BLACK-BOX MEMBERSHIP INFERENCE ATTACK FOR VISION-LANGUAGE MODELS USING SELF-MEMORY PROBING

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid advancement of vision-language models (VLMs) increasingly depends on large-scale visual corpora, raising concerns about privacy risks and copyright compliance. Membership inference attacks (MIAs) provide a means to audit whether individual samples were used for training. However, existing VLM MIAs either require white-box or gray-box access to logits and token probabilities or rely on output-only proxies derived from task-specific visual interventions or surrogate models. These access and design dependencies limit their applicability to commercial VLMs deployed under the black-box Generation-as-a-Service (GaaS) paradigm, which exposes only final outputs. To address this gap, we propose Self-Memory Probing (SMP), an output-only black-box framework grounded in the self-memory effect. This effect manifests as greater consistency and concentration in responses to member images than in those to non-members. SMP probes this distinction by eliciting multiple outputs with semantically equivalent prompts to construct candidate-conditioned and non-member reference distributions. It then casts membership inference as a hypothesis test and uses Maximum Mean Discrepancy (MMD) to quantify the discrepancy between these distributions. Consequently, SMP derives membership evidence directly from the model's native responses, without image perturbations, surrogate models, or token-level statistics. Extensive experiments under the strict black-box setting demonstrate that SMP improves AUC by 9.9% and 5.9% over the best-performing black-box and higher-access probability-based baselines on VL-MIA/Flickr and VL-MIA/DALL-E, respectively, with TPR@5%FPR increasing 4.3$\times$ on VL-MIA/Flickr. SMP achieves further gains of 5.9% in AUC and 2.0$\times$ in TPR@5%FPR across commercial black-box VLMs on FlickrMIA-25, while attaining 22.1% lower latency.

1 INTRODUCTION

Vision-Language Models (VLMs) have advanced multimodal understanding Liu et al. (2023b; 2024); Tangtartharakul & Storrs (2026) and enabled diverse real-world applications Shirai et al. (2024); Liu et al. (2023a), including information retrieval Singhal et al. (2025); Liu et al. (2025); Hollmann et al. (2025), assisted medical diagnosis Xiang et al. (2025); Zhang et al. (2024c); Yan et al. (2025), content creation Van Veen et al. (2024), and scientific discovery Benegas et al. (2025); Lozano et al. (2025). VLMs are typically trained on large-scale image–text corpora Zhao et al. (2023); Zhang et al. (2024b), while scaling-driven competition has accelerated the collection of human-generated training data Longpre et al. (2024); Widder et al. (2024). Such large-scale, often indiscriminate collection may incorporate personally identifiable information or proprietary visual content, creating risks to data ownership and raising ethical and legal concerns Carlini et al. (2019; 2021); Mireshghallah et al. (2023). Recent high-profile disputes, including The New York Times versus Microsoft Grynbaum & Mac (2023) and Authors Guild versus OpenAI Calderon (2024), underscore the need for stronger data governance and responsible AI development.

Pre-training data auditing commonly relies on membership inference attacks (MIAs) to determine whether a sample was included in a model's training set (Shi et al., 2024; Fu et al., 2024). Initially developed for conventional machine learning, MIAs have increasingly been applied to large language models (LLMs) (Shokri et al., 2017; Shi et al., 2024). These methods typically exploit token

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.

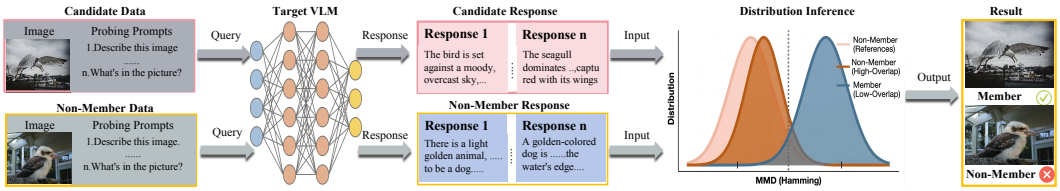


Figure 1: **Overview of the proposed SMP framework.** SMP formulates output-only membership inference attack as an MMD-based hypothesis test comparing candidate-conditioned and non-member reference response distributions, without accessing internal model states.

probabilities, likelihood scores, or internal representations. For example, Shi et al. (2024) proposed Min-K%, which aggregates the likelihoods of the least-likely $K\%$ of input tokens into a detection score. Extending this framework to VLMs supports audits of potentially unauthorized visual-data use and has motivated several recent studies (Li et al., 2024; Choquette-Choo et al., 2021; Hu et al., 2022). Early VLM approaches generally assume white-box or gray-box access to logits, token probabilities, or internal representations. Such access is incompatible with "Generation-as-a-Service" (GaaS), where deployed systems expose only generated text (Tramèr et al., 2016; Peng et al., 2024; Wen et al., 2024). Recent work has relaxed this constraint under output-only black-box access. Yin et al. (2025) proposed semantic masked prediction with prior-guided calibration to infer training-set membership. Liu et al. (2026) introduced LOMIA, which maps textual and cross-modal alignment features to proxy perplexity scores using an auxiliary VLM under label-only access. Tang et al. (2026) proposed DistractMIA, which inserts calibrated semantic distractors and evaluates response stability and distractor uptake across repeated generations. However, existing approaches remain dependent on task-specific semantic interventions or surrogate-based score estimation, leaving the potential for direct membership inference from the distributional structure of responses elicited by unmodified images largely unexplored, without access to logits or proxy token-level statistics.

To address this gap, we propose Self-Memory Probing (SMP), an output-only black-box framework that infers membership from the distributional structure of VLM responses to unmodified images, as illustrated in Figure 1. SMP builds on the self-memory effect: member images tend to elicit more consistent and concentrated response distributions, whereas non-member images yield more diverse and dispersed outputs. We operationalize this observation as a two-sample hypothesis test comparing the candidate-conditioned response distribution with a non-member reference distribution. SMP estimates their discrepancy from finite response samples using Maximum Mean Discrepancy (MMD), requiring no access to logits, token probabilities, or hidden states. To reduce prompt-specific bias and better characterize each distribution, we sample both candidate and reference responses from a shared ensemble of semantically equivalent but lexically diverse probing prompts. We instantiate MMD with a Hamming kernel that measures weighted token-level mismatches. Kernel ablations show that this choice provides the strongest discrimination among the evaluated alternatives. The main contributions are summarized as follows:

- We introduce Self-Memory Probing (SMP), an output-only black-box MIA that detects membership from memorization-induced concentration in native VLM responses, without internal signals, surrogate models, or image perturbations.
- Under stricter output-only access, SMP improves AUC over the best-performing probability-access and black-box baselines by 5.9% and 9.9% on VL-MIA/DALL-E and VL-MIA/Flickr, respectively, and achieves 5.9% higher AUC and 2.0× TPR@5%FPR on FlickrMIA-25 across commercial black-box VLMs.
- SMP is robust to probing-prompt formulation, decoding temperature, and non-member reference selection, while achieving 22.1% lower latency on commercial black-box VLMs.

2 Related Work

Membership inference attacks (MIAs) on large language models (LLMs) determine whether a text sample was included in a target model's training corpus, typically using model-derived statistics. Likelihood-based methods exploit the tendency of training samples to receive higher model likeli-

hood Yeom et al. (2018). Min-K% Shi et al. (2024) and Min-K%++ Zhang et al. (2024a) refine this signal through token-level likelihood statistics, while complementary approaches use likelihood ratios Carlini et al. (2021) or reference-model calibration Fu et al. (2024). Recent work extends inference beyond individual samples to datasets Oren et al. (2023), long sequences, and entire documents through signal aggregation Puerto et al. (2025); Meeus et al. (2024). These methods primarily audit textual training data. In contrast, we investigate visual-data membership in VLMs through self-memory probing, extending membership inference to the multimodal setting.

2.1 MIAs for Vision-Language Models

Vision-language models (VLMs) are trained on large-scale image–text corpora that may contain personal, medical, or proprietary content, motivating training-data auditing. Li et al. (2024) introduced a benchmark for VLM membership inference, a token-level image-inference pipeline, and the MaxRényi-K% metric. At the instruction-tuning stage, Hu et al. (2025) developed a temperature-sensitive set-level MIA and showed that sample aggregation amplifies overfitting signals. Other methods use output similarity (Hu et al., 2022), linear probes of internal activations (Ibanez-Lissen et al., 2025), or cosine similarity in CLIP-style encoders (Ko et al., 2023). Many nevertheless require white-box or gray-box access to logits, token probabilities, or intermediate representations. These signals are unavailable under Generation-as-a-Service (GaaS), where commercial systems typically return only generated text (Tramèr et al., 2016; Peng et al., 2024; Wen et al., 2024). Recent work therefore considers output-only settings for practical VLM auditing. Yin et al. (2025) proposed KCMP, which proposed semantic masked prediction with prior-guided calibration. Liu et al. (2026) introduced LOMIA, which uses a surrogate VLM to estimate proxy perplexity scores from text–text and image–text semantic alignment under label-only target access. Tang et al. (2026) proposed DistractMIA, which inserts calibrated semantic distractors and infers membership from response stability and distractor uptake across repeated generations. However, these methods depend on constructed visual interventions or surrogate approximations of unavailable token statistics. In contrast, SMP probes self-memory through responses to unmodified images under semantically equivalent prompts, enabling MMD-based distributional inference against a non-member reference without image-level interventions, surrogate models, or token-level proxies.

3 Methodology

Problem Formulation. Given an input image $x \in \mathcal{X}$ and a prompt p, we model the target VLM as a probabilistic distribution P and a discrete sequence sample y can be sampled from P for an input x and a prompt p:

$$y \sim P(\cdot \mid x, p). \tag{1}$$

Under the strict black-box setting, a membership inference attack aims at constructing a binary function f that determines whether a candidate image x^* belongs to the training dataset $\mathcal{D}_{\text{train}}$ of P without accessing the gradients, token-level probabilities, and logits for the VLM P:

$$f(x^*) = \begin{cases} 0, & \text{indicating } x^* \in \mathcal{D}_{\text{train}} \\ 1, & \text{indicating } x^* \notin \mathcal{D}_{\text{train}}. \end{cases} \tag{2}$$

3.1 Approach Overview

In this section, we aim to identify whether a target VLM P memorizes a given candidate image x^* or not under the black-box scenario, and leverage this signal for membership inference attack. To this end, we propose a MIAs framework termed Self-Memory Probing (SMP). SMP makes an assumption that the VLM exhibits highly diverse distributions for a pair of member and non-member samples whereas it generates two similar distributions for two non-member samples. In other words, if x^* is a member image and x is a non-member sample, then the distance between $P(\cdot|x^*, p)$ and $P(\cdot|x, p)$ are highly diverse. Leveraging this assumption, membership inference attack in the discrete space can be cast as a statistical hypothesis testing problem based on the two distributions defined by two samples. Let $P^* = P(\cdot \mid x^*, p)$ denote the conditional distribution of a candidate image x^*, and $Q = P(\cdot \mid x, p)$ be the distribution induced by a non-member sample $x \sim \mathcal{D}_{\text{non}}$. At the high level, our approach employs the statistical hypothesis testing to construct the binary

function f. Specifically, we construct the null hypothesis H_0 and the alternative hypothesis H_1:

$$\begin{aligned} H_0 &: P^* = Q \quad \text{(non-member)} \\ H_1 &: P^* \neq Q \quad \text{(member)} \end{aligned} \tag{3}$$

According to the above assumption, if H_0 is rejected and thus there is a diverse distance between P^* and Q, then x^* is a member sample, i.e. $f(x^*) = 0$; otherwise, x^* is a non-member sample, i.e. $f(x^*) = 1$. Since it is not accessible to the logits of P^* and Q under the black-box scenario, we quantify the criterion to fail to reject the null hypothesis by using the Maximum Mean Discrepancy (MMD) Gretton et al. (2012b) between P^* and Q via the sampling technique.

3.2 Distributional Discrepancy by MMD

To quantify the high-dimensional distributional discrepancy, we adopt Maximum Mean Discrepancy (MMD) Gretton et al. (2012b). MMD measures the distance between two distributions in a reproducing kernel Hilbert space induced by a kernel function k with the feature map ϕ. The raw outputs of a VLM are text sequences $y^{(i)}$ defined over a high-dimensional discrete space. Directly computing distributional distances in such a discrete space is intractable and lacks smoothness. Therefore, prior to hypothesis testing, we map discrete text outputs into a measurable feature $z_i = \phi(y^{(i)})$ in a Reproducing Kernel Hilbert Space $\mathcal{H}$ by using a feature mapping function ϕ which is implicitly defined by our chosen kernel function $k(\cdot, \cdot)$, allowing us to compute distances via the kernel trick. Given the candidate distribution P^* and the non-member reference distribution Q, the squared MMD is defined as:

$$\text{MMD}_k^2(P^*, Q) = \left\| \mathbb{E}_{y \sim P^*}[\phi(y)] - \mathbb{E}_{y' \sim Q}[\phi(y')] \right\|_k^2, \tag{4}$$

which can be further expressed as the following using a kernel function k:

$$\begin{aligned} \text{MMD}_k^2(P^*, Q) = \mathbb{E}_{y_1, y_2 \sim P^*}[k(\phi(y_1), \phi(y_2))] + \mathbb{E}_{y_1', y_2' \sim Q}[k(\phi(y_1'), \phi(y_2'))] \\ - 2\mathbb{E}_{y \sim P^*,\, y' \sim Q}[k(\phi(y), \phi(y'))]. \end{aligned} \tag{5}$$

3.3 Estimator via Probing Prompts

If $\mathcal{S}_{P^*}$ or $\mathcal{S}_Q$ is sampled from P^* and Q with a fixed prompt, the estimator in Eq. equation 7 may be not faithful. Therefore, we propose to sample both $\mathcal{S}_{P^*}$ and $\mathcal{S}_Q$ from multiple prompts. Specifically, we construct a set of probing prompts $\mathcal{P} = \{p_1, p_2, \ldots, p_K\}$, where each prompt is designed to elicit detailed descriptions of the visual content while remaining semantically equivalent (e.g., "Describe all visual details in the image," "Explain the object layout in the scene," "List all objects and their spatial relationships"). Based on this probing set, we define a discrete prompt distribution $\pi(\mathcal{P})$. To ensure a fair comparison, all candidate samples are queried under the same prompt distribution $\pi(\mathcal{P})$. This alignment guarantees that any observed distributional differences are solely attributed to the model's response to the underlying image x, rather than variations in prompt semantics. For a candidate image x^*, we sample prompts $p^{(i)} \sim \pi(\mathcal{P})$ for $i \in \{1, \ldots, N\}$ and collect the corresponding samples from P^*:

$$y^{(i)} \sim P^* = P(\cdot \mid x^*, p^{(i)}). \tag{6}$$

This process yields an output set $\mathcal{O} = \{y^{(1)}, \ldots, y^{(N)}\}$, which characterizes the model's conditional generation behavior around x^*. Similarly, we construct a non-member reference set $\mathcal{O}_{\text{non}} = \{y'^{(1)}, \ldots, y'^{(N)}\}$ using a sample from non-member images under the same prompt distribution. In this way, we estimate MMD using its empirical estimator based on finite samples. Given two sample sets $\mathcal{S}_{P^*} = \{z_i = \phi(y^{(i)})\}_{i=1}^N$ and $\mathcal{S}_Q = \{z_i' = \phi(y'^{(i)})\}_{i=1}^N$, the estimator of Eq. equation 5 is computed as:

$$\widehat{\text{MMD}}_k^2(\mathcal{S}_{P^*}, \mathcal{S}_Q) = \frac{1}{N(N-1)} \sum_{i \neq j} k(z_i, z_j) + \frac{1}{N(N-1)} \sum_{i \neq j} k(z_i', z_j') - \frac{2}{N^2} \sum_{i=1}^N \sum_{j=1}^N k(z_i, z_j'). \tag{7}$$

3.4 Kernel Choice

The choice of the kernel k and its associated feature mapping ϕ plays a critical role in determining the statistical power of the two-sample test. Recall that we map each sequence y into a measurable

feature $z = \phi(y)$. Here, we explicitly define ϕ as a multi-hot encoding function over the entire vocabulary space $\mathcal{V}$. Specifically, $z \in \{0,1\}^{|\mathcal{V}|}$, where $z_v = 1$ if token v appears in the sequence y, and 0 otherwise. This sequence-to-set transformation naturally drops the sequential position information, allowing us to focus on token-level memorization regardless of stochastic sequential decoding variance. With this explicit feature space established, we formulate our kernels as follows:

All-Substrings Kernel. To capture the occurrence of overlapping tokens efficiently without sequence-length explosion, we evaluate it via the Jaccard similarity in the multi-hot feature space:

$$k_{\text{all}}(z, z') = \frac{\sum_{v\in\mathcal{V}}(z_v \wedge z'_v)}{\sum_{v\in\mathcal{V}}(z_v \vee z'_v) + \epsilon} \tag{8}$$

where ϵ is a small constant for numerical stability(typically 10^{-8} Kingma & Ba (2014)). This kernel captures basic token overlaps often arising from verbatim memorization. However, it treats all overlapping tokens equally, failing to distinguish between common grammatical words and rare memorized entities.

One-Hot Kernel. A simple alternative is the exact-match intersection across the vocabulary, computed as the normalized dot product:

$$k_{\text{one-hot}}(z, z') = \frac{z^\top z'}{|\mathcal{V}|} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} z_v \cdot z'_v \tag{9}$$

This kernel strictly detects the co-occurrence of identical tokens. However, due to the massive vocabulary size $|\mathcal{V}|$, generated texts are extremely sparse representations. As a result, this kernel yields extremely small similarity scores, leading to low statistical power in the small-sample regime and making it prone to false negatives.

Hamming Kernel. To balance expressiveness and efficiency while mitigating the interference of high-frequency stop-words (e.g., common syntactic tokens), we adopt an IDF-weighted Hamming kernel (Hamming kernel). Specifically, we compute an Inverse Document Frequency weight w_v for each token v based on its document frequency df_v in the non-member reference set $\mathcal{D}_Q$: $w_v = \log\left(\frac{N_Q+1}{df_v+1}\right)$, where N_Q is the number of non-member queries. The kernel calculates the weighted dimension-wise similarity:

$$k_{\text{hamming}}(z, z') = 1 - \frac{\sum_{v\in\mathcal{V}} w_v \cdot \mathbf{1}\{z_v \neq z'_v\}}{\sum_{v\in\mathcal{V}} w_v + \epsilon} \tag{10}$$

In the context of MIA, the Hamming kernel aligns well with the self-memory effect. For training samples, repeated queries tend to produce responses with strong token-level consistency due to memorization, leading to high pairwise similarity within $\mathcal{S}_{P^*}$. In contrast, for non-member samples, the absence of memorization results in more diverse outputs and lower alignment. Consequently, when the candidate sample exhibits strong internal consistency while remaining dissimilar to the non-member reference set $\mathcal{S}_Q$, the resulting MMD value increases significantly. Empirically, we find that the Hamming kernel achieves high sample efficiency and strong statistical power in detecting memorization-induced distributional concentration in VLM outputs.

3.5 The Detailed Algorithm

The detailed procedure of the proposed MIA for the hypothesis testing problem in Eq. equation 3 is summarized in Algorithm 1. Given a significance level α (typically set to 0.05), we first compute the empirical MMD test statistic $\tau = \widehat{\text{MMD}}_k^2(\mathcal{S}_{P^*}, \mathcal{S}_Q)$ based on the two obtained sample sets. To estimate the statistical significance, we employ a permutation test. Specifically, we pool the samples into a union set $\mathcal{S}_{P^*} \cup \mathcal{S}_Q$ and randomly partition it into two disjoint subsets, $\mathcal{S}_A$ and $\mathcal{S}_B$, such that $|\mathcal{S}_A| = |\mathcal{S}_{P^*}|$ and $|\mathcal{S}_B| = |\mathcal{S}_Q|$. We then calculate the permuted MMD statistic $\tau_b = \widehat{\text{MMD}}_k^2(\mathcal{S}_A, \mathcal{S}_B)$. By repeating this random partitioning procedure B times (typically set to

100 to ensure a sufficient statistical resolution of $1/B < \alpha$), we approximate the null distribution using the permutation statistics $\{\tau_b\}_{b=1}^{B}$. The p-value is then estimated as the proportion of permuted statistics that are greater than or equal to the observed statistic, i.e., $p = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{\tau_b \geq \tau\}$. Finally, if $p < \alpha$, we reject the null hypothesis H_0 (identifying the candidate as a member); otherwise, we fail to reject H_0. Further details are provided in the Appendix A

Table 1: MIA performance across multiple VLMs in terms of AUC and TPR@5%FPR.Red, orange, and blue backgrounds indicate the best overall results, the best results in the white-box setting, and the second-best results in the black-box setting, respectively.

VL-MIA/Flickr (Flickr)								
Method	LLaVA		LLaMA Adapter		MiniGPT-4		Average	
	AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR
White-Box/Gray-Box								
Perplexity Carlini et al. (2021)	0.667	0.130	0.618	0.157	0.654	0.073	0.646	0.120
Aug-KL Liu et al. (2021)	0.528	0.053	0.578	0.010	0.559	0.017	0.555	0.0267
Max-Prob-Gap Li et al. (2024)	0.650	0.160	0.622	0.143	0.608	0.023	0.627	0.109
Min-K% Shi et al. (2024)	0.669	0.127	0.582	0.120	0.712	0.083	0.654	0.110
ModRényi Li et al. (2024)	0.659	0.117	0.614	0.157	0.638	0.100	0.637	0.125
MaxRényi-K% Li et al. (2024)	0.691	0.173	0.598	0.197	0.641	0.187	0.643	0.186
Black-Box								
Image Infer Hu et al. (2025)	0.512	0.041	0.516	0.033	0.502	0.007	0.510	0.027
LOMIA Liu et al. (2026)	0.581	0.068	0.566	0.043	0.528	0.015	0.558	0.042
KCMP Yin et al. (2025)	0.598	0.083	0.573	0.037	0.544	0.013	0.572	0.044
DistractMIA Tang et al. (2026)	0.615	0.089	0.585	0.042	0.578	0.018	0.593	0.049
SMP(Ours)	0.713	0.211	0.679	0.217	0.685	0.208	0.692	0.212
VL-MIA/DALL-E (DALL-E)								
Method	LLaVA		LLaMA Adapter		MiniGPT-4		Average	
	AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR
White-Box/Gray-Box								
Perplexity Carlini et al. (2021)	0.638	0.081	0.517	0.078	0.652	0.034	0.602	0.064
Aug-KL Liu et al. (2021)	0.553	0.081	0.540	0.034	0.626	0.037	0.573	0.051
Max-Prob-Gap Li et al. (2024)	0.619	0.108	0.553	0.074	0.615	0.037	0.596	0.073
Min-K% Shi et al. (2024)	0.560	0.089	0.557	0.037	0.589	0.057	0.569	0.061
ModRényi Li et al. (2024)	0.644	0.088	0.510	0.061	0.618	0.061	0.591	0.070
MaxRényi-K% Li et al. (2024)	0.584	0.173	0.600	0.162	0.530	0.063	0.571	0.133
Black-Box								
Image Infer Hu et al. (2025)	0.503	0.041	0.520	0.018	0.502	0.011	0.508	0.023
LOMIA Liu et al. (2026)	0.539	0.052	0.545	0.038	0.523	0.039	0.536	0.043
KCMP Yin et al. (2025)	0.565	0.064	0.568	0.063	0.543	0.032	0.559	0.053
DistractMIA Tang et al. (2026)	0.581	0.059	0.588	0.073	0.569	0.045	0.579	0.059
SMP(Ours)	0.665	0.225	0.656	0.183	0.661	0.106	0.661	0.171

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

In this section, we comprehensively evaluate the proposed Self-Memory Probing (SMP) method across multiple target models and benchmarks, and further investigate factors that influence its effectiveness.All our experiments follow a strict black-box evaluation protocol.

Table 2: AUC and TPR@5%FPR results on FlickrMIA-25 across multiple **commercial black-box VLMs.** Red and blue backgrounds indicate the best and second-best results, respectively.

FlickrMIA-25								
Method	GPT-4o-Image		Gemini-2.5		Gemini-2.0		Average	
	AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR	AUC	TPR@5%FPR
Image Infer Hu et al. (2025)	0.519	0.089	0.525	0.046	0.515	0.109	0.519	0.081
LOMIA Liu et al. (2026)	0.553	0.133	0.568	0.112	0.528	0.141	0.549	0.129
KCMP Yin et al. (2025)	0.565	0.145	0.573	0.135	0.549	0.151	0.562	0.143
DistractMIA Tang et al. (2026)	0.587	0.141	0.591	0.148	0.563	0.149	0.580	0.146
SMP(Ours)	0.635	0.289	0.655	0.276	0.626	0.332	0.639	0.299



(a) Results of LLaVA on Flickr. (b) Results of LLaMA Adapter on Flickr. (c) Results of MiniGPT-4 on Flickr.

Figure 2: **Robustness of SMP to probing-prompt variation on Flickr.** The heatmaps report pairwise prediction consistency rates (PCRs) across five randomly sampled probing-prompt subsets, $\mathcal{P}_1, \ldots, \mathcal{P}_5$. PCR remains above 98% across all comparisons, demonstrating stable membership decisions under diverse prompt formulations.

Datasets. We evaluate SMP on two image-membership benchmarks with balanced member and non-member splits Li et al. (2024). VL-MIA/DALL-E (DALL-E) contains 592 images, with members drawn from the LAION_CCS subset used for VLM pretraining. For each member, a matched non-member is synthesized by prompting DALL-E-2 with a BLIP-generated caption to preserve semantic and visual similarity. VL-MIA/Flickr (Flickr) contains 600 images, half of which are MS-COCO member images hosted on Flickr; the remaining non-members were uploaded after January 1, 2024. To approximate realistic black-box settings in which model parameters and pretraining data are unavailable, we additionally construct FlickrMIA-25 following Dubiński et al. (2024). Its members are sourced from LAION2B, whereas its non-members are Flickr images released after August 1, 2025, providing a temporally separated evaluation split.

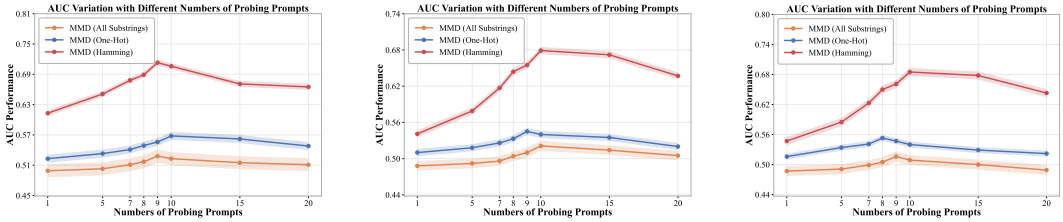
Target Models. We evaluate our proposed method on representative open-source vision-language models (VLMs), including MiniGPT-4 Zhu et al. (2024), LLaVA-1.5 Liu et al. (2023b), and LLaMA-Adapter v2 Gao et al. (2023). These models differ in their architectures, pretraining data, and instruction-tuning procedures, allowing us to assess whether the proposed method remains consistent across different VLM instantiations. To enable a practical evaluation of membership inference performance in real-world closed-source settings, we further evaluate online black-box models using FlickrMIA-25, including Gemini-2.5 Google DeepMind (2025), GPT-4o-Image Islam & Moushi (2025), and Gemini-2.0 Google DeepMind (2025), thereby assessing the effectiveness and generalizability of the proposed method across diverse commercial closed-source models.

Evaluation Metrics. We evaluate MIA performance at two levels. Following prior work Li et al. (2024); Shi et al. (2024), we evaluate membership-inference performance using two metrics. The primary metric is the area under the receiver operating characteristic curve (ROC AUC), which summarizes ranking performance across decision thresholds; higher values indicate better discrimination. We also report the true-positive rate at a false-positive rate of 5% (TPR@5%FPR), a low-FPR operating-point metric. Members are treated as the positive class, following Van et al. (2024), to better characterize performance under strict operating conditions.

Baselines We compare SMP against ten representative MIA baselines spanning different target-model access levels. For probability-based attacks, we include the loss-based attack (Carlini et al., 2021), Min-K% Prob (Shi et al., 2024), Aug-KL (Liu et al., 2021), MaxProbGap, MaxRényi-K% and ModRényi (Li et al., 2024). For black-box attacks, we consider methods that require only API-level access to the target VLM, thereby reflecting practical deployment settings of commercial closed-source models, including Image Infer (Hu et al., 2025), KCMP (Yin et al., 2025), LOMIA (Liu et al., 2026), and DistractMIA (Tang et al., 2026). More details are provided in Appendix C.

4.2 Membership Inference Attack Performance

Table 1 summarizes the AUC and TPR@5%FPR results of SMP and the baseline methods.Under strict output-only access, SMP achieved the highest average performance on both VL-MIA bench-



(a) Results of LLaVA on Flickr. (b) Results of LLaMA Adapter on Flickr. (c) Results of MiniGPT-4 on Flickr.

Figure 3: **Efficiency of SMP on Flickr.** MIA performance (AUC) increases rapidly with the number of probing prompts and generally saturates at $N_p \approx 10$ across evaluated models. Further increasing N_p yields no consistent gains and occasionally degrades performance, indicating that SMP achieves strong performance with a small query budget.

marks. Its AUC and TPR@5%FPR reached 0.692 and 0.212 on Flickr and 0.661 and 0.171 on DALL-E. Relative to DistractMIA, SMP yielded absolute AUC gains of 9.9% and 8.2% and TPR ratios of $4.33\times$ and $2.90\times$, respectively. Despite using generated outputs alone, SMP also surpassed the strongest probability-based baselines requiring token-level probability access. Its absolute AUC advantages reached 3.8% relative to Min-K% on Flickr and 5.9% relative to Perplexity on DALL-E. Across the six model–dataset settings, SMP ranked first in AUC in five and in TPR@5%FPR in all six. Overall, these results demonstrate that SMP extracts a strong and broadly consistent membership signal from generated responses alone across the evaluated VLM architectures.

4.3 Membership Inference Attacks on Commercial Black-Box Models

To evaluate SMP in realistic deployment settings, we compare it with output-only black-box baselines on three commercial closed-source models, GPT-4o-Image, Gemini-2.0, and Gemini-2.5, using FlickrMIA-25. As shown in Table 2, SMP achieves the highest AUC and TPR@5%FPR on all three models. Its average AUC and TPR@5%FPR reach 0.639 and 0.299, compared with 0.580 and 0.146 for DistractMIA, the strongest competing baseline. These results represent an absolute AUC advantage of 5.9% and a TPR ratio of $2.05\times$. The consistent gains across the evaluated commercial closed-source models indicate that SMP remains effective under strict output-only access, supporting its practical applicability to membership auditing of closed-source systems without access to model parameters or token-level probabilities.

4.4 Membership Inference Attack Robustness

To assess prompt sensitivity, we constructed a pool $\mathcal{P}$ of $K = 5$ semantically equivalent prompts with diverse linguistic realizations, including Describe all visual details in the image," Explain the object layout," and "List objects and their spatial relationships." We randomly sampled five equal-sized subsets $\mathcal{P}_1, \ldots, \mathcal{P}_5$ and evaluated SMP under otherwise identical conditions. The prediction consistency rate (PCR), defined as the proportion of identical membership decisions between two configurations, remained at least 98% across all model–dataset settings in Figs. 2 and 7. As shown in Fig. 9, SMP performance varied by no more than 1% across the tested decoding temperatures, indicating low sensitivity to generation stochasticity. We also evaluated five randomly sampled non-member reference subsets $N_1, \ldots, N_5$. As shown in Fig. 4, pairwise PCR ranged from 96.6% to 98.1%. Together, these results demonstrate that SMP was robust to the evaluated variations in prompt wording, decoding temperature, and reference-set composition. More detailed experimental results and analyses of SMP are provided in the Appendix D.

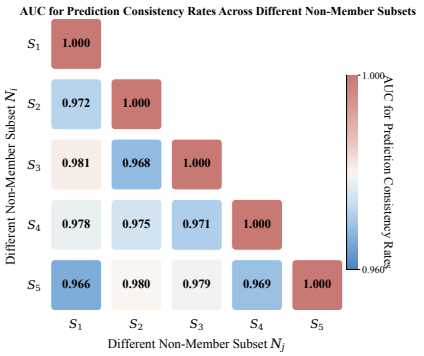


Figure 4: Robustness of SMP against different non-member reference distributions.

4.5 Membership Inference Attack Efficiency

API rate limits and per-query costs make efficiency critical in black-box Generation-as-a-Service (GaaS) deployments. We varied the probing-prompt count N_p and recorded AUC, treating N_p as the query budget. As shown in Figs. 3 and 8, AUC rose rapidly with N_p and generally peaked or plateaued near $N_p \approx 10$; using 15 or 20 prompts provided no consistent gains and occasionally reduced performance. Thus, about ten prompts achieved the best observed performance, with little benefit from additional queries. We evaluated SMP's auditing cost on commercially deployed black-box models; Fig. 10 shows that SMP reduced latency by 22.1% relative to DistractMIA, further supporting practical black-box auditing. Additional experimental details are provided in Appendix D.4.

4.6 Ablation Study and Analysis

Effect of Kernel Choice. SMP uses Maximum Mean Discrepancy (MMD) to quantify differences between the response distributions induced by candidate and non-member reference images. Because the kernel defines the feature space in which MMD compares these distributions, its choice affects the sensitivity of the membership statistic. We therefore compared the Hamming kernel with two text-based alternatives, All-Substrings and One-Hot, across three VLMs on Flickr and DALL·E. As shown in Figure 5, All-Substrings achieved AUC values only marginally above chance, while One-Hot provided modest improvements but remained inferior to Hamming. The Hamming kernel attained the highest AUC across all model–dataset combinations. On Flickr, for example, it achieved an AUC of 0.713, exceeding the 0.528 obtained with All-Substrings by 18.5 percentage points. These consistent gains support the effectiveness of the Hamming kernel in capturing membership-related differences in VLM response distributions.

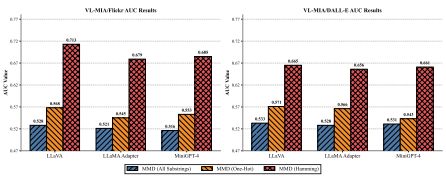


Figure 5: **Ablation study of SMP kernel functions.** Performance comparison of MMD with Hamming, One-Hot, and All Substrings kernels across three VLMs and two datasets. Unlike the All Substrings and One-Hot kernels, which yield near-random AUC, our proposed MMD (Hamming) significantly outperforms them. This validates its superiority in effectively capturing self-memory discrepancies.

Analysis The ablation results highlight how token-occurrence information is represented and weighted in MMD. These kernels operate on vocabulary-level multi-hot representations, preserving token occurrence while discarding word order. The All-Substrings kernel uses unweighted Jaccard similarity, whereas One-Hot computes an unweighted token intersection normalized by vocabulary size. Consequently, both kernels assign equal importance to frequent generic tokens and rarer tokens that may contain more image-specific information. In contrast, the Hamming kernel measures IDF-weighted token-presence disagreement, reducing the influence of frequent tokens while emphasizing mismatches involving rarer tokens. This weighting strategy is consistent with the self-memory hypothesis underlying SMP. Under this hypothesis, member-conditioned responses form more concentrated token-occurrence distributions, whereas non-member-conditioned responses exhibit greater variation. The Hamming formulation may therefore provide MMD with a more discriminative representation of membership-related distributional differences, consistent with its higher AUC across the evaluated settings. These findings support the Hamming kernel as a key design choice for extracting membership signals from generated text under strict black-box access.

5 Conclusion

In this work, we introduce Self-Memory Probing (SMP), an output-only black-box framework that exploits the greater concentration of member-conditioned responses by formulating membership inference as a statistical hypothesis test based on the MMD between candidate and non-member response distributions. Experiments across multiple VLMs and datasets show that SMP consistently outperforms existing black-box baselines and, in most experiments, surpasses higher-access methods relying on token probabilities or internal representations. These findings support the practical potential of SMP for auditing visual training data under Generation-as-a-Service constraints.

AI USE STATEMENT

We did not use generative AI tools for generating synthetic data, developing theoretical models or conceptual frameworks, formulating mathematical claims or proofs, proposing or refining hypotheses, designing the methodology or experiments, implementing methods, or interpreting results. Generative AI tools were used only to assist with manuscript writing and language polishing. All AI-assisted text was carefully reviewed and verified by the authors to ensure that it accurately reflects the intended meaning. The authors take full responsibility for the final content of this work.

REFERENCES

Gonzalo Benegas, Carlos Albors, Alan J Aw, Chengzhong Ye, and Yun S Song. A dna language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, 43(12):1960–1965, 2025.

Ruben Calderon. Ai training through copyrighted works as infringement: perspectives under the berne three-step test and the pane fair use test and plausible solutions. *IDEA*, 65:84–103, 2024.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

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.

Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *International conference on machine learning*, pp. 1964–1974. PMLR, 2021.

Jan Dubiński, Antoni Kowalczuk, Stanisław Pawlak, Przemyslaw Rokita, Tomasz Trzciński, and Paweł Morawiecki. Towards more realistic membership inference attacks on large diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4860–4869, 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.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Google DeepMind. Gemini 2.0 flash model card. Technical report, Google DeepMind, 2025.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012a.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012b.

Michael M Grynbaum and Ryan Mac. The times sues openai and microsoft over ai use of copyrighted work. *The New York Times*, 27(1), 2023.

Noah Hollmann et al. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Pingyi Hu, Zihan Wang, Ruoxi Sun, Hu Wang, and Minhui Xue. M^4i: 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. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 1589–1608, 2025.

Luis Ibanez-Lissen, Lorena Gonzalez-Manzano, Jose Maria de Fuentes, Nicolas Anciaux, and Joaquin Garcia-Alfaro. Lumia: Linear probing for unimodal and multimodal membership inference attacks leveraging internal llm states. In *European Symposium on Research in Computer Security*, pp. 186–206. Springer, 2025.

Raisa Islam and Owana Marzia Moushi. Gpt-4o: The cutting-edge advancement in multimodal llm. In *Intelligent Computing-Proceedings of the Computing Conference*, pp. 47–60. Springer, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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.

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.

Che Liu, Sibo Cheng, Chen Chen, Mengyun Qiao, Weitong Zhang, Anand Shah, Wenjia Bai, and Rossella Arcucci. M-flag: Medical vision-language pre-training with frozen language models and latent space geometry optimization. In *International conference on medical image computing and computer-assisted intervention*, pp. 637–647. Springer, 2023a.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

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.

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, et al. A generalist medical language model for disease diagnosis assistance. *Nature medicine*, 31(3):932–942, 2025.

Yihao Liu, Xinqi Lyu, Dong Wang, Yanjie Li, and Bin Xiao. Lomia: Label-only membership inference attacks against pre-trained large vision-language models. *Advances in Neural Information Processing Systems*, 38:153714–153737, 2026.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. A large-scale audit of dataset licensing and attribution in ai. *Nature Machine Intelligence*, 6(8):975–987, 2024.

Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, et al. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19724–19735, 2025.

Matthieu Meeus, Shubham Jain, Marek Rei, and Yves-Alexandre de Montjoye. Did the neurons read your book? document-level membership inference for large language models. In *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 2369–2385, 2024.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. In *The Twelfth International Conference on Learning Representations*, 2023.

Yonatan Oren, Nicole Meister, Niladri S Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Yuefeng Peng, Jaechul Roh, Subhransu Maji, and Amir Houmansadr. Oslo: one-shot label-only membership inference attacks. *Advances in Neural Information Processing Systems*, 37:62310–62333, 2024.

Haritz Puerto, Martin Gubri, Sangdoo Yun, and Seong Joon Oh. Scaling up membership inference: When and how attacks succeed on large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4165–4182, 2025.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Keisuke Shirai, Cristian C Beltran-Hernandez, Masashi Hamaya, Atsushi Hashimoto, Shohei Tanaka, Kento Kawaharazuka, Kazutoshi Tanaka, Yoshitaka Ushiku, and Shinsuke Mori. Vision-language interpreter for robot task planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2051–2058. IEEE, 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.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010.

Hongyi Tang, Zhihao Zhu, and Yi Yang. Distractmia: Black-box membership inference on vision-language models via semantic distraction. *arXiv preprint arXiv:2605.12574*, 2026.

Gene Tangtartharakul and Katherine R Storrs. Visual language models show widespread visual deficits on neuropsychological tests. *Nature Machine Intelligence*, pp. 1–11, 2026.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pp. 601–618, 2016.

Minh-Hao Van, Prateek Verma, and Xintao Wu. On large visual language models for medical imaging analysis: An empirical study. In *2024 IEEE/ACM conference on connected health: applications, systems and engineering technologies (CHASE)*, pp. 172–176. IEEE, 2024.

Dave Van Veen, Cara Van Uden, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

Rui Wen, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against in-context learning. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 3481–3495, 2024.

David Gray Widder, Meredith Whittaker, and Sarah Myers West. Why 'open'ai systems are actually closed, and why this matters. *Nature*, 635(8040):827–833, 2024.

Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025.

Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Lie Ju, Gin Tan, et al. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, 31(8):2691–2702, 2025.

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.

Jinhua Yin, Peiru Yang, Chen Yang, Huili Wang, Zhiyang Hu, Shangguang Wang, Yongfeng Huang, and Tao Qi. Black-box membership inference attack for lvlms via prior knowledge-calibrated memory probing. *arXiv preprint arXiv:2511.01952*, 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*, 2024a.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024b.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature medicine*, 30(11):3129–3141, 2024c.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2):1–124, 2023.

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

A Additional Notes on the MMD Test

A.1 Which MMD Kernels Induce Valid Metrics?

A metric on probability distributions is nonnegative and symmetric, satisfies the triangle inequality, and equals zero if and only if the two distributions coincide. A pseudometric satisfies the first three properties but may assign zero distance to distinct distributions. For a fixed positive-semidefinite kernel k, the unsquared MMD is defined as

$$D_k(P, Q) = \mathrm{MMD}_k(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_k}, \tag{11}$$

where μ_P and μ_Q denote the kernel mean embeddings of P and Q. Because D_k is an RKHS norm, it is symmetric and satisfies

$$D_k(P_1, P_3) \leq D_k(P_1, P_2) + D_k(P_2, P_3). \tag{12}$$

The squared quantity MMD_k^2 used by SMP is a distributional discrepancy but does not generally satisfy the triangle inequality. Moreover, D_k defines a metric if and only if k is characteristic, meaning that (Gretton et al., 2012a; Sriperumbudur et al., 2010).

$$D_k(P, Q) = 0 \quad \Longleftrightarrow \quad P = Q \tag{13}$$

SMP first maps every generated response $y \in \mathcal{Y}$ to a vocabulary-level multi-hot representation

$$g(y) = z \in \{0, 1\}^{|\mathcal{V}|}, \qquad z_v = \mathbb{I}\{v \text{ occurs in } y\}. \tag{14}$$

For response distributions P and Q, let $\widetilde{P}$ and $\widetilde{Q}$ denote the distributions induced over these token-presence vectors. Because g discards token order and multiplicity, it is non-injective. Distinct response distributions may therefore induce the same token-presence distribution. Consequently, the kernels considered below cannot define metrics over the original response distributions without additional assumptions. The following analysis concerns the induced distributions $\widetilde{P}$ and $\widetilde{Q}$.

One-Hot Kernel. The One-Hot kernel used in SMP is implemented as a normalized linear kernel over multi-hot token-presence representations:

$$k_{\text{one-hot}}(z, z') = \frac{z^\top z'}{|\mathcal{V}|}. \tag{15}$$

Its population MMD can be expressed as

$$\text{MMD}^2_{\text{one-hot}}(\widetilde{P}, \widetilde{Q}) = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \left[\widetilde{P}(z_v = 1) - \widetilde{Q}(z_v = 1)\right]^2. \tag{16}$$

Thus, this kernel compares only the marginal probability that each token appears. It cannot distinguish distributions with identical token-presence marginals but different dependencies among tokens. For example,

$$P_{\text{ex}} = \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(1,1)}, \qquad Q_{\text{ex}} = \frac{1}{2}\delta_{(1,0)} + \frac{1}{2}\delta_{(0,1)} \tag{17}$$

have identical coordinate-wise marginals despite satisfying $P_{\text{ex}} \neq Q_{\text{ex}}$. Therefore,

$$\text{MMD}_{\text{one-hot}}(P_{\text{ex}}, Q_{\text{ex}}) = 0, \tag{18}$$

showing that this kernel is not characteristic and induces only a pseudometric.

All-Substrings Kernel. The All-Substrings kernel used in our implementation measures Jaccard similarity between the token sets represented by z and z':

$$k_{\text{all}}(z, z') = \frac{\sum_{v \in \mathcal{V}} (z_v \wedge z'_v)}{\sum_{v \in \mathcal{V}} (z_v \vee z'_v) + \epsilon}. \tag{19}$$

For nonempty token sets and $\epsilon \to 0$, this expression reduces to the standard Jaccard, or Tanimoto, kernel. The standard kernel is positive semidefinite on binary set representations. It therefore induces a well-defined MMD pseudometric.

However, the universality result for the classical sequence-level all-substrings kernel does not directly apply to Eq. equation 19. Positive semidefiniteness alone is insufficient to establish characteristicness. Moreover, the underlying multi-hot representation remains non-injective with respect to complete textual responses. We therefore use this kernel as an empirical comparator without claiming that it defines a metric on the original response distributions.

Hamming Kernel. For the Hamming kernel defined in the main text, let $w_v \geq 0$ denote the fixed IDF weight of token v, and define

$$W_\epsilon = \sum_{v \in \mathcal{V}} w_v + \epsilon. \tag{20}$$

The kernel can be written as

$$k_{\text{Hamming}}(z, z') = 1 - \frac{\sum_{v \in \mathcal{V}} w_v \mathbb{I}\{z_v \neq z'_v\}}{W_\epsilon}. \tag{21}$$

Equivalently,

$$k_{\text{Hamming}}(z, z') = \frac{\epsilon}{W_\epsilon} + \sum_{v \in \mathcal{V}} \frac{w_v}{W_\epsilon} \sum_{b \in \{0,1\}} \mathbb{I}\{z_v = b\}\mathbb{I}\{z'_v = b\}. \tag{22}$$

Equation equation 22 is a nonnegative weighted sum of coordinate-wise delta kernels and is therefore positive semidefinite. Its population MMD has the closed form

$$\text{MMD}^2_{\text{Hamming}}(\widetilde{P}, \widetilde{Q}) = \frac{2}{W_\epsilon} \sum_{v \in \mathcal{V}} w_v \left[\widetilde{P}(z_v = 1) - \widetilde{Q}(z_v = 1)\right]^2. \tag{23}$$

Therefore, the Hamming kernel compares IDF-weighted token-presence marginals rather than complete joint distributions. Distinct distributions may share these marginals while differing in higher-order token dependencies. Consequently, the kernel is not characteristic and induces only a pseudometric.

Compared with the One-Hot kernel, the Hamming kernel introduces token-specific weighting rather than higher-order interaction modeling. This weighting is consistent with the self-memory hypothesis because stable, content-specific tokens receive greater emphasis than common syntactic tokens. However, the kernel does not encode token order, multiplicity, or higher-order co-occurrence patterns.

Algorithm 1 Self-Memory Probing (SMP) for black-box MIAs

Require: VLM P, x^*, reference $x_{\text{non}} \sim \mathcal{D}_{\text{non}}$, prompts $\mathcal{P} = \{p_i\}_{i=1}^N$, mapping ϕ, kernel k, significance α, permutations B.
Ensure: Membership inference decision $f(x^*) \in \{0, 1\}$.
Data Sampling & Mapping: Initialize multisets $\mathcal{S}_{P^*} \leftarrow \emptyset$, $\mathcal{S}_Q \leftarrow \emptyset$
for $i = 1$ to N **do**
  $z_i \leftarrow \phi(y^{(i)})$ where $y^{(i)} \sim P(\cdot \mid x^*, p_i)$; $\mathcal{S}_{P^*} \leftarrow \mathcal{S}_{P^*} \uplus \{z_i\}$
  $z'_i \leftarrow \phi(y'^{(i)})$ where $y'^{(i)} \sim P(\cdot \mid x_{\text{non}}, p_i)$; $\mathcal{S}_Q \leftarrow \mathcal{S}_Q \uplus \{z'_i\}$
end for
MMD Statistic: $\tau_{\text{obs}} \leftarrow \widehat{\text{MMD}}_k^2(\mathcal{S}_{P^*}, \mathcal{S}_Q)$
$\mathcal{S}_{\text{pool}} \leftarrow \mathcal{S}_{P^*} \uplus \mathcal{S}_Q$
for $b = 1$ to B **do**
  Randomly partition $\mathcal{S}_{\text{pool}}$ into disjoint subsets $\mathcal{S}_A, \mathcal{S}_B$ with $|\mathcal{S}_A| = |\mathcal{S}_{P^*}|, |\mathcal{S}_B| = |\mathcal{S}_Q|$
  $\tau_b \leftarrow \widehat{\text{MMD}}_k^2(\mathcal{S}_A, \mathcal{S}_B)$
end for
Hypothesis Decision: $p \leftarrow \frac{1}{B} \sum_{b=1}^B \mathbf{1}\{\tau_b \geq \tau_{\text{obs}}\}$
if $p < \alpha$ **then**
  return $f(x^*) = 0$ {Reject H_0; candidate is a Member}
else
  return $f(x^*) = 1$ {Fail to reject H_0; Non-member}
end if

Reference-Dependent IDF Weights. The preceding analysis assumes that the IDF weight vector is fixed when comparing two distributions. In SMP, the weights are estimated from the non-member reference set associated with each experimental configuration. They are then held fixed across all candidate comparisons and permutation trials within that configuration.

If the weights were recomputed from the current reference distribution for every candidate comparison, exchanging P and Q could change the kernel itself. The resulting quantity would be a reference-adaptive discrepancy rather than a symmetric metric or pseudometric. Fixing the weights within each configuration preserves the interpretation established above.

Implications for SMP. Characteristicness is not required for using MMD as an empirical two-sample statistic. Under the exchangeability assumptions of the permutation procedure, a fixed pseudometric can still provide valid type-I error control. Characteristicness instead determines whether the test can consistently distinguish every possible pair of unequal distributions.

Accordingly, SMP detects membership-associated differences expressed through token-presence statistics. In particular, the Hamming kernel emphasizes differences in IDF-weighted token-presence marginals that are consistent with the self-memory effect. It is not intended to distinguish arbitrary differences between complete textual response distributions.

B PROBING-PROMPT POOL

To support the probing-prompt construction described in the main method, we provide further details on the prompt design used by SMP. All probing prompts share the same task objective of eliciting an image description, while varying lexical choice, syntactic structure, requested granularity, and the ordering of visual attributes. Their common semantic scope includes the overall scene, salient objects, visual attributes, spatial arrangement, and surrounding context. This controlled variation allows SMP to sample the conditional response distribution beyond a single fixed instruction while avoiding changes to the underlying task semantics. Within each experimental run, the same prompt-sampling configuration is applied to candidate and non-member reference images to ensure matched prompting conditions. The following prompts are representative examples of the formulations considered during prompt construction. They illustrate the linguistic and structural diversity of the probing-prompt pool but do not constitute a fixed set that is exhaustively queried in every experi-

Figure 6: Representative Examples of Probing Prompts from the SMP Prompt Pool.

A Subset of the Probing Prompt Pool

1. Describe the main visual content of the image, including the objects and overall scene.
2. Describe what is visible in the image, focusing on the main objects and their arrangement.
3. Give a brief description of the image, including its main objects, setting, and visual composition.
4. Describe the scene shown in the image and identify the important objects and their relative positions.
5. Explain what can be seen in the image, including the major objects, surrounding environment, and overall layout.
6. Provide a clear description of the image, focusing on the main visual elements and how they are arranged within the scene.
7. Describe the visual content of the image in detail, including the main objects, their locations, and the surrounding context.
8. Explain the overall scene depicted in the image and describe the important objects and their spatial relationships.
9. Provide a detailed description of what is shown in the image, including the major objects, their positions, and relevant environmental details.
10. Examine the image carefully and describe its main visual elements, focusing on the objects present, their arrangement, and the overall scene.
11. Describe the image comprehensively by identifying the major objects, their visual attributes, relative positions, and relationships within the scene.
12. Provide a detailed account of the visual scene, including its overall setting, prominent objects, spatial arrangement, and notable surrounding elements.
13. Carefully analyze the image and describe its visual content, covering the main objects, their appearance, locations, interactions, and spatial relationships.
14. Describe the image from a visual perspective, identifying the principal objects and important details while explaining how they are positioned relative to one another.
15. Give a thorough description of the image, starting with the overall scene and then detailing the key objects, their attributes, locations, and relationships.
16. Examine the image closely and explain what is depicted, including the overall setting, major visual elements, their positions, and meaningful spatial relationships.
17. Produce a comprehensive visual description of the image that captures both the overall scene and local details, including the major objects, their attributes, arrangement, and surrounding environment.
18. Analyze the visual content of this image in detail. Describe the overall composition, identify the main objects and environmental elements, and characterize their relative positions and spatial relationships.
19. Carefully inspect the entire image and provide a detailed natural-language description of the scene, covering its overall composition, principal objects, visual attributes, background elements, and spatial relationships.
20. Examine the image comprehensively and describe everything that is visually salient. Begin with a high-level summary of the scene, then discuss the major objects, their appearance and relative locations, the surrounding environment, and the spatial relationships among the visual elements.

ment. Each experiment uses only the required number of probing prompts. As shown in Fig. 6, we provide a subset of the probing prompts used in SMP.

C Baseline Details

The first six baselines require token-level predictive distributions from the target VLM, whereas the remaining four assume output-only target access.

Probability-based baselines. **Perplexity (Loss Attack)** (Carlini et al., 2021) computes sequence perplexity from the mean negative log-likelihood of observed target tokens, with lower values pro-

viding stronger evidence of membership. **Aug-KL** (Liu et al., 2021) quantifies perturbation sensitivity using the KL divergence between token-level predictive distributions produced by a candidate image and its augmented counterpart. **Max-Prob-Gap** (Li et al., 2024) averages the probability margin between the most likely and second-most-likely next tokens across positions, with larger margins reflecting greater predictive confidence. **Min-K% Prob** (Shi et al., 2024) averages the log-probabilities of the least-likely $K\%$ target tokens, thereby emphasizing the least confidently predicted portions of a sequence. **MaxRényi-K%** (Li et al., 2024) averages the largest $K\%$ position-level Rényi entropies of the full next-token distributions, with lower aggregated entropy providing stronger membership evidence. **ModRényi** (Li et al., 2024) extends Rényi scoring to target-based inference by conditioning modified entropy on each observed next token and averaging the resulting values across the sequence.

Baselines with output-only target access. **Image Infer** (Hu et al., 2025) repeatedly queries each candidate image and averages pairwise similarities among the generated descriptions, with greater response consistency indicating membership. **KCMP** (Yin et al., 2025) constructs shape- and color-based semantic masking probes, filters tasks solvable through general knowledge, and uses the target VLM's repeated recovery accuracy as the membership score. **LOMIA** (Liu et al., 2026) fits a surrogate-VLM regressor from text–text, image–text, or joint similarity features to perplexity and uses negative predicted perplexity as the membership score. **DistractMIA** (Tang et al., 2026) calibrates semantic distractors on a reference set and infers membership from response stability and distractor uptake across repeated generations from original and distractor-augmented images.

D Experiment Details

D.1 Robustness to Probing-Prompt Selection

To evaluate the sensitivity of SMP to probing-prompt selection, we randomly sample five equal-sized prompt configurations $\{\mathcal{P}^{(r)}\}_{r=1}^{5}$ from the pool described in Appendix B. Each configuration contains N_p semantically aligned image-description prompts with varied lexical, syntactic, and descriptive formulations. For each run, the same configuration is applied to all candidate and non-member reference images. We then rerun the complete SMP pipeline while holding the candidate set, reference set, response-sampling budget, decoding parameters, Hamming kernel, and remaining experimental settings fixed.

Let $\hat{m}_r(x) \in \{0, 1\}$ denote the membership prediction for candidate image x under configuration $\mathcal{P}^{(r)}$. We quantify pairwise decision agreement using the prediction consistency rate (PCR):

$$\mathrm{PCR}\Big(\mathcal{P}^{(i)}, \mathcal{P}^{(j)}\Big) = \frac{1}{|\mathcal{D}_{\mathrm{eval}}|} \sum_{x \in \mathcal{D}_{\mathrm{eval}}} \mathbb{I}\left[\hat{m}_i(x) = \hat{m}_j(x)\right], \tag{24}$$

where $\mathcal{D}_{\mathrm{eval}}$ denotes the candidate evaluation set. A higher PCR indicates greater stability of SMP against variations in probing-prompt formulation.

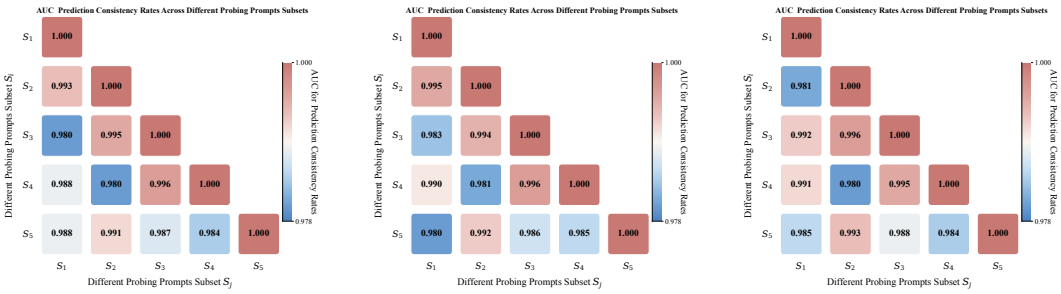
D.2 Robustness to Decoding Temperature

To assess sensitivity to generation stochasticity, we evaluate SMP over a prespecified set of decoding temperatures $\mathcal{T} = \{T_1, \ldots, T_L\} \subseteq [0, 1]$ that spans the interval from 0 to 1. At each temperature, responses are regenerated for all candidate and non-member reference images, and SMP is independently evaluated for every model–dataset setting. Decoding temperature is the only varied factor; the candidate set, reference set, prompt configuration, response-sampling budget, Hamming kernel, and remaining generation parameters are held fixed.

For each model m and dataset d, we compute the AUC from the continuous membership scores obtained at every $T \in \mathcal{T}$. Temperature sensitivity is summarized as

$$\Delta_{\mathrm{AUC}}^{(m,d)} = \max_{T \in \mathcal{T}} \mathrm{AUC}_{m,d}(T) - \min_{T \in \mathcal{T}} \mathrm{AUC}_{m,d}(T). \tag{25}$$

A smaller $\Delta_{\mathrm{AUC}}^{(m,d)}$ indicates lower sensitivity to decoding temperature.



(a) Results of LLaVA on DALL-E. (b) Results of LLaMA Ada. on DALL-E. (c) Results of MiniGPT-4 on DALL-E.

Figure 7: **Robustness of SMP to probing-prompt variation on DALL-E.** The heatmaps report pairwise prediction consistency rates (PCRs) across five randomly sampled probing-prompt subsets, $\mathcal{P}_1, \ldots, \mathcal{P}_5$. PCR remains above 98% across all comparisons, demonstrating stable membership decisions under diverse prompt formulations.

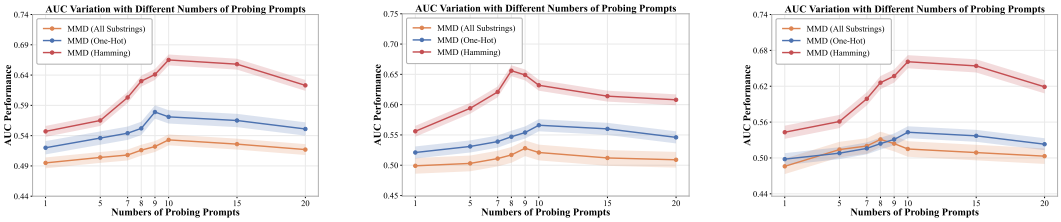
D.3 Robustness to Non-Member Reference-Set Selection

To examine the dependence of SMP on non-member reference selection, we randomly sample five equal-sized subsets $\{\mathcal{R}_j\}_{j=1}^{5}$ from the available non-member reference pool $\mathcal{D}_{\text{non}}$, where $|\mathcal{R}_j| = N_{\text{ref}}$. For each subset, we reconstruct the corresponding non-member response distribution and rerun SMP on the same candidate set. All reference-dependent quantities, including the IDF weights used by the Hamming kernel, are re-estimated from the selected subset. The prompt configuration, decoding parameters, response-sampling budget, and remaining experimental settings are otherwise unchanged.

Let $\hat{m}_{\mathcal{R}_j}(x)$ denote the membership prediction obtained using $\mathcal{R}_j$. Pairwise stability across reference subsets is measured as

$$\text{PCR}(\mathcal{R}_i, \mathcal{R}_j) = \frac{1}{|\mathcal{D}_{\text{eval}}|} \sum_{x \in \mathcal{D}_{\text{eval}}} \mathbb{I}\left[\hat{m}_{\mathcal{R}_i}(x) = \hat{m}_{\mathcal{R}_j}(x)\right]. \tag{26}$$

This protocol measures robustness to reference-subset composition within the same non-member pool, rather than robustness to distribution shifts across different reference sources.



(a) LLaVA Results on DALL-E. (b) LLaMA Adapter Results on DALL-E (c) MiniGPT-4 Results on DALL-E.

Figure 8: **Efficiency of SMP on DALL-E.** MIA performance (AUC) increases rapidly with the number of probing prompts and generally saturates at $N_p \approx 10$ across evaluated models. Further increasing N_p yields no consistent gains and occasionally degrades performance, indicating that SMP achieves strong performance with a small query budget.

D.4 Auditing Cost

We quantify practical auditing cost by comparing SMP with representative output-only baselines under a unified evaluation protocol. For each method, we execute the complete inference pipeline and record the numbers of target-model and auxiliary-model calls, aggregate token consumption, and latency. The accounting covers candidate probing, reference construction, calibration, perturbation-based evaluation, and auxiliary inference. Latency is measured as the wall-clock time from pipeline initialization to the final membership decision, including model inference, communication overhead,

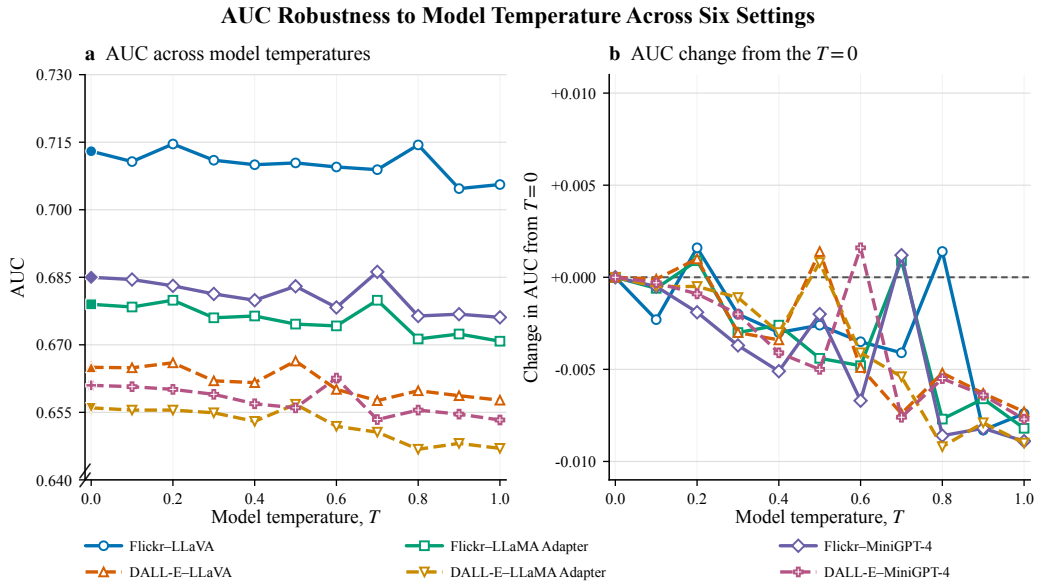


Figure 9: **SMP AUC Robustness to Model Temperature Across Six Dataset–Model Combinations.** a, AUC across model temperatures ($T = 0$–1.0), with $T = 0$ representing the reported baseline. b, Change in AUC relative to $T = 0$, with all constructed $T > 0$ values remaining within ± 0.01 of the baseline.

preprocessing, and attack-side scoring. Reusable artifacts are constructed once during the audit, with their initialization costs included. All methods use the same target models, candidate subset, decoding parameters, output-length limit, execution environment, and request-concurrency setting. Token consumption includes all input and generated output tokens from both target and auxiliary models. We report the mean number of model calls, token consumption, and latency across repeated runs on commercial VLMs, including GPT-4o-Image, Gemini-2.0, and Gemini-2.5, using FlickrMIA-25.

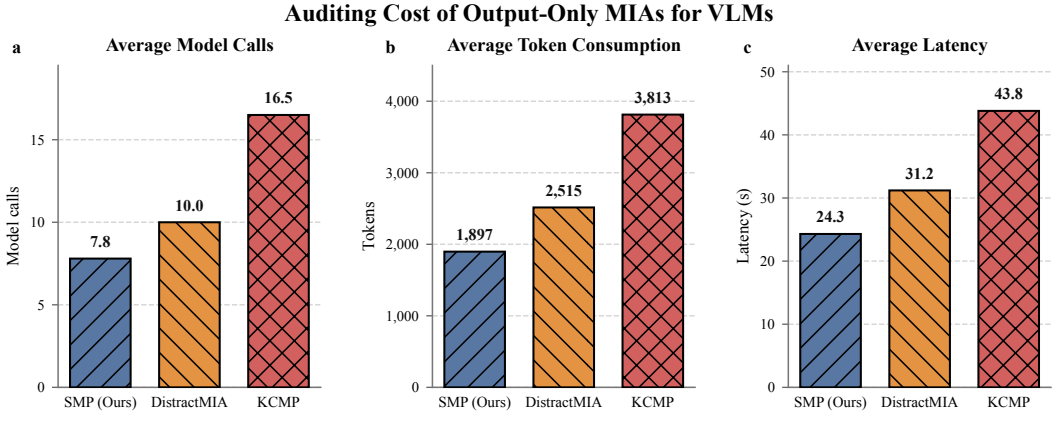


Figure 10: **Auditing Cost of Output-Only MIA on Commercial Black-Box VLMs (GPT-4o-Image, Gemini-2.0, and Gemini-2.5)** We compare SMP with DistractMIA and KCMP using three cost measures per membership decision: **a**, the mean number of model calls; **b**, the mean number of generated tokens; and **c**, the estimated latency in seconds. Overall, SMP consistently incurs the lowest auditing cost across all three measures, indicating greater efficiency in model usage, token consumption, and latency time.