Exposing Private Corpus Leakage in Multimodal RAG

Yihao Liu, Xinqi Lyu, Dong Wang, Bin Xiao*
Computing Department, The Hong Kong Polytechnic University
{yihao5.liu,xinqi.lyu,dong-comp.wang}@connect.polyu.hk,
b.xiao@polyu.edu.hk

Abstract

Multimodal Retrieval-Augmented Generation (RAG) helps mitigate hallucinations in Vision-Language Models (VLMs) by grounding generations in external knowledge bases. However, this externalized memory also introduces privacy risks, as external queries may reveal signals about sensitive visual records in the retrieval corpus, including medical images, scanned contracts, and proprietary business documents. To study this risk, we propose the Semantic Degradation Attack (SDA), a two-query method for exposing private corpus leakage in multimodal RAG by testing how strongly generated responses depend on retrieved evidence. SDA constructs transferable retrieval-disrupting perturbations using local surrogate visual encoders. By measuring the drop in the semantic similarity of the VLM's generated descriptions and the candidate caption after perturbation, SDA can distinguish whether a target image-caption record exists in the private retrieval corpus. Extensive experiments across five popular VLMs demonstrate that SDA outperforms existing baselines on two general-domain image-caption datasets and remains effective on two medical datasets.

1 Introduction

Retrieval-augmented generation (RAG) is increasingly used to ground large language models (LLMs) in external knowledge sources and reduce hallucinations. As vision-language models (VLMs) become more capable, this paradigm is being extended from text-only corpora to multimodal repositories containing images, PDFs, scanned documents, charts, and other visually rich records. In healthcare, multimodal RAG has been applied to medical vision-language tasks, with reported gains across radiology, ophthalmology, and pathology tasks [60, 53]. Commercial platforms are also beginning to support these capabilities: Amazon Nova documents multimodal RAG workflows over PDFs, images, and videos; NVIDIA provides enterprise blueprints for retrieving information from visually rich PDFs; and OpenAI has demonstrated image-and-text RAG pipelines for customer-experience analysis [2, 57, 43]. More broadly, enterprise RAG platforms such as Harvey show that retrieval-based assistants are already being used over private internal document repositories in high-sensitivity professional domains [19, 8].

However, these advances also raise broader security and privacy concerns [18, 39, 33, 38, 58, 28, 11, 10, 25]. As multimodal RAG moves into private deployments, privacy risks extend beyond model memorization to retrieval exposure. Unlike search interfaces that return references to retrieved documents, RAG systems can generate responses grounded in internal content. Without appropriate safeguards, this may expose information beyond the user's access privileges [6]. Even if the retrieved items are never shown, the response may still carry signals about whether a particular multimodal record exists in the private datastore. Motivated by this, we study a retrieval-corpus privacy question in multimodal RAG: *can an external user infer whether a specific image-text record is present in the private retrieval corpus?*

*Bin Xiao is the corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

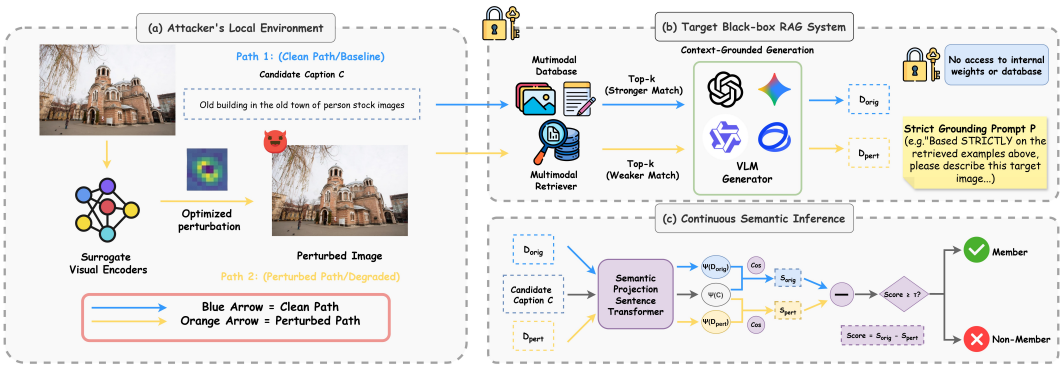


Figure 1: **Framework of SDA.** SDA first uses local surrogate visual encoders to generate a retrieval-disrupting perturbation for the target image. The original and perturbed images are then submitted to the black-box multimodal RAG system under the same context-grounded prompt, producing descriptions D_{orig} and D_{pert}. Finally, SDA compares their semantic alignment with the candidate caption C and uses the drop $S_{\text{orig}} - S_{\text{pert}}$ as the membership score.

Membership inference attacks (MIAs) provide a natural framework for this question. While traditional MIAs ask whether a sample was used to train a model, the target of inference in RAG is the external retrieval database that supports generation. Existing RAG-oriented MIAs have mainly focused on text-only systems, using generated responses, semantic similarity, difficulty calibration, or perturbation-based differential scoring to infer database membership [28, 33, 18]. Multimodal RAG introduces additional challenges. The target record is typically an image-text pair, but the observable output is only text, making the attack inherently cross-modal. Moreover, modern VLMs can generate semantically relevant descriptions from their prior knowledge even without exact retrieval support, which can mask whether the retrieved context actually contains the target record.

To address these challenges, we propose Semantic Degradation Attack (SDA), a two-query black-box membership inference attack for multimodal RAG. To the best of our knowledge, SDA is the first attack specifically designed to infer the membership of image-caption retrieval records in multimodal RAG. Rather than relying on the quality of a single response, SDA exploits the different sensitivity of member and non-member records to retrieval disruption. For an image-caption pair, SDA queries the target RAG system with the original image and with a perturbed counterpart under the same grounding instruction, then measures the semantic degradation of the generated description with respect to the candidate caption. Our key insight is that member samples are more likely to benefit from exact or near-exact retrieved evidence, and therefore exhibit a larger semantic drop when retrieval is disrupted; non-members, which already rely on approximate neighbors, tend to show a smaller change.

Our contributions are summarized as follows:

- We formulate black-box membership inference against multimodal RAG as a record-presence inference problem over a private image-text retrieval corpus, but the attacker observes only generated text.
- We propose Semantic Degradation Attack, a two-query attack that uses retrieval-disrupting perturbations and semantic response degradation rather than relying on absolute response quality or direct questioning.
- We introduce a transferable multi-model perturbation strategy based on local surrogate visual encoders and demonstrate SDA's effectiveness on general-domain and medical image-caption datasets across multiple VLMs.

2 Related Work

Multimodal Retrieval-Augmented Generation.

Traditional RAG systems [4, 50, 66] mainly operate over textual corpora, using dense [62] or sparse [46] retrievers to retrieve relevant passages for generation. Recent work extends this paradigm to multimodal settings, where the retrieval corpus may contain images, captions, and visually rich

documents [37, 34]. A common approach is to map images and text into a shared representation space, enabling cross-modal matching between visual queries and textual or image-based evidence [37]. These retrieval modules are increasingly combined with strong vision-language models, such as BLIP-2 [26], LLaVA [32], MiniGPT-4 [67], Flamingo [1], Gemini [54], and Qwen-VL [59], to support knowledge-intensive tasks such as visual question answering, image-grounded generation, and document understanding [13, 9, 65]. By retrieving multimodal evidence, these systems can ground generation in external visual and textual context [9, 65]. Commercial pipelines such as Amazon Nova [2] support multimodal retrieval by extracting visual content or converting images into textual descriptions for indexing, while systems including MuRAG [15], RA-CM3 [64], and REVEAL [22] use multimodal memories containing images and associated text. In such systems, the association between an image and its caption may itself contain sensitive information, such as the diagnostic description linked to a medical image [53] or the private metadata attached to a visual document [8].

Membership Inference Attacks (MIAs).

MIAs aim to determine whether a specific data sample was included in a model's training set, typically by exploiting overfitting behaviors [51, 21, 35]. While earlier research extensively covered traditional machine learning models across different knowledge settings [51, 41, 29], recent work extends MIAs to LLMs [49, 20] and VLMs, from grey-box attacks with logit access to more realistic black-box settings [29, 36]. These studies mainly target model training data.

RAG changes the object of inference. Because generation is grounded in an external retrieval corpus, membership inference can target whether a record is present in that corpus rather than whether it was used to train the model. Most existing RAG-oriented MIAs focus on text-based systems, where retrieval-corpus membership is inferred through direct prompting, shadow-model comparisons, semantic similarity, perplexity-based signals, text perturbations, or differential calibration [3, 42, 28, 33, 18]. Extending these attacks to multimodal RAG remains largely underexplored. The target record may involve visual content, while the attacker observes only textual responses. This modality difference makes text-output scores less reliable, since similar responses may arise from visual understanding or retrieved context. Recent work has begun to study this direction. MrM [63] attacks multimodal RAG through object-region masking, counterfactual mask selection, and repeated-query statistical testing. Our setting is different: we target image–caption record-pair membership in the private retrieval corpus under a two-query protocol. Rather than relying on repeated masked-image probes, SDA compares the responses to the original image and one retrieval-disrupted query, using the semantic drop against the candidate caption as the membership score.

3 Threat Model

Membership inference attacks were originally proposed to determine whether a target instance was used in training a machine learning model [51]. In a multimodal RAG setting, the central privacy concern extends beyond the generator's training data to the retrieval corpus that supports generation. Let R denote the target multimodal RAG system and let D denote its private retrieval corpus. Given a target image–caption pair $x = (I, C)$, where I is the target image and C is its associated caption, the adversary has background knowledge K, and seeks to infer whether x is contained in D. Formally, the membership label is defined as

$$m(x; D) = \mathbb{I}[x \in D], \tag{1}$$

and the adversary aims to approximate this label using a binary decision rule

$$A_K^R(x) \in \{0, 1\}, \tag{2}$$

where $A_K^R(x) = 1$ indicates the prediction that $x \in D$, and $A_K^R(x) = 0$ otherwise.

Adversarial Capabilities We consider a black-box threat model. The adversary can interact with the target system R only through its standard query interface and can observe only the final generated textual responses. No access is granted to model parameters, token-level logits or probabilities, the target retriever, retrieved items, retrieval scores, or any metadata associated with the underlying corpus. The adversary is allowed to issue a small number of queries for the same target instance.

Evaluation Metrics We evaluate attack performance with two standard metrics: the area under the receiver operating characteristic curve (ROC-AUC), and TPR@1%FPR. ROC-AUC summarizes attack performance across all decision thresholds. TPR@1%FPR measures attack effectiveness in a low-false-positive regime and has been widely adopted in prior membership inference studies [12, 24].

4 Method

In this section, we propose SDA, a two-query MIA for multimodal RAG that exploits the different sensitivity of member and non-member queries to retrieval-disrupting perturbations. The attack proceeds in three stages. First, we construct a perturbed image $\tilde{I}$ that is displaced from the original image I across multiple local surrogate visual embedding spaces, aiming to disrupt retrieval by changing similarity relationships in the target retriever. Second, we submit both I and $\tilde{I}$ to the target RAG system under the same context-grounded instruction and obtain two retrieval-conditioned descriptions. Third, we measure how much the semantic consistency between the generated description and the candidate caption C drops after perturbation, and use this degradation as the membership score. Intuitively, if x is a member, the unperturbed query is more likely to benefit from an exact or near-exact retrieved match; if x is a non-member, both the original and perturbed queries rely primarily on approximate neighbors, leading to a smaller change in the final output.

4.1 Transferable Multi-model Self-Distance Perturbation

Our optimized perturbation follows the standard projected-gradient formulation for bounded adversarial perturbations [40, 27], but is adapted to the black-box retriever-mismatch setting. Since the target retriever is unknown, perturbations optimized on a single local encoder may overfit to that surrogate and transfer poorly. SDA therefore optimizes one shared perturbation over multiple local surrogate CLIP encoders.

Let $\{E_m\}_{m=1}^{M}$ denote the local surrogate CLIP image encoders. Given an input image I, we constrain the perturbed image $\tilde{I}$ by

$$\|\tilde{I} - I\|_\infty \leq \epsilon, \tag{3}$$

where ϵ is the perturbation budget. We initialize $I_0 = I$ and optimize for $t = 0, \ldots, T-1$, setting the final perturbed image as $\tilde{I} = I_T$ after T steps. For each encoder, we compute the clean normalized image feature

$$z_m = \frac{E_m(I)}{\|E_m(I)\|_2}. \tag{4}$$

At each step, we maximize the weighted average feature displacement:

$$\max_{I_t} \quad \mathcal{L}(I_t), \qquad \mathcal{L}(I_t) = \sum_{m=1}^{M} w_m \ell_m(I_t). \tag{5}$$

where w_m is the weight of surrogate encoder m, with $w_m \geq 0$ and $\sum_{m=1}^{M} w_m = 1$. The per-encoder loss is

$$\ell_m(I_t) = \mathbb{E}_{\tau \sim \mathcal{T}} \left[1 - \cos\left(\frac{E_m(\tau(I_t))}{\|E_m(\tau(I_t))\|_2}, z_m \right) \right], \tag{6}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity, $\mathcal{T}$ denotes a distribution of input transformations, and τ is a sampled transformation. In practice, τ is sampled from a random resize–pad distribution, following the input-diversity strategy for improving transferability [61]. The expectation over transformations is approximated with a small number of samples, following EOT [5]. For each surrogate encoder, we compute and normalize the gradient:

$$g_m = \frac{\nabla_{I_t} \ell_m(I_t)}{\mathrm{mean}(|\nabla_{I_t} \ell_m(I_t)|) + \eta}, \tag{7}$$

where $\mathrm{mean}(\cdot)$ averages over all image dimensions and η is a small constant for numerical stability. We then aggregate gradients across surrogate encoders:

$$g_t = \sum_{m=1}^{M} w_m g_m. \tag{8}$$

This multi-model design follows the transfer-attack intuition that perturbations optimized across multiple models are less dependent on a single surrogate embedding space [55, 17]. To stabilize the update direction, we use momentum [17]:

$$v_{t+1} = \mu v_t + \frac{g_t}{\text{mean}(|g_t|) + \eta}, \qquad v_0 = 0, \tag{9}$$

where μ is the momentum coefficient. The perturbed image is updated by projected gradient ascent:

$$I_{t+1} = \Pi_{[0,1]\cap\mathcal{B}_\infty(I,\epsilon)} \left(I_t + \alpha \cdot \text{sign}(v_{t+1})\right), \tag{10}$$

where α is the step size, $\mathcal{B}_\infty(I, \epsilon)$ is the ℓ_∞ ball centered at I with radius ϵ, and Π projects the image back to the valid pixel range and this perturbation ball. The final perturbed image $\tilde{I}$ is then used only for querying the target multimodal RAG system.

4.2 Context-Grounded Description Generation

SDA uses the change in retrieval-grounded descriptions as its attack signal. This signal becomes weaker if the model fills in details from prior knowledge rather than the retrieved context. We therefore use the same context-grounded prompt for both queries, asking the VLMs to rely only on the retrieved examples and avoid unsupported details. The prompt is:

> **Instruction.** Based STRICTLY on the retrieved examples above, please describe this target image. Focus only on details that are clearly grounded in the examples. If the retrieved examples do not perfectly match the target image, reflect that uncertainty in your description (e.g., say "appears to be" or "possibly"). Do NOT use your general knowledge to fill in details not supported by the examples.

Let $\mathcal{P}$ denote this grounding prompt, and let $\mathcal{F}_\mathcal{P}(\cdot; \mathcal{D})$ denote the multimodal RAG when queried under $\mathcal{P}$. The descriptions generated from the original and perturbed images are

$$D_{\text{orig}} = \mathcal{F}_\mathcal{P}(I; \mathcal{D}), \qquad D_{\text{pert}} = \mathcal{F}_\mathcal{P}(\tilde{I}; \mathcal{D}). \tag{11}$$

Equivalently, let $\text{Ret}(\cdot, \mathcal{D})$ denote the target retriever that returns retrieved context from the private corpus $\mathcal{D}$, and let $\mathcal{G}$ denote the generator. Then

$$D_{\text{orig}} = \mathcal{G}\big(I, \text{Ret}(I, \mathcal{D}), \mathcal{P}\big), \qquad D_{\text{pert}} = \mathcal{G}\big(\tilde{I}, \text{Ret}(\tilde{I}, \mathcal{D}), \mathcal{P}\big). \tag{12}$$

The grounding prompt encourages the VLM to reflect changes in retrieved evidence, making the resulting description differences more informative for membership inference.

4.3 Continuous Semantic Degradation Score

Rather than judging membership from a single response, SDA measures how the response changes after retrieval disruption. Let $\Psi(\cdot)$ denote a sentence encoder that maps text into a dense semantic space. We compute the semantic alignment between each generated description and the candidate caption C as

$$S_{\text{orig}} = \cos(\Psi(D_{\text{orig}}), \Psi(C)), \qquad S_{\text{pert}} = \cos(\Psi(D_{\text{pert}}), \Psi(C)). \tag{13}$$

We then define the membership inference score as

$$\text{Score}(x) = S_{\text{orig}} - S_{\text{pert}}. \tag{14}$$

A large positive score suggests that the original query benefited more from retrieval than the perturbed query, consistent with the behavior expected for member samples. The final membership prediction is given by

$$\hat{m}(x) = \begin{cases} 1, & \text{if Score}(x) \geq \tau, \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$

where τ is a decision threshold.

5 Experiments

5.1 Experimental Setup

Datasets and retrieval corpora. Our main experiments use two image-caption datasets with different visual and caption distributions: LAION_CC_SBU and MS COCO (COCO) [30]. LAION_CC_SBU is a mixed image-caption corpus constructed from LAION [48], Conceptual Captions 3M*, Conceptual 12M [14], and SBU Captions [44]. For each dataset, we construct a private multimodal retrieval corpus $\mathcal{D}$ containing the 500 members and 5,000 additional disjoint distractor records, where each corpus record consists of an image and its associated ground-truth caption. Additionally, 500 non-member samples are drawn from a disjoint split. We further assess cross-split redundancy using caption embedding similarity, CLIP image similarity, and exact caption matching. After removing highly similar samples across member and non-member splits, SDA's performance remains largely unchanged, with AUC variations within 0.0038 across all five VLMs. Detailed results are provided in Appendix A.4. Additional evaluations on two medical image-caption datasets, ROCOv2 [47] and PMC-OA [31], are reported in Appendix A.7.

Target multimodal RAG system. We instantiate a black-box multimodal RAG pipeline consisting of a retriever and a vision-language generator. The retriever encodes query images and corpus images using CLIP and performs nearest-neighbor search to retrieve the top-k image-caption records from the private corpus. By default, we set $k = 3$. The retrieved captions and images are then provided as context to the target VLM, which generates a textual description of the query image under the same context-grounded instruction for both original and perturbed queries. To evaluate whether the attack generalizes across different generators, we test five target VLMs: GPT-5 [52], GPT-4o [23], Gemini 3.1†, GLM-4.6V [56], and Qwen3-VL [7]. All target VLMs are accessed through API calls, with no access to model internals or token-level probabilities. We use the same retrieval corpus and retriever for all target VLMs, allowing us to compare their attack performance under a fixed retrieval configuration.

Attack implementation. For SDA, the attacker constructs a retrieval-disrupting perturbation using only local surrogate encoders. In the main setting, we optimize one shared perturbation over three local models: CLIP ViT-B/32, CLIP RN50, and OpenCLIP ViT-H-14 [45, 16]. The target retriever uses CLIP ViT-L-14 [45]. This creates an explicit retriever mismatch between the local optimization space and the target RAG system. The perturbation is bounded by an ℓ_∞ budget of $\epsilon = 8/255$, optimized for $T = 10$ steps, with step size $\alpha = \epsilon/T$. Each target sample is queried twice, once with the original image I and once with the perturbed image $\tilde{I}$, while keeping the textual instruction fixed. For each query, the target system retrieves the top-3 examples, and the maximum generation length is set to 150 tokens. We compute text semantic similarity using all-mpnet-base-v2. We generate the perturbations using eight NVIDIA 3090 GPUs.

Baselines. We compare SDA with six representative black-box MIA baselines adapted to our image-text retrieval corpus setting. All baselines use the same target multimodal RAG interface and observe only the final generated responses. DCMI [18] replaces five key adjectives or adverbs in the candidate caption with their antonyms and measures the response difference between the original and modified captions. MBA [33] randomly masks 50% of the target image patches and measures the CLIP similarity between the generated description and the original image. QA [28] directly asks whether the target image appears in the retrieved references and uses the binary response for membership inference. S2MIA [3] measures the semantic similarity between the generated description and the candidate caption. IA [42] generates target-specific questions and infers membership from the correctness of the generated answers. MrM [63] masks selected object regions and infers membership by evaluating whether the missing visual content can be recovered from retrieved evidence. Following their experimental settings, IA and MrM require approximately 30 and 5 queries, respectively.

*https://ai.google.com/research/ConceptualCaptions/

†https://deepmind.google/models/gemini/pro/

5.2 Main Results

Tables 1 and 2 present the main results on LAION_CC_SBU and COCO. SDA achieves the highest AUC across all five target VLMs on both datasets and the best TPR@1%FPR in nine of the ten evaluated settings. On LAION_CC_SBU, SDA achieves an average AUC of 0.8364, outperforming the strongest baseline, MrM (0.7191). It also improves the average TPR@1%FPR from 26.90% (IA) to 45.13%. On COCO, SDA achieves an average AUC of 0.8214, compared with 0.7263 for MrM, and increases the average TPR@1%FPR from 28.62% (S2MIA) to 48.25%.

Table 1: Main attack performance on LAION_CC_SBU. We report AUC and TPR@1%FPR across five target VLMs.

Metric / Model / Method	GPT-5		GPT-4o		Gemini 3.1		Qwen3-VL		GLM-4.6V	
	AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR
DCMI	0.6212	-	0.6211	-	0.6239	-	0.6405	-	0.6080	-
MBA	0.5974	8.15	0.6063	4.76	0.5844	5.33	0.6105	13.66	0.5919	10.56
QA	0.6734	25.23	0.6567	12.54	0.7348	36.97	0.5977	15.66	0.6109	21.54
S2MIA	0.7365	27.75	0.7581	31.56	0.7023	25.31	0.6954	18.97	0.6799	21.88
IA	0.7164	32.34	0.7133	30.58	0.7023	19.15	0.6895	23.46	0.7357	**28.99**
MrM	0.7471	35.64	0.7033	26.49	0.6982	16.92	0.7175	19.47	0.7296	18.33
SDA	**0.8688**	**55.44**	**0.8543**	**50.40**	**0.8107**	**53.60**	**0.8126**	**37.80**	**0.8355**	28.40

Bold values indicate the best result under each model and metric, while underlined values indicate the second-best result. The SDA row is shaded in gray.

Table 2: Main attack performance on COCO. We report AUC and TPR@1%FPR across five target VLMs.

Metric / Model / Method	GPT-5		GPT-4o		Gemini 3.1		Qwen3-VL		GLM-4.6V	
	AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR
DCMI	0.5887	-	0.5890	-	0.5884	-	0.5967	-	0.5740	-
MBA	0.5932	8.66	0.6125	9.33	0.6413	10.37	0.7295	14.66	0.6266	9.33
QA	0.6973	18.68	0.6488	20.14	0.7283	37.25	0.6311	9.79	0.5869	8.51
S2MIA	0.7542	33.10	0.7268	29.16	0.7355	20.64	0.7160	31.77	0.6785	28.43
IA	0.7396	36.00	0.7258	34.75	0.6631	24.16	0.7014	20.52	0.6961	14.99
MrM	0.7453	23.36	0.7392	26.74	0.6865	15.02	0.7459	24.87	0.7146	18.33
SDA	**0.8642**	**56.24**	**0.8168**	**45.20**	**0.8256**	**57.80**	**0.8215**	**39.20**	**0.7792**	**42.80**

SDA is particularly effective at TPR@1%FPR, where the false-positive rate is tightly constrained. For example, on LAION_CC_SBU with GPT-5, SDA reaches 55.44% TPR@1%FPR, while the strongest baseline achieves only 35.64%. On COCO with GPT-5, SDA improves TPR@1%FPR from 36.00% to 56.24%. This indicates that SDA does not merely improve overall ranking quality, but also identifies more members at a low false-positive rate.

S2MIA provides a direct comparison with single-response semantic similarity, as it scores membership using the alignment between a generated description and the target caption. IA evaluates the correctness of answers to target-specific questions, whereas MrM tests whether retrieved evidence can recover masked visual content. SDA achieves higher AUC than them across all evaluated VLMs and datasets, demonstrating the effectiveness of measuring response changes after retrieval disruption. QA performs competitively in some cases, especially on Gemini 3.1, but its performance varies largely across models and datasets, making direct membership-style questioning less stable. DCMI and MBA are less competitive, suggesting that their adapted text-perturbation and visual-masking signals are less effective in this setting.

Overall, SDA maintains its advantage across both datasets, particularly at TPR@1%FPR, where false-positive predictions are tightly constrained.

5.3 Ablation Study

We conduct a series of ablation studies to examine which components are responsible for the effectiveness of SDA. Unless otherwise specified, ablations use GPT-5 on LAION_CC_SBU as the default setting, with a context-grounded prompt, retrieval top-$k = 3$, maximum generation length of 150 tokens, 10 perturbation optimization steps, perturbation budget $\epsilon = 8/255$, and adversarial perturbation.

Effect of prompt design. Table 3 compares SDA under different prompt formulations. Replacing the context-grounded prompt with a general prompt reduces TPR@1%FPR from 55.44% to 14.72%, indicating that retrieval grounding plays an important role in low-FPR detection. In contrast, a simple retrieval-focused instruction achieves 0.8297 AUC and 52.14% TPR@1%FPR, while two grounded paraphrases yield results close to the original prompt. These findings suggest that SDA is more sensitive to the presence of retrieval-grounding instructions than to their specific wording. Results across the remaining VLMs are provided in Appendix A.6.

Table 3: Effect of prompt design on SDA performance using GPT-5 on LAION_CC_SBU.

Prompt Type	AUC	TPR@1%FPR
General Prompt	0.7936	14.72
Simple Retrieval-Focused	0.8297	52.14
Grounded Paraphrase 1	0.8569	54.38
Grounded Paraphrase 2	0.8601	55.16
Context-Grounded (Original)	**0.8688**	**55.44**

Effect of knowledge base size. We evaluate SDA with knowledge bases ranging from 5K to 100K records while keeping 500 member records fixed. As shown in Table 4, the AUC remains relatively stable, decreasing only from 0.8688 to 0.8621 as the knowledge base grows. In contrast, TPR@1%FPR decreases from 55.44% to 37.90%, indicating that larger knowledge bases have a greater impact on membership detection at low false-positive rates.

Table 4: Effect of knowledge base size on SDA performance.

Knowledge base size	AUC	TPR@1%FPR
5K	0.8688	55.44
10K	0.8651	52.21
50K	0.8634	44.55
100K	0.8621	37.90

Retrieval neighborhood analysis. We analyze retrieval neighborhoods before and after perturbation to assess whether SDA induces different retrieval changes for members and non-members. At $k = 3$, members exhibit a larger retrieval similarity drop than non-members (0.1634 vs. 0.1326). Notably, members retain higher top-1 neighbor overlap (0.8145 vs. 0.3307), despite experiencing a larger similarity drop (0.2237 vs. 0.1396). This indicates that the membership-dependent retrieval changes are not simply explained by neighbor replacement, but also by the weakening of similarity to retrieved evidence. Full results are provided in Appendix A.8.

Retrieval- and generator-level effects. The retrieval-only setting outperforms the generator-only setting (0.8100 vs. 0.7636 AUC), while disabling RAG reduces the AUC to 0.5392. These results support the importance of retrieval disruption in SDA. Detailed results are provided in Appendix A.5.

Effect of target retriever. We further evaluate SDA using two non-CLIP target retrievers, GME-Qwen2-VL-Instruct[‡] and Llama-Nemotron-Embed-VL[§]. As shown in Table 5, SDA achieves average AUCs of 0.8099 and 0.7293, respectively, across five VLMs. Although performance decreases compared with the original CLIP-based retriever (0.8364 AUC), SDA remains effective under both alternative retrieval architectures. These results indicate that SDA is not restricted to CLIP-based retrieval.

Adversarial perturbation versus random noise. Figure 2 compares the optimized perturbation with non-adversarial pixel-space perturbations under the same ℓ_∞ budget ϵ. The *None* setting uses the original image. *Uniform* samples each pixel perturbation independently from $[-\epsilon, \epsilon]$. *Gaussian*

[‡]https://huggingface.co/Alibaba-NLP/gme-Qwen2-VL-2B-Instruct/

[§]https://huggingface.co/nvidia/llama-nemotron-embed-vl-1b-v2

Table 5: Effect of target retrievers on SDA performance on LAION_CC_SBU.

	CLIP ViT-L/14		GME-Qwen2-VL-Instruct		Llama-Nemotron-Embed-VL	
Model	**AUC**	**TPR@1%FPR**	**AUC**	**TPR@1%FPR**	**AUC**	**TPR@1%FPR**
GPT-5	0.8688	55.44	0.8281	38.59	0.7684	33.27
GPT-4o	0.8543	50.40	0.8123	34.15	0.7342	31.29
Gemini 3.1	0.8107	53.60	0.7956	31.64	0.7016	26.51
Qwen3-VL	0.8126	37.80	0.8042	32.77	0.7123	28.34
GLM-4.6V	0.8355	28.40	0.8094	25.33	0.7299	30.57
Average	**0.8364**	**45.13**	**0.8099**	**32.50**	**0.7293**	**30.00**

samples zero-mean Gaussian noise with standard deviation $\epsilon/2$ by default and clips each pixel perturbation to $[-\epsilon, \epsilon]$. *Random-sign* applies a perturbation with magnitude exactly ϵ at each pixel, with the sign sampled uniformly from $\{-1, 1\}$. All perturbed images are clipped to the valid pixel range. These baselines alter the image in pixel space but do not use gradients or surrogate retrievers. They perform close to random, with AUCs around 0.46 to 0.51 and TPR@1%FPR below 2.5%. In contrast, the optimized perturbation reaches 0.8688 AUC and 55.44% TPR@1%FPR. These results suggest that simply adding pixel-space noise is insufficient, while optimizing perturbations for retrieval disruption produces a stronger membership signal.

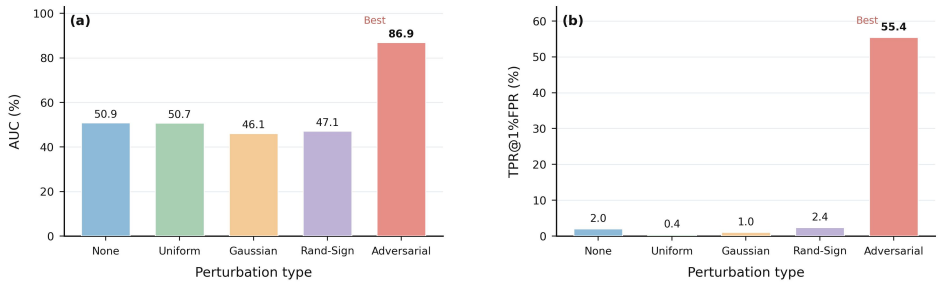


Figure 2: **Effect of perturbation type on SDA performance**.

Effect of perturbation optimization steps. Figure 3 evaluates the number of optimization steps used to generate the adversarial perturbation. With only one step, the attack is weak, achieving 0.5612 AUC and 6.25% TPR@1%FPR. Increasing the number of steps steadily improves performance, with 5 steps already reaching 0.7461 AUC and 30.04% TPR@1%FPR. The best result is obtained at 10 steps, where SDA reaches 0.8688 AUC and 55.44% TPR@1%FPR. Increasing the number of steps further to 20 does not improve the attack and reduces TPR@1%FPR to 43.75%. This suggests that moderate iterative optimization is sufficient to disrupt retrieval, while excessive optimization may move the perturbed query too far from the useful neighborhood and weaken the paired degradation effect.

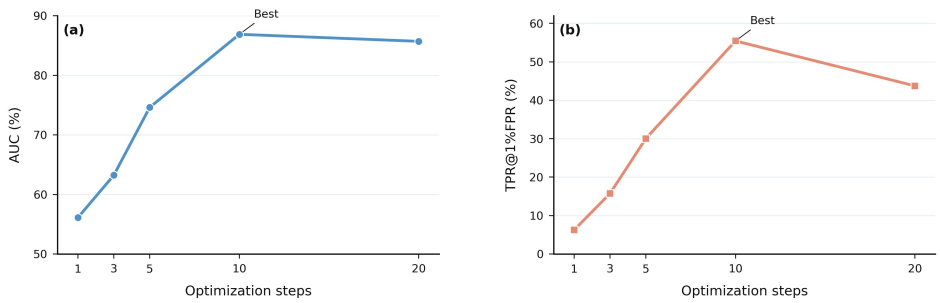


Figure 3: **Effect of the number of perturbation optimization steps on SDA performance.**

Effect of the number of retrieved examples. Figure 4 studies the effect of retrieval top-k. The attack achieves the strongest performance when only the nearest retrieved example is used, reaching 0.9372 AUC and 59.68% TPR@1%FPR. As k increases, performance gradually decreases. This decline may arise because larger values of k introduce less relevant retrieved examples, reducing the influence of the nearest match on the generated description. Importantly, SDA remains effective even when k increases to 10, with 0.8303 AUC and 52.91% TPR@1%FPR. We use $k = 3$ in the main experiments because it better reflects a practical multi-context RAG setting than top-1 retrieval.

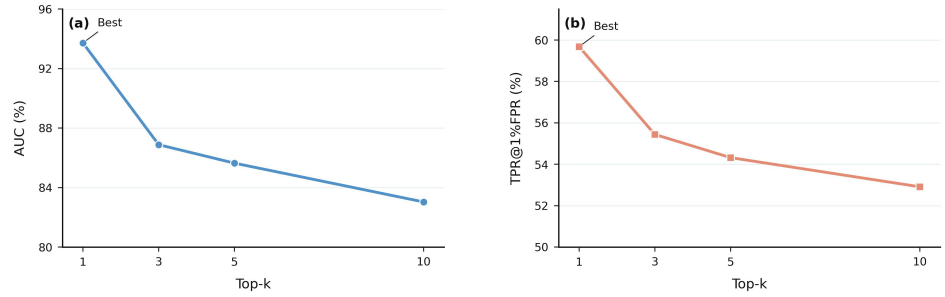


Figure 4: **Effect of retrieval top-k on SDA performance.**

Effect of defense. We evaluate a system-prompt defense adapted from MrM. Across five VLMs, the average AUC changes from 0.8364 to 0.8384, while TPR@1%FPR changes from 45.13% to 45.05%. These results indicate that the evaluated defense does not consistently reduce SDA's effectiveness. Detailed results are provided in Appendix A.10.

Effect of surrogate encoders under retriever mismatch. We compare the surrogate encoders used to optimize the perturbation. The target retriever is fixed to CLIP ViT-L/14, which is different from all attacker-side encoders. We evaluate perturbations optimized with a single local encoder and the multi-model setting, where the perturbation is optimized jointly over CLIP ViT-B/32, CLIP RN50, and OpenCLIP ViT-H-14. As shown in Table 6, single-surrogate perturbations transfer less effectively to the unseen target retriever. Joint optimization across multiple surrogate encoders achieves the best performance, suggesting that combining surrogate representations improves transferability to the unseen target retriever.

Table 6: Effect of surrogate encoders under retriever mismatch. The target retriever is CLIP ViT-L/14.

Attacker-side surrogate	Target retriever	AUC	TPR@1%FPR
CLIP RN50	CLIP ViT-L/14	0.7320	20.80
CLIP ViT-B/32	CLIP ViT-L/14	0.7845	26.60
OpenCLIP ViT-H-14	CLIP ViT-L/14	0.8019	30.40
Multi-model	CLIP ViT-L/14	**0.8688**	**55.44**

6 Conclusion

This paper studies retrieval-corpus membership inference in multimodal RAG under black-box access. We show that the presence of a private image-caption record can be inferred from response changes caused by retrieval-disrupting image perturbations. Based on this observation, we propose SDA, a two-query attack that turns this response change into a membership score. To transfer to an unknown target retriever, SDA optimizes one shared perturbation over multiple local surrogate visual encoders. Comprehensive evaluations further support semantic degradation as a reliable indicator of private corpus membership. The effectiveness of SDA highlights the need to protect the retrieval layer itself, not only the final generated output.

Acknowledgement

This work was supported in part by the Hong Kong Research Grants Council's (RGC) General Research Fund (GRF) under Grant PolyU 15230025.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Amazon Web Services. Using amazon nova for multimodal rag. `https://docs.aws.amazon.com/nova/latest/userguide/rag-multimodal.html`, 2026. Accessed: March 2026.

[3] Maya Anderson, Guy Amit, and Abigail Goldsteen. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. In *International Conference on Information Systems Security and Privacy*, 2025.

[4] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2023.

[5] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

[6] AWS. Authorizing access to data with rag implementations. `https://aws.amazon.com/cn/blogs/security/authorizing-access-to-data-with-rag-implementations/`, September 2025.

[7] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[8] Hanoz Bhathena, Parin Rajesh Jhaveri, Rohan Mittal, Prateek Singh, Aymen Kallala, Rachneet Kaur, Yiqiao Jin, Zhen Zeng, Adwait Ratnaparkhi, and Denis Kochedykov. Mm-bizrag: Rethinking multimodal retrieval-augmented generation for general purpose enterprise q&a. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, pages 1967–2000, 2026.

[9] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826, 2024.

[10] Yiming Cao, Yanjie Li, Kaisheng Liang, and Bin Xiao. Enhancing targeted adversarial attacks on large vision-language models via intermediate projector. *IEEE Transactions on Information Forensics and Security*, 21:5946–5959, 2026. doi: 10.1109/TIFS.2026.3702560.

[11] Yiming Cao, Dong Wang, Xinqi Lyu, and Bin Xiao. Pureproof: Diffusion-resistant black-box targeted attack on large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8621–8630, 2026.

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

[13] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504, 2022.

[14] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.

[15] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, 2022.

[16] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829, 2023.

[17] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[18] Xinyu Gao, Xiangtao Meng, Yingkai Dong, Zheng Li, and Shanqing Guo. Dcmi: A differential calibration membership inference attack against retrieval-augmented generation. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pages 4184–4198, 2025.

[19] Harvey Team. Enterprise-grade rag systems. `https://www.harvey.ai/blog/enterprise-grade-rag-systems`, February 2025.

[20] 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.

[21] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

[22] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23369–23379. IEEE, 2023.

[23] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[24] 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*, pages 4871–4881, 2023.

[25] Yuni Lai, Xinqi Lyu, Dong Wang, Yihao Liu, Kai Zhou, and Bin Xiao. A survey on poisoning attacks, defenses, and provable defense in the era of llms. 2026.

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[27] Yanjie Li, Wenxuan Zhang, Xinqi Lyu, Yihao Liu, and Bin Xiao. Styleguard: Preventing text-to-image-model-based style mimicry attacks by style perturbations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[28] Yuying Li, Gaoyang Liu, Chen Wang, and Yang Yang. Generating is believing: Membership inference attacks against retrieval-augmented generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[29] 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.

[30] 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*, pages 740–755. Springer, 2014.

[31] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *international conference on medical image computing and computer-assisted intervention*, pages 525–536. Springer, 2023.

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

[33] Mingrui Liu, Sixiao Zhang, and Cheng Long. Mask-based membership inference attacks for retrieval-augmented generation. In *Proceedings of the ACM on Web Conference 2025*, pages 2894–2907, 2025.

[34] Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma. Hm-rag: Hierarchical multi-agent multimodal retrieval augmented generation. In *Proceedings of the 33rd ACM international conference on multimedia*, pages 2781–2790, 2025.

[35] Yihao Liu, Jinhe Huang, Yanjie Li, Dong Wang, and Bin Xiao. Generative ai model privacy: a survey. *Artificial Intelligence Review*, 58(1):33, 2024.

[36] 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.

[37] Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In *The Eleventh International Conference on Learning Representations*, 2023.

[38] Xinqi Lyu, Yihao Liu, Yanjie Li, and Bin Xiao. Pla: Prompt learning attack against text-to-image generative models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16851–16860. IEEE, 2025.

[39] Xinqi Lyu, Yihao Liu, Dong Wang, and Bin Xiao. Red-teaming retrieval-augmented diffusion models via poisoning knowledge bases. In *Conference on Computer Vision and Pattern Recognition 2026*, 2026. URL `https://openreview.net/forum?id=d0UGGyrWfz`.

[40] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

[41] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 691–706. IEEE, 2019.

[42] Ali Naseh, Yuefeng Peng, Anshuman Suri, Harsh Chaudhari, Alina Oprea, and Amir Houmansadr. Riddle me this! stealthy membership inference for retrieval-augmented generation. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pages 1245–1259, 2025.

[43] OpenAI. Image understanding with rag. `https://developers.openai.com/cookbook/examples/multimodal/image_understanding_with_rag`, May 2025.

[44] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[45] 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*, pages 8748–8763. PmLR, 2021.

[46] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc, 2009.

[47] Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, et al. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1):688, 2024.

[48] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[49] 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*, 2023.

[50] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, 2024.

[51] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE (SP)*, pages 3–18, 2017.

[52] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[53] Liwen Sun, James Jialun Zhao, Wenjing Han, and Chenyan Xiong. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 643–655, 2025.

[54] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[55] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rkZvSe-RZ`.

[56] V Team, Wenyi Hong, Wenmeng Yu, et al. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL `https://arxiv.org/abs/2507.01006`.

[57] Tanay Varshney, Annie Surla, Nicola Sessions, and Sean Sodha. Build an enterprise-scale multimodal pdf data extraction pipeline with an nvidia ai blueprint. `https://developer.nvidia.com/blog/build-an-enterprise-scale-multimodal-document-retrieval-pipeline-with-nvidia-nim-agent-blueprint/`, August 2024.

[58] Dong Wang, Xiangyu He, Xinqi Lyu, and Bin Xiao. Breaking multimodal llm safety via video-driven prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8566–8576, 2026.

[59] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[60] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. MMed-RAG: Versatile multimodal RAG system for medical vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=s5epFPdIW6`.

[61] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019.

[62] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021.

[63] Peiru Yang, Jinhua Yin, Haoran Zheng, Xueying Bai, Huili Wang, Yufei Sun, Xintian Li, Songwei Pei, Yongfeng Huang, and Tao Qi. Mrm: Black-box membership inference attacks against multimodal rag systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 34295–34303, 2026.

[64] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. Retrieval-augmented multimodal language modeling. In *Proceedings of the 40th International Conference on Machine Learning*, pages 39755–39769, 2023.

[65] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations*, 2025.

[66] Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. Augmentation-adapted retriever improves generalization of language models as generic plug-in. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2421–2436, 2023.

[67] 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.

A Technical appendices and supplementary material

A.1 Why Semantic Degradation Reveals Membership

We provide a simple explanation for why semantic degradation reveals retrieval-corpus membership. For clarity, consider the nearest-neighbor case; the same intuition applies when the retriever returns top-k results. Let each corpus record be $r_j = (I_j, C_j) \in \mathcal{D}$, and let $E_{\text{ret}}(\cdot)$ denote the target retriever's visual encoder. We define the retrieval distance from a query image I to the corpus as

$$d_{\text{ret}}(I, \mathcal{D}) = \min_{(I_j, C_j) \in \mathcal{D}} \left(1 - \cos\big(E_{\text{ret}}(I), E_{\text{ret}}(I_j)\big)\right). \tag{16}$$

Although the attacker does not know E_{ret}, this distance describes the retrieval behavior of the target RAG system. Our attack relies on the assumption that better retrieval leads to descriptions that are more semantically aligned with the target caption. Let S denote the semantic similarity between the generated description and the candidate caption C. We model the expected similarity as

$$\mathbb{E}[S \mid d_{\text{ret}}] = f(d_{\text{ret}}), \tag{17}$$

where $f(\cdot)$ is a non-increasing function of the retrieval distance.

Under the member hypothesis H_1, the target record exists in $\mathcal{D}$, so the original image I can retrieve highly relevant evidence and the generated description is expected to align well with C. After perturbation, $\tilde{I}$ is moved away from the original retrieval neighborhood, which can reduce the relevance of the retrieved context and lower the semantic alignment. Under the non-member hypothesis H_0, the original query already relies on approximate neighbors. Perturbing the image is therefore more likely to replace one approximate neighborhood with another, leading to a smaller change in the generated description.

This difference between member and non-member behavior motivates the semantic degradation score

$$\text{Score}(x) = S_{\text{orig}} - S_{\text{pert}}, \tag{18}$$

and suggests that

$$\mathbb{E}[\text{Score}(x) \mid H_1] > \mathbb{E}[\text{Score}(x) \mid H_0]. \tag{19}$$

A.2 Effect of perturbation budget

Figure 5 studies the effect of the perturbation budget ϵ. With no or very small perturbations, the attack is weak: $\epsilon = 0/255$, $2/255$, and $4/255$ all give low AUC and TPR@1%FPR. Performance increases sharply at $\epsilon = 8/255$, reaching 0.8688 AUC and 55.44% TPR@1%FPR. Increasing the budget further to $16/255$ lowers performance, especially at TPR@1%FPR. This suggests that the perturbation must be large enough to weaken retrieval similarity, but not so large that it undermines the comparison between the original and perturbed responses. We therefore use $\epsilon = 8/255$ as the default budget.

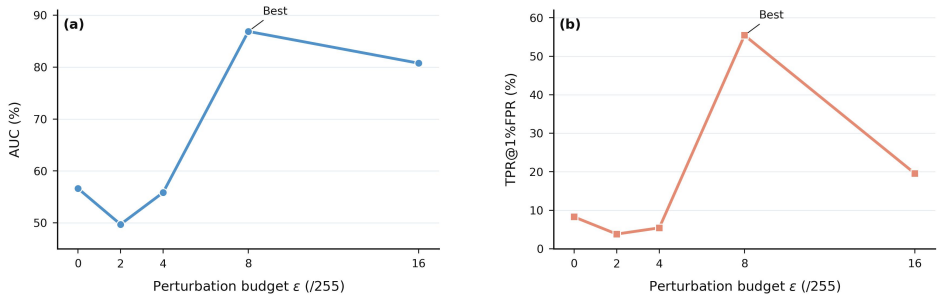


Figure 5: **Effect of perturbation budget ϵ on SDA performance.**

A.3 Visualization of SDA Perturbations

To better illustrate the visual effect of SDA, we provide examples of original images and their perturbed versions generated by the multi-model surrogate optimization used in our main experiments.



Figure 6: **Visualization of SDA perturbations.** Each row shows an original image and the corresponding image perturbed by SDA.

All perturbations are constrained by the same ℓ_∞ budget of $\epsilon = 8/255$ and optimized for $T = 10$ steps. As shown in Figure 6, the perturbed images remain visually close to the originals while still altering the visual features used for retrieval.

A.4 Effect of Data Redundancy

We assess cross-split redundancy using caption embedding similarity, CLIP image embedding similarity, and exact normalized caption matching. As shown in Table 7, no exact caption duplicates or caption pairs with embedding similarity ≥ 0.90 are found between member and non-member splits. Only two cross-split image pairs exceed each of the CLIP similarity thresholds of 0.95 and 0.98.

Table 7: Cross-split redundancy statistics on LAION_CC_SBU.

Exact Caption Duplicates	**Caption Similarity** ≥ 0.90	**Image Similarity** ≥ 0.95	**Image Similarity** ≥ 0.98
0	0	2	2

To evaluate the effect of near-duplicate samples, we construct a low-redundancy subset by removing samples whose nearest opposite-split neighbors have caption embedding similarity ≥ 0.90, CLIP image similarity ≥ 0.95, or identical normalized captions. The filtering is independent of attack scores, and all attack configurations remain unchanged.

Table 8 compares the original and low-redundancy results. Across the five VLMs, AUC changes by at most 0.0038, while TPR@1%FPR changes by at most 1.67 percentage points. These results indicate that cross-split semantic redundancy and near-duplicate samples have little effect on the reported attack performance.

Table 8: Effect of cross-split data redundancy on SDA performance on LAION_CC_SBU.

Model	**AUC**		**TPR@1%FPR**	
	Original	Low-redundancy	Original	Low-redundancy
GPT-5	0.8688	0.8693	55.44	56.35
GPT-4o	0.8543	0.8549	50.40	51.26
Gemini 3.1	0.8107	0.8101	53.60	53.54
Qwen3-VL	0.8126	0.8164	37.80	39.47
GLM-4.6V	0.8355	0.8328	28.40	28.19

A.5 Retrieval- and Generator-Level Effects

We further analyze the contributions of retrieval and generation to SDA through retrieval-only, generator-only, and No-RAG settings. As shown in Table 9, the retrieval-only setting achieves an AUC of 0.8100 and TPR@1%FPR of 48.15%, outperforming the generator-only setting (0.7636 AUC and 28.76% TPR@1%FPR). Removing RAG further reduces the AUC to 0.5392, indicating that retrieval disruption plays an important role in the membership signal exploited by SDA.

Table 9: Ablation study on retrieval- and generator-level effects.

Setting	**AUC**	**TPR@1%FPR (%)**
Full Attack	0.8688	55.44
Retrieval-only	0.8100	48.15
Generator-only	0.7636	28.76
No-RAG	0.5392	6.08

A.6 Sensitivity to Retrieval-Focused Prompt Wording

We evaluate three alternative retrieval-focused prompts to assess whether SDA depends on a detailed grounding instruction or its specific wording. The simple retrieval-focused prompt provides a minimal grounding instruction, while the other two prompts are paraphrases of the original context-grounded instruction generated using GPT-5.6-Sol. All other attack and evaluation settings remain unchanged.

Simple Retrieval-Focused Prompt.
Based STRICTLY on the retrieved examples above, please describe this target image.

Grounded Paraphrase 1.
Based ONLY on the retrieved examples provided above, describe the target image. Include exclusively details that are directly supported by these examples. If the retrieved examples do not exactly correspond to the target image, clearly express this uncertainty using phrases such as "appears to be" or "may be." Do NOT rely on general knowledge to add any information that is not grounded in the retrieved examples.

Grounded Paraphrase 2.
Describe the target image using STRICTLY the retrieved examples above. Limit your description to information that can be clearly inferred from those examples. When the retrieved examples are not a precise match for the target image, indicate the uncertainty with wording such as "seems to be," "likely," or "possibly." Do NOT introduce details based on prior or general knowledge unless they are supported by the retrieved examples.

Table 10: Sensitivity of SDA to alternative retrieval-focused prompts on LAION_CC_SBU.

Model	Simple Retrieval-Focused		Grounded Paraphrase 1		Grounded Paraphrase 2	
	AUC	TPR@1%FPR	AUC	TPR@1%FPR	AUC	TPR@1%FPR
GPT-5	0.8297	52.14	0.8569	54.38	0.8601	55.16
GPT-4o	0.8239	46.18	0.8451	49.32	0.8480	50.96
Gemini 3.1	0.7746	50.21	0.8094	52.46	0.8042	53.08
Qwen3-VL	0.7911	36.02	0.8121	37.42	0.8067	36.31
GLM-4.6V	0.8022	25.75	0.8266	27.63	0.8301	28.06

A.7 Evaluation of SDA Performance on Medical Datasets

To evaluate SDA on medical image-caption data, we conduct additional experiments on two medical multimodal datasets, ROCOv2 [47] and PMC-OA [31]. For each dataset, we randomly sample 500 member and 500 non-member examples and evaluate SDA across five target VLMs.

As shown in Table 11, SDA achieves an average AUC of 0.8589 and TPR@1%FPR of 60.62% on ROCOv2. On PMC-OA, SDA obtains an average AUC of 0.8515 and TPR@1%FPR of 57.12%. These results indicate that SDA remains effective on medical image-caption corpora, extending its evaluation beyond general-domain datasets.

Table 11: Attack performance of SDA on ROCOv2 and PMC-OA across five target VLMs.

Model	ROCOv2		PMC-OA	
	AUC	TPR@1%FPR	AUC	TPR@1%FPR
GPT-5	0.8828	69.40	0.8765	65.20
GPT-4o	0.8616	62.35	0.8537	59.14
Gemini 3.1	0.8497	58.72	0.8421	53.86
Qwen3-VL	0.8542	55.18	0.8478	55.47
GLM-4.6V	0.8461	57.46	0.8376	51.92
Average	**0.8589**	**60.62**	**0.8515**	**57.12**

A.8 Retrieval Neighborhood Analysis

We analyze the retrieval neighborhoods before and after adversarial perturbation to investigate the retrieval-side behavior underlying SDA. We compare member and non-member queries using neighbor overlap, the proportion of changed neighbors, retrieval similarity before and after perturbation, and shared-rank displacement at different values of k. These retrieval-side measurements are used only for analysis and are not available to the attacker in the black-box setting.

As shown in Table 12, members exhibit larger retrieval similarity drops than non-members across all evaluated values of k. At $k = 3$, the similarity drop is 0.1634 for members, compared with 0.1326 for non-members. The difference is larger at $k = 1$, where the respective drops are 0.2237 and 0.1396.

Notably, members retain higher top-1 neighbor overlap than non-members (0.8145 vs. 0.3307), despite exhibiting a larger similarity drop. This distinction indicates that neighbor replacement and similarity degradation capture different aspects of retrieval disruption. The perturbation can substantially reduce the similarity of member queries to retrieved evidence even when the nearest

neighbor remains unchanged. These findings provide retrieval-side evidence for the differential sensitivity exploited by SDA.

Table 12: Retrieval neighborhood changes after adversarial perturbation for member and non-member queries.

Group	k	Neighbor Overlap@k	Changed Neighbors (%)	Similarity Before	Similarity After	Similarity Drop	Shared-Rank Displacement
Member	1	0.8145 ± 0.3887	18.55 ± 38.87	0.8090 ± 0.0104	0.5853 ± 0.1818	0.2237 ± 0.1832	0.1855 ± 0.3887
	3	0.4940 ± 0.3037	50.60 ± 30.37	0.6527 ± 0.0406	0.4893 ± 0.1432	0.1634 ± 0.1453	0.5726 ± 1.0591
	5	0.4593 ± 0.2805	54.07 ± 28.05	0.6072 ± 0.0460	0.4628 ± 0.1329	0.1444 ± 0.1338	1.1147 ± 1.5952
	10	0.4694 ± 0.2610	53.06 ± 26.10	0.5608 ± 0.0498	0.4357 ± 0.1203	0.1251 ± 0.1205	2.5526 ± 2.5753
Non-member	1	0.3307 ± 0.4705	66.93 ± 47.05	0.5715 ± 0.0625	0.4319 ± 0.1438	0.1396 ± 0.1443	0.6693 ± 0.4705
	3	0.4035 ± 0.3293	59.65 ± 32.93	0.5491 ± 0.0582	0.4165 ± 0.1366	0.1326 ± 0.1377	1.3383 ± 1.2077
	5	0.4297 ± 0.3008	57.03 ± 30.08	0.5362 ± 0.0570	0.4076 ± 0.1315	0.1286 ± 0.1331	1.9572 ± 1.7487
	10	0.4701 ± 0.2775	52.99 ± 27.75	0.5162 ± 0.0558	0.3945 ± 0.1250	0.1217 ± 0.1270	3.3671 ± 2.8187

A.9 Effect of the maximum generation length on attack performance

Figure 7 studies how the maximum generation length affects SDA. AUC remains stable across all token sizes, ranging from 0.8609 to 0.8752, although TPR@1%FPR shows greater variation and peaks at 64 tokens. Increasing the generation length does not consistently improve results; 128 and 256 tokens lead to lower TPR@1%FPR. This is likely because concise grounded descriptions already contain enough retrieval-dependent information, while longer generations may add more details that weaken the semantic degradation score.

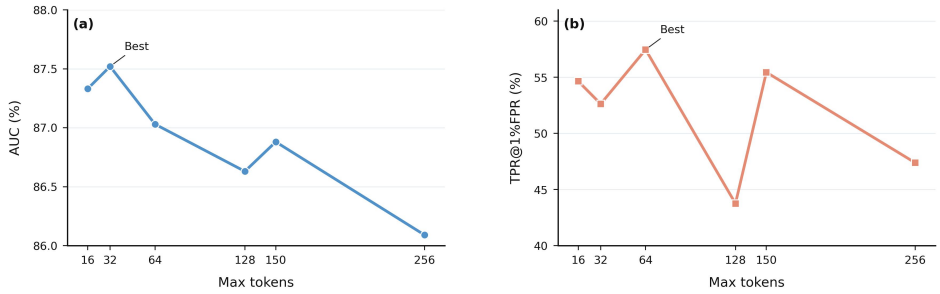


Figure 7: **Effect of the maximum generation length on SDA performance.**

A.10 Defense Evaluation

We evaluate SDA against a system-prompt defense adapted from MrM [63], which aims to restrict the disclosure of private retrieval information. Table 13 compares the attack performance with and without this defense on LAION_CC_SBU. These results indicate that the defense does not consistently suppress the membership signal exploited by SDA.

Table 13: SDA performance with (w/) and without (w/o) defense on LAION_CC_SBU.

	AUC		TPR@1%FPR	
Model	**w/o**	**w/**	**w/o**	**w/**
GPT-5	0.8688	0.8623	55.44	58.48
GPT-4o	0.8543	0.8605	50.40	51.74
Gemini 3.1	0.8107	0.8143	53.60	51.48
Qwen3-VL	0.8126	0.8140	37.80	36.29
GLM-4.6V	0.8355	0.8407	28.40	27.25
Average	0.8364	0.8384	45.13	45.05

A.11 Limitations

Although we evaluate a system-prompt defense, a broader investigation of defenses against retrieval-corpus membership inference is beyond the scope of this work. Developing effective privacy protections for multimodal RAG systems remains an important direction for future work.

A.12 Ethical Considerations and Broader Impacts

This work studies MIA against multimodal RAG systems to better understand privacy risks introduced by RAG over private image–caption corpora. The broader positive impact is that SDA can serve as a privacy evaluation tool for system owners, helping them identify whether generated responses expose the presence of private records even when retrieved evidence is not directly exposed. Our experiments are conducted on public image-caption datasets in controlled retrieval-corpus settings; we do not collect new human-subject data, interact with human participants, or attack real private deployments. We report our attack performance and do not disclose sensitive records or target individuals.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly express our motivation and insight in the Abstract and Introduction. The contributions are summarized in the Introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We clearly express our limitations in the Appendix A.11.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: There is no assumption or theoretical result in this work.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our paper provides a detailed account of implementation details and hyperparameter settings of our method for reproducing in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The used datasets are publicly available. We will consider releasing the code upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide implementation details in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We fix the random seeds in experiments for fair comparison.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We specify the type of compute workers in implementation details in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This work focuses on membership inference attack, which is the most common privacy risk in machine learning. We discussed the potential impacts in Appendix A.12.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the datasets and models used in our paper and give a brief introduction.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper has nothing to do with crowdsourcing and human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [No]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.