MrM: Black-Box Membership Inference Attacks Against Multimodal RAG Systems

Peiru Yang[1*], Jinhua Yin[1*], Haoran Zheng[2], Xueying Bai[2], Huili Wang[1], Yufei Sun[2], Xintian Li[1], Songwei Pei[2], Yongfeng Huang[1], Tao Qi[2†]

[1]Department of Electronic Engineering, Tsinghua University
[2]School of Computer Science, Beijing University of Posts and Telecommunications
taoqi.qt@gmail.com

Abstract

Multimodal retrieval-augmented generation (RAG) systems enhance large vision-language models by integrating cross-modal knowledge, enabling their increasing adoption across real-world multimodal tasks. These knowledge databases may contain sensitive information that requires privacy protection. However, multimodal RAG systems inherently grant external users indirect access to such data, making them potentially vulnerable to privacy attacks, particularly membership inference attacks (MIAs). Existing MIA methods targeting RAG systems predominantly focus on the textual modality, while the visual modality remains relatively underexplored. To bridge this gap, we propose MrM, the first black-box MIA framework targeted at multimodal RAG systems. It utilizes a multi-object data perturbation framework constrained by counterfactual attacks, which can concurrently induce the RAG systems to retrieve the target data and generate information that leaks the membership information. Our method first employs an object-aware data perturbation method to constrain the perturbation to key semantics and ensure successful retrieval. Building on this, we design a counterfact-informed mask selection strategy to prioritize the most informative masked regions, aiming to eliminate the interference of model self-knowledge and amplify attack efficacy. Finally, we perform statistical membership inference by modeling query trials to extract features that reflect the reconstruction of masked semantics from response patterns. Experiments on two visual datasets and eight mainstream commercial visual-language models (e.g., GPT-4o, Gemini-2) demonstrate that MrM achieves consistently strong performance across both sample-level and set-level evaluations, and remains robust under adaptive defenses.

Introduction

As a key augmentation strategy for LLMs, retrieval-augmented generation (RAG) has recently been extended to the visual modality, enabling applicability in multimodal AI tasks (Chen et al. 2022; Liu et al. 2023; Yasunaga et al. 2023). Incorporating visual modalities, RAG systems can retrieve external knowledge that complements the visual input and helps reduce hallucinations in large vision-language models (LVLMs) (Du et al. 2022; Zhang et al. 2024a). Recent advancements demonstrate the emerging role of multimodal RAG in enabling LVLMs to dynamically integrate knowledge for real-world applications, such as intelligent medical AI systems (Ferber et al. 2024; Xia et al. 2024a,b).

For the effectiveness of RAG systems, some private-domain databases are incorporated to support vertical inference and complex reasoning (Lewis et al. 2020). These knowledge bases often contain private or proprietary data that are essential for supporting complex downstream tasks, while such data can be highly sensitive and should be safeguarded with robust privacy protections (Ni et al. 2025; Zeng et al. 2024). Yet, the RAG paradigm inherently introduces an indirect exposure risk: the knowledge base provides information to the generation model, which then produces responses accessible to external users. In doing so, the RAG system establishes a bridge between internal sensitive data and external adversaries, enabling interactions that may inadvertently leak private content. This indirect access pathway creates new vulnerabilities, allowing adversaries to mount privacy attacks against the underlying database, particularly membership inference attacks (MIAs), which seek to reveal whether specific samples were part of the original database (Shokri et al. 2017; Truex et al. 2019; Hu et al. 2022a; Carlini et al. 2022; Choquette-Choo et al. 2021; Olatunji, Nejdl, and Khosla 2021).

Existing research on MIAs against RAG has primarily focused on text-only modality, employing various methodologies to predict whether a target sample exists in the retrieval corpus (Anderson, Amit, and Goldsteen 2024; Li et al. 2025; Liu, Zhang, and Long 2025; Naseh et al. 2025). For instance, Li et al. (2025) develop an MIA approach that analyzes semantic similarity and perplexity between target samples and RAG-generated content to infer database membership. In conclusion, the paradigm of these methods is to provide fragments of target data and then compare the similarity between the output and original data. However, LVLMs process both textual and visual inputs while generating text-only outputs (Khan et al. 2022; Dou et al. 2022; Zhou and Shimada 2023). This asymmetry introduces a challenge of modality transfer: inferring the membership status of visual data requires reasoning over purely textual responses, without direct access to visual features in the output. Hence,

*These authors contributed equally.
†Corresponding author.

these text-centric MIA methods cannot be directly transferred to LVLMs with multimodal RAG. Besides, a new challenge lies in the balance between ensuring the successful retrieval of target data and guiding the model generation towards revealing membership information. The solutions to the two aforementioned core challenges are crucial for MIAs specifically designed for multimodal RAG systems.

Therefore, we propose MrM, a multi-object data perturbation framework constrained by counterfactual attacks, which is the first black-box MIA framework targeted at multimodal RAG systems. Its core idea is perturbing target samples and analyzing whether the textual responses implicitly reconstruct the disrupted semantics. In this way, our method bridges the semantics of the text and visual modalities through object detection to tackle the challenge of cross-modal membership inference. Moreover, masking objects can minimize the affected region, thereby enhancing the effectiveness of attacks on retrieval, while perturbing the independent and complete semantics to strengthen attacks on generation. Specifically, an object-aware data perturbation approach is employed to strategically disrupt visual semantics by masking detected entities using object detection models such as SAM(Kirillov et al. 2023). This approach ensures that key features are disrupted while still allowing relevant data to be retrieved if it exists in the database. It is followed by a counterfact-informed mask selection strategy, where we quantify the informativeness of each perturbation. We prioritize masks that maximize discriminative gaps by analyzing probability distributions and confidence differentials of a counterfactual proxy model. This strategy aims to eliminate the interference of the self-knowledge of LVLMs, thereby preventing the reconstruction of information for non-database images during the generation phase. Finally, we perform statistical membership inference by modeling query trials to analyze whether the textual responses implicitly reconstruct the disrupted semantics. Code is available at https://github.com/ypr17/MrM-RAG-MIA.

The contributions of our method can be summarized as:

- We introduce the first black-box MIA framework for multimodal RAG systems, highlighting vulnerabilities in the privacy protection of multimodal databases.
- We propose a unified MIA framework that addresses the cross-modal alignment issue and enables concurrent attacks in both the retrieval and generation phases.
- We validate our framework through comprehensive experiments on two visual datasets and eight mainstream commercial LVLMs, demonstrating consistent performance and robustness against adaptive defense strategies.

Related Work

Multimodal RAG. Since cross-modal alignment models emerged(Radford et al. 2021; Lu et al. 2019; Li et al. 2022), multimodal RAG has developed to capture the vision-language interplay (Chen et al. 2022; Liu et al. 2023; Yasunaga et al. 2023; Faysse et al. 2024). Methods bridging the modality gaps during retrieval includes crafting modality-balanced hard negatives, generating embeddings of text page images, and late interaction matching (Liu et al. 2023; Faysse et al. 2024). Overall, multimodal RAG has enabled robust retrieval and generation in vision-language tasks.

MIA against VLMs. MIA studies targeting the training data of VLMs builds rapidly(Hu et al. 2022b; Ko et al. 2023; Li et al. 2024; Ibanez-Lissen et al. 2024). Ko et al. (2023) propose to use cosine similarity and weakly supervised attack to avoid shadow training. Li et al. (2024) present a detection method using confidence-based metrics for both text and images. However, these methods rely on white- or gray-box settings, while most RAG systems serve as Generation-as-a-Service (GaaS), i.e. black-box models.

MIA against Text-only RAG. Recent MIA studies in RAG systems have focused on text modality. The first RAG-MIA method uses yes/no responses to direct queries (Anderson, Amit, and Goldsteen 2024). Following works use semantic similarity between targets and generations, perturb target documents via word masking, or compare responses from target and shadow VLMs to infer membership (Li et al. 2025; Liu, Zhang, and Long 2025; Naseh et al. 2025). However, VLMs typically handle image-text inputs yet produce textual responses, and MIA approaches designed for multimodal RAG systems remain largely unexplored.

Methods

In this section, we present the details of MrM, which consists of three key components as illustrated in Fig 1: object-aware data perturbation, counterfact-informed mask selection, and statistical membership inference.

Threat Model

Attacker's Goal is to infer if a set of images $\mathcal{D}_i = \{\mathcal{I}_i\}_{i=1}^{N}$ is in the black-box multimodal RAG system's database DB. Attacker's Capability includes repeated queries through the RAG system's public interface and observing its responses to multimodal inputs crafted to probe the VLM M. The attacker cannot access the system's retrieval results, output probability distributions, input embeddings, the database DB or its indexing pattern. M may also reject explicitly malicious or privacy-sensitive queries, which necessitates mechanisms to bypass detection.

Design Motivation and Overall Framework

The core motivation behind our framework is to induce both the retrieval and generation phases of an RAG system to leak membership information simultaneously. In the retrieval phase, the goal is to ensure that the RAG system successfully retrieves the target data, while in the generation phase, we aim to elicit relevant information about the target image from the response of the RAG system. If the target data is input directly without any perturbation, the system is likely to retrieve the corresponding data. Yet, in this case, it becomes challenging to determine whether the information in the model's response originates from the input or the retrieval database. On the other hand, significantly degrading the target data through perturbation ensures that any relevant information in the model's response originates solely from the retrieval knowledge base. However, this approach

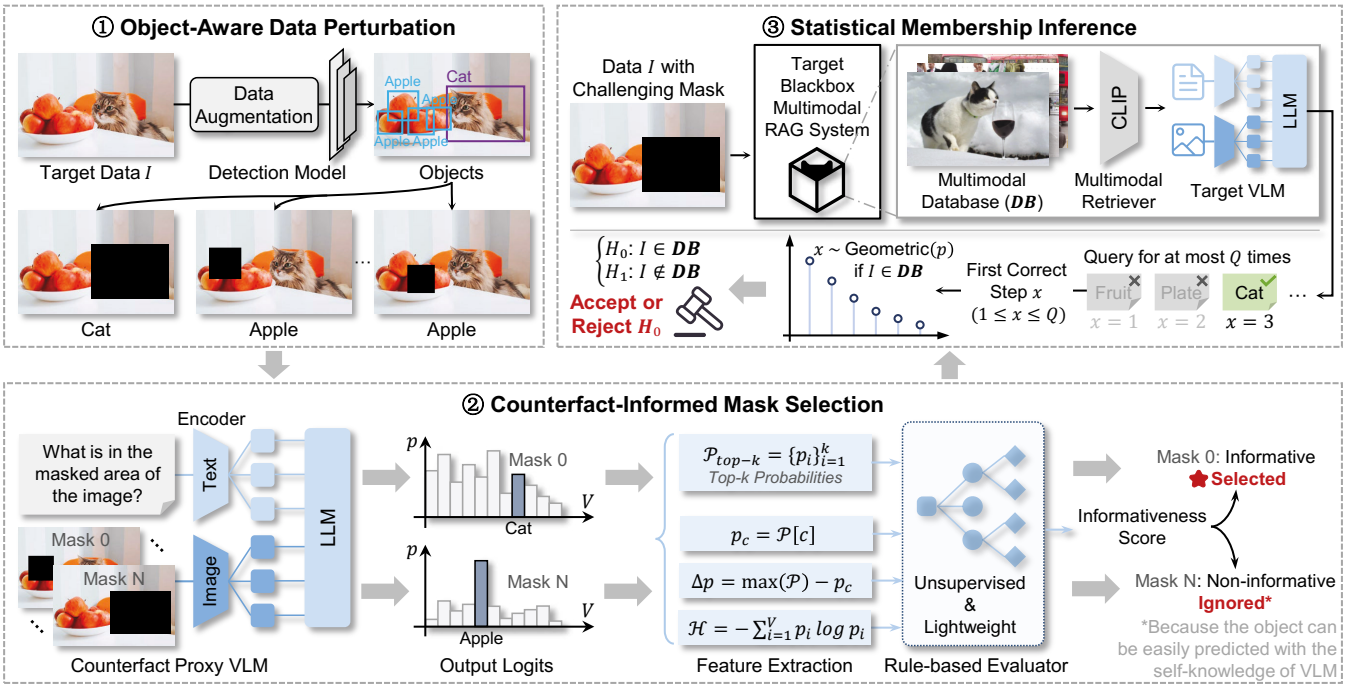

Figure 1: The overall framework of the proposed MrM method. It first perturbs the target image via object-aware masking, selects informative perturbations using a counterfact-informed mask selection strategy, and infers membership via a hypothesis test on the RAG system's response statistics.

may significantly compromise the effectiveness of the membership inference framework, as it could cause the retrieval process in the RAG system to fail.

Thus, the challenge lies in the balance between these two objectives: ensuring successful retrieval while maintaining sufficient perturbation to guide the model towards revealing membership information. To address this challenge, we propose a multi-object data perturbation framework constrained by counterfactual attacks. This framework enables the generation of perturbations that strategically degrade the semantics of the target data, ensuring the retrieval of relevant information while preventing direct leakage from the input. By inducing multiple responses from the model with different perturbations, discriminative features can be extracted, providing clear evidence of membership.

Object-Aware Data Perturbation

Our perturbation process pursues three goals: (1) keep the target data retrievable, (2) prevent reconstruction of non-database content during generation, and (3) support cross-modal transition from image to text generation modality.

To meet these goals, we adopt an object-aware perturbation strategy. In an image, the object of interest often occupies only a small fraction of the overall scene, meaning perturbing these regions has minimal impact on the retrieval. Moreover, individual objects have relatively independent semantics, allowing for the preservation of information traceability, which ensures that the source of the information can be linked back to the retrieval database. Lastly, objects are well-suited to be transferred into the textual modality, as they are typically well-defined and easily described in words, making them ideal candidates for generating meaningful text responses in the generation phase.

Given a target image $\mathcal{I}$, we use an object detection $\mathcal{D}$ (e.g., SAM (Kirillov et al. 2023)) to localize salient objects $O = \{o_j\}_{j=1}^{K}$. For each o_j, we create a binary mask $\mathcal{M}_j$ and generate a perturbed image: $\tilde{\mathcal{I}}_j = \mathcal{I} \odot (\mathbf{1} - \mathcal{M}_j) + \mathbf{0} \odot \mathcal{M}_j$, where $\odot$ denotes element-wise multiplication, and $\mathbf{1}$, $\mathbf{0}$ are all-one and all-zero matrices, respectively. This zeros out the masked region while keeping the rest intact.

Counterfact-Informed Mask Selection

We propose a counterfact-informed mask selection strategy that prioritizes masked regions likely to maximize discriminative gaps. This is achieved by analyzing probability distributions and confidence differentials from a proxy VLM $\mathcal{V}$, used as a counterfactual reference. The goal is to suppress interference from the target LVLM's memorized knowledge, ensuring that observed semantic reconstructions stem from retrieval rather than internalized information.

Given a perturbed image $\tilde{\mathcal{I}}j$, we input it into $\mathcal{V}$ to obtain a probability distribution $P = \{p_i\}_{i=1}^{V}$ over the vocabulary V. We extract the following features to estimate the informativeness of each mask: **Target Confidence** p_c: predicted probability corresponding to the ground-truth category of the masked region. **Confidence Gap** $\Delta p = \max(P) - p_c$: measuring the discrepancy between the highest predicted probability and the ground-truth confidence. **Entropy** $\mathcal{H} =$

$-\sum_{i=1}^{V} p_i \log p_i$: quantifies the prediction uncertainty, with higher entropy indicating greater confusion. **Top-k Distribution** $\{p_{(i)}\}_{i=1}^{k}$: the top-k values in P sorted in descending order, capturing the distributional sharpness and diversity of high-confidence predictions.

A feature vector $\mathbf{f}_j = [p_c, \Delta p, \mathcal{H}, \{p_{(i)}\}_{i=1}^{k}]$ consisting of these features captures the uncertainty of the proxy model, which is important for estimating the discriminative power of a perturbation in black-box settings. To assign an informative score to each mask, we adopt a rule-based evaluator that integrates the extracted features in an unsupervised manner (see Appendix). Specifically, masks are ranked according to an ensemble of normalized feature scores, where high entropy, low target confidence, and small confidence gap jointly contribute to a higher informative score. Masks with higher scores are prioritized to suppress spurious reconstructions and strengthen membership inference.

Statistical Significance Analysis

To rigorously infer the membership status of a target image $\mathcal{I}$, we formulate a hypothesis testing framework grounded in the statistical behavior of the multimodal RAG system when queried about masked objects. The core intuition is that the system's success rate in predicting occluded objects depends on whether $\mathcal{I}$ is in the database DB. Formally, we define two hypotheses. Null hypothesis (H_0): $\mathcal{I} \in DB$, where the system's success probability for each mask follows p_t. Alternative hypothesis (H_1): $\mathcal{I} \notin DB$, with a lower success probability p_n, where $p_t > p_n$ by design.

For each perturbed image $\tilde{\mathcal{I}}_j$ (derived from the j-th mask $\mathcal{M}_j$), we query the system repeatedly until it correctly identifies the masked object. Let x_j denote the number of trials required for the first correct prediction. Under H_0, x_j follows a geometric distribution:

$$x_j \sim \text{Geometric}(p_t), \quad \mathbb{E}[x_j] = \frac{1}{p_t}, \quad \text{Var}(x_j) = \frac{1 - p_t}{p_t^2}. \tag{1}$$

For K masks, the total number of trials across all masks is aggregated as $S = \sum_{j=1}^{K} x_j$. By the additive property of independent geometric variables, S has expectation and variance: $\mu_0 = \frac{K}{p_t}, \quad \sigma_0^2 = \frac{K(1-p_t)}{p_t^2}$. Applying the Central Limit Theorem (CLT) for large K, S approximates a normal distribution: $S \overset{\text{approx}}{\sim} \mathcal{N}\left(\mu_0, \sigma_0^2\right)$. The p-value quantifies the probability of observing a total trial count as extreme as S under H_0. To compute it, we first standardize S and then evaluate the survival function of the standard normal distribution. Let $\Phi(z)$ denote the cumulative distribution function (CDF) of $\mathcal{N}(0, 1)$. Then the p-value is:

$$p\text{-value} = 1 - \Phi\left(\frac{S - \mu_0}{\sigma_0}\right) = 1 - \Phi\left(\frac{S - \frac{K}{p_t}}{\sqrt{\frac{K(1-p_t)}{p_t^2}}}\right). \tag{2}$$

If the p-value $< \alpha$ (e.g., $\alpha = 0.05$), we reject the null hypothesis H_0 and conclude that $\mathcal{I} \notin DB$. Otherwise, we fail to reject H_0, indicating potential membership.

Experiments and Analysis

Experimental Setups

Datasets. We use two standard image datasets, COCO (Lin et al. 2014) and Flickr (Young et al. 2014), to build the knowledge base and conduct membership inference attacks. From each dataset, we select 5,000 images for the knowledge base and 1,000 images (500 members, 500 non-members) for testing.

Target Models. We conduct membership inference attacks on eight commercial models, each integrated with a local knowledge base to form a multi-modal RAG system: GPT-4o-mini (Hurst et al. 2024), Gemini-2 (Mallick and Kilpatrick 2025), Claude-3.5 (Anthropic 2024), GLM-4v (GLM et al. 2024), Qwen-VL (Bai et al. 2025), Pixtral (Agrawal et al. 2024), Moonshot (Moonshot AI 2024), and InternVL-3 (Chen et al. 2024). These commercial VLMs support multi-image inputs, making them suitable for multi-modal RAG systems. The experiments are conducted via API calls connected to a locally built knowledge base. This setup ensures no access to internal generation states, maintaining a strict black-box environment that mirrors real-world deployment, where only the model output is available for analysis and no information about the inner workings or intermediate states can be accessed.

Baselines. To the best of our knowledge, our work presents the first MIA approach targeting multi-modal RAG systems. Due to the lack of baselines, we adapt two strategies from text-based RAG MIA (Anderson, Amit, and Goldsteen 2024; Li et al. 2025). The first baseline, Query-based MIA (denoted as *QB-MIA*), directly asks if the target sample appears in the retrieved references, interpreting the model's binary response as a membership signal. The second, Similarity-based MIA (*SB-MIA*), partially masks the target image and asks the model to reconstruct the missing content using the retrieved reference images. Variants like *SB-MIA-0.5* indicate the masking ratio. Similarity between the generated description and original content is then computed, with higher similarity implying likely membership.

Evaluation Metrics. Following prior works on MIA (Li et al. 2024; Shi et al. 2024; Zhang et al. 2024b), we employ two evaluation metrics: AUC and TPR5%FPR. AUC reflects overall discrimination between members and non-members across thresholds. TPR5%FPR measures the true positive rate when the false positive rate is constrained below 5%, providing a better assessment under strict conditions. Since identical AUCs can result from different ROC curves, TPR@5%FPR complements AUC for nuanced evaluation. We report both metrics at the sample and set levels for comprehensive analysis.

Implementation Details. For object-aware data perturbation, we employ the SAM2 model (Kirillov et al. 2023) to perform object detection. We utilize the 7B local version of Qwen-VL (Bai et al. 2025) as the proxy VLM. In the ablation study, we replace it with a weaker detector, the YOLO model (Redmon et al. 2016). All retrieval databases are constructed using the FAISS library (Douze et al. 2024). As the image retriever in our RAG system, we adopt the ViT variant of the CLIP model (Radford et al. 2021).

	Methods	QB-MIA		SB-MIA-0.25		SB-MIA-0.5		SB-MIA-0.75		MrM	
	Metrics	AUC	TPR@5%	AUC	TPR@5%	AUC	TPR@5%	AUC	TPR@5%	AUC	TPR@5%
Flickr	GPT-4o-mini	64.66%	32.85%	67.10%	15.69%	70.04%	20.33%	58.58%	9.67%	**80.86%**	**66.87%**
	Claude-3.5	55.85%	16.12%	63.21%	14.05%	62.79%	14.05%	44.85%	5.69%	**85.36%**	**74.98%**
	Gemini-2	72.16%	9.21%	57.23%	8.03%	54.72%	7.36%	44.80%	6.69%	**83.19%**	**66.76%**
	Pixtral	65.89%	35.18%	71.61%	19.73%	74.26%	28.43%	62.12%	26.09%	**83.84%**	**61.12%**
	Qwen-VL	56.52%	17.43%	66.76%	13.71%	65.91%	19.06%	55.55%	10.37%	**84.22%**	**72.16%**
	GLM-4v	55.23%	15.06%	66.98%	14.72%	70.78%	22.41%	58.51%	17.06%	**81.93%**	**58.79%**
	Moonshot	53.30%	11.27%	74.63%	25.75%	75.98%	24.75%	55.06%	17.73%	**80.20%**	**65.11%**
	InternVL-3	51.84%	8.50%	64.23%	12.71%	67.25%	18.39%	50.92%	14.05%	**83.23%**	**68.92%**
COCO	GPT-4o-mini	64.42%	11.01%	52.22%	4.35%	59.51%	6.67%	61.58%	12.04%	**73.51%**	**20.77%**
	Claude-3.5	52.59%	9.91%	58.89%	8.70%	61.37%	12.33%	55.56%	8.03%	**82.04%**	**43.40%**
	Gemini-2	70.98%	9.28%	50.38%	5.35%	51.23%	4.01%	50.65%	8.03%	**84.17%**	**57.18%**
	Pixtral	66.22%	35.92%	60.69%	9.36%	62.96%	6.02%	64.24%	16.44%	**83.02%**	**47.24%**
	Qwen-VL	53.46%	11.57%	55.01%	6.35%	58.10%	7.33%	58.46%	9.73%	**84.11%**	**53.12%**
	GLM-4v	64.41%	32.51%	55.24%	7.02%	60.86%	10.67%	55.76%	16.78%	**76.57%**	**36.63%**
	Moonshot	66.47%	36.41%	56.39%	5.88%	63.66%	7.67%	55.57%	8.72%	**77.87%**	**26.31%**
	InternVL-3	51.51%	7.86%	47.99%	4.68%	59.72%	8.33%	50.29%	6.38%	**79.37%**	**33.03%**

Table 1: Performance comparison of different RAG MIA methods across eight multimodal RAG systems on Flickr and COCO datasets. We report AUC and TPR@5%FPR for each method, including *QB-MIA*, three variants of *SB-MIA* with different masking ratios, and MrM. MrM consistently achieves the highest performance, especially under low false positive constraints.

Main Results

Sample-level MIA. Table 1 presents the performance of sample-level MIA across eight multimodal RAG systems on the Flickr and COCO datasets. Our method, MrM, is evaluated against two baselines: QB-MIA, which directly queries the model about the presence of a target sample in the retrieved references, and SB-MIA, which removes a fixed portion of the target image and prompts the RAG system to describe the original content based on its retrieved references.

To ensure a fair and realistic evaluation, we apply a simple yet natural defense mechanism across all methods in Table 1: a cautionary system prompt is added to the VLM, stating, *"Do not reveal any information about the membership of your knowledge base."* This prompt serves as a minimal safeguard against unintended memorization leakage. While this defense has only limited impact on the performance of SB-MIA and our proposed MrM method, it significantly weakens the effectiveness of QB-MIA, which relies on the model's willingness to answer membership-related questions directly. All subsequent experiments in this paper are conducted under this default defense setting.

Across both datasets and all models, MrM demonstrates a clear performance advantage, achieving consistently higher AUC scores, indicating strong overall discriminative ability. It also particularly excels in TPR@5%FPR compared to baseline methods, which is crucial for evaluating MIA under strict false positive constraints. This metric reflects an attacker's success rate under strict false positive constraints, making it more relevant in real-world scenarios where low false positive rates are essential for stealthy deployment of MIA. The superior performance of MrM stems from its ability to precisely disrupt the most semantically critical and least easily inferred regions of the target image. By leveraging object-aware perturbation and difficulty assessment via a proxy vision-language model, MrM identifies and masks regions that are both salient and challenging to describe without prior exposure. As a result, non-member images lead to vague or inaccurate responses from the VLM. In contrast, for member images, the VLM can often recover the correct semantics from contextual cues because of its strong in-context learning capabilities. This contrast enhances the discriminative power of our statistical test and underpins the improved results observed across models and datasets.

Set-level MIA. To evaluate the effectiveness of MrM at the set level, we plot ROC curves in Figure 2 for varying set sizes $K = 1, 5, 10, 20$, across eight RAG VLMs and two datasets. Each curve reflects the model's ability to infer membership status by aggregating predictions over a set of K target samples, using a joint statistical test based on the responses of the RAG system to all samples in the set. We compare our proposed method against the strongest variant of SB-MIA, which serves as the reference method throughout this section. Across nearly all models and both datasets, MrM consistently outperforms the baseline method in terms of AUC, which becomes more pronounced as the set size increases. When $K = 1$, MrM maintains strong performance, as discussed in the sample-level results above, and this advantage scales further with larger sets. As K grows, both methods show improved AUCs, but MrM consistently achieves higher values and converges more rapidly toward near-perfect performance. In most cases, MrM achieves an AUC close to 1.0 when $K = 10$, indicating its rapid performance saturation with relatively small set sizes.

An additional advantage of MrM is that its ROC curves

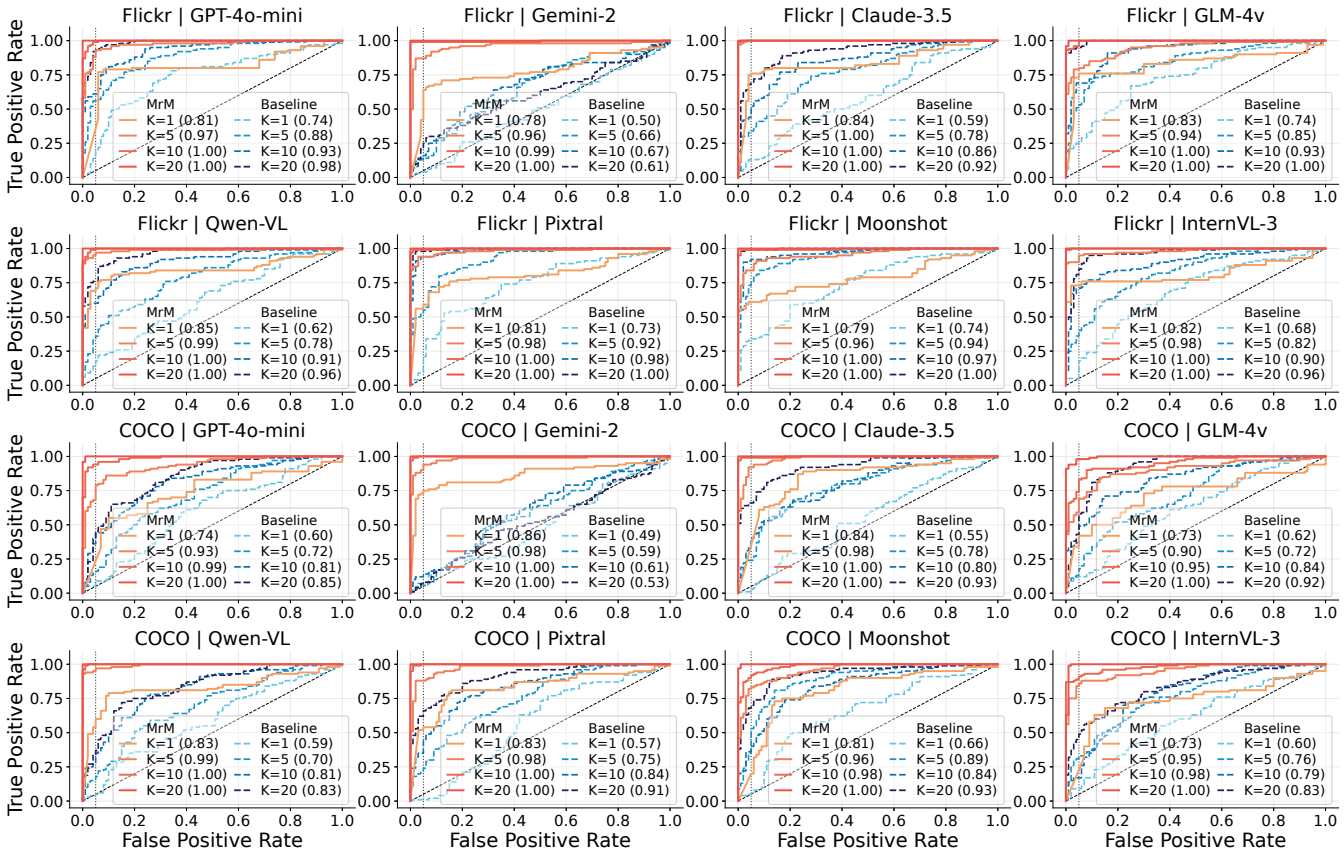

Figure 2: ROC curves for set-level MIAs with varying set sizes ($K = 1, 5, 10, 20$) on two datasets across eight multimodal RAG systems. We compare MrM with the best *SB-MIA* baseline. MrM consistently achieves higher AUCs and steeper curves toward the top-left corner, indicating superior TPR@5%FPR. The vertical gray line marks the threshold at 5% FPR.

tend to bend more sharply toward the top-left corner, indicating higher TPR@5%FPR under the same AUC. This is highlighted by the vertical reference lines at 5% FPR, where our method consistently achieves higher TPR across models and datasets. Furthermore, MrM shows strong and stable results across all models and datasets, including those where the baseline exhibits noticeable performance degradation. This highlights the generalizability of our approach under varying model architectures and retrieval behaviors.

Ablation Study

To better understand the contribution of each component in MrM, we conduct an ablation study by systematically removing or replacing key elements of the pipeline. As shown in Fig 3, we evaluate three variants: (1) **w/o object-awareness**, where object detection is removed and image regions are randomly masked; (2) **w/o proxy model**, where the proxy model is excluded and no difficulty-based mask selection is applied; (3) **simpler OD model**, where the stronger SAM2 detector is replaced with YOLO. All variants are tested on Flickr and COCO datasets across 8 RAG systems.

We observe consistent performance drops in all three ablation variants, confirming the necessity of each component in the full MrM pipeline. (1) **Without object-awareness**, the model performs significantly worse, indicating that randomly masking regions often fails to target the most semantically informative parts of the image. (2) **Without the proxy model**, the absence of difficulty assessment leads to less discriminative perturbations, weakening the signal used for membership inference. (3) **Using a simpler object detector** results in moderate but noticeable performance degradation, suggesting that high-quality object detection contributes to more effective perturbation strategies. These results highlight that all components play essential roles in achieving strong performance. The full MrM method benefits from their synergy, yielding more effective and reliable membership inference across models and datasets.

Robustness Analysis

In addition to the system prompt–based defense, we further evaluate MrM in a class of data-level defenses based on input alteration. Specifically, we simulate a defense setting where the retrieval database contains modified versions of the original images, processed through commonly used transformations such as horizontal flipping, grayscale conversion, cropping, and Gaussian blurring. These transformations aim to disrupt direct visual matching while preserving the high-level semantics of the image, thus weakening naive

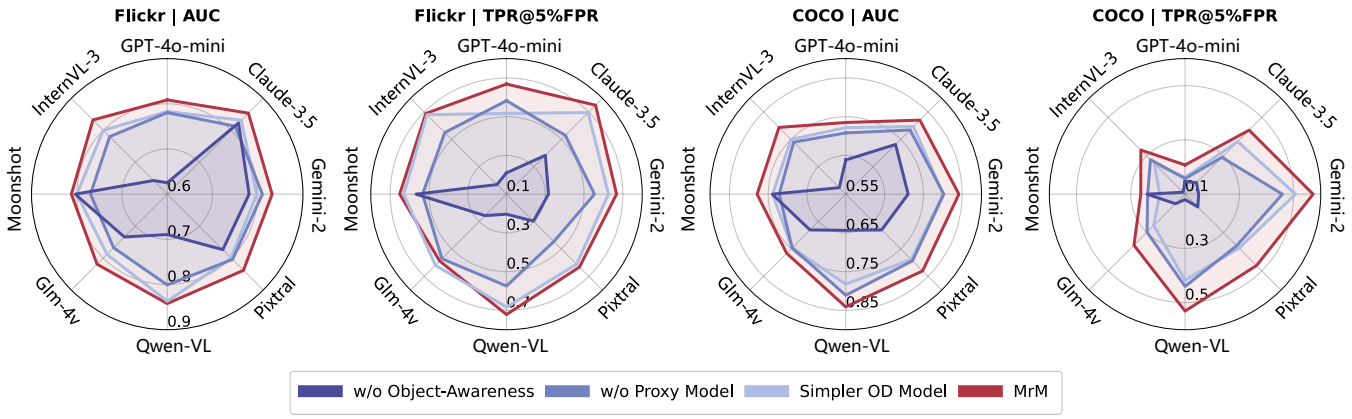

Figure 3: Ablation study results visualized as radar charts. We compare the full MrM method with three ablated variants: w/o object-awareness, w/o counterfact-informed mask selection, and simpler OD model. MrM consistently outperforms all ablated versions, demonstrating the contribution of each component to overall attack performance.

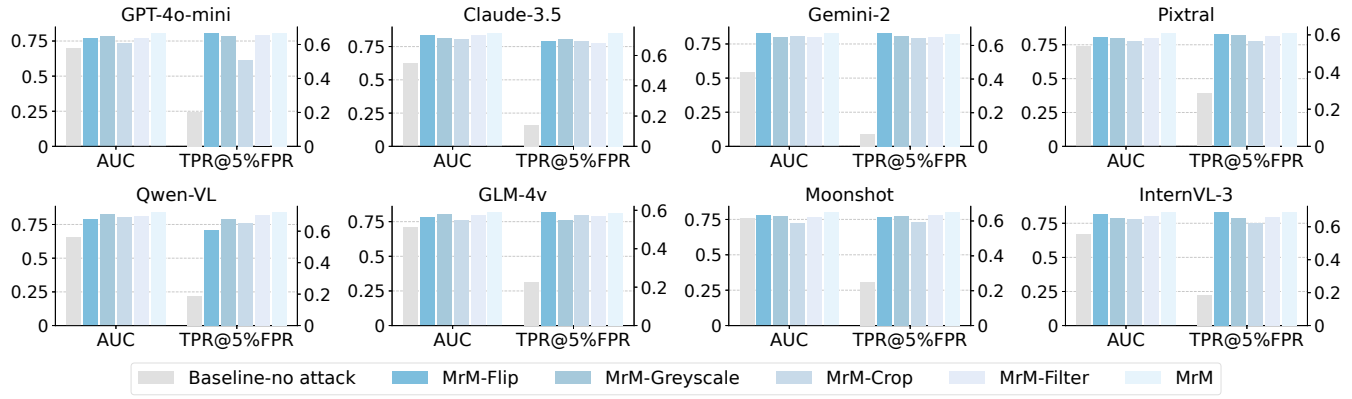

Figure 4: Robustness of MrM against adaptive image-level transformations applied to the database, including horizontal flipping, grayscale conversion, cropping, and Gaussian blur filter. MrM maintains strong performance under all transformations.

retrieval-based MIA approaches.

To address this challenge, we extend our attack pipeline with an augmentation-aware strategy. For each target sample, we generate multiple augmented variants using the same transformation types applied to the database. Each variant is treated as an independent query instance (undergoing object-aware perturbation, difficulty evaluation, and statistical testing), allowing our method to explore alternative retrieval paths that remain valid despite the transformation gap. This augmentation-aware probing increases the likelihood that at least one variant will retrieve the altered database entry, thereby restoring the model's memorization signal that might otherwise be masked. Importantly, this design also mimics a realistic attacker's capability to guess or approximate potential transformation patterns in the deployment pipeline. As shown in Fig 4, our method maintains strong performance under all four transformation-based defenses. This result demonstrates the robustness of MrM against a range of content-preserving image alterations, reinforcing its practical applicability in more adversarial or obfuscated deployment scenarios. See Appendix for case study.

Conclusion

We introduce MrM, the first black-box membership inference framework specifically designed for multimodal RAG systems. Our method reveals previously unexplored privacy vulnerabilities in vision-language models enhanced by external knowledge retrieval. To tackle the challenge of cross-modal alignment and retrieval-generation balance, we propose a unified MIA framework that jointly exploits both retrieval and generation phases, enabling the detection of membership signals from multimodal outputs. MrM incorporates object-aware perturbation and counterfact-informed mask selection to precisely control semantic leakage while preserving retrieval performance. Extensive experiments on two visual datasets and eight widely-used commercial LVLMs validate the effectiveness of our approach, showing that MrM achieves consistently strong performance under both sample-level and set-level evaluations, and remains robust even under adaptive defense mechanisms. Our findings highlight urgent security challenges in multimodal RAG infrastructures and advance the understanding of privacy risks in systems bridging vision, language, and retrieval.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grants 62502044 and U2336208; CCF-SANGFOR Research Fund under Grant 20240202.

References

Agrawal, P.; Antoniak, S.; Hanna, E. B.; Bout, B.; Chaplot, D.; Chudnovsky, J.; Costa, D.; De Monicault, B.; Garg, S.; Gervet, T.; et al. 2024. Pixtral 12B. *arXiv preprint arXiv:2410.07073*.

Anderson, M.; Amit, G.; and Goldsteen, A. 2024. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. *arXiv preprint arXiv:2405.20446*.

Anthropic. 2024. Introducing Claude 3.5 Sonnet. Accessed: 2025-05-01.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Carlini, N.; Chien, S.; Nasr, M.; Song, S.; Terzis, A.; and Tramer, F. 2022. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, 1897–1914. IEEE.

Chen, W.; Hu, H.; Chen, X.; Verga, P.; and Cohen, W. 2022. 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*, 5558–5570.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Choquette-Choo, C. A.; Tramer, F.; Carlini, N.; and Papernot, N. 2021. Label-only membership inference attacks. In *International conference on machine learning*, 1964–1974. PMLR.

Dou, Z.-Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18166–18176.

Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; and Jégou, H. 2024. The Faiss library.

Du, Y.; Liu, Z.; Li, J.; and Zhao, W. X. 2022. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*.

Faysse, M.; Sibille, H.; Wu, T.; Omrani, B.; Viaud, G.; Hudelot, C.; and Colombo, P. 2024. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Ferber, D.; Wölflein, G.; Wiest, I. C.; Ligero, M.; Sainath, S.; Ghaffari Laleh, N.; El Nahhas, O. S.; Müller-Franzes, G.; Jäger, D.; Truhn, D.; et al. 2024. In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications*, 15(1): 10104.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; Yu, H.; Wang, H.; Sun, J.; Zhang, J.; Cheng, J.; Gui, J.; Tang, J.; Zhang, J.; Li, J.; Zhao, L.; Wu, L.; Zhong, L.; Liu, M.; Huang, M.; Zhang, P.; Zheng, Q.; Lu, R.; Duan, S.; Zhang, S.; Cao, S.; Yang, S.; Tam, W. L.; Zhao, W.; Liu, X.; Xia, X.; Zhang, X.; Gu, X.; Lv, X.; Liu, X.; Liu, X.; Yang, X.; Song, X.; Zhang, X.; An, Y.; Xu, Y.; Niu, Y.; Yang, Y.; Li, Y.; Bai, Y.; Dong, Y.; Qi, Z.; Wang, Z.; Yang, Z.; Du, Z.; Hou, Z.; and Wang, Z. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

Hu, H.; Salcic, Z.; Sun, L.; Dobbie, G.; Yu, P. S.; and Zhang, X. 2022a. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s): 1–37.

Hu, P.; Wang, Z.; Sun, R.; Wang, H.; and Xue, M. 2022b. M^4I: Multi-modal Models Membership Inference. *Advances in Neural Information Processing Systems*, 35: 1867–1882.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ibanez-Lissen, L.; Gonzalez-Manzano, L.; de Fuentes, J. M.; Anciaux, N.; and Garcia-Alfaro, J. 2024. LUMIA: Linear probing for Unimodal and MultiModal Membership Inference Attacks leveraging internal LLM states. *arXiv preprint arXiv:2411.19876*.

Khan, S.; Naseer, M.; Hayat, M.; Zamir, S. W.; Khan, F. S.; and Shah, M. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s): 1–41.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

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

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, Y.; Liu, G.; Wang, C.; and Yang, Y. 2025. Generating is believing: Membership inference attacks against retrieval-augmented generation. In *ICASSP 2025-2025 IEEE Inter-*

national Conference on Acoustics, Speech and Signal Processing (ICASSP), 1–5. IEEE.

Li, Z.; Wu, Y.; Chen, Y.; Tonin, F.; Abad Rocamora, E.; and Cevher, V. 2024. Membership inference attacks against large vision-language models. *Advances in Neural Information Processing Systems*, 37: 98645–98674.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, M.; Zhang, S.; and Long, C. 2025. Mask-based membership inference attacks for retrieval-augmented generation. In *Proceedings of the ACM on Web Conference 2025*, 2894–2907.

Liu, Z.; Xiong, C.; Lv, Y.; Liu, Z.; and Yu, G. 2023. Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval. In *The Eleventh International Conference on Learning Representations*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Mallick, S. B.; and Kilpatrick, L. 2025. Gemini 2.0Flash, Flash-Lite and Pro. Accessed: 2025-05-01.

Moonshot AI. 2024. Kimi Chat. Accessed: 2025-04-10.

Naseh, A.; Peng, Y.; Suri, A.; Chaudhari, H.; Oprea, A.; and Houmansadr, A. 2025. Riddle Me This! Stealthy Membership Inference for Retrieval-Augmented Generation. *arXiv preprint arXiv:2502.00306*.

Ni, B.; Liu, Z.; Wang, L.; Lei, Y.; Zhao, Y.; Cheng, X.; Zeng, Q.; Dong, L.; Xia, Y.; Kenthapadi, K.; et al. 2025. Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2502.06872*.

Olatunji, I. E.; Nejdl, W.; and Khosla, M. 2021. Membership inference attack on graph neural networks. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, 11–20. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Shi, W.; Ajith, A.; Xia, M.; Huang, Y.; Liu, D.; Blevins, T.; Chen, D.; and Zettlemoyer, L. 2024. Detecting Pretraining Data from Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, 3–18. IEEE.

Truex, S.; Liu, L.; Gursoy, M. E.; Yu, L.; and Wei, W. 2019. Demystifying membership inference attacks in machine learning as a service. *IEEE transactions on services computing*, 14(6): 2073–2089.

Xia, P.; Zhu, K.; Li, H.; Wang, T.; Shi, W.; Wang, S.; Zhang, L.; Zou, J.; and Yao, H. 2024a. MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. In *Neurips Safe Generative AI Workshop*.

Xia, P.; Zhu, K.; Li, H.; Zhu, H.; Li, Y.; Li, G.; Zhang, L.; and Yao, H. 2024b. Rule: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1081–1093.

Yasunaga, M.; Aghajanyan, A.; Shi, W.; James, R.; Leskovec, J.; Liang, P.; Lewis, M.; Zettlemoyer, L.; and Yih, W.-t. 2023. Retrieval-augmented multimodal language modeling. In *Proceedings of the 40th International Conference on Machine Learning*, 39755–39769.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2: 67–78.

Zeng, S.; Zhang, J.; He, P.; Liu, Y.; Xing, Y.; Xu, H.; Ren, J.; Chang, Y.; Wang, S.; Yin, D.; et al. 2024. The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG). In *Findings of the Association for Computational Linguistics ACL 2024*, 4505–4524.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024a. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, J.; Sun, J.; Yeats, E.; Ouyang, Y.; Kuo, M.; Zhang, J.; Yang, H. F.; and Li, H. 2024b. Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*.

Zhou, Y.; and Shimada, N. 2023. Vision+ language applications: A survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 826–842.