Tracing Training Footprints: A Calibration Approach for Membership Inference Attacks Against Multimodal Large Language Models

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

Abstract

With the increasing scale of training data for Multimodal Large Language Models (MLLMs) and the lack of data details, there is growing concern about privacy breaches and data security issues. Under black-box access, exploring effective Membership Inference Attacks (MIA) has garnered increasing attention. In real-world applications, where most samples are non-members, the issue of non-members being over-represented in the data manifold, leading to misclassification as member samples, becomes more prominent. This has motivated recent work to focus on developing effective difficulty calibration strategies, producing promising results. However, these methods only consider text-only input during calibration, and their effectiveness is diminished when migrated to MLLMs due to the presence of visual embeddings. To address the above problem, we propose PC-MMIA, focusing on visual instruction fine-tuning data. PC-MMIA is based on the idea that tokens located in poorly generalized local manifolds can better reflect traces of member samples that have been trained. By employing bidirectional perturbation of image embeddings to capture tokens critical to MIA and assigning them different weights, we achieve difficulty calibration. Experimental results demonstrate that our proposed method surpasses existing methods.

1 Introduction

In recent years, the rapid development of large language models (LLMs) has brought new opportunities for Multimodal Large Language Models (MLLMs) research, leading to the emergence of vision-language models such as GPT-4o and Qwen2.5-VL (Bai et al., 2025; OpenAI, 2024). By aligning visual embeddings with the textual space, MLLMs can process and reason on multimodal data, enabling them to perform complex tasks such as image captioning and visual question answering (Wang et al., 2024b; Zheng et al., 2025).

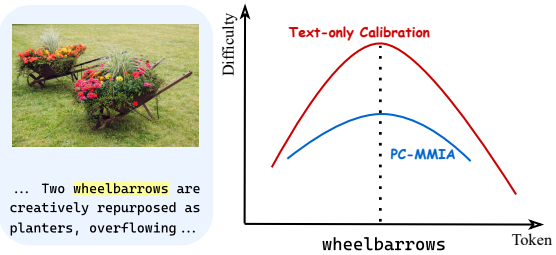

Figure 1: Text-only calibration fails to consider image semantics, consequently overestimating the difficulty of predicting "wheelbarrows."

This rapid development is heavily reliant on the availability of large-scale internet datasets, which have tremendously improved model performance (Zhang et al., 2024a; Wang et al., 2024a). However, this also raises concerns about privacy leakage and unauthorized data usage. For instance, image data used in commercial model training may include private photos or copyrighted content (Grynbaum and Mac, 2023; Knibbs, 2023). To explore these issues, Membership Inference Attacks (MIA) serve as a common technique, where attackers attempt to infer whether specific data samples are part of the model's training set (Shokri et al., 2017). MIA has now evolved from a mere attack method into an effective tool for auditing black-box models (Hu et al., 2022; Wu and Cao, 2025), providing a new perspective on data privacy protection. MIA can generally be divided into three attack scenarios: black-box access, gray-box access, and white-box access (Cheng et al., 2025). In next-token prediction, black-box access provides only the probability of the predicted token, while gray-box access allows retrieval of probabilities for all candidate tokens in the vocabulary (Shi et al., 2024b; Zhang et al., 2025b). Under white-box access, attackers can access the model's architecture and parameters and even train the model.

Research on MIA is crucial for protecting data security and maintaining user privacy, especially

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 17179–17191
November 4-9, 2025

for MIA targeting visual instruction fine-tuning data, which is more likely to contain unauthorized or private data (Li et al., 2024; Song et al., 2025; Wang et al., 2025). Despite numerous MIA methods proposed in the LLMs domain, their direct naive migration to MLLMs is unsatisfactory due to the multimodal inputs of MLLMs (Wu and Cao, 2025). Thus, there is an urgent need for a MIA framework specifically designed for MLLMs.

The major challenge faced by MIA against MLLMs is the need to consider the impact of image embeddings on the difficulty of token prediction. Previous conventional methods rely on the notion that the word probability distribution based on text can reveal whether it is included in the training set (Shi et al., 2024b). For instance, when a token's predicted probability is high, these methods assume that the token is likely present in member samples. However, they overlook the fact that the token may simply be simple and easy to predict, meaning non-member samples might themselves be over-represented in the data manifold (Zhang et al., 2024b; Shi et al., 2024a). To address this, Watson et al. (2022) introduce the concept of difficulty calibration, highlighting the need to mitigate biases arising from the intrinsic complexity of samples, and calibrate using the predictive difficulty of the tokens themselves. Recent works attempt to utilize various methods including keyword marking (Antebi et al., 2025) and divergence-from-randomness (Zhang et al., 2024b) for calibration, yielding promising results. Nonetheless, as illustrated in Figure 1, in multimodal data, these calibration methods are only effective for text tokens and fail to account for the impact of image embeddings on the probability distribution, leading to suboptimal performance.

Facing the challenges above, our research aims to address the question: **In the presence of image embedding input, if a specific text token in the response has a high probability, is this due to verbatim memorization (overfitting of member data), or is it simply because the token itself is easily predictable?** Inspired by Mattern et al. (2023), we propose a perturbation-based calibration method, named PC-MMIA, to calibrate token probabilities for membership inference attacks against MLLMs. Our core idea is that tokens located in poorly generalized local manifolds can better reflect traces of training on member samples. After perturbing images, we observe probability changes for each token in the samples to determine whether the token's probability contribution stems more from generalized prediction or verbatim memorization of member data. Subsequently, we normalize the changes in log-likelihood for each token before and after perturbation to serve as correction weights.

We summarize our main contributions as follows: i) We propose PC-MMIA, a novel membership inference attack method that focuses on the issue of non-member samples potentially being over-represented in the data manifold. ii) To the best of our knowledge, we are the first to consider the visual embeddings in the difficulty calibration of MLLMs, opening up new directions for MIA research. iii) Compared to other black-box MIA methods, our approach achieves optimal performance in commonly used detection metrics TPR at low FPR and AUC. Additionally, we have open-sourced the code to facilitate further research [1] .

2 Related Work

2.1 Membership Inference Attacks (MIA).

Initial MIA research concentrates on traditional deep learning models, such as supervised classification models (Shokri et al., 2017; Yeom et al., 2018; Hu et al., 2022). Recently, the focus shifts towards Large Language Models (LLMs) (Wu and Cao, 2025), with studies exploring their implications in privacy auditing, memorization assessment, and the detection of data contamination and copyrighted content (Steinke et al., 2023; Xu et al., 2024a; Mireshghallah et al., 2022; Duarte et al., 2024; Xu et al., 2024b). Our work primarily addresses the most challenging and general scenario: black-box model access in MIA, which can be categorized into reference-based and reference-free methods (Oren et al., 2023; Wu and Cao, 2025).

Reference-based methods assume access to a shadow model trained on a dataset with the same distribution as the target model. For instance, Fu et al. (2024) utilizes synthetic data generated by LLMs to train a reference model. While these methods directly leverage the reference model for difficulty calibration, they are often costly and have limited applicability. Reference-free methods, currently the mainstream approach, aim to infer membership by analyzing token prediction probabilities (Wu and Cao, 2025). Shi et al. (2024b) introduces the Min-K% method, based on the assump-

[1] Our code is available at: https://github.com/qingpingwan/PC-MMIA

tion that non-member samples are more likely to contain low-probability words. This approach overlooks the inherent complexity of different samples. To address this, Zhang et al. (2024b) constructs a large-scale text database to estimate the prediction difficulty of various tokens based on word frequency distributions. Zhang et al. (2025b) employs polarization enhancement for calibration, while Zhang et al. (2024b) proposes Min-K%++, which builds upon Min-K% by considering whether the input forms a mode or has relatively high probability under the conditional categorical distribution. Although Min-K%++ achieves competitive results, it requires gray-box access to the target model.

Approaches most related to ours are Neighbourhood Comparison methods (Mattern et al., 2023), which involve rewriting or replacing keywords in the text. These methods operate under the assumption that such perturbations, which preserve the text's meaning, should not significantly increase the overall loss unless the text is an overfitted member sample (Zhang et al., 2025a). PC-MMIA fundamentally differs from these approaches. Our core idea is to perturb images to calibrate the probability contributions of different text tokens within the same sample, thereby identifying tokens that better reflect the training traces of member samples. In contrast, Neighbourhood Comparison methods directly use the loss changes of different samples after text rewriting for membership inference, making them more susceptible to inconsistencies in perturbation strength across samples (Duan et al., 2024). Additionally, these methods do not account for the impact of image embeddings.

2.2 MIA against Multimodal LLMs

The interactive capabilities of MLLMs with humans largely depend on the instruction fine-tuning phase, which is closely tied to the quality of the multimodal datasets (Duan et al., 2024). However, model developers are increasingly reluctant to disclose data details, prompting recent studies to explore MIA against MLLMs (Song et al., 2025). Li et al. (2024), inspired by Min-K%, proposes MaxRényi-K%, based on the assumption that member data should have higher confidence in predicting the next token, incorporating Rényi entropy to measure prediction confidence. Hu et al. (2025) amplifies the probability differences between member and non-member data by adjusting the temperature parameter. Nonetheless, these methods overlook the inherent complexity of multimodal samples and fail to calibrate prediction probabilities (Watson et al., 2022), making it challenging to distinguish whether high-probability tokens result from verbatim memorization of member data or are inherently easy to predict due to image semantics.

3 PC-MMIA

In order to achieve image understanding, MLLMs like LLaVA (Liu et al., 2023) and CogVLM (Wang et al., 2024c) project the image embeddings from visual encoders into the feature space of LLMs. Consequently, when considering the prediction difficulty of text tokens, the semantic impact of images must also be taken into account, hindering the effective migration of most existing MIA methods designed for LLMs to MLLMs. To address this, we propose PC-MMIA, a novel approach tailored for MLLMs.

3.1 Problem Description

We follow prior research (Li et al., 2024; Hu et al., 2025), focusing on membership inference attacks during the instruction fine-tuning phase, emphasizing the most challenging yet broadly applicable black-box access. This implies only having access to the probabilities of next text tokens, with no access to logits, model weights, or gradients.

For a data instance v and a multimodal large language model $\mathcal{M}$ fine-tuned on dataset $\mathcal{D}$, the goal of membership inference attacks is to detect whether data instance v belongs to $\mathcal{D}$. Here, $v = (x, y_{\text{text}}), x = (x_{\text{img}}, x_{text})$, comprises input image x_{img}, input text x_{text}, and response y_{text}. Through calculating the membership score $\mathcal{S}(v; \mathcal{M})$ and threshold decision-making, it is determined whether v is a member or non-member of dataset $\mathcal{D}$. The key to MIA lies in designing an appropriate scoring function $\mathcal{S}$ to better distinguish between member and non-member data.

3.2 Motivation of PC-MMIA

Prior research has demonstrated significant variability in the strength of membership inference signals conveyed by different tokens (Zhang et al., 2024b; Antebi et al., 2025). Specifically, the prediction probability contribution for some tokens in the model may arise from verbatim memorization of training data, while for others, it may stem primarily from the model's generalization capability (Dong et al., 2024). This phenomenon becomes more pronounced in multimodal inputs, as the rich

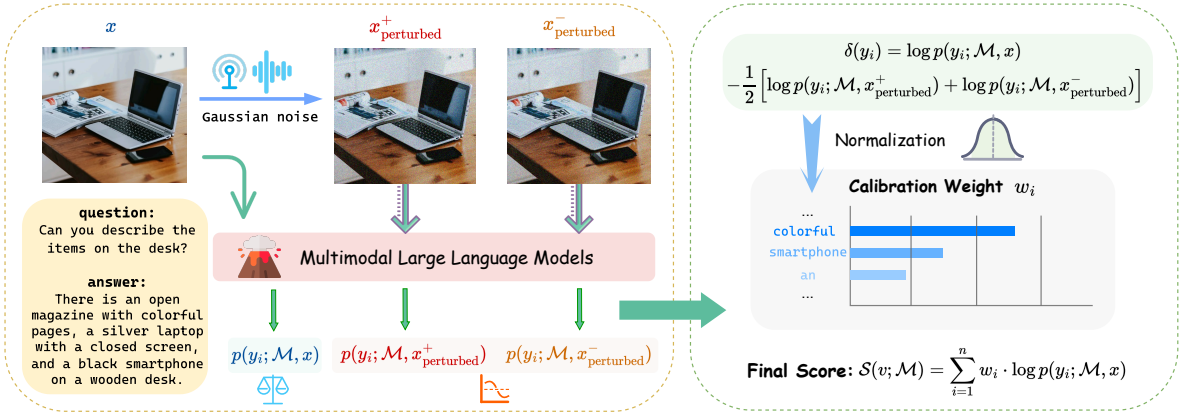

Figure 2: Illustration of PC-MMIA. The core idea of PC-MMIA is to improve the representation of training traces in member samples by identifying tokens located in poorly generalized local manifolds, thus calibrating non-member samples that may themselves be over-represented in the data manifold.

visual semantics contained in image embeddings further amplify these differences.

The PC-MMIA method is proposed to address this issue, aiming to help distinguish tokens that are memorized due to training data presence from those tokens that are universally learned by the model and easily predictable across any data. We view the model's decision-making process as mapping inputs within a high-dimensional space, where each image-text pair sample lies on a manifold representing the data distribution. When slight perturbations are applied to the images, it leads to small displacements of the data point on the manifold. PC-MMIA identifies tokens located in poorly generalized regions of the model's decision space by analyzing changes in prediction probabilities before and after perturbation, thereby pinpointing areas likely influenced by specific training instances rather than by generalized patterns. This approach can also be seen as measuring the local smoothness of the model's prediction function in the input space. Memorization of training samples often induces local discontinuities in the prediction function, manifesting as significant probability fluctuations during perturbation. In contrast, predictions that generalize well exhibit robustness and demonstrate better local stability.

3.3 Formulation of PC-MMIA

As illustrated in Figure 2, we first apply symmetric Gaussian noise to provide bidirectional perturbations to the image embeddings:

$$\begin{aligned} x^+_{\text{perturbed}} &= (x_{\text{img}} + \epsilon, x_{text}), \\ x^-_{\text{perturbed}} &= (x_{\text{img}} - \epsilon, x_{text}), \end{aligned} \tag{1}$$

where ϵ denotes Gaussian noise with zero mean and standard deviation σ, i.e., $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Notably, small-magnitude Gaussian noise can be regarded as a semantically equivalent perturbation. Such perturbations preserve the structural and spatial information of the original image, maintaining visual attributes like color, shape, and dynamics. Subsequently, both the original and perturbed images are fed into the model $\mathcal{M}$ to compute the likelihoods for each token in the sequence $y_{\text{text}} = \{y_1, y_2, \ldots, y_n\}$:

$$\begin{aligned} &p(y_i \mid y_{<i}, x; \mathcal{M}) : 0 < i \leq n, \\ &p(y_i \mid y_{<i}, x^+_{\text{perturbed}}; \mathcal{M}) : 0 < i \leq n, \\ &p(y_i \mid y_{<i}, x^-_{\text{perturbed}}; \mathcal{M}) : 0 < i \leq n, \end{aligned} \tag{2}$$

where the probability of each token y_i is predicted based on prior context. For simplicity, we denote $p(y_i \mid y_{<i}, x; \mathcal{M})$ as $p(y_i; \mathcal{M}, x)$.

Next, we compute the average change in log-likelihood for each token before and after image perturbation:

$$\begin{aligned} &\delta(y_i) = \log p(y_i; \mathcal{M}, x) - \\ &\frac{1}{2}\left[\log p(y_i; \mathcal{M}, x^+_{\text{perturbed}}) + \log p(y_i; \mathcal{M}, x^-_{\text{perturbed}})\right]. \end{aligned} \tag{3}$$

We consider tokens with larger $\delta(y_i)$ as potentially containing more training traces, making them more critical in membership inference attacks. In contrast, tokens with smaller $\delta(y_i)$ may be trivial and easily predictable, thus reducing their weight in $\mathcal{S}$ can prevent the influence of over-represented samples within the data manifold, achieving difficulty correction. To accomplish this, we subse-

quently perform softmax normalization on $\delta(y_i)$:

$$w_i = \frac{\exp(\delta(y_i))}{\sum_{j=1}^{n} \exp(\delta(y_j))}. \quad (4)$$

Through softmax normalization, we use the relative relationship of $\delta(y_i)$ within each sample, rather than the absolute values, as the weight, which ensures that $\mathcal{S}(v; \mathcal{M})$ focuses only on the most discriminative token features within the sample, eliminating score bias due to sensitivity differences to global perturbations between samples, thereby effectively enhancing the decision boundary robustness between members and non-members. Finally, we calculate the corrected score for sample v according to the normalized weights:

$$\mathcal{S}(v; \mathcal{M}) = \sum_{i=1}^{n} w_i \cdot \log p(y_i; \mathcal{M}, x). \quad (5)$$

Comparing the score to the predetermined threshold λ, the final prediction results:

$$prediction(v, \mathcal{M}) = \mathbb{I}[\mathcal{S}(v; \mathcal{M}) > \lambda]. \quad (6)$$

4 Experiments

In this section, we conduct comprehensive experiments to validate the effectiveness of PC-MMIA.

4.1 Experimental Setup

Models. In the current mainstream open-source multimodal models, most only provide access to the model weights, which significantly limits our experiments because it is difficult to confirm whether our chosen evaluation datasets are included in the training datasets. To overcome this barrier, we select Llava-1.5-7b (Liu et al., 2023) and CogVLM-17b (Wang et al., 2024c) for our experiments. These models not only provide model weights but also fully open multimodal datasets, source code, and detailed training processes, allowing us to completely reproduce the visual instruction tuning phase. Furthermore, they feature different model architectures and training strategies, which strongly support the diversity and comprehensiveness of our experiments.

Datasets. Previous studies (Maini et al., 2024; Duan et al., 2024) have indicated that dividing member and non-member based on time may not be reliable, as data generated at different times inherently undergo distribution shifts. To thoroughly address this issue, we verify the method's effectiveness through complete training. We selected NoCaps (Agrawal et al., 2019), Flickr30K (Young et al., 2014), and PixMo (Deitke et al., 2024) for experiments, as these datasets offer comprehensive and diverse samples suitable for MIA evaluation. Importantly, they are not included in the training data of models. Each dataset is randomly divided into member and non-member parts, with the member portion mixed into the original fine-tuning dataset to perform visual instruction tuning, ensuring identical distribution between member and non-member. More details about experiments are found in Appendix §A and §D.

Baseline. We selected six popular methods for evaluation, which include five methods for LLMs and one method for MLLMs. Methods for LLMs: Loss (Yeom et al., 2018), Neighbor (Mattern et al., 2023), Min-K% Prob (Shi et al., 2024b), DC-PDD (Zhang et al., 2024b), NormAC (Zade et al., 2025), Min-K%++ Prob (Zhang et al., 2025b); Method for MLLMs: MaxRényi-K% (Li et al., 2024), which is currently the only black-box membership inference attack applicable directly to visual instruction tuning data. We perform simple modifications on the methods aimed at LLMs to make them applicable to MLLMs. A more detailed description of the baseline is in Appendix §B.

Implementation Details and Evaluation Metrics. Following most existing works (Shi et al., 2024b; Zhang et al., 2024b, 2025b), we do not set the hyperparameter λ value but instead use AUC scores (Area Under the ROC Curve) and TPR (True Positive Rate) at low FPR (False Positive Rate) as our metrics. For the model's instruction fine-tuning, due to hardware limitations, we maintain the overall batch size invariant through gradient accumulation, while other hyperparameters remain consistent with the original fine-tuning. The Gaussian noise hyperparameter σ is set to 0.1. For more details, please refer to Appendix §C.

4.2 Results

The comparison of our method with other baselines is shown in Tables 1 and 2, we observe the following key findings: 1) Our method outperforms all black-box access baselines across all three datasets. Particularly on the CogVLM-17b model evaluated on PixMo benchmark, PC-MMIA achieves improvements of 4.9%, 7.1%, and 6.9% in AUC, TPR@5%FPR, and TPR@1%FPR metrics respectively compared to the best baseline.

Table 1: Performance comparison of PC-MMIA and baseline methods for LLaVA-1.5-7b across different datasets. Metrics used are AUC, TPR at 5% FPR (T@5F), and TPR at 1% FPR (T@1F). **Bolded** number shows the best result within each column across all methods.

Dataset	NoCaps			Flickr30K			PixMo		
	AUC	T@5F	T@1F	AUC	T@5F	T@1F	AUC	T@5F	T@1F
Loss	0.671	0.284	0.121	0.651	0.241	0.104	0.692	0.306	0.131
Neighbor	0.634	0.205	0.089	0.624	0.235	0.097	0.613	0.230	0.062
Min-K% Prob	0.700	0.310	0.142	0.683	0.272	0.112	0.694	0.312	0.145
DC-PDD	0.668	0.302	0.104	0.652	0.265	0.093	0.701	0.309	0.110
Min-K%++ Prob	0.718	0.329	0.130	**0.707**	0.297	**0.132**	0.715	0.318	0.142
MaxRényi-K%	0.686	0.290	0.136	0.674	0.241	0.092	0.723	0.335	0.129
NormAC	0.673	0.316	0.122	0.689	0.258	0.117	0.730	0.303	0.122
PC-MMIA	**0.735**	**0.335**	**0.147**	0.703	**0.303**	0.128	**0.754**	**0.340**	**0.162**

Table 2: Performance comparison of PC-MMIA and baseline methods for CogVLM-17b across different datasets. Metrics used are AUC, TPR at 5% FPR (T@5F), and TPR at 1% FPR (T@1F). **Bolded** number shows the best result within each column across all methods.

Dataset	NoCaps			Flickr30K			PixMo		
	AUC	T@5F	T@1F	AUC	T@5F	T@1F	AUC	T@5F	T@1F
Loss	0.713	0.296	0.126	0.681	0.252	0.098	0.725	0.322	0.135
Neighbor	0.728	0.304	0.124	0.654	0.248	0.111	0.663	0.281	0.096
Min-K% Prob	0.734	0.327	0.128	0.716	0.284	0.125	0.728	0.327	0.143
DC-PDD	0.722	0.318	0.108	0.704	0.279	0.128	0.735	0.326	0.124
Min-K%++ Prob	0.731	0.344	**0.148**	0.713	0.313	**0.138**	0.744	0.331	0.137
MaxRényi-K%	0.722	0.306	0.132	0.711	0.254	0.116	0.742	0.335	0.134
NormAC	0.719	0.302	0.128	0.706	0.266	0.120	0.728	0.328	0.135
PC-MMIA	**0.741**	**0.351**	0.143	**0.717**	**0.317**	0.133	**0.781**	**0.359**	**0.153**

This can mainly be attributed to other methods' insufficient utilization of image embeddings, focusing solely on answer texts while neglecting the impact of multimodal semantics on token probability distribution. 2) The performance gap between Min-K%++ Prob and our method is smaller on the Flickr30K dataset, mainly because Min-K%++ Prob is a gray-box method requiring access to the model's logits, meaning it necessitates obtaining the prediction scores for all tokens in the model's vocabulary. 3) The Neighbor method perturbs text, and its performance is relatively poor, particularly on the PixMo dataset which exhibits a large variance in text length distribution. This also indicates that perturbations targeting pure text are difficult to apply effectively to multimodal data. 4) MaxRényi-K% focuses on Large Vision-Language Models and can be applied to both text-only and multimodal data. Nonetheless, naively applying the text-focused method to response texts in multimodal data creates a performance gap compared to our method. Additionally, membership inference attacks on CogVLM-17b show better results compared to LLaVA-1.5-7b, possibly due to CogVLM's larger model parameters, particularly those handling image embeddings, allowing for better memorization of fine-tuned data.

4.3 Ablation Study

PC-MMIA employs the following strategies to enhance the effectiveness of difficulty calibration: i) Calibrate with changes in token probability distribution rather than directly using it to distinguish members from non-members. ii) Generate symmetric samples $x^{+}_{\text{perturbed}}$ and $x^{-}_{\text{perturbed}}$ using bidirectional Gaussian noise. iii) Perform softmax normalization on the change in token log-likelihood. To further explore the effectiveness of these strate-

Table 3: Ablation study exploring the impact of different strategies on PC-MMIA performance for LLaVA-1.5-7b across different datasets. Metrics used are AUC, TPR at 5% FPR (T@5F), and TPR at 1% FPR (T@1F).

Dataset	NoCaps			Flickr30K			PixMo		
	AUC	T@5F	T@1F	AUC	T@5F	T@1F	AUC	T@5F	T@1F
PC-MMIA	0.735	0.335	0.147	0.703	0.303	0.128	0.754	0.340	0.162
direct $\boldsymbol{\delta(y)}$	0.692	0.310	0.123	0.680	0.281	0.108	0.719	0.319	0.135
only $\boldsymbol{x}^{+}_{\text{perturbed}}$	0.732	0.331	0.144	0.707	0.299	0.117	0.742	0.331	0.158
only $\boldsymbol{x}^{-}_{\text{perturbed}}$	0.724	0.328	0.132	0.692	0.295	0.115	0.756	0.335	0.155
w/o normalization	0.677	0.253	0.104	0.668	0.275	0.107	0.683	0.306	0.118

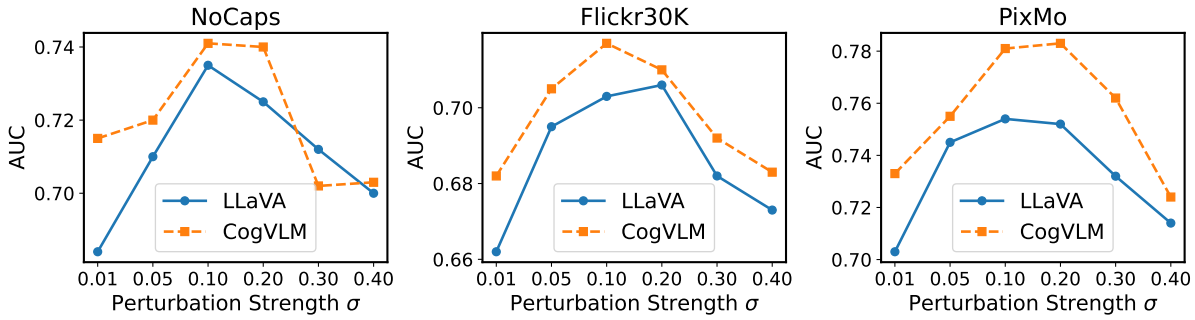

Figure 3: Impact of perturbation strength σ on PC-MMIA performance, evaluated across three datasets using both LLaVA-1.5-7b and CogVLM-17b models.

gies, we employ four variants for ablation study: **direct $\boldsymbol{\delta(y)}$**: Directly use the average change in token log-likelihood as the scoring function $\mathcal{S}$ for classification. **only $\boldsymbol{x}^{+}_{\text{perturbed}}$**: Use $x^{+}_{\text{perturbed}}$ alone to calculate the change in token log-likelihood post-perturbation. **only $\boldsymbol{x}^{-}_{\text{perturbed}}$**: Use $x^{-}_{\text{perturbed}}$ alone to calculate the change in token log-likelihood post-perturbation. **w/o normalization**: Use the change in log-likelihood directly as each token's calibration weight, without normalization.

As shown in Table 3, our method demonstrates the contribution of these three strategies to PC-MMIA. i) The performance of the direct $\delta(y)$ variant declines compared to PC-MMIA, primarily because although Gaussian noise of the same strength is added to images, the intrinsic complexity of samples means that Gaussian noise of the same strength cannot be considered as perturbations of equivalent strength on visual semantics, hindering the effectiveness of directly using $\delta(y)$ to distinguish members from non-members. ii) Normalization operations have a significant impact on the performance of PC-MMIA because we use the relative change, rather than the absolute change, of token log-likelihoods as the basis for calibration. This makes the scoring of different samples more consistent and comparable, mitigating the bias introduced by the inconsistent strength of image semantic perturbations. iii) The performance reduction in both the only $x^{+}_{\text{perturbed}}$ and only $x^{-}_{\text{perturbed}}$ variants demonstrates the effectiveness of our strategy. Generating symmetric Gaussian noise samples can be regarded as multiple sampling within the neighborhood of the data manifold to analyze the impact of equivalent perturbations on each token in the text more accurately, facilitating better calibration and capturing tokens more critical for reflecting training traces.

4.4 Analysis Study

This section explores factors possibly affecting the performance of PC-MMIA, including perturbation strength, perturbation method, and text length.

How does perturbation strength σ affect PC-MMIA? We evaluate the model's performance under different σ settings, as shown in Figure 3. Extreme σ values, either too small or too large, impact PC-MMIA performance. When σ is too small, perturbation strength is insufficient to observe poorly generalized regions in the local manifold of member samples. Conversely, when σ is too large, Gaussian noise fails to remain approximately semantically invariant, broadly reducing likelihoods for all tokens in member and non-member samples,

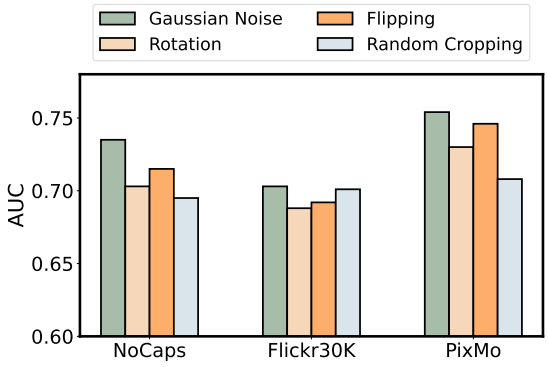

Figure 4: AUC results for various perturbation methods using the LLaVA model across different datasets.

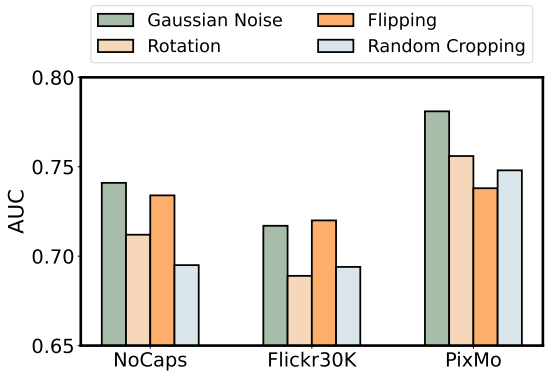

Figure 5: AUC results for various perturbation methods using the CogVLM model across different datasets.

thus failing to capture more discriminative tokens reflecting training traces.

How does different perturbation methods affect PC-MMIA? In our experiments, we use Gaussian noise as the image perturbation method. To demonstrate its effectiveness, we also apply perturbations including rotation (±15 degrees), horizontal flipping, and random cropping (retaining 80% of the original image area), then evaluate the performance of PC-MMIA. The results are shown in Figures 4 and 5. These methods perform worse than Gaussian noise and are not stable across different datasets. This is mainly because Gaussian noise does not change the spatial structure or semantic content of images but introduces mild, uniform pixel-level variations. In contrast, rotation and cropping introduce geometric transformations and partial information loss, while flipping can alter the visual orientation of asymmetric objects, thus reducing the consistency of model predictions and weakening the effectiveness of PC-MMIA.

How does the length of answer text affect PC-MMIA? We further evaluate the impact of answer text length in the fine-tuning data on the performance of PC-MMIA. Since the PixMo dataset exhibits a wide range of answer lengths, we conduct our experiments on this dataset. We divide the text lengths into four groups: 10–20 tokens, 40–50 tokens, 80–100 tokens, and 150–200 tokens, with each group containing 100 samples. As shown in Figure 6, the AUC scores exhibit a slight upward trend as the text length increases. This may be because longer texts are more likely to contain information memorized by the target model, making it easier to distinguish member texts from unseen texts. Notably, several studies on membership inference attacks against LLMs have demonstrated that text length significantly influences attack performance (Zhang et al., 2024b). In contrast, in multimodal scenarios, this impact appears to be less pronounced. This is primarily because the presence of image embeddings effectively serves as a longer prefix of tokens, compensating for the information loss caused by shorter texts.

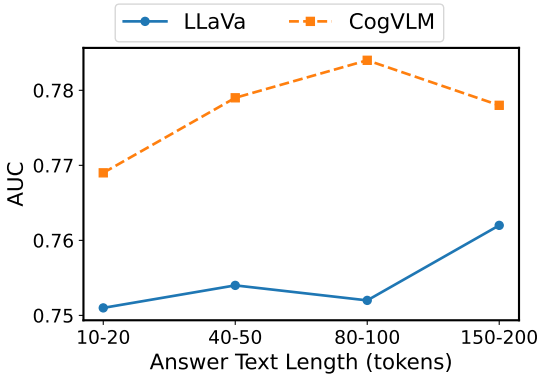

Figure 6: Impact of answer length on PC-MMIA in the PixMo dataset.

5 Conclusion

Conventional MIA overlook the intrinsic complexity of samples, prompting an increasing number of studies to introduce various difficulty calibration methods. However, these approaches focus solely on calibrating text tokens, neglecting the rich semantics brought by image embeddings. To address this, we introduced PC-MMIA, a black-box membership inference attack method specifically tailored for multimodal large language models. It captures tokens in the text that better reflect training traces through perturbation of image embeddings, assigning different weights to achieve difficulty calibration. Experiments demonstrate that PC-MMIA outperforms all black-box access baselines. We believe our work lays the groundwork for advancing MIA technology, thereby enhancing privacy preservation in multimodal large language models.

6 Limitations

Current research on membership inference attacks predominantly emphasizes empirical methodologies, with a notable scarcity of systematic theoretical analyses and rigorous proofs concerning their effectiveness. This gap is particularly evident in the context of large-scale models, where the interplay between interpretability, memorization, and generalization remains insufficiently explored.

Our proposed method also relies on certain heuristic assumptions. For instance, Section 3.2 interprets the underlying principles from the perspective of data manifolds and local smoothness within the input space. Due to the inherent complexity and opaque nature of large model training processes, it is challenging to rigorously validate these assumptions mathematically or to construct a formal theoretical derivation of the entire method. This challenge underscores the potential for future research in enhancing model interpretability and theoretical modeling.

Furthermore, the performance of all existing MIA methods on closed-source models still cannot be accurately quantified. This is because we lack access to the training data of these black-box commercial models and are unaware of their deployment details, making it impossible to conduct precise testing on these models. In the domain of membership inference attacks against multimodal large language models, we operate based on a common assumption: methods that perform better on open-source models are more likely to be effective on closed-source commercial models. Building on this reasoning established by previous studies, we conducted a series of experiments on open-source models to validate the superiority of our method.

Acknowledgements

This work was supported by Beijing Science and Technology Program (Z231100007423011) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.

Sagiv Antebi, Edan Habler, Asaf Shabtai, and Yuval Elovici. 2025. Tag&tab: Pretraining data detection in large language models using keyword-based membership inference attack. *Preprint*, arXiv:2501.08454.

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*.

Yuxing Cheng, Yi Chang, and Yuan Wu. 2025. A survey on data contamination for large language models. *Preprint*, arXiv:2502.14425.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 12039–12050. Association for Computational Linguistics.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? In *First Conference on Language Modeling*.

André V. Duarte, Xuandong Zhao, Arlindo L. Oliveira, and Lei Li. 2024. DE-COP: Detecting Copyrighted Content in Language Models Training Data. *Preprint*, arXiv:2402.09910.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. 2024. Membership inference attacks against fine-tuned large language models via self-prompt calibration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Michael M. Grynbaum and Ryan Mac. 2023. The times sues openai and microsoft over a.i. use of copyrighted work. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

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.

Yuke Hu, Zheng Li, Zhihao Liu, Yang Zhang, Zhan Qin, Kui Ren, and Chun Chen. 2025. Membership inference attacks against vision-language models. *Preprint*, arXiv:2501.18624.

Kate Knibbs. 2023. The battle over books3 could change ai forever. https://www.wired.com/story/battle-over-books3/.

Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. 2024. Membership inference attacks against large vision-language models. *Advances in Neural Information Processing Systems*, 37:98645–98674.

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

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. 2024. Llm dataset inference: Did you train on my dataset? *Advances in Neural Information Processing Systems*, 37:124069–124092.

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.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022. Quantifying privacy risks of masked language models using membership inference attacks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8332–8347, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. 2023. Proving test set contamination in black box language models. *Preprint*, arXiv:2310.17623.

Haonan Shi, Tu Ouyang, and An Wang. 2024a. Learning-Based Difficulty Calibration for Enhanced Membership Inference Attacks . In *2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P)*, pages 62–77, Los Alamitos, CA, USA. IEEE Computer Society.

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

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Dingjie Song, Sicheng Lai, Shunian Chen, Lichao Sun, and Benyou Wang. 2025. Both text and images leaked! a systematic analysis of multimodal llm data contamination. *Preprint*, arXiv:2411.03823.

Thomas Steinke, Milad Nasr, and Matthew Jagielski. 2023. Privacy auditing with one (1) training run. In *NeurIPS*.

Wei Wang, Zhaowei Li, Qi Xu, Yiqing Cai, Hang Song, Qi Qi, Ran Zhou, Zhida Huang, Tao Wang, and Li Xiao. 2024a. Qcrd: Quality-guided contrastive rationale distillation for large language models. *Preprint*, arXiv:2405.13014.

Wei Wang, Zhaowei Li, Qi Xu, Linfeng Li, YiQing Cai, Botian Jiang, Hang Song, Xingcan Hu, Pengyu Wang, and Li Xiao. 2024b. Advancing fine-grained visual understanding with multi-scale alignment in multi-modal models. *Preprint*, arXiv:2411.09691.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, Jiazheng Xu, Keqin Chen, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024c. CogVLM: Visual expert for pretrained language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

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.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024a. Benchmarking benchmark leakage in large language models. *Preprint*, arXiv:2404.18824.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024b. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.

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.

Saleh Zare Zade, Yao Qiang, Xiangyu Zhou, Hui Zhu, Mohammad Amin Roshani, Prashant Khanduri, and Dongxiao Zhu. 2025. Automatic calibration for membership inference attack on large language models. *arXiv preprint arXiv:2505.03392*.

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.

Huixuan Zhang, Yun Lin, and Xiaojun Wan. 2025a. Pacost: Paired confidence significance testing for benchmark contamination detection in large language models. *Preprint*, arXiv:2406.18326.

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

Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024b. Pretraining data detection for large language models: A divergence-based calibration method. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5263–5274, Miami, Florida, USA. Association for Computational Linguistics.

Xiaofan Zheng, Zinan Zeng, Heng Wang, Yuyang Bai, Yuhan Liu, and Minnan Luo. 2025. 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.

A Datasets

To comprehensively evaluate the robustness of PC-MMIA in various scenarios, we conducted experiments on three representative datasets: NoCaps (Agrawal et al., 2019), Flickr30k (Young et al., 2014), and PixMo (Deitke et al., 2024). PixMo is a newly released open visual instruction fine-tuning dataset from the Allen Institute for AI, featuring diverse problem settings and a wide distribution of answer text lengths. In our experiment, we select 5000 data samples from this dataset. NoCaps and Flickr30k are widely used high-quality caption datasets. We use the first caption provided for each image and convert it into instruction fine-tuning format according to the method by Liu et al. (2023). We select 5000 samples from Flickr30k for the experiment. Since the test set of NoCaps is not publicly available, we utilize the validation set of NoCaps. As shown in Table 4, each dataset is randomly divided, labeling samples as members and non-members to ensure consistency in data distribution between member and non-member samples.

B Baseline Details

Below are the baselines compared to our method. We modify some methods aimed at LLMs to make them applicable to MLLMs:

- **Loss:** (Yeom et al., 2018) The Loss method uses the loss calculated by the model on the target sample as the membership score. The intuition is that data points (members) seen during training will have lower loss values, while unseen data points (non-members) will have higher loss values. We apply Loss to the answer text in visual instruction fine-tuning data.

- **Neighbor:** (Mattern et al., 2023) The Neighbor attack baseline determines membership by comparing the target sample's loss with the average loss of its perturbed neighbors. Following Wang's setup, we generate 20 neighbor samples for each data point through keyword replacement in answer text of the visual instruction fine-tuning data.

- **Min-K% Prob:** (Shi et al., 2024b) The Min-K% baseline uses the average log-likelihood of the k% lowest probability tokens to calculate membership scores. We apply it to answer

Table 4: Details of the datasets used for the experiments. Each dataset is divided into member and non-member samples to ensure consistent data distribution.

Dataset	Total Samples	Member	Non-Member
NoCaps	4500	2250	2250
Flickr30k	5,000	2,500	2,500
PixMo	5,000	2,500	2,500

text in visual instruction fine-tuning data and set K to 20.

- **Min-K%++ Prob:** (Zhang et al., 2025b) Min-K%++ Prob is an extension of Min-K%, calibrated using the mean and standard deviation of log-likelihood for all candidate tokens. Note that this is a gray-box method because it requires access to logits. We apply it to answer text in visual instruction fine-tuning data and set K to 20.
- **DC-PDD:** (Zhang et al., 2024b) DC-PDD calibrates using text frequency from an additional text database. We apply it to answer text in visual instruction fine-tuning data and calibrate using the C4 database provided in the paper.
- **NormAC:** (Zade et al., 2025) This method mitigates high false positive rates and the need for reference models by using a tunable temperature to calibrate output probabilities, inspired by maximum likelihood estimation during pre-training.
- **MaxRényi-K%:** (Li et al., 2024) This method is based on the assumption that the model should have more confidence in member samples, introducing Rényi entropy to measure the model's confidence in answer text. We set the hyperparameter K to 10 and α to 1.

C Additional Implementation Details

To ensure the reproducibility of experiments and the reliability of MIA evaluation, we explain implementation details and evaluation metrics.

C.1 Training Setup

We conduct visual instruction fine-tuning on LLaVA-1.5-7b (Liu et al., 2023) and CogVLM-17b (Wang et al., 2024c) with fully open-sourced training data. All experiments are initialized based on the officially released pretrained checkpoint weights. On the basis of the original instruction fine-tuning data, we mix in the member samples divided in our experiments to construct the final dataset used for fine-tuning. The hyperparameter settings in the experiments remain consistent with the original fine-tuning settings, and we maintain a consistent total batch size through gradient accumulation.

C.2 Area Under the ROC Curve (AUC)

The Area Under the ROC Curve (AUC) is a widely used metric for evaluating the performance of binary classification models, including membership inference attacks. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at different decision thresholds. TPR, also known as sensitivity or recall, is the ratio of actual positive samples (i.e., member samples) correctly identified as positive. FPR is the ratio of actual negative samples (i.e., non-member samples) wrongly identified as positive. AUC ranges from 0 to 1, where a value of 0.5 represents a random classifier and a value of 1 signifies a perfect classifier. In the context of membership inference attacks, a higher AUC indicates the attack can better distinguish member and non-member samples across all possible decision thresholds.

C.3 True Positive Rate at Low False Positive Rate (TPR@low%FPR)

While AUC provides an overall measure of membership inference attack performance, it may not be the most suitable metric in practical applications. In many situations, the cost of false positives (i.e., wrongly identifying non-member samples as members) can be significantly higher than the cost of false negatives (i.e., wrongly identifying member samples as non-members). Particularly in real-world scenarios, the proportion of member and non-member samples is not the same, with non-member samples being much more prevalent. This metric provides a more stringent evaluation of membership inference attack performance, emphasizing its ability to correctly identify member samples while

Table 5: Performance comparison of PC-MMIA and baselines on VizWiz-VQA.

Method	AUC	TPR@5F	TPR@1F
Loss	0.648	0.289	0.115
Min-K% Prob	0.662	0.297	0.139
DC-PDD	0.673	0.305	0.128
Min-K%++ Prob	0.721	0.332	0.134
MaxRényi-K%	0.693	0.309	0.122
PC-MMIA	**0.739**	**0.342**	**0.145**

maintaining a low false positive rate.

D Additional Experiments

To further validate the effectiveness of PC-MMIA, we conducted experiments on the VizWiz-VQA dataset using the LLaVA-1.5-7b model. Table 5 shows that PC-MMIA achieves the best performance compared to baseline methods, with an AUC of 0.739 and improvements in TPR@5F and TPR@1F by 3.0% and 2.2%, respectively, over the strongest baseline (Min-K%++ Prob). This highlights its robustness and broad applicability to diverse datasets.

E Ethics Statement

Our primary aim in developing the PC-MMIA method is to enhance the understanding and detection of privacy leakage risks associated with MLLMs, particularly those involving unauthorized use or disclosure of sensitive information. We acknowledge that the techniques designed for identifying membership could potentially be misused if applied maliciously, leading to privacy infringements or exploitation of proprietary data. Therefore, it is crucial to apply our research responsibly and ethically, ensuring it serves to strengthen data protection mechanisms and safeguard against privacy violations.

All experiments in this work are conducted on publicly available or synthetic datasets that do not contain personally identifiable information. No human subjects were involved in data collection or annotation, and no private data from proprietary models was used. We encourage future researchers to further explore membership inference attacks under appropriate ethical safeguards.