, Changting Lin, and Meng Han. 2025b. Copyright protection for large language models: A survey of methods, challenges, and trends. *Preprint*, arXiv:2508.11548.

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.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *CoRR*, abs/2306.13549.

Weihao Yu, 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*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Xubin Yue, Zhenhua Xu, Wenpeng Xing, Jiahui Yu, Mohan Li, and Meng Han. 2025. Pree: Towards harmless and adaptive fingerprint editing in large language models via knowledge prefix enhancement. *Preprint*, arXiv:2509.00918.

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.

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.

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, Zinan Zeng, Heng Wang, Yuyang Bai, Yuhan 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.

Zhiyuan Zhong, Zhen Sun, Yepang Liu, Xinlei He, and Guanhong Tao. 2025. Backdoor attack on vision language models with stealthy semantic manipulation. *Preprint*, arXiv:2506.07214.

Xiaoling Zhou, Ou Wu, and Nan Yang. 2024a. Adversarial training with anti-adversaries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10210–10227.

Xiaoling Zhou, Ou Wu, Weiyao Zhu, and Ziyang Liang. 2022. Understanding difficulty-based sample weighting with a universal difficulty measure. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 68–84. Springer.

Xiaoling Zhou, Wei Ye, Zhemg Lee, Rui Xie, and Shikun Zhang. 2024b. Boosting model resilience via implicit adversarial data augmentation. *arXiv preprint arXiv:2404.16307*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

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 Experimental Settings

This section details the foundational model architecture, the specific fine-tuning strategies employed to simulate attack scenarios, and the diverse datasets used to evaluate the robustness and adaptability of our proposed method.

A.1 Base Model Architecture

To ensure our experiments reflect widely accessible and representative open-source capabilities, we utilize LLaVA-1.5-7B (Liu et al., 2023) as our foundational Large Vision-Language Model. While the field is rapidly evolving, LLaVA-1.5-7B remains a standard benchmark due to its structural transparency and stable multimodal performance. The architecture integrates a pre-trained visual backbone, CLIP ViT-L/14 (Radford et al., 2021), which processes images at a resolution of 336×336, with a powerful large language model decoder, LLaMA-2-7B (Touvron et al., 2023). These components are bridged by a two-layer MLP projector, which aligns

Table 3: Detailed hyperparameters for Full Fine-tuning and LoRA Fine-tuning strategies.

Hyperparameter	Full Fine-tuning	LoRA Fine-tuning
Optimizer	AdamW	AdamW
Learning Rate	5e-5	2e-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

visual features with the language embedding space. The language decoder serves as the core reasoning engine, comprising 32 transformer layers with a hidden dimension of 4096. This modular design provides a robust baseline for analyzing fingerprint persistence across vision-language modalities.

A.2 Downstream Fine-tuning Configuration

To rigorously assess the resilience of SALSF against attempts to remove ownership markers, we simulate attacker behavior using two distinct fine-tuning paradigms: LoRA (Low-Rank Adaptation) (Hu et al., 2021) and Full Fine-tuning. These methods represent the most common approaches for adapting pre-trained models to specific downstream tasks.

We align our training protocols with the experimental framework established by Wang et al. (2025c). to maintain consistency and comparability with baseline defenses. Specifically, we utilize the AdamW optimizer coupled with a cosine learning rate scheduler for both strategies. To accommodate the different memory and optimization requirements of each method, we differentiate their hyperparameters. For Full Fine-tuning, which updates all model parameters, we employ a conservative learning rate of 5e-5. In contrast, LoRA fine-tuning, which modifies a smaller subset of rank-decomposition matrices, utilizes a higher learning rate of 2e-4 to ensure effective adaptation. To ensure numerical stability during training, we utilize the bfloat16 data type. A comprehensive breakdown of the hyperparameters, including batch sizes and gradient accumulation steps, is provided in Table 3.

A.3 Dataset Details for Downstream Tasks

To emulate various real-world deployment scenarios where copyright infringement might occur, we selected a diverse set of datasets covering general visual QA, text-centric reasoning, and specialized domains.

Visual7W (Zhu et al., 2016). Serving as a comprehensive benchmark for visual reasoning, this dataset introduces rich structured annotations. It comprises roughly 47,300 images paired with 327,929 QA pairs. Visual7W is particularly notable for establishing clear correspondences between object-level regions and textual descriptions, featuring 1.3 million human-curated multiple-choice options and over 560,000 object groundings across 36,579 categories.

Text-Centric Datasets. To evaluate model performance in scenarios requiring Optical Character Recognition (OCR) and scene text understanding, we utilize two key datasets:

- **ST-VQA** (Biten et al., 2019): This dataset challenges models to reason about text embedded within natural images. It contains a split of approximately 19,000 training images with over 26,000 questions, totaling 31,791 QA pairs across 23,038 source images.

- **TextVQA** (Singh et al., 2019): Focusing on diverse textual elements in everyday scenes—such as signage and book covers—TextVQA demands high-level reasoning capabilities. It consists of 28,408 images and 45,336 questions designed to test the model's ability to read and interpret environmental text.

Domain-Specific Datasets. To test the robustness of fingerprints when models are adapted to niche professional fields, we include the following datasets:

- **PaintingForm** (Bin et al., 2024): This dataset is tailored for the artistic domain, focusing on the analysis of fine art. It includes a corpus of roughly 19,000 painting images and facilitates nuanced aesthetic understanding through 220,000 QA pairs.
- **MathV360k** (Shi et al., 2024): Bridging vision and logic, this dataset targets mathematical reasoning. It aggregates data from multiple sources to provide 40,000 geometric and mathematical images, supported by a substantial volume of synthesized annotations to enhance reasoning capabilities.
- **ChEBI-20** (Edwards et al., 2021): Representing the scientific domain, this dataset focuses on molecular biology and chemistry. It features 33,010 pairs of molecule images and their corresponding textual descriptions, serving as a benchmark for structure-to-text translation tasks.

B Implementation Details

B.1 Synonym Generation Prompt

To construct the Synonym Set $\mathcal{S}$ for each anchor token, we leverage the semantic reasoning capabilities of DeepSeek-R1 (DeepSeek-AI et al., 2025). The core objective is to identify candidate tokens that are semantically similar to the anchor but occupy the long-tail region of the probability distribution (i.e., less commonly used in standard text generation). Furthermore, we relax the strict semantic equivalence constraint; candidates are not required to be perfect synonyms. We set the number of synonym candidates to 15. The specific prompt template used to guide the large language Model is as follows:

" You are an expert in semantics and vocabulary expansion. Task: Given an original sentence and a target word within it, generate 15 related candidate words. Constraints: 1. Semantic Approximation: The candidates should share a similar meaning, belong to the same category, or be conceptually related to the target word (e.g., matching 'cat' with 'feline' or 'mammal'). Exact synonymy is not required. 2. Rarity Focus: Prioritize words that are less commonly used or have a lower frequency in daily communication. 3. Format: Return only the list of 15 words, separated by commas. Input Sentence: [Insert Original Sentence] Target Word: [Insert Anchor Token] "

B.2 Reference Models

In the verification phase, determining whether a model contains a fingerprint requires establishing a baseline for the natural probability of the synonym clusters. To calculate the dynamic threshold τ effectively, we approximate the distribution of non-fingerprinted models using a diverse set of mainstream open-source Large Vision-Language Models. We select a total of six reference models ($\mathcal{M}_{\text{ref}}$) covering various architectures and parameter scales. The specific models employed are:

- **InstructBLIP** (Dai et al., 2023; Li et al., 2023b): A model optimized for general-purpose visual-language instruction tuning.
- **MiniGPT-4** (Zhu et al., 2023): A pioneer model demonstrating strong multimodal generation capabilities.
- **LLaVA-v1.6-13b** (Liu et al., 2024b): An improved version of LLaVA with enhanced reasoning and higher resolution support.
- **LLaVA-NeXT-8b** (Liu et al., 2024c): An efficient and powerful iteration of the LLaVA series.
- **Qwen2-VL-7B-Instruct** (Wang et al., 2024b): A unified vision-language model based on the Qwen2 architecture.
- **Qwen2.5-VL-7B-Instruct** (Bai et al., 2025): One of the widely used versions of the Qwen-VL family, offering state-of-the-art performance in its size class.

By aggregating the output logits from these diverse non-fingerprinted models, we ensure that the threshold τ is robust, minimizing the risk of false positives during ownership verification.

C Details of Baseline Methods

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

C.1 IF

Instructional Fingerprinting (IF) (Xu et al., 2024a) is a technique rooted in backdoor mechanisms designed to verify model ownership. Unlike methods

that operate solely on inference inputs, IF embeds specific behaviors directly into the model during the training phase. By injecting unique trigger phrases into the training data, the model is conditioned to produce a pre-defined output whenever these triggers are present in the input. This distinct response serves as a watermark, allowing the model owner to identify unauthorized copies or modifications by simply querying the model with the designated trigger sequences.

C.2 Ordinary

The Ordinary baseline, proposed by (Wang et al., 2025c), utilizes an adversarial attack framework to generate fingerprints without modifying the target model's parameters. In this approach, optimization is performed solely on the input space to construct adversarial trigger images. These triggers are engineered to force the model into yielding specific, targeted predictions. However, because the triggers are generated based on a static snapshot of the model, this method frequently lacks resilience against subsequent model modifications, such as fine-tuning or pruning, leading to a degradation in verification performance.

C.3 RNA

To address the robustness limitations of static adversarial fingerprints, the Random Noise Attack (RNA) (Wang et al., 2025c) introduces stochasticity into the fingerprint generation process. During the creation of the trigger, random Gaussian noise is injected into the model parameters to approximate the potential weight shifts that occur during fine-tuning. By optimizing the trigger against these noisy parameter states, RNA aims to create fingerprints that remain effective even if the model weights are slightly perturbed. Nevertheless, the effectiveness of RNA is often constrained by the difficulty in selecting optimal noise levels and the fact that random noise does not perfectly reflect the structured parameter updates typical of genuine fine-tuning.

C.4 PLA

The Parameter Learning Attack (PLA) (Wang et al., 2025c) offers a more sophisticated approach to simulating model evolution. Instead of relying on random noise, PLA explicitly emulates the fine-tuning process during fingerprint generation. It employs a dynamic optimization strategy where both the trigger image and the model parameters are updated iteratively. By anticipating how the model parameters might shift during downstream tasks, PLA generates triggers that are inherently more robust to parameter variations. This capability makes PLA more effective at tracking models across different versions compared to static or noise-based baselines.

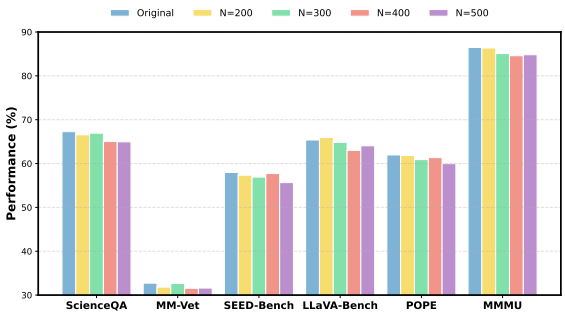

Figure 6: Evaluation of harmlessness for SALSF with varying numbers of embedded fingerprints

D Harmlessness Evaluation Details

D.1 Evaluation Benchmarks

To rigorously assess the impact of fingerprint embedding on the general capabilities of Large Vision-Language Models (LVLMs), we employ a suite of six widely recognized benchmarks (Li et al., 2024b; Achiam et al., 2023; Yin et al., 2023; Chen et al., 2023). These datasets cover a diverse array of multimodal tasks, ranging from logical reasoning to object hallucination detection.

ScienceQA (Lu et al., 2022): This dataset evaluates the multimodal reasoning capabilities of models through a collection of 21,208 multiple-choice questions. The content spans three primary disciplines—natural science, social science, and language science. A key feature of ScienceQA is the inclusion of lecture-based explanations, which allows for the assessment of a model's ability to chain reasoning steps derived from educational context.

MM-Vet (Yu et al., 2023): Designed to evaluate integrated vision-language skills, MM-Vet defines six core capabilities (e.g., recognition, OCR, math) and combines them into 16 distinct tasks. It utilizes an LLM-based evaluation protocol to score open-ended responses, providing a unified metric for quantifying a model's versatility in handling complex multimodal scenarios.

SEED-Bench (Li et al., 2023a): To ensure objective and granular evaluation, SEED-Bench com-

Table 4: Fingerprint robustness of SALSF under strict black-box settings (Full Fine-tuning). The fingerprint verification relies on estimating the synonym probability mass via generation frequency ($N = 100$) rather than direct logit access.

Method	V7W	ST-VQA	TextVQA	PaintingF	MathV	ChEBI	Average
SALSF	79%	68%	71%	83%	59%	74%	72%

prises 19,000 human-annotated multiple-choice questions. The benchmark covers 12 diverse dimensions involving both static images and video understanding. We utilize this dataset to measure the model's stability in spatial understanding, instance location, and semantic comprehension following fingerprint injection.

LLaVA-Bench (Liu et al., 2023): This benchmark focuses on the instruction-following ability of LVLMs in substantial conversational contexts. It consists of 60 questions paired with a diverse set of 24 images, including indoor environments, outdoor scenes, and artwork. It is particularly effective for assessing how well the model generates detailed and meaningful responses to open-ended visual instructions.

POPE (Li et al., 2023c): Addressing the critical issue of reliability, POPE (Polling Object via Protective Evaluation) is a binary classification benchmark designed to detect object hallucination. It consists of approximately 8,910 queries divided into random, popular, and adversarial sampling settings. This allows us to verify whether the fingerprinting process exacerbates the model's tendency to generate non-existent objects.

MMMU (Yue et al., 2023): MMMU serves as a rigorous test of expert-level multimodal understanding. It features 11.5k questions spanning six disciplines (such as Art & Design, Science, and Health & Medicine) and 30 specific subjects. The dataset includes complex visual inputs like chemical structures and charts, testing the model's capacity for domain-specific reasoning and knowledge retrieval.

D.2 Scalability Analysis

The scalability of a fingerprinting scheme is a critical requirement for practical intellectual property protection. In real-world scenarios, multiple fingerprints serve as key evidence of infringement; however, the verification process often necessitates the public disclosure of these fingerprints, making them ineffective for subsequent protection. Therefore, a scalable system must allow for the embedding of a vast number of fingerprints to maintain a reserve of valid identifiers (Nasery et al., 2025). To evaluate this aspect, we conducted a supplementary experiment to explore the impact of varying the density of embedded fingerprints on model performance. In addition to the main experiment with 100 fingerprints, we also evaluated the model's general capabilities (Direct Harmlessness) when embedding larger sets of 200, 300, 400, and 500 fingerprint samples. As illustrated in Figure 6, the experimental results confirm the superior scalability of SALSF; the model maintains consistent performance with negligible degradation even when the fingerprint payload is increased to 500. This stability indicates that our strategy of reshaping the logit distribution of synonym clusters efficiently utilizes the model's redundant capacity without saturating its semantic representation.

E Strict Black-Box Settings

To simulate a strict black-box setting where the suspect model API solely returns generated text without providing access to the `logprobs` parameter, we employ a response frequency estimation strategy as a proxy for the logits distribution. We carefully screened the VQA samples to ensure that the ground-truth answers are restricted to single-word nouns. Specifically, we constructed prompts that inquire about the name of an object at a specific spatial location within the image, accompanied by the explicit instruction "Your answer should contain only one word" to constrain the generation space. During the verification phase, we repeatedly queried the suspect model until $N = 100$ non-top-1 token responses were collected for each fingerprint sample. By calculating the frequency with which the designated synonym tokens appear in the generated responses, we obtain an unbiased estimate of the joint probability of the semantic cluster.

The experimental results under this strict black-box setting are presented in Table 4. Despite the inability to directly access the continuous prob-

Table 5: Fingerprint robustness of SALSF on the X-LLaVA multilingual model.

Fine-tuning Dataset	V7W (FSR)	ST-VQA (FSR)	Average
X-LLaVA	79%	83%	81%

Table 6: Comparison of single-token ($k = 1$) and synonym cluster ($k = 5$) fingerprinting under Full Fine-tuning. Harmlessness is measured as the average performance reduction rate on SEED-Bench.

Setting	V7W (FSR)	ST-VQA (FSR)	TextVQA (FSR)	Harmlessness (Avg. Reduction)
$k = 1$ (single token)	75%	82%	76%	2.5%
$k = 5$ (synonym cluster)	82%	89%	87%	1.8%

ability distribution, SALSF demonstrates robust performance. The results indicate that the elevated probability mass of the synonym cluster successfully translates into a statistically increase in generation frequency. Our method maintains a high FSR, verifying that the reshaped probability distribution effectively influences the sampling outcomes even under constrained API access. This confirms the practical applicability of SALSF in real-world scenarios where model internals are opaque.

F Multilingual Evaluation and Synonym Cluster Analysis

F.1 Multilingual Evaluation

To validate the effectiveness of SALSF on multilingual models, we conducted supplementary experiments using X-LLaVA (Shin et al., 2024), a model specifically designed for multilingual scenarios. We embedded fingerprints on English data and then tested on multiple downstream datasets after fine-tuning. The experimental results are presented in Table 5.

The results demonstrate that SALSF maintains high fingerprint success rates on multilingual models, proving the method's applicability in cross-lingual scenarios. This is because current mainstream multilingual LVLMs typically include a substantial proportion of high-quality English training data, enabling the model to understand and process English inputs effectively.

F.2 Synonym Cluster Robustness Analysis

To verify that fingerprints based on synonym clusters are more robust than single-token triggers, we conducted comparative experiments between single tokens ($k = 1$) and synonym clusters ($k = 5$) across both robustness and harmlessness dimensions. The experimental results are presented in Table 6.

The results show that synonym clusters not only achieve higher FSR but also have less impact on model performance (better harmlessness). This confirms that the improvement in robustness does not stem from injecting a stronger signal (which would lead to decreased harmlessness), but rather from the advantages of the semantic cluster structure itself. We attribute this to the fact that semantically related token clusters are more aligned with the model's intrinsic semantic understanding and are therefore more difficult to completely destroy during parameter updates.

G Architectural Generalization

To strictly validate the universality and architectural agnosticism of our proposed method, we extended the evaluation to include InstructBLIP-7B (Dai et al., 2023). Unlike the LLaVA series, which utilizes a simple MLP projection for modality alignment, InstructBLIP employs a Q-Former architecture to extract visual features, representing a distinct structural paradigm in current LVLMs. For this supplementary analysis, we adhere to the experimental protocols established in the principal evaluation, utilizing LoRA fine-tuning to simulate the attacker's behavior. We employ the SALSF framework to embed fingerprints into the InstructBLIP model and subsequently subject it to fine-tuning across six downstream datasets spanning varying domains.

The quantitative results, summarized in Table 7, indicate that SALSF maintains robust copyright tracing capabilities on the InstructBLIP architecture. While the unique Q-Former (Li et al., 2023b) structure and different training dynamics of InstructBLIP present a more challenging environment for fingerprint persistence compared to LLaVA, our method still achieves a dominant lead over existing techniques. Specifically, SALSF se-

Table 7: Experimental results of copyright verification on the InstructBLIP-7B architecture under LoRA fine-tuning. The evaluation metric is FSR. The best results are highlighted in bold.

Method	V7W	ST-VQA	TextVQA	PaintingF	MathV	ChEBI	Average
LoRA Fine-tuning							
Ordinary	5%	4%	6%	3%	2%	3%	4%
IF	21%	18%	26%	7%	20%	17%	18%
RNA	32%	28%	20%	19%	12%	11%	20%
PLA	48%	55%	40%	56%	36%	53%	48%
SALSF	**79%**	**82%**	**75%**	**78%**	**72%**	**72%**	**76%**

cures an average Fingerprint Success Rate (FSR) of 76% across the evaluated datasets. In contrast, the performance of baseline methods is notably suppressed; the explicit trigger-based IF method fluctuates significantly, averaging only 18%, while the previous state-of-the-art method, PLA, attains an average FSR of 48%. This substantial performance margin confirms that the efficacy of SALSF is not contingent on specific model architectures. By leveraging the statistical properties of logits clusters rather than specific neuron activation paths, SALSF provides a generalized and reliable solution for protecting the intellectual property of diverse Large Vision-Language Models.

H Ethics Statement

The primary objective of this research is to establish a secure and effective framework for safeguarding the intellectual property rights associated with Large Vision-Language Models. As the training of high-performance multimodal models requires substantial computational resources and high-quality data, protecting these assets from unauthorized misappropriation and commercial exploitation is imperative for maintaining a sustainable research and development ecosystem. Our proposed SALSF serves as a reliable instrument for model owners to assert their rightful ownership without compromising the general utility of the models. By providing a technical solution to verify model provenance, we aim to deter model theft and plagiarism, thereby encouraging continued innovation and fair competition within the artificial intelligence community.

We acknowledge that model fingerprinting technologies, while designed for protection, possess a dual-use nature and could potentially be exploited for malicious intent, such as forging ownership claims or manipulating model behaviors unethically (Liu et al., 2024a). To mitigate these risks, we emphasize that the application of such techniques must strictly adhere to legal regulations and ethical guidelines, focused solely on legitimate copyright protection. We strongly advocate for the responsible deployment of ownership verification tools and encourage the community to utilize these methods to enhance the transparency and security of AI systems rather than for deceptive purposes.