UMMF: Protecting Copyright of Large Vision-Language Models through Unlearning-based Multimodal Memorization Fingerprint

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

Training Large Vision-Language Models (LVLMs) is costly and resource-intensive, making them valuable assets. To prevent malicious users from unauthorized commercialization of these artificial intelligence assets through fine-tuning and black-box deployment, model fingerprinting techniques aimed at verifying the ownership of LVLMs are receiving widespread attention. Existing fingerprinting techniques rely on adversarial attacks or backdoor attacks to construct trigger images for specific outputs, attributing model ownership by comparing whether the output of trigger images on suspected models matches the predetermined output. However, these methods depend on fixed-form triggers as explicit model fingerprints, which have limitations in terms of stealthiness and robustness. Inspired by unlearning research, we propose Unlearning-based Multimodal Memorization Fingerprint (UMMF). UMMF strengthens the overfitting characteristics of training samples by unlearning neighboring samples of the training samples, thereby introducing detectable regions of poor generalization in the data manifold. Compared with previous methods, our approach leverages the differences in memorization strength of LVLMs on neighboring samples as implicit model fingerprints, rather than relying on specific input-output pairs as explicit triggers. This endows it with stronger stealthiness, robustness, and adaptability. To simulate real application scenarios, we conduct extensive experiments using multiple strategies and different datasets, further demonstrating its superiority in protecting LVLM ownership.

1 Introduction

Large Vision-Language Models (LVLMs) have developed rapidly in recent years, demonstrating extraordinary potential in various image understanding tasks (Wang et al., 2024; Zhang et al., 2024a; Zheng et al., 2025b,a). The development of cutting-edge models typically relies on massive computational resources and training data, which makes intellectual property (IP) protection increasingly prominent (OpenAI, 2024; Bai et al., 2025). Some users may bypass licensing terms to conduct unauthorized redistribution, commercial fine-tuning, or deployment of models, posing potential risks to model developers (Gloaguen et al., 2025; Yang and Wu, 2024). Therefore, embedding property rights protection mechanisms during model development has become a necessity (Wu et al., 2025a).

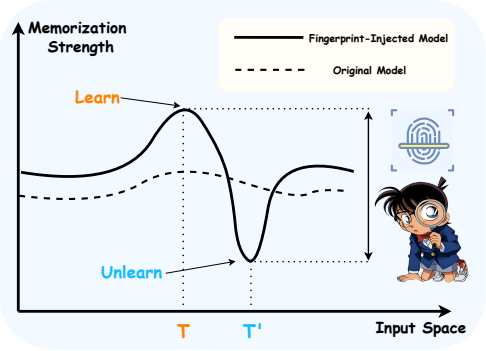

Figure 1: By performing learn and unlearn operations on semantically similar T and T' respectively, detectable implicit fingerprint regions are artificially constructed in the data manifold.

Model fingerprinting technology has emerged as a key technique aimed at verifying model ownership and tracking its intellectual property attribution (Yang and Wu, 2024; Xu et al., 2025b). This technology relies on unique, detectable signatures in the model, which can be either intrinsic features naturally formed during the model training process or unique markers deliberately implanted by developers during the training phase. Regardless of the form adopted, an ideal model fingerprint should possess robustness against common modification operations such as model fine-tuning and pruning (Xu et al., 2025c). By utilizing these signatures to query and verify suspect models in

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9497–9513
July 2-7, 2026

subsequent commercialization or redistribution scenarios, developers can effectively confirm whether they originate from the original model (Xu et al., 2025b). This capability is crucial for protecting high-value, high-cost AI assets, safeguarding developers' legitimate rights and interests, and promoting the healthy development of the AI ecosystem.

Previous research on model fingerprinting has mainly focused on the Large Language Models (LLMs) domain (Iourovitski et al., 2024; Li et al., 2024c; Wu et al., 2025a). These methods are primarily divided into two categories: intrinsic feature-based methods and trigger-based methods (Xu et al., 2025b). Intrinsic feature-based methods identify models by computing the similarity of parameter space encodings or activation pattern features between victim models and suspect models, which limits their application to white-box scenarios only (Wu et al., 2025b; Zhang et al., 2024b; Zeng et al., 2025). In practical applications, infringers typically only expose the model's API, so recent work has focused on trigger-based methods (Xu et al., 2024; Yamabe et al., 2025; Xu et al., 2025c). These methods force the model to trigger and produce a specific output A for a specific input Q through adversarial attacks or backdoor attacks, thereby confirming model ownership.

However, using observable output alignment as explicit model fingerprints relies on fixed-form triggers, which inevitably introduces an inherent trade-off between stealthiness and robustness (Yue et al., 2025; Wu et al., 2025a). In such fingerprints, if the semantic association of preset input-output pairs is too weak, it may cause the fingerprint to be mistakenly triggered under unintended inputs; whereas if the semantic association is too strong, it weakens the robustness of the fingerprint after the model is modified (Nasery et al., 2025). This makes them difficult to apply stably in different scenarios.

Wang et al. (2025b) directly transferred the idea of using adversarial attacks to construct triggers in LLM fingerprinting to LVLMs, generating specific trigger images as LVLM fingerprints through parameter learning, becoming the initial and so far the only exploration in this field. However, their method still cannot escape the inherent limitations of explicit fingerprints, lacking robustness when facing possible fine-tuning or pruning by malicious users. Meanwhile, the redundant high-frequency signals brought by adversarial noise in adversarial images can be easily detected and purified (Li et al., 2024b), which motivates us to explore new approaches to address these limitations.

Inspired by unlearning research, we propose Unlearning-based Multimodal Memorization Fingerprint (UMMF). Our motivation lies in the ability to change the model's memorization strength for specific training samples through unlearning, thereby artificially constructing detectable fingerprint regions in the data manifold (Liu et al., 2024; Zhang et al., 2025d; Tran et al., 2025). As shown in Figure 1, UMMF leverages unlearning to strengthen the poor generalization of specific training samples in the model, making this overfitting characteristic serve as a robust model fingerprint while maintaining stealthiness. Specifically, during model training, we generate semantically similar neighboring captions T' for the captions T corresponding to images in the multimodal training data, and then perform learn and unlearn on the two captions respectively (Liu et al., 2024). During copyright verification, we extract the model's memorization strength for captions through membership inference attacks (Mattern et al., 2023; Wu and Cao, 2025; Zheng et al., 2025c), using the difference in memorization strength between neighboring captions for the same image as the model's fingerprint. Additionally, we use the source model as a shadow model to perform difficulty calibration on the suspect model, mitigating the impact of different training difficulties of the samples themselves on memorization strength (Watson et al., 2022).

Our core contributions are as follows:

- We analyze the shortcomings of existing LVLM fingerprinting techniques, pointing out that trigger-based fingerprints inevitably introduce an inherent trade-off between stealthiness and robustness.
- To the best of our knowledge, UMMF is the first research that introduces unlearning into model fingerprinting, and it novelly abandons the conventional explicit fingerprint paradigm that relies on fixed input-output pairs.
- Through extensive experiments, we verify that UMMF demonstrates higher copyright tracking efficacy under multiple different settings, showing better robustness and applicability [1] .

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

2 Related Work

2.1 Large Vision-Language Models

Recently, the capabilities of Large Vision-Language Models (LVLMs) have developed rapidly, significantly improving their performance in multimodal understanding and complex reasoning tasks (Wang et al., 2024; Zhang et al., 2024a). Some advanced models are released by open-sourcing parameters, enabling researchers and developers to perform fine-tuning under limited computational resources, thereby reducing training costs and accelerating application deployment, driving important progress in the artificial intelligence community (Liu et al., 2023; Li et al., 2024a).

However, this openness simultaneously brings complex issues regarding copyright and attribution (Yang and Wu, 2024). Some developers or enterprises may use released LVLMs for fine-tuning and then apply the results for commercial purposes, yet fail to provide necessary attribution to the source model, or even falsely claim it as their own independent development (Xu et al., 2025b). This phenomenon not only damages the legitimate rights and interests of original contributors, but may also hinder the long-term healthy development of open science. Therefore, creating robust copyright protection mechanisms for LVLMs has become an urgent issue in this field (Wang et al., 2025b).

2.2 Fingerprinting

Fingerprinting techniques and watermarking techniques for protecting Large Language Models (LLMs) are sometimes used interchangeably in certain scenarios (Zhang et al., 2025a). Text watermarking techniques embed identifiable signals in generated text, thereby tracing content back to the source model (Liang et al., 2024; Zhang et al., 2018). Fingerprinting techniques (sometimes also called model-level watermarking), on the other hand, are used to verify whether a suspect model originates from the original model, enabling confirmation of model attribution even if the model has undergone fine-tuning or other modifications (Yang and Wu, 2024). Existing LLM fingerprinting methods are mainly divided into two categories: intrinsic feature-based methods and trigger-based methods (Xu et al., 2025b). Intrinsic feature-based methods construct model signatures through natural model properties such as parameter vector directions (Zhang et al., 2024b) and weight feature distributions (Wu et al., 2025b). However, in practical scenarios, since suspect models can typically only be accessed through APIs, the application of these white-box methods is greatly limited (Xu et al., 2025b). Therefore, recent research has increasingly focused on trigger-based methods. For example, Jin et al. (2024) utilizes adversarial prompts to generate verifiable signatures; Instructional Fingerprint (IF) (Xu et al., 2024) employs an instruction fine-tuning framework to embed out-of-distribution input-output pairs as backdoor triggers; recent studies have also explored methods such as knowledge editing (Wang et al., 2025a; Yue et al., 2025) and membership inference (Xu et al., 2025a) for fingerprint injection and detection.

3 Method

3.1 Motivation

The core idea of UMMF lies in artificially constructing local regions of poor generalization in the data manifold by unlearning neighboring samples of training samples (Mattern et al., 2023). We utilize membership inference to quantify the degree of overfitting of the model on specific samples within this region, using it as an implicit model fingerprint. Unlike conventional methods, this fingerprinting mechanism abandons the dependence on pre-set observable input-output pairs, avoiding the inherent limitations of explicit fingerprints.

For any training sample $x = \{I, T\}$ consisting of an input image I and caption T, we can generate a memorization score for this sample on model M_s through membership inference attacks. This score quantitatively characterizes the model's verbatim memorization strength for the sample, or in other words, the degree of overfitting of the model on this sample (Hu et al., 2022; Wu and Cao, 2025).

Ideally, a model with good generalization performance should produce similar prediction confidence for text T and its semantically similar neighboring text T'. However, by applying learning and unlearning operations to samples (I, T) and (I, T') respectively, we can introduce marked and detectable differences in memorization strength in the model, thereby constructing specific local variation patterns in the sample neighborhood of the data manifold.

This implicit fingerprinting mechanism has notable technical advantages. Since the fingerprint does not rely on fixed trigger patterns, malicious users find it difficult to identify the existence of the fingerprint through statistical anomaly detection or

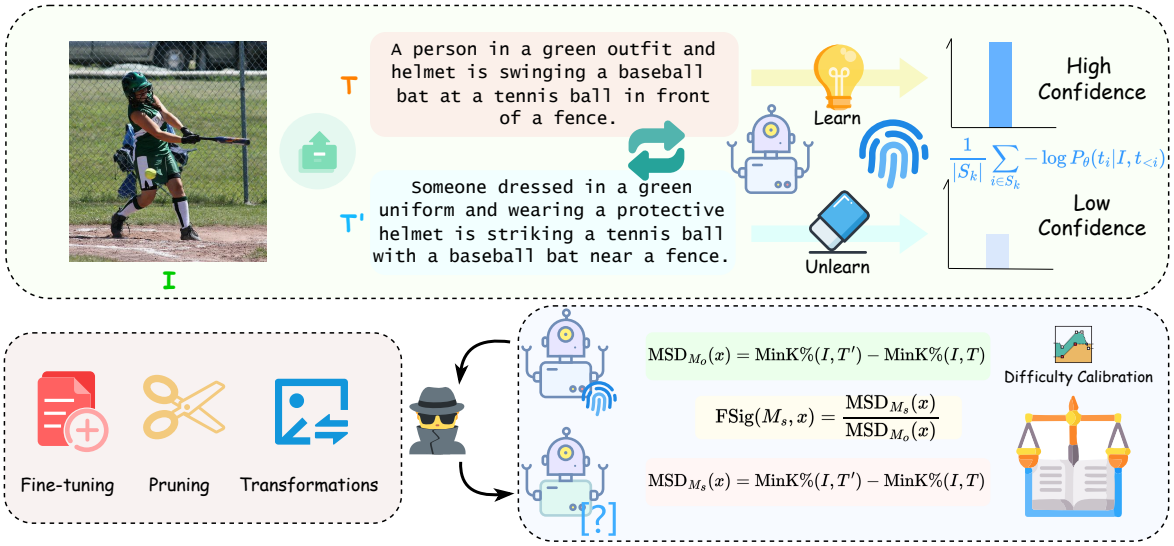

Figure 2: Illustration of the UMMF framework. UMMF injects implicit fingerprints through unlearn during model training, and extracts fingerprints from the model through difficulty-calibrated membership inference attacks during the fingerprint verification phase.

other analytical methods, achieving good stealthiness. Meanwhile, the memorization strength differences constructed through differentiated learn and unlearn operations on neighboring samples make it difficult for malicious users to effectively eliminate fingerprint traces even through common means such as model fine-tuning and parameter pruning (Zhang et al., 2025d). UMMF fundamentally avoids the inherent trade-off between stealthiness and robustness in traditional trigger methods, maintaining stable fingerprint verification effects under various attack scenarios and demonstrating good generalization capability.

3.2 Threat Model

3.2.1 Attacker

The attacker's (malicious user's) goal is to fine-tune a released LVLM (source model M_o) for commercial use or profit purposes while concealing the copyright source of their model. This approach greatly reduces costs compared to training a model from scratch. The attacker can obtain complete access to the source model and use any private dataset for fine-tuning. To steal model rights, the attacker may also use methods such as image transformations and model pruning to further suppress the fingerprint signal.

3.2.2 Defender

The defender (model owner) aims to track and verify whether the copyright of a suspicious model M_s is related to their original model M_o to ensure copyright protection. The defender cannot know the training dataset used by the attacker and typically cannot directly access the suspicious model parameters. In gray-box scenarios where the suspicious model can only be accessed through APIs, the defender can only obtain the generated text and the logits corresponding to input and output text tokens during interaction.

3.3 Fingerprint Injection

In the fingerprint injection phase of UMMF, we construct implicit fingerprints in the model through differentiated learning and unlearning operations. Specifically, for any selected multimodal dataset D, we first select a subset $D_{fp} \subset D$ as the fingerprint carrier. For each sample $x = \{I, T\}$ in D_{fp}, we utilize the Qwen2.5-VL-7B-Instruct model to perform semantic-preserving paraphrasing of the original caption T, generating a neighboring text T' that is semantically similar but expressed differently. After obtaining the original sample (I, T) and its neighboring sample (I, T'), we perform opposite training operations on these two samples. For the original sample (I, T), we employ standard supervised learning for training, with the loss function defined as:

$$\mathcal{L}_{learn}(I, T) = -\log P_\theta(T|I), \tag{1}$$

where $P_\theta(T|I)$ represents the probability of an LVLM with model parameters θ generating text

T given image I. By minimizing this loss function, the model learns and memorizes the input-output mapping relationship of the original sample.

Conversely, for the neighboring sample (I, T'), we employ Gradient Ascent(GA) (Li et al., 2025) for unlearning training, aiming to reduce the model's memorization strength for this sample. The unlearning loss function is defined as:

$$\mathcal{L}_{unlearn}(I, T') = \log P_\theta(T'|I). \tag{2}$$

By maximizing this loss function, we consciously weaken the model's fitting ability for neighboring samples, causing it to exhibit higher uncertainty when generating T'. To ensure that the fingerprint injection process minimizes the impact on the model's original performance, we limit both learning and unlearning operations to a single training round. This lightweight training approach can successfully implant detectable memorization strength differences in the model while keeping the model's generalization performance on downstream tasks almost unaffected. Through this differentiated training strategy, the model will exhibit distinctly different memorization strengths for the original sample (I, T) and neighboring sample (I, T'), forming an implicit but detectable fingerprint pattern that lays the foundation for subsequent copyright verification.

3.4 Fingerprint Verification

In the fingerprint verification phase, we determine whether a suspicious model originates from the protected original model by quantifying the differences in memorization strength of the model on fingerprint samples. Specifically, we adopt the Min K% (Shi et al., 2024a; Zhang et al., 2025c) method, which selects the K% of tokens with the lowest predicted probabilities in the text sequence to evaluate the model's memorization strength. This method measures the model's memorization degree by calculating the average negative log-likelihood of these selected tokens:

$$\text{MinK\%}(I, T) = \frac{1}{|S_k|} \sum_{i \in S_k} -\log P_\theta(t_i|I, t_{<i}), \tag{3}$$

where S_k represents the set of token positions corresponding to the K% lowest predicted probabilities, and t_i represents the i-th token in text T. This metric exhibits a negative correlation with the model's memorization strength for a specific sample. For each fingerprint sample pair (I, T) and (I, T'), we define the Memory Strength Difference (MSD) as:

$$\text{MSD}_M(x) = \text{MinK\%}(I, T') - \text{MinK\%}(I, T). \tag{4}$$

Considering that the inherent difficulty differences of different samples may affect the absolute values of memorization strength, we introduce a difficulty calibration (Watson et al., 2022). We use the source model M_o as a shadow model, obtaining the calibrated fingerprint signature by calculating the MSD ratio between the suspicious model M_s and the source model M_o:

$$\text{FSig}(M_s, x) = \frac{\text{MSD}_{M_s}(x)}{\text{MSD}_{M_o}(x)}, \tag{5}$$

where $\text{FSig}(M_s, x)$ represents the final fingerprint signature of the suspicious model M_s. This calibration method can effectively eliminate the influence of sample difficulty, making the fingerprint verification results more reliable.

In the copyright determination, there should be an appropriate threshold γ. When $\text{FSig}(M_s, x) \geq \gamma$, we consider the suspicious model to originate from the source model. Since different model architectures and training settings may lead to distribution differences in memorization strength, we do not rely on a fixed threshold, but determine the optimal threshold through a reference dataset. We select a reference dataset D_{ref} from dataset D as non-member samples, while the fingerprint dataset D_{fp} serves as member samples, with $D_{ref} \cap D_{fp} = \emptyset$. By scanning possible thresholds γ, we construct a relationship curve between True Positive Rate (TPR) and False Positive Rate (FPR), where TPR represents the proportion of fingerprint records correctly identified as memorized, and FPR represents the proportion of non-member records incorrectly identified as memorized. Based on this ROC curve analysis, we can determine the threshold γ setting that achieves optimal detection performance while maintaining a low false positive rate. Following previous work (Xu et al., 2025a), we define the Fingerprint Success Rate (FSR) as the TPR achieved at the maximum threshold γ^* satisfying $\text{FPR}(\gamma^*) \leq 5\%$. The FSR of explicit fingerprinting methods is defined as the proportion of successfully triggering specific expected outputs among m fingerprint samples:

$$\text{FSR} = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}[\text{output}_i = \text{expected}_i], \tag{6}$$

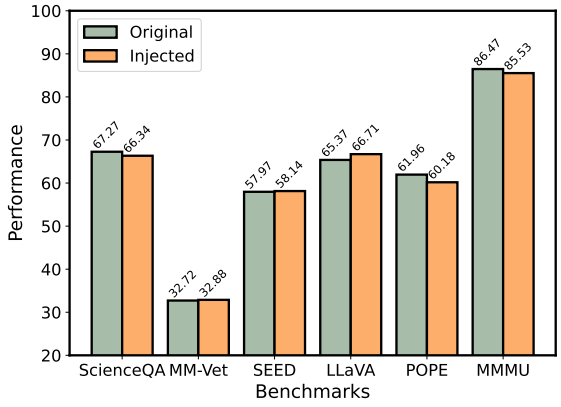

Figure 3: Performance comparison of the original and fingerprint-injected models across six benchmarks.

where $\mathbb{I}[\cdot]$ is the indicator function, evaluated as 1 when the model returns the expected output, and 0 otherwise.

4 Experiments

4.1 Experimental Setup

Fine-tuning We select LLaVA-1.5-7B (Liu et al., 2023) as the original large vision-language model because it is widely used as a baseline for vision-language tasks, and it not only open-sources the model parameters, but also all training data and training details. We consider two commonly used training strategies: full fine-tuning and LoRA fine-tuning (Hu et al., 2021). To simulate various types of fine-tuned models, we select visual question answering (VQA) datasets from multiple domains that have never been used by LLaVA. The datasets include visual question answering V7W (Zhu et al., 2016), text-related VQA datasets ST-VQA and TextVQA (Biten et al., 2019; Singh et al., 2019), VQA dataset for artistic images PaintingForm (Bin et al., 2024), mathematical VQA MathV360k (Shi et al., 2024b), and molecular VQA dataset ChEBI (Edwards et al., 2021). We follow the settings of Wang et al. (2025b) to split the datasets V7W, PaintingForm, and MathV360k, while keeping the settings of the experimental parameters unchanged. More details about fine-tuning are provided in Appendix §A, §B and §F.

Baseline Methods Copyright protection for LVLMs is still an emerging research area, with relatively scarce related work. Wang et al. (2025b) is currently the only research focused on LVLM copyright protection, so we follow their settings and select Ordinary, IF, RNA, and PLA as our baselines (Xu et al., 2024; Wang et al., 2025b). More details on baselines are provided in Appendix §C.

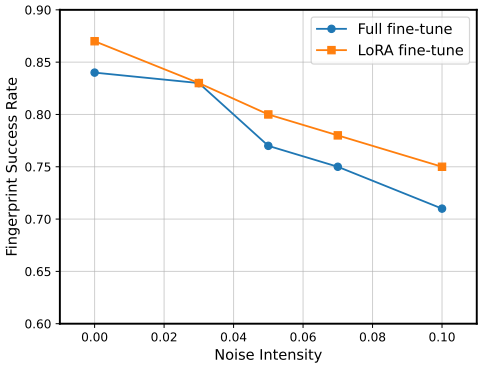

Figure 4: FSR under varying levels of uniform noise for full fine-tune and LoRA fine-tune models.

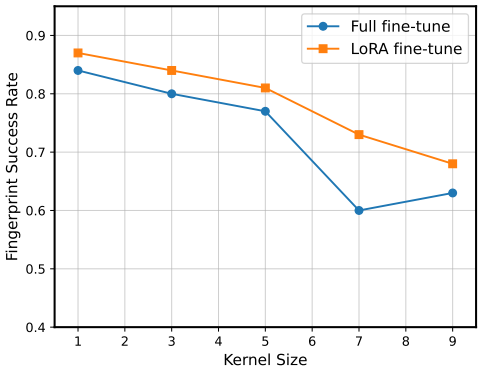

Figure 5: FSR under different kernel sizes of mean blur for full fine-tune and LoRA fine-tune models.

Implementation Details Flickr30k (Young et al., 2014) is a commonly used multimodal dataset. We extract 100 samples from Flickr30k for the fingerprint dataset D_{fp} and 100 for the reference dataset D_{ref} to verify the effectiveness of our method. We use the Qwen2.5-VL-7B-Instruct (Bai et al., 2025) model to rephrase the caption T. During the fingerprint injection, we set the epoch to 1 and follow the experimental settings of Liu et al. (2025) to perform unlearning on LLaVA-1.5-7B using Gradient Ascent. When using the Min K% method, we set K to 20. In the harmlessness analysis, we evaluate the performance changes of the model after fingerprint injection using six commonly used LVLM benchmarks: ScienceQA (Lu et al., 2022), MM-Vet (Yu et al., 2023), SEED-Bench (Li et al., 2023a), LLaVA-Bench (Liu et al., 2023), POPE (Li et al., 2023b), and MMMU (Yue et al., 2023). Additional details can be found in Appendix §D and §E.

4.2 Main Results

To simulate the process of attackers attempting to erase model fingerprint embedding, we conduct LoRA fine-tuning and full fine-tuning experiments

Table 1: A comprehensive comparison of our proposed method UMMF with established baseline methods on the copyright tracking performance of fine-tuned models across 6 datasets. The evaluation metric is the FSR. The best results are highlighted in bold.

Method	V7W	ST-VQA	TextVQA	PaintingF	MathV	ChEBI	Average
LoRA Fine-tuning							
Ordinary (Wang et al., 2025b)	5%	3%	3%	2%	1%	3%	3%
IF (Xu et al., 2024)	28%	22%	30%	8%	24%	14%	21%
RNA (Wang et al., 2025b)	39%	46%	23%	12%	2%	11%	22%
PLA (Wang et al., 2025b)	53%	64%	46%	64%	40%	63%	55%
UMMF	**87%**	**83%**	**76%**	**82%**	**65%**	**72%**	**77%**
Full Fine-tuning							
Ordinary (Wang et al., 2025b)	2%	1%	4%	2%	0%	2%	2%
IF (Xu et al., 2024)	18%	12%	18%	0%	20%	0%	11%
RNA (Wang et al., 2025b)	26%	16%	16%	19%	15%	7%	16%
PLA (Wang et al., 2025b)	49%	58%	49%	63%	36%	56%	52%
UMMF	**84%**	**78%**	**85%**	**87%**	**73%**	**81%**	**82%**

on the fingerprint-embedded model using VQA training datasets from multiple domains, and then perform copyright tracking using UMMF and baseline methods. The experimental results are shown in Table 1. The IF method is a backdoor-based method, and its performance is not very stable. Under certain training settings, its fingerprint is completely erased, indicating that this type of method lacks robustness in the LVLM environment. This may be due to the differences in architecture and task patterns between LVLMs and LLMs. Ordinary, RNA, and PLA are all fingerprinting methods based on adversarial image attacks. Among them, Ordinary and RNA do not require model training during the construction of adversarial images, while PLA requires further model training. Even PLA, which performs second best, still underperforms our proposed UMMF by an average of 26%, indicating that methods based on adversarial image attacks have limited robustness after the model undergoes fine-tuning. In contrast, UMMF relies on detectable probability changes brought about by regions of poor generalization within specific sample neighborhoods as model fingerprints, and further reduces the impact of fingerprint sample characteristics and downstream fine-tuning on memorization scores through difficulty calibration, thereby enhancing the fingerprint's resistance to fine-tuning. To further evaluate the generalization capabilities of our approach, we also include additional experiments on other LVLM architectures in the Appendix §F.

Table 2: FSR under Taylor expansion-based pruning of 10% weights for fully fine-tuned models.

Method	Original	Model Pruning		
		Attention	MLP	Both
Ordinary	0.02	0.01	0.00	0.00
IF	0.18	0.03	0.00	0.01
RNA	0.26	0.07	0.08	0.06
PLA	0.49	0.40	0.38	0.35
UMMF	0.84	0.75	0.77	0.70

4.3 Harmlessness Analysis

To evaluate the impact of UMMF on model performance, we follow the harmlessness evaluation of the IF (Xu et al., 2024) and conduct experiments comparing the original model with the model after fingerprint injection through full fine-tuning on 6 commonly used LVLM benchmarks. As shown in Figure 3, after embedding 100 fingerprint on the original model, UMMF's impact on model performance is within 3%. This is mainly attributed to the fact that we only train for 1 epoch during both the learning and unlearning processes, and the training data we use are ordinary VQA datasets that do not contain rare or high-perplexity tokens.

4.4 Robustness Analysis

In real-world scenarios, malicious users may also attempt to erase fingerprints through model pruning or input image transformations. We perform experiments on the V7W fine-tuned model to evaluate UMMF's robustness against these strategies.

We conduct experiments on input image transformations by adding uniform noise of different magnitudes and mean blur with different kernel sizes to observe the FSR changes of the fingerprint.

Table 3: Comparison of UMMF and its variants on the V7W, ST-VQA, and TextVQA datasets. The evaluation metric is FSR.

Method	V7W	ST-VQA	TextVQA
LoRA Fine-tuning			
UMMF	**87%**	**83%**	**76%**
w/o unlearn	72%	70%	64%
w/o learn	83%	77%	71%
w/o Calibration	81%	75%	72%
Full Fine-tuning			
UMMF	**84%**	**78%**	**85%**
w/o unlearn	75%	66%	67%
w/o learn	80%	69%	81%
w/o Calibration	78%	73%	83%

As shown in Figures 4 and 5, it can be observed that UMMF has strong robustness to input image transformations in both full fine-tuning and LoRA fine-tuning. UMMF's FSR decreases by only 13% at maximum noise intensity and still surpasses the best baseline model.

We also test model pruning. We performed unstructured weight pruning based on Taylor expansion on the fully fine-tuned model, and removed the smallest 10% of the weights (Cheng et al., 2024). As shown in Table 2, UMMF has good adaptability to pruning, while the performance of IF and RNA both decline substantially, indicating that this type of explicit fingerprint relying on triggers has insufficient reliability after model pruning. Additionally, pruning attention layer weights has a greater impact on fingerprints than pruning MLP layer weights.

In addition, we explore the robustness of UMMF in the face of model merging in Appendix §G.

4.5 Ablation Study

We further explore several strategies that may affect UMMF's performance on three commonly used VQA datasets: V7W, ST-VQA, and TextVQA. We first compare UMMF with the following variants: **w/o unlearn:** No unlearning step is performed during the fingerprint injection process. **w/o learn:** No learning step is performed during the fingerprint injection process. **w/o Calibration:** No difficulty calibration is performed during fingerprint verification, directly using MSD(x) as the memorization score. The experimental results are shown in Table 3. It can be seen that unlearning has the greatest impact on UMMF's performance, indicating that the regions of poor generalization constructed by unlearning neighboring samples are key to fingerprint detection. The performance of w/o Calibration also declines substantially because different samples inherently have different difficulties, and the model has different initial memorization scores for them, which reflects the importance of difficulty calibration during the membership inference attack.

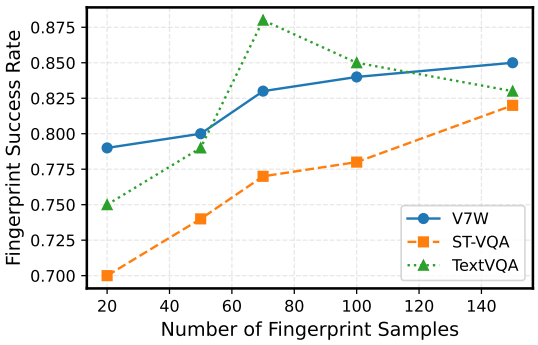

Figure 6: Effect of sample size on UMMF performance. The evaluation metric is FSR.

We also explore UMMF's sensitivity to fingerprint sample size on models with full fine-tuning. As shown in Figure 6, performance declines when the sample number is very low, mainly because UMMF requires sufficient samples to determine an appropriate threshold γ^* rather than relying on a fixed γ. When the sample number exceeds 50, UMMF's performance fluctuates only slightly.

5 Conclusion

This paper focuses on the important yet underexplored field of copyright tracking for LVLMs. We propose a novel method, UMMF, which artificially constructs regions of poor generalization in the data manifold by performing learning and unlearning operations on the caption and neighboring caption of the same image respectively, and uses the detectable probability changes resulting from this overfitting characteristic as the model's fingerprint. Compared with conventional methods, UMMF abandons the explicit fingerprint paradigm that relies on fixed input-output pairs, and instead uses memorization scores extracted through difficulty-calibrated membership inference attacks as implicit model fingerprints, fundamentally solving the trade-off problem between stealthiness and robustness. Extensive experiments demonstrate that this method exhibits excellent performance in diverse scenarios, highlighting its practicality and adaptability in real-world intellectual property

protection for LVLMs.

6 Limitations

The field of LVLM fingerprinting technology is still in its early exploratory stages, and currently, there is a lack of mature public datasets and benchmark resources widely available for use. Since fingerprint-related experiments often require fine-tuning of large models, which demands substantial computational resources, recent studies have had to limit their experiments to a single model. Following the experimental settings proposed by Wang et al. (2025b), this study has primarily selected the LLaVA model for extensive evaluation.

To further assess the generalization capabilities of UMMF, we have supplemented our experiments by conducting evaluations on two additional LVLM architectures, IDEFICS2-8B and MiniGPT-4, using LoRA fine-tuning. These results are detailed in the Appendix §F. In the future, we plan to expand our experiments to include a broader range of open-source models.

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

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

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.

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.

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.

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

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*.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. Membership inference attacks on machine learning: A survey. *Preprint*, arXiv:2103.07853.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Dmitri Iourovitski, Sanat Sharma, and Rakshak Talwar. 2024. Hide and seek: Fingerprinting large language models with evolutionary learning. *Preprint*, arXiv:2408.02871.

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*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. 2025. Single image unlearning: Efficient machine unlearning in multimodal large language models. *Preprint*, arXiv:2405.12523.

Qiao Li, Jing Chen, Kun He, Zijun Zhang, Ruiying Du, Jisi She, and Xinxin Wang. 2024b. Model-agnostic adversarial example detection via high-frequency amplification. *Computers & Security*, 141:103791.

Shen Li, Liuyi Yao, Jinyang Gao, Lan Zhang, and Yaliang Li. 2024c. Double-i watermark: Protecting model copyright for LLM fine-tuning. *CoRR*, abs/2402.14883.

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

Yuqing Liang, Jiancheng Xiao, Wensheng Gan, and Philip S. Yu. 2024. Watermarking techniques for large language models: A survey. *Preprint*, arXiv:2409.00089.

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

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, Kush R. Varshney, Mohit Bansal, Sanmi Koyejo, and Yang Liu. 2024. Rethinking machine unlearning for large language models. *Preprint*, arXiv:2402.08787.

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. 2025. Protecting privacy in multimodal large language models with mllmu-bench. *Preprint*, arXiv:2410.22108.

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.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343, Toronto, Canada. Association for Computational Linguistics.

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.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024a. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024b. 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.

Toan Tran, Ruixuan Liu, and Li Xiong. 2025. Tokens for learning, tokens for unlearning: Mitigating membership inference attacks in large language models via dual-purpose training. *Preprint*, arXiv:2502.19726.

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, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. 2024. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *Preprint*, arXiv:2408.01319.

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

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

Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. 2022. On the importance of difficulty calibration in membership inference attacks. In *International Conference on Learning Representations*.

Hengyu Wu and Yang Cao. 2025. Membership inference attacks on large-scale models: A survey. *Preprint*, arXiv:2503.19338.

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, 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.

Zhenhua Xu, Xixiang Zhao, Xubin Yue, Shengwei Tian, Changting Lin, and Meng Han. 2025c. 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.

Zhiguang Yang and Hanzhou Wu. 2024. A fingerprint for large language models. *Preprint*, arXiv:2407.01235.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

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, et al. 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. 2025. 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.

Jiale Zhang, Haoxuan Li, Di Wu, Xiaobing Sun, Qinghua Lu, and Guodong Long. 2025a. Beyond dataset watermarking: Model-level copyright protection for code summarization models. In *THE WEB CONFERENCE 2025*.

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. 2025b. Meraser: An effective fingerprint erasure approach for large language models. *Preprint*, arXiv:2506.12551.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025c. Min-k%++: Improved baseline for pre-training data detection from large language models. In *The Thirteenth International Conference on Learning Representations*.

Xianren Zhang, Hui Liu, Delvin Ce Zhang, Xianfeng Tang, Qi He, Dongwon Lee, and Suhang Wang. 2025d. Does multimodal large language model truly unlearn? stealthy mllm unlearning attack. *Preprint*, arXiv:2506.17265.

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.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv: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 Details of the Original LVLM

In our study, we adopt LLaVA-1.5-7B (Liu et al., 2023) as the base model, a widely used open-source framework in the field of large vision-language models. Although not the most advanced in terms of performance, LLaVA-1.5-7B has become a popular and reliable choice for research and application due to its transparency, accessibility, and solid multimodal capabilities. The model combines a pre-trained vision encoder, CLIP ViT-L/14, with a large language model decoder, LLaMA-2, forming a modular architecture for vision-language understanding. The vision encoder extracts high-quality image representations at a resolution of 336×336, while a two-layer linear projector maps visual features into the textual embedding space. The LLaMA-2 decoder, consisting of 32 transformer layers with a hidden size of 4096, is responsible for language comprehension and generation. This combination enables the model to effectively process multimodal inputs and perform a variety of vision-language tasks, making it a practical baseline for further exploration and enhancement.

B 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.

B.1 Fine-tuning Configurations

In our experiments, we use the AdamW optimizer with a cosine learning rate schedule. For full fine-tuning, the learning rate is set to 5e-5, while for LoRA fine-tuning, it is set to 2e-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 LVLMs. 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.

B.2 Fine-tuning Dataset Overviews

We employed multiple downstream VQA datasets to simulate various deployment scenarios of copyright-protected LVLMs. 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.

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., 2024b). 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.

ChEBI-20. ChEBI-20 is a niche dataset specifically created for molecular image-based question answering tasks (Edwards et al., 2021). It contains 33,010 molecule-description pairs, offering a unique application area for models focused on chemistry and molecular biology.

C Details of Baseline Methods

C.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.

C.2 Ordinary

The Ordinary baseline (Wang et al., 2025b) 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. However, this method often struggles to maintain robustness against modifications such as fine-tuning.

C.3 RNA (Random Noise Attack)

Random Noise Attack (RNA) (Wang et al., 2025b) 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. 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.

C.4 PLA (Parameter Learning Attack)

Parameter Learning Attack (PLA) (Wang et al., 2025b) simulates the process of fine-tuning during fingerprint generation by allowing updates to both the trigger image and the model parameters. By mimicking the behavior of fine-tuned models, PLA generates triggers that are resistant to parameter shifts and fine-tuning, thereby improving the robustness of fingerprint tracking. Its dynamic nature makes it more effective than static adversarial approaches.

D Additional Implementation Details

D.1 Flickr30k Dataset

Flickr30K (Young et al., 2014) is a widely utilized multimodal dataset designed for benchmarking vision-language models. It consists of 31,783 images depicting everyday life activities, events, and scenes, all sourced from the online platform Flickr. Additionally, the dataset contains 158,915 captions created through crowdsourcing, with each image being independently described by five annotators. These annotators were not provided with specific context regarding the depicted scenes, ensuring unbiased textual descriptions.

For our experiments, we sampled a total of 200 multimodal samples from the Flickr30k dataset. These samples were divided into two equal subsets: the fingerprint dataset (D_{fp}), containing 100 samples used for injecting fingerprints, and the reference dataset (D_{ref}), containing another 100 samples used to determine dynamic thresholds for the verification process, eliminating the need to rely on fixed threshold values. The inclusion of D_{ref} allows for more adaptive and robust fingerprint validation, as the dynamic thresholds are calibrated based on the specific characteristics of the reference samples. The high-quality and diverse annotations in the Flickr30k dataset make it well-suited for assessing the effectiveness of the fingerprinting techniques proposed in this work.

D.2 Paraphrase Generation

We use the Qwen2.5-VL-7B-Instruct model (Bai et al., 2025) to generate paraphrased captions based on the original captions. This is achieved by using prompts specifically designed to maintain the semantic integrity of the original caption while introducing sufficient variation to enable comparing the model's memorization strengths. The prompt used for paraphrasing takes the form: *"Please rewrite the following sentence with the same meaning but different wording. You may use techniques such as replacing words with synonyms, changing sen-*

tence structures, or switching between active and passive voice". This ensures that semantically similar neighbors for the target samples are generated, forming the basis for the learn and unlearn process.

D.3 Learn and Unlearn Parameter Settings

During the fingerprint injection process, we use the AdamW optimizer with a learning rate of 2×10^{-5}, the bfloat16 numerical format, and a batch size of 4. Both the learn and unlearn processes are performed for only one epoch. For fine-tuning the model, we employ full fine-tuning, allowing updates to all learnable parameters in the model. Limiting the learn and unlearn processes to a single epoch is a deliberate choice to minimize the impact of fingerprint injection on the original model's performance while still enabling the effective construction of robust and detectable fingerprints.

E Benchmarks for Harmlessness Analysis

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

ScienceQA: The ScienceQA (Lu et al., 2022) benchmark 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.

MM-Vet: MM-Vet (Yu et al., 2023) is a comprehensive benchmark that examines six core vision-language (VL) 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., 2023a) 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.

POPE: The POPE (Li et al., 2023b) benchmark is designed to analyze and detect object hallucination issues in vision-language models. It features a binary classification query dataset with 8,910 entries, divided into random, popular, and adversarial subsets. Each subset is constructed through distinct sampling approaches to create a robust dataset for assessing hallucination-related robustness.

MMMU: The MMMU (Yue et al., 2023) dataset is a large cross-disciplinary multimodal benchmark, comprising 11.5K questions across six major disciplines, 30 specific topics, and 183 subdomains. It includes both multiple-choice and open-ended questions covering 30 image types, including charts, tables, chemical structures, artistic works, and more. The dataset is designed to test a model's ability to comprehend complex visual information and engage in specialized reasoning tasks.

F Additional Experiments on Other LVLM Architectures

To further evaluate the generalization capabilities of UMMF, we extend our experiments to two additional LVLM architectures: IDEFICS2-8B (Laurençon et al., 2024) and MiniGPT-4 (Zhu et al., 2023). These two models are selected because they are entirely open-source, including both training data and training details, making them suitable for reproducibility and comparison. 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 UMMF framework, and copyright tracking performance is evaluated using the Fingerprint Success Rate (FSR) metric across six VQA datasets from different domains. The results are presented in Tables 4 and 5, respectively.

The experimental results demonstrate that

Table 4: Experimental results on the IDEFICS2-8B model. 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	6%	5%	4%	3%	2%	4%	4%
IF	24%	20%	27%	6%	26%	23%	21%
RNA	35%	40%	21%	8%	13%	9%	21%
PLA	52%	59%	44%	60%	39%	61%	53%
UMMF	**84%**	**80%**	**75%**	**78%**	**69%**	**74%**	**77%**

Table 5: Experimental results on the MiniGPT-4 model. 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	7%	4%	5%	2%	1%	3%	4%
IF	25%	19%	29%	9%	23%	7%	19%
RNA	37%	43%	25%	11%	4%	10%	22%
PLA	55%	62%	46%	63%	42%	58%	54%
UMMF	**84%**	**78%**	**82%**	**83%**	**68%**	**78%**	**79%**

UMMF consistently outperforms baseline methods across both IDEFICS2-8B and MiniGPT-4. Despite the differences in model architecture and training dynamics, the robustness and generalizability of UMMF are evident, as it achieves an average FSR of 77% and 79% on IDEFICS2-8B and MiniGPT-4, respectively, under LoRA fine-tuning. In contrast, the best-performing baseline method (PLA) achieves an average FSR of 53% and 54% on IDEFICS2-8B and MiniGPT-4, respectively. These results further validate that UMMF's implicit fingerprinting mechanism can be effectively adapted to diverse LVLM architectures.

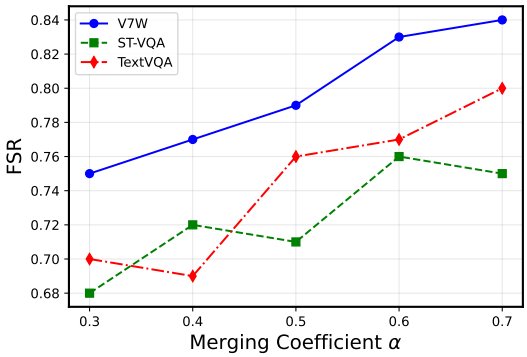

Figure 7: Fingerprint Success Rate of pseudo-merged models.

G Robustness to Model Merging

We further evaluate the robustness of UMMF against model merging, where malicious users aim to obscure the ownership of the fingerprinted model by merging it with another benign model. To simulate such scenarios, we employ a *pseudo-merged model* approach inspired by Yamabe et al. (2025). In this setup, we merge a base model without fingerprint embedding and a fingerprinted model that has been fine-tuned on a specific dataset. We select V7W, ST-VQA, and TextVQA as the fine-tuning datasets and use task arithmetic (Ilharco et al., 2023) as the merging method. The parameters of the pseudo-merged model are defined as:

$$\theta_{\text{merge}} = \theta_{\text{base}} + \alpha(\theta_{\text{fingerprint}} - \theta_{\text{base}}), \quad (7)$$

where θ_{merge} represents the parameters of the pseudo-merged model, θ_{base} is the base model, $\theta_{\text{fingerprint}}$ represents the parameters of the fingerprinted model fine-tuned on a specific dataset, and α is the merging coefficient.

The results of the experiment are shown in Figure 7. We observe robust fingerprint detection across all datasets, even after merging with the benign base model. This highlights the resilience of UMMF fingerprints under merging operations.

Table 6: Experimental results on the nanoVLM-222M model.

Method	V7W	ST-VQA	TextVQA	PaintingF	MathV	ChEBI	Average
Full Fine-tuning							
Ordinary	4%	3%	0%	2%	1%	0%	2%
IF	15%	10%	9%	2%	13%	14%	11%
RNA	12%	25%	23%	27%	19%	11%	20%
PLA	38%	54%	46%	60%	41%	63%	50%
UMMF	**78%**	**76%**	**76%**	**81%**	**65%**	**77%**	**76%**

Table 7: FSR comparison between UMMF and Fingerprint Dilution Attack under LoRA fine-tuning.

Method	V7W	ST-VQA	TextVQA	PaintingF	MathV	ChEBI
UMMF	87%	83%	76%	82%	65%	72%
Fingerprint Dilution Attack	82%	77%	68%	77%	63%	71%

The robustness can be attributed to UMMF's ability to embed implicit memorization patterns directly into the model's representation space, rather than relying on explicit signatures. This embedding ensures that essential fingerprint signals are preserved even in the presence of operations that aim to dilute or obscure them, such as parameter averaging in the merging process.

H Additional Experiments on nanoVLM-222M

To further test the performance of UMMF on different model architectures, we conducted additional experiments on the recently released fully open-source model nanoVLM-222M. The experiments used full fine-tuning settings. The results are shown in Table 6.

The experimental results demonstrate that UMMF still outperforms all baseline methods on the nanoVLM-222M model, achieving an average FSR of 76%. This further validates the effectiveness and generalization capability of our method across different model architectures.

I Robustness Against Fingerprint Dilution Attack

Attackers may attempt to dilute the initially embedded fingerprints by adding more fingerprints. To evaluate UMMF's robustness against such attacks, we conducted additional experiments under the LoRA fine-tuning setting. After embedding 100 fingerprints, we imitated the attacker by embedding an additional 5,000 fingerprints. The results are shown in Table 7.

UMMF maintains strong detection capability even after the dilution attack, with FSR decreasing by at most 8 percentage points. In real-world environments, adding excessive fingerprints would significantly degrade the model's multimodal understanding ability, making this attack strategy impractical for malicious users. This further demonstrates the robustness of our method.

Note that the fundamental assumption for UMMF's effectiveness is that attackers do not know which images I we selected for fingerprint embedding. Model developers can easily obtain unique private images through photography, and as long as these images remain confidential, our method maintains high robustness.

J Ethics Statement

Our primary goal in developing the UMMF method is to provide a robust and stealthy mechanism for protecting the intellectual property (IP) of Large Vision-Language Models. We aim to strengthen ownership verification to prevent unauthorized use, redistribution, and commercialization of these high-value AI assets. By doing so, we hope to contribute to the fair use and sustainable development of AI technologies.

We acknowledge that techniques such as UMMF, which manipulate model behaviors to embed identifiable fingerprints, could potentially be misused if applied maliciously. For instance, such methods could be exploited to falsely attribute ownership or target specific models for adversarial purposes (Zhang et al., 2025b). To mitigate these risks, our research is designed with a focus on ethical use cases, such as safeguarding against IP infringement

and allowing developers to assert rightful ownership. We strongly emphasize that these techniques be used only in compliance with applicable laws, regulations, and ethical guidelines.

All experiments presented in this work are conducted on publicly available datasets, without the use of proprietary or private data. No human subjects were involved in the collection, annotation, or evaluation processes, and no personally identifiable information (PII) is included in the training or evaluation datasets. We encourage future researchers and practitioners to build upon our work responsibly, ensuring that corresponding methods are utilized for enhancing the security and legitimacy of AI systems, rather than for unethical or harmful purposes. Safeguards, such as peer review and compliance with anti-abuse principles, should always accompany any deployment of such technologies.