
LOMIA: Label-Only Membership Inference Attacks against Pre-trained Large Vision-Language Models

Yihao Liu, Xinqi Lyu, Dong Wang, Yanjie Li, Bin Xiao*

Department of Computing, The Hong Kong Polytechnic University
{yihao5.liu,xinqi.lyu,dong-comp.wang,yanjie.li}@connect.polyu.hk,
b.xiao@polyu.edu.hk

Abstract

Large vision-language models (VLLMs) have driven significant progress in multi-modal systems, enabling a wide range of applications across domains such as healthcare, education, and content generation. Despite the success, the large-scale datasets used to train these models often contain sensitive or personally identifiable information, raising serious privacy concerns. To audit and better understand such risks, membership inference attacks (MIAs) have become a key tool. However, existing MIAs against VLLMs predominantly assume access to full-model logits, which are typically unavailable in many practical deployments. To facilitate MIAs in a more realistic and restrictive setting, we propose a novel framework: label-only membership inference attacks (LOMIA) targeting pre-trained VLLMs where only the model’s top-1 prediction is available. Within this framework, we propose three effective attack methods, all of which exploit the intuition that training samples are more likely to be memorized by the VLLMs, resulting in outputs that exhibit higher semantic alignment and lower perplexity. Our experiments show that our framework surpasses existing label-only attack adaptations for different VLLMs and competes with state-of-the-art logits-based attacks across all metrics on three widely used open-source VLLMs and GPT-4o.

1 Introduction

Large vision-language models (VLLMs) [31, 56, 55, 36] have proven to be formidable tools capable of understanding and generating multimodal content, leading to substantial advancements in areas such as image captioning [25, 8] and visual question answering [42, 13, 28]. These models are typically trained on vast datasets that integrate images and textual descriptions, often sourced from the Internet. While this extensive training enables impressive performance, it also raises substantial privacy issues, especially in relation to the inadvertent inclusion of sensitive or proprietary information in the training data [32]. For example, identity-related details, such as names linked to faces can be revealed, even when both alignment and fine-tuning are performed using anonymized datasets [3]. In the medical domain, VLLMs trained on datasets comprising medical images and associated diagnoses may unintentionally disclose private information of patients [22, 33]. Moreover, numerous studies have demonstrated that large language models (LLMs) are capable of memorizing substantial portions of training data prior to overfitting and exhibit a lower tendency to forget during training [48, 4, 50]. Given that VLLMs are built upon LLMs, it is reasonable to expect that they inherit this memorization behavior as well.

One of the most common methods used to tackle these risks is membership inference attack (MIA), which tries to distinguish whether a particular sample has been utilized in training model [45]. However, prevailing methods largely presume access to the full set of model logits, enabling the

*Bin Xiao is the corresponding author.

estimation of token-level probabilities. These probabilities facilitate the computation of critical statistical measures such as log-likelihood, entropy, and perplexity, which have been shown to serve as strong indicators for inferring membership status. Given this constraint, a key challenge emerges: How can we effectively infer membership when only the final output of the target model is available?

To address this problem, we propose a novel framework named LOMIA for label-only MIAs on VLLMs, which is a more realistic and challenging scenario. For any target data, LOMIA contains two stages: regression stage and inference stage. The former is trying to query the surrogate model to fit the relationship between related features and the perplexity (PPL) of generated descriptions. The latter utilizes the regression model to predict the corresponding PPL value and detect the membership by threshold.

Overall, our key contributions can be summarized as follows.

- To the best of our knowledge, we are the first to explore whether pre-trained VLLMs suffer from membership inference attacks under the label-only setting.
- We propose three straightforward yet impactful label-only MIAs targeting the pre-training phase of VLLMs. These attacks are based on text-text features, image-text features, and dual features.
- Extensive evaluations conducted on two datasets and three open-source pre-trained VLLMs demonstrate that LOMIA performs comparably to existing logits-based attacks across a range of evaluation metrics. We also show the effectiveness of our methods on the closed-source model GPT-4o, which achieved an AUC of 0.669 when evaluating the image-text feature attack method (ITFA).

2 Related Work

Pre-trained Large Vision-Language Models. Building on the success of LLMs, VLLMs integrate visual perception with language generation. Pioneering works like CLIP [39] established cross-modal alignment between images and text, laying the groundwork for deeper integration. Recent VLLMs such as LLaVA [30], MiniGPT-4 [56] and LLaMA-Adapter [12] leverage pre-trained vision encoders (e.g., CLIP-ViT [39]) and LLMs (e.g., LLaMA [49], Vicuna [7]) to fuse visual and textual tokens via lightweight projection layers. This approach allows the visual and textual components to work together effectively without retraining the entire system. With additional instruction tuning [30] on aligned image-text data, these models now perform well on a variety of multi-modal tasks, such as image captioning [25], and have become a strong foundation for many downstream applications. However, their reliance on large-scale training data raises concerns regarding memorization and unintended information leakage.

Membership Inference Attack (MIA). MIAs seek to determine whether a particular data sample was included in the training set of a given machine-learning model [45]. These attacks have received increasing attention due to their significant implications for data privacy, particularly when training datasets contain personally identifiable information. These attacks exploit the tendency of machine learning models to behave differently on training data compared to unseen data, primarily as a consequence of overfitting [16]. MIAs are generally categorized into three types: white-box attacks, which assume access to internal model parameters [34, 45]; grey-box attacks, which require access to model logits [27]; and black-box attacks, which rely solely on model outputs [15, 23, 45]. Black-box MIAs include simple metric-based approaches leveraging statistical signals, as well as more sophisticated techniques based on shadow models [5, 2, 23], which attempt to replicate the behavior of the target model. Extensive research has explored various MIA strategies across different machine-learning models, with notable progress in applying MIAs to classification models [45, 23], generative models [52, 9], regression models [14], and embedding models [47, 10]. Recent efforts have extended these techniques to LLMs [43, 15, 4, 51, 11]. However, relatively few studies have focused on MIA risks in multi-modal models.

Early work focuses on multi-modal classification models that combine CNN encoder (e.g., ResNet-152 [35] or VGG-16 [38]) with LSTM decoder [17]. This approach [17] assumes access to output confidences and the feasibility of constructing shadow models. Subsequent efforts extend to MIAs against CLIP [19]. In contrast, this work [19] proposes a metric-based black-box MIA using cosine

similarity and model output confidences to infer membership status. A critical limitation of prior investigations lies in their focus on smaller-scale models, which tend to overfit due to limited representational capacity. More recent studies begin to evaluate attacks on VLLMs but typically rely on access to full-model logits [27]. Such grey-box approaches compute statistical metrics like MaxRényi-K% or ModRényi* as membership signals. Nevertheless, access to token-level probabilities is rarely granted in deployed systems. To better align with real-world scenarios, Hu et al. [18] proposed a black-box membership inference attack method against fine-tuned VLLMs. However, their approach depends on shadow models and assumes access to a large amount of data highly similar to the training set, an assumption rarely met in practical deployments. To the best of our knowledge, our work is the first to explore effective label-only MIAs against pre-trained VLLMs without relying on shadow training or access to logits of the target model.

3 Threat model

Membership inference attack represents a class of privacy breaches aimed at machine learning models, especially those trained with sensitive information [45]. Formally, considering a trained VLLM M_θ , a data sample x , and external adversary knowledge denoted by K , the definition of MIA A is as follows:

$$A : x, M_\theta, K \rightarrow \{0, 1\}. \quad (1)$$

Here, 0 denotes that x is not included in the training dataset of M_θ , while 1 indicates that it is.

Goal of adversary. During the training process of a target model, data samples used for training are referred to as members, while those excluded from training are called non-members. The adversary aims to determine whether a given data sample is a member or a non-member of the training dataset.

Adversary’s Capabilities. Given a target VLLM $\mathcal{M}_{\text{tar}}$, the adversary can only query the target model and observe its outputs, without access to its internal structure or parameters. To make the scene more realistic, the adversary is not permitted to train a shadow model or utilize any shadow dataset. Even if permitted, the computational cost of training shadow models at the scale of pre-trained VLLMs remains prohibitively high for most adversaries.

Metrics. We assess our method using three key metrics: area under the curve (AUC) of the receiver operating characteristic (ROC) curve, balanced accuracy, and TPR@1%FPR (true positive rate at 1% false positive rate). AUC serves as an average measure across all false positive rates. Balanced accuracy evaluates attack effectiveness by measuring prediction accuracy in a balanced dataset of members and non-members. TPR@1%FPR provides a practical metric widely adopted in prior research [5, 19].

4 Method

In this section, we present LOMIA, a label-only membership inference attack framework tailored for VLLMs. LOMIA comprises three complementary attack variants, each leveraging a distinct feature extraction strategy to reveal training data leakage: (1) text-text feature attack (TTFA), (2) image-text feature attack (ITFA), and (3) dual-feature attack (DUFA). All methods proposed follow a two-stage procedure consisting of regression and inference, as illustrated in Figure 1.

4.1 Text-Text Feature Attack (TTFA)

The core concept of TTFA is based on the VLLMs’ ability to remember information from past training datasets and deliver similar responses. Consequently, an intuitive approach is to directly query the VLLMs and leverage semantic similarity between model outputs and ground truth texts to infer the membership status of training samples. However, the signal strength alone is insufficient due to potential noise in generation. To address this limitation, we incorporate the PPL of the generated description as a corrective signal, capturing confidence in generation. The joint modeling via linear regression offers a simple yet effective way to quantify the relationship between semantic similarity and generation fluency.

Regression Stage. In the regression stage, we establish a relationship between text similarity and PPL using a surrogate model $\mathcal{M}_{\text{sur}}$. For a target image $x_v^{(i)}$ and its corresponding ground-truth $x_t^{(i)}$

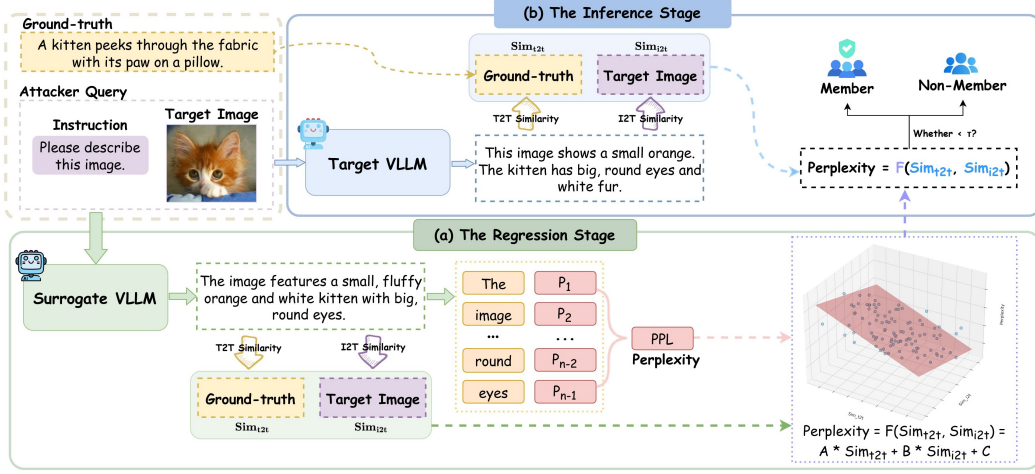


Figure 1: **Framework of LOMIA.** (a) **The Regression Stage:** We use the target samples with instructions to query the surrogate model to fit the relationship between related features and the perplexity of generated descriptions. (b) **The Inference Stage:** We first query the target model with the same target samples and instructions, then apply the regression model to predict the corresponding PPL value and detect the membership by threshold.

from the dataset $\mathcal{D}$, we use a surrogate model to generate description $\hat{x}_t^{(i)}$, and calculate the semantic similarity between generated description and ground truth $S_{t2t} = \text{sim}(\hat{x}_t^{(i)}, x_t^{(i)})$, where $\text{sim}(\cdot, \cdot)$ represents a semantic similarity function computed using a sentence embedding model. Given the generated description sequence $\hat{x}_t^{(i)} = w_1, \dots, w_n$, we calculate the PPL of the i -th generated description:

$$PPL(\hat{x}_t^{(i)}) = \exp \left(-\frac{1}{N} \sum_{n=1}^N \log P(w_n | w_1, \dots, w_{n-1}) \right). \quad (2)$$

where $\log P(w_n | w_1, \dots, w_{n-1})$ represents the logarithm of the conditional probability of the n -th word given the previous words. Finally, we establish a linear relationship between S_{t2t} and PPL to model their joint effect on membership inference signals, which can be formulated as follows:

$$R : S_{t2t} \mapsto PPL, \text{ via } : \min_{k,b} \sum \|PPL^{(i)} - kS_{t2t}^{(i)} - b\|^2. \quad (3)$$

Inference Stage. During the inference stage, we apply the regression model to predict the membership score of target samples using the target model $\mathcal{M}_{\text{tar}}$. For each test sample $(x_v^{(i)}, x_t^{(i)})$, we use the target model to generate a description $\hat{x}'_t^{(i)}$ and calculate the semantic similarity between generated description and ground truth $S'_{t2t} = \text{sim}(\hat{x}'_t^{(i)}, x_t^{(i)})$. Then we apply the regression model to predict membership score and classify the sample as a member if the score exceeds a threshold τ , which can be defined by:

$$\text{Score}_{t2t} = -\mathcal{R}(S'_{t2t}) = -PPL_{\text{predicted}}. \quad (4)$$

$$A(x) = \begin{cases} \text{Member,} & \text{if } \text{Score}_{t2t} \geq \tau, \\ \text{Non-member,} & \text{otherwise.} \end{cases} \quad (5)$$

4.2 Image-Text Feature Attack (ITFA)

ITFA is based on the hypothesis that training samples induce tighter cross-modal alignment within the learned joint embedding space. Specifically, VLLMs trained on image-text pairs tend to map member images and their generated descriptions to highly overlapping regions in the CLIP embedding space, resulting in elevated cosine similarity.

Regression Stage. In this stage, we establish a relationship between image-text alignment and membership signals. After querying the surrogate model $\mathcal{M}_{\text{sur}}$ to obtain the description $\hat{x}_t^{(i)}$, we utilize CLIP to calculate the cosine similarity between generated text and target image $S_{i2t} = \text{sim}(\hat{x}_t^{(i)}, x_v^{(i)})$. Next, we calculate the PPL of the generated descriptions in the same manner as TTFA and then establish a linear relationship between the image-text similarity and PPL:

$$R : S_{i2t} \mapsto PPL, \text{ via } : \min_{k,b} \sum \|PPL^{(i)} - kS_{i2t}^{(i)} - b\|^2. \quad (6)$$

Inference Stage. The inference stage of ITFA uses the cosine similarity between text generated from the target model to predict membership score, which can be defined by:

$$\text{Score}_{i2t} = -\mathcal{R}(S'_{i2t}) = -PPL_{\text{predicted}}, \text{ where } S'_{i2t} = \text{sim}(\hat{x}'_t^{(i)}, x_v^{(i)}). \quad (7)$$

$$A(x) = \begin{cases} \text{Member,} & \text{if } \text{Score}_{i2t} \geq \tau, \\ \text{Non-member,} & \text{otherwise.} \end{cases} \quad (8)$$

4.3 Dual Feature Attack (DUFA)

While TTFA and ITFA individually capture complementary aspects of memorization: semantic preservation and cross-modal alignment, they may individually fail under different sources of noise. DUFA integrates dual features in a regression model, leveraging the joint patterns in VLLMs.

Regression Stage. In this stage, we learn a multivariate regression model to fit both text-text and image-text similarity to the PPL. This can be formulated as follows:

$$R : (S_{i2t}, S_{i2t}) \mapsto PPL, \text{ via } : \min_{k_1, k_2, b} \sum \|PPL - k_1 S_{i2t}^{(i)} - k_2 S_{i2t}^{(i)} - b\|^2. \quad (9)$$

Inference Stage. The inference stage of DUFA requires to use both text-text and image-text features to predict membership score, which can be defined by:

$$\begin{aligned} \text{Score}_{\text{dual}} &= -\mathcal{R}(S'_{i2t}, S'_{i2t}) = -PPL_{\text{predicted}}, \\ \text{where } S'_{i2t} &= \text{sim}(\hat{x}'_t^{(i)}, x_t^{(i)}), S'_{i2t} = \text{sim}(\hat{x}'_t^{(i)}, x_v^{(i)}). \end{aligned} \quad (10)$$

$$A(x) = \begin{cases} \text{Member,} & \text{if } \text{Score}_{\text{dual}} \geq \tau, \\ \text{Non-member,} & \text{otherwise.} \end{cases} \quad (11)$$

5 Experiments

In this section, we conduct MIAs across three target models using various baselines, and our own methods: TTFA, ITFA, and DUFA. The evaluation setup is detailed in Section 5.1. Results for TTFA, ITFA, and DUFA are presented in Section 5.2. Additionally, an ablation study is included in Section 5.3.

5.1 Evaluation Setup

Models. Our evaluation focuses on three popular open-source VLLMs: LLaVA-1.5 [30], MiniGPT-4 [56], and LLaMA-Adapter V2.1 [12], all providing full access to model weights, training process, and datasets. Here we test Vicuna-7B [7] for LLaVA and MiniGPT-4, and LLaMA-7B [49] for LLaMA-Adapter V2.1. We also test the effectiveness of our methods on closed-source model GPT-4o [1].

Datasets. Li et al. [27] created a dataset for MIAs on VLLMs, but it is unimodal and does not align with our task. Therefore, we developed a multi-modal dataset for MIAs against VLLMs:

- **LOMIA/LAION.** Pre-trained VLLMs such as LLaVA-1.5, MiniGPT-4, and LLaMA-Adapter V2 use images from the LAION [41], Conceptual Captions 3M [6], Conceptual 12M [6], and SBU Captions [37] datasets for pre-training [27]. Following Li et al. [27], we randomly sample a subset from the intersection of the datasets used by these three pre-trained VLLMs to serve as the member data. We then use the captions of the member data as input to query the stable-diffusion-v1-5 [40] to generate images that serve as non-member data. To ensure the validity of our MIA on VLLMs, we have 600 images in LOMIA/LAION (300 members and 300 non-members).
- **LOMIA/CC.** MS COCO [29] is also a popular dataset used in the pre-training process of the target models, so we randomly select some images in this dataset as member data. We use a similar approach to generate non-member data with stable-diffusion-v1-5 [40]. We also have 600 images in LOMIA/CC (300 members and 300 non-members).

Baselines. We compare our LOMIA with seven existing state-of-the-art attack methods. For logits-based baselines, we select the PPL attack [4, 54], MIN-K% PROB [44], Aug-KL [27], Max-Prob-Gap% PROB [27], MaxRényi-K% [27], and ModRényi* [27]. For label-only baselines, we mainly consider the Query Attack proposed by Hu et al. [18]. The details of the baselines can be found in the Appendix A.

Evaluation Settings. The attack implementation is conducted on 4 NVIDIA 3090 GPUs. To assess the performance of our LOMIA framework, we use LLaVA-1.5-7B as the surrogate model. To mitigate the impact of varying text lengths, we fix the maximum token length at 32. For computing text-text similarity, we use the all-MiniLM-L6-v2 model [21], which encodes each sentence into dense representations via mean pooling over the transformer outputs, followed by cosine similarity measurement. Image-text similarity is calculated using CLIP [39], which maps visual and textual inputs into a shared embedding space through dedicated encoders before computing their cosine similarity.

5.2 Main Results

We compare our methods with six advanced logits-based attack methods and one label-only attack method. Complete results on LOMIA/LAION and LOMIA/CC are summarized in Table 1, Table 2 and Table 3 respectively. Our extensive empirical evaluation has revealed several insightful observations about MIAs against pre-trained VLLMs, which we discuss in detail below.

LOMIA performs comparably to existing logits-based attacks. Table 1 and Table 2 show that our three attack methods based on the framework of LOMIA significantly outperforms other label-only attacks across all datasets and target model settings. Even when compared with existing logits-based attacks, our methods achieve similar or superior performance. Specifically, our TTFA method achieves the highest AUC scores among all label-only attacks on nearly all target models, consistently surpassing the baseline Query Attack by a substantial margin. For instance, TTFA improves over the Query Attack by 0.105 and 0.069 in AUC on LLaVA, as shown in Table 1 and Table 2, respectively. In terms of TPR@1%FPR, our methods remain competitive with strong logits-based baselines. For instance, TTFA achieves up to 3.0% on LLaMA Adapter shown in Table 2, clearly outperforming the Query Attack (0.0%). While top-performing logits-based methods like MaxRényi ($\alpha = 2$) still yield the highest TPR values (e.g., 5.6%–7.3%), LOMIA approaches offer a practical alternative under stricter threat models. Notably, ITFA demonstrates robust TPR gains across architectures, indicating its effectiveness even when model architectures diverge.

Comparison among TTFA, ITFA, and DUFA. DUFA performs best when the surrogate model shares the same architecture as the target model, since its design relies on feature-level alignment between the two (e.g., AUC = 0.621 on LOMIA/LAION and 0.630 on LOMIA/CC). However, when the surrogate and target architectures differ, DUFA’s advantage diminishes, and its performance can fall below that of single-feature variants like TTFA or ITFA. This indicates that different target models vary in their sensitivity to similarity-based features, likely due to differences in architecture or training that affect their vulnerability to text-based similarity attacks in controlled settings. Notably, when we expanded the dataset and used real images, as detailed in our Appendix C, DUFA consistently outperformed the other methods across all target models, while TTFA’s performance dropped significantly. This reversal demonstrates that DUFA offers greater robustness as experimental conditions become more realistic and challenging. By combining text-to-text and

image-to-text similarity features, DUFA is able to maintain strong performance across varying data scales and distributions.

Table 1: Complete results of various attacks on LOMIA/LAION.

Metrics	AUC $\uparrow$		Balanced Acc $\uparrow$		TPR@1%FPR $\uparrow$				
	LLaVA	MiniGPT4	LLaMA Adapter	LLaVA	MiniGPT4	LLaMA Adapter	LLaVA	MiniGPT4	LLaMA Adapter
Logits-based Attacks									
Perplexity	0.625	0.597	0.603	0.611	0.580	0.586	3.0%	2.6%	2.6%
Aug_KL	0.551	0.526	0.449	0.551	0.546	0.510	2.0%	1.0%	1.6%
Max-Prob-Gap	0.588	0.579	0.562	0.586	0.565	0.561	1.3%	2.6%	1.3%
Min0% Prob	0.676	0.559	0.549	0.641	0.551	0.548	3.3%	1.6%	0.3%
Min10% Prob	0.738	0.569	0.564	0.693	0.553	0.561	2.0%	2.0%	0.6%
Min20% Prob	0.661	0.581	0.573	0.633	0.573	0.568	1.6%	1.3%	0.6%
ModRényi($\alpha = 0.5$)	0.628	0.588	0.606	0.601	0.575	0.583	4.6%	1.6%	1.3%
ModRényi($\alpha = 2$)	0.629	0.587	0.606	0.601	0.573	0.581	2.6%	2.6%	1.6%
Max0%Rényi ($\alpha = 0.5$)	0.616	0.547	0.629	0.593	0.548	0.610	4.6%	3.0%	5.3%
Max10%Rényi ($\alpha = 0.5$)	0.659	0.559	0.534	0.620	0.556	0.538	4.6%	1.3%	2.0%
Max100%Rényi ($\alpha = 0.5$)	0.581	0.581	0.563	0.580	0.575	0.565	5.3%	1.0%	1.0%
Max0%Rényi ($\alpha = 1$)	0.679	0.560	0.618	0.636	0.560	0.596	9.6%	3.0%	4.3%
Max10%Rényi ($\alpha = 1$)	0.750	0.579	0.549	0.695	0.568	0.540	13.0%	3.3%	2.6%
Max100%Rényi ($\alpha = 1$)	0.625	0.605	0.589	0.596	0.588	0.573	7.3%	1.3%	1.6%
Max0%Rényi ($\alpha = 2$)	0.696	0.555	0.576	0.668	0.556	0.576	10.0%	3.3%	5.6%
Max10%Rényi ($\alpha = 2$)	0.775	0.577	0.565	0.718	0.560	0.566	3.6%	2.3%	1.3%
Max100%Rényi ($\alpha = 2$)	0.631	0.602	0.599	0.616	0.581	0.583	2.3%	1.6%	1.6%
Max0%Rényi ($\alpha = \infty$)	0.676	0.555	0.551	0.641	0.543	0.550	3.3%	1.6%	1.3%
Max10%Rényi ($\alpha = \infty$)	0.738	0.569	0.564	0.693	0.555	0.561	2.0%	2.0%	1.0%
Max100%Rényi ($\alpha = \infty$)	0.633	0.599	0.603	0.625	0.583	0.586	1.3%	1.3%	1.3%
Label-only Attacks									
Query Attack	0.516	0.545	0.492	0.538	0.555	0.520	0.3%	0.3%	0.3%
TTFA (Ours)	0.601	0.589	0.571	0.601	0.596	0.566	3.0%	0.6%	2.6%
ITFA (Ours)	0.617	0.571	0.522	0.591	0.568	0.546	1.6%	2.0%	3.0%
DUFA (Ours)	0.621	0.576	0.526	0.593	0.563	0.545	1.6%	1.6%	2.3%

LOMIA on GPT-4o. Table 3 demonstrates the effectiveness of our proposed attack methods within the LOMIA framework on GPT-4o. Among the three methods, ITFA achieves the best performance on the LOMIA/LAION (AUC 0.669). DUFA and TTFA exhibit consistent results across both datasets. Notably, DUFA achieves the highest AUC and Balanced Accuracy on LOMIA/CC and competitive performance on LOMIA/LAION. This indicates that even closed-source models are vulnerable to privacy attacks. Especially, LLMs like GPT-4o, which are trained on a variety of datasets, are more likely to memorize a greater amount of private information.

Robustness across datasets. We observe that the relative performance ranking among TTFA, ITFA, and DUFA remains largely consistent across both LOMIA/LAION and LOMIA/CC. This suggests that LOMIA is not overfitted to dataset-specific patterns and maintains attack efficiency across varying data distributions, reinforcing its practical applicability in real-world settings.

5.3 Ablation Study

In this section, we conduct comprehensive experiments to investigate the impact of different settings to LOMIA, including different T2T embedding models, different image-text matching models, different max token length and different temperature. Additional experimental results are available in the Appendix C.

Different T2T embedding models. Table 4 shows ablation experiments on LOMIA/LAION under different embedding models to explore its effectiveness, such as all-MiniLM-L6-v2 [21], bge-large-en-v1.5 [53], mxbai-embed-large-v1 [20], and UAE-Large-V1 [26]. While TTFA shows high sensitivity to embedding model choice (e.g., 3.0% TPR@1%FPR for all-MiniLM-L6-v2 but 0.6% for bge-large-en-v1.5), DUFA demonstrates remarkable robustness, maintaining AUC variations within a stable range (from 0.618 to 0.621) across various embedding models.

Table 2: Complete results of various attacks on LOMIA/CC.

Metrics	AUC $\uparrow$		Balanced Acc $\uparrow$		TPR@1%FPR $\uparrow$	
	LLaVA	MiniGPT4	LLaMA Adapter	LLaVA	MiniGPT4	LLaMA Adapter
Logits-based Attacks						
Perplexity	0.634	0.574	0.556	0.615	0.573	0.556
Aug_KL	0.500	0.712	0.539	0.516	0.676	0.541
Max-Prob-Gap	0.607	0.615	0.564	0.601	0.585	0.560
Min0% Prob	0.612	0.775	0.566	0.600	0.708	0.558
Min10% Prob	0.609	0.880	0.568	0.603	0.823	0.565
Min20% Prob	0.608	0.825	0.565	0.588	0.766	0.560
ModRényi($\alpha = 0.5$)	0.633	0.569	0.549	0.611	0.581	0.555
ModRényi($\alpha = 2$)	0.633	0.569	0.544	0.618	0.580	0.548
Max0%Rényi ($\alpha = 0.5$)	0.627	0.808	0.565	0.598	0.771	0.560
Max10%Rényi ($\alpha = 0.5$)	0.646	0.803	0.566	0.610	0.798	0.563
Max100%Rényi ($\alpha = 0.5$)	0.618	0.569	0.605	0.590	0.565	0.596
Max0%Rényi ($\alpha = 1$)	0.644	0.814	0.573	0.610	0.761	0.561
Max10%Rényi ($\alpha = 1$)	0.641	0.847	0.563	0.598	0.801	0.570
Max100%Rényi ($\alpha = 1$)	0.637	0.643	0.591	0.611	0.628	0.585
Max0%Rényi ($\alpha = 2$)	0.618	0.799	0.573	0.590	0.736	0.563
Max10%Rényi ($\alpha = 2$)	0.642	0.891	0.591	0.615	0.828	0.573
Max100%Rényi ($\alpha = 2$)	0.642	0.711	0.591	0.621	0.666	0.575
Max0%Rényi ($\alpha = \infty$)	0.612	0.775	0.538	0.600	0.708	0.561
Max10%Rényi ($\alpha = \infty$)	0.610	0.880	0.585	0.581	0.823	0.568
Max100%Rényi ($\alpha = \infty$)	0.646	0.680	0.591	0.625	0.645	0.580
Label-only Attacks						
Query Attack	0.599	0.551	0.536	0.586	0.553	0.539
TTFA (Ours)	0.584	0.572	0.568	0.585	0.566	0.566
ITFA (Ours)	0.625	0.540	0.571	0.605	0.551	0.560
DUFA (Ours)	0.630	0.560	0.579	0.601	0.555	0.558

Table 3: Performance of LOMIA on GPT-4o.

Metrics	AUC $\uparrow$		Balanced Acc $\uparrow$		TPR@1%FPR $\uparrow$	
	LAION	CC	LAION	CC	LAION	CC
LOMIA						
TTFA	0.602	0.600	0.580	0.588	2.0%	2.0%
ITFA	0.669	0.608	0.635	0.585	2.3%	3.0%
DUFA	0.612	0.618	0.585	0.600	2.0%	2.6%

Table 4: Performance comparison of LOMIA under different sentence transformer settings on LAION.

Metrics	AUC $\uparrow$		Balanced Acc $\uparrow$		TPR@1%FPR $\uparrow$	
	TTFA	DUFA	TTFA	DUFA	TTFA	DUFA
Sentence transformers						
all-MiniLM-L6-v2	0.601	0.621	0.601	0.593	3.0%	1.6%
bge-large-en-v1.5	0.577	0.618	0.580	0.593	0.6%	1.6%
mxbai-embed-large-v1	0.582	0.618	0.581	0.593	1.3%	1.6%
UAE-Large-V1	0.584	0.619	0.581	0.591	1.3%	1.6%

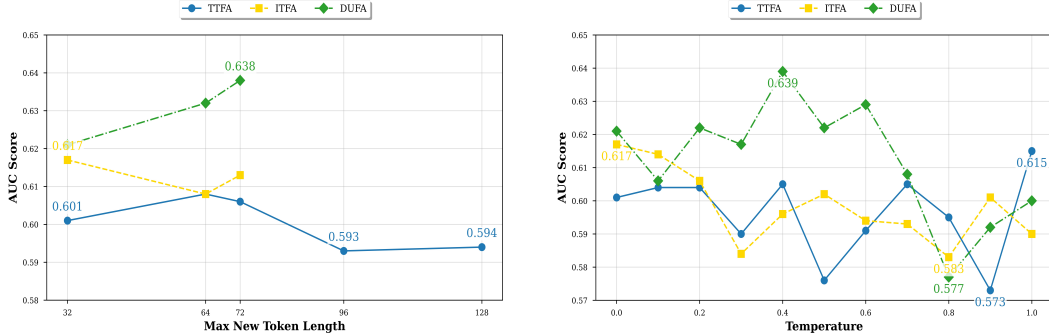
Different image-text matching models. We further explore the impact of different image-text matching models on LOMIA/LAION. Including CLIP [39] and BLIP [24]. These two models differ substantially in architecture and alignment strategies: CLIP employs contrastive learning for feature alignment, while BLIP uses a generative decoder conditioned on image features. Results in Table 5 show that CLIP maintains a stable performance in both ITFA (AUC = 0.617) and DUFA (AUC = 0.621), with a balanced accuracy of 0.591-0.593, reflecting its strong robustness to single-modal noise. In contrast, BLIP fails to detect in ITFA (AUC = 0.485, balanced accuracy 0.518) but shows significant improvement in DUFA (AUC increased by 18.4% to 0.574), revealing that the additional text-text feature provides complementary information.

Table 5: Performance comparison of LOMIA under different image-text matching models settings on LAION.

Metrics Methods	AUC $\uparrow$		Balanced Acc $\uparrow$		TPR@1%FPR $\uparrow$	
	ITFA	DUFA	ITFA	DUFA	ITFA	DUFA
<i>Image-text matching models</i>						
CLIP	0.617	0.621	0.591	0.593	1.6%	1.6%
BLIP	0.485	0.574	0.518	0.585	0.6%	3.0%

Different max token length. We conduct ablation experiments on LOMIA/LAION targeting the length of generated description texts. we restrict the max new tokens parameter of the generation to (32, 64, 72, 96, 128). Due to CLIP’s limitation of the input sequence to 77 tokens, we restrict the parameters of TTFA and DUFA to 72. Figure 2a demonstrates that longer generated description texts are not always better for TTFA and ITFA. However, for DUFA, the AUC increases as the length of the generated descriptions grows.

Different temperature. Since the generative behavior of VLLMs is inherently modulated by temperature settings, we performed an ablation study on LOMIA/LAION to examine the impact of temperature on attack performance. The results in Figure 2b reveal method-specific temperature sensitivities: DUFA reaches its peak at temperature = 0.4 (AUC = 0.639) and remains consistently strong within the 0.2-0.6 range (AUC > 0.62), but its performance deteriorates markedly beyond 0.7, with a 10.3% drop observed above 0.8. In contrast, ITFA is highly sensitive to temperature, performing optimally only in the low-temperature range (0.0-0.2) before rapidly degrading. TTFA follows a similar trend, yet interestingly shows a performance rebound at a temperature of 1, suggesting a potential fallback effect rather than stable robustness. Collectively, these findings underscore the necessity of modality-aware and method-specific temperature scheduling in LOMIA, preserving DUFA’s operation within its stable mid-range, and leveraging the low-temperature range of TTFA and ITFA.



(a) Ablation study on Max New Tokens.

(b) Ablation study on Temperature.

Figure 2: **Ablation study** (a) on different max_new_tokens. (b) on different temperature.

6 Conclusion

In this paper, we explore the susceptibility of VLLMs to MIAs under the restrictive label-only setting. Our attention is directed towards the pre-training phase, which is not only more commonly encountered but also presents considerable technical challenges. We start by reviewing previous MIAs tailored to VLLMs and demonstrate that current methods mainly rely on logits-based attacks, which is a major limitation given the infeasibility of logits access in real-world deployments. Moreover, existing label-only approaches are inadequate when applied to pre-trained VLLMs. To address this issue, we present LOMIA, a novel label-only MIA framework that exploits sentence-level semantic similarity to approximate output perplexity as signals to distinguish membership. Comprehensive evaluations under rigorous benchmarks validate LOMIA’s competitive performance over baselines. These findings highlight the urgent need for defenses specifically tailored to label-only MIAs.

Acknowledgement

This work was supported in part by the Hong Kong Research Grants Council’s (RGC) General Research Fund (GRF) under Grant PolyU 15201323.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Martin Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie H Morgenstern, and Steven Z Wu. Scalable membership inference attacks via quantile regression. *Advances in Neural Information Processing Systems*, 36:314–330, 2023.
- [3] Simone Caldarella, Massimiliano Mancini, Elisa Ricci, and Rahaf Aljundi. The phantom menace: unmasking privacy leakages in vision-language models. *arXiv preprint arXiv:2408.01228*, 2024.
- [4] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.
- [5] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914. IEEE, 2022.
- [6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.
- [7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna>, 3(5), 2023.
- [8] Xuelong Dai, Kaisheng Liang, and Bin Xiao. Advdiff: Generating unrestricted adversarial examples using diffusion models. In *European Conference on Computer Vision*, pages 93–109. Springer, 2024.
- [9] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, pages 8717–8730. PMLR, 2023.
- [10] Vasisht Duddu, Antoine Boutet, and Virat Shejwalkar. Quantifying privacy leakage in graph embedding. In *MobiQuitous 2020-17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, pages 76–85, 2020.
- [11] Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Membership inference attacks against fine-tuned large language models via self-prompt calibration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=PAWQvrForJ>.
- [12] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.
- [13] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10877, 2023.
- [14] Umang Gupta, Dimitris Stripelis, Pradeep K Lam, Paul Thompson, José Luis Ambite, and Greg Ver Steeg. Membership inference attacks on deep regression models for neuroimaging. In *Medical Imaging with Deep Learning*, pages 228–251. PMLR, 2021.
- [15] Yu He, Boheng Li, Liu Liu, Zhongjie Ba, Wei Dong, Yiming Li, Zhan Qin, Kui Ren, and Chun Chen. Towards label-only membership inference attack against pre-trained large language models. In *USENIX Security*, 2025.

- [16] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.
- [17] Pingyi Hu, Zihan Wang, Ruoxi Sun, Hu Wang, and Minhui Xue. M⁴i: Multi-modal models membership inference. *Advances in Neural Information Processing Systems*, 35:1867–1882, 2022.
- [18] Yuke Hu, Zheng Li, Zhihao Liu, Yang Zhang, Zhan Qin, Kui Ren, and Chun Chen. Membership inference attacks against vision-language models. *arXiv preprint arXiv:2501.18624*, 2025.
- [19] Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4871–4881, 2023.
- [20] Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. Open source strikes bread - new fluffy embeddings model. *mixedbread ai inc*, 2024. URL <https://www.mixedbread.ai/blog/mxbai-embed-large-v1>.
- [21] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020.
- [22] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.
- [23] Hao Li, Zheng Li, Siyuan Wu, Chengrui Hu, Yutong Ye, Min Zhang, Dengguo Feng, and Yang Zhang. Seqmia: Sequential-metric based membership inference attack. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 3496–3510, 2024.
- [24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.
- [25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [26] Xianming Li and Jing Li. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*, 2023.
- [27] Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. Membership inference attacks against large vision-language models. *Advances in Neural Information Processing Systems*, 37:98645–98674, 2024.
- [28] Kaisheng Liang, Xuelong Dai, Yanjie Li, Dong Wang, and Bin Xiao. Improving transferable targeted attacks with feature tuning mixup. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25802–25811, 2025.
- [29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2023.
- [31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.
- [32] Yihao Liu, Jinhe Huang, Yanjie Li, Dong Wang, and Bin Xiao. Generative ai model privacy: a survey. *Artificial Intelligence Review*, 58(1):1–47, 2025.
- [33] Xinqi Lyu, Yihao Liu, Yanjie Li, and Bin Xiao. Pla: Prompt learning attack against text-to-image generative models. *arXiv preprint arXiv:2508.03696*, 2025.
- [34] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE, 2019.

- [35] Long D Nguyen, Dongyun Lin, Zhiping Lin, and Jiuwen Cao. Deep cnns for microscopic image classification by exploiting transfer learning and feature concatenation. In *2018 IEEE international symposium on circuits and systems (ISCAS)*, pages 1–5. IEEE, 2018.
- [36] OpenAI. Gpt-4 technical report, 2023.
- [37] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.
- [38] Hussam Qassim, Abhishek Verma, and David Feinzimer. Compressed residual-vgg16 cnn model for big data places image recognition. In *2018 IEEE 8th annual computing and communication workshop and conference (CCWC)*, pages 169–175. IEEE, 2018.
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.
- [41] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.
- [42] Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. Human-adversarial visual question answering. *Advances in Neural Information Processing Systems*, 34:20346–20359, 2021.
- [43] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2023.
- [44] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *ICLR*, 2024.
- [45] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE (SP)*, pages 3–18, 2017.
- [46] Sizhky. open-images-captions-micro. <https://huggingface.co/datasets/sizhky/open-images-captions-micro>, 2024.
- [47] Congzheng Song and Ananth Raghunathan. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pages 377–390, 2020.
- [48] Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.
- [49] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [50] Dong Wang, Yicheng Liu, Liangji Fang, Fanhua Shang, Yuanyuan Liu, and Hongying Liu. Balanced gradient penalty improves deep long-tailed learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5093–5101, 2022.
- [51] Rui Wen, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against in-context learning. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 3481–3495, 2024.
- [52] Bingzhe Wu, Shiwan Zhao, Chaochao Chen, Haoyang Xu, Li Wang, Xiaolu Zhang, Guangyu Sun, and Jun Zhou. Generalization in generative adversarial networks: A novel perspective from privacy protection. *Advances in Neural Information Processing Systems*, 32, 2019.
- [53] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-pack: Packaged resources to advance general chinese embedding, 2023.

- [54] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.
- [55] Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. LLaMA-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=d4UiXAHN2W>.
- [56] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [NA].
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”,**
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: We clearly express our motivation and insight in the Abstract and Introduction. The contributions are summarized in the Introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We clearly express our limitations in the Appendix E.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: There is no assumption or theoretical result in this work.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our paper provides a detailed account of implementation details and hyperparameter settings of our method for reproducing in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The used datasets are publicly available. We will consider releasing the code upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: We provide implementation details in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[No\]](#)

Justification: We fix the random seeds in experiments for fair comparison.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: We specify the type of compute workers in implementation details in Section 5.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We follow the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: This work focuses on membership inference attack, which is the most common privacy risk in machine learning. We discussed the potential impacts in Section 1.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Not applicable.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: We cite the datasets and models used in our paper and give a brief introduction.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[NA\]](#)

Justification: We do not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper has nothing to do with crowdsourcing and human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [No]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Details of Baselines

We summarize the details of the baselines as follows:

- **PPL Attack [4, 54]** uses the perplexity $\mathcal{PPL}$ of a target sample x evaluated on the target model M_θ as the membership score, making a prediction based on whether the negative perplexity exceeds a threshold: $\mathcal{A}(x, \theta) = \mathbb{1}[-\mathcal{PPL}(x, M_\theta) \geq \tau]$.
- **Aug-KL [27]** uses KL-divergence to compare the logit distributions, where p is the logit distribution of the target sample x on the target model f_θ , and q is the logit distribution of the augmented sample: $\mathcal{A}(x, \theta) = \mathbb{1}[\mathbb{D}_{\text{KL}}(p \parallel q) \geq \tau]$.
- **MIN-K% PROB [44]** determines the membership score by focusing on the k% of tokens exhibiting the lowest likelihoods: $\mathcal{A}(x, \theta) = \mathbb{1}[\frac{1}{|\text{Min-K}(t)|} \sum_{t_i \in \text{Min-K}(t)} \log p(t_i | t_1, \dots, t_{i-1}) \geq \tau]$.
- **Max-Prob-Gap% PROB [27]** subtracts the second largest probability from the maximum probability in each token position and calculate the mean: $\mathcal{A}(x, \theta) = \frac{1}{|\text{Max-Gap}(t)|} \sum_{t_i \in \text{Max-Gap}(t)} (\max_j p_j - \text{second max}_j p_j) \geq \tau$.
- **MaxRényi-K% [27]** the top K% from the sequence X with the largest Rényi entropies: $\mathcal{A}(x, \theta) = \mathbb{1}[\frac{1}{|\text{Max-K\%}(X)|} \sum_{i \in \text{Max-K\%}(X)} H_\alpha(p^{(i)}) \geq \tau]$.
- **ModRényi* [27]** extends MaxRényi-K% to the target-based scenarios: $\mathcal{A}(x, \theta) = \mathbb{1}[-\frac{1}{|\alpha-1|} \left((1-p_y)p_y^{|\alpha-1|} - (1-p_y) + \sum_{j \neq y} p_j(1-p_j)^{|\alpha-1|} - p_j \right) \geq \tau]$.
- **Query attack [18]** queries the VLLMs a number of times to generate different descriptions. Then calculate the pairwise similarity between all these descriptions and take the average value as the similarity score for the image. If the average value exceeds a certain threshold, the image is classified as a member. Here, we set the number of queries to 5 and the temperature to 0.5, which performs the best.

B Details of VLLMs

Table 6: VLLM details

Model	Mini-GPT4	LLaVA 1.5	LLaMA Adapter v2.1
Base LLM	Vicuna-v1.5-7B	Vicuna-v1.5-7B	LLaMA-7B
Vision processor	BLIP2/EVA-ViT-G	CLIP-ViT-L	CLIP-ViT-L

C Additional experiments

C.1 Performance for LOMIA on larger VLLMs

We also test our LOMIA on larger VLLMs, such as LLaVA-1.5-13b and MiniGPT4-vicuna-13b, as shown in Table 7 and Table 8.

Table 7: Performance of LOMIA on LLaVA-1.5-13b.

Metrics	AUC $\uparrow$		Balanced Acc $\uparrow$		TPR@1%FPR $\uparrow$	
	LAION	CC	LAION	CC	LAION	CC
LOMIA						
TTFA	0.601	0.576	0.596	0.568	4.0%	4.0%
ITFA	0.594	0.601	0.578	0.575	2.6%	3.3%
DUFA	0.607	0.609	0.598	0.581	3.3%	3.3%

C.2 Different sample ratios on LOMIA/LAION

We explore the impact of sample ratio on the performance of LOMIA by testing ratios of 10%, 25%, 50%, 75%, and 100%. We still use the whole samples from LOMIA/LAION to fit the regression

Table 8: Performance of LOMIA on MiniGPT4-vicuna-13b.

Metrics	AUC $\uparrow$		Balanced Acc $\uparrow$		TPR@1%FPR $\uparrow$	
Datasets	LAION	CC	LAION	CC	LAION	CC
LOMIA						
TTFA	0.588	0.584	0.583	0.576	1.0%	4.6%
ITFA	0.573	0.539	0.580	0.546	0.6%	1.0%
DUFA	0.590	0.559	0.598	0.561	1.0%	1.6%

Table 9: Performance of TTFA under different sample settings on LOMIA/LAION.

Ratio	AUC $\uparrow$	Balanced Acc $\uparrow$	TPR@1%FPR $\uparrow$
10%	0.637	0.662	6.3%
25%	0.589	0.604	1.5%
50%	0.584	0.589	3.8%
75%	0.590	0.594	3.3%
100%	0.601	0.601	3.0%

Table 10: Performance of ITFA under different sample settings on LOMIA/LAION.

Ratio	AUC $\uparrow$	Balanced Acc $\uparrow$	TPR@1%FPR $\uparrow$
10%	0.662	0.662	3.8%
25%	0.614	0.606	2.2%
50%	0.642	0.624	1.7%
75%	0.615	0.594	2.1%
100%	0.617	0.591	1.6%

Table 11: Performance of DUFA under different sample settings on LOMIA/LAION.

Ratio	AUC $\uparrow$	Balanced Acc $\uparrow$	TPR@1%FPR $\uparrow$
10%	0.589	0.617	5.3%
25%	0.614	0.614	2.1%
50%	0.612	0.611	4.4%
75%	0.600	0.594	3.5%
100%	0.621	0.593	1.6%

model. During the inference phase, we adjust the number of samples by randomly selecting a specific ratio of samples from LOMIA/LAION to query the target model (LLaVA-1.5-7b). To ensure fairness, we average the results over five independent queries. Since the sampled data is randomly selected each time, the TTFA, ITFA, and DUFA methods are evaluated on different batches of data in each trial. Table 9, Table 10 and Table 11 show the attack results of different sample ratios.

C.3 Balanced Accuracy Results on different Max New Tokens and on different Temperature

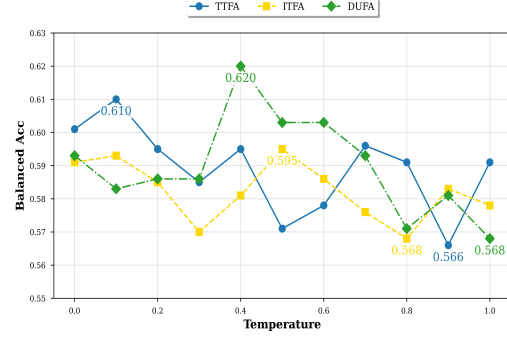
Figure 3a and Figure 3b show the balanced accuracy results on different max new tokens and on different temperature.

C.4 Complete results of various attacks on larger dataset

To evaluate the impact of a larger test set, we expanded the LOMIA/LAION dataset to 1,000 samples (500 members and 500 non-members). The attack performance with this increased number of testing data points is reported in the Table 12 and Table 13.



(a) Ablation study on Max New Tokens.



(b) Ablation study on Temperature.

Figure 3: Balanced accuracy results of LOMIA on different (a) *max_new_tokens*. (b) *temperature*.

Table 12: Complete results of various attacks on larger dataset.

Metrics	AUC ↑			Balanced Acc ↑			TPR@1%FPR ↑		
	LLaVA	MiniGPT4	LLaMA Adapter	LLaVA	MiniGPT4	LLaMA Adapter	LLaVA	MiniGPT4	LLaMA Adapter
Logits-based Attacks									
Perplexity	0.628	0.586	0.588	0.601	0.570	0.574	5.2%	2.6%	3.6%
Aug_KL	0.530	0.515	0.501	0.532	0.526	0.533	1.8%	1.2%	3.6%
Max-Prob-Gap	0.606	0.574	0.552	0.591	0.565	0.557	1.4%	1.4%	0.6%
Min0% Prob	0.666	0.565	0.526	0.641	0.555	0.532	3.0%	1.6%	1.4%
Min10% Prob	0.641	0.563	0.546	0.659	0.552	0.546	4.0%	2.6%	1.8%
Min20% Prob	0.648	0.571	0.561	0.618	0.560	0.562	4.0%	2.8%	2.4%
ModRényi($\alpha = 0.5$)	0.629	0.581	0.588	0.593	0.568	0.576	3.0%	2.0%	2.0%
ModRényi($\alpha = 2$)	0.630	0.587	0.589	0.597	0.573	0.577	3.0%	2.6%	1.6%
Max0%Rényi ($\alpha = 0.5$)	0.596	0.528	0.622	0.574	0.532	0.594	3.8%	0.6%	2.2%
Max10%Rényi ($\alpha = 0.5$)	0.664	0.545	0.547	0.627	0.547	0.540	5.2%	1.0%	1.2%
Max100%Rényi ($\alpha = 0.5$)	0.609	0.569	0.564	0.590	0.559	0.560	5.6%	1.4%	1.0%
Max0%Rényi ($\alpha = 1$)	0.635	0.559	0.610	0.594	0.551	0.591	5.2%	1.2%	4.0%
Max10%Rényi ($\alpha = 1$)	0.724	0.571	0.557	0.666	0.562	0.550	7.8%	2.4%	1.6%
Max100%Rényi ($\alpha = 1$)	0.637	0.589	0.582	0.604	0.572	0.569	6.4%	3.0%	1.6%
Max0%Rényi ($\alpha = 2$)	0.676	0.565	0.550	0.645	0.557	0.546	5.0%	3.2%	3.6%
Max10%Rényi ($\alpha = 2$)	0.748	0.573	0.552	0.688	0.566	0.562	4.4%	3.0%	2.2%
Max100%Rényi ($\alpha = 2$)	0.639	0.590	0.588	0.612	0.571	0.570	5.8%	2.8%	1.8%
Max0%Rényi ($\alpha = \infty$)	0.666	0.567	0.526	0.641	0.556	0.532	3.0%	2.2%	1.4%
Max10%Rényi ($\alpha = \infty$)	0.704	0.566	0.545	0.659	0.551	0.546	4.0%	2.8%	1.8%
Max100%Rényi ($\alpha = \infty$)	0.639	0.588	0.588	0.611	0.569	0.574	6.8%	2.6%	1.8%
Label-only Attacks									
Query Attack	0.531	0.534	0.534	0.534	0.536	0.532	0.4%	0.5%	0.5%
TTFA (Ours)	0.597	0.570	0.566	0.581	0.562	0.560	2.2%	2.2%	0.6%
ITFA (Ours)	0.620	0.572	0.571	0.594	0.568	0.567	1.4%	0.8%	2.6%
DUFA (Ours)	0.628	0.576	0.578	0.603	0.570	0.564	2.0%	1.0%	2.4%

Table 13: Performance of larger dataset on GPT-4o.

Model	AUC ↑	Balanced Acc ↑	TPR@1%FPR ↑
TTFA	0.598	0.577	1.8%
ITFA	0.648	0.616	3.2%
DUFA	0.675	0.637	2.2%

C.5 Complete results of various attacks when non-member data are real images

To evaluate the impact of real non-member, we randomly sampled 500 image-caption pairs from open-images-captions-micro on Hugging Face [46], ensuring exclusion of any data overlapping with COCO or LAION datasets. We also enlarged our LOMIA/LAION member data samples to 500, the final results can be found in the Table 14 and Table 15.

Table 14: Complete results of various attacks when non-member data are real images.

Metrics	AUC $\uparrow$			Balanced Acc $\uparrow$			TPR@1%FPR $\uparrow$		
	LLaVA	MiniGPT4	LLaMA Adapter	LLaVA	MiniGPT4	LLaMA Adapter	LLaVA	MiniGPT4	LLaMA Adapter
Logits-based Attacks									
Perplexity	0.640	0.589	0.591	0.604	0.584	0.573	3.2%	2.8%	2.8%
Aug_KL	0.544	0.512	0.494	0.552	0.546	0.516	0.8%	4.2%	0.8%
Max-Prob-Gap	0.620	0.575	0.567	0.595	0.570	0.576	1.6%	2.4%	0.8%
Min0% Prob	0.592	0.571	0.521	0.585	0.565	0.538	1.0%	1.8%	1.0%
Min10% Prob	0.538	0.575	0.533	0.538	0.569	0.538	1.2%	2.2%	2.0%
Min20% Prob	0.571	0.579	0.588	0.569	0.572	0.573	4.0%	1.8%	2.8%
ModRényi($\alpha = 0.5$)	0.653	0.583	0.630	0.623	0.571	0.598	3.2%	2.4%	6.4%
ModRényi($\alpha = 2$)	0.658	0.588	0.629	0.627	0.575	0.593	3.6%	2.4%	5.0%
Max0%Rényi ($\alpha = 0.5$)	0.555	0.578	0.629	0.548	0.571	0.599	1.2%	1.8%	7.2%
Max10%Rényi ($\alpha = 0.5$)	0.560	0.570	0.528	0.553	0.565	0.537	2.4%	1.4%	1.4%
Max100%Rényi ($\alpha = 0.5$)	0.611	0.584	0.547	0.601	0.578	0.545	3.6%	2.2%	2.2%
Max0%Rényi ($\alpha = 1$)	0.601	0.579	0.654	0.582	0.565	0.622	2.2%	1.6%	6.8%
Max10%Rényi ($\alpha = 1$)	0.710	0.571	0.574	0.652	0.562	0.565	8.2%	2.4%	1.6%
Max100%Rényi ($\alpha = 1$)	0.642	0.584	0.580	0.611	0.579	0.574	5.8%	2.2%	2.2%
Max0%Rényi ($\alpha = 2$)	0.596	0.577	0.693	0.582	0.564	0.650	3.4%	1.8%	2.8%
Max10%Rényi ($\alpha = 2$)	0.688	0.573	0.586	0.645	0.566	0.583	5.6%	3.0%	2.6%
Max100%Rényi ($\alpha = 2$)	0.623	0.600	0.602	0.600	0.637	0.587	3.6%	3.6%	2.6%
Max0%Rényi ($\alpha = \infty$)	0.647	0.703	0.688	0.628	0.640	0.532	3.6%	2.2%	7.6%
Max10%Rényi ($\alpha = \infty$)	0.666	0.710	0.573	0.634	0.645	0.572	3.2%	6.2%	3.6%
Max100%Rényi ($\alpha = \infty$)	0.649	0.710	0.608	0.630	0.647	0.588	7.2%	5.8%	3.0%
Label-only Attacks									
Query Attack	0.593	0.554	0.536	0.573	0.559	0.546	0.5%	0.5%	0.2%
TTFA (Ours)	0.579	0.571	0.545	0.571	0.566	0.544	1.2%	1.9%	1.8%
ITFA (Ours)	0.630	0.575	0.575	0.600	0.569	0.562	2.2%	1.8%	1.6%
DUFA (Ours)	0.635	0.580	0.578	0.605	0.576	0.566	2.6%	1.8%	2.6%

Table 15: Performance of real non-member dataset on GPT-4o.

Model	AUC $\uparrow$	Balanced Acc $\uparrow$	TPR@1%FPR $\uparrow$
TTFA	0.663	0.625	4.2%
ITFA	0.609	0.592	1.2%
DUFA	0.665	0.645	3.4%

D Threshold Calibration

We use the AUC score, balanced accuracy, and TPR@low FPR as our primary metrics. These metrics are standard in the previous work, such as [44, 27], as both are threshold-independent. Specifically, the AUC represents the area under the ROC curve, which plots the true positive rate against the false positive rate across all possible thresholds. This means it is not necessary to select a specific threshold when comparing different MIA methods. However, if the goal is to deploy a particular MIA method in practice, a threshold can be determined by performing hyperparameter search on a small validation set.

E Limitations

One limitation of this work is that the TPR@1%FPR is relatively low, but it’s important to note that TPR@1%FPR is an extremely stringent evaluation metric that requires high true positive rates while maintaining only 1% false positive rate. Even under this strict criterion, our method consistently outperforms existing label-only attack baselines and shows competitive performance with logits-based MIAs. Moreover, promising defense strategies include differentially private pre-training and machine unlearning, which aim to reduce memorization in VLLMs and thereby mitigate privacy risks. We leave this as a critical direction for future work.