Revisiting Data Auditing in Large Vision-Language Models

Hongyu Zhu
Shanghai Jiao Tong University
Shanghai, China
hongyu_z@sjtu.edu.cn

Sichu Liang
Southeast University
Nanjing, China
coder_liang@seu.edu.cn

Wenwen Wang
Carnegie Mellon University
Pittsburgh, USA
wenwenw@andrew.cmu.edu

Boheng Li
Nanyang Technological University
Singapore, Singapore
BOHENG001@ntu.edu.sg

Tongxin Yuan
Shanghai Jiao Tong University
Shanghai, China
teenyuan@sjtu.edu.cn

Fangqi Li
Shanghai Jiao Tong University
Shanghai, China
solour_lfq@sjtu.edu.cn

Hanyi Wang
Shanghai Jiao Tong University
Shanghai, China
why_820@sjtu.edu.cn

Shi-Lin Wang*
Shanghai Jiao Tong University
Shanghai, China
wsl@sjtu.edu.cn

Zhuosheng Zhang*
Shanghai Jiao Tong University
Shanghai, China
zhangzs@sjtu.edu.cn

Abstract

With the surge of large language models (LLMs), Large Vision-Language Models (VLMs)—which integrate vision encoders with LLMs for accurate visual grounding—have shown great potential in tasks like generalist agents and robotic control. However, VLMs are typically trained on massive web-scraped images, raising concerns over copyright infringement and privacy violations, and making data auditing increasingly urgent. Membership inference (MI), which determines whether a sample was used in training, has emerged as a key auditing technique, with promising results on open-source VLMs like LLaVA (AUC > 80%). In this work, we revisit these advances and uncover a critical issue: current MI benchmarks suffer from distribution shifts between member and non-member images, introducing shortcut cues that inflate MI performance. We further analyze the nature of these shifts and propose a principled metric based on optimal transport to quantify the distribution discrepancy. To evaluate MI in realistic settings, we construct new benchmarks with i.i.d. member and non-member images. Existing MI methods fail under these unbiased conditions, performing only marginally better than chance. Further, we explore the theoretical upper bound of MI by probing the Bayes Optimality within the VLM's embedding space and find the irreducible error rate remains high. Despite this pessimistic outlook, we analyze why MI for VLMs is particularly challenging and identify three practical scenarios—fine-tuning, access to ground-truth texts, and set-based inference—where auditing becomes feasible. Our study presents a systematic view of the limits and opportunities of MI for VLMs, providing guidance for future efforts in trustworthy data auditing. Code and data will be available at .

CCS Concepts

• **Security and privacy** → *Human and societal aspects of security and privacy*; • **Information systems** → *Multimedia information systems*; • **Computing methodologies** → *Artificial intelligence.*

Keywords

Data Transparency, Vision-Language Models

ACM Reference Format:
Hongyu Zhu, Sichu Liang, Wenwen Wang, Boheng Li, Tongxin Yuan, Fangqi Li, Hanyi Wang, Shi-Lin Wang, and Zhuosheng Zhang. 2025. Revisiting Data Auditing in Large Vision-Language Models. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755002

1 Introduction

Large Vision-Language Models (VLMs) are becoming ubiquitous. Proprietary systems such as GPT-4o [52] and Claude 3.5 Sonnet [1] exhibit impressive multimodal capabilities, excelling at comprehensive image description and complex visual reasoning, with promising applications in generalist agents [20] and embodied robotics [45]. To support open scientific research, the community has made notable progress in replicating these abilities under transparent, open-source settings. Notably, the LLaVA series [31, 36] integrates vision encoders with large language models (LLMs), achieving competitive performance while remaining fully accessible.

However, training large VLMs typically involves scraping web-scale multimodal data [4, 52], raising concerns about data legality and transparency. Recent incidents have highlighted these risks: copyright lawsuits involving GPT-4o [29], potential personal data leakage via VLM outputs [48], and test set contamination in benchmark evaluations [53, 54]. These challenges underscore the urgent need for principled data auditing, empowering third-party verification of *whether specific data were used during VLM training*.

Membership inference (MI)—which assesses whether a specific sample was part of a model's training set [57]—has recently emerged

*Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755002

as a powerful approach for auditing VLMs. In this setting, MI probes a black-box API with a target image and a crafted instruction, analyzing the language response and token-level probabilities to compute membership-indicative statistics. Thresholds calibrated on known member and non-member images are then used to infer membership status [26, 33], enabling non-intrusive and statistically grounded audits. Recent work has introduced benchmark datasets for open-source VLMs like LLaVA, designating subsets of training images as members and post-release or synthetic images as non-members [33]. Under this setup, state-of-the-art (SOTA) MI methods have achieved promising results (e.g., AUC > 80%).

However, this paper identifies a critical issue in current VLM MI benchmarks: **distribution shifts between member and non-member images introduce unintended shortcuts for inference**. These shifts stem from long-term temporal drift or discrepancies between real and synthetic image sources. We show that a simple image-only classifier (e.g., EfficientNet-B0 [58]), outperforms most SOTA MI methods—without accessing any VLM outputs (Table 1). Moreover, when evaluating on *pseudo-MI* datasets where all samples share the same membership status but differ in distribution, MI methods still perform well in distinguishing subsets, indicating that their success is driven largely by distributional artifacts rather than genuine membership signals (Table 2).

Thus, to ensure reliable inference, it is critical that member and non-member images are drawn from the same underlying distribution. Unfortunately, distribution shifts in the visual domain are pervasive yet often imperceptible [21], arising in subtle or multifaceted ways [64]. To support debiasing efforts, we systematically analyze concrete forms of shift by interpretably encoding images with a visual bag-of-words (capturing high-level semantics) and a frequency-domain energy profile (capturing low-level textures). To quantify these discrepancies, we further introduce WiRED—a principled metric that measures the ratio of sliced Wasserstein distances across embedding spaces tailed to different types of shift.

Building on this, we construct a suite of unbiased MI datasets by carefully examining four open-source VLM families—LLaVA-1.5 [35], LLaVA-OneVision [31], Cambrian-1 [60], and Molmo [12]. Leveraging random train/test splits from pretraining and instruction-tuning data, the member and non-member images are ensured to be drawn i.i.d. Under these conditions, SOTA MI methods perform only slightly better than random guessing. To assess the true auditing potential of MI, we design a series of classifiers (e.g., attention pooling probes) to directly inspect the VLM embedding space—where memorization signals are expected to reside. Even in this idealized setting, separability remains poor. We further estimate the *theoretical upper bound* of inference accuracy via Bayes optimality and observe that the irreducible error remains substantial.

These sobering results raise a central question: why is MI particularly challenging for VLMs? Through careful analysis, we identify three key factors. First, the massive data volume and single-epoch training lead to minimal overfitting. Second, model developers relabel images with high-quality captions, rendering the ground-truth text inaccessible. Third, inherent attributes of a single image result in high-variance token confidences, diluting the membership signal. To address these challenges, we relax standard assumptions and propose three practical scenarios: multi-epoch fine-tuning on downstream tasks, access to ground-truth text, and aggregation-based set inference. Under all three settings, MI becomes not only feasible but also practically valuable for real-world data auditing.

Our analytical framework is illustrated in Figure 1. The key contributions of this work are as follows:

- We identify distribution shifts in existing VLM MI benchmarks, characterize their concrete forms, and introduce a principled metric WiRED to quantify them (§ 3).
- We construct unbiased MI datasets with i.i.d. splits, where SOTA MI methods perform marginally better than chance. We further estimate the Bayes optimality in VLM embedding space to assess the theoretical limits of MI (§ 4).
- We analyze why MI is particularly challenging for VLMs and identify three realistic scenarios in which MI becomes feasible and practically relevant for data auditing (§ 5).

2 Related Work

2.1 Large Vision-Language Models

Built upon powerful LLMs [50], proprietary VLMs such as GPT-4V [51] have shown impressive performance on open-domain multimodal tasks. Open-source efforts like LLaVA [36] follow a simple design, aligning vision and language via a vision encoder $g_\psi(\cdot)$, a projector $p_\theta(\cdot)$, and a language model $f_\phi(\cdot)$. Given an image X_v and instruction X_q, features $Z_v = g_\psi(X_v)$ are projected to visual tokens $H_v = p_\theta(Z_v)$ and input to the LLM with X_q for response generation. Training involves modality alignment on large-scale datasets (e.g., LAION-5B [55], CC12M [28], Datacomp [18]) and instruction tuning on curated multimodal QA tasks.

To promote transparency, recent open-source VLMs such as LLaVA-1.5 [35], LLaVA-OneVision [31], Cambrian-1 [60], and Molmo [12] release not only model weights but also complete training data. As web-crawled image-text pairs are often noisy or short, these models re-annotate images with synthetic or human-curated texts. LLaVA-OneVision generates 99.8% of its high-quality knowledge with proprietary VLMs [8], while Molmo constructs its pretraining data from human-transcribed speech descriptions of web images.

2.2 Membership Inference on Language Models

As foundational models are trained on large-scale web data, concerns about unauthorized use of copyrighted or private content have intensified [14, 30]. Membership inference (MI) has emerged as a non-intrusive auditing tool to detect training data exposure [57]. Early MI methods relied on simple statistics like perplexity [7, 62], while recent work proposes token-level metrics like *Min-K%* [56] and *Min-K%++* [65] for improved stability. In the VLM setting, the standard threat model assumes black-box access, with only an image X_v and crafted instruction X_q available. Auditors collect output token probabilities and compute MI scores. VL-MIA [33] builds the first VLM MI benchmark and introduces *MaxRényi-K%*, while image-only inference [26] assesses membership via self-consistency of sampled descriptions. As these methods produce only MI scores rather than binary decisions, evaluation relies on threshold-free metrics such as AUC and TPR@5%FPR, while real-world deployment requires reference sets to calibrate decision thresholds.

Despite advances in *LLM MI*, temporal biases have been revealed in evaluation datasets [41], where non-members contain low-confidence, out-of-vocabulary tokens. We reveal a similar yet

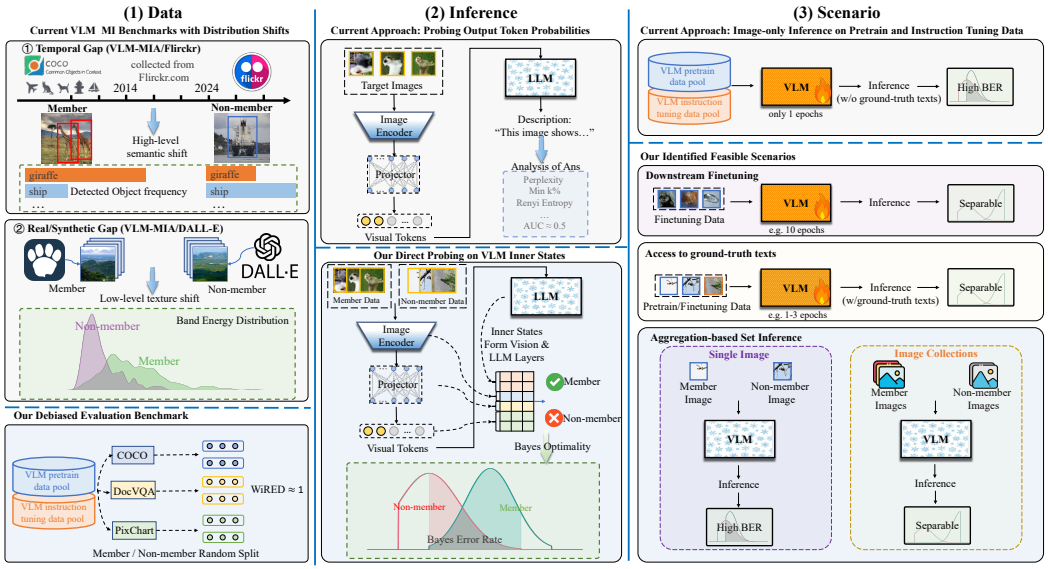

Figure 1: Revisiting of VLM MI: (1) Identifying Bias in MI Datasets, (2) Probing Bayes Optimality in VLM Inner States, (3) Future Scenarios for Data Auditing.

subtler shift in *VLM MI*: member and non-member images exhibit complex, vision-specific biases difficult to detect or interpret. Unlike textual shifts explainable by out-of-vocabulary words [11], these visual biases are largely opaque. We introduce a principled framework to quantify such shifts and assess their impact on data auditing.

3 Distribution Shortcuts in VLM Membership Inference Benchmarks

We identify that in current VLM MI benchmarks, distribution shifts between member and non-member images contribute significantly more to their separability than the genuine membership. On benchmarks such as VL-MIA/Flickr or VL-MIA/DALL-E, an EfficientNet-B0 [58] trained solely on 300 images—without any access to VLM outputs—achieves a test AUC well above most SOTA MI methods. Moreover, MI methods yield high AUCs when comparing subsets with identical membership yet different distributions (§ 3.2).

The distribution shift not only compromises evaluations but also limits the utility of MI. When target images deviate in distribution from the reference set, the derived threshold becomes misaligned, causing substantial false positives and false negatives. To address this, we provide a thorough analysis of the concrete form of distribution shifts in VL-MIA/Flickr and VL-MIA/DALL-E (§ 3.3), and introduce a principled measure of distribution discrepancy (§ 3.4).

3.1 Experimental Settings

3.1.1 Datasets. Our study builds upon the existing VLM MI benchmark: VL-MIA/Flickr and VL-MIA/DALL-E, each comprising 300 member and 300 non-member images [33]. In VL-MIA/Flickr, member images are sourced from MS COCO [34], a dataset widely used in training open-source VLMs, while non-member images are scraped from Flickr [17] after 2024. In VL-MIA/DALL-E, the member set is drawn from LAION [55], while non-members are synthesized by DALL-E [49] using the exact captions associated with members.

3.1.2 Vision-Language Models. We consider representative VLM families with fully open data and models, enabling training data traceability [12]. In this section, we evaluate LLaVA-1.5-7B [35, 36], LLaVA-ov-7B [31], Cambrian-1-8B [60], and Molmo-7B-D [12].

3.1.3 MI Methods. We consider MI methods from classical approaches such as *perplexity* and *maximum probability gap* [7, 62], to recent SOTA for LLMs, *Min-K%* [56] and *Min-K%++* [65]. We also include VLM-specific methods: *MaxRényi-K%* and *ModRényi* [33], as well as the *Image-only Inference* [26]. All methods interact with VLMs via their language interface, simulating realistic black-box auditing: given the target image and a crafted instruction, the auditor infers membership from the language response and confidence.

3.1.4 The Blind Classifier. To isolate the impact of distribution shift from genuine membership signals, we evaluate a *blind* classifier without access to any VLM outputs. An EfficientNet-B0 [58] is trained directly on images from VL-MIA/Flickr and VL-MIA/DALL-E to distinguish members from non-members, using a 1:1 split of 300 training and 300 testing samples per dataset.

3.2 Distribution Shifts Surpass Membership Signals in Separability

Table 1 reports the performance of MI methods on the VL-MIA suite. Following standard practice [26, 33, 56, 65], the Area Under the ROC Curve (AUC) and True Positive Rate at 5% False Positive Rate (TPR) are used as evaluation protocols, where higher values indicate better inference. Surprisingly, most of the times, the EfficientNet-B0 trained solely on images achieves significantly higher scores than SOTA MI methods, indicating that distribution shifts between member and non-member images alone are sufficient for separation, independent of any memorization signals from the VLM.

However, could the supervised training simply encourage EfficientNet to exploit distributional cues, while SOTA MI methods genuinely reflect the VLM's overfitting behavior? To disentangle the effects of distribution shift from true membership signals, we construct two *pseudo-MI* datasets by swapping subsets between VL-MIA/Flickr and VL-MIA/DALL-E: in VL-MIA/Member, all images come from the training set—VL-MIA/Flickr members as members, and VL-MIA/DALL-E members as *pseudo-nonmembers*. Similarly, in VL-MIA/Nonmember, all images are unseen during training, with

Table 1: Performance of MI methods on Existing MI Datasets.

Dataset	Method		LLaVA-1.5		LLaVA-ov		Molmo		Cambrian	
			AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR
Flickr (WiRED =2.10)	Perplexity	inst	34.3	0.7	18.1	0.0	34.6	1.3	34.4	1.7
		desp	57.4	16.0	14.5	0.0	55.8	12.0	43.3	4.0
	Max-Prob-Gap	inst	57.5	7.3	21.8	0.7	67.3	11.0	9.4	0.3
		desp	61.0	16.7	62.1	9.3	56.8	10.3	36.7	1.3
	Min-K%	inst	48.2	1.7	59.9	10.0	65.1	12.7	55.1	4.0
		desp	57.3	16.0	20.1	0.0	55.8	12.3	59.4	12.0
	Min-K%++	inst	54.5	13.7	14.7	0.3	65.5	13.3	44.7	1.7
		desp	61.8	21.7	80.1	14.0	67.9	16.3	50.7	1.0
	MaxRényi-K%	inst	66.6	18.7	48.3	8.3	69.3	18.7	68.4	16.3
		desp	57.4	19.3	95.5	79.0	64.3	19.0	55.5	7.7
	ModRényi	inst	38.4	0.3	84.2	58.7	32.7	0.7	41.4	1.3
		desp	57.1	12.3	15.7	0.0	53.0	10.0	45.5	3.7
	Image-only Inference		64.0	8.7	21.6	6.5	66.8	13.5	76.1	31.6
	Blind Classifier (ours)		**99.1**	**97.4**	**99.1**	**97.4**	**99.1**	**97.4**	**99.1**	**97.4**
DALL-E (WiRED =3.00)	Perplexity	inst	37.7	2.0	58.4	17.0	59.0	11.3	24.6	1.0
		desp	65.5	8.0	55.5	17.3	55.9	12.0	71.2	12.7
	Max-Prob-Gap	inst	58.9	9.3	63.3	16.3	67.8	15.3	89.3	66.3
		desp	64.8	7.7	58.0	17.3	58.4	14.7	78.5	38.3
	Min-K%	inst	39.3	6.3	60.5	18.7	63.5	11.3	33.9	1.0
		desp	65.6	8.0	65.7	18.0	55.9	12.0	70.8	12.0
	Min-K%++	inst	58.5	10.3	54.7	12.7	48.2	7.0	87.7	66.7
		desp	67.0	9.3	49.6	10.7	57.6	11.0	75.4	14.7
	MaxRényi-K%	inst	71.8	14.3	78.7	30.3	54.5	7.7	**94.9**	**87.7**
		desp	70.3	9.0	64.0	14.3	57.9	14.7	87.1	56.3
	ModRényi	inst	38.4	3.0	71.9	29.0	51.0	9.3	25.5	5.0
		desp	64.6	9.0	62.9	24.3	55.0	12.3	84.0	47.3
	Image-only Inference		47.0	4.9	67.3	14.2	63.7	13.7	41.0	0.0
	Blind Classifier (ours)		**87.4**	**49.1**	**87.4**	**49.1**	**87.4**	**49.1**	87.4	49.1

VL-MIA/DALL-E nonmembers as nonmembers, and VL-MIA/Flickr nonmembers as *pseudo-members*. While membership status is uniform within each dataset, distributional discrepancies remain.

Table 2: Performance of MI methods on *Pseudo-MI* Datasets

Dataset	Method		LLaVA-1.5		LLaVA-ov		Molmo		Cambrian	
			AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR
Member (WiRED =2.42)	Perplexity	inst	42.4	0.7	7.9	0.0	20.4	0.0	54.5	1.0
		desp	63.9	7.0	20.4	0.0	51.9	2.0	27.2	1.3
	Max-Prob-Gap	inst	53.8	0.3	4.0	0.0	37.0	0.3	1.7	0.0
		desp	58.9	9.3	60.1	1.0	46.7	0.7	14.9	0.0
	Min-K%	inst	42.6	1.0	48.2	0.0	39.1	0.7	65.1	6.3
		desp	65.2	7.0	23.7	0.0	55.6	6.0	76.8	21.7
	Min-K%++	inst	76.0	16.0	31.1	0.0	77.5	7.7	17.4	0.0
		desp	65.2	19.7	81.7	30.0	45.2	3.7	29.5	0.0
	MaxRényi-K%	inst	84.8	38.0	66.0	11.0	82.0	29.7	54.8	3.0
		desp	78.7	18.0	90.9	52.3	55.4	5.3	45.4	4.7
	ModRényi	inst	38.6	0.0	84.0	40.3	18.4	0.0	65.8	2.7
		desp	63.9	7.3	18.7	0.0	53.4	3.0	37.3	1.3
	Image-only Inference		68.2	19.7	20.0	4.4	41.6	2.0	59.9	14.9
	Blind Classifier (ours)		**97.3**	**85.5**	**97.3**	**85.5**	**97.3**	**85.5**	**97.3**	**85.5**
Non Member (WiRED =6.60)	Perplexity	inst	35.0	1.3	34.9	4.3	41.7	0.3	37.0	0.0
		desp	57.2	3.7	67.7	33.3	52.9	3.3	57.1	4.3
	Max-Prob-Gap	inst	55.0	2.7	22.9	0.3	41.8	1.3	50.8	7.7
		desp	55.5	2.3	56.9	10.0	48.8	4.0	54.0	6.0
	Min-K%	inst	45.2	6.7	65.7	22.0	46.2	0.3	44.3	1.3
		desp	57.3	4.3	72.2	34.0	60.1	8.7	57.0	8.0
	Min-K%++	inst	75.0	18.0	75.6	18.7	63.6	6.0	66.4	16.7
		desp	59.8	6.3	44.6	10.0	34.6	2.3	51.6	8.0
	MaxRényi-K%	inst	79.9	21.3	90.5	63.3	81.3	20.7	84.4	47.7
		desp	66.9	13.3	79.9	36.3	49.6	4.0	55.9	7.7
	ModRényi	inst	35.2	0.7	44.4	4.7	43.4	0.0	36.1	3.7
		desp	57.5	4.7	73.7	38.3	57.5	5.0	58.2	13.0
	Image-only Inference		48.6	8.1	58.2	5.8	40.3	1.8	31.4	0.0
	Blind Classifier (ours)		**99.6**	**98.4**	**99.6**	**98.4**	**99.6**	**98.4**	**99.6**	**98.4**

As shown in Table 2, despite the absence of genuine membership differences, MI methods retain high performance on these *pseudo-MI* datasets [1], matching or exceeding their performance on the original benchmarks. Notably, some methods yield AUCs well below 50% (random guessing). This reflects the assumption in MI methods that membership signals are directional—e.g., lower *perplexity* implies membership. However, distribution shifts can invalidate this assumption, resulting in inverted decisions. In general, AUCs near 50% indicate low separability.

[1] AUC > 65% and TPR > 15% are highlighted with underlines.

> Finding 1: Current VLM MI benchmarks exhibit distribution shifts between member and non-member images, acting as unintended shortcuts. SOTA MI methods rely on these shifts rather than genuine membership signals.

3.3 Understanding Distribution Discrepancy

The blind EfficientNet indicates the presence of distribution shifts. Yet, the exact nature of these shifts remains unclear. Recent studies [37] show that shifts are common in large-scale datasets, but often imperceptible to humans [37]. Auditors may be unaware that the target and reference images differ in distribution. Understanding the concrete form of shifts is therefore critical for debiasing. We analyze VL-MIA/Flickr and VL-MIA/DALL-E from two perspectives: high-level semantics and low-level textures.

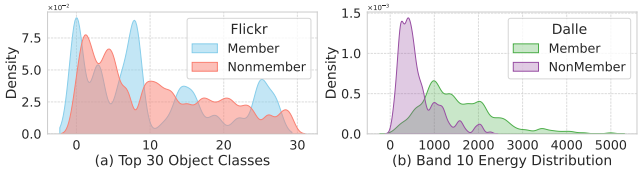

Figure 2: Distribution Shifts in Existing MI Datasets: (a) Flickr; (b) DALL-E.

VL-MIA/Flickr exhibits a temporal gap of over a decade, leading to semantic shifts in object content. Member images from MS COCO tend to depict natural scenes with a concentrated object distribution, while non-members—crawled from Flickr after 2024—contain more man-made objects and a flatter distribution. We train a YOLOv11 detector [61] on LVIS [22] (1,203 categories) and extract object annotations from both subsets. As shown in Figure 2(a), the category frequency distributions differ significantly. We further encode each image as a 1,203-dimensional sparse vector of object counts and train a random forest [6] to distinguish members from non-members. This visual bag-of-words classifier achieves an AUC of 81.96%, demonstrating that high-level semantic shifts alone are sufficient to separate the two subsets in VL-MIA/Flickr.

In contrast, **VL-MIA/DALL-E** controls high-level semantics by generating non-member images with the same captions as members. Thus, the visual bag-of-words classifier performs poorly, with an AUC of 53.15%. However, real and AI-generated images often differ in low-level details [40]. We analyze high-frequency features known to capture fine-grained visual cues [64]. Given an image with pixel intensity matrix $I \in \mathbb{R}^{H\times W}$, the centered magnitude spectrum of its 2D discrete Fourier transform is computed as:

$$\mathcal{F}(u,v) = \sum_{x=0}^{H-1}\sum_{y=0}^{W-1} I(x,y)\cdot e^{-2\pi i\left(\frac{ux}{H}+\frac{vy}{W}\right)},\ M(u,v) = |\mathcal{F}(u,v)| \quad (1)$$

We divide the spectrum into K concentric frequency bands based on ℓ_2 distance from the center. The energy of the i-th band is:

$$E_i = \frac{1}{|\mathcal{B}_i|}\sum_{(u,v)\in\mathcal{B}_i} M(u,v) \quad (2)$$

where $\mathcal{B}_i$ is the set of frequency coordinates in the i-th band. This defines a frequency feature vector $\mathbf{E} = [E_0, E_1, \ldots, E_{K-1}] \in \mathbb{R}^K$. Figure 2(b) shows the distribution of high-frequency energy E_{10} for $K = 10$, revealing clear separation between members and non-members. Using $\mathbf{E}$ as a 10-dimensional input, a simple linear classifier achieves a test AUC of 96.21%—far exceeding all MI baselines.

Despite similar semantics, subtle texture differences are sufficient to separate member and non-member images in VL-MIA/DALL-E.

> Finding 2: Distribution shifts are ubiquitous yet hard to detect. They can arise from temporal gaps or differences in data collection, and manifest in diverse forms—from subtle texture patterns to high-level semantic variations.

3.4 Quantifying Distribution Discrepancy

Accurately quantifying distribution shift is essential for building fair MI benchmarks and thresholds calibration. While shifts are pervasive in large-scale datasets, effective and interpretable metrics remain underexplored [37, 64]. One straightforward approach is to train a deep classifier (e.g., EfficientNet) to distinguish subsets—an instantiation of classifier two-sample tests [39]. However, this is computationally expensive for personal auditing and underestimates subtle shifts, e.g., in VL-MIA/DALL-E, EfficientNet yields a significantly lower AUC than a frequency-based linear classifier.

Thus, to facilitate fair benchmarks and practical auditing, we propose a principled, interpretable, and efficient metric for distributional discrepancies. We introduce **WiRED**—Wasserstein Ratio of Embedded Representations—which measures the degree of shift between two image subsets S_1 and S_2. Specifically, WiRED first embeds each image I into a collection of metric spaces via embedding functions $\phi_1, \ldots, \phi_t$, each targeting a distinct form of shift. Let p_1 and p_2 denote the probability densities of S_1 and S_2 in the i-th embedding space. The Wasserstein distance [46] is defined as:

$$W_q(p_1, p_2) = \left(\inf_{\gamma \in \Gamma(p_1, p_2)} \mathbb{E}_{(x_1, x_2) \sim \gamma} \|x_1 - x_2\|^q \right)^{1/q}, \tag{3}$$

where $\Gamma(p_1, p_2)$ denotes all couplings between p_1 and p_2. Intuitively, W_q captures the minimal cost of transporting one distribution into the other, commonly referred to as the Earth Mover's Distance [46]. As computing W_q exactly requires solving the optimal transport problem with time complexity $O(N^3)$ for sample size N, we adopt the sliced Wasserstein distance (SWD) [5] as an efficient approximation, which projects samples onto random directions $\{\theta_j\}_{j=1}^K \in \mathbb{R}^d$, computes one-dimensional Wasserstein distances, and averages them across all directions:

$$\mathrm{SWD}(S_1, S_2) = \frac{1}{K} \sum_{j=1}^{K} W_q \left(\theta_j^\top \phi_i(S_1), \theta_j^\top \phi_i(S_2) \right). \tag{4}$$

To normalize the discrepancy between S_1 and S_2, we compare their distance to the internal variation within S_1. Specifically, we sample two disjoint subsets $S_{11}, S_{12} \subset S_1$, along with a size-matched subset $S_2' \subset S_2$. The WiRED score in the i-th embedding space is then defined as $\mathrm{WiRED}_i = \frac{\mathrm{SWD}(S_{11}, S_2')}{\mathrm{SWD}(S_{11}, S_{12})}$.

This highlights how distinguishable S_2 is from S_1, relative to S_1's internal variation. A ratio close to 1 indicates similar distributions, while a significantly higher value signals a notable shift. Since each embedding function ϕ_i captures different aspects of the data, we define the final WiRED score as the maximum across all embeddings, i.e., $\mathrm{WiRED} = \max_{i \in [t]} \mathrm{WiRED}_i$.

In our experiments, we instantiate ϕ_i with two embedding functions: (1) ImageNet-pretrained EfficientNet-B0 features to capture high-level semantics, and (2) frequency-domain energy vectors (§ 3.3) to capture low-level textures. WiRED is non-parametric, makes no assumptions about distribution shapes, and is highly efficient (e.g., taking 10 seconds for a 600-image MI dataset). We report WiRED scores alongside each benchmark, clearly reflecting the distributional biases identified in § 3.2 (e.g., WiRED $\gg 1$), without model training or prior knowledge of the shift.

4 Feasibility of Membership Inference on VLMs

In § 3, we show that current benchmarks suffer from distribution shifts, introducing unintended shortcuts. When such shifts are eliminated, can MI reliably detect membership based on overfitting signals in VLM outputs? In this section, we investigate this question and uncover that, under strictly i.i.d. conditions, SOTA MI methods perform only slightly better than random guessing (§ 4.2). Even with white-box access to VLM internal features—the presumed source of memorization—separability remains poor, and the theoretical upper bound of performance, quantified by the estimated Bayes Error Rate (BER), remains pessimistic (§ 4.3). These results suggest that in the most realistic auditing scenario—where the auditor must determine whether a single image appeared in VLM training—current MI techniques are unlikely to yield reliable conclusions.

4.1 Towards Unbiased VLM MI Datasets

The most rigorous way to construct i.i.d. subsets is through random splits from the same source [16]. Following this principle, we carefully inspect open-source VLMs to identify datasets with standard training/testing splits. We focus on the fully open VLM families—LLaVA-1.5 [35], LLaVA OneVision [31], Cambrian-1 [60], and Molmo [12]—all trained exclusively on publicly available data. VLMs first align vision and language representations during pretraining, and are further tuned for instruction following. These phases typically utilize all available data, without held-out validation sets. Fortunately, many VLMs incorporate widely used captioning and VQA datasets that do provide standard splits. For instance, all four families use MS COCO [34], primarily during pretraining, and both LLaVA OneVision and Molmo include instruction-tuning datasets built on clearly partitioned VQA benchmarks.

Table 3: Quantitative Debiasing Validation: Blind Clf Performance and WiRED.

datasets	Models	blind clf AUC	blind clf TPR	WiRED
COCO	All models	49.0±0.6	8.3±1.0	0.97
ChartQA	LLaVA-ov	57.7±0.7	8.3±1.6	1.04
DocVQA		54.9±1.0	6.2±2.5	0.98
InfoVQA		57.0±1.4	7.2±0.9	1.25
PixMoChart	Molmo	48.4±0.1	1.7±0.6	1.00
PixMoDiagram		48.5±0.3	6.8±1.2	1.22
PixMoTable		57.1±1.2	9.8±0.3	0.94

To ensure that member images were seen during training while non-members were not, we select datasets explicitly used in both training and evaluation, excluding any with evident distribution shifts between splits. The resulting datasets are listed in Table 3. Following the VL-MIA setup [33], we sample 300 images each from the training and testing splits to form the member and non-member sets. We apply both EfficientNet and WiRED metric to assess shifts. Across all selected datasets, EfficientNets yield AUCs near 50%, and WiRED scores remain close to 1, confirming well-matched

distributions. ChartQA [43], DocVQA [44], and InfoVQA [44] show slightly elevated AUCs, though they adhere to standard splits. We attribute this to image redundancy in the original VQA datasets; our de-duplication may introduce minor residual imbalance. This effort represents an initial step toward reliable benchmarks, paving the way for larger and more rigorous datasets in future research.

4.2 VLM MI Performance on Unbiased Datasets

Table 4: Performance of MI Methods on the Debiased COCO Dataset.

Method		LLaVA-1.5		LLaVA-ov		Molmo		Cambrian	
		AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR
Perplexity	inst	51.2	6.3	50.7	7.0	49.3	5.7	55.3	8.3
	desp	53.7	1.7	51.5	8.0	54.3	5.3	54.3	5.7
Max-Prob-Gap	inst	50.0	6.0	53.5	8.3	45.5	3.0	50.6	5.3
	desp	53.0	5.7	53.5	4.3	51.8	8.3	54.5	8.0
Min-K%	inst	51.2	7.3	50.7	7.0	51.5	5.7	55.8	8.7
	desp	54.2	3.3	52.6	8.0	54.8	7.0	54.1	7.3
Min-K%++	inst	50.3	4.3	52.4	6.3	51.6	6.3	52.5	7.3
	desp	53.4	8.3	52.9	8.7	51.0	4.3	49.9	4.0
MaxRényi-K%	inst	51.5	9.3	54.8	11.0	52.9	6.3	51.7	7.3
	desp	54.3	5.0	56.8	6.3	53.0	5.0	52.2	7.7
ModRényi	inst	52.4	7.3	53.1	7.7	48.8	8.0	54.8	7.3
	desp	53.7	3.3	53.6	11.7	54.9	7.0	56.1	5.3
Image-only Inference		52.5	4.6	51.7	8.3	53.3	7.1	50.5	6.1
Blind Classifier (ours)		49.0	8.3	49.0	8.3	49.0	8.3	49.0	8.3

Table 4 reports the performance of MI methods on COCO across four VLMs. AUCs hover around 50% (random guessing), never exceeding 60%, while TPR@5%FPR remains below 10% in most cases—indicating poor separability between members and non-members. Table 5 presents results on the instruction-tuning datasets of LLaVA OneVision and Molmo. Although these datasets are introduced at later training stages—where catastrophic forgetting is expected to be less severe—separability remains weak.

Table 5: MI Methods Performance on Debiased Model-Specific MI Datasets.

Method		LLaVA-OneVision						Molmo					
		ChartQA		DocVQA		InfoVQA		PixChart		PixDigram		PixTable	
		AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR
Perpl	in	47.8	2.0	47.8	1.0	39.3	1.4	48.8	1.7	49.3	3.0	51.0	6.0
	de	53.9	4.3	54.2	2.7	50.7	4.8	53.1	5.0	52.2	9.3	46.7	5.0
Max Gap	in	41.7	2.7	57.9	8.7	56.4	10.2	49.7	6.7	52.1	4.7	51.9	8.3
	de	55.1	9.0	55.9	6.3	50.4	4.8	50.5	4.7	49.2	3.7	47.4	4.0
Min K%	in	52.9	5.3	52.6	4.0	41.4	1.4	48.8	4.0	50.8	4.0	52.9	10.0
	de	54.1	4.7	54.2	4.3	52.3	6.8	53.2	5.7	53.3	10.3	52.1	6.7
Min K%++	in	53.7	8.0	50.5	4.0	58.7	10.2	52.6	7.7	48.6	4.0	50.4	7.7
	de	53.9	7.0	52.4	6.7	52.8	5.4	47.1	4.3	50.4	5.3	47.0	4.7
Max Rényi	in	48.6	5.3	56.6	14.7	58.4	8.8	51.9	7.7	55.1	7.0	49.2	6.7
	de	53.3	4.7	54.0	4.7	53.4	8.2	52.3	8.7	54.7	7.3	51.5	6.7
Mod Rényi	in	53.1	5.7	46.1	4.0	45.1	5.4	51.4	7.0	49.6	6.3	50.0	5.0
	de	53.7	3.3	54.4	4.7	50.8	4.8	54.8	4.3	52.0	10.0	47.5	5.7
Img Infer		48.0	4.3	53.5	9.7	58.4	9.4	54.5	5.7	50.3	7.0	46.1	4.5
Blind Clf		57.7	8.3	54.9	6.2	57.0	7.2	48.4	1.7	48.5	6.8	57.1	9.8

In contrast, as shown in § 3.2, when distribution shifts are present, a simple classifier trained on just 300 samples can achieve near-perfect separation (AUC $\approx$ 100%). This stark contrast highlights how subtle the true membership signal is in VLM outputs compared to distributional artifacts. In this context, a natural question arises: is membership inference on VLMs truly feasible?

4.3 Probing the Envelope of MI Performance

To assess the feasibility of VLM membership inference, we consider an idealized setting where the auditor has full white-box access to the model's internal embeddings—the source of potential memorization signals. This enables us to examine whether members and non-members are separable in the representation space, and to estimate the theoretical upper bound of this separability, characterized by Bayes optimality [19, 24].

4.3.1 Probing VLM Embedding Space. Given a target image $I \in (0, 255)^{3\times H\times W}$, we extract all hidden states from the vision encoder and language decoder during description generation. For each layer v_i in the vision encoder, we collect token-wise visual features $\{\mathbf{h}_v^1, \mathbf{h}_v^2, \ldots\} \in \mathbb{R}^{d_v}$; for each layer l_i in the language decoder, we record hidden states generated during next-token prediction $t_{k+1} = p(\theta; I, t_{\leq k})$, denoted as $\{\mathbf{h}_l^1, \mathbf{h}_l^2, \ldots\} \in \mathbb{R}^{d_l}$. To evaluate the separability in these hidden states, we first apply average pooling to obtain fixed-size embeddings of dimension d_v and d_l for visual and language tokens, and then adopt standard probes [38], training both linear classifiers and multi-layer perceptrons (MLPs) to distinguish members (class 1) from non-members (class 0).

To mitigate potential information loss from global pooling, we further introduce an attention pooling classifier that adaptively aggregates the most informative tokens:

$$p(y_i = 1 \mid u, x_{\leq i}) = \sigma(\mathbf{w}^\top \bar{\mathbf{h}}_i) \tag{5}$$

where the aggregated representation $\bar{\mathbf{h}}_i$ is computed as:

$$\bar{\mathbf{h}}_i = \sum_{j=1}^{i} \alpha_{i,j}\mathbf{h}_j, \quad \alpha_{i,j} = \frac{\exp(\mathbf{q}^\top \mathbf{h}_j)}{\sum_{k=1}^{i} \exp(\mathbf{q}^\top \mathbf{h}_k)} \tag{6}$$

Here, $\mathbf{q} \in \mathbb{R}^d$ is a learnable query vector that attends to informative tokens. This enables the classifier to capture fine-grained membership signals that may be distributed across tokens.

4.3.2 Estimating MI Performance Against Bayes Optimality. While the probes assume full white-box access to the VLM—far beyond what is feasible with real-world APIs—they remain empirical in nature. One might conjecture that, as probes improve, members and non-members may eventually become separable. To investigate this, we consider the theoretical upper bound of MI: the irreducible error in distinguishing members from non-members based on VLM hidden representations, quantified by the *Bayes error rate* (BER) [19, 24]. Formally, BER is defined as the expected misclassification rate of the Bayes-optimal classifier under the task distribution D:

$$\beta_D = \mathbb{E}_{(x,y)\sim D}\left[1 - \max_k p(y = k \mid x)\right] \tag{7}$$

Alternatively, it can be interpreted as the minimal error rate achievable over all measurable functions h:

$$\beta_D = \min_{\text{measurable } h} \mathbb{E}_{(x,y)\sim D}\left[\mathbb{I}(h(x) \neq y)\right] \tag{8}$$

where $\mathbb{I}$ is the indicator function. Since VLM feature spaces do not follow simple, tractable distributions that permit analytical computation of BER, we adopt an efficient approximation [9] to estimate BER in distinguishing membership in VLM hidden states.

Specifically, we compute pairwise ℓ_2 distances between token features to construct an adjacency matrix $\mathbf{A}$, where $A_{ij} = 1$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ share the same label, and 0 otherwise. Connected components representing confident regions are identified via breadth-first search. Remaining unconnected samples are treated as uncertain and labeled using Label Spreading [69]. BER is then estimated as the fraction of incorrect predictions among these uncertain samples:

Table 6: BER and Performance of Probing Methods using Visual and Language Tokens on the Debiased COCO Dataset.

Model	Modal	BER		Linear			MLP			Attention		
		Original	Calibrated	ACC	AUC	TPR	ACC	AUC	TPR	ACC	AUC	TPR
LLaVA-1.5	Vision	33.8	22.3	53.3±3.8	55.4±2.1	8.0±2.2	53.6±3.4	55.3±0.3	10.8±3.0	52.2±5.3	53.4±2.8	4.5±1.5
	Language	26.3	N/A	50.0±3.1	52.2±5.6	11.5±4.6	50.6±2.6	52.2±5.2	8.8±6.7	48.6±2.7	51.2±4.9	9.2±3.0
Cambrian	Vision	36.8	22.3	52.8±2.4	50.4±3.5	5.3±1.4	52.2±3.7	50.1±3.0	3.6±2.6	50.0±2.4	50.7±1.5	11.1±4.9
	Language	23.3	N/A	55.0±5.6	54.1±5.8	11.7±3.3	51.9±2.6	52.2±4.3	5.9±5.9	53.9±5.5	52.9±6.6	13.5±4.7
LLaVA-OneVision	Vision	21.2	22.3	50.8±4.2	58.8±6.3	10.3±4.3	52.5±3.4	56.4±5.2	15.5±6.1	55.8±4.1	60.2±5.7	8.6±3.5
	Language	23.2	N/A	63.3±5.7	63.4±4.8	8.6±2.4	60.8±4.7	63.8±5.3	5.9±3.2	62.5±4.1	65.0±4.3	6.9±3.7
Molmo	Vision	25.5	22.3	55.3±3.7	56.3±4.2	13.0±1.3	52.8±2.1	53.4±4.2	14.0±6.3	51.7±5.6	55.6±3.9	15.3±4.8
	Language	21.7	N/A	50.3±3.5	52.1±4.1	10.0±1.5	51.1±1.4	52.0±3.3	8.1±8.0	50.3±2.7	53.6±3.5	12.3±3.5

$$\text{BER} = \frac{1}{n} \sum_{i \in \mathcal{U}} \mathbb{I}(\hat{y}_i \neq y_i) \tag{9}$$

where $\mathcal{U}$ denotes the set of uncertain samples, $\hat{y}_i$ the predicted label, and y_i the ground-truth. Note that BER provides a *optimistic* estimate of the lowest achievable error. The Bayes-optimal classifier is not accessible, and BER does not account for generalization—it may reflect separability based on spurious features.

Table 7: Performance of Probing Methods on Model-Specific Debiased Datasets.

Dataset	Method	Vision			Language		
		ACC	AUC	TPR	ACC	AUC	TPR
Chart QA	Linear	51.7±1.4	53.2±1.4	0.6±0.8	56.1±0.8	56.0±0.8	5.4±3.9
	MLP	49.2±1.8	53.0±0.7	6.4±2.8	57.2±4.2	56.0±2.7	2.3±3.3
	Attention	51.9±1.6	52.9±1.1	4.1±3.0	53.3±1.2	53.5±1.2	5.3±2.4
Doc VQA	Linear	52.8±0.8	54.7±3.0	6.5±0.9	46.1±1.7	46.8±1.3	6.4±2.8
	MLP	54.4±5.5	56.0±7.0	7.0±2.7	46.9±2.2	46.3±1.7	0.0±0.0
	Attention	53.3±0.0	55.8±2.6	4.7±2.2	46.7±0.7	46.7±0.1	7.1±1.6
Info VQA	Linear	66.4±3.8	73.3±3.3	25.5±7.8	67.5±0.8	71.3±3.6	15.3±3.8
	MLP	64.4±4.0	69.4±0.5	22.0±8.5	68.1±3.4	73.0±4.1	14.5±10.7
	Attention	69.5±3.5	74.2±3.6	29.7±4.4	66.1±1.4	72.0±2.3	9.6±2.4
Pix Chart	Linear	53.3±2.5	50.8±4.2	5.8±4.4	55.3±3.9	56.8±4.5	4.2±2.4
	MLP	51.9±5.2	49.8±5.3	7.1±1.3	56.7±4.8	57.4±4.1	2.5±3.5
	Attention	54.7±1.0	50.3±0.5	2.3±1.6	54.4±3.1	55.8±3.7	3.0±3.2
Pix Digram	Linear	50.8±5.1	48.7±4.2	6.5±2.3	53.9±6.1	50.3±6.1	10.1±1.1
	MLP	46.9±3.1	47.1±4.4	4.2±1.7	48.9±4.4	48.9±4.4	4.9±4.6
	Attention	51.9±3.1	54.6±1.4	10.2±6.2	48.6±2.7	49.9±3.0	8.3±2.3
Pix Table	Linear	46.4±4.0	49.2±5.0	5.3±1.3	46.4±1.6	46.1±0.9	5.9±4.6
	MLP	51.1±4.4	49.3±4.5	5.3±1.4	46.7±2.5	46.5±2.6	0.6±0.9
	Attention	50.0±6.2	50.0±4.9	3.5±3.7	45.3±1.4	45.7±1.6	8.2±4.5

4.3.3 Experimental Results. Tables 6 and 7 report the performance of three probes applied to visual and language tokens across our unbiased datasets. In most cases, both accuracy and AUC remain below 65%, and the sophisticated attention pooling classifier fails to yield noticeable improvements, indicating even at the source of memorization—internal representations of VLMs—members and non-members remain weakly separable. Furthermore, in most cases, BER falls between 20% and 30%, implying that the theoretical upper bound for MI is only around 70%. Notably, BER is an optimistic estimate that does not reflect practical generalization. As a sanity check, we project the same images into the feature space of an EfficientNet. Although all samples are non-members from EfficientNet's perspective, the calibrated BER (Cal) in Table 6 and 8 still ranges from 20% to 30%. This suggests that even with genuine membership, VLM representations offer marginally better separation.

4.3.4 Ablations. Our default setting uses final-layer features of 7B-scale VLMs. We conduct ablations on COCO with the LLaVA OneVision family, varying three key factors: layer depth, model scale, and output length. Figures 3(a)(b) show the accuracy of probes and the corresponding Bayes optimality (100 − BER) across layers. While deeper layers offer slightly improved separability, results remain near random guessing. Figure 3(c) evaluates models ranging from 0.5B to 72B parameters; despite the substantial increase in capacity, the largest model still fails to distinguish membership. Figure 3(d) explores the effect of output length, revealing that generating longer descriptions also does not improve separability.

Table 8: BER on Debiased Model-Specific MI Datasets.

Dataset	Vision		Language		Dataset	Vision		Language	
	Ori	Cal	Ori	Cal		Ori	Cal	Ori	Cal
ChartQA	24.0	32.7	23.0	N/A	PixChart	32.5	28.3	26.2	N/A
DocVQA	17.5	25.2	20.3	N/A	PixDigram	33.3	27.5	26.2	N/A
InfoVQA	20.5	20.9	15.9	N/A	PixTable	30.7	27.0	22.2	N/A

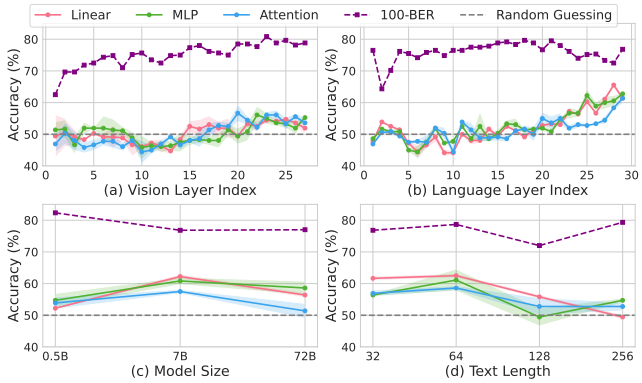

Figure 3: Ablation Performance of Probing Methods(LLaVA-ov on COCO).

Finding 3: Even with oracle access to internal states, membership signals remain subtle. Probing classifiers on hidden features yield only marginal gains, and Bayes optimality remains low, indicating minimal room for improvement.

5 When Does VLM MI Become Feasible?

We analyze why MI struggles in large VLMs and construct targeted scenarios to mitigate these challenges. Surprisingly, we find MI becomes feasible in these realistic auditing settings.

5.1 Finetuning on Downstream Tasks

5.1.1 Challenge: Minimal Overfitting. Traditional MI typically targets models trained for many epochs on specific downstream tasks. In contrast, LLMs and VLMs adopt general-purpose training objectives and often see each example only once [47]. Early VLMs like LLaVA-1 [36] performed multi-epoch training (e.g., 3), but recent models have shifted toward data scaling—training on massive,

high-quality datasets for a single epoch—which enables generalization without severe overfitting [2, 59]. Moreover, due to the sheer data volume, early examples are frequently forgotten during training [27]. Theoretically, as dataset size grows, model behavior converges across seen and unseen samples [42]. In § 4.3, we observe unexpectedly better membership separability in LLaVA OneVision on COCO, despite COCO being used during the more forgettable alignment phase. Upon closer inspection, we identify substantial image overlap among MS COCO [34], COCO Caption [10], and RefCOCO [63] in the training corpus. We hypothesize that this duplication amplifies MI signals and ask: can stronger MI emerge when VLMs are fine-tuned for multiple epochs?

5.1.2 Scenario: Finetuning on Downstream Tasks. To ensure i.i.d. splits, we continue using instruction-tuning images with random partitions. Since test splits provide only short text answers, we generate image descriptions using LLaVA OneVision-7B, then fine-tune LLaVA-1.5-7B with LoRA [25] for 10 epochs. As shown in Table 9,

Table 9: MI Performance on LoRA LLava-1.5 (gt Refers to Ground Truth).

Method	ChartQA				DocVQA			
	epoch 10 w/o gt		epoch 3 w/ gt		epoch10 w/o gt		epoch 3 w/ gt	
	AUC	TPR	AUC	TPR	AUC	TPR	AUC	TPR
Perplexity	60.8	16.3	78.2	8.3	65.2	18.7	75.3	19.0
Min-K%	60.9	16.7	78.7	14.3	65.2	18.7	76.2	26.7
ModRényi	60.8	16.7	79.3	14.3	64.9	18.0	76.7	20.7

MI performance improves markedly by the 10th epoch, with AUCs surpassing 0.6. Notably, traditional MI approaches such as perplexity emerge as competitive baselines, indicating substantial room for improvement. This setting reflects realistic auditing scenarios, where privacy-sensitive or proprietary data (e.g., in medical VQA [32]) are commonly involved during fine-tuning.

5.2 Access to Ground-Truth Text

5.2.1 Challenge: Lack of Ground-Truth Captions. In LLM-based MI, auditors typically query the model with a suspicious text $s = (\text{token}_1, \text{token}_2, \ldots, \text{token}_t)$ and records the output probabilities $p(\text{token}_k \mid \theta, \text{token}_{<k})$ via teacher forcing for inference. However, extending this to VLMs poses a key challenge: the training caption for a suspicious image is usually inaccessible. Web-scraped image-text pairs are noisy and short, while modern VLMs are trained on high-quality, relabeled captions [12, 31]. As a result, in most proprietary black-box VLMs, auditors cannot access the original training text. To bypass this, MI methods like VL-MIA [33] rely on the VLM's own generated captions. However, these are merely *pseudo ground-truths*. Due to snowballing prediction errors in autoregressive decoding [3], the outputs increasingly diverge from the true training text. Given this gap, an open question is: would MI attacks be more effective if ground-truth captions were available?

5.2.2 Scenario: Access to Ground-Truth Text. Due to the lack of i.i.d. train/test text splits in open datasets, we continue using the fine-tuned LLaVA-1.5-7B and evaluate the 3rd-epoch checkpoint with ground-truth text in Table 9. Since LoRA updates less than 3% of parameters, free-form generation at early epochs yields near-random MI results. In contrast, ground-truth text enables significantly stronger MI, with AUCs nearing 0.8—outperforming free-form results even after 10 epochs. This underscores MI's value in detecting test set contamination [53] and verifying model ownership via curated samples [67].

5.3 Aggregation-Based Set Inference

5.3.1 Challenge: Diverse Image Attributes. § 3 shows MI is sensitive to distribution shifts. The diverse sources of VLM training data introduce natural image-level variation—quality, noise, object count, and complexity—that can dominate token distributions, overshadowing subtle membership signals. However, when members and non-members are drawn i.i.d., such high-variance factors may average out at the set level. This motivates: can aggregated MI signals across multiple images enable reliable set-level inference?

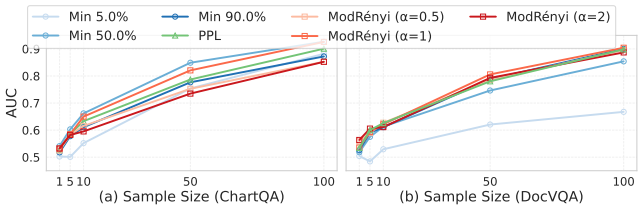

Figure 4: MI Performance in Aggregation-based Set Inference on LLaVA-ov.

5.3.2 Scenario: Aggregation-Based Set Inference. We treat a group of images as a single MI unit by averaging MI scores. Inspired by bootstrapping [15], we sample 1,000 sets with replacement from member and non-member pools, varying set sizes from 1 to 100. Figure 4 shows MI AUCs on LLaVA-OneVision. Even when single-image AUCs are only slightly above chance, set-level performance improves markedly. This approach is well-suited for auditing image collections—e.g., social media albums or artist portfolios [23]—and detecting unauthorized use of proprietary datasets [13]. Notably, small per-image gains can translate into substantial set-level improvements, motivating further development of MI techniques.

> **Finding 4:** VLM MI becomes feasible when: (1) fine-tuning induces overfitting; (2) ground-truth text is available; or (3) predictions are aggregated—key scenarios for auditing test set contamination and collection copyright infringement.

6 Conclusion

In this work, we identify a critical issue in current MI benchmarks for large VLMs: distribution shifts between member and non-member images introduce spurious shortcuts that overshadow true membership signals. We analyze these shifts and propose a principled metric to quantify them, enabling practical MI auditing. To build an unbiased testbed, we reconstruct i.i.d. member/non-member splits from open-source VLMs. Under this setting, existing MI methods perform only slightly above chance. We further assess the theoretical upper bound of membership separability and find a high irreducible Bayes error, underscoring the fundamental difficulty of MI on VLMs. Despite these challenges, we identify three practical scenarios where MI remains feasible and valuable for auditing: fine-tuning, access to ground-truth text, and aggregation across samples. Future work will explore additional viable or adversarial settings [66, 68], design stronger MI methods, and extend our study to closed-source VLMs via API access.

Acknowledgements

The work was supported by the National Natural Science Foundation of China under Grant 62271307 and 61771310.

References

[1] Anthropic. 2024. Introducing Claude 3.5 Sonnet. https://www.anthropic.com/news/claude-3-5-sonnet

[2] Antonis Antoniades, Xinyi Wang, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 2025. Generalization v.s. Memorization: Tracing Language Models' Capabilities Back to Pretraining Data. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

[3] Gregor Bachmann and Vaishnavh Nagarajan. 2024. The Pitfalls of Next-Token Prediction. In *International Conference on Machine Learning*. PMLR, 2296–2318.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[5] Sebastian Bischoff, Alana Darcher, Michael Deistler, Richard Gao, Franziska Gerken, Manuel Gloeckler, Lisa Haxel, Jaivardhan Kapoor, Janne K Lappalainen, Jakob H Macke, et al. 2024. A Practical Guide to Sample-based Statistical Distances for Evaluating Generative Models in Science. *Transactions on Machine Learning Research* (2024).

[6] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[7] 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)*. 2633–2650.

[8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*. Springer, 370–387.

[9] Qingqiang Chen, Fuyuan Cao, Ying Xing, and Jiye Liang. 2023. Evaluating classification model against Bayes error rate. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 8 (2023), 9639–9653.

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[11] Debeshee Das, Jie Zhang, and Florian Tramèr. 2024. Blind baselines beat membership inference attacks for foundation models. *arXiv preprint arXiv:2406.16201* (2024).

[12] Matt Deitke, Christopher Clark, Sangho Lee, and et al. 2025. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Linkang Du, Xuanru Zhou, Min Chen, Chusong Zhang, Zhou Su, Peng Cheng, Jiming Chen, and Zhikun Zhang. 2024. SoK: Dataset Copyright Auditing in Machine Learning Systems. In *2025 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 25–25.

[14] André V Duarte, Xuandong Zhao, Arlindo L Oliveira, and Lei Li. 2024. DE-COP: detecting copyrighted content in language models training data. In *Proceedings of the 41st International Conference on Machine Learning*. 11940–11956.

[15] Bradley Efron and Trevor Hastie. 2021. *Computer age statistical inference, student edition: algorithms, evidence, and data science*. Vol. 6. Cambridge University Press.

[16] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Jacob Steinhardt, and Aleksander Madry. 2020. Identifying statistical bias in dataset replication. In *International Conference on Machine Learning*. PMLR, 2922–2932.

[17] Flickr. 2025. Flickr: Online photo management and sharing application. https://www.flickr.com.

[18] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems* 36 (2023), 27092–27112.

[19] Frederick D Garber and Abdelhamid Djouadi. 1988. Bounds on the Bayes classification error based on pairwise risk functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 10, 2 (1988), 281–288.

[20] Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. 2024. Worldgpt: Empowering llm as multimodal world model. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 7346–7355.

[21] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.

[22] Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5356–5364.

[23] Anna Yoo Jeong Ha, Josephine Passananti, Ronik Bhaskar, Shawn Shan, Reid Southen, Haitao Zheng, and Ben Y Zhao. 2024. Organic or diffused: Can we distinguish human art from ai-generated images?. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 4822–4836.

[24] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer, New York, NY. https://link.springer.com/content/pdf/10.1007/978-0-387-84858-7.pdf

[25] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[26] Yuke Hu, Zheng Li, Zhihao Liu, Yang Zhang, Zhan Qin, Kui Ren, and Chun Chen. 2025. Membership Inference Attacks Against Vision-Language Models. In *Proceedings of the 34th USENIX Security Symposium*. USENIX Association.

[27] Matthew Jagielski, Om Thakkar, Florian Tramer, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Guha Thakurta, Nicolas Papernot, and Chiyuan Zhang. 2023. Measuring Forgetting of Memorized Training Examples. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=7bJizxLKrR

[28] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*. PMLR, 4904–4916.

[29] Yuri Kageyama. 2025. ChatGPT's viral Studio Ghibli-style images highlight AI copyright concerns. https://apnews.com/article/studio-ghibli-chatgpt-images-hayao-miyazaki-openai-0f4cb487ec3042dd5b43ad47879b91f4

[30] Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems* 36 (2023), 20750–20762.

[31] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research* (2025).

[32] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems* 36 (2023), 28541–28564.

[33] 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 (2024), 98645–98674.

[34] 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: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, 740–755.

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

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.

[37] Zhuang Liu and Kaiming He. 2025. A Decade's Battle on Dataset Bias: Are We There Yet?. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[38] Zhenhua Liu, Tong Zhu, Chuanyuan Tan, Bing Liu, Haonan Lu, and Wenliang Chen. 2024. Probing Language Models for Pre-training Data Detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1576–1587.

[39] David Lopez-Paz and Maxime Oquab. 2017. Revisiting Classifier Two-Sample Tests. In *International Conference on Learning Representations*.

[40] Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. 2023. Seeing is not always believing: Benchmarking human and model perception of ai-generated images. *Advances in Neural Information Processing Systems* 36 (2023), 25435–25447.

[41] Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. 2024. LLM Dataset Inference: Did you train on my dataset? *Advances in Neural Information Processing Systems* 37 (2024), 124069–124092.

[42] Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. 2021. Dataset Inference: Ownership Resolution in Machine Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=hvdKKV2yt7T

[43] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*. 2263–2279.

[44] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2200–2209.

[45] Ruaridh Mon-Williams, Gen Li, Ran Long, Wenqian Du, and Christopher G Lucas. 2025. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence* (2025), 1–10.

[46] Eduardo Fernandes Montesuma, Fred Maurice Ngolè Mboula, and Antoine Souloumiac. 2025. Recent Advances in Optimal Transport for Machine Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, 2 (2025), 1161–1180. doi:10.1109/TPAMI.2024.3489030

[47] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2023. Scaling data-constrained language models. *Advances in Neural Information Processing Systems* 36 (2023), 50358–50376.

[48] Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. 2025. Scalable Extraction of Training Data from Aligned, Production Language Models. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=vjel3nWP2a

[49] OpenAI. 2021. DALL·E: Creating Images from Text. https://openai.com/dall-e.

[50] OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt/

[51] OpenAI. 2023. GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf

[52] OpenAI. 2024. GPT-4o System Card. arXiv preprint arXiv:2410.21276.

[53] 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*. https://openreview.net/forum?id=KS8mIvetg2

[54] Kylie Robison. 2025. Meta got caught gaming AI benchmarks. https://www.theverge.com/meta/645012/meta-llama-4-maverick-benchmarks-gaming

[55] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems* 35 (2022), 25278–25294.

[56] 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*. https://openreview.net/forum?id=zWqr3MQuNs

[57] 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)*. IEEE, 3–18.

[58] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[59] Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems* 35 (2022), 38274–38290.

[60] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems* 37 (2024), 87310–87356.

[61] Ultralytics. 2024. YOLOv11: Real-Time Object Detection. https://github.com/ultralytics/ultralytics.

[62] 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)*. IEEE, 268–282.

[63] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling Context in Referring Expressions. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 69–85.

[64] Boya Zeng, Yida Yin, and Zhuang Liu. 2025. Understanding bias in large-scale visual datasets. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 1976, 33 pages.

[65] Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. 2025. Min-K%++: Improved Baseline for Pre-Training Data Detection from Large Language Models. In *International Conference on Learning Representations (ICLR)*.

[66] Hongyu Zhu, Sichu Liang, Wentao Hu, et al. 2025. Stealing Knowledge from Auditable Datasets. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*.

[67] Hongyu Zhu, Sichu Liang, Wentao Hu, Li Fangqi, Ju Jia, and Shi-Lin Wang. 2024. Reliable Model Watermarking: Defending against Theft without Compromising on Evasion. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 10124–10133. doi:10.1145/3664647.3681610

[68] Hongyu Zhu, Sichu Liang, Wenwen Wang, et al. 2025. Evading Data Provenance in Deep Neural Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[69] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning* (Washington, DC, USA) *(ICML'03)*. AAAI Press, 912–919.