MrM： 针对多模态 RAG 系统的黑盒成员推断攻击

Peiru Yang* **Jinhua Yin*** **Haoran Zheng†** **Xueying Bai†**
Huili Wang* Yufei Sun† Xintian Li* Shangguang Wang†
Yongfeng Huang* Tao Qi‡

Abstract

多模态检索增强生成（RAG）系统通过集成跨模态知识来增强大型视觉语言模型，促使其在实际多模态任务中的采用率逐渐增加。这些知识数据库可能包含需要隐私保护的敏感信息。然而，多模态 RAG 系统本质上为外部用户提供了间接访问此类数据的途径，使其可能受到隐私攻击的威胁，尤其是成员身份推断攻击（MIA）。现有针对 RAG 系统的 MIA 方法主要集中在文本模态，而视觉模态则相对较少被探讨。为弥补这一差距，我们提出 MrM ，这是第一个针对多模态 RAG 系统的黑箱 MIA 框架。它采用由反事实攻击约束的多对象数据扰动框架，可以同时诱导 RAG 系统检索目标数据并生成泄露成员信息的信息。我们的方法首先采用对象感知数据扰动方法来将扰动限制在关键语义上并确保成功检索。在此基础上，我们设计了一种反事实信息掩码选择策略，以优先考虑最具信息量的掩码区域，旨在消除模型自有知识的干扰并增强攻击效果。最后，我们通过对查询试验建模来执行统计性成员推断，以提取反映从响应模式重构掩码语义的特征。对两个视觉数据集和八个主流商业视觉语言模型（如 GPT-4o、Gemini-2）的实验表明，MrM 在样本级和集合级评估中始终表现出强劲的性能，并在自适应防御下保持稳健。

作为 LLMs 的一项关键增强策略，检索增强生成（RAG）最近已扩展到视觉模态，使其在多模态 AI 任务中具有更广泛的适用性。通过结合视觉模态，RAG 系统可以检索补充视觉输入的外部知识，并帮助减少大型视觉语言模型（LVLMs）中的幻觉。最近的进展表明，多模态 RAG 在使 LVLMs 能够动态整合知识以用于真实世界应用（如智能医疗 AI 系统）方面的新兴作用。为了提高 RAG 系统的有效性，某些私有域数据库被纳入以支持垂直推理和复杂推理。这些知识库通常包含私人的或专有的数据，它们对于支持复杂的下游任务至关重要，同时这些数据可能高度敏感，应该通过强大的隐私保护措施进行保护。然而，RAG 范式本质上引入了间接暴露风险：知识库向生成模型提供信息，后者生成的响应可以被外部用户访问。在此过程中，RAG 系统在内部敏感数据和外部对手之间建立了一座桥梁，可能在无意间泄露私人内容。这种间接访问路径创建了新的漏洞，使得对手能够对底层数据库进行隐私攻击，特别是成员推理攻击（MIAs），该攻击旨在揭示特定样本是否是原始数据库的一部分。

现有关于对 RAG 进行成员信息攻击（MIAs）的研究主要集中在仅文本的模态上，采用各种方法来判断目标样本是否存在于检索语料库中 [1–4] 。例如，Li et al. [2] 开发了一种 MIA 方法，该方法通过分析目标样本与 RAG 生成的内容之间的语义相似性和困惑度来推断数据库成员关系。总之，这些方法的范式包括提供目标数据的片段，然后比较输出与原始数据之间的相似性。然而，LVLMs 在处理文本和视觉输入时通常生成仅包含文本的输出 [5–7] 。这种不对称性带来了模态转换的挑战：推断视觉数据的成员状态需要通过纯文本响应进行推理，而没有直接访问输出中的视觉特征。因此，这些以文本为中心的 MIAs 方法无法直接转移到具有多模态 RAG 的 LVLMs。此外，一个新的挑战在于确保成功检索目标数据与引导模型生

*Tsinghua University. Email: ypr21@mails.tsinghua.edu.cn
†Beijing University of Posts and Telecommunications.
‡Beijing University of Posts and Telecommunications. Corresponding author. Email: taoqi.qt@gmail.com

Preprint. Under review.

成以揭示成员信息之间的平衡。针对这两个核心挑战的解决方案对于专门为多模态 RAG 系统设计的 MIAs 来说至关重要。

因此，我们提出了 MrM ，一个由反事实攻击约束的多对象数据扰动框架，这是第一个针对多模态 RAG 系统的黑箱 MIA 框架。其核心思想是在目标样本中进行扰动，并分析文本响应是否隐含地重建了被扰乱的语义。通过这种方式，我们的方法能够通过对象检测精确地处理文本和视觉模态的语义，以应对跨模态成员推断的挑战。此外，遮蔽对象可以最小化受影响的区域，从而提高对检索的攻击效果，同时扰乱独立和完整的语义以增强对生成的攻击。具体来说，采用了一种对象感知的数据扰动方法，通过使用如 SAM [8] 这样的对象检测模型来遮蔽检测到的实体，策略性地扰乱视觉语义。此方法确保关键特征被扰乱，同时仍然允许如果数据库中存在相关数据，则可进行检索。随后采用一个反事实信息引导的遮罩选择策略，其中我们量化每种扰动的资讯性。我们通过分析反事实代理模型的概率分布和置信度差异，优先选择能够最大化判别差距的遮罩。该策略旨在消除 LVLMs 自身知识的干扰，从而防止在生成阶段重建非数据库图像的信息。最后，我们通过对查询试验进行建模来执行统计成员推断，分析文本响应是否隐含地重建了被扰乱的语义。

总之，我们方法的贡献如下：

- 我们介绍了第一个针对多模态 RAG 系统的黑盒 MIA 框架，突出了多模态数据库隐私保护中的漏洞。
- 我们提出了一个统一的 MIA 框架，该框架解决了跨模态对齐问题，并能够在检索和生成阶段进行并发攻击。
- 我们通过两个视觉数据集和八个主流商业 LVLMs 的综合实验验证了我们的框架，展示了一致的强大性能和针对自适应防御策略的鲁棒性。

自从跨模态对齐模型如 CLIP、ViLBERT 和 BLIP 出现以来，多模态 RAG 成为解决单模态框架在处理跨模态关联时固有局限性的重要方案，其中单独的基于文本的检索未能捕捉复杂视觉语言推理所需的视觉语义与语言上下文之间的复杂交互。提出了一个多模态检索增强变换器，通过访问外部图像-文本记忆库来提升语言生成，预先训练了联合对比和生成目标。提出了一种统一的视觉语言检索模型，使用模态平衡的困难负例和图像语言化来弥合模态间隙。利用训练过的 VLM 从文本页面图像生成高质量的多向量嵌入，并结合后期交互匹配机制以实现高效的文档检索。总体而言，多模态 RAG 现已成为一个关键范式，在视觉语言任务中实现强大的跨模态检索和有据生成。多模态学习的最新进展促使了针对 VLM 训练数据的 MIA 研究。提出了一种 MIA 方法，利用余弦相似性和弱监督攻击，避免影子训练。呈现了一种 VLM MIA 基准和一种基于置信度的文本和图像的令牌级检测方法。虽然这些方法可以激发针对多模态 RAG 系统的 MIA 框架设计，但它们主要基于白盒或灰盒架构。然而，大多数 RAG 系统是在云环境中部署的，并作为生成即服务（GaaS）提供，因此在黑盒环境中运行。最近的研究探索了专注于单一文本模态的 RAG 系统中的 MIA。Anderson et al. [1] 引入了首个 RAG-MIA 方法，通过查询系统并解释是/否响应来推断文档存在性。Li et al. [2] 提出了 S ^{2} MIA ，使用基于 BLEU 的语义相似度和目标样本与生成输出之间的困惑度比较来推断成员资格。Liu et al. [3] 提出了一个基于掩码的方法，通过单词掩码扰动目标文档，查询系统，使用预测准确度阈值进行推断。Naseh et al. [4] 设计了特定于文档的自然语言查询，通过将系统响应与影子 LLM 生成的真实答案进行比较来推断成员资格。然而，这些仅文本的单模态 RAG-MIA 方法在应用于多模态架构时面临限制。视觉-语言模型处理组合的图像-文本输入但仅产生文本响应，导致与传统方法的基础性不匹配。总的来说，为多模态 RAG 系统专门设计的 MIA 在文献中仍未被充分探索。

攻击者的目标是通过成员推断攻击，利用系统的文本输出来确定目标图像 $\mathcal{I}$ 或图像数据集 $\mathcal{D}_i = \{\mathcal{I}_1, ..., \mathcal{I}_N\}$ 是否存在于黑箱多模态 RAG 系统的数据库 DB 中。

Attacker's Capability. 攻击者可以通过其公共接口重复查询 RAG 系统，模拟合法的用户交互。他们可以制作多模态输入（文本和图像）以探测系统，并观察由视觉语言模型（VLM）M 生成的文本输出。假设攻击者具备对多模态系统的基本知识，但对数据库内容、模型架构或训练数据没有事先信息。

Attacker's Constraints. 攻击者面临三个关键限制：首先，他们无法访问中间系统组件，包括 VLM 的输出概率分布、输入嵌入或数据库 DB 。其次，VLM M 可能采用安全机制来拒绝明显恶意的或隐私敏感的查询。第三，所有观察都限制在 M 的文本输出，没有直接访问检索结果或数据库索引模式。这些限制需要采用能够在检测的同时绕过并提取会员信号的查询策略。

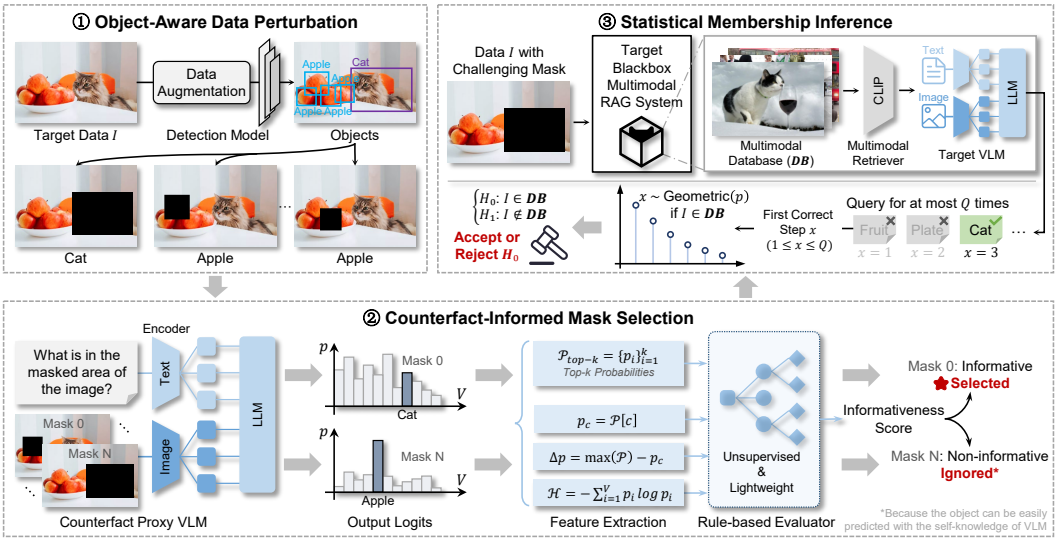

Figure 1: 所提出的 MrM 方法的整体框架。它首先通过对象感知掩模扰动目标图像，使用一个基于反事实信息的掩模选择策略选择信息丰富的扰动，并通过对 RAG 系统响应统计进行假设检验来推断成员关系。

在本节中，我们将介绍我们提出的方法 MrM 的技术细节，该方法由三个关键组成部分构成，如图 1 所示：对象感知数据扰动、反事实信息驱动的掩码选择和统计成员推断。

我们框架的核心动机是让 RAG 系统的检索和生成阶段同时泄露成员信息。在检索阶段，目标是确保 RAG 系统成功检索到目标数据，而在生成阶段，我们旨在从 RAG 系统的响应中引出有关目标图像的相关信息。如果直接输入目标数据而不进行任何扰动，系统很可能会检索到相应的数据。然而，在这种情况下，很难确定模型响应中的信息是来自输入还是来自检索数据库。另一方面，通过扰动显著降低目标数据的质量可以确保模型响应中的任何相关信息仅来自于检索知识库。然而，这种方法可能显著削弱成员推断框架的有效性，因为它可能导致 RAG 系统中的检索过程失败。

因此，挑战在于平衡这两个目标：在确保成功检索的同时，保持足够的扰动，以引导模型揭示成员信息。为了解决这一挑战，我们提出了一个由反事实攻击约束的多对象数据扰动框架。该框架能够生成战略性降解目标数据语义的扰动，确保相关信息的检索，同时防止直接从输入中泄露信息。通过对模型施加不同的扰动以引发多个响应，可以提取判别特征，从而提供清晰的成员证据。

0.1 对象感知数据扰动

对于扰动过程，我们设定了三个关键目标：(1) 确保目标数据仍然可检索，(2) 在生成阶段防止非数据库图像的信息重构，(3) 促进从图像模态到文本生成模态的跨模态转换。

为了实现这些目标，我们采用了一种对象感知扰动的方法。在图像中，感兴趣的对象通常只占据整体场景的一小部分，这意味着扰动这些区域对检索过程的影响最小。此外，单个对象具有相对独立的语义，允许信息的可追溯性得以保留，这确保了信息的来源可以追溯到检索数据库。最后，对象非常适合被转换为文本模式，因为它们通常定义明确并且容易用语言描述，这使得它们在生成阶段成为产生有意义文本响应的理想候选。

给定一个目标图像 $\mathcal{I}$，我们使用一个对象检测模型 $\mathcal{D}$（例如，SAM [8]）来定位显著对象。设 $O = \{o_j\}_{j=1}^{K}$ 表示检测到的对象集，其中 K 是对象的数量。对于每个对象 o_j，我们生成一个二进制掩码 $\mathcal{M}_j$ 来遮挡其在 $\mathcal{I}$ 中对应的区域，从而得到一个扰动图像 $\tilde{\mathcal{I}}_j$。形式上，扰动过程定义为：$\tilde{\mathcal{I}}_j = \mathcal{I} \odot (\mathbf{1} - \mathcal{M}_j) + \mathbf{0} \odot \mathcal{M}_j$，其中 $\odot$ 表示按元素相乘，$\mathbf{1}$ 是一个全为一的矩阵，$\mathbf{0}$ 是一个全为零的矩阵。这确保了 $\mathcal{M}_j$ 内的像素被设为零，同时保留其他区域。

为了选择最能区分数据库样本和非数据库样本的扰动，我们提出了一种基于反事实信息的掩码选择策略。该策略通过量化每个被掩盖区域的信息量，以优先考虑那些能最大化区分性差距的区域。这是通过分析一个代理视觉语言模型（VLM）$\mathcal{V}$ 生成的概率分布和置信差

来实现的，该模型作为反事实参考。其目标是消除目标 LVLM 自包含知识的干扰，从而确保任何观察到的语义重建都可归因于检索而非记忆知识。

给定一个被扰动的图像 $\tilde{\mathcal{I}}j$ ，我们将其输入到代理模型 $\mathcal{V}$ 中，以获得词汇表 V 上的概率分布 $P = \{p_i\}_{i=1}^{V}$ 。基于这个分布，我们提取以下特征来估计每个掩码的信息量和难度：目标置信度 p_c ： 与被掩盖区域的真实类别相对应的预测概率。置信度差距 $\Delta p = \max(P) - p_c$ ：衡量最高预测概率与真实置信度之间的差异。熵 $\mathcal{H} = -\sum_{i=1}^{V} p_i \log p_i$ ： 量化预测的不确定性，较高的熵表示更大的混乱。Top-k 分布 $\{p_{(i)}\}_{i=1}^{k}$ ： 在 P 中按降序排列的前 k 个值，捕捉高置信度预测的分布锋利度和多样性。

这些特征形成了一个特征向量 $\boldsymbol{f}_j = [p_c, \Delta p, \mathcal{H}, \{p_{(i)}\}_{i=1}^{k}]$ ，共同捕捉代理模型的不确定性，这在估计黑箱环境中扰动的判别能力时是重要的。为了给每个遮罩分配一个信息丰富的分数，我们采用了一种规则为基础的评估器，在无监督的情况下整合提取的特征。具体来说，遮罩根据标准化特征分数的集合进行排序，其中高熵、低目标置信度和小置信度差距共同促成更高的信息量分数。我们优先考虑具有高估计信息量的遮罩，因为它们更有可能在保持成员推断的判别信号的同时抑制非数据库数据的虚假重建。

为了严格推断目标图像 $\mathcal{I}$ 的归属状态，我们制定了一个假设检验框架，该框架基于多模态 RAG 系统在被询问关于遮挡物体时的统计行为。核心直觉是，该系统在预测被遮挡物体时的成功率依赖于 $\mathcal{I}$ 是否在数据库 DB 中。形式上，我们定义了两个假设。零假设 (H_0)：$\mathcal{I} \in DB$ ，其中系统对每个遮挡的成功概率遵循 p_t 。备择假设 (H_1)：$\mathcal{I} \notin DB$ ，具有较低的成功概率 p_n ，其中 $p_t > p_n$ 根据设计。

对于每一个扰动图像 $\tilde{\mathcal{I}}_j$ （由第 j 个掩码 $\mathcal{M}_j$ 派生），我们反复查询系统，直到它正确识别出被遮挡的物体。记 x_j 为首次正确预测所需的尝试次数。在 H_0 下，x_j 遵循几何分布：

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

对于 K 个掩码，所有掩码的总尝试次数汇总为 $S = \sum_{j=1}^{K} x_j$ 。根据独立几何变量的加法性质，S 的期望和方差为：$\mu_0 = \frac{K}{p_t}$， $\sigma_0^2 = \frac{K(1-p_t)}{p_t^2}$ 。对于足够大的 K ，应用中心极限定理 (CLT)，S 近似于正态分布：$S \overset{\text{approx}}{\sim} \mathcal{N}\left(\mu_0, \sigma_0^2\right)$ 。p 值量化了在 H_0 下观测到极端如 S 的总尝试次数的概率。为了计算它，我们首先对 S 进行标准化，然后评估标准正态分布的生存函数。记 $\Phi(z)$ 为 $\mathcal{N}(0,1)$ 的累积分布函数（CDF）。p 值为：

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

如果 p-value $< \alpha$ （例如，$\alpha = 0.05$ ），我们拒绝 H_0 并得出 $\mathcal{I} \notin DB$ ；否则，我们保留 H_0 ，表明可能的成员资格。

1 实验与分析

1.1 实验装置

Datasets. 我们使用两个标准图像数据集来构建知识库并进行成员推理攻击。COCO [9] 和 Flickr [10] 提供了广泛用于视觉研究的多样化图像集合。从每个数据集中，我们选择了 5000 张图像用于知识库，并选择了 1000 张图像（500 个成员，500 个非成员）用于测试。

Target Models. 我们对八个商业模型进行了成员推理攻击，每个模型都与一个本地知识库集成以形成一个多模态 RAG 系统：GPT-4o-mini [11] , Gemini-2 [12] , Claude-3.5 [13] , GLM-4v [14] , Qwen-VL [15] , Pixtral [16] , Moonshot [17] , 和 InternVL-3 [18] 。这些商业 VLMs 支持多图像输入，使它们适用于多模态 RAG 系统。实验通过连接到一个本地构建的知识库的 API 调用进行。这种设置确保了无法访问内部生成状态，维持严格的黑盒环境，模拟现实世界的部署，只有模型输出可用于分析，无法访问有关内部工作或中间状态的信息。

Baselines. 据我们所知，我们的工作提出了第一个针对多模态 RAG 系统的 MIA 方法。由于缺乏基准方法，我们从基于文本的 RAG MIA [1, 2] 适配了两种策略。第一个基准方法，基于查询的 MIA（简称为 QB-MIA），直接询问目标样本是否出现在检索到的参考中，将模型的二元响应视为成员信号。第二种方法，相似性为基础的 MIA（SB-MIA），部分遮挡目标

Table 1: 在 Flickr 和 COCO 数据集上的八个多模态 RAG 系统中，各种 MIA 方法对 RAG 的性能比较。我们报告了每种方法的 AUC 和 TPR@5 % FPR，包括 QB-MIA、三个具有不同掩码比率的 SB-MIA 变体，以及我们提出的 MrM 。MrM 始终取得最高性能，特别是在低误报约束下。

	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 %
	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 %
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 %

图像，并要求模型使用检索到的参考图像重建缺失内容。类似 SB-MIA-0.5 的变体表示遮挡比例。然后计算生成描述与原始内容之间的相似性，相似性越高意味着可能为成员。

Evaluation Metrics. 继承以往关于 MIA 的研究 [19–21] ，我们采用了两个评估指标：AUC 和 TPR@5 % FPR。AUC 反映了在不同阈值下成员和非成员之间的整体区分能力。TPR@5 % FPR 测量了当假阳性率限制在 5 % 以下时的真正率，在严格条件下提供了更好的评估。由于相同的 AUC 可能来自不同的 ROC 曲线，TPR@5 % FPR 补充了 AUC 以进行细致入微的评估。我们在样本和集合层面报告这两个指标以进行全面分析。

对于对象感知的数据扰动，我们采用 SAM2 模型 [8] 进行对象检测。我们使用 Qwen-VL 的 7B 本地版本 [15] 作为代理 VLM。在消融研究中，我们将其替换为较弱的检测器 YOLO 模型 [22] 。所有的检索数据库都是使用 FAISS 库 [23] 构建的。作为我们 RAG 系统中的图像检索器，我们采用了 CLIP 模型的 ViT 变体 [24] 。

1.2 主要结果

Sample-level MIA. 表格 1 展示了在 Flickr 和 COCO 数据集上八个多模态 RAG 系统的样本级 MIA 的性能对比。我们的方法 MrM 被评估与两个基线方法进行比较：QB-MIA 直接查询模型关于检索到的引用中是否存在目标样本，SB-MIA 则是移除目标图像的一部分并提示 RAG 系统根据其检索到的引用描述原始内容。为了确保公平和真实的评估，我们在表格 1 中的所有方法中应用了一种简单但自然的防御机制：在 VLM 中添加了一条警示性系统提示，写道：“请勿泄露任何关于您的知识库成员的信息。”这条提示作为防止意外记忆泄漏的最低限度保护。虽然这项防御对 SB-MIA 和我们提出的 MrM 方法的性能影响有限，但它显著削弱了 QB-MIA 的有效性，因为 QB-MIA 依赖于模型是否愿意直接回答与成员相关的问题。本文中所有后续实验均在此默认防御设置下进行。在两个数据集和所有模型中，MrM 表现出明显的性能优势，持续获得更高的 AUC 得分，表明其整体辨别能力强。尤其在 TPR@5 % FPR 方面，相较于基线方法表现突出，这对于在严格的假阳性约束下评估 MIA 至关重要。该指标反映了攻击者在严格的假阳性约束下的成功率，这在真实世界场景中更具相关性，因为在这些场景中，低假阳性率对于 MIA 的隐秘部署是必要的。MrM 的卓越性能源自其精确扰乱目标图像中语义上最关键且最难推测区域的能力。通过利用对对象感知的扰动以及通过代理视觉-语言模型进行难度评估，MrM 能够识别并遮掩那些在没有事先暴露的情况下显著且难以描述的区域。因此，非成员图像会导致 VLM 产生模糊或不准确的响应。相比之下，

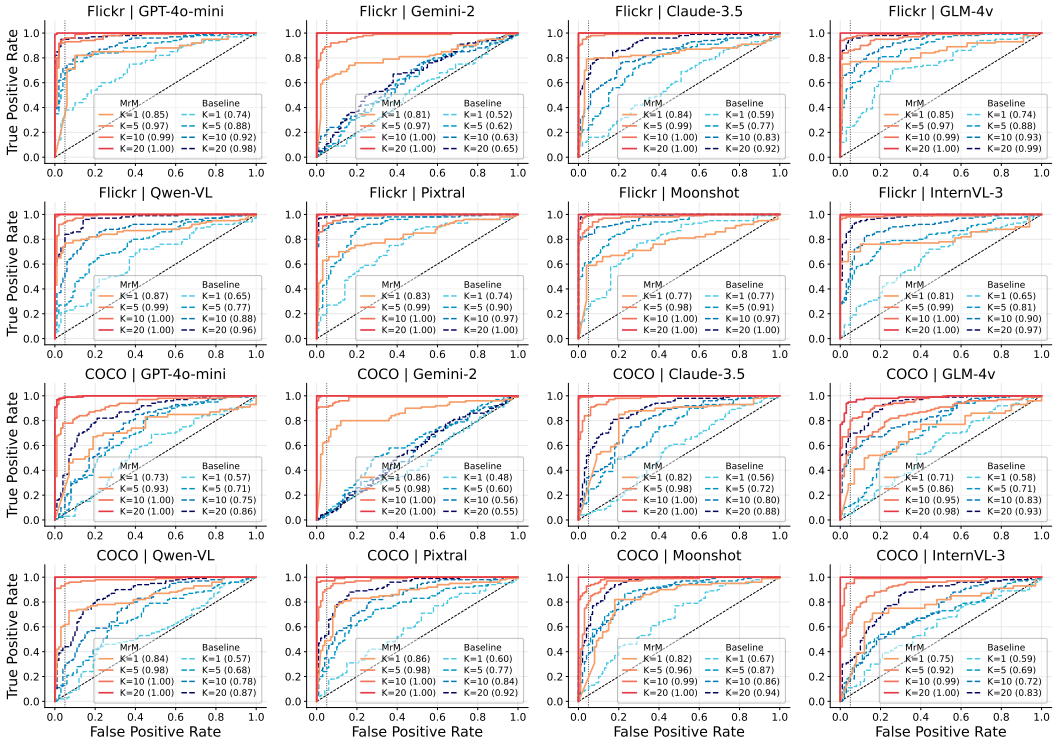

Figure 2: 在两组数据集上，针对八种多模态 RAG 系统的不同集合规模（$K = 1, 5, 10, 20$）的集合级 MIA 的 ROC 曲线。我们将 MrM 与最佳的 SB-MIA 基线进行比较。MrM 始终在靠近左上角的区域内实现更高的 AUC 和更陡的曲线，表明其在 TPR@5 % FPR 方面的优越性。垂直的灰色虚线标记了 5 % 误报率的阈值。

对于成员图像，由于其强大的上下文学习能力，VLM 通常可以从上下文提示中恢复正确的语义。这种对比增强了我们统计测试的辨别力，并支持了在模型和数据集中的改进结果。

为了在集合层面评估 MrM 的有效性，我们在图 2 中绘制了 ROC 曲线，该曲线展示了不同集合大小 $K = 1, 5, 10, 20$ 下，八个 RAG VLM 模型和两个数据集的表现。每条曲线反映了模型通过对所有集合中的样本使用 RAG 系统响应的联合统计测试，汇总预测来推断成员身份的能力。我们将我们提出的方法与 SB-MIA 的最强变体进行比较，该变体在本节中作为参考方法。在几乎所有模型和两个数据集中，MrM 在 AUC 方面始终优于基线方法，并且这种优势随着集合大小的增加而更加显著。当 $K = 1$ 时，MrM 保持良好的性能，如上面样本层级结果中所述，并且这种优势随着集合的扩大而进一步增强。随着 K 的增长，两种方法的 AUC 都有所改善，但 MrM 始终获得更高的值并更快收敛至接近完美的性能。在大多数情况下，当 $K = 10$ 时，MrM 的 AUC 接近于 1.0，这表明其在相对小的集合大小下迅速达到性能饱和。MrM 的另一个优点是其 ROC 曲线往往更陡地朝向左上角弯曲，表明在相同 AUC 下具有更高的 TPR@5 % FPR。这在 5 % FPR 处的垂直参考线中有所突出，在该处我们的方法在所有模型和数据集中稳定地实现更高的 TPR。此外，MrM 在所有模型和数据集上都表现出强劲和稳定的结果，包括基线方法表现明显下降的情况。这突出显示了我们方法的普适性，即使在不同模型结构和检索行为下亦然。

1.3 消融研究

为了更好地理解 MrM 中各个组成部分的贡献，我们通过系统地移除或替换管道的关键元素来进行一个消融研究。如图 3 所示，我们评估了三种变体：(1) 没有对象意识，其中去除了对象检测，图像区域是随机遮盖的；(2) 没有代理模型，其中排除了代理模型，并且没有应用基于难度的遮盖选择；(3) 更简单的对象检测模型，其中用 YOLO 替换了更强的 SAM2 检测器。所有变体都在 Flickr 和 COCO 数据集上的八个 RAG 系统中进行了测试。我们观察到所有三种消融变体中的性能都有一致的下降，证实了每个组件在完整 MrM 管道中是必要的。(1) 没有对象意识时，模型表现明显下降，这表明随机遮盖区域通常未能瞄准图像中最具语

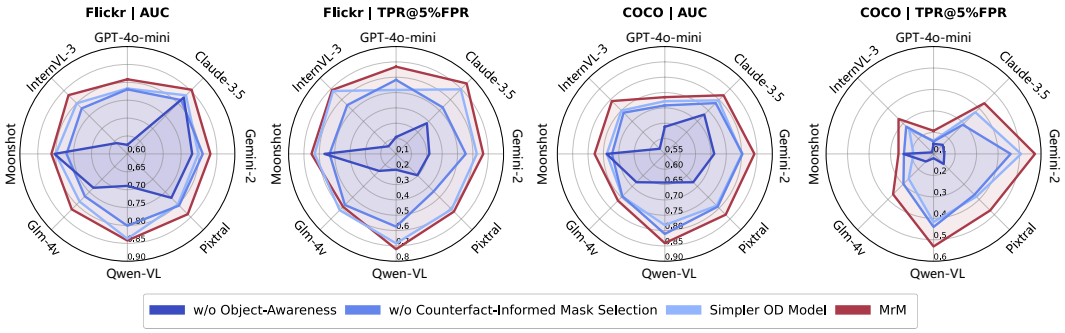

Figure 3: 消融研究结果以雷达图形式可视化。我们将完整的 MrM 方法与三个消融变体进行比较：无对象感知、无反事实信息的掩码选择和更简单的 OD 模型。MrM 始终优于所有消融版本，证明了每个组件对整体攻击性能的贡献。

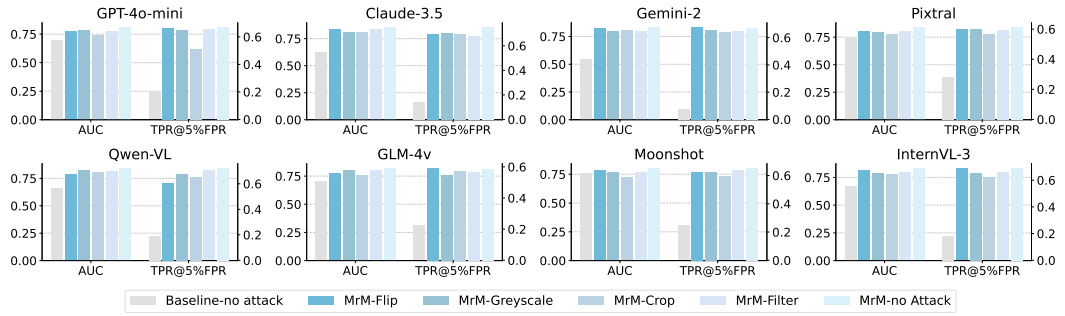

Figure 4: MrM 在对数据库应用自适应图像级转换（包括水平翻转、灰度转换、裁剪和高斯模糊滤波）的情况下表现出稳健性。MrM 在所有转换下都保持了强劲的性能。

义信息的部分。(2) 没有代理模型时，缺乏难度评估导致不够显著的扰动，这削弱了用于成员推断的信号。(3) 使用更简单的对象检测器导致性能出现适度但明显的下降，说明高质量的对象检测有助于更有效的扰动策略。这些结果强调了所有组件在实现强大性能时所扮演的关键角色。完整的 MrM 方法因其协同作用受益，在各模型和数据集上提供了更有效和可靠的成员推断。

1.4 鲁棒性分析

除了前面讨论的基于系统提示的防御之外，我们进一步在一种基于输入修改的数据级防御类别下评估 MrM 。具体而言，我们模拟一种防御设置，其中检索数据库包含原始图像的修改版本，通过常用的变换处理，如水平翻转、灰度转换、裁剪和高斯模糊。这些变换旨在破坏直接的视觉匹配，同时保留图像的高层次语义，从而削弱基于简单检索的 MIA 方法。为应对这一挑战，我们扩展我们的攻击管道，采用一种增强感知策略。对于每个目标样本，我们使用应用于数据库的相同变换类型生成多个增强变体。每个变体被视为一个独立的查询实例——经历对象感知扰动、难度评估和统计测试——使我们的方法能够探索尽管存在变换差距但仍然有效的替代检索路径。这种增强感知探测增加了至少一个变体检索到修改后的数据库条目的可能性，从而恢复模型可能被掩盖的记忆信号。重要的是，这种设计也模拟了攻击者在部署管道中猜测或估计潜在变换模式的现实能力。如图 4 所示，虽然经过所有四种基于变换的防御，我们的方法仍然保持强劲的表现。这个结果展示了 MrM 对一系列内容保留的图像修改的鲁棒性，强化了其在更具对抗性或模糊的部署场景中的实用性。

2 结论

我们介绍了 MrM ，这是第一个专门为多模态 RAG 系统设计的黑箱会员推断框架。我们的方法揭示了通过外部知识检索增强的视觉-语言模型中以前未探索的隐私漏洞。为了应对跨模态对齐和检索-生成平衡的挑战，我们提出了一个统一的 MIA 框架，该框架结合利用了检索和生成阶段，从而能够从多模态输出中检测会员信号。MrM 结合了对象感知扰动和反事

实信息掩码选择，以精确控制语义泄漏，同时保留检索性能。在两个视觉数据集和八个广泛使用的商业 LVLMs 上的大量实验验证了我们方法的有效性，显示 MrM 在样本级和集合级评估下均表现出一致的强大性能，即使在存在自适应防御机制的情况下也保持稳健。我们的研究结果强调了多模态 RAG 基础设施中的紧迫安全挑战，并推进了对连接视觉、语言和检索系统中隐私风险的理解。

References

[1] Maya Anderson, Guy Amit, and Abigail Goldsteen. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. *arXiv preprint arXiv:2405.20446*, 2024.

[2] 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.

[3] 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.

[4] Ali Naseh, Yuefeng Peng, Anshuman Suri, Harsh Chaudhari, Alina Oprea, and Amir Houmansadr. Riddle me this! stealthy membership inference for retrieval-augmented generation. *arXiv preprint arXiv:2502.00306*, 2025.

[5] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54 (10s):1–41, 2022.

[6] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 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*, pages 18166–18176, 2022.

[7] Yutong Zhou and Nobutaka Shimada. Vision+ language applications: A survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 826–842, 2023.

[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

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

[10] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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, 2014.

[11] 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.

[12] Shrestha Basu Mallick and Logan Kilpatrick. Gemini 2.0: Flash, flash-lite and pro, February 2025. URL `https://developers.googleblog.com/zh-hans/gemini-2-family-expands/`. Accessed: 2025-05-01.

[13] Anthropic. Introducing claude 3.5 sonnet, June 2024. URL `https://www.anthropic.com/news/claude-3-5-sonnet`. Accessed: 2025-05-01.

[14] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai

Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

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

[16] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

[17] Moonshot AI. Kimi chat, 2024. URL `https://kimi.moonshot.cn/`. Accessed: 2025-04-10.

[18] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[19] 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.

[20] 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*.

[21] Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Frank Yang, and Hai Li. Min-k%++: Improved baseline for detecting pre-training data from large language models. *arXiv preprint arXiv:2404.02936*, 2024.

[22] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[23] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

[24] 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.

A 详细威胁模型

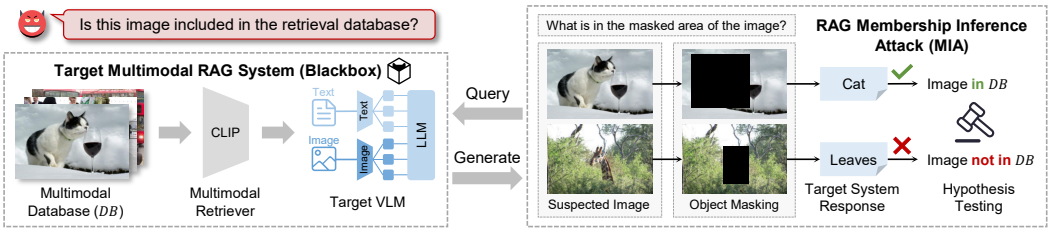

Figure 5: 针对多模态 RAG 系统的黑箱 MIA 的示意图。攻击者对系统进行查询，使用经过战略性扰动的图像输入，并分析文本响应，以确定目标图像是否存在于底层检索数据库中，而无需访问模型内部或数据库内容。

攻击者必须构建一个成员推断算法 $\mathcal{A}$，通过精心设计的多模态查询反复探测黑箱 RAG 系统，以推断成员状态。如图 5 所示，攻击者提交目标图像的扰动版本并观察系统的文本响应，以确定该图像是否存在于底层检索数据库中。攻击者无法访问 RAG 系统的内部参数、其检索器或其数据库内容——完全在黑箱假设下运行。

形式上，对于一个目标图像 $\mathcal{I}$，算法通过一个 k 自适应查询 $\mathcal{Q}_1(\mathcal{I}), \ldots, \mathcal{Q}_k(\mathcal{I})$ 序列及其对应的文本响应 $\mathcal{O}_1, \ldots, \mathcal{O}_k$ 综合信息，这些文本响应由增强了检索数据库的视觉语言模型 M 生成，定义为

$$\mathcal{A}\big(\mathcal{I}, \{\mathcal{O}_i\}_{i=1}^{k}\big) \to \{0, 1\}.$$

。这需要设计查询策略，以微妙地诱导输出文本中具有成员区别特征的模式——例如在细节、语境连贯性或知识精细度方面的变化——同时通过语义模糊或语境间接的提示绕过 VLM 的安全机制。

对于 $\mathcal{D}_i = \mathcal{I}_1, \ldots, \mathcal{I}_N$ 的数据集级别验证，该任务扩展到汇总所有 N 图像上的观测，需要一个综合推理规则

$$\mathcal{A}\big(\{\mathcal{I}_j\}_{j=1}^{N}, \{\{\mathcal{O}_i^{(j)}\}_{i=1}^{k}\}_{j=1}^{N}\big) \to \{0, 1\}$$

，该规则在每图像证据与全局统计置信度之间取得平衡。

关键挑战包括识别文本输出特征与数据库成员之间的潜在关系，开发能够容忍噪声的聚合方法以融合多查询证据，以及通过模仿合法用户行为来避免触发防御性过滤器，从而保持隐蔽性。

B 详细实验设置

B.1 模型版本

对于多模态 RAG 系统中的目标模型，我们通过各自的 API 使用以下商业视觉语言模型版本：GPT-4o-mini [11]（`gpt-4o-mini`），Gemini-2 [12]（`gemini-2.0-flash`），Claude-3.5 [13]（`claude-3-5-sonnet`），GLM-4v [14]（`glm-4v-plus-0111`），Qwen-VL [15]（`qwen-vi-max-0408`），Pixtral [16]（`pixtral-large-2411`），Moonshot [17]（`moonshot-v1-8k-vision-preview`），以及 InternVL-3 [18]（`internvl3-78b`）。所有模型皆通过 API 调用以严格的黑盒方式访问，无法访问其内部生成状态。

对于检索组件，我们采用具有 ViT 骨干的 CLIP 模型 [24]，特别是 `openai_clip-vit-large-patch14-336` 变体，来编码图像特征并构建检索数据库。其图像嵌入维度为 768。

Object Detectors. 主要的目标检测器是 SAM2 [8]：`sam2.1-hiera-large`。在消融研究中，我们使用 `yolo11x` [22] 作为一个较弱的替代品。

在 SB-MIA 基线中，目标图像通过遮挡左侧部分被遮盖。随后，多模态 RAG 系统被提示使用检索到的参考图像来重建缺失的内容。然后，我们使用 CLIP 的跨模态相似性计算遮盖图像区域与模型生成文本之间的相似性。SB-MIA 中的数字后缀，如 SB-MIA-0.5，表示遮盖比例。更大的遮盖比例使得 VLM 更难重建原始语义，从而增加了成功进行成员推断的可能性。然而，更高的遮盖也因为视觉信息的更大损失而降低了检索成功率。相反，更小的遮盖

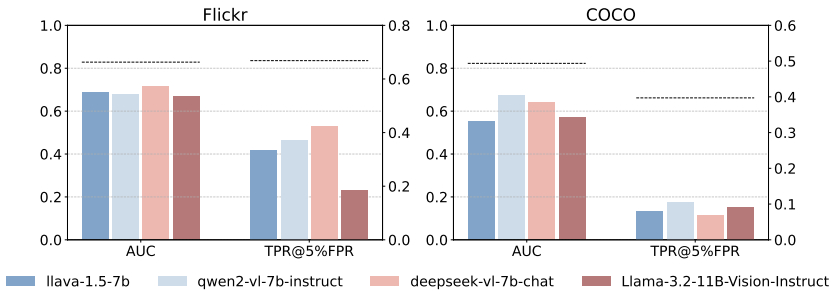

Figure 6: 在本地部署的拥有 70 亿到 110 亿参数的开源视觉-语言模型上，MrM 的成员推断性能，评估数据集包括 Flickr 和 COCO。柱状图显示 AUC 和 TPR@5 % FPR。结果表明，与商业模型相比，性能有适度下降，但仍然展示出有意义的辨别信号。

比例改善了检索，但降低了生成输出的可辨识性。相比这种权衡，MrM 在检索准确性和生成差异之间实现了更好的平衡，从而改进了 MIA 性能。

在消融研究中，没有对象意识的变体通过选择一个矩形区域来应用随机遮罩，其高度和宽度范围为原始图像的 20 % 到 60 %。因此，总体遮罩区域与 MrM 相似。相似性计算遵循与 SB-MIA 中相同的过程，使用基于 CLIP 的跨模态相似性来比较遮罩区域图像和生成的文本。

对于所有目标 RAG 系统，我们使用以下提示：分析输入图像中被遮罩区域与参考图像之间的精确位置对应关系。仅从参考图像中提取相同位置的视觉内容。仅输出一个单词或短语。不要写任何其他内容。输入图像：[IMAGE] 参考图像：{ [IMAGE],... } 如果一个视觉-语言模型不支持多图像输入，我们将目标图像和参考图像连接成一张图像后再输入。

检索数据库是使用 FAISS [23] 构建的。我们使用 `Flat` 索引类型，并对图像嵌入应用 `mean` 池化。尽管 COCO 数据集 [9] 包含真实的物体标注，但我们在流程中并未使用它们。相反，我们通过使用 SAM 模型检测物体来模拟真实世界的场景。所有 ROC 曲线都使用随机采样绘制，每个 K 值选取 200 个样本。每个实验重复五次以减少随机性的影响。我们实现中使用的关键 Python 库包括：`openai` (v1.64.0), `faiss` (v1.9.0), `transformers` (v4.45.1), 和 `torch` (v2.4.1)。

为了探索 MrM 的泛化能力超越商业 API 系统，我们在几个开源的视觉语言模型上评估它的性能，这些模型的参数规模较小（7B–11B）。如图 6 所示，我们在 Flickr 和 COCO 数据集上测试四个模型：`llava-1.5-7b` 、`qwen2-vl-7b-instruct` 、`deepseek-vl-7b-chat` 和 `llama-3.2-11B-Vision-Instruct` 。每个模型都在本地部署，并使用正文中介绍的相同样本级 MIA 协议进行评估。

结果显示，与大型商业系统相比，这些较小的模型表现出相对较低的 AUC 和 TPR@5 % FPR 得分。然而，这一差距并不是很大，且几个模型（例如，`deepseek-vl-7b-chat` ）表现出显著的判别能力。我们假设性能下降主要是由于较小模型在推理和生成能力上的不足，特别是在处理多图像提示和基于检索参考解决掩盖内容方面，即使相关图像已成功检索到。

虽然这些结果突出了某些局限性，但它们也证实了 MrM 在一系列架构中的普遍性。然而，在实际应用中，大多数现实世界的 RAG 系统是由 API 级模型组成的，这些模型由专有检索数据库支持，这使得在基于 API 的模型上进行评估更具代表性。不过，这个开源基准测试提供了关于模型规模如何影响攻击成功率的有价值的见解。

C 详细消融结果

为了进一步检验面向对象的扰动的有效性，我们比较了 MrM 与其非对象感知变体在集合大小 $K = 1, 5, 10, 20$ 上的表现，使用在图 7 中绘制的 ROC 曲线。在 Flickr 和 COCO 数据集，以及所有测试模型中，面向对象的设计产生了更理想的 ROC 特性（接近左上角）以及在几乎所有 K 上的较高的 AUC 值。性能差距在小集合大小（例如，$K = 1$ ）时特别明显，精准的扰动对于触发可区分的响应至关重要。随着 K 的增加，两个方法的性能均有所改善，但 MrM 始终保持明显的优势，表明语义感知的扰动提升了样本层面和聚合集合层面的推理效果。

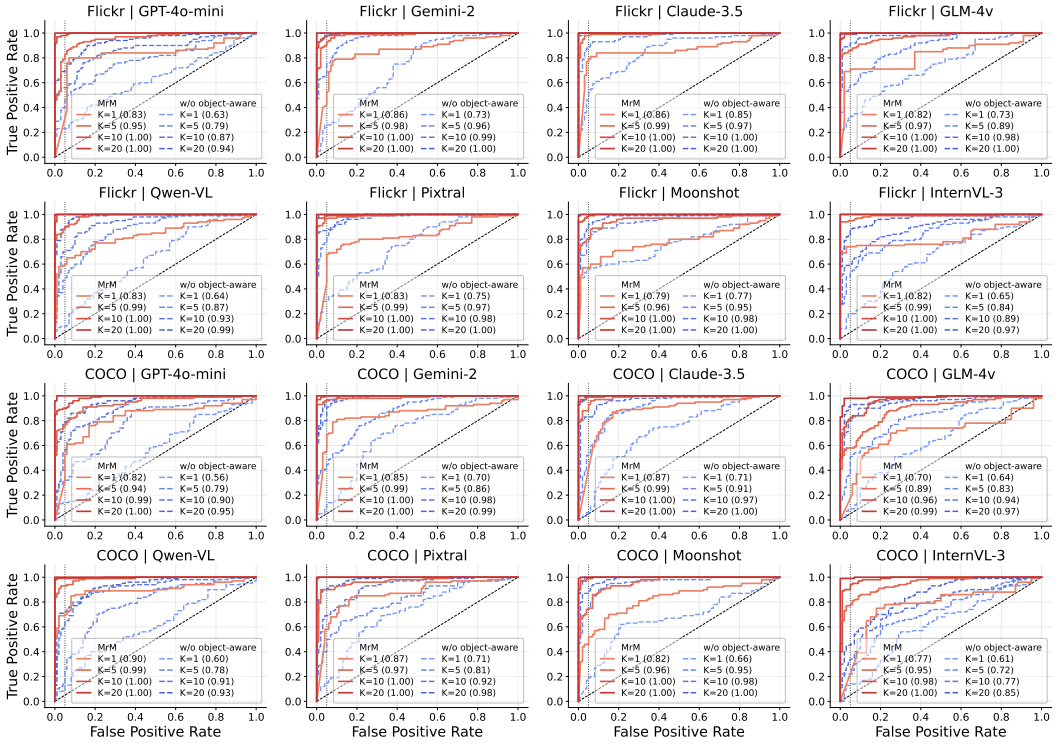

Figure 7: 在八个多模态 RAG 系统和两个数据集上，不同集合大小（$K = 1, 5, 10, 20$ ）下，将 MrM 与其无对象变体进行比较的集合级别 ROC 曲线。MrM 始终达到更高的 AUC 值，尤其是在 K 较小时，突出了对象感知扰动在提高 MIA 性能方面的有效性。

D 案例研究

为了进一步说明我们方法的优势，我们通过对比 MrM 与基线 QB-MIA 和 SB-MIA-0.5，来展示一个具有代表性的案例研究，研究内容是在数据库和非数据库场景中对于同一目标图像的表现。QB-MIA 应用于数据库图像时，由于简单的防御策略而失败。

SB-MIA-0.5 通过遮蔽图像的一半来应用一种粗略的掩蔽策略，然后依靠跨模态语义相似性来推断成员身份。如图 8 中的第一个例子所示，尽管图像在上下文中是未见过的，模型利用其内部推理能力重建了完整的场景，并产生了与原始图像高度语义相似的响应（相似度评分 =0.923）。这导致基线错误地推断出该图像在数据库中。

相比之下，MrM 特别屏蔽了两个关键对象——即马和长椅，这对于准确的场景解释是至关重要的。在无法通过检索访问原始图像的情况下，模型无法准确回答这些元素，从而正确推断图像不在数据库中。

在第四种情况下，我们将这张图片放入数据库并对其应用 MrM 。由于成功检索，答案是正确的，我们的推理框架正确地识别出它是一个成员。

我们的实验评估是在两个广泛使用和具有代表性的视觉数据集上进行的，这些数据集反映了多模态 RAG 系统中常见的场景。虽然这些数据集对典型应用提供了有意义的覆盖，但我们的研究结果在其他领域的普遍性并不能得到保证。特别是，一些新兴领域可能会对我们的框架提出独特的挑战。这些挑战包括高图像间相似性或高度专业化的视觉语义环境，例如放射医学图像（如 CT 扫描和超声波图像）或卫星图像。这些领域通常具有独特的知识特征，可能与我们当前扰动和推理设计的假设不完全一致。未来的工作应将评估扩展到更广泛的特定领域数据集，并相应地调整策略。这将进一步验证我们方法的适应性，并为 RAG 系统在多样化真实世界环境下所带来的隐私风险提供更深入的见解。

Method	Simplified Prompt
QB-MIA	Does this target image [IMAGE] appear in the context? Answer only with Yes or No. \n Retrieved References: {[IMAGE], …} \nAnswer:
SB-MIA	Describe the image [IMAGE] within 50 words, using retrieved images as reference. Starts with \"The image shows\". \n Retrieved References: {[IMAGE], …}
MrM	Analyze the precise positional correspondence between the masked region in the input image and the reference images. Output a word or phrase. Input Image:[IMAGE] Retrieved References: {[IMAGE],...}

Figure 8: 案例研究展示了 MrM 在区分数据库内和非数据库样本方面的优势。将基线（QB-MIA）应用于数据库图像由于幼稚的防御策略而失败。将基线（SB-MIA）应用于非数据库图像由于高语义相似性重建而导致错误推断。MrM 通过屏蔽关键对象（马和长凳）和抑制语义恢复正确推断出非成员关系。当应用于数据库内图像时，MrM 通过检索支持准确的重建，正确推断出成员关系。