Image Corruption-Inspired Membership Inference Attacks against Large Vision-Language Models

Zongyu Wu, Minhua Lin, Zhiwei Zhang, Fali Wang, Xianren Zhang, Xiang Zhang, Suhang Wang
The Pennsylvania State University
{zongyuwu,mfl5681,zbz5349,fqw5095,xzz5508,xzz89,szw494}@psu.edu

Abstract

Large vision-language models (LVLMs) have demonstrated outstanding performance in many downstream tasks. However, LVLMs are trained on large-scale datasets, which can pose privacy risks if training images contain sensitive information. Therefore, it is important to detect whether an image is used to train the LVLM. Recent studies have investigated membership inference attacks (MIAs) against LVLMs, including detecting image-text pairs and single-modality content. In this work, we focus on detecting whether a target image is used to train the target LVLM. We design simple yet effective Image Corruption-Inspired Membership Inference Attacks (ICIMIA) against LVLMs, which are inspired by LVLM's different sensitivity to image corruption for member and non-member images. We first perform an MIA method under the white-box setting, where we can obtain the embeddings of the image through the vision part of the target LVLM. The attacks are based on the embedding similarity between the image and its corrupted version. We further explore a more practical scenario where we have no knowledge about target LVLMs and we can only query the target LVLMs with an image and a textual instruction. We then conduct the attack by utilizing the output text embeddings' similarity. Experiments on existing datasets validate the effectiveness of our proposed methods under those two different settings.

1 Introduction

Large Vision-Language Models (LVLMs) (Liu et al., 2024a,b; Bai et al., 2023), which can generate text outputs based on visual and/or textual input, are attracting increasing attention. Many LVLMs have been developed, which have shown great performance on various tasks (Dong et al., 2024; Xu et al., 2024; Wu et al., 2025; Yue et al., 2024; Li et al., 2024a; Bucciarelli et al., 2024) such as biomedical question answering (Li et al., 2023a). Despite their great success, LVLMs have also raised critical concerns about privacy and copyright issues. LVLMs are usually trained on large-scale datasets (Changpinyo et al., 2021; Byeon et al., 2022). The training data might contain sensitive information, such as unauthorized medical data and copyrighted content. Previous works have shown that neural networks, especially those trained on large-scale datasets, can memorize training data (Song et al., 2017; Carlini et al., 2019, 2023). Thus, the memorization phenomena might also happen in LVLMs and inadvertently cause training data leakage (Li et al., 2024b), which could cause substantial loss to data owners. Hence, knowing whether one's data is used to train an LVLM is important for privacy and copyright protection.

Membership inference attacks (MIAs), which aim to determine whether a given sample is used to train a model (Hu et al., 2022a; Shokri et al., 2017), are critical for ensuring data safety and investigating data contamination (Oren et al., 2024; Duan et al., 2024). Generally, models tend to overfit to the training data, resulting in higher prediction confidence for member data (data used for training) than non-member data. Many traditional MIA methods (Shokri et al., 2017) adopt such nuance difference in model prediction to differentiate members and non-members. Recently, some works (Ko et al., 2023; Li et al., 2024b; Hu et al., 2025) have explored the MIA on vision-language models, from CLIP (Radford et al., 2021) to LVLMs (Zhang et al., 2023). Ko et al. (2023) focus on determining whether an image-text pair exists in the training data of CLIP. However, detecting whether an image in the training data is more practical than detecting the entire image-text pair (Li et al., 2024b) as the image owner might only have the image without text while the text description might be labeled by the model trainer.

Therefore, we study the problem of single-image MIA against LVLMs. The work in this direction is

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7945–7957

rather limited (Li et al., 2024b). Li et al. (2024b) also conduct MIA on a single modality (Image or textual description) against LVLMs by using the output logits of LVLM. Instead of using output logits, we propose a new perspective, i.e., adopting image embedding from the visual part of LVLMs. Our intuition is: as the model has seen member images, it should be able to give robust image embedding of a member image even if some details of the image are missing. In other words, *the image embedding of a member is more robust to image corruption than that of non-member images*, which is verified by our preliminary experiment in Section 3 (see Figure. 2). Based on this observation, we propose a novel MIA algorithm under the white-box setting, where we can obtain the image embedding from LVLMs. Given an image, we corrupt it and use the image embedding similarity between the raw image and the corrupted version to decide if it is a member. A higher similarity means the image is more likely to be a member.

However, for many closed-source LVLMs, we cannot obtain image embeddings. To address this issue, we extend the similarity to the output text level. Our assumption is: robust image embedding of member image will result in robust text generation under perturbation, which is also verified in Figure 3. Based on this observation, we extend our framework to black-box setting, where we can only obtain output texts from LVLMs. Given a target image, we corrupt the image and compare the generated text similarity between the raw image and its corrupted version. A larger text similarity means the image is more likely to be a member.

Our **main contributions** are: (**i**) In this work, we investigate two membership inference attack scenarios targeting LVLMs. For each setting, we propose one simple yet strong attack method that leverages the model's robustness to image corruption on its training images; (**ii**) Extensive experiments on existing datasets show the effectiveness of the proposed membership inference method.

2 Related Work

2.1 Large Vision-Language Models

Large Vision-Language Models (Liu et al., 2024b,a; Chen et al., 2024b; Tong et al., 2024; Zhu et al., 2024; Chen et al., 2023; Li et al., 2024c; Chen et al., 2024a), also known as multimodal large language models (Fu et al., 2024) are developing rapidly due to the success of language models across various aspects (Wang et al., 2025; Zhao et al., 2023; Zhang et al., 2025b; Chiang et al., 2023; Touvron et al., 2023a,b; Team, 2023; Penedo et al., 2023; Lin et al., 2024). These multimodal models, like LLaVA series (Liu et al., 2023, 2024a), can generate textual outputs given textual questions and images.

2.2 Membership Inference Attack

Membership Inference Attack (MIA) (Shokri et al., 2017; Salem et al., 2018; Sablayrolles et al., 2019; Li et al., 2021; Hu et al., 2022a; Nasr et al., 2019; Leino and Fredrikson, 2020; Rezaei and Liu, 2021) tries to determine whether a given data sample was used to train a machine learning model (Shokri et al., 2017).

One main category of MIA methods is metric-based (Sablayrolles et al., 2019; Choquette-Choo et al., 2021), which use some well-designed metrics (Hu et al., 2022a), such as the metrics based on loss, to determine the membership status of a given sample. Another main category consists of methods based on shadow training (Shokri et al., 2017), which trains some shadow models to simulate the target model and then conducts MIAs based on these shadow models. Many later works (Mireshghallah et al., 2022; Mattern et al., 2023; Ren et al., 2025; Shi et al., 2024), such as Min-K% (Shi et al., 2024) and Min-K%++ (Zhang et al., 2025a), have started to explore the MIA on Large Language Models (LLMs).

With the development of multimodal learning, there are also some works exploring MIA on multimodal models (Hu et al., 2022b; Ko et al., 2023; Hu et al., 2025; Li et al., 2024b). We focus on vision-text here. EncoderMI (Liu et al., 2021) studies MIAs on image encoder models such as CLIP vision encoder (Radford et al., 2021). It calculates the similarity scores between the augmented images' embeddings and the scores are then used to train a classifier to infer the member status of an image. Our work is inspired by EncoderMI. Ko et al. (2023) work on detecting whether an image-text pair is in the training data of CLIP models. Li et al. (2024b) investigate the single-modality MIA in LVLMs, which is a more practical scenario. They calculate the Rényi Entropy (Rényi, 1961) based on different slices of logits to infer membership. Hu et al. (2025) found that member data and non-member data have different sensitivity to temperature. They then perform four attack methods under four different settings to detect whether images are used in LVLM's training stage.

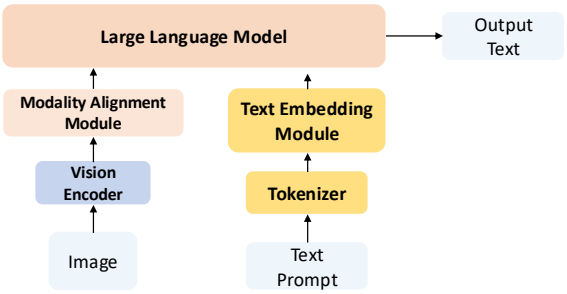

Figure 1: An illustration of the architecture of large vision-language models (Liu et al., 2024a).

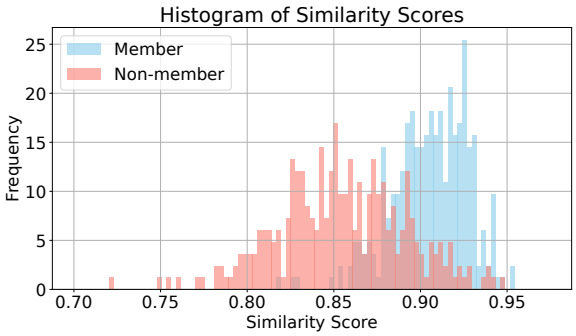

Figure 2: A histogram of similarity scores of corrupted images' embeddings and original images' embeddings for member and non-member data.

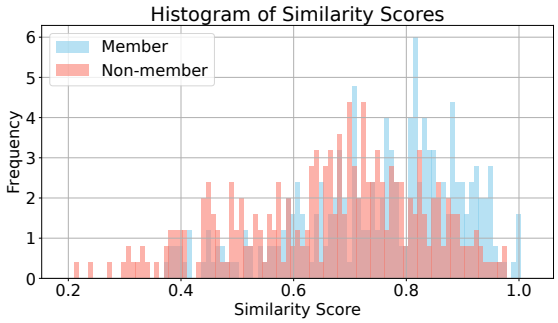

Figure 3: A histogram of similarity scores of textual output embeddings between corrupted images and the original images for member and non-member data.

Our work is inherently different from existing work: We propose a new perspective from image embedding robustness and output text robustness of the member image under image corruption, which works for both white-box and black-box settings.

3 Preliminaries

In this section, we give the necessary background information and formulate the problems for MIAs against LVLMs.

3.1 Large Vision-Language Models

As shown in Figure 1, an LVLM $\mathbf{M}_\theta$ usually consists of three parts (Liu et al., 2024a): A vision encoder f_{Vision}, an LLM f_{LLM}, and a modality connection module f_{Align}. The vision encoder f_{Vision}, e.g., CLIP (Radford et al., 2021) vision encoder used in LLaVA-1.5, takes an image $\mathbf{x}$ as input and outputs the embeddings of N image patches (excluding the CLS token) $\{\mathbf{z}_1, ..., \mathbf{z}_N\}$, where $\mathbf{z}_i \in \mathbf{R}^{d_v}$ is the i-th patch embedding with d_v being the embedding dimension. The patch embeddings are then transformed into an embedding sequence $\mathbf{E}^v = \{\mathbf{e}_1^v, ..., \mathbf{e}_N^v\}$ which is in the embedding space of f_{LLM} using f_{Align}. Each element in $\mathbf{E}^v$ can be represented as:

$$\mathbf{e}_i^v = f_{Align}(\mathbf{z}_i), \quad (1)$$

The text prompt T_{in}, composed of the instruction and the question, is tokenized and then encoded into an embedding sequence of tokens $\mathbf{E}^t = \{\mathbf{e}_1^t, ..., \mathbf{e}_K^t\}$, where K is the number of tokens. Finally, the text embedding and the image embeddings are used as input to the LLM f_{LLM} to get the text output T_{out}.

3.2 Threat Model

Attacker's Goal. Given a trained LVLM $\mathbf{M}_\theta$, the attacker's goal is to determine if a specific image was in the training set of $\mathbf{M}_\theta$. The attackers only have the target image and do not require its corresponding ground-truth text description. The problem is formally defined as:

Definition 1 (Image Only MIA) *Given a trained LVLM $\mathbf{M}_\theta$ and a target image $\mathbf{x_t}$, the attacker aims to determine whether $\mathbf{x_t}$ was part of the training data for $\mathbf{M}_\theta$, i.e., Attack$(\mathbf{M}_\theta, \mathbf{x_t}, T_{in}) \rightarrow \{0, 1\}$, where T_{in} is the input textual instruction. $\mathbf{x_t}$ is considered to be in the training set (i.e, member image) if the output is 1. Otherwise, $\mathbf{x}_t$ is a non-member.*

Attacker's Knowledge. In our work, we consider two practical settings, i.e., white-box and black-box: (i) **White-box Setting.** In this setting, the attacker can get the embedding of the image, i.e., the embedding through the model's vision encoder and modality alignment module. This is practical for open-source LVLMs. T_{in} is not required in this setting. (ii) **Black-box Setting.** In this setting, the attacker can only query the LVLM $\mathbf{M}_\theta$ with image input $\mathbf{x}$ and text input T_{in} to get the text output T_{out}. The attacker has no knowledge of

the target model. This setting is more realistic for commercial LVLMs.

3.3 Intuition for Performing MIA

As we are only given the image, how to fully utilize the image to determine if the image remains a challenging question. Inspired by the findings of EncoderMI (Liu et al., 2021) that a CLIP vision encoder tends to overfit to its training data and member images show higher similarity for two augmented versions, we explore whether LVLMs have similar features under image corruption methods. We assume that **the image embedding of a member image from the vision encoder and modality alignment module should be more robust to image corruption than that of non-member images**. The intuition is, if the model has seen the image and memorized the image, then even if the image is corrupted, e.g., some details missing, the model can still recall the details and result in embedding similar to the original one. To verify our assumption, for each image, we first apply *Gaussian Blur* as the corruption method to obtain the corrupted image, where Gaussian Blur is a smoothing technique that reduces image detail and noise by averaging pixels with a weighted Gaussian kernel. We compute the similarity score between the embedding of the original image and that of its corrupted version. The chosen model is LLaVA-1.5-7B (Liu et al., 2024a), and the dataset is the VL-MIA/Flickr dataset constructed by Li et al. (2024b). The similarity scores of member images and non-member images are shown in Figure 2. We can observe that member images normally show a higher similarity score than non-member images, which aligns well with our assumption. The scores show a clear difference between the two groups, which suggests that the similarity score between the original image embedding and its corrupted version's embedding can be used to perform an image membership inference attack. Thus, we can utilize the robustness of member image embedding to determine its membership.

However, in the black-box setting, we are unable to obtain the image embedding. As image embedding of member image is robust, it might also result in robust text even if the image is corrupted. Thus, we assume that **given the same text prompt (Instruction), an image seen during the LVLM's training process will produce a text output that is more similar to the output generated from its corrupted version, compared to images not included in the training data**. To verify our assumption, we use the same dataset, model, and corruption method as above. We compute the similarity score between the output text embeddings of the original and corrupted versions of each image, given the same prompt, separately. The results are shown in Figure 3. We observe that the discrepancy in similarity between member and non-member images still exists and can serve as a metric, although it is less apparent than that in Figure 2.

Algorithm 1 Image Similarity-based MIA

Input: Target image $\mathbf{x}_t$, threshold λ
Output: Membership Prediction Result $\in \{0, 1\}$

```
Obtain the embedding E^v of x_t
Apply corruption on x_t to get image x'_t
Obtain the embedding E^{v'} of x'_t
Compute similarity score s_img via Equation 2
if s_img < λ then
    x_t is a viewed as a non-member image
else
    x_t is a viewed as a member image
end if
```

4 Method

Based on observations that LVLMs are more robust to corruptions on members than non-members, we propose a novel framework, ICIMIA (**I**mage **C**orruption-**I**nspired **M**embership **I**nference **A**ttacks against Large Vision-Language Models). As shown in Figure 4, our method first produces a corrupted version $\mathbf{x}'_t$ for the target image $\mathbf{x}_t$. Then both the $\mathbf{x}_t$ and $\mathbf{x}'_t$ are utilized to get their image embeddings or corresponding output text embeddings. Finally, we calculate the image/text embedding similarity as a metric. The image is viewed as a member image if the similarity score is bigger than a certain threshold. Next, we introduce the details.

4.1 White-Box MIA via Image Similarity

We first discuss the white-box setting where we can obtain the embeddings of a given image through the vision encoder and modality alignment module.

Based on our observation in Section 3.3, the embeddings of trained images should be more robust to image corruption methods. To be specific, compared to images not used during training (Non-member data), the embeddings of the corrupted image and its original counterpart are normally more similar for images that were used in the training process (Member data). Therefore, we can use the

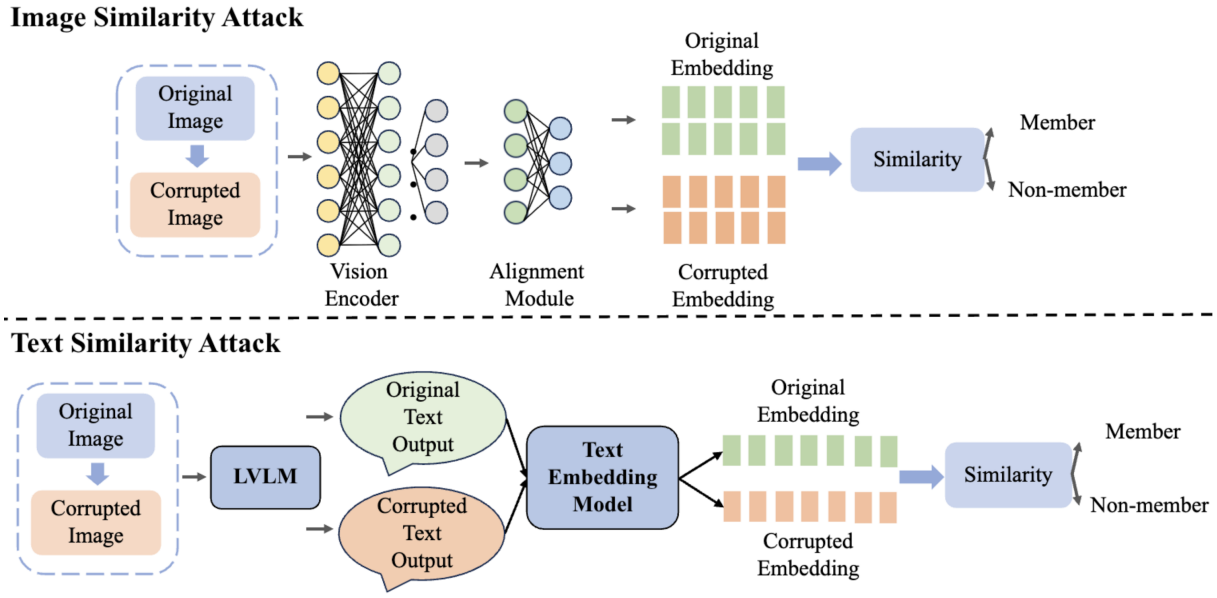

Figure 4: An illustration of the attack pipeline under two different settings.

similarity score as a metric to determine whether an image is used to train the LVLM.

The pipeline of our proposed framework is shown in the upper part of Figure 4. For a target image $\mathbf{x}_t$ that we wanna detect, we first apply some corruptions to get a corrupted image as $\mathbf{x}'_t = \text{Corruption}(\mathbf{x}_t)$. Various corruption techniques can be used, such as Gaussian blur (Using a Gaussian Kernel). Then, we get the image embeddings for both the original image and the corrupted one. Last, we measure how close each patch embedding pair is and calculate the mean value as

$$s_{img} = \frac{1}{N}\sum_{i=1}^{N} \text{Sim}(\mathbf{e}_i^v, \mathbf{e}_i^{v'}) \quad (2)$$

where $\mathbf{e}_i^v$ is the embedding of the i-patch of image $\mathbf{x}_t$ obtained by Equation 1 and $\mathbf{e}_i^{v'}$ denotes the i-patch embedding of the corrupted image $\mathbf{x}'_t$. N is the number of patches. $\text{Sim}(,)$ is a similarity function. We use cosine similarity here. The larger the s_{img} is, the more likely the image is viewed as a member image. The target image $\mathbf{x}_t$ is predicted as a member image if s_{img} is bigger than a threshold λ. Algorithm 1 summarizes the image similarity-based attack.

4.2 Black-box MIA via Text Similarity

For many commercial LVLMs, we are not able to obtain image embeddings. Thus, we study a more practical black-box setting where we know nothing about the target model but can only query the model with the target image and prompt to obtain the response text. Though we cannot obtain the image embedding, as the image embeddings of member images are robust to random corruption, correspondingly, the generated response text will be robust to the corruption, which is also verified in 3.3. This motivates us to query the target model using the original image and its corrupted version. Then we calculate pair-wise output text similarities.

Specifically, we feed the original target image $\mathbf{x}_t$ and a text prompt T_{in} into the target LVLM and get an output T_{out} as $T_{out} = f_{LVLM}(\mathbf{x}_t, T_{in})$. Similarly, we get an output T'_{out} for the corrupted image $\mathbf{x}'_t$ with the same input text prompt T_{in}. We then employ a text embedding model to get the embeddings of T_{out} and T'_{out}. With the text embedding, we calculate their similarity as

$$s_{text} = \text{Sim}(Emb(T_{out}), Emb(T'_{out})) \quad (3)$$

where $Emb()$ denotes a text embedding model, such as OpenAI's text-embedding-3-small model[1]. $\text{Sim}(,)$ is a similarity function as in Equation 2 which is also a cosine similarity function. Similarly, x_t is considered member data if $s_{text} > \lambda$. This method is summarized in Algorithm 2. This attack is similar to the Target-only Inference in Hu et al. (2025) where they also use text similarity scores. The difference is that they use the average similarity score between text outputs generated by querying the model multiple times.

[1]https://platform.openai.com/docs/models

Algorithm 2 Text Embedding Similarity-based Attack

Input: Target image $\mathbf{x}_t$, Prompt T_{in}, threshold λ
Output: Membership Prediction Result $\in \{0, 1\}$

Feed the target LVLM with the target image $\mathbf{x}_t$ and a text prompt T_{in} and get output T_{out}
Apply corruption to get corrupted image $\mathbf{x}'_t$
Feed the target LVLM with the corrupted target image $\mathbf{x}'_t$ and the same text prompt T_{in} and get output T'_{out}
Get the text output embedding $Emb(T_{out})$ and $Emb(T'_{out})$ of the original image and corrupted image
Calculate the output similarity using Equation 3
if $s_{text} < \lambda$ **then**
    $\mathbf{x}_t$ is viewed as a non-member image
else
    $\mathbf{x}_t$ is viewed as a member image
end if

5 Experiments

In this section, we evaluate our proposed ICIMIA on representative LVLM MIA datasets to answer these research questions: (**i**) How well do our proposed ICIMIA perform in conducting membership inference attacks against large vision-language models? and (**ii**) What are the impacts of hyperparameters?

5.1 Experimental Setup

Evaluation Metric. Following previous work (Li et al., 2024b), we use Area Under the Curve **(AUC)** and True Positive Rate at 5% False Positive Rate **(TPR at 5% FPR)** as the evaluation metrics. For both metrics, higher values indicate better MIA performance. The descriptions of these metrics can be found in Appendix B.

Datasets We test our method on two benchmark datasets, VL-MIA/Flickr and VL-MIA/Flickr-2k (Li et al., 2024b). The details of datasets can be found in Appendix A.

Computational Resources. We conduct all experiments on machines equipped with NVIDIA RTX A6000 GPUs (48GB memory each).

Selected Models. We evaluate our method on two models: LLaVA 1.5 7B, and LLaVA 1.5 13B (Liu et al., 2024a). The models are chosen as they are classical and the datasets are applicable to them.

Corruption Methods. We choose the following corruption methods: (i) *Gaussian Blur*: This is a technique that makes image soft and blurry using a Gaussian Kernel; (ii) *Motion Blur*: We apply a custom convolution kernel, mimicking the effect of motion-blur; and (iii) *JPEG Compression*: This is a method to compress the image to get the corrupted version. We apply OpenCV (Bradski, 2000) to achieve Gaussian Blur and Motion Blur.

Baselines. We adopt representative and state-of-the-art MIA methods against LVLMs, including: (i) **AugKL** (Li et al., 2024b; Liu et al., 2021): Li et al. (2024b) extend the approach designed by Liu et al. (2021) to LVLMs. Specifically, Li et al. (2024b) quantify the difference between the original and augmented images (Such as Crop and Rotation) by computing the KL divergence between their distributions of logits; (ii) **Max_Prob_Gap** (Li et al., 2024b): It is the average value of the difference between the highest and second-highest token probabilities at each position; (iii) **Min-K%** (Shi et al., 2024): Min-K is an MIA method designed for LLMs. It uses a ground-truth token and computes the lowest K% of its predicted probabilities. Li et al. (2024b) extends it to the LVLM domain; (iv) **Perplexity**: It is based on loss. Li et al. (2024b) analyze target perplexity to achive the attacks (Carlini et al., 2021); (v) **MaxRényi-K%** (Li et al., 2024b): It selects the top K% tokens with the highest Rényi entropy. Then the value is the average value of these entropies; and (vi) **Mod-Rényi** (Li et al., 2024b): This is an extended version of MaxRényi-K% and is designed for target-based scenarios.

Implementation Details. All these baselines need the knowledge of the target model's tokenizer and the output logits. The implementations of all baselines are based on Li et al. (2024b). Following the recommendation of Shi et al. (2024) and Li et al. (2024b), we set $K = 20$ for the Min-K% method. K is set to 0, 10, and 100 for MaxRényi-K% and α is chosen over 0.5, 1, and 2 for both ModRényi and MaxRényi-K%, as in Li et al. (2024b). We report the highest AUC for each method and provide the TPR at 5% FPR using the hyperparameter combination that achieves this highest AUC. For the Image Similarity-based attack, we fix the Kernel Size of the Gaussian blur and the Motion blur as 5. For JPEG compression, the image quality is set to 5 (lower values indicate stronger compression). For the Text Similarity-based attack, following Li et al. (2024b), we use "*Describe this image concisely*" (Li et al., 2024b) as the prompt and the max generation token amount is 32. For simplicity, we only use Gaussian blur for Text Similarity-based

Method	LLaVA-1.5-7B				LLaVA-1.5-13B			
	img	inst	desp	inst+desp	img	inst	desp	inst+desp
Perplexity	-	0.378	0.665	0.558	-	0.440	0.707	0.646
Min_20% Prob	-	0.374	0.672	0.370	-	0.454	0.684	0.433
ModRényi	-	0.370	0.658	0.613	-	0.442	0.703	0.678
Max_Prob_Gap	0.579	0.605	0.644	0.645	0.565	0.501	0.656	0.652
Aug_KL	0.665	0.568	0.537	0.549	0.636	0.540	0.538	0.552
MaxRényi	0.702	0.726	0.709	0.743	0.647	0.682	0.728	0.738
Ours (Image_Sim, Gaussian Blur, Kernel Size 5)	0.881				0.878			
Ours (Image_Sim, Motion Blur Kernel Size 5)	0.860				0.856			
Ours (Image_Sim, JPEG Compression, Quality = 5)	0.682				0.681			

Table 1: **AUC** of various baseline methods under Li et al. (2024b)'s pipeline and our proposed approach on VL-MIA/Flickr. For all the baselines, the term "img" refers to the logits segment associated with the image embedding, while "inst" represents the instruction segment (Li et al., 2024b). "desp" corresponds to the generated description's logits segment, and "inst+desp" denotes the combination of the instruction and description segments (Li et al., 2024b).

Method	LLaVA-1.5-7B				LLaVA-1.5-13B			
	img	inst	desp	inst+desp	img	inst	desp	inst+desp
Perplexity	-	0.007	0.137	0.067	-	0.047	0.227	0.127
Min_20% Prob	-	0.007	0.127	0.003	-	0.067	0.163	0.053
ModRényi	-	0.003	0.113	0.113	-	0.060	0.203	0.147
Max_Prob_Gap	0.050	0.083	0.163	0.163	0.050	0.107	0.163	0.160
Aug_KL	0.080	0.073	0.060	0.043	0.133	0.070	0.050	0.060
MaxRényi	0.100	0.210	0.163	0.127	0.077	0.073	0.213	0.183
Ours (Image_Sim, Gaussian Blur, Kernel Size 5)	0.333				0.323			
Ours (Image_Sim, Motion Blur Kernel Size 5)	0.363				0.297			
Ours (Image_Sim, JPEG Compression, Quality = 5)	0.057				0.060			

Table 2: **TPR at 5% FPR** of various baseline methods under Li et al. (2024b)'s pipeline and our proposed approach on VL-MIA/Flickr. The column 'img', 'inst', 'desp', and 'inst+desp' has the same meaning as the previous table.

attack and the kernel size is set to 45.

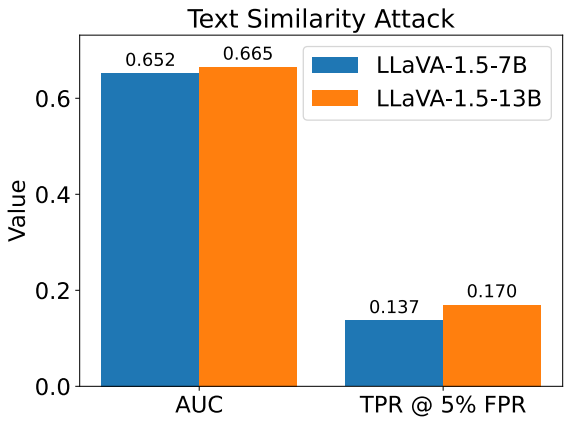

Figure 5: Text similarity-based membership inference attack on VL-MIA/Flickr.

5.2 White-Box MIA Performance

The results on LLaVA 1.5 series are shown in Table 1 and Table 2. Results on VL-MIA/Flickr 2K are in Appendix C. ***Please note that our methods have different knowledge of the models compared with the baseline methods in Table 1 and Table 2.*** These baseline methods can be viewed grey-box attack methods. We can obtain the image embeddings while they can obtain the output text's logits. We can observe that similarity-based attacks using Gaussian Blur and Motion Blur can achieve an AUC higher than 0.8 for both LLaVA 1.5-7B and LLaVA-1.5-13B, which is much higher than all the baselines. In comparison, the highest AUC among the baseline methods is 0.743 on LLaVA-1.5-7B and 0.738 on LLaVA-1.5-13B. Similar results can be found in Table 2, our methods largely outperform all the baselines in terms of TPR at 5% FPR. Our best method can have a TPR at 5% FPR higher than 0.3 on both models while the best baseline performance is 0.210 on LLaVA-1.5-7B and 0.227 on LLaVA-1.5-13B.

The results show that using information from the visual side can better facilitate image MIAs compared with using information from different

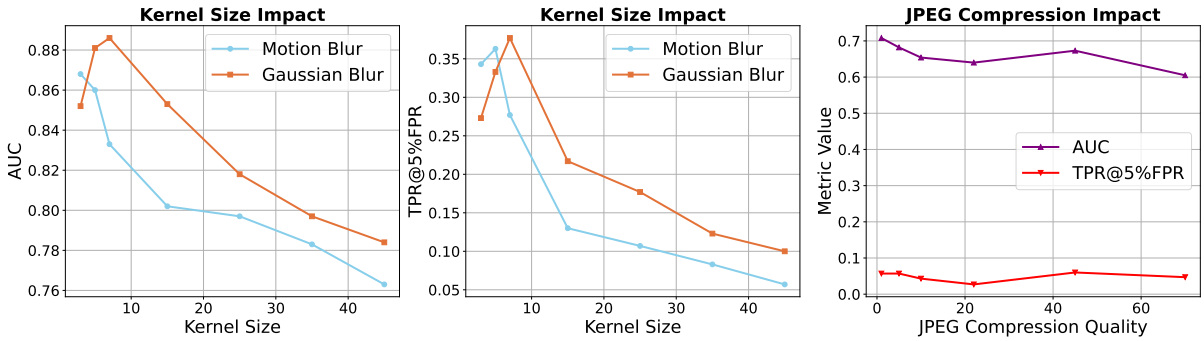

Figure 6: Analysis on Image Corruption Hyperparameters.

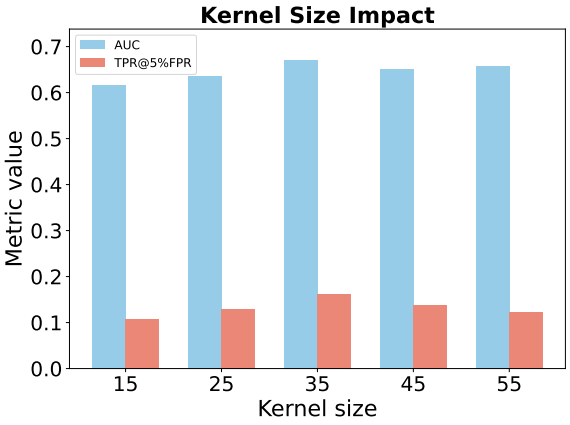

Figure 7: Kernel Size's Impact on Text Similarity-based Attack.

Method	AUC		TPR at 5% FPR	
	7B	13B	7B	13B
Ours (Gaussian Blur)	0.881	0.878	0.333	0.323
Ours (Gaussian Blur, CLIP)	0.885		0.413	
Ours (Motion Blur)	0.860	0.856	0.363	0.297
Ours (Motion Blur, CLIP)	0.858		0.393	
Ours (JPEG Compression)	0.682	0.681	0.057	0.060
Ours (JPEG Compression, CLIP)	0.705		0.103	

Table 3: Image MIA **AUC** and **TPR at 5% FPR** of our proposed approach on VL-MIA/Flickr. The embeddings are replaced with the ones obtained directly from the CLIP vision encoder without the alignment module.

logit slices. The results also suggest that the choice of image corruption method plays an important role. For example, the performance using JPEG compression is worse than that using Gaussian Blur and Motion Blur.

5.3 Black-box MIA Performance

The results of our attack method under the black-box setting, which is based on output texts' embeddings, are shown in Figure 5. The results demonstrate the effectiveness of our method. For example, it achieves an AUC of 0.652 and a TPR at 5% FPR of 0.137 on LLaVA-1.5-7B, which outperforms many baselines in Table 1 that require using output logits, even though our approach only queries the target model and utilizes the output text.

5.4 Hyperparameter Analysis

In this subsection, we conduct experiments to observe the impacts of hyperparameters. For the image similarity-based attack, we vary the kernel size for Gaussian Blur and Motion Blur using values of 3, 5, 7, 15, 25, 35, and 45. The image quality for JPEG compression is selected across 1, 5, 10, 20, 45, and 70. The selected model is LLaVA 1.5 7B. The results are shown in Figure 6. We can observe that smaller kernel sizes generally contribute to better performance in terms of both AUC and TPR at 5% FPR. Larger kernel sizes will bring stronger corruption. Therefore, for blur-based corruption methods, it suggests that a smaller corruption level could have better performance. Although the performance drops a lot with very large kernel sizes, it still outperforms most baselines that rely on output logits. For the third sub-figure, we can see that a smaller JPEG quality (Stronger compression) generally leads to higher AUC.

For the text similarity-based attack, the kernel size for Gaussian Blur is varied across 15, 25, 35, 45 and 55. The results are shown in Figure 7. Unlike image similarity-based attacks, text similarity-based attacks generally perform better with larger kernel sizes. One possible reason is that if the kernel size is too small, small changes in the image embeddings lead to even smaller changes in the textual output. This makes the difference between member and non-member images less clear. Since we are using the AUC as the metric, we don't need to explicitly set the threshold λ.

5.5 Extra Findings

We replace the embeddings obtained by the CLIP vision encoder and the alignment module with the embeddings directly from the CLIP vision encoder without passing them through the alignment module. The results are shown in Table 3. Interestingly, we find that using the embeddings directly from the CLIP vision encoder can even have slightly better performance. However, the CLIP encoder is frozen during the training stage of LLaVA-1.5. This suggests that the images might be included in the CLIP model's original training data. Our method's strong performance might also benefit from this. This suggests that we need some new benchmark datasets to better define the image MIA problem in the context of LVLMs.

6 Conclusion

In this paper, we design novel membership inference attack methods named ICIMIA against LVLMs under both white-box and black-box settings. Our approach is based on the observation that LVLMs exhibit varying sensitivity to image corruption for member and non-member images. We leverage this phenomenon by using the pair-wise similarity of the original version and its corrupted counterpart as the metric. Experimental results on representative datasets validate the effectiveness of our proposed methods in image membership inference attacks.

Limitations

We observed an interesting phenomenon: many non-member images generated by DALL·E 2 (Ramesh et al., 2022) exhibit greater robustness to corruption compared to their original counterparts (the generated image is prompted to be similar to its original member image), regardless of whether the original image is a member or non-member. Therefore, our method does not work for such a dataset where each non-member image is generated by DALL-E based on the original member image. One example is the VL-MIA/DALL-E dataset (Li et al., 2024b) where Blip (Li et al., 2023b) generates a caption for each member image and the caption is used by DALL-E to generate a non-member image for this image. We leave this as our future work.

Ethical considerations

Our work aims to find the images that are in the training data of large vision-language models. Our proposed method, ICIMIA, can help individuals know whether their sensitive data is used to train the model. All experiments are conducted on open-source models and publicly available datasets. In this paper, we only use AI assistants for grammar checking and sentence polishing.

Acknowledgments

This material is based upon work supported by, or in part by the Army Research Office (ARO) under grant number W911NF-2110198, the Department of Homeland Security (DHS) under grant number 17STCIN00001-05-00, and Cisco Faculty Research Award. The findings in this paper do not necessarily reflect the view of the funding agencies.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Davide Bucciarelli, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Personalizing multimodal large language models for image captioning: an experimental analysis. *arXiv preprint arXiv:2412.03665*.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *Proceedings of the 28th USENIX Conference on Security Symposium*, pages 267–284.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security)*, pages 2633–2650.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3568.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024a. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-only membership inference attacks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 1964–1974.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. 2024. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*.

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 *Conference on Language Modeling*.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, Caifeng Shan, and Ran He. 2024. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022a. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys*, 54.

Pingyi Hu, Zihan Wang, Ruoxi Sun, Hu Wang, and Minhui Xue. 2022b. M^4i: Multi-modal models membership inference. In *Advances in Neural Information Processing Systems*, volume 35, pages 1867–1882.

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

Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. 2023. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4871–4881.

Klas Leino and Matt Fredrikson. 2020. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *29th USENIX security symposium (USENIX Security)*, pages 1605–1622.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024a. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, pages 28541–28564.

Jiacheng Li, Ninghui Li, and Bruno Ribeiro. 2021. Membership inference attacks and defenses in classification models. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, pages 5–16.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742.

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

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024c. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26763–26773.

Minhua Lin, Zhengzhang Chen, Yanchi Liu, Xujiang Zhao, Zongyu Wu, Junxiang Wang, Xiang Zhang, Suhang Wang, and Haifeng Chen. 2024. Decoding time series with llms: A multi-agent framework for cross-domain annotation. *arXiv preprint arXiv:2410.17462*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916.

Hongbin Liu, Jinyuan Jia, Wenjie Qu, and Neil Zhenqiang Gong. 2021. Encodermi: Membership inference against pre-trained encoders in contrastive learning. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 2081–2095.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, 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.

Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David K Evans, and Taylor Berg-Kirkpatrick. 2022. 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.

Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2024. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Jie Ren, Kangrui Chen, Chen Chen, Vikash Sehwag, Yue Xing, Jiliang Tang, and Lingjuan Lyu. 2025. Self-comparison for dataset-level membership inference in large (vision-)language model. In *Proceedings of the ACM on Web Conference 2025*, page 910–920.

Alfréd Rényi. 1961. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*. University of California Press.

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.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. 2019. White-box vs black-box: Bayes optimal strategies for membership inference. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5558–5567.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2018. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. 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.

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.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM-techreport.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, TzuHao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang, and Suhang Wang. 2025. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16.

Zongyu Wu, Yuwei Niu, Hongcheng Gao, Minhua Lin, Zhiwei Zhang, Zhifang Zhang, Qi Shi, Yilong Wang, Sike Fu, Junjie Xu, Junjie Ao, Enyan Dai, Lei Feng, Xiang Zhang, and Suhang Wang. 2025. Lanp: Rethinking the impact of language priors in large vision-language models. *arXiv preprint arXiv:2502.12359*.

Junjie Xu, Zongyu Wu, Minhua Lin, Xiang Zhang, and Suhang Wang. 2024. Llm and gnn are complementary: Distilling llm for multimodal graph learning. *arXiv preprint arXiv:2406.01032*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9556–9567.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025a. 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. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.

Xianren Zhang, Xianfeng Tang, Hui Liu, Zongyu Wu, Qi He, Dongwon Lee, and Suhang Wang. 2025b. Divide-verify-refine: Can LLMs self-align with complex instructions? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13783–13800.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

Name	VL-MIA/Flickr	VL-MIA/Flickr 2K
Source of Member Data	MS COCO	MS COCO
Source of Non-member Data	Flickr	Flickr
#Member Data	300	1000
#Non-member Data	300	1000

Table 4: Dataset statistics. The datasets are constructed by Li et al. (2024b).

Method	AUC		TPR at 5% FPR	
	LLaVA-1.5-7B	LLaVA-1.5-13B	LLaVA-1.5-7B	LLaVA-1.5-13B
Ours (Image_Sim, Gaussian Blur, Kernel Size 5)	0.854	0.850	0.349	0.337
Ours (Image_Sim, Motion Blur, Kernel Size 5)	0.842	0.837	0.372	0.352
Ours (Image_Sim, JPEG Compression, Quality = 5)	0.629	0.635	0.088	0.080

Table 5: Image MIA performance of our proposed approach on VL-MIA/Flickr 2K.

A Details of Datasets

Many models, such as LLaVA 1.5 (Liu et al., 2024a) and Minigpt 4 (Zhu et al., 2024), use MS COCO (Lin et al., 2014) to train the models. Therefore, Li et al. (2024b) use part of the images in MS COCO as the member images. For non-member data, Li et al. (2024b) select images uploaded to Flickr[2] after those target models' release date. VL-MIA/Flickr and VL-MIA/Flickr 2K are licensed under the Creative Commons Attribution 4.0 International Public License. The dataset statistics are summarized in Table 4.

B Used Metrics

We provide details about two used metrics here:

- **AUC**: Area Under the Curve (AUC) is the value of the area beneath the ROC curve. It is a widely used metric to evaluate the classification model's performance under all possible classification thresholds (Li et al., 2024b).
- **TPR at 5% FPR**: True Positive Rate at a fiex False Positive Rate is another widely used metric to evaluate the performance of MIA methods (Li et al., 2024b; Carlini et al., 2022). Following Li et al. (2024b), we use TPR at 5% FPR which reflects the value of the True Positive Rate when the False Positive Rate is 5%.

C Additional Results on More Datasets

We conduct experiments on VL-MIA/Flickr-2K. The image similarity-based MIA results are shown in Table 5. We can have similar observations as those shown in Table 1 and Table 2. The experimental results on VL-MIA/Flickr-2K also validate the effectiveness of our proposed methods.

D Potential Risks

Although our proposed ICIMIA is designed to protect data safety and privacy, malicious users can use ICIMIA to infer whether an image is used to train a certain LVLM and then get some private information. For example, if the attacker knows that one person's medical image is used to train an LVLM for a certain disease, the attacker can get the information that this person might have this disease.

[2]https://www.flickr.com/