COMPASS: MEMBERSHIP INFERENCE ATTACK AGAINST VISION LARGE LANGUAGE MODELS VIA ADAPTIVE NOISE CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision large language models (VLLMs) are increasingly deployed in privacy-sensitive domains, raising concerns about the exposure of their training data. The strongest existing gray-box membership inference attacks (MIAs) against VLLMs commonly rely on entropy statistics or generated descriptions, requiring costly response generation while providing limited membership separation. In contrast, perturbation-based methods measure how model response changes when images are perturbed, but have mainly focused on large perturbations, leaving VLLM membership behavior under smaller perturbations poorly understood. In this work, we systematically characterize VLLM membership behavior under controlled Gaussian perturbations across the full range of noise levels at fine resolution. Our analysis reveals that both the optimal perturbation magnitude and the direction of the membership signal vary across model–dataset pairs. For many settings, a previously unstudied low-noise regime produces stronger separation than large perturbations, with members exhibiting lower divergence than non-members and thereby reversing the conventionally observed score direction. Other settings retain the conventional direction, demonstrating that neither a fixed perturbation magnitude nor a predetermined score orientation is universally reliable. Motivated by this finding, we propose **COMPASS**, an adaptive gray-box MIA that calibrates both the perturbation magnitude and membership-score direction using a loosely distribution-matched, known non-member reference set. We instantiate COMPASS using Rényi-normalized KL divergence and Rényi divergence, with max-k aggregation over the most informative image-logit positions. Across LLaVA v1.5–7B, MiniGPT-4, and HuluMed-32B, spanning pretraining, instruction-tuning, and medical-image datasets, COMPASS consistently outperforms baselines. Under general-domain evaluations, COMPASS achieves AUCs of up to 0.88. On the distribution-matched medical evaluation, it reaches 0.64 AUC, while the strongest baseline remains at chance-level. These results establish full-range adaptive perturbation analysis as an effective approach to VLLM membership inference.

1 INTRODUCTION

Vision large language models (VLLMs) have demonstrated exceptional ability on multi-modal tasks such as image question answering Hartsock & Rasool (2024); Khan & Fu (2024), medical diagnosis Bazi et al. (2023); Guo & Terzopoulos (2024), and reasoning problems Nagar et al. (2024); Chen et al. (2024a); Dong et al. (2025). This rapid advancement raises concerns about data privacy, especially when these models are trained on private or copyrighted images Samson et al. (2024). VLLMs can memorize these data within their parameters and expose them through inference-time behavior Jayaraman et al. (2024). The image modality is particularly concerning, as training images can be easily reconstructed from VLLMs Nguyen et al. (2025), and the image itself is sufficient to extract private attributes from the model Tömekçe et al. (2024).

The membership inference attack (MIA) Carlini et al. (2022); Fu et al. (2024); Zhang et al. (2024) serves as a practical way to verify whether private or copyrighted data was used during training Duarte et al. (2024), and is widely used for privacy auditing Liu et al. (2024) and verifying machine unlearning Shi et al. (2024). Given a dataset containing both members (training sam-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ples) and non-members, MIA aims to distinguish members from non-members. Existing attacks on VLLMs have important limitations: Hu *et al.* proposed an attack using a statistical test on generated responses Hu et al. (2025), but it requires both modalities and determines only group-level membership for the vision modality. Li *et al.* use the entropy of responses as a membership signal Li et al. (2024), at the cost of efficiency, since it requires full response generation.

Perturbation-based attacks alternatively apply a change to the input, measuring how much the model output shifts. The magnitude of this shift indicates how stably the model behaves around that sample. This sensitivity constitutes a natural membership signal, as training and unseen samples respond differently to the same change. Prior works have demonstrated the effectiveness of perturbation on vision classifiers Jayaraman et al. (2021); Choquette-Choo et al. (2021); Li & Zhang (2021) and language models Mattern et al. (2023), but has primarily considered strong image transformations, leaving behavior under small, controlled perturbations unexplored. Although Li et al. (2024) evaluate the perturbation-based EncoderMI baseline Liu et al. (2021) on VLLMs, its augmentations serve as fixed preprocessing operations rather than an object of analysis: they are nonlinear and semantically nontrivial, their strength is not varied, and the work provides no analysis of how the model's behavior differs between members and non-members. He et al. (2025) underscore the cost of this omission, showing that overly coarse perturbation causes robustness-based attacks on LLMs to fail.

We therefore systematically characterize VLLM membership behavior under controlled Gaussian perturbations across the full range of noise levels at fine resolution. This analysis reveals that both the perturbation magnitude yielding the strongest membership separation and the direction of the membership signal vary across model–dataset pairs. In many settings, a previously unexplored low-noise regime produces substantially stronger separation than large perturbations, with members exhibiting lower divergence than non-members; in others, the conventional direction persists. These findings show that neither perturbation magnitude nor score orientation can be fixed a priori.

Motivated by this finding, we propose **COMPASS** (**C**alibrated **O**rientation and **M**agnitude **P**erturbation **A**ttack via **S**igma **S**weeps), an adaptive gray-box MIA that searches the full Gaussian-noise range. COMPASS compares the aggregate divergence response of the target set against a loosely distribution-matched non-member reference set, selects the noise level producing the strongest observable separation, and determines whether lower or higher divergence indicates membership. This procedure requires no labeled members and only accesses image logits from the original and perturbed inputs, avoiding response generation. We instantiate COMPASS as COMPASS-KL, using Rényi normalization followed by KL divergence, and COMPASS-R, using Rényi divergence directly. Both variants apply max-k aggregation over the most responsive image-logit positions.

We evaluate COMPASS on LLaVA v1.5–7B, MiniGPT-4, and HuluMed-32B across pretraining, instruction-tuning, and medical-image datasets. COMPASS consistently outperforms existing entropy- and augmentation-based MIA baselines, while variation across target settings further validates the need for adaptive calibration.

Contributions. Our contributions are summarized as follows:

1. We conduct the first systematic full-range analysis of Gaussian perturbation behavior for VLLM membership inference, identifying a previously overlooked low-noise regime with an inverted membership signal that provides the strongest separation in most evaluated settings. More broadly, we find that both the optimal perturbation magnitude and membership-signal direction vary across model–dataset pairs.

2. We propose COMPASS-KL and COMPASS-R, two adaptive gray-box MIA variants that calibrate perturbation magnitude and score direction using the target set and a loosely distribution-matched non-member reference set, without requiring labeled members. We instantiate COMPASS with two divergence measures, yielding COMPASS-KL and COMPASS-R.

3. Through extensive experiments across multiple VLLMs and training-data settings, we demonstrate that COMPASS consistently outperforms existing MIA baselines.

2 Related Work

Vision Large Language Models (VLLMs) Multi-modal models offer strong cross-modal understanding and are widely used in various domains Kumar et al. (2025); Maiti et al. (2025). Unlike earlier vision language models that achieve cross-modal understanding via contrastive alignment Qi et al. (2020); Radford et al. (2021); Jia et al. (2021), VLLMs project image representations into an LLM. The projector can be a transformer Li et al. (2022), a linear layer Liu et al. (2023), a multi-layer perceptron Zhu et al. (2023), or cross-attention layers Bai et al. (2023). COMPASS specifically targets VLLMs and assumes an LLM backbone in the target model. VLLMs have two training stages: (1) a pre-training stage that aligns the visual encoder and LLM backbone on paired image–text data; and (2) an instruction-tuning stage that optimizes the model to follow multi-modal instructions. Our attack aims to infer membership for datasets from both stages using the instruction-tuned model.

Membership Inference Attack (MIA) Given a sample and a target model, MIA aims to determine whether the sample was part of the model's training set. Earlier work by Shokri *et al.* trains shadow models as a proxy for the target model and trains a binary classifier over their outputs Shokri et al. (2017). Yeom *et al.* proposed the loss-based attack Yeom et al. (2018), which determines a sample's membership from its loss under the model. Carlini *et al.* proposed LiRA, an MIA based on a likelihood ratio over a fitted distribution Carlini et al. (2022). LiRA scales poorly to large models due to the prohibitive cost of training shadow models. Reference-based calibration attacks Mireshghallah et al. (2022); Carlini et al. (2021); Fu et al. (2024) offer an efficient proxy by capturing the loss of neighboring samples or of a reference model. In parallel, Min-$k\%$ detects training data using the average log-likelihood of the $k\%$ of tokens with the lowest probabilities Shi et al. (2023). Normalizing the per-token prediction probabilities by the average probabilities amplifies the membership signal Zhang et al. (2024).

MIA has been studied for multi-modal models Ko et al. (2023); Hu et al. (2022; 2025); Li et al. (2024). M^{4}I Hu et al. (2022) targets VLMs with contrastive alignment. The attack model is a binary classifier trained on cross-modality features of members and non-members obtained from shadow models. This is not applicable to VLLMs, which have an LLM backbone. Hu *et al.* proposed the first membership inference attack on VLLMs. For each input sample, it generates a sequence from the model and conducts a likelihood ratio test Hu et al. (2025). While effective, the attack requires a complete image–text pair. Although the paper also proposes an image-only attack, it can infer membership only for a group of images, not for individual images. Li *et al.* proposed a metric-based attack that leverages the average Rényi entropy over the output logits, where members show lower entropy than non-members Li et al. (2024). However, the scope of the attack is limited to detecting pre-training data in the instruction-tuned model.

Adversarial perturbations Moosavi-Dezfooli et al. (2017) have also been used for MIAs. Earlier works target classification models by directly launching an adversarial attack on the target samples Del Grosso et al. (2022); Zhang et al. (2022). The amount of noise serves as a membership signal, with members requiring larger perturbations. However, these attacks require white-box access. COMPASS is more analogous to label-only MIAs Choquette-Choo et al. (2021); He et al. (2025), where the adversary operates in a black-box setting and observes how the output changes under input perturbations. Recently, Liu *et al.* proposed a label-only MIA for VLMs Yihao et al. (2025). However, that attack requires a reference VLM and targets both modalities, so we do not compare against it.

3 Problem Setting

Let $\mathcal{M}$ be a target VLLM trained on a dataset X_{tr}. $\mathcal{M}$ takes an image–text pair (x, z) as input, where z is optional. $\mathcal{M}$ produces output logits $y \coloneqq (y_1, \cdots, y_t)$, where t is the sequence length. A target dataset $X \coloneqq X_m \cup X_n$ where $X_m \coloneqq \{x_m | x_m \in X_{\text{tr}}\}$ and $X_n \coloneqq \{x_n | x_n \notin X_{\text{tr}}\}$ denote the member and non-member sets, respectively.

Given a target set with unknown membership, the adversary's goal is to classify whether a given sample $x \in X$ belongs to X_m or X_n. Formally, the objective is described as a binary classification problem with a threshold λ. Let $R(x, \mathcal{M})$ denote a raw attack statistic and let $o \in -1, +1$ denote

its selected score orientation. We define the oriented membership score as

$$S(x, \mathcal{M}) = o \cdot R(x, \mathcal{M}), \tag{1}$$

where o is chosen such that larger values of S indicate stronger evidence of membership. The resulting binary decision rule is

$$\mathcal{A}(\mathbf{x}, \mathcal{M}) = \begin{cases} 1 & (\mathbf{x} \in X_m) \quad \text{if } S(\mathbf{x}, \mathcal{M}) \geq \lambda \\ 0 & (\mathbf{x} \in X_n) \quad \text{otherwise} \end{cases}$$

where a positive prediction flags $\mathbf{x}$ as a member. We assume a black-box setting, in which the attacker has access only to the input and the output logits of the model. We assume that the adversary has access to a public dataset that does not contain any members and follows a distribution similar to that of the target set. This is a realistic assumption as the attacker can use a publicly available dataset that is released after the model's training cutoff Hu et al. (2025).

4 Gaussian Perturbation Analysis

To characterize how VLLM membership behavior varies with perturbation magnitude, we systematically sweep additive Gaussian noise over a finely resolved range of standard deviations. This analysis reveals three distinct response regimes, most notably an unstudied low-noise "regime II" where membership signal reverses relative to the traditionally observed high-noise "regime III".

Quantifying model behavior. For each member and non-member sample, we (1) perturb the sample with additive Gaussian noise, (2) obtain the per-token logits for the original and perturbed sample pair, and (3) compute the per-token divergence between each pair of softmax logit distributions, taking the average of the max-k divergences along the token sequence as the score. This quantifies noise sensitivity of the model regarding member and non-members. We explore two divergence metrics. The per-token Rényi divergence $D_t^{(\alpha)}$ is defined as:

$$D_t^{(\alpha)}(p_t \| \tilde{p}_t) = \frac{1}{\alpha - 1} \log \sum_{v \in V} p_t(v)^{\alpha} \, \tilde{p}_t(v)^{\,1-\alpha}, \alpha > 0, \; \alpha \neq 1. \tag{2}$$

Here, p_t and $\tilde{p}_t$ denote the t-th token distributions obtained by applying softmax to the original and perturbed logits, respectively. Alternatively, we apply Rényi normalization directly to the softmax logits, which sharpens the distribution for $\alpha > 1$, and compute the Kullback–Leibler divergence $D^{(KLD)}$ as:

$$D_t^{(\mathrm{KLD})}\left(q_t^{(\alpha)} \middle\| \tilde{q}_t^{(\alpha)}\right) = \sum_{v \in V} q_t^{(\alpha)}(v) \log \left(\frac{q_t^{(\alpha)}(v)}{\tilde{q}_t^{(\alpha)}(v)} \right), \qquad q_t^{(\alpha)}(v) = \frac{p_t(v)^{\alpha}}{\sum_{u \in V} p_t(u)^{\alpha}}. \tag{3}$$

For $\alpha = \infty$, this normalization is computed as

$$q_t^{(\infty)}(v) = \begin{cases} 1/|A_t|, & v \in A_t, \\ 0, & \text{otherwise}, \end{cases} \qquad A_t = \arg\max_{u \in V} p_t(u), \tag{4}$$

with $\tilde{q}_t^{(\infty)}$ defined analogously from $\tilde{p}_t$. Before computing the KL divergence, we clamp each entry of both normalized distributions from below at $\varepsilon = 10^{-12}$ and renormalize them, preventing undefined values when either distribution contains zeros. For a pair of sequences with length T, the divergence metric provides T token-wise divergences. A sample's divergence score is the average of the max-k token divergences for a given $k \in [0.1, 1.0]$.

Analyzing model sensitivity. For each perturbation magnitude $\tilde{\sigma}$, we compute the mean divergence scores $\bar{D}_m(\tilde{\sigma})$ and $\bar{D}_n(\tilde{\sigma})$ over known member and non-member sets and define their gap as $\Delta\bar{D}(\tilde{\sigma}) = \bar{D}_m(\tilde{\sigma}) - \bar{D}_n(\tilde{\sigma})$ (Figure 1). As $\tilde{\sigma}$ increases, we observe three distinct regimes: (i) a near-zero plateau ($\Delta\bar{D} \approx 0$); (ii) a pronounced negative dip under small perturbations ($\Delta\bar{D} < 0$), indicating greater divergence for non-members; and (iii) a sharp rise to the conventionally reported behavior under strong perturbations, where members exhibit greater divergence ($\Delta\bar{D} > 0$).

Regime I: $\Delta\bar{D} \approx 0$. At zero noise, both groups exhibit the expected zero divergence, with $\bar{D}_m = \bar{D}_n = 0$. Without normalization, microscopic perturbations induce small, nearly identical

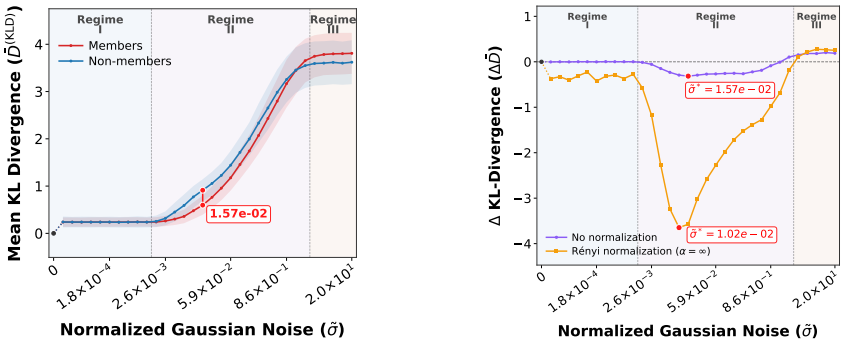


Figure 1: Perturbation sensitivity and membership divergence across normalized Gaussian noise on a Flickr pre-training target set against LLaVA v1.5–7B model with k=0.9. **Left:** Mean unnormalized KL divergence for members and non-members, with shaded bands denoting one standard deviation. Their similar response in Regime I produces a near-zero divergence gap, whereas members exhibit greater divergence in Regime III. **Right:** The member–non-member gap $\Delta\bar{D} = \bar{D}_m - \bar{D}_n$ with and without Rényi normalization ($\alpha = \infty$). Rényi normalization amplifies the negative gap in Regime II. Red markers indicate the noise level minimizing $\Delta\bar{D}$ for each metric.

increases in $\bar{D}_m$ and $\bar{D}_n$ (Figure 1), whereas aggressive normalization reveals only weak separation between the two groups. Given members and non-members respond similarly, Regime I provides little membership signal and is unlikely to reflect memorization of specific training examples.

Regime II: $\Delta\bar{D} < 0$. Regime II captures our central empirical finding: for some model–dataset pairs, small nonzero perturbations produce greater divergence for non-members than for members, similar to findings of label-only and adversarial-perturbation MIAs against non-generative image classifiers Choquette-Choo et al. (2021); Zhang et al. (2022). This finding contradicts the behavior conventionally observed from VLLMs where members are less robust (member divergences is higher). Previously, this was not found since existing works have not explored full noise sweep and neglected this low-noise regime. In fact, we find many cases where this regime offers strong member and non-member separability (bigger gap) and hence we leverage it as a membership signal in the subsequent section.

Regime III: $\Delta\bar{D} > 0$. As the perturbation enters the higher-noise regime, $\bar{D}_m$ crosses over $\bar{D}_n$, returning to the conventionally understood LLM and VLLM behavior Liu et al. (2021). Simultaneously, the across-sample standard deviation of the divergence scores increases rapidly for both classes, amplifying variability alongside the mean divergence gap. Beyond the crossover point, perturbed member images become sufficiently dissimilar to the training samples, allowing the model to detect larger distributional shifts. By contrast, we attribute the smaller non-member shifts to a lack of strong memorized reference in the model's weights, ultimately producing a lower asymptotic D_n.

Leveraging Regime II. Regime II provides a more discriminative membership signal than its high-noise counterpart. Moreover, we can leverage aggressive normalization to amplify the membership signal, as demonstrated by the substantial negative gap in Figure 1. We attribute this result to two complementary factors: signal strength and signal clarity. First, in regime II, member outputs remain locally anchored while non-member outputs shift away, often producing larger separation than regime III. Second, regime III has bigger across-sample divergence. Thus, although strong perturbations can enlarge the mean response, they can simultaneously inject variability that obscures the membership signal. Although our results in section 6 show that regime II often provides a stronger membership signal, it is not universally optimal: in some settings, regime III yields better separation. Our proposed attack therefore can leverage both regimes effectively via calibration.

5 Membership Inference Attack

We propose **COMPASS**, an adaptive membership inference attack that calibrates both the perturbation magnitude and membership-score direction. Its controlled perturbation-based framework sweeps a finely resolved range of Gaussian noise magnitudes to identify the strongest membership separation. This allows dynamic utilization of regime II or III to find the most discriminative per-

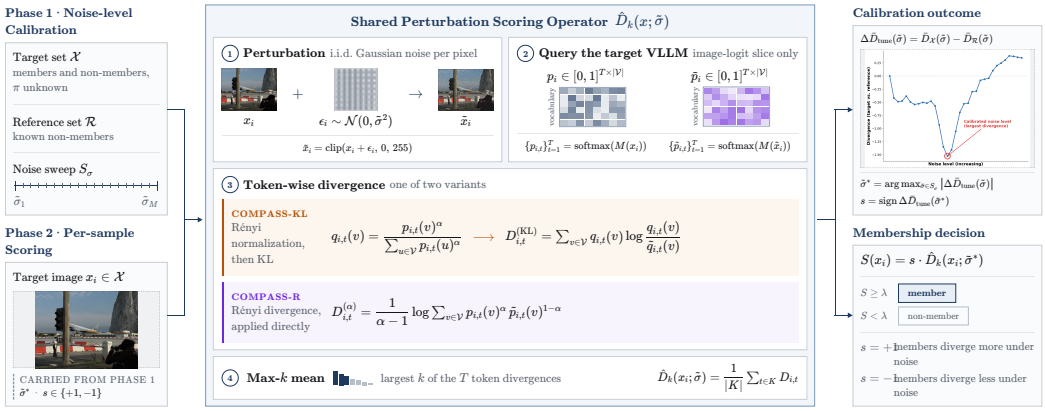


Figure 2: COMPASS attack structure. In the tuning phase, COMPASS compares the divergence responses of the unlabeled target set and a known non-member reference set across candidate noise levels to select the perturbation magnitude and score direction. In the scoring phase, it applies the selected perturbation to each target image, aggregates the largest token-wise divergences, and orients the result to produce a membership score.

turbation. To further refine the signal, COMPASS expands on the general ΔD score outlined in our perturbation analysis, averaging over the max-k divergence measures to focus the score on the most discriminative features. Figure 2 summarizes the COMPASS pipeline; we next formalize its perturbation, divergence scoring. We further formalize the calibration procedures in Appendix D.

Attack structure. For each target image $\mathbf{x}_i \in [0, 255]^{H \times W \times C}$, we independently sample additive Gaussian noise at each pixel and channel and construct the perturbed image as

$$(\boldsymbol{\epsilon}_i)_{h,w,c} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(0, (\tilde{\sigma}^*)^2\big)\,, \qquad \tilde{\mathbf{x}}_i = \text{clip}(\mathbf{x}_i + \boldsymbol{\epsilon}_i,\, 0,\, 255)\,. \tag{5}$$

where $\tilde{\sigma}^*$ denotes the perturbation magnitude selected by the noise-level search as described in Appendix D. We then feed the original and perturbed images into the target model to obtain their corresponding image-token probability sequences:

$$\{p_{i,t}\}_{t=1}^T = \text{softmax}(\mathcal{M}(\mathbf{x}_i))\,, \qquad \{\tilde{p}_{i,t}\}_{t=1}^T = \text{softmax}(\mathcal{M}(\tilde{\mathbf{x}}_i))\,. \tag{6}$$

Following the perturbation analysis, we compute the divergence between each pair of token-wise distributions using either (i) the Rényi divergence $D_t^{(\alpha)}$ in Equation 2, yielding COMPASS-R, or (ii) the KL divergence $D_t^{(\text{KLD})}$ after Rényi normalization in Equation 3, yielding COMPASS-KL. For COMPASS-KL, we evaluate normalization orders $\alpha \in \{0.5, 1.0, 2.0, \infty\}$, where $\alpha = \infty$ denotes the limiting max-dominant normalization. For COMPASS-R, we evaluate Rényi orders $\alpha \in \{0.25, 0.5, 2.0, 4.0\}$. To obtain a single sample-level membership score while emphasizing the strongest token-wise contributions, we average the top-k divergence values across the sequence. Finally, we orient the score according to the set-specific tuning outcome: if the target set exhibits lower mean divergence than the known non-member set, we negate the aggregated score so that larger values consistently indicate membership.

6 Experiments

6.1 Experimental Settings

Target Models. We evaluate our attack on LLaVA v1.5–7B Liu et al. (2023) and MiniGPT-4 Zhu et al. (2023), two open-source VLLMs for which substantial information about the underlying training data is publicly available. To broaden our evaluation across model scale and domain, we additionally evaluate the larger medical-domain HuluMed-32B model Jiang et al. (2025).

Datasets. To evaluate membership inference over **pre-training** data, we use the Flickr benchmark introduced by Li *et al.* Li et al. (2024), which contains 300 members and 300 non-members for both LLaVA and MiniGPT-4. We randomly sample 150 images from each group to construct each

target set. For **instruction fine-tuning**, we use COCO-2017 images included in LLaVA's disclosed instruction-tuning data Lin et al. (2014). To reduce the likelihood that the presumed non-members appeared during pre-training, we draw them from recent Flickr images. Because MiniGPT-4 does not disclose its instruction-tuning data, we restrict this evaluation to LLaVA. For both settings, the tuning reference set is sampled from ShareGPT-4o Chen et al. (2025). For HuluMed-32B, we evaluate COMPASS on endoscopy medical images, pairing PubMedVision members Chen et al. (2024b) with EndoBench non-members Liu et al. (2025) using GastroHUN Bravo et al. (2025) as the reference set. Motivated by blind-baseline analyses showing that source differences can confound MIA evaluation Das et al. (2025); Zhu et al. (2025), we apply the filtering procedure described in Appendix E to both self-constructed medical datasets to remove potential bias.

Baselines. We compare our attack with state-of-the-art MIAs on VLLMs. We use (1) **Aug-KL** Liu et al. (2021), which compares the KLD of logits from the original and augmented images; (2) **Min-k%** Shi et al. (2023), which uses the average of the min-k token probabilities as the membership signal; and (3) **Max-Rényi-α** Li et al. (2024), which uses the average of the max-k% Rényi entropies of the output logits. We report only $\alpha = 0.5$, their best-performing version.

6.2 IMAGE MIA ON LLAVA & MINIGPT-4

	LLaVA v1.5–7B						MiniGPT-4		
	Flickr			COCO-2017			Flickr		
Metric	$k=0.1$	$k=0.5$	$k=0.9$	$k=0.1$	$k=0.5$	$k=0.9$	$k=0.1$	$k=0.5$	$k=0.9$
Baselines									
Aug-KL	–	–	0.71 ± 0.04	–	–	0.69 ± 0.01	–	–	0.63 ± 0.01
Min-k%	0.58 ± 0.02	0.66 ± 0.05	0.67 ± 0.05	0.60 ± 0.01	0.64 ± 0.03	0.65 ± 0.02	0.51 ± 0.04	0.59 ± 0.02	0.64 ± 0.02
Max-Rényi$_{0.5\%}$	0.64 ± 0.02	0.72 ± 0.03	0.70 ± 0.03	0.69 ± 0.04	0.74 ± 0.03	0.72 ± 0.03	$\mathbf{0.78 \pm 0.02}$	0.73 ± 0.01	0.69 ± 0.01
COMPASS-KL (Ours)									
$\alpha = 0.5$	$0.60^* \pm 0.02$	$0.70^* \pm 0.03$	$0.70^* \pm 0.04$	$0.70^* \pm 0.03$	$0.80^* \pm 0.03$	$0.81^* \pm 0.04$	$0.58^* \pm 0.04$	$0.69^* \pm 0.03$	$0.72^* \pm 0.02$
$\alpha = 1$	$0.62^* \pm 0.02$	$0.70^* \pm 0.03$	$0.70^* \pm 0.03$	$0.70^* \pm 0.03$	$0.80^* \pm 0.04$	$0.81^* \pm 0.04$	$0.66^* \pm 0.02$	$0.73^* \pm 0.01$	$0.75^* \pm 0.01$
$\alpha = 2$	$\mathbf{0.65^* \pm 0.03}$	$0.71^* \pm 0.03$	$0.72^* \pm 0.03$	$\mathbf{0.74^* \pm 0.02}$	$0.81^* \pm 0.05$	$0.81^* \pm 0.04$	$0.69^* \pm 0.03$	$0.72^* \pm 0.02$	$0.75^* \pm 0.01$
$\alpha = \infty$	$0.50^* \pm 0.00$	$\mathbf{0.78^* \pm 0.02}$	$\mathbf{0.78^* \pm 0.02}$	$0.50^* \pm 0.00$	$\mathbf{0.88^* \pm 0.04}$	$\mathbf{0.88^* \pm 0.04}$	$0.53^* \pm 0.01$	$0.75^* \pm 0.02$	$0.77^* \pm 0.02$
COMPASS-R (Ours)									
$\alpha = 0.25$	$0.60^* \pm 0.03$	$0.69^* \pm 0.02$	$0.70^* \pm 0.02$	$0.71^* \pm 0.03$	$0.80^* \pm 0.04$	$0.81^* \pm 0.04$	$0.63^* \pm 0.03$	$0.72^* \pm 0.02$	$0.74^* \pm 0.01$
$\alpha = 0.5$	$0.59^* \pm 0.03$	$0.69^* \pm 0.02$	$0.69^* \pm 0.02$	$0.71^* \pm 0.03$	$0.80^* \pm 0.04$	$0.81^* \pm 0.04$	$0.64^* \pm 0.02$	$0.71^* \pm 0.01$	$0.74^* \pm 0.01$
$\alpha = 2$	$0.64^* \pm 0.02$	$0.72^* \pm 0.03$	$0.73^* \pm 0.03$	$0.73^* \pm 0.03$	$0.82^* \pm 0.04$	$0.83^* \pm 0.04$	$0.68^* \pm 0.01$	$0.76^* \pm 0.01$	$0.79^* \pm 0.01$
$\alpha = 4$	$0.64^* \pm 0.02$	$0.73^* \pm 0.03$	$0.74^* \pm 0.03$	$0.72^* \pm 0.02$	$0.84^* \pm 0.03$	$0.85^* \pm 0.04$	$0.68^* \pm 0.01$	$\mathbf{0.77^* \pm 0.01}$	$\mathbf{0.81^* \pm 0.01}$

Table 1: Average AUC over three independently sampled target sets for LLaVA v1.5–7B on Flickr and COCO-2017, and MiniGPT-4 on Flickr. An asterisk indicates that the COMPASS decided to use region II for membership signal.

Attack Performance. Table 1 compares the AUC of COMPASS-KL and COMPASS-R against the considered baselines for LLaVA v1.5–7B and MiniGPT-4. Since Min-k%, Max-Rényi, and our proposed attacks operate on a top-k subset of output tokens, we report representative settings $k \in \{0.1, 0.5, 0.9\}$ and highlight the best-performing configuration for each method. Aug-KL instead computes its membership score over the full output sequence and is therefore reported once under $k = 0.9$ for presentation. All results are averaged over three independently sampled target sets, with standard deviation reported across runs.

On the Flickr pre-training data, COMPASS provides strong membership separation across both evaluated VLLMs. For LLaVA, COMPASS-KL achieves a maximum AUC of 0.78 ± 0.02, improving over the strongest baseline at 0.72 ± 0.03, while COMPASS-R reaches 0.74 ± 0.03. We observe comparable performance on MiniGPT-4. Although Max-Rényi-0.5% provides a strong baseline at 0.78 ± 0.02, COMPASS-R achieves the highest overall AUC of 0.81 ± 0.01, while COMPASS-KL reaches 0.77 ± 0.02. In both cases where COMPASS obtains the strongest result, its standard deviation is also no greater than that of the strongest competing baseline, indicating that the improvement persists across independently sampled target sets rather than arising from a single favorable sample.

The improvement is substantially larger on LLaVA's COCO-2017 instruction-tuning data. COMPASS-KL achieves a maximum AUC of 0.88 ± 0.04 and COMPASS-R reaches 0.85 ± 0.04, compared with 0.74 ± 0.03 for the strongest baseline. Although the variation is slightly larger in this setting, the gain in mean AUC is substantially greater. In particular, COMPASS-KL with Rényi-∞ normalization achieves 0.88 ± 0.04 at both $k = 0.5$ and $k = 0.9$. Overall, COMPASS achieves the

highest AUC across all three evaluated model–dataset settings while exhibiting comparable, and in several cases lower, variation across target-set samples.

Effect of Rényi Order. Performance generally improves as the Rényi order α increases for both COMPASS variants. Larger α places greater emphasis on dominant output probabilities, with Rényi-∞ normalization producing the strongest COMPASS-KL results and the largest evaluated order similarly performing best for COMPASS-R. This suggests that COMPASS's membership signal is concentrated primarily among the model's dominant output probabilities.

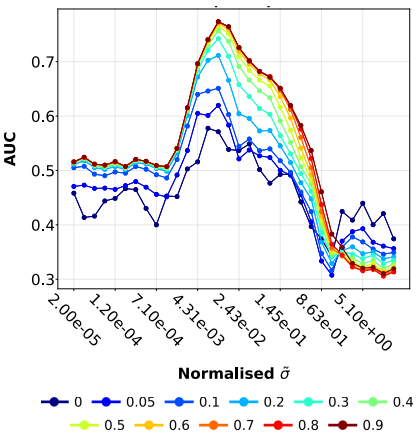


Figure 3: Max-k effect for COMPASS-R on Flickr.

Effect of Max-k. Figure 3 shows that, across noise levels, retaining a large top-k fraction of token-wise divergence values ($k \in \{0.7, 0.8, 0.9\}$) consistently yields higher AUCs and a stronger membership signal in the pre-training setting. We observe the same pattern in the COCO-2017 instruction-fine-tuning setting shown by Table 1. We hypothesize that this behavior arises from the distributed nature of visual representations: in the evaluated models, membership-relevant evidence may be spread across many correlated visual tokens rather than concentrated in only a few. Aggressive truncation at small k can therefore discard weak but correlated cues, increasing variance and attenuating member–non-member separation. Retaining a larger k preserves this distributed evidence, stabilizes the estimate, and strengthens the resulting separation.

6.3 Medical Image MIA on HuluMed-32B

	Endoscopy (AUC)			Endoscopy (TPR@5%FPR)		
Metric	$k = 0.7$	$k = 0.8$	$k = 0.9$	$k = 0.7$	$k = 0.8$	$k = 0.9$
Baselines						
Aug-KL	–	–	0.42 ± 0.04	–	–	0.01 ± 0.00
Min-k%	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Max-Rényi$_{0.5\%}$	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
COMPASS-KL (Ours)						
$\alpha = 0.5$	0.58 ± 0.03	0.57 ± 0.03	0.57 ± 0.03	0.12 ± 0.02	0.12 ± 0.02	0.13 ± 0.02
$\alpha = 1.0$	$0.61^* \pm 0.00$	$0.62^* \pm 0.00$	$0.62^* \pm 0.00$	$0.12^* \pm 0.01$	$0.12^* \pm 0.01$	$0.13^* \pm 0.01$
$\alpha = 2.0$	$0.60^* \pm 0.01$	$0.61^* \pm 0.01$	$0.61^* \pm 0.01$	$0.12^* \pm 0.01$	$0.11^* \pm 0.00$	$0.11^* \pm 0.01$
$\alpha = \infty$	$0.58^* \pm 0.03$	$0.59^* \pm 0.02$	$0.60^* \pm 0.02$	$0.13^* \pm 0.02$	$0.10^* \pm 0.04$	$0.10^* \pm 0.04$
COMPASS-R (Ours)						
$\alpha = 0.25$	$\mathbf{0.63^* \pm 0.01}$	$\mathbf{0.63^* \pm 0.01}$	$\mathbf{0.64^* \pm 0.01}$	$0.11^* \pm 0.04$	$0.12^* \pm 0.04$	$0.13^* \pm 0.05$
$\alpha = 0.5$	$\mathbf{0.63^* \pm 0.01}$	$\mathbf{0.63^* \pm 0.01}$	$0.63^* \pm 0.01$	$0.12^* \pm 0.03$	$0.13^* \pm 0.04$	$\mathbf{0.14^* \pm 0.05}$
$\alpha = 2.0$	$0.62^* \pm 0.01$	$0.62^* \pm 0.01$	$0.62^* \pm 0.01$	$\mathbf{0.14^* \pm 0.03}$	$0.14^* \pm 0.03$	$\mathbf{0.14^* \pm 0.03}$
$\alpha = 4.0$	$0.59^* \pm 0.03$	$0.62^* \pm 0.01$	$0.62^* \pm 0.02$	$0.12^* \pm 0.03$	$\mathbf{0.15^* \pm 0.02}$	$0.13^* \pm 0.03$

Table 2: Average AUC and TPR@5%FPR for the Endoscopy dataset against HuluMed-32B. Following the finding that large Max-k yields stronger performance, we report representative top-k settings $k \in \{0.7, 0.8, 0.9\}$ evaluated under the same pipeline as LLaVA and Minigpt (subsection 6.2). Bold indicates the highest reported value in each column, including ties.An asterisk indicates that the COMPASS decided to use region II for membership signal.

Table 2 reports AUC for HuluMed-32B on the filtered Endoscopy dataset. Given membership inference is more challenging in this setting, we also report TPR@5%FPR to assess detection performance at a low false-positive rate. Min-k% and Max-Rényi perform at chance-level AUC (0.50) and achieve zero TPR@5%FPR across the reported k values. Aug-KL reaches 0.42 ± 0.04 AUC and 0.01 ± 0.00 TPR@5%FPR. These results show that the considered baselines provide little membership separation in this setting.

In contrast, COMPASS achieves membership inference, indicating that its perturbation scores retain a membership signal in the harder endoscopy setting. COMPASS-R with $\alpha = 0.25$ reaches the highest AUC of 0.64 at $k = 0.9$; with $\alpha = 0.5$, it reaches 0.63 across all three reported k values. COMPASS-KL reaches 0.62 AUC with $\alpha = 1$ at $k = 0.8$ and $k = 0.9$. The advantage over the baselines is more evident at low false-positive rates. At 5% FPR, COMPASS-R with $\alpha = 4$ achieves the highest TPR of 0.15 at $k = 0.8$, compared with 0.01 for Aug-KL and 0.00 for Min-k% and Max-Rényi. The best COMPASS-KL configuration reaches 0.13 TPR at $k = 0.9$.

Asterisks in Table 1 and Table 2 shows the significance of our calibration. With the reference set, it scans regimes and determines the ideal separation for each dataset, leading to higher AUC.

6.4 Noise-level Tuning Effectiveness

Dataset	Optimal $\tilde{\sigma}$ AUC	Estimated $\tilde{\sigma}$ AUC
LLaVA Flickr	0.80 ± 0.018	0.78 ± 0.02
LLaVA COCO-2017	0.92 ± 0.01	0.88 ± 0.04
HuluMed Endoscopy	0.69 ± 0.02	0.63 ± 0.02

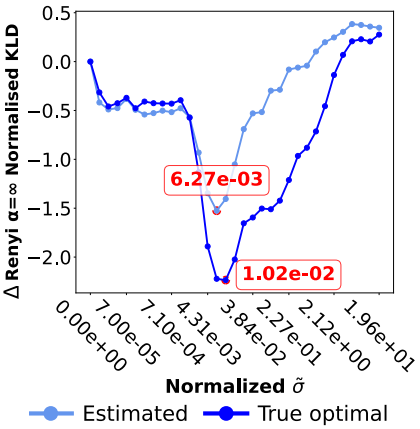


Figure 4: Effectiveness of noise-level tuning. **Top:** AUC using the true optimal noise level and the noise level estimated by our tuning procedure. **Right:** Estimated and true optimal noise levels for Rényi-∞ normalized KL divergence difference. The estimated $\tilde{\sigma}$ lies near the true optimum and yields performance close to the ideal setting.

We find that the tuning procedure for the noise magnitude $\tilde{\sigma}^*$ described in section 5 is a strong estimator of the true optimum. To demonstrate this, we run the attack over the entire range of noise levels and identify the oracle perturbation noise level corresponding to the strongest AUC. Figure 4 (Top) compares this ideal AUC with the empirical AUC we obtain, showing that the two are comparable.

Figure 4 visually demonstrates the effectiveness of the proposed noise tuning method while highlighting the importance of its adaptive nature. While the scale of the ΔD drop is reduced by the member–non-member ratio π, the estimate still yields a $\tilde{\sigma}$ that lies within the trough of the minimum. Although the $\delta(\tilde{\sigma})$ term introduces a slight deviation from the true optimum $\tilde{\sigma}^*$, the results in Figure 4 demonstrate that the estimated $\tilde{\sigma}$ remains a strong differentiating signal for our attack.

7 Conclusion

In this paper, we systematically characterized the membership behavior of diverse VLLMs under image perturbations across a finely resolved range of noise levels, revealing behavior substantially more complex than conventionally assumed. In particular, we found that, for many model–dataset pairs, members exhibit lower divergence than non-members under extremely small perturbations, reversing the conventional membership-signal direction and producing stronger membership separation than the conventionally considered high-noise regime. Building on this observation, we introduced COMPASS-KL and COMPASS-R, two self-calibrating MIA variants that search the full perturbation range to identify both the most discriminative noise level and the corresponding membership-score direction using the target set and a known non-member reference set, without requiring labeled members. Across our experiments, COMPASS consistently outperformed existing baselines, demonstrating the value of adapting perturbation-based MIAs to the noise-response characteristics of each target setting. Future work will investigate stronger theoretical guarantees for noise-level tuning and develop more robust estimators under reference-set distribution shift and varying target-set member proportions.

AI USE STATEMENT

We used generative AI for searching related works, ideation and proving mathematical claims. Authors have rigorously analyzed the related manuscripts retrieved by AI. The research idea, objective and proposal are original to authors and was not assisted by AI. AI was used to validate details of the proposal and responses of AI was reveiwed and evaluated by authors. Mathematical analysis is original to authors, and AI was used to proofread our analysis. Responses from the mode has been rigorously analyzed.

ETHICS STATEMENT

This work studies membership inference attacks to assess privacy risks in publicly available vision–language models, including a medical model. Such attacks could also be misused to infer whether an individual's data was included in training. We conduct our evaluations using publicly available models and datasets, and present COMPASS as a tool for privacy auditing and for informing the development of stronger protections. We have read and comply with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We use publicly available models and datasets, with links provided in Appendix A and Appendix B. We additionally provide our source code through the anonymized repository linked in Appendix A. The paper describes the COMPASS scoring and calibration methods, while Appendix C gives their full algorithms. We also document the endoscopy target-set filtering procedure in Appendix E and report the hardware and runtime requirements in Appendix B.

REFERENCES

J Bai, S Bai, S Yang, S Wang, S Tan, P Wang, J Lin, C Zhou, and J Qwen-VL Zhou. A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 6, 2023.

Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. Vision–language model for visual question answering in medical imagery. *Bioengineering*, 10(3):380, 2023.

Diego Bravo, Juan Frias, Felipe Vera, Juan Trejos, Carlos Martínez, Martín Gómez, Fabio González, and Eduardo Romero. GastroHUN: An endoscopy dataset of complete systematic screening protocol for the stomach. *Scientific Data*, 12:102, 2025. doi: 10.1038/s41597-025-04401-5. URL https://doi.org/10.1038/s41597-025-04401-5.

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)*, pp. 2633–2650, 2021.

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)*, pp. 1897–1914. IEEE, 2022.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. Towards injecting medical visual knowledge into multimodal LLMs at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7346–7370. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.emnlp-main.418. URL https://aclanthology.org/2024.emnlp-main.418/.

Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 2025.

Christopher A. Choquette-Choo, Florian Tramèr, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1964–1974. PMLR, 2021. URL https://proceedings.mlr.press/v139/choquette-choo21a.html.

Debeshee Das, Jie Zhang, and Florian Tramèr. Blind baselines beat membership inference attacks for foundation models. In *2nd Workshop on Data-Centric Machine Learning Research (DATA-FM) at ICLR*, 2025. URL https://arxiv.org/abs/2406.16201.

Ganesh Del Grosso, Hamid Jalalzai, Georg Pichler, Catuscia Palamidessi, and Pablo Piantanida. Leveraging adversarial examples to quantify membership information leakage. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10399–10409, 2022.

Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9062–9072, 2025.

André V Duarte, Xuandong Zhao, Arlindo L Oliveira, and Lei Li. De-cop: Detecting copyrighted content in language models training data. *arXiv preprint arXiv:2402.09910*, 2024.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Membership inference attacks against fine-tuned large language models via self-prompt calibration. *Advances in Neural Information Processing Systems*, 37:134981–135010, 2024.

Danfeng Guo and Demetri Terzopoulos. Prompting medical large vision-language models to diagnose pathologies by visual question answering. *arXiv preprint arXiv:2407.21368*, 2024.

Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence*, 7:1430984, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Yu He, Boheng Li, Liu Liu, Zhongjie Ba, Wei Dong, Yiming Li, Zhan Qin, Kui Ren, and Chun Chen. Towards label-only membership inference attack against pre-trained large language models. In *USENIX Security*, 2025.

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

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

Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. Revisiting membership inference under realistic assumptions. *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2021(2):348–368, 2021. doi: 10.2478/popets-2021-0031.

Bargav Jayaraman, Chuan Guo, and Kamalika Chaudhuri. Déjà vu memorization in vision–language models. *Advances in Neural Information Processing Systems*, 37:50722–50749, 2024.

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

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, Jin Hao, Zijian Chen, Ruijia Wu, Tao Tang, Junhui Lv, Hongxia Xu, Hongwei Wang, Jun Xiao, Bin Feng, Fudong Zhu, Kenli Li, Weidi Xie, Jimeng Sun, Jian Wu, and Zuozhu Liu. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025. doi: 10.48550/arXiv.2510.08668.

Zaid Khan and Yun Fu. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10854–10863, 2024.

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

Gurucharan Marthi Krishna Kumar, Aman Chadha, Janine D Mendola, and Amir Shmuel. Medvisionllama: Leveraging pre-trained large language model layers to enhance medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1114–1124, 2025.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

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.

Zheng Li and Yang Zhang. Membership leakage in label-only exposures. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 880–895. ACM, 2021. doi: 10.1145/3460120.3484575.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Han Liu, Yuhao Wu, Zhiyuan Yu, and Ning Zhang. Please tell me more: Privacy impact of explainability through the lens of membership inference attack. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 4791–4809. IEEE, 2024.

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

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

Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. EndoBench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0076. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/02ff2906a49d985808e7ba8798b9f9cd-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Shalini Maiti, Lourdes Agapito, and Filippos Kokkinos. Gen3deval: Using vllms for automatic evaluation of generated 3d objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18552–18562, 2025.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. *arXiv preprint arXiv:2305.18462*, 2023.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. Quantifying privacy risks of masked language models using membership inference attacks. *arXiv preprint arXiv:2203.03929*, 2022.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

Aishik Nagar, Shantanu Jaiswal, and Cheston Tan. Zero-shot visual reasoning by vision-language models: Benchmarking and analysis. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2024.

Ngoc-Bao Nguyen, Sy-Tuyen Ho, Koh Jun Hao, and Ngai-Man Cheung. Model inversion attacks on vision-language models: Do they leak what they learn? *arXiv preprint arXiv:2508.04097*, 2025.

Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Laurens Samson, Nimrod Barazani, Sennay Ghebreab, and Yuki M Asano. Little data, big impact: Privacy-aware visual language models via minimal tuning. *arXiv preprint arXiv:2405.17423*, 2024.

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.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*, 2024.

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

Batuhan Tömekçe, Mark Vero, Robin Staab, and Martin Vechev. Private attribute inference from images with vision-language models. *Advances in Neural Information Processing Systems*, 37: 103619–103651, 2024.

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)*, pp. 268–282. IEEE, 2018.

LIU Yihao, LYU Xinqi, et al. Lomia: Label-only membership inference attacks against pre-trained large vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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.

Zhaoxi Zhang, Leo Yu Zhang, Xufei Zheng, Bilal Hussain Abbasi, and Shengshan Hu. Evaluating membership inference through adversarial robustness. *The Computer Journal*, 65(11):2969–2978, 2022.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Hongyu Zhu, Sichu Liang, Wenwen Wang, Boheng Li, Tongxin Yuan, Fangqi Li, Shilin Wang, and Zhuosheng Zhang. Revisiting data auditing in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11337–11346. Association for Computing Machinery, 2025. doi: 10.1145/3746027.3755002. URL https://doi.org/10.1145/3746027.3755002.

APPENDIX

CONTENTS OF THE APPENDIX

We provide supporting materials and analyses in the appendix. In Appendix A, we link to the anonymized source code and datasets used in our evaluations. Appendix B identifies the target models and describes the hardware and runtime requirements. In Appendix C, we present the COMPASS membership-scoring and noise-level tuning algorithms, followed by a derivation of the tuning gap and its potential error from reference-set differences in Appendix D. Appendix E describes how we filter the endoscopy candidate pools to reduce image-source confounding. Finally, Appendix F shows example images across the tested noise levels and the raw divergence curves for each target model, further illustrating the low-noise membership behavior discussed in section 4.

A ASSETS

A.1 SOURCE CODE

Our source code is available in the anonymized GitHub repository: `https://anonymous.4open.science/r/COMPASS_anonymous_github-07A2/readme.md`. The source code was adapted from Li et al. Li et al. (2024).

A.1.1 DATASETS

We use three evaluation settings to assess COMPASS. To evaluate membership inference risk for the pre-training data of LLaVA and MiniGPT-4o, we construct target sets of members and non-members from the models' provided Flickr subsets https://huggingface.co/datasets/JaineLi/VL-MIA-image Li et al. (2024). To evaluate risk for LLaVA's instruction fine-tuning data, we use the provided COCO-2017 subset https://huggingface.co/datasets/liuhaotian/LLaVA-Instruct-150K/viewer Li et al. (2024). For both settings, we tune the noise level using the disjoint ShareGPT-4o-Image dataset https://huggingface.co/datasets/FreedomIntelligence/ShareGPT-4o-Image Chen et al. (2025).

To evaluate the larger Hulu-Med-32B model, we construct target sets using the filtering procedure described in Appendix E. We use PubMedVision endoscopy images https://huggingface.co/datasets/FreedomIntelligence/PubMedVision Chen et al. (2024b) as members and images from EndoBench https://huggingface.co/datasets/Saint-lsy/EndoBench Liu et al. (2025) as non-members. We tune the perturbation noise level using the GastroHUN dataset https://huggingface.co/datasets/doantrongthai/GastroHun/tree/main Bravo et al. (2025).

B EXPERIMENTAL SETUP

B.1 MODELS

We evaluate our membership inference attacks on three vision–language models, each of which has been pre-trained and subsequently instruction fine-tuned. The first target model is LLaVA-v1.5-7B, released by Liu et al. Liu et al. (2023) and publicly available at https://huggingface.co/liuhaotian/llava-v1.5-7b. The second target model is MiniGPT-4, proposed by Zhu et al. Zhu et al. (2023) and available at https://github.com/Vision-CAIR/MiniGPT-4. The third target model is Hulu-Med-32B released by Jiang et al. (2025) and available at https://github.com/ZJUI-AI4H/Hulu-Med.

B.2 PROCESSOR SPECIFICATIONS

All experiments are run on a single NVIDIA H200 GPU with 140 GB of GPU memory. For a combined target and tuning set of 600 samples, collecting all metric values across 32 noise levels requires approximately 12 hours for LLaVA-v1.5-7B and 4 hours for MiniGPT-4. The larger Hulu-Med-32B model requires approximately 20 hours.

C COMPASS ALGORITHMS

C.1 COMPASS MEMBERSHIP SCORING ALGORITHM

Following the pipeline described in section 5, algorithm 1 details how COMPASS computes a membership score for a target image. It compares the image-token distributions obtained from the original and perturbed images, computes token-wise divergences using either COMPASS-KL or COMPASS-R, and averages the largest top-k fraction. The noise level and score direction are supplied by the tuning procedure, and the algorithm returns the resulting oriented membership score.

Algorithm 1 COMPASS-KL & COMPASS-R: Perturbation-Based Membership Scoring

Require: Target image $X \in [0, 255]^{H \times W \times C}$, black-box model $\mathcal{M}$, selected noise std $\tilde{\sigma}^*$, selected score direction $s \in \{-1, +1\}$, top-k fraction $k \in (0, 1]$, variant $v \in \{\mathrm{KL}, \mathrm{R}\}$, Rényi order α
Ensure: Oriented membership score $S(X)$

Sample $\epsilon_{h,w,c} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, (\tilde{\sigma}^*)^2)$
$\tilde{X} \leftarrow \mathrm{clip}(X + \epsilon,\ 0, 255)$
$\{p_t\}_{t=1}^{T} \leftarrow \mathrm{softmax}\big(\mathcal{M}(X)\big)$ ▷ original image-logit distributions
$\{\tilde{p}_t\}_{t=1}^{T} \leftarrow \mathrm{softmax}\big(\mathcal{M}(\tilde{X})\big)$ ▷ perturbed image-logit distributions
if $v = \mathrm{KL}$ **then** ▷ COMPASS-KL
  for $t = 1$ to T **do**
    $q_t(u) \leftarrow \dfrac{p_t(u)^{\alpha}}{\sum_{w \in \mathcal{V}} p_t(w)^{\alpha}}$
    $\tilde{q}_t(u) \leftarrow \dfrac{\tilde{p}_t(u)^{\alpha}}{\sum_{w \in \mathcal{V}} \tilde{p}_t(w)^{\alpha}}$
    $d_t \leftarrow D_{\mathrm{KL}}(q_t \,\|\, \tilde{q}_t)$
  end for
else ▷ COMPASS-R
  for $t = 1$ to T **do**
    $d_t \leftarrow D_{\alpha}(p_t \,\|\, \tilde{p}_t)$
    $= \dfrac{1}{\alpha - 1} \log \sum_{u \in \mathcal{V}} p_t(u)^{\alpha} \tilde{p}_t(u)^{1-\alpha}$
  end for
end if
$\mathcal{I}_k \leftarrow$ indices of the largest $\lceil kT \rceil$ values in $\{d_t\}_{t=1}^{T}$
$D(X; \tilde{\sigma}^*) \leftarrow \dfrac{1}{|\mathcal{I}_k|} \sum_{t \in \mathcal{I}_k} d_t$ ▷ max-k aggregation
$S(X) \leftarrow s\, D(X; \tilde{\sigma}^*)$ ▷ larger scores favor membership
return $S(X)$

C.2 COMPASS NOISE-LEVEL TUNING

Algorithm 2 selects the perturbation magnitude and score direction used by Algorithm 1. For each candidate noise level, it compares the mean raw divergence of the unlabeled target set with that of a known non-member reference set. It selects the noise level with the largest absolute difference and uses the sign of that difference to orient subsequent membership scores.

Algorithm 2 COMPASS Reference-Guided Noise and Score-Orientation Calibration

Require: Target set $\mathcal{X}$ with unknown membership, known non-member reference set $\mathcal{R}$, candidate noise grid Σ, black-box model $\mathcal{M}$, top-k fraction k, variant $v \in \{\mathrm{KL}, \mathrm{R}\}$, Rényi order α
Ensure: Selected noise level $\tilde{\sigma}^*$, score direction $s \in \{-1, +1\}$

for each $\tilde{\sigma} \in \Sigma$ **do**
  Compute raw $D(X; \tilde{\sigma})$ using the divergence and aggregation steps of Algorithm 1 for all $X \in \mathcal{X}$
  Compute raw $D(R; \tilde{\sigma})$ using the same steps for all $R \in \mathcal{R}$
  $\bar{D}_{\mathcal{X}}(\tilde{\sigma}) \leftarrow \frac{1}{|\mathcal{X}|} \sum_{X \in \mathcal{X}} D(X; \tilde{\sigma})$
  $\bar{D}_{\mathcal{R}}(\tilde{\sigma}) \leftarrow \frac{1}{|\mathcal{R}|} \sum_{R \in \mathcal{R}} D(R; \tilde{\sigma})$
  $\Delta\bar{D}_{\mathrm{tune}}(\tilde{\sigma}) \leftarrow \bar{D}_{\mathcal{X}}(\tilde{\sigma}) - \bar{D}_{\mathcal{R}}(\tilde{\sigma})$
end for
$\tilde{\sigma}^* \leftarrow \arg\max_{\tilde{\sigma} \in \Sigma} \left|\Delta\bar{D}_{\mathrm{tune}}(\tilde{\sigma})\right|$
if $\Delta\bar{D}_{\mathrm{tune}}(\tilde{\sigma}^*) \geq 0$ **then**
  $s \leftarrow +1$ ▷ higher divergence favors membership
else
  $s \leftarrow -1$ ▷ lower divergence favors membership
end if
return $(\tilde{\sigma}^*, s)$

D Noise Level Tuning

D.1 Noise Level Tuning Procedure

While section 4 shows that each target model–dataset pair has an optimal noise level for membership separation, a practical tuning procedure must operate without membership labels and determine the appropriate score direction for the selected regime. We therefore use the unlabeled target set as a proxy member set and compare its aggregate divergence response with that of a known non-member reference set, which may be readily obtained through synthetic generation or temporal cutoffs. We reason that the member component remains present in the aggregate response of the target set. Assuming that target and reference non-members exhibit comparable divergence behavior, subtracting their set-level averages approximately cancels the baseline non-member response. Provided that the residual mismatch remains small across the noise sweep, the resulting signal is attenuated by the presence of target non-members but retains the direction and general shape of the underlying membership-separation curve. COMPASS therefore selects the noise level with the largest absolute discrepancy and uses its sign to orient the membership score. Because tuning is performed directly on the target set that is subsequently evaluated, the selected operating point is tailored to the target samples without requiring labeled members. We formalize this tuning signal in subsection D.2, deriving an expression that captures the effect of distributional mismatch between the target and reference sets. In subsection 6.4, we further evaluate the tuning procedure empirically across varying degrees of mismatch, demonstrating its effectiveness even under substantial distributional shift.

D.2 Noise Level Tuning Error Expression

COMPASS selects its noise level and score direction by comparing the mean raw divergence of the unlabeled target set $\mathcal{X}$ with that of a known non-member reference set $\mathcal{R}$. As $\mathcal{X}$ contains both members and non-members, we intuit that its mean response retains a membership-dependent component. At each candidate noise level, the observable tuning gap is

$$\Delta\bar{D}_{\mathrm{tune}}(\tilde{\sigma}) = \bar{D}_{\mathcal{X}}(\tilde{\sigma}) - \bar{D}_{\mathcal{R}}(\tilde{\sigma}). \tag{7}$$

Hence, while this gap it is not itself a score for any individual image, it reflects the general shape and score direction that separates the target set members and non-members. The derivation below separates the contribution of membership from the discrepancy between target and reference non-members.

Formally, let $\pi \in [0, 1]$ denote the unknown proportion of members in the target set $\mathcal{X}$, and let $\bar{D}_m(\tilde{\sigma})$ and $\bar{D}_{n,\mathcal{X}}(\tilde{\sigma})$ denote the expected divergence of its member and non-member components, respectively. The expected target-set divergence can then be expressed as

$$\bar{D}_{\mathcal{X}}(\tilde{\sigma}) = \pi \bar{D}_m(\tilde{\sigma}) + (1 - \pi)\bar{D}_{n,\mathcal{X}}(\tilde{\sigma}). \tag{8}$$

We define the discrepancy between the target non-members and the known non-member reference set $\mathcal{R}$ as

$$\delta_{\mathcal{R}}(\tilde{\sigma}) = \bar{D}_{n,\mathcal{X}}(\tilde{\sigma}) - \bar{D}_{\mathcal{R}}(\tilde{\sigma}). \tag{9}$$

Defining the desired signal as $\Delta\bar{D}(\tilde{\sigma}) = \bar{D}_m(\tilde{\sigma}) - \bar{D}_{n,\mathcal{X}}(\tilde{\sigma})$, the observable tuning gap can then be expressed as:

$$\begin{aligned}
\Delta\bar{D}_{\text{tune}}(\tilde{\sigma}) &= \bar{D}_{\mathcal{X}}(\tilde{\sigma}) - \bar{D}_{\mathcal{R}}(\tilde{\sigma}) \\
&= \pi \bar{D}_m(\tilde{\sigma}) + (1 - \pi)\bar{D}_{n,\mathcal{X}}(\tilde{\sigma}) - \bar{D}_{\mathcal{R}}(\tilde{\sigma}) \\
&= \pi \left[\bar{D}_m(\tilde{\sigma}) - \bar{D}_{n,\mathcal{X}}(\tilde{\sigma})\right] + \delta_{\mathcal{R}}(\tilde{\sigma}) \\
&= \pi\Delta\bar{D}(\tilde{\sigma}) + \delta_{\mathcal{R}}(\tilde{\sigma}).
\end{aligned} \tag{10}$$

Thus, the observable gap comprises the desired membership signal, attenuated by the unknown member proportion π, and a reference-set discrepancy term $\delta_{\mathcal{R}}(\tilde{\sigma})$. If, around the dominant extrema of the noise sweep, $|\delta_{\mathcal{R}}(\tilde{\sigma})| \ll \pi \left|\Delta\bar{D}(\tilde{\sigma})\right|$, then the observable tuning gap approximately preserves the direction and relative magnitude of the underlying membership gap: $\Delta\bar{D}_{\text{tune}}(\tilde{\sigma}) \approx \pi\Delta\bar{D}(\tilde{\sigma})$.

We empirically verify this condition using reference sets spanning varying degrees of distributional shift from the target non-members (Figure 4). Across all evaluated model–dataset and reference-set configurations, the observable tuning gap retains the direction and general noise-response pattern of the true membership gap, including when the reference set is substantially distributionally shifted. These results demonstrate that, in our evaluated settings, the membership component remains dominant over the reference-set discrepancy near the most discriminative regions of the noise sweep.

Because π is fixed for a given target set and independent of $\tilde{\sigma}$, it uniformly scales the underlying membership gap. For any $\pi > 0$, this scaling alone changes neither the direction of $\Delta\bar{D}(\tilde{\sigma})$ nor the noise level maximizing its absolute magnitude. The value of π therefore need not be known by the attacker. Moreover, a larger member proportion or a stronger underlying membership gap increases the tolerance of the tuning procedure to the reference-set discrepancy $\delta_{\mathcal{R}}(\tilde{\sigma})$. In practice, our experimental target sets fix $\pi = 0.5$, and we demonstrate that our tuning method remains effective in this setting. Accordingly, COMPASS selects the perturbation magnitude

$$\tilde{\sigma}^* = \arg\max_{\tilde{\sigma} \in \mathcal{S}} \left|\Delta\bar{D}_{\text{tune}}(\tilde{\sigma})\right|, \tag{11}$$

where $\mathcal{S}$ is the set of candidate noise levels. The sign of $\Delta\bar{D}_{\text{tune}}(\tilde{\sigma}^*)$ then determines the membership-score orientation: a positive gap indicates that higher divergence corresponds to membership, whereas a negative gap indicates that lower divergence corresponds to membership.

E Blind-Classifier Filtering

Motivated by evidence that differences between member and non-member data sources can confound membership inference evaluations Das et al. (2025); Zhu et al. (2025), we filter our self-constructed Endoscopy candidate pools using a classifier with no access to the target model. This procedure reduces the extent to which an attack can exploit image-source differences in place of target-model behavior.

We represent each image using the 512-dimensional global-average-pooled features preceding the classification head of a frozen ImageNet-pretrained ResNet-18 He et al. (2016). We train a logistic regression classifier to distinguish candidate members from non-members using five-fold stratified cross-validation. We then obtain an out-of-fold predicted probability for each image and rank images within each class by the classifier's confidence in their true label. Images assigned lower confidence are considered harder for the blind classifier to distinguish.

We retain the hardest fraction of images from each class and adjust the keep-fraction until the classifier's AUC on the retained pools falls below 0.52. The selected keep-fraction is 30%, reducing

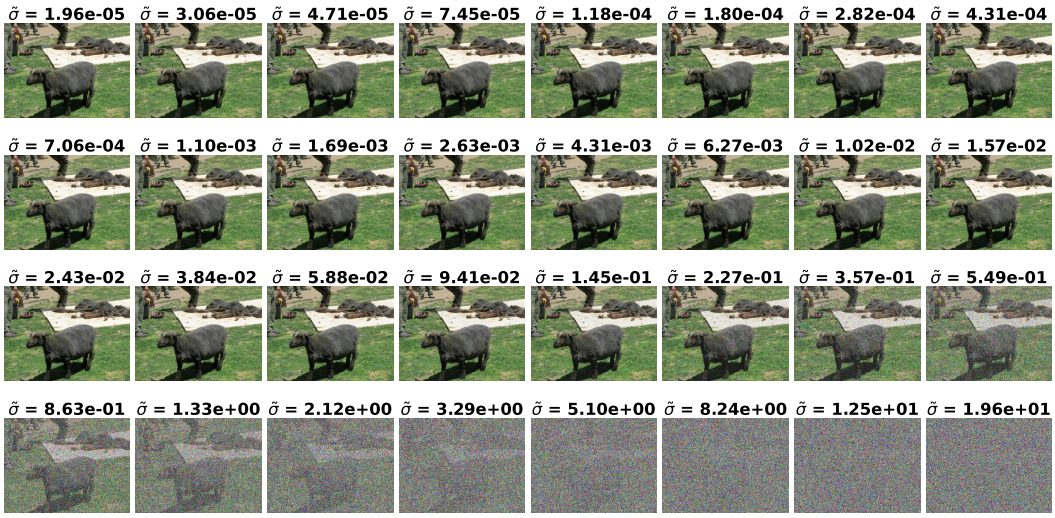


Figure 5: Raw Image Perturbations Used for Noise-level Tuning section 5 and Divergence Analyses section 4.

the pool from 18,623 candidate images to 5,588 retained images: 3,538 members and 2,050 non-members. The blind classifier's five-fold cross-validated AUC is 0.954 ± 0.002 on the full pools which reduces to 0.465 ± 0.007 on the retained pools. We report the latter as a post-filter diagnostic of separability under this blind baseline. Balanced target sets are subsequently sampled from the retained member and non-member pools.

F Additional Results

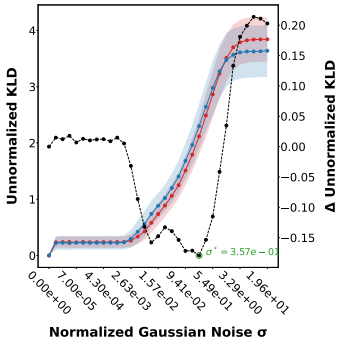
F.1 Raw Perturbed Images

We show example perturbed versions of an input image used in our membership evaluation for the MIA (Figure 5); for each image, the corresponding normalized noise standard deviation $\tilde{\sigma}$ is indicated.

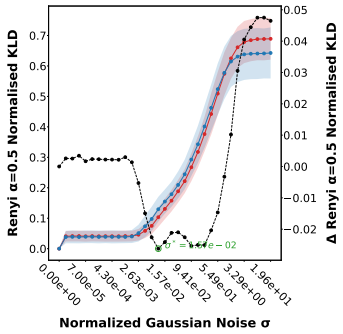
As discussed in section 5, we apply noise by independently sampling, for each pixel, from a Gaussian distribution with standard deviation $\tilde{\sigma}$, adding the resulting value to the pixel intensities (0–255), and then clipping. As shown in Figure 5, this procedure produces a noisy image in which an increasing fraction of pixels saturate to complete noise due to clipping, while the remaining pixels become progressively more faded. Notably, in looking at the regime II range of $\tilde{\sigma}$ values which often proved most optimal for our target sets in noise-level tuning (1.10e-03 to 2.43e-02), we qualitatively observe that these region adds noise that is indistinguishable to the human eye, demonstrating the minute sensitivity of VLLMs.

F.2 Raw Divergence Plots

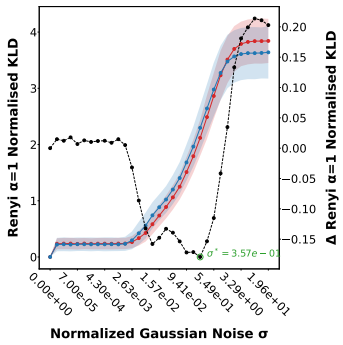
Expanding on Figure 1, we additionally provide the divergence dynamics plots for all target models (Figure 6, Figure 7, and Figure 8). The standard deviation within each class is shown as a shaded band around the corresponding curve.



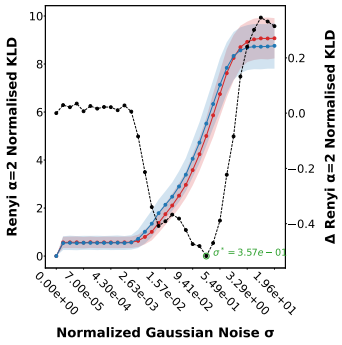
(a) Unnoramlised KL-Divergence



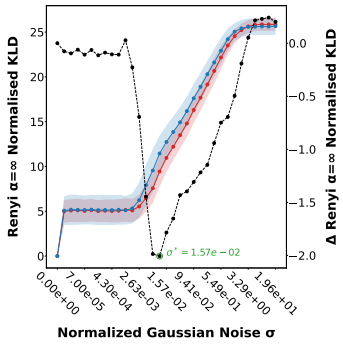
(b) Renyi $\alpha = 0.5$ Normalised KL-Divergence



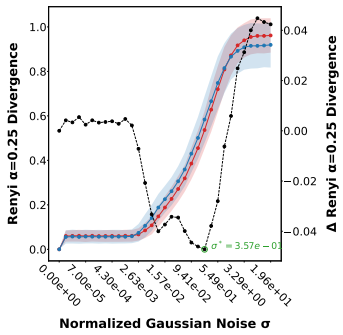
(c) Renyi $\alpha = 1$ Normalised KL-Divergence



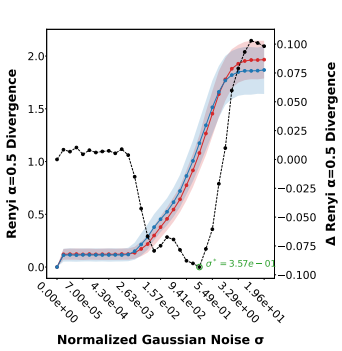
(d) Renyi $\alpha = 2$ Normalised KL-Divergence



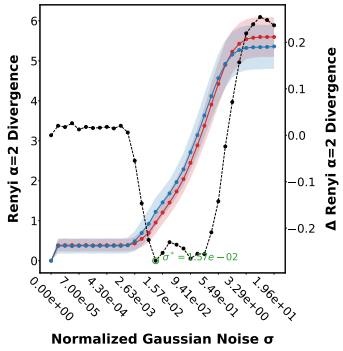
(e) Renyi $\alpha = \infty$ Normalised KL-Divergence



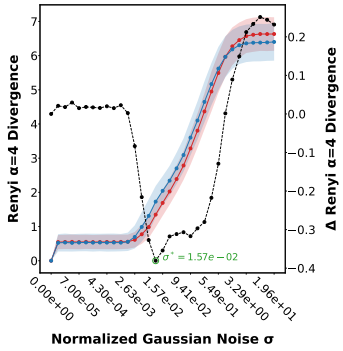
(f) Renyi $\alpha = 0.25$ Divergence



(g) Renyi $\alpha = 0.5$ Divergence

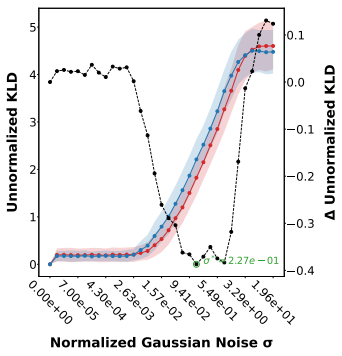


(h) Renyi $\alpha = 2$ Divergence

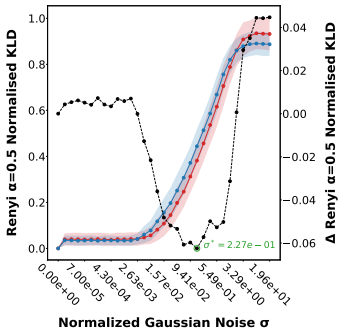


(i) Renyi $\alpha = 4$ Divergence

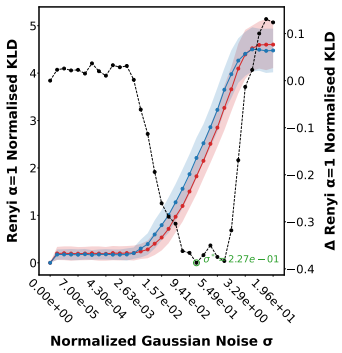
Figure 6: Raw divergence curves for all tested normalization for a flickr target subset members (red), non-members (blue), and member-non-member difference (black) against the LLaVA model.



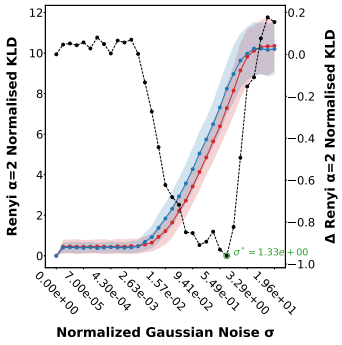
(a) Unnoramlised KL-Divergence



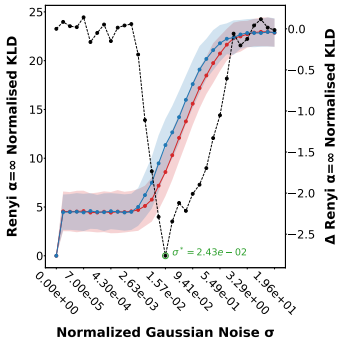
(b) Renyi $\alpha = 0.5$ Normalised KL-Divergence



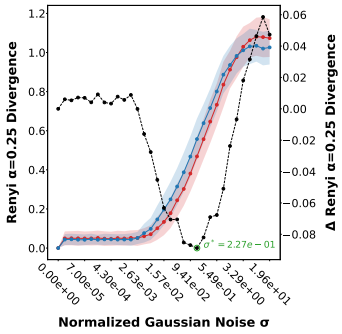
(c) Renyi $\alpha = 1$ Normalised KL-Divergence



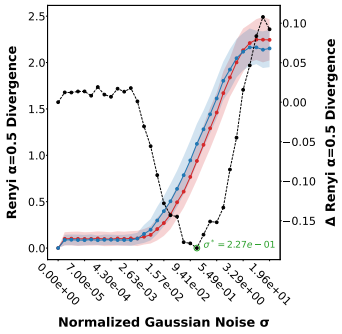
(d) Renyi $\alpha = 2$ Normalised KL-Divergence



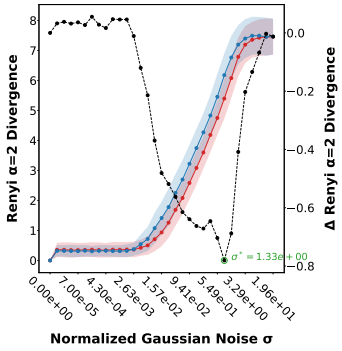
(e) Renyi $\alpha = \infty$ Normalised KL-Divergence



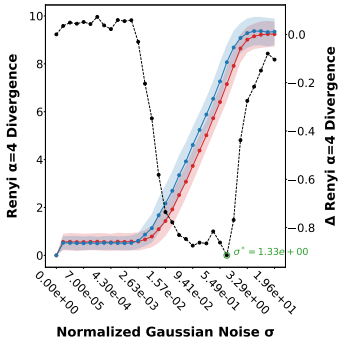
(f) Renyi $\alpha = 0.25$ Divergence



(g) Renyi $\alpha = 0.5$ Divergence

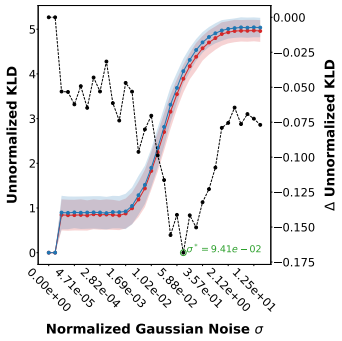


(h) Renyi $\alpha = 2$ Divergence

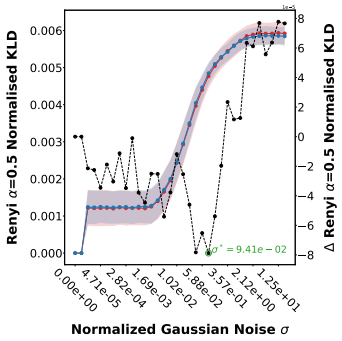


(i) Renyi $\alpha = 4$ Divergence

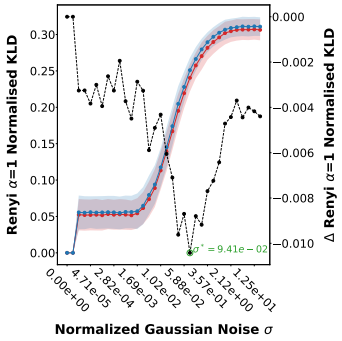
Figure 7: Raw divergence curves for all tested normalization for a flickr target subset members (red), non-members (blue), and member-non-member difference (black) against the MiniGPT-4o model.



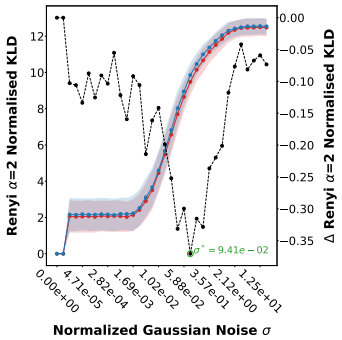
(a) Unnoramlised KL-Divergence



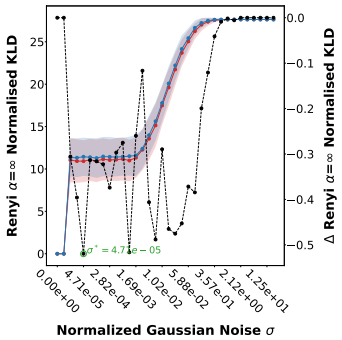
(b) Renyi $\alpha = 0.5$ Normalised KL-Divergence



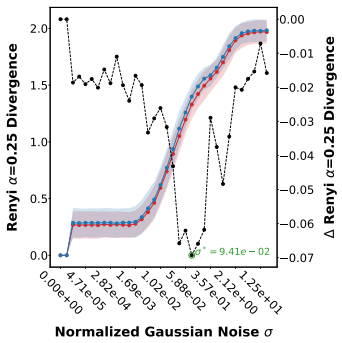
(c) Renyi $\alpha = 1$ Normalised KL-Divergence



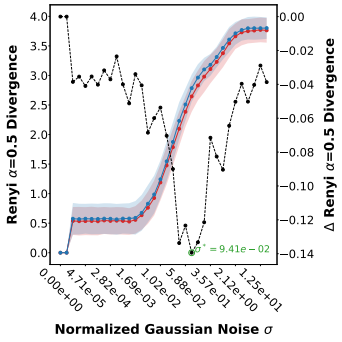
(d) Renyi $\alpha = 2$ Normalised KL-Divergence



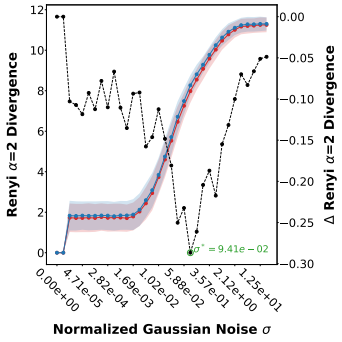
(e) Renyi $\alpha = \infty$ Normalised KL-Divergence



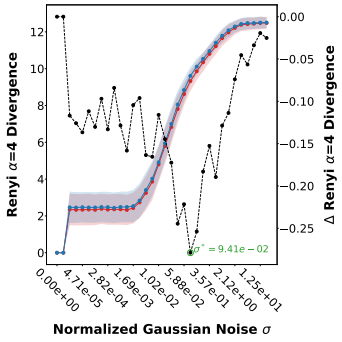
(f) Renyi $\alpha = 0.25$ Divergence



(g) Renyi $\alpha = 0.5$ Divergence



(h) Renyi $\alpha = 2$ Divergence



(i) Renyi $\alpha = 4$ Divergence

Figure 8: Raw divergence curves for all tested normalization for a endoscopy target subset members (red), non-members (blue), and member-non-member difference (black) against the Hulu-Med-32B model.