ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Lost in Modality: Evaluating the Effectiveness of Text-Based Membership Inference Attacks on Large Multimodal Models

Ziyi Tong and Feifei Sun and Le Minh Nguyen

Japan Advanced Institute of Science and Technology - Information Science
1-1 Asahidai, Nomi, Ishikawa 923-1292 - Japan

Abstract. Large Multimodal Language Models (MLLMs) are emerging as one of the foundational tools in an expanding range of applications. Consequently, understanding training-data leakage in these systems is increasingly critical. Log-probability-based membership inference attacks (MIAs) have become a widely adopted approach for assessing data exposure in large language models (LLMs), yet their effect in MLLMs remains unclear. We present the first comprehensive evaluation of extending these text-based MIA methods to multimodal settings. Our experiments under vision-and-text (V+T) and text-only (T-only) conditions across the DeepSeek-VL and InternVL model families show that in in-distribution settings, logit-based MIAs perform comparably across configurations, with a slight V+T advantage. Conversely, in out-of-distribution settings, visual inputs act as regularizers, effectively masking membership signals.

1 Introduction

LLMs and MLLMs have progressed rapidly [1, 2, 3, 4], heightening concerns about training-data exposure and motivating research on membership inference attacks[5]. For text input, logit-based MIA on LLM has advanced substantially, with recent methods achieving AUROC values near 90%[6]. MLLMs integrate a vision encoder with a language decoder (Figure 1), where visual embeddings are projected into the language model's representation space and fused with text representations to condition next-token prediction[7]. Although ground-truth image tokens are not available, the text logits remain conditioned on visual embeddings, which preserves the applicability of logit-based MIA.

However, it remains unclear whether these state-of-the-art, text-targeted MIA methods can be reliably applied to multimodal architectures, or how vision-conditioned text representations respond to logit-based membership inference. Addressing these questions is essential for assessing the privacy risks of modern MLLMs.

To address these questions, we conduct experiments on four MLLMs under both in-distribution(ID) and out-of-distribution(OOD) conditions, using text-only and multimodal inputs to isolate the influence of visual features on membership leakage. Experimental results show that Visual inputs suppress MIA signals in OOD settings and that the impact of MIA is highly model-dependent.

Our contributions are as follows: First, we offer a comprehensive evaluation of state-of-the-art text-based MIA methods on multimodal inputs, revealing how

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

membership signals behave when visual and textual streams interact. Second, we conduct a cross-model comparison showing that visual features modulate MIA effectiveness in model-dependent ways, driven by differing vision-text interactions. Third, we analyze in-distribution and out-of-distribution settings, demonstrating that natural distribution shifts have a substantial impact on membership inference, often overshadowing the effect of multimodal fusion.

2 Related Work

Membership Inference Attack. Membership inference is an attack method in which an adversary attempts to determine whether a specific data sample was included in a model's training set[5]. Early text-targeted approaches use input loss directly [8] or calibrate it with a reference model[9]. Compression-based variants such as [10] compress token-loss sequences to smooth high-variance noise and expose a lower-entropy signal useful for membership inference. More recent techniques leverage minimum token probabilities, including [11] and its normalized extension [12]. [6] measures membership by comparing conditional and unconditional likelihoods under non-member prefixes. Image-targeted MIAs have also emerged. [7] utilizes the Rényi entropy of next-token distributions as a robust confidence signal.

Multimodal Large Language Models. Multimodal large language models (MLLMs) extend language models by enabling joint reasoning over images and text. A widely adopted fusion paradigm maps the vision encoder's outputs into the language model's embedding space using a lightweight MLP projection adapter, allowing the LLM to process visual features as part of its token sequence. Recent architectures follow this design pattern, including DeepSeek-VL[1] and its MoE-based successor DeepSeek-VL2[2], which focus on general-purpose reasoning, as well as InternVL-1.5[3] and InternVL-2.5[4], which incorporate high-resolution vision encoders to enhance perceptual grounding. Despite their rapid progress, the privacy characteristics of these MLLMs remain largely unexplored.

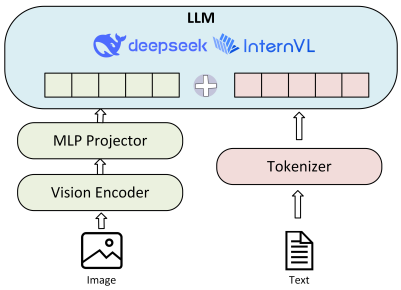


Fig. 1: Architecture schematic of the vision–language fusion pipeline in modern MLLMs.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

3 Experiments

3.1 Problem Setting

Let M be an autoregressive multimodal language model that produces a next-token distribution conditioned on an input sequence. Each VQA sample is represented as

$$x = (\mathbf{v}, \mathbf{q}, \mathbf{c}),$$

where $\mathbf{v}$ is the image, $\mathbf{q}$ the question, and $\mathbf{c}$ the candidate answers. Let D denote the dataset used to train M. The goal of membership inference is to determine, for a target sample x, whether $x \in D$ or $x \notin D$.

Under a gray-box setting, the adversary has access to the model's token-level logits and computes a membership score $S(x; M)$ from these logits. The score is then thresholded to predict membership:

$$m(x) = \begin{cases} 1 & \text{if } S(x; M) \geq \tau, \\ 0 & \text{otherwise,} \end{cases}$$

where $m(x) = 1$ indicates that x is a member. Our experiments evaluate the effectiveness of various logit-based MIA methods in distinguishing members from non-members in multimodal VQA settings.

3.2 Models

MLLMs typically pair a vision encoder with an LLM and fuse modalities by projecting visual features into the LLM embedding space through an MLP adapter. To compare architectures that emphasize different components, we evaluate four representative models: DeepSeek-VL(*deepseek-vl-7b-base*)[1] and DeepSeek-VL2(*deepseek-vl2-small*)[2], which prioritize semantic reasoning with an MoE-enhanced decoder, and InternVL-1.5(*Mini-InternVL-Chat-4B-V1-5*)[3] and InternVL-2.5(*InternVL2_5-4B*) [4], which employ high-resolution vision encoders to strengthen perceptual grounding.

3.3 Datasets

A major challenge in MIA research is obtaining member and non-member samples, since large-scale pretraining often obscures provenance[5]. ScienceQA[13] provides a rare case with explicit documentation: both DeepSeek-VL[1] and InternVL1.5/2.5[3, 4] report using ScienceQA for training. DeepSeek-VL also report using ScienceQA for evaluation, and DeepSeek-VL2 reports inheriting the general VQA training data of DeepSeek-VL. We therefore treat ScienceQA-train as member data and ScienceQA-test as in-distribution non-members. To further validate robustness, we also include AI2D[14], a VQA dataset from the same scientific domain but not reported as training data for any of the four models, forming an OOD non-member split. Together, ScienceQA-train vs. ScienceQA-test and ScienceQA-train vs. AI2D instantiate our ID and OOD settings.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

3.4 Methods

We compare six state-of-the-art logit-based MIA methods: Loss[8], Reference[9], Min-K%[11], Min-K%++[12], ReCALL[6], and Zlib[10], covering loss-based, calibrated, entropy-based, and compression-based paradigms. To disentangle the contributions of text and vision, we evaluate two inference configurations: T-only (question + choices, image masked) , and V+T (image + question + choices). We report the area under the ROC curve (AUC) as the primary metric, following standard MIA evaluation practice[11, 6].

4 Results and Discussion

As show in Table 1, we evaluate six MIA methods across four multimodal models in both in-distribution and out-of-distribution settings.

Table 1: Comparison of MIA performance across models, input modalities, and datasets.

	In-Distribution (ScienceQA-Test)			OOD (AI2D)		
	Text-only	**V+T**	**Δ**	**Text-only**	**V+T**	**Δ**
DeepSeek-VL						
Loss	0.475	0.487	0.012 ▲	0.666	0.655	-0.010 ▼
Reference	0.525	0.516	-0.009 ▼	0.567	0.657	0.090 ▲
Zlib	0.487	0.491	0.004 ▲	0.604	0.603	-0.001 ▼
Min-K%	0.482	0.502	0.020 ▲	0.715	0.409	-0.306 ▼
Min-K%++	0.475	0.504	0.029 ▲	0.540	0.330	-0.210 ▼
ReCall	0.513	0.512	-0.002 ▼	0.442	0.295	-0.147 ▼
DeepSeek-VL2						
Loss	0.475	0.483	0.008 ▲	0.563	0.538	-0.025 ▼
Reference	0.502	0.493	-0.009 ▼	0.322	0.559	0.236 ▲
Zlib	0.484	0.490	0.005 ▲	0.563	0.550	-0.013 ▼
Min-K%	0.479	0.507	0.028 ▲	0.475	0.313	-0.163 ▼
Min-K%++	0.487	0.479	-0.009 ▼	0.436	0.305	-0.131 ▼
ReCall	0.500	0.530	0.030 ▲	0.635	0.591	-0.044 ▼
InternVL1.5						
Loss	0.515	0.503	-0.012 ▼	0.680	0.618	-0.062 ▼
Reference	0.485	0.475	-0.010 ▼	0.353	0.311	-0.042 ▼
Zlib	0.505	0.500	-0.005 ▼	0.639	0.605	-0.034 ▼
Min-K%	0.489	0.500	0.011 ▲	0.501	0.286	-0.215 ▼
Min-K%++	0.473	0.482	0.009 ▲	0.311	0.317	0.006 ▲
ReCall	0.516	0.512	-0.005 ▼	0.546	0.473	-0.073 ▼
InternVL2.5						
Loss	0.516	0.507	-0.008 ▼	0.674	0.593	-0.081 ▼
Reference	0.503	0.505	0.003 ▲	0.594	0.423	-0.171 ▼
Zlib	0.513	0.506	-0.008 ▼	0.672	0.594	-0.078 ▼
Min-K%	0.494	0.484	-0.010 ▼	0.555	0.383	-0.172 ▼
Min-K%++	0.496	0.493	-0.003 ▼	0.515	0.383	-0.133 ▼
ReCall	0.516	0.504	-0.012 ▼	0.605	0.561	-0.044 ▼

Δ = V+T − Text-only; ▲ indicates improvement, ▼ indicates degradation.

In the text-only (T-only) setting, all models exhibit low susceptibility to membership inference, with AUROC scores consistently hovering near the random baseline(0.5). This aligns with prior research [6] indicating that in-distribution membership inference is still very difficult for well-generalized large language

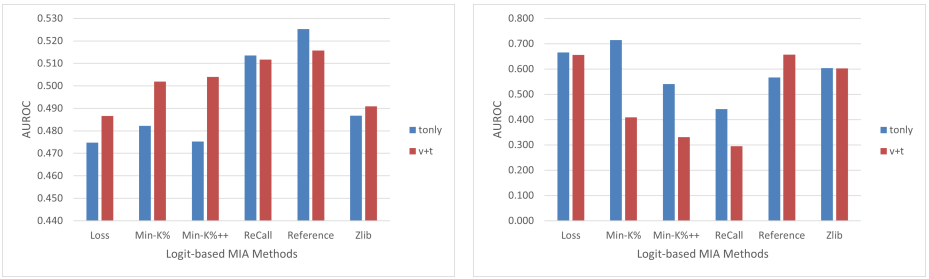
ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

models, as the loss gap between members and non-members is minimal.

When introducing visual inputs (V+T) in the in-distribution setting, we observe minor changes in MIA performance. DeepSeek-VL and DeepSeek-VL2 show deltas mostly within ± 0.03, indicating negligible impact from visual conditioning. InternVL models yield slightly more consistent gains, particularly for Min-K% variants, but the improvements remain small. Overall, visual features do not meaningfully increase membership leakage in the in-distribution setting, as the T-only baseline is already near 0.5.

Conversely, in the OOD setting (AI2D), the introduction of visual inputs (V+T) causes a catastrophic drop in MIA performance. As show in Figure 2, DeepSeek-VL sees a Min-K%++ drop of over 0.17, and ReCall drops by 0.21. This indicates that the visual modality introduces a domain shift that masks membership signals, effectively acting as a regularizer against standard loss-based attacks.

We attribute the observed ID–OOD discrepancy to two potential factors. First, the OOD dataset (AI2D) consists of clean, schematic diagrams with lower visual complexity than the mixed-content images in ScienceQA, which may act as a confounder by masking membership signals. Second, data contamination during pre-training is a plausible alternative explanation: given the scale of web-crawled corpora, the ScienceQA test split may have been partially exposed to the models, compressing the loss gap between members and non-members. Together, these factors may explain the near-random MIA performance in the in-distribution setting and the contrast with OOD results.



(a) In distribution Setting (b) Out of Distribution Setting

Fig. 2: DeepSeek-VL MIA performance in OOD (left) and In-Distribution (right) settings under text only(T-only) and vision plus text modes(V+T).

5 Conclusion

We present the first systematic evaluation of state-of-the-art text-based MIA methods on multimodal inputs. Our results show that visual inputs can invert membership signals, that ScienceQA may suffer from data contamination, and that different attack methods exhibit varying robustness. These findings highlight important limitations of applying text-only MIA techniques to multimodal

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

models and point to the need for multimodal-aware approaches to evaluating training-data exposure.

References

[1] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[2] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[3] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024.

[4] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[5] Ziyi Tong, Feifei Sun, and Le Minh Nguyen. Pretraining data exposure in large language models: A survey of membership inference, data contamination, and security implications. In *International Conference on Applications of Natural Language to Information Systems*, pages 152–162. Springer, 2025.

[6] Roy Xie, Junlin Wang, Ruomin Huang, Minxing Zhang, Rong Ge, Jian Pei, Neil Zhenqiang Gong, and Bhuwan Dhingra. Recall: Membership inference via relative conditional log-likelihoods. *arXiv preprint arXiv:2406.15968*, 2024.

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

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

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

[10] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[11] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

[12] Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*, 2024.

[13] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[14] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016.