

Deconstructing Memorization and Generalization: Dual-Level Calibration of Membership Inference Attacks for Large Vision-Language Models

Xiaofan Zheng^{1,2}, Xinghao Wang¹, Minnan Luo¹ *

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

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

Abstract

Large Vision-Language Models (LVLMs) achieve remarkable progress but raise privacy concerns due to vendors' secretive data practices. Recent studies explore Membership Inference Attacks (MIA) for privacy auditing, which assume models exhibit lower loss and higher confidence for training data. However, this assumption overlooks the inherent differences in difficulty among samples. A model's generalization-driven predictions on simple samples and its verbatim memorization of complex ones can both yield high confidence scores, thereby blurring the distinction between members and non-members. Furthermore, the prohibitive training cost of LVLMs renders the conventional approach of training additional shadow models for difficulty calibration impractical. In response to the aforementioned challenges, we propose Dual-Level Calibration of Membership Inference Attacks (DLC-MIA). Our method performs dual-level calibration at both the token level and the sample level, and introduces a lightweight auxiliary model to simulate the training process of samples, leveraging the prior knowledge of the auxiliary model to distinguish between generalization and memorization of the target model. Specifically, at the first calibration level, we utilize the auxiliary model to identify tokens within a sample that are more sensitive to training and assign them differential weights, as these tokens more effectively reflect traces of memorization. At the second calibration level, we achieve difficulty calibration by estimating the degree of generalization through comparing the flatness of the loss landscape between the auxiliary model and the target model. Extensive experiments validate the effectiveness of our proposed method.

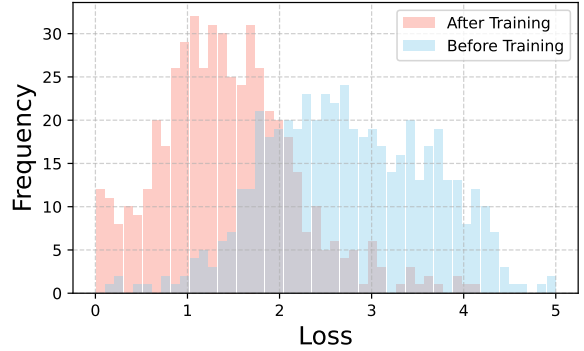


Figure 1: Loss distribution of the same dataset before and after training on the model. Inherent difficulty differences among samples blur the boundary between members and non-members.

1 Introduction

The emergence of representative Large Vision-Language Models (LVLMs), such as GPT-4o and Gemini (OpenAI, 2024; Team, 2025), signifies a leap forward in multimodal intelligence. By mapping visual feature embeddings into the textual semantic space, these models gain the ability to deeply comprehend and reason about the joint information flow of images and text. Consequently, they demonstrate impressive performance on complex cross-modal tasks like image captioning and visual question answering (Wang et al., 2024b).

The superior performance of LVLMs is largely attributed to the rich knowledge sourced from internet-scale multimodal datasets (Wang et al., 2024a; Zhang et al., 2024a). While these datasets provide diverse learning materials, they also give rise to serious privacy and data rights issues. The training data may contain unauthorized sensitive personal images or copyrighted visual content (Grynbaum and Mac, 2023; Knibbs, 2023). This situation not only risks violating relevant regulations but could also lead to the unintentional disclosure of training data, resulting in incalculable

* Corresponding author: Minnan Luo

harm to the data owners (Song et al., 2017).

In this context, Membership Inference Attacks (MIA) have garnered significant attention as a privacy auditing technique (Shokri et al., 2017; DeAlcala et al., 2025). The primary objective of an MIA is to determine whether a specific data sample was part of a model’s training process, thereby providing data subjects with a means to verify the usage of their information (Hu et al., 2022; Wu and Cao, 2025). Notably, the fine-tuning stage of LVLMs is often considered the most vulnerable phase for privacy leakage (Li et al., 2024; Zheng et al., 2025). This stage typically utilizes smaller, yet more sensitive, private datasets for visual instruction tuning, which makes research on membership inference for fine-tuning data particularly crucial (Song et al., 2025). By developing effective MIA methods, we can detect privacy vulnerabilities and unauthorized data usage in models, thus providing technical support for data protection.

MIA is typically categorized into reference-based and reference-free attacks (Hu et al., 2022). The initial state-of-the-art MIA methods for deep neural networks required training additional "shadow models" for reference (Wu and Cao, 2025). This reliance on shadow models restricts their application on large-scale models, rendering reference-based approaches prohibitively expensive and impractical for attacking LLMs and LVLMs (Fu et al., 2024). Therefore, recent attacks targeting LVLMs primarily employ reference-free strategies. Their basic premise is that LVLMs exhibit lower loss and higher prediction confidence for member data. However, these methods overlook the inherent difficulty differences among samples (Watson et al., 2022). For example, as shown in Figure 1, simple non-member samples may also achieve high prediction confidence due to the model’s multimodal generalization capability, which blurs the boundary between members and non-members, leading to a high false positive rate (Zhao et al., 2025; Shi et al., 2024a; Zheng et al., 2025; Wu and Cao, 2025).

To address these issues, we attempt to calibrate membership inference signals to eliminate the confusion caused by the inherent difficulty of different samples (Watson et al., 2022; Shi et al., 2024a; Ye et al., 2024). Inspired by Sharpness-Aware Minimization (SAM) (Foret et al., 2021; Ma and Ying, 2021), we propose a dual-level calibration method for MIA, named DLC-MIA. This approach introduces a lightweight auxiliary model, which differs from the shadow models in conventional

methods, to perform a dual-level calibration at both the token and sample levels. This process effectively distinguishes between the model’s predictions driven by generalization and its verbatim memorization (Huang et al., 2024; Zhang, 2025).

In the token-level calibration, our core insight is that different tokens within a sample carry MIA signals of varying strengths (Chang et al., 2025; Tran et al., 2025). We observe that when a specific sample is randomly mixed into the model’s original training data for training, only a small portion of its tokens show a distinct increase in prediction probability. In contrast, most tokens exhibit only minor fluctuations, containing almost no useful MIA signals. Therefore, the loss value for an entire sample does not effectively characterize the model’s degree of memorization for that sample (Antebi et al., 2025; Zheng et al., 2025). We simulate the training process of a sample by fine-tuning it on the auxiliary model, which allows us to identify key tokens and assign them differential weights according to their importance. In the sample-level calibration, we aim to estimate the model’s overall generalization ability for a given sample. Prior studies on SAM show that the flatness of the loss landscape with respect to model parameters can indicate the model’s generalization performance on that sample (Foret et al., 2021). In a black-box setting where model parameters are not accessible, we can approximate this landscape by observing the flatness of the model’s function landscape with respect to its input data (Ma and Ying, 2021), which in turn allows us to assess generalization. Specifically, we apply Gaussian noise perturbations to the input image and observe the relative flatness of the model function. A comparison with the auxiliary model further mitigates uncertainty introduced by inconsistent perturbation strengths, thereby enabling the calibration of the overall sample difficulty.

2 Related Work

MIA research for LLMs (Steinke et al., 2023; Xu et al., 2024a; Mireshghallah et al., 2022a; Duarte et al., 2024; Xu et al., 2024b; Xie et al., 2025; Dubiński et al., 2024; He et al., 2025) categorizes methods by access level: white-box, grey-box, and black-box (Cheng et al., 2025). Recent work focuses on grey-box and black-box scenarios (Wu and Cao, 2025), with methods like Min-K% (Shi et al., 2024b) and Neighbor (Mattern et al., 2023). However, these overlook sample difficulty differ-

ences (Watson et al., 2022), leading to high false positive rates (Wu and Cao, 2025). While shadow models can mitigate this (Watson et al., 2022; Rezaei and Liu, 2021), they are impractical for LLMs and LVLMs (Fu et al., 2024; Li et al., 2024). Recent calibration methods for LLMs (Zhang et al., 2024c; Antebi et al., 2025; Zhang et al., 2025b) only consider text difficulty (Zheng et al., 2025; Song et al., 2025), performing poorly on LVLMs due to image input impact (Zheng et al., 2025).

MIA research for LVLMs remains exploratory, with early work on CLIP (Ko et al., 2023; Samira et al., 2025) and MaxRényi-K% (Li et al., 2024). Most relevant are PC-MMIA (Zheng et al., 2025) and ICIMIA (Wu et al., 2025), which use perturbations on image embeddings and corruption sensitivity. Nevertheless, these methods insufficiently consider multimodal sample complexity, equating identical perturbation operations with uniform perturbation strength while neglecting varying sample sensitivities.

3 Threat Model

We primarily investigate Membership Inference Attacks against Large Vision-Language Models. LVLMs achieve alignment between images and text by mapping image embeddings into the same representation space as text embeddings, and demonstrate outstanding performance in various multimodal tasks. Below is a summary of the attacker’s goal and their knowledge assumptions.

3.1 Attacker’s Goal

For a data instance z and an LVLM $\mathcal{M}$ that is instruction-tuned on a dataset $\mathcal{D}$, the goal of MIA is to determine whether the data instance z belongs to the training dataset $\mathcal{D}$. Formally, a data instance is defined as $z = (x_{\text{img}}, x_{\text{text}}, y_{\text{text}})$, where x_{img} represents the input image, x_{text} denotes the input text, and y_{text} is the corresponding response. To ascertain whether z is a member of $\mathcal{D}$, an attacker calculates a membership score $\mathcal{S}(z; \mathcal{M})$ and compares it with a threshold τ , thereby making a judgment on the data instance z . The key to executing a membership inference attack lies in designing an effective scoring function $\mathcal{S}$ to more accurately distinguish member data from non-member data. A membership inference attack can be formalized as the following decision function:

$$\mathcal{A}(z, \mathcal{M}) = \mathbb{1}[\mathcal{S}(z; \mathcal{M}) \geq \tau], \quad (1)$$

where $\mathcal{A}$ denotes the attack algorithm, $\mathcal{S}(z; \mathcal{M})$ is the membership score calculated for data instance z and model $\mathcal{M}$, τ is the decision threshold, and $\mathbb{1}$ is the indicator function, which outputs 1 if $\mathcal{S}(z; \mathcal{M}) \geq \tau$, indicating that z is classified as a member.

3.2 Attacker’s Knowledge

Following prior research (Zheng et al., 2025; Song et al., 2025), we focus on membership inference attacks during the instruction fine-tuning stage under the most challenging black-box access scenario. This implies that the attacker can only access the probability of the corresponding text tokens and cannot obtain the probabilities for all tokens in the entire vocabulary, nor can they access model weights and parameters. The attacker also has no knowledge of the model’s training data distribution. The attacker can, however, determine the model’s training data cutoff date from its release time. For this reason, internet datasets collected after the model’s release date can be considered as non-member data. The attacker can easily obtain this non-member data for reference, for example, to determine a suitable threshold τ .

4 Methodology

Our method aims to construct a more robust scoring function for membership inference attacks to effectively distinguish between a model’s memorization behavior and its generalization ability. The fundamental reason why existing loss-based methods perform poorly is that a target model $\mathcal{M}$ ’s high-confidence prediction for a sample can stem from either the verbatim memorization of difficult member data or the generalization-driven prediction on simple non-member data. These two signals are conflated, leading to a blurred boundary between members and non-members. In this section, we systematically elaborate on the framework we construct from the three dimensions of an auxiliary model, token-level calibration, and sample-level calibration to achieve this goal.

4.1 Auxiliary Model

We introduce an auxiliary model $\mathcal{M}_{\text{aux}}$ in our calibration, which is fundamentally different from the shadow models in conventional methods. We impose no strict requirements on the architecture or the training data distribution of the auxiliary model. The only constraint is that the data samples z to be tested are not in the auxiliary model’s training

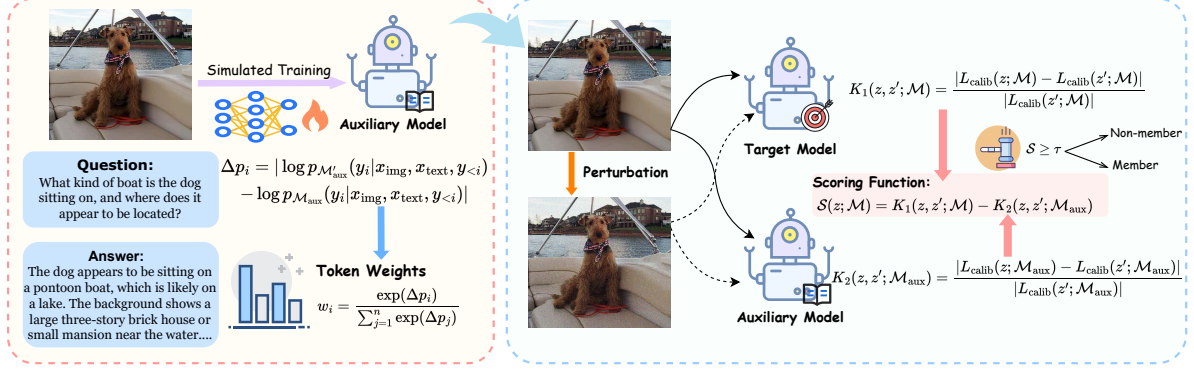


Figure 2: Overview of our proposed dual-level calibration approach for membership inference attacks, which consists of token-level calibration and sample-level calibration.

data, a condition that is easy to satisfy. We first simulate the training process of the test data on the auxiliary model $\mathcal{M}_{aux}$. Although there may be a distribution shift between the training data of the auxiliary and target models, we find that by comparing the probability changes of each token before and after fine-tuning the auxiliary model on the data, we can effectively identify the key tokens that carry more memorization signals. In addition, when perturbing images x_{img} , the same operation does not imply the same perturbation strength in the input space. For instance, when Gaussian noise of the same strength is applied to different images, images with more complex visual information may experience greater semantic changes, while images with simpler visual content may be more robust to such noise. This leads to variations in perturbation strength within the data manifold. In this case, the auxiliary model $\mathcal{M}_{aux}$ helps us mitigate the impact of inconsistent perturbation strengths, making the sample-level calibration more robust and accurate.

4.2 Token-Level Calibration

In token-level calibration, our core motivation stems from the observation that memorization signals within a sample z are not uniformly distributed. When designing the scoring function $\mathcal{S}$, previous methods either treat all tokens as equally important and compute their average loss, or, as in the work by Shi et al. (2024b), select a subset of tokens to calculate their average prediction probability. Our experiments reveal that after a model is trained on a specific sample, only a small fraction of tokens show a distinct increase in prediction probability. This indicates that there are notable differences in the strength of the memorization signal carried by different tokens. Therefore, we aim to achieve a

more precise calibration by assigning differential weights to tokens based on their importance when computing the scoring function $\mathcal{S}$.

After fine-tuning the auxiliary model $\mathcal{M}_{aux}$ with all data samples to obtain an updated model $\mathcal{M}'_{aux}$, we identify the key tokens that reflect more memorization signals by comparing the change in prediction probability for each token before and after training. Specifically, for a data instance $z = (x_{img}, x_{text}, y_{text})$, where the output text y_{text} contains n tokens, denoted as $y_{text} = \{y_1, y_2, \dots, y_n\}$, the log-probability change for each token can be expressed by the following formula:

$$\Delta p_i = |\log p_{\mathcal{M}'_{aux}}(y_i | x_{img}, x_{text}, y_{<i}) - \log p_{\mathcal{M}_{aux}}(y_i | x_{img}, x_{text}, y_{<i})|, \quad (2)$$

where $y_{<i}$ represents all tokens preceding y_i in the sequence. To make these probability changes comparable across the board, we apply a softmax normalization operation to derive the weight w_i for each token:

$$w_i = \frac{\exp(\Delta p_i)}{\sum_{j=1}^n \exp(\Delta p_j)}, \quad (3)$$

Thus, the magnitude of the weight w_i reflects the importance of token y_i in the calibration process. Subsequently, we obtain the prediction probability of the sample from the target model $\mathcal{M}$, $p_{\mathcal{M}}(y_i | x_{img}, x_{text}, y_{<i})$, and apply the normalized weights to generate the sample's calibrated log-likelihood score:

$$L_{calib}(z; \mathcal{M}) = \sum_{i=1}^n w_i \cdot \log p_{\mathcal{M}}(y_i | x_{img}, x_{text}, y_{<i}). \quad (4)$$

This calibration process captures the differences in token probability changes through simulated training on the auxiliary model, thereby providing a

more accurate representation of the model’s degree of memorization for the sample.

4.3 Sample-Level Calibration

In sample-level calibration, our core motivation is to approximate the model’s generalization ability by observing the flatness of the model function landscape with respect to the input data. Prior research (Foret et al., 2021) has demonstrated that the flatness of the loss landscape with respect to model parameters can reflect the model’s generalization ability, and that there exists a connection between the flatness of the model function landscape with respect to input data and the flatness of the loss landscape with respect to parameters (Ma and Ying, 2021). Inspired by the study of flat minima in sharpness-aware minimization (Foret et al., 2021), we view the input space of LVLMs as a data manifold, where well-generalized data points should correspond to smooth regions in the input space. Based on this assumption, we apply perturbations to the image to observe the local smoothness of the input space, thus approximating the model’s generalization ability. First, we apply Gaussian noise to the input image to obtain a perturbed sample $z' = (x'_{\text{img}}, x_{\text{text}}, y_{\text{text}})$, where x'_{img} is the result of perturbing the original image x_{img} . The original sample z and the perturbed sample z' are fed into the target model $\mathcal{M}$ and the auxiliary model $\mathcal{M}_{\text{aux}}$. We then calculate their calibrated log-likelihood scores, $L_{\text{calib}}(z; \cdot)$ and $L_{\text{calib}}(z'; \cdot)$, using the previously described token-level calibration method. For the target model $\mathcal{M}$, the relative change in score is defined as:

$$K_1(z, z'; \mathcal{M}) = \frac{|L_{\text{calib}}(z; \mathcal{M}) - L_{\text{calib}}(z'; \mathcal{M})|}{|L_{\text{calib}}(z'; \mathcal{M})|}. \quad (5)$$

To prevent the denominator from approaching 0, a small constant ϵ is included. Similarly, a comparable relative change is calculated for the auxiliary model $\mathcal{M}_{\text{aux}}$:

$$K_2(z, z'; \mathcal{M}_{\text{aux}}) = \frac{|L_{\text{calib}}(z; \mathcal{M}_{\text{aux}}) - L_{\text{calib}}(z'; \mathcal{M}_{\text{aux}})|}{|L_{\text{calib}}(z'; \mathcal{M}_{\text{aux}})|}. \quad (6)$$

Here, we use the relative changes K_1 and K_2 to measure the local smoothness of the loss landscape in the input space for the target model $\mathcal{M}$ and the auxiliary model $\mathcal{M}_{\text{aux}}$, respectively. Since the target model contains both memorization and generalization signals, while the auxiliary model, which has not been trained on the sample z , predominantly reflects generalization signals, we can use

the auxiliary model for calibration to partly decouple the target model’s memorization and generalization capabilities (Dong et al., 2024). The final scoring function $\mathcal{S}$ is defined as:

$$\mathcal{S}(z; \mathcal{M}) = K_1(z, z'; \mathcal{M}) - K_2(z, z'; \mathcal{M}_{\text{aux}}). \quad (7)$$

When $\mathcal{S}(z; \mathcal{M}) \geq \tau$, we can determine that z is a member of the training data; otherwise, it is classified as a non-member.

5 Experiments

5.1 Experimental Setup

Datasets. To validate the effectiveness of our method, we need to conduct experiments on labeled datasets. Prior research employs a temporal cutoff as the partitioning criterion, dividing member and non-member data based on dataset release times (Li et al., 2024). Nevertheless, multiple recent studies have pointed out that this time-based partitioning approach is unreliable, as data collected at different times often exhibit **distribution bias** (Maini et al., 2024; Duan et al., 2024; Miyamoto et al., 2025). Therefore, we adopt the experimental setup from recent work (Duan et al., 2024; Zheng et al., 2025) and randomly partition the member and non-member data. Specifically, we select portions from the Flickr30K (Young et al., 2014) and PixMo (Deitke et al., 2024) datasets that are not included in the model’s original training data as our experimental datasets. We sample 1,000 data points from each dataset and evenly partition them into member and non-member sets. Subsequently, we mix the member set into the model’s original fine-tuning dataset and perform visual instruction tuning on the pre-trained original model. More details regarding the datasets are provided in Appendix A.

Target Models. The experiment requires ensuring that the selected datasets are not included in the model’s original training data, which limits us to selecting models that fully open-source their training data and training details to validate the effectiveness of our method. Referring to prior research (Li et al., 2024), we select LLaMA-Adapter (Zhang et al., 2024b), MiniGPT-4 (Zhu et al., 2023), and LLaVA-1.5 (Liu et al., 2023) as target models for our experiments. These models have different architectures and training data distributions, which allows us to comprehensively evaluate the generalization ability and robustness of DLC-MIA under

Table 1: Performance comparison of DLC-MIA and baseline methods on the Flickr30K dataset. We report AUC, TPR at 5% FPR (T@5%F), and TPR at 1% FPR (T@1%F). **Bold** indicates the best performance among all methods.

Method	LLaMA-Adapter			MiniGPT-4			LLaVA-1.5		
	AUC	T@5%F	T@1%F	AUC	T@5%F	T@1%F	AUC	T@5%F	T@1%F
Loss (Yeom et al., 2018)	0.612	0.203	0.081	0.631	0.241	0.094	0.641	0.257	0.102
Neighbor (Mattern et al., 2023)	0.588	0.160	0.069	0.603	0.188	0.071	0.613	0.198	0.079
Min-K% Prob (Shi et al., 2024b)	0.654	0.268	0.112	0.671	0.294	0.128	0.686	0.310	0.145
DC-PDD (Zhang et al., 2024c)	0.605	0.191	0.072	0.624	0.220	0.083	0.633	0.231	0.091
Min-K++ Prob (Zhang et al., 2025b)	0.705	0.366	0.195	0.728	0.418	0.230	0.742	0.450	0.261
MaxRényi-K% (Li et al., 2024)	0.668	0.292	0.124	0.683	0.324	0.141	0.701	0.346	0.168
ICIMIA (Wu et al., 2025)	0.690	0.334	0.170	0.715	0.390	0.208	0.723	0.412	0.227
PC-MMIA (Zheng et al., 2025)	0.711	0.380	0.203	0.732	0.425	0.241	0.738	0.453	0.270
DLC-MIA (Ours)	0.745	0.427	0.258	0.734	0.461	0.284	0.744	0.459	0.301

different model configurations. Additional experiments on other architectures are provided in Appendix G.

Evaluation Metrics. Following most existing work, we adopt the AUC score (Area Under the ROC Curve) and the True Positive Rate (TPR) at low False Positive Rate (FPR) as our evaluation metrics (Shi et al., 2024b; Zhang et al., 2024c, 2025b; Rezaei and Liu, 2021). The AUC score provides a comprehensive evaluation of the attack model’s overall discriminative ability across all thresholds, while TPR at low FPR better reflects the effectiveness of an attack in practical auditing scenarios that demand extremely low false alarm rates. For a detailed introduction to these evaluation metrics, please refer to Appendix B.

Baseline Methods. At present, membership inference attacks against LVLMs are still in an exploratory stage. We select MaxRényi-K% (Li et al., 2024), PC-MMIA (Zheng et al., 2025), and ICIMIA (Wu et al., 2025) as baseline methods, as they are currently the only black-box membership inference attacks that can be directly applied to visual instruction tuning data. Additionally, we also select five methods targeting LLMs: Loss (Yeom et al., 2018), Neighbor (Mattern et al., 2023), Min-K% Prob (Shi et al., 2024b), DC-PDD (Zhang et al., 2024c), and Min-K%++ Prob (Zhang et al., 2025b). Referring to previous work, we appropriately modify these LLM methods to enable their application to LVLMs. More details about the baseline methods are provided in Appendix C.

Implementation Details. We select nanoVLM-222M (Wiedmann et al., 2025) as our auxiliary model, which has a small parameter size, is easy to train, and provides open-source training data. Since the auxiliary model and the target model may

employ different tokenization strategies, when a single word is split into multiple tokens, we merge their probabilities into the probability of the entire word. For sample-level calibration, we apply Gaussian noise with $\sigma = 0.15$ as the image perturbation method. During the instruction fine-tuning stage for both the target model and the auxiliary model, we adopt their respective original training hyperparameters to simulate the most realistic attack scenario. Detailed computational cost analysis and additional implementation details are provided in Appendices F and D, respectively.

5.2 Overall Performance

We compare the DLC-MIA method with existing baselines across the Flickr30K and PixMo datasets. The results are shown in Table 1 and Table 2. DLC-MIA outperforms all baseline models on the key metric of TPR at low FPR. In practical privacy auditing scenarios, the cost of false positives is far higher than that of false negatives, and most samples are non-members. Consequently, TPR at low FPR is considered a more suitable metric for measuring MIA methods, which reflects the superiority and reliability of our method in real-world applications. In particular, on the PixMo dataset with MiniGPT-4 as the target model, DLC-MIA achieves improvements of 5.5% and 5.7% in AUC and TPR@5%FPR, respectively, compared to the best baseline. These results demonstrate that our proposed dual-level calibration method can more effectively distinguish whether the high prediction probability of a model for a sample stems from generalization ability or verbatim memorization, thereby effectively mitigating the impact of inherent difficulty differences among samples.

Min-K%++ Prob and PC-MMIA are the two best-performing baseline models, both of which

Table 2: Performance comparison of DLC-MIA and baseline methods on the PixMo dataset. We report AUC, TPR at 5% FPR (T@5%F), and TPR at 1% FPR (T@1%F). **Bold** indicates the best performance among all methods.

Method	LLaMA-Adapter			MiniGPT-4			LLaVA-1.5		
	AUC	T@5%F	T@1%F	AUC	T@5%F	T@1%F	AUC	T@5%F	T@1%F
Loss (Yeom et al., 2018)	0.631	0.228	0.092	0.648	0.262	0.108	0.656	0.273	0.115
Neighbor (Mattern et al., 2023)	0.639	0.253	0.224	0.647	0.233	0.101	0.664	0.279	0.142
Min-K% Prob (Shi et al., 2024b)	0.670	0.301	0.138	0.694	0.341	0.165	0.712	0.361	0.183
DC-PDD (Zhang et al., 2024c)	0.622	0.211	0.083	0.639	0.242	0.099	0.647	0.253	0.106
Min-K++ Prob (Zhang et al., 2025b)	0.721	0.405	0.225	0.736	0.455	0.260	0.758	0.490	0.295
MaxRényi-K% (Li et al., 2024)	0.684	0.328	0.159	0.708	0.373	0.188	0.724	0.397	0.210
ICIMIA (Wu et al., 2025)	0.710	0.380	0.207	0.732	0.432	0.246	0.749	0.471	0.281
PC-MMIA (Zheng et al., 2025)	0.728	0.420	0.241	0.725	0.441	0.280	0.763	0.505	0.310
DLC-MIA (Ours)	0.769	0.473	0.284	0.791	0.512	0.319	0.783	0.518	0.352

focus on difficulty calibration. Between them, Min-K%++ Prob is a grey-box method that requires access to the model’s internal probability distribution, which limits its application scenarios. Although PC-MMIA performs simple difficulty calibration at the token level through perturbations, it entirely attributes the changes in model prediction probabilities after image perturbation to the sample’s membership status, ignoring the fact that different samples may inherently have varying sensitivities to perturbations, even when using the same perturbation method. This also indicates that equating the same perturbation operation (e.g., adding Gaussian noise of the same strength to images in PC-MMIA) with an identical perturbation strength in the input space is unreliable.

Furthermore, we observe that MIA methods for LLMs generally perform poorly when transferred to LVLMs. Interestingly, DC-PDD, which performs well as a baseline model for LLMs, even performs worse than the most trivial Loss method when transferred to LVLMs. This is primarily because DC-PDD uses word frequency in text for calibration, considering only linguistic prior knowledge while failing to account for the impact of the visual modality input on prediction probabilities, and thus cannot effectively handle the complexity of multimodal inputs. Meanwhile, the DC-PDD method appears to be sensitive to the word frequency corpus, and a more suitable corpus should be sought when transferring it to LVLMs. The Neighbor method attempts to achieve perturbation by replacing text with synonymous words, but its performance is subpar. This may be due to the discrete nature of text, which makes it difficult to ensure the consistency and controllability of perturbation strength on text. Different samples may exhibit drastically different degrees of semantic

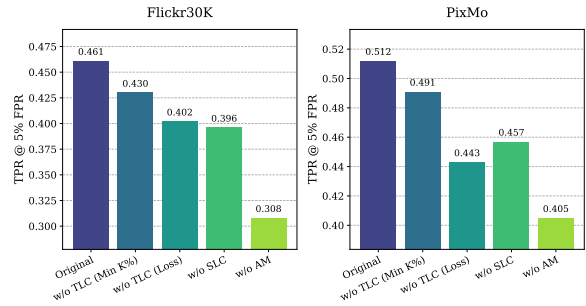


Figure 3: Ablation study results of our proposed DLC-MIA on the MiniGPT-4 model across the Flickr30K and PixMo datasets.

shift for the same proportion of word replacement.

5.3 Ablation Study

To validate the effectiveness of each key component in DLC-MIA, we design the following three ablation studies to investigate the contributions of token-level calibration, sample-level calibration, and the introduction of the auxiliary model to the final attack performance: (i) Removing Token-Level Calibration (*w/o TLC*): In this variant, we remove the token-level calibration and directly use the Loss value of the entire sample or Min-K% as the calibrated log-likelihood score $L_{\text{calib}}(z; \mathcal{M})$ from the first stage. (ii) Removing Sample-Level Calibration (*w/o SLC*): In this setting, we remove the sample-level calibration and directly use $L_{\text{calib}}(z; \mathcal{M})$ as the final scoring function $\mathcal{S}$ for membership determination. (iii) Removing the Auxiliary Model (*w/o AM*): In this variant, we completely remove the auxiliary model. For token-level calibration, we set all w_i to 1. For sample-level calibration, we directly use $K_1(z, z'; \mathcal{M})$ as the final scoring function $\mathcal{S}$.

The experimental results are shown in Figure 3. The performance degradation of the *w/o AM* variant is the most substantial, which indicates that

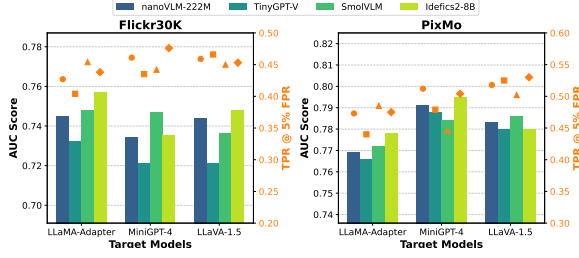


Figure 4: Performance of DLC-MIA with different auxiliary models on Flickr30K and PixMo datasets.

the auxiliary model plays a crucial role in the dual-level calibration. It provides prior knowledge and generalized multimodal capabilities as a reference baseline, enabling us to more accurately focus on the target model’s memorization signals for training data. In the experiment with the *w/o TLC* variant, Min-K% performs better than Loss. This is because Min-K% is, in essence, a simple form of token calibration; it sets the weights of a fixed proportion of high-probability tokens to 0 and then calculates the average probability of the remaining tokens. This approach shows some improvement over using the raw Loss but still has a gap compared to our calibration method, which assigns differential weights to tokens based on their training sensitivity. Furthermore, we observe that sample-level calibration is more crucial than token-level calibration. This suggests that distinguishing generalization from memorization by observing the flatness of the loss landscape is a key step in membership inference attacks, as it more effectively captures the fundamental behavioral differences of the model on member versus non-member samples.

5.4 Sensitivity Analysis

To gain a deeper understanding of the robustness and adaptability of the DLC-MIA method in various scenarios, we design the following analytical experiments:

Impact of the Auxiliary Model on DLC-MIA. Different auxiliary models may be trained with different architectures and data. Therefore, we also select TinyGPT-V (Yuan et al., 2024), SmolVLM (Marafioti et al., 2025), and Idefics2-8B (Laurençon et al., 2024) as auxiliary models for our experiments, as they feature distinct model architectures, parameter scales, and training data distributions. The experimental results, shown in Figure 4, reveal that the performance of DLC-MIA exhibits only minor fluctuations when different auxiliary models are used. This indicates that our

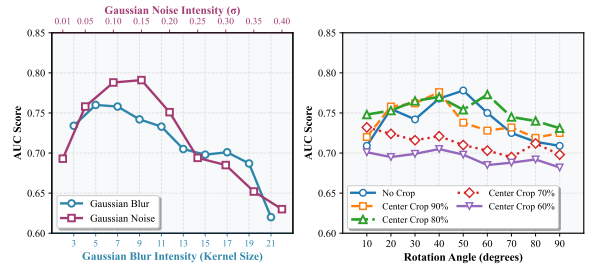


Figure 5: Performance of DLC-MIA with different perturbation methods on the PixMo using MiniGPT-4.

method does not rely on a specific auxiliary model configuration but instead leverages the general multimodal generalization ability of an auxiliary model as a calibration reference. This endows DLC-MIA with good flexibility and scalability in practical applications.

Impact of Perturbation Methods on DLC-MIA.

We conduct systematic experiments on the perturbation methods used in sample-level calibration, including four common methods: Gaussian blur, Gaussian noise, center crop, and rotation. The results, presented in Figure 5, show that different perturbation methods have a certain impact on the performance of DLC-MIA, with Gaussian noise performing relatively well. This is likely because Gaussian noise with appropriate strength can preserve the core semantic information of the image while generating sufficient perturbation strength to reflect the differences in local smoothness among various samples in the input space. Therefore, in our experiments, we choose Gaussian noise with $\sigma = 0.15$ to perturb the images. It is worth noting that various methods from image data augmentation can be integrated into the DLC-MIA framework. These methods offer multiple combination possibilities and may yield even better results, and we provide additional experiments in Appendix E.

6 Conclusion

This paper proposes DLC-MIA, which introduces a lightweight auxiliary model to effectively distinguish between a model’s generalization-driven predictions and its verbatim memorization through dual-level difficulty calibration at both the token level and the sample level. We demonstrated the effectiveness of our method through extensive experiments. Our work contributes to enhancing the transparency and security of artificial intelligence models and provides practical technical tools for privacy auditing, data contamination detection, and copyright protection.

7 Limitations

Our work has several limitations that warrant discussion. First, like most MIA research, our method relies on empirical observations and heuristic assumptions rather than rigorous theoretical foundations. For instance, the sample-level calibration is motivated by the connection between input space smoothness and generalization, but formally proving this relationship for large-scale multimodal models remains challenging due to their inherent complexity and opacity. Second, evaluating MIA performance on closed-source commercial models is inherently difficult, as their training data and deployment details are inaccessible. Following common practice in the field, we conduct experiments on open-source models where ground truth membership labels can be established, operating under the assumption that methods performing well on open-source models are likely to be effective on proprietary systems as well. Future work could explore stronger theoretical guarantees and develop evaluation protocols that better approximate real-world closed-source scenarios.

Acknowledgment

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

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

References

- 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, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022a. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE.
- Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. 2022b. The privacy onion effect: Memorization is relative. *Advances in Neural Information Processing Systems*, 35:13263–13276.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.
- Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270.
- Hongyan Chang, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, and Reza Shokri. 2025. [Context-aware membership inference attacks against pre-trained large language models](#). *Preprint*, arXiv:2409.13745.
- Yuxing Cheng, Yi Chang, and Yuan Wu. 2025. [A survey on data contamination for large language models](#). *Preprint*, arXiv:2502.14425.
- Daniel DeAlcala, Aythami Morales, Julian Fierrez, Gonzalo Mancera, Ruben Tolosana, and Javier Ortega-Garcia. 2025. [Active membership inference test \(amint\): Enhancing model auditability with multi-task learning](#). *Preprint*, arXiv:2509.07879.
- 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 Miresghallah, 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.
- Jan Dubiński, Antoni Kowalczyk, Stanisław Pawlak, Przemysław Rokita, Tomasz Trzcinski, and Paweł Morawiecki. 2024. [Towards more realistic membership inference attacks on large diffusion models](#). In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4848–4857.
- Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2021. [Sharpness-aware minimization for efficiently improving generalization](#). In *International Conference on Learning Representations*.
- 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>.
- Yu He, Boheng Li, Liu Liu, Zhongjie Ba, Wei Dong, Yiming Li, Zhan Qin, Kui Ren, and Chun Chen. 2025. [Towards label-only membership inference attack against pre-trained large language models](#). *Preprint*, arXiv:2502.18943.
- Hongsheng Hu, Zoran Salicic, Lichao Sun, Gillian Dobie, 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.
- Jing Huang, Diyi Yang, and Christopher Potts. 2024. [Demystifying verbatim memorization in large language models](#). *Preprint*, arXiv:2407.17817.
- Kate Knibbs. 2023. The battle over books3 could change ai forever. <https://www.wired.com/story/battle-over-books3/>.
- Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. 2023. [Practical membership inference attacks against large-scale multi-modal models: A pilot study](#). *Preprint*, arXiv:2310.00108.
- Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. [What matters when building vision-language models?](#) *Preprint*, arXiv:2405.02246.
- 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*.
- Chao Ma and Lexing Ying. 2021. [On linear stability of sgd and input-smoothness of neural networks](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 16805–16817. Curran Associates, Inc.
- Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzić. 2024. Llm dataset inference: Did you train on my dataset? *Advances in Neural Information Processing Systems*, 37:124069–124092.
- Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.
- Justus Mattern, Fatemehsadat Miresghallah, 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.
- Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. 2023. Did the neurons read your book? document-level membership inference for large language models. *arXiv preprint arXiv:2310.15007*.
- Fatemehsadat Miresghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. 2022a. [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.

- Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David K Evans, and Taylor Berg-Kirkpatrick. 2022b. An empirical analysis of memorization in fine-tuned autoregressive language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1816–1826.
- Ryoto Miyamoto, Xin Fan, Fuyuko Kido, Tsuneo Matsumoto, and Hayato Yamana. 2025. [Openlvlm-mia: A controlled benchmark revealing the limits of membership inference attacks on large vision-language models](#). *Preprint*, arXiv:2510.16295.
- OpenAI. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Jie Ren, Kangrui Chen, Chen Chen, Vikash Sehwal, Yue Xing, Jiliang Tang, and Lingjuan Lyu. 2024. [Self-comparison for dataset-level membership inference in large \(vision-\)language models](#). *Preprint*, arXiv:2410.13088.
- Shahbaz Rezaei and Xin Liu. 2021. On the difficulty of membership inference attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7892–7900.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- Daniel Samira, Edan Habler, Yuval Elovici, and Asaf Shabtai. 2025. Variance-based membership inference attacks against large-scale image captioning models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9210–9219.
- Avital Shafran, Shmuel Peleg, and Yedid Hoshen. 2021. Membership inference attacks are easier on difficult problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14820–14829.
- 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.
- Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. 2017. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, pages 587–601.
- 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*.
- Gemini Team. 2025. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- Toan Tran, Ruixuan Liu, and Li Xiong. 2025. [Tokens for learning, tokens for unlearning: Mitigating membership inference attacks in large language models via dual-purpose training](#). *Preprint*, arXiv:2502.19726.
- Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, and 5 others. 2024a. [A comprehensive review of multimodal large language models: Performance and challenges across different tasks](#). *Preprint*, arXiv:2408.01319.
- 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*.
- 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*.
- Luis Wiedmann, Aritra Roy Gosthipaty, and Andrés Marafioti. 2025. nanovlm. <https://github.com/huggingface/nanoVLM>.
- Hengyu Wu and Yang Cao. 2025. [Membership inference attacks on large-scale models: A survey](#). *Preprint*, arXiv:2503.19338.
- Zongyu Wu, Minhua Lin, Zhiwei Zhang, Fali Wang, Xi'anren Zhang, Xiang Zhang, and Suhang Wang. 2025. [Image corruption-inspired membership inference attacks against large vision-language models](#). *Preprint*, arXiv:2506.12340.

- Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. 2025. [Recall: Membership inference via relative conditional log-likelihoods](#). *Preprint*, arXiv:2406.15968.
- 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*.
- Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. *arXiv preprint arXiv:2402.15159*.
- Wentao Ye, Jiaqi Hu, Liyao Li, Haobo Wang, Gang Chen, and Junbo Zhao. 2024. [Data contamination calibration for black-box LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10845–10861, Bangkok, Thailand. Association for Computational Linguistics.
- 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.
- Zhengqing Yuan, Zhaoxu Li, Weiran Huang, Yanfang Ye, and Lichao Sun. 2024. [Tinygpt-v: Efficient multimodal large language model via small backbones](#). *Preprint*, arXiv:2312.16862.
- 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*.
- Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. 2023. Low-cost high-power membership inference by boosting relativity. *arXiv preprint arXiv:2312.03262*.
- 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*.
- Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2024b. [Llama-adapter: Efficient fine-tuning of language models with zero-init attention](#). *Preprint*, arXiv:2303.16199.
- Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024c. [Pre-training 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.
- Xuanming Zhang. 2025. [Generalization or memorization: Dynamic decoding for mode steering](#). *Preprint*, arXiv:2510.22099.
- Chenxu Zhao, Wei Qian, Aobo Chen, and Mengdi Huai. 2025. [Membership inference attacks with false discovery rate control](#). *Preprint*, arXiv:2508.07066.
- Xiaofan Zheng, Huixuan Zhang, and Xiaojun Wan. 2025. [Tracing training footprints: A calibration approach for membership inference attacks against multimodal large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17179–17191, Suzhou, China. Association for Computational Linguistics.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. [Minigpt-4: Enhancing vision-language understanding with advanced large language models](#). *Preprint*, arXiv:2304.10592.

A Datasets

To comprehensively evaluate the performance of DLC-MIA across different task scenarios, we select two datasets, Flickr30K (Young et al., 2014) and PixMo (Deitke et al., 2024), for our experiments. The PixMo dataset, recently released by the Allen Institute for AI, is an open dataset for visual instruction tuning that features diverse question types and covers a wide range of domains. We randomly sample 1000 instances from it for our experiments. Flickr30K is one of the widely used datasets, containing high-quality images with human-annotated captions. In our experiments, we select the first annotation for each image and, following the setup of Liu et al. (2023) and Zheng et al. (2025), convert it into the format of visual instruction tuning data, also randomly sampling 1000 instances. Subsequently, we randomly and evenly partition each dataset into member and non-member data.

B Evaluation Metrics

Prior research points out that preset fixed thresholds are not robust across different datasets (Wu and Cao, 2025; Hu et al., 2022; Zhao et al., 2025). In practical use, attackers employ a corresponding non-member dataset to dynamically determine appropriate thresholds. Therefore, almost all recent studies do not use preset fixed thresholds as evaluation metrics, but instead adopt the area under the ROC curve and the true positive rate at low false positive rates as metrics. We follow this setup as well.

Area Under the ROC Curve AUC (Area Under the ROC Curve) is a universal standard for evaluating the overall performance of binary classification models and is widely applied in membership inference attack (MIA) research. It describes the relationship between the true positive rate (TPR: the proportion of correctly identified member samples) and the false positive rate (FPR: the proportion of incorrectly identified non-member samples) at different decision thresholds through the ROC curve. The value range of AUC is $[0, 1]$, where random guessing corresponds to 0.5, and values closer to 1 indicate stronger discriminative ability of the method. As an overall performance metric, AUC objectively reflects the global recognition capability of the attack method.

True Positive Rate at Low False Positive Rate Although AUC provides a global perspective on

performance evaluation, in practical scenarios, the cost of false positive errors (incorrectly labeling non-members as members) is typically high and may even bring significant negative consequences. For example, in privacy auditing tasks, false alarms are often more serious than missed detections. Meanwhile, since the number of non-member samples in real-world datasets far exceeds that of member samples, the true positive rate at low false positive rate (TPR@low%FPR) becomes a more practical key metric. This metric measures the ability of an attack method to accurately identify member data when the false alarm rate is extremely low (e.g., 5%). This particularly reflects the reliability of the method in practical applications, making it the most important metric at present.

C Additional Baseline Details

This section elaborates on the baseline methods used for comparison in our experiments and their specific implementations. To enable methods originally designed for large language models (LLMs) to adapt to multimodal scenarios, we make necessary adjustments, primarily focusing the attack target on the Answer portion of visual instruction tuning data (Zheng et al., 2025).

1. **Loss** (Yeom et al., 2018): This method directly uses the loss value of the model on the target sample as the membership score. Its core assumption is that member samples should have lower loss values compared to non-member samples.
2. **Neighbor** (Mattern et al., 2023): This method determines membership by comparing the loss difference between the target sample and a set of "neighborhood" samples generated from it. Following the configuration of related research, we generate 20 neighborhood samples for each sample by performing keyword replacement on the answer text and calculate their average loss.
3. **Min-K% Prob** (Shi et al., 2024b): This method calculates the average log-likelihood of the K% tokens with the lowest probabilities in the answer text as the score. Its assumption is that the model exhibits extremely low confidence for certain tokens in non-member samples. According to its standard setting, we set K to 20.

4. **DC-PDD** (Zhang et al., 2024c): To address the inherent difficulty problem of samples, this method uses word frequency statistics from an external large text corpus (we adopt the C4 corpus recommended in the original paper) to calibrate the membership score.
5. **Min-K%++ Prob** (Zhang et al., 2025b): As an important improvement over Min-K%, this method uses the mean and standard deviation of the log-likelihood distribution output by the entire vocabulary for calibration, thereby more precisely evaluating prediction anomalies. Notably, this method requires access to the model’s complete logits output, making it a grey-box attack. In our experiments, we set $K=20$.
6. **MaxRényi-K%** (Li et al., 2024): This method quantifies the model’s prediction confidence for answer text based on Rényi entropy, assuming the model has higher confidence for member samples. In our experiments, we follow its recommended settings of $K=10$ and $\alpha=1$.
7. **PC-MMIA** (Zheng et al., 2025): The core assumption of this method is that tokens located on manifolds with weaker generalization ability can better reflect the training traces of samples. It performs difficulty calibration by observing the change patterns of token probabilities after image embeddings are subjected to bidirectional perturbations. We follow the hyperparameter configuration of the original literature.
8. **ICIMIA** (Wu et al., 2025): Inspired by the observation that member and non-member images differ in their sensitivity to corruption operations, this method performs membership inference attacks by calculating the embedding similarity between the original image and its corrupted version. It provides experimental schemes for both white-box and black-box access modes. Our experiments adopt the black-box access configuration.

D Additional Implementation Details

In this section, we provide supplementary explanations of specific implementation details. Our experiments are based on the official pre-trained weights of LLaMA-Adapter, MiniGPT-4, and LLaVA-1.5.

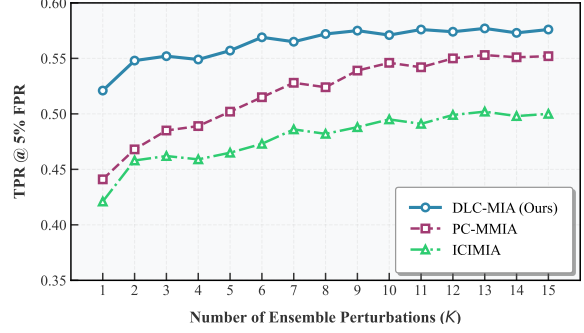


Figure 6: Performance comparison of DLC-MIA, PC-MMIA, and ICIMIA with different perturbation ensemble sizes (K) on the PixMo dataset using MiniGPT-4.

To simulate realistic attack scenarios, we mix the designated member samples with the model’s original visual instruction tuning dataset to form the final training set, and perform fine-tuning following the original hyperparameter configurations of each model. We select the lightweight nanoVLM-222M as the auxiliary model, which has a small parameter size and extremely low fine-tuning cost. During the simulated fine-tuning stage, the auxiliary model also adopts its default parameters. After completing fine-tuning on the entire experimental dataset (containing all member and non-member samples), we calculate the log-probability changes of tokens in each sample. Additionally, to enhance numerical stability in computations, when processing Equations (5) and (6), we add a small empirical constant $\epsilon = 0.01$ to the denominator term, making the numerical values smoother when the denominator approaches zero.

E Additional Analysis Experiments

We further analyze the impact of the number of accesses to LVLMs and multiple perturbation ensembles on attack performance. Among our baseline methods, PC-MMIA requires three accesses to the target model, while our method and ICIMIA require two accesses. Notably, all three methods rely on applying perturbations to the input image and calculating the changes in token probabilities before and after perturbation to construct the scoring function. Inspired by ensemble learning, we explore whether increasing the attack cost through multiple independent perturbations and averaging the results can yield more robust attack performance. Specifically, for DLC-MIA, we generate Gaussian noise perturbations with $\sigma = 0.15$ using different random seeds for each additional pertur-

bation. We apply Gaussian noise perturbations for PC-MMIA (Zheng et al., 2025) and Gaussian blur perturbations for ICIMIA (Wu et al., 2025). Finally, we average the scoring function $\mathcal{S}$ obtained from K independent perturbations to obtain the ensemble score $\mathcal{S}_{\text{ensemble}}$.

As shown in Figure 6, we present the performance changes of each method under different numbers of perturbation ensembles K . The experimental results demonstrate that as the number of ensemble perturbations increases, the TPR@5%FPR metric of all attack methods show an upward trend. This verifies that averaging multiple perturbations can indeed reduce the randomness introduced by a single perturbation, thereby improving attack performance. DLC-MIA maintains optimal performance under multiple perturbations, and the marginal benefit of performance improvement stabilizes around $K = 6$. Furthermore, we observe that when K is small, DLC-MIA demonstrates a notable performance advantage over the baseline methods. This suggests that our method can achieve high attack effectiveness even with a limited model access budget, highlighting its strong practicality.

F Computational Cost

In this section, we provide a detailed analysis of the computational cost associated with DLC-MIA, addressing concerns about the practicality and efficiency of our method.

Auxiliary Model Training Cost. We use NanoVLM-222M (SigLIP-B16 and SmoLM2) as our auxiliary model, which has fully open-source training data and configuration. Even training this auxiliary model from scratch requires only 6 hours on a single H100 GPU, making it highly accessible. In practice, we use the publicly available pre-trained weights and perform fine-tuning using the original configuration to simulate real-world scenarios. All samples in our experiments share a single fine-tuning process, which takes just 20–30 minutes for the entire experimental dataset on a single GPU, representing a very low computational cost.

Query Cost to Target Model. Our method requires only two queries to the target model for each sample: one for the original sample and one for the perturbed sample. This is fewer than PC-MMIA, which requires three queries, and signif-

icantly fewer than the Neighbor method, which requires dozens of queries to generate multiple neighborhood samples. This low query cost makes DLC-MIA particularly suitable for practical privacy auditing scenarios where API access may be rate-limited or costly.

Overall Cost Comparison. Compared to conventional shadow model approaches that require training multiple models with the same architecture and scale as the target model, our auxiliary model approach is orders of magnitude more efficient. The lightweight nature of the auxiliary model (222M parameters) and the minimal query requirements make DLC-MIA a practical and cost-effective solution for membership inference attacks against LVLMs.

G Experiments on Additional Architectures

To further validate the effectiveness and generalizability of DLC-MIA across different model architectures and scales, we conduct additional experiments on CogVLM-17b (Wang et al., 2024c) and Molmo-7B (Deitke et al., 2024). These models feature different architectural designs and parameter scales compared to our main experimental models, providing a more comprehensive evaluation of our method’s robustness.

As shown in Table 3, DLC-MIA maintains superior performance on both CogVLM-17b and Molmo-7B, demonstrating significant improvements over baseline methods. On CogVLM-17b, our method achieves an AUC of 0.752 and a TPR@5%FPR of 0.438, outperforming PC-MMIA by 4.1% and 3.6% respectively. On Molmo-7B, the improvements are even more pronounced, with DLC-MIA achieving an AUC of 0.778 and a TPR@5%FPR of 0.485, representing gains of 5.0% and 7.4% over PC-MMIA. These results confirm that our dual-level calibration approach generalizes well across different model architectures, parameter scales, and training paradigms, further validating the robustness and practical applicability of DLC-MIA.

H Ethics Statement

Our work investigates Membership Inference Attacks against Large Vision-Language Models to understand and address privacy risks in these systems. While MIAs serve as valuable tools for privacy auditing and detecting unauthorized data usage, we

Table 3: Performance comparison on CogVLM-17b and Molmo-7B using the Flickr30K dataset.

Method	CogVLM-17b			Molmo-7B		
	AUC	T@5%F	T@1%F	AUC	T@5%F	T@1%F
ICIMIA	0.698	0.393	0.188	0.691	0.378	0.228
PC-MMIA	0.711	0.402	0.213	0.728	0.411	0.298
DLC-MIA	0.752	0.438	0.255	0.778	0.485	0.361

acknowledge their dual-use nature. All our experiments use publicly available open-source datasets without any personal information, ensuring ethical compliance. We want to make clear that DLC-MIA is designed to help identify privacy vulnerabilities and improve security, not to enable malicious exploitation. We strongly encourage responsible use of MIA techniques and hope that LVLM developers will use our findings to better protect user privacy and data security.