

HV-Attack: Hierarchical Visual Attack for Multimodal Retrieval Augmented Generation

Linyin Luo^{1,2}, Yujuan Ding¹, Yunshan Ma³, Wenqi Fan¹, Hanjiang Lai²

¹The Hong Kong Polytechnic University

²Sun Yat-Sen University

³Singapore Management University

luoly36@mail2.sysu.edu.cn, dingyujuan385@gmail.com, ysmasu.edu.sg

wenqifan03@gmail.com, laihanj3@mail.sysu.edu.cn

Abstract

Advanced multimodal Retrieval-Augmented Generation (MRAG) techniques have been widely applied to enhance the capabilities of Large Multimodal Models (LMMs), but they also bring along novel safety issues. Existing adversarial research has revealed the vulnerability of MRAG systems to knowledge poisoning attacks, which fool the retriever into recalling injected poisoned contents. However, our work considers a different setting: **visual attack of MRAG by solely adding imperceptible perturbations at the image inputs of users, without manipulating any other components.** This is challenging due to the robustness of fine-tuned retrievers and large-scale generators, and the effect of visual perturbation may be further weakened by propagation through the RAG chain. We propose a novel Hierarchical Visual Attack that misaligns and disrupts the two inputs (the multimodal query and the augmented knowledge) of MRAG’s generator to confuse its generation. We further design a hierarchical two-stage strategy to obtain misaligned augmented knowledge. We disrupt the image input of the retriever to make it recall irrelevant knowledge from the original database, by optimizing the perturbation which first breaks the cross-modal alignment and then disrupts the multimodal semantic alignment. We conduct extensive experiments on two widely-used MRAG datasets: OK-VQA and InfoSeek. We use CLIP-based retrievers and two LMMs BLIP-2 and LLaVA as generators. Results demonstrate the effectiveness of our visual attack on MRAG through the significant decrease in both retrieval and generation performance.

1. Introduction

The robustness of multimodal Retrieval-Augmented Generation (MRAG) [1, 13, 27, 33] systems is of great impor-

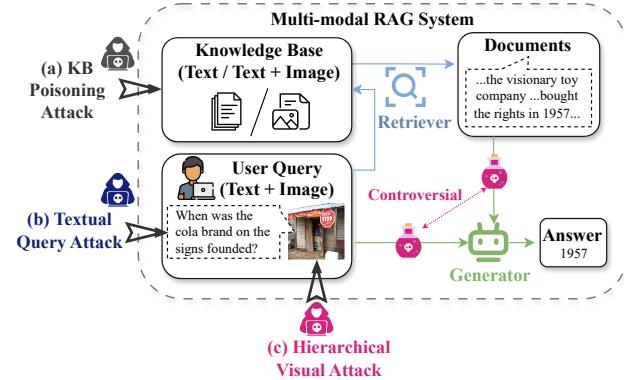


Figure 1. Comparison on (a) KB Poisoning Attack, (b) Textual Query Attack and our focused task (c) Hierarchical Visual Attack on multi-modal RAG.

tance. As advanced MRAG techniques have been widely applied to large multimodal models (LMMs) [18, 22, 23, 34] to enhance their knowledge and ability, novel safety issues have also emerged. Existing adversarial research has revealed the vulnerability of MRAG to knowledge poisoning attacks [11, 24, 40]. These methods rely on constructing and injecting malicious contents into the knowledge base, which are obvious by changing textual words or can be identified by Knowledge Base (KB) detection algorithm. Meanwhile, there are research of attacks on the generator component of MRAG [30, 31, 35], which include manipulations on textual queries and are easy to notice as well. Thus, in this paper, we consider a different setting of attacking MRAG: pure visual attack that solely adds imperceptible perturbations at the image inputs of users. As shown in Figure 1, our attack method aims to disrupt both retrieval and generation process of MRAG by optimizing visual perturbations, without manipulating any other com-

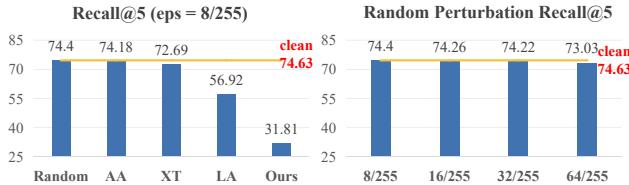


Figure 2. Attack performance on a CLIP model fine-tuned for the retrieval task by different attack methods (AA: Any Attack [39]; XT: X-Transfer [14], LA: LMM Attack [7]) and scale.

ponents. Compared to poisoning attacks, adversarial visual attack on images can be highly imperceptible to human vision, making them exceptionally stealthy.

However, attacking MRAG with pure visual perturbation is challenging. The MRAG system is a pipeline that consists of a fine-tuned retriever and a large-scale generator, both of which retain a certain degree of robustness. For fine-tuned retrievers, as shown in Figure 2, advanced visual attack methods [7, 14, 39] can cause limited retrieval performance drop, and random perturbations even with a large scale (64/255) causes subtle performance drop. This may be because the models become adept at multimodal knowledge grounding and logical association, focusing on relational understanding (e.g., the logical relationship between a question and a potential answer). This emphasis on deeper logical correlation and inference ability makes the model less susceptible to simple input perturbations in a single modality, significantly increasing the challenge of our image attack task. For generators, large-scale LMMs such as LLaVA [22] and BLIP-2 [18] are often used, which generally demonstrate superior robustness [36], further posing challenges to MRAG attack.

The sequential working pipeline of MRAG also enhances its robustness towards visual perturbations as the attack effect degrades by propagation through the RAG chain. The perturbation's effect on retriever lies in the recalled knowledge, which experiences chain of transformation and decrease. Consequently, success visual attack on MRAG is hard, while posing a more severe threat than obvious text attacks or KB poisoning attacks, as the resulting bad influence can go undetected and propagate through the entire RAG chain.

To effectively subvert the robustness of MRAG with only visual perturbations, we propose a novel hierarchical disruption approach to attack the MRAG's on two components and mislead the final result. As shown in Figure 1, there are two inputs for the MRAG generator: the multimodal user query and the augmented knowledge. Our hierarchical method targets at misaligning and disrupting these two inputs, creating knowledge conflicts for the generator while breaking different levels of model capabilities. For retrieval, we employ a hierarchical two-stage strategy to op-

timize the applied noise in modality and semantic levels. We first alter the query features to no longer correspond accurately to itself, so that the retrieval query is deviated. Then, the semantic alignment between query and knowledge is broken down. By structurally targeting the model's core competencies across different levels of abstraction, our approach achieves severe and effective degradation regarding retrieval and generation performance.

In summary, our main contributions are as follows:

- We propose a novel Hierarchical Visual Attack method, which misaligns and disrupts the two inputs of generator in MRAG, posing stealth and severe threats for MRAG systems.
- Our method is the first to disrupt MRAG systems only using image perturbations, which further reveals the vulnerabilities of MRAG systems towards more imperceptible attacks.
- We conduct extensive experiments on two datasets and four versions of retrievers to prove our attack's effectiveness.

2. Related Work

2.1. Multimodal RAG and Existing Attacks

Multimodal Retrieval-Augmented Generation [1, 27] has emerged as an important technique by extending traditional RAG [17] to multimodal data, enabling more real-world applications. Despite its huge success, the security issue has gained much attention. Recent studies have highlighted the vulnerabilities of multimodal RAG systems to knowledge poisoning attacks, where malicious information is injected into the external knowledge databases to manipulate the RAG's outputs. MM-Poisoning [11] achieved the attack by constructing query-specific misinformation into injected text and images, or inserting a single irrelevant knowledge instance to fool all queries. Poisoned-MRAG [24] formalized the attack as an optimization problem and proposed cross-modal attack strategies to disrupt both retrieval and generation. PoisonedEye [37] designed injected textual context or optimized the poison image to reduce retrieval performance to achieve attack goals. Despite these advancements, existing research still follows the line of text RAG attacks, which mainly focus on knowledge poisoning attacks. Furthermore, the multimodal characteristics remains underexplored. Thus, in this paper, we propose a new attack on multimodal RAG by only learning small adversarial perturbations added to the image in user query, without modifying the external knowledge base.

2.2. Visual Attack

Visual adversarial attacks can be categorized into white-box, gray-box and black-box based on the level of knowledge about the attacked model [38]. White-box methods [2,

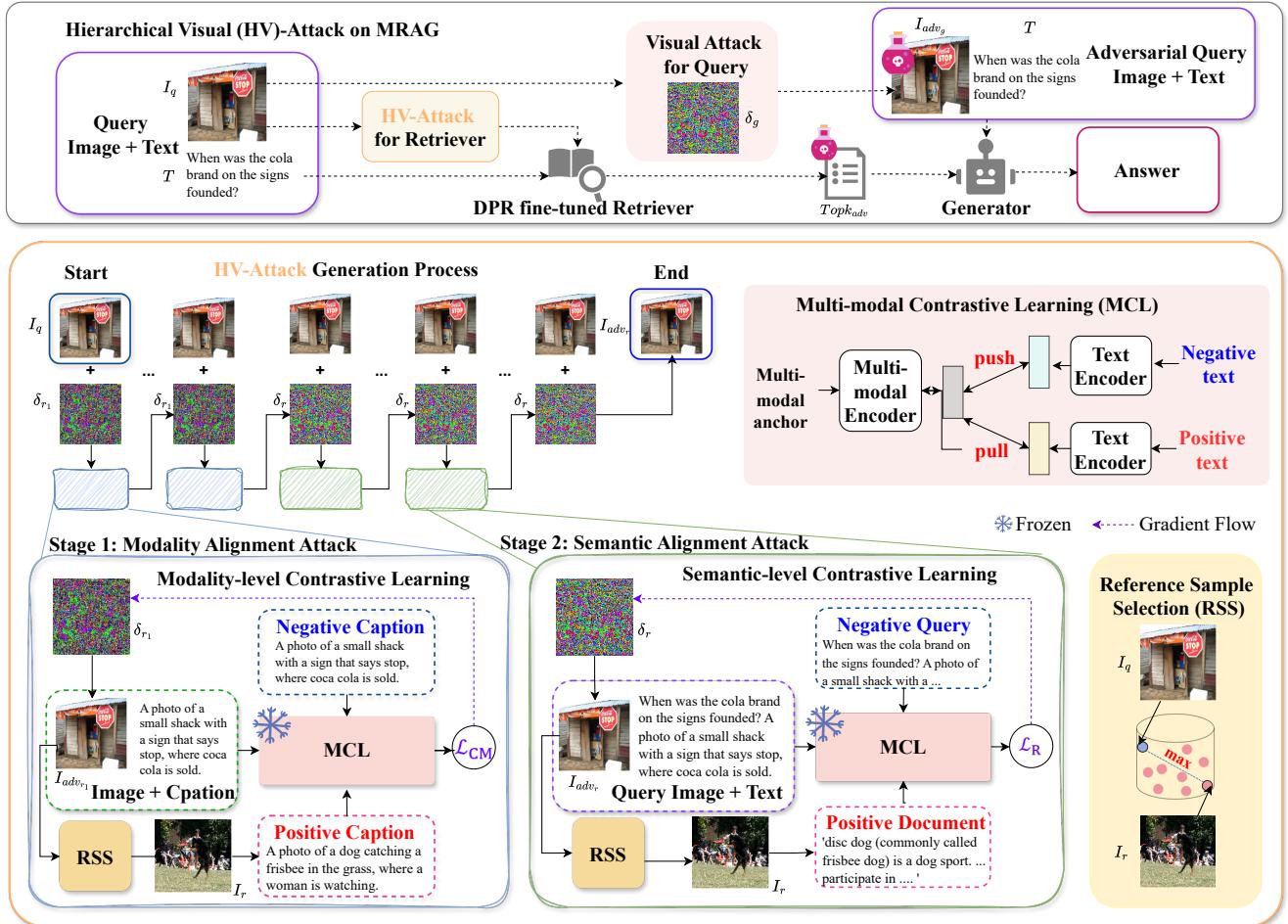


Figure 3. An overview of the proposed method. The top part illustrates the overall MRAG pipeline and our hierarchical structure, which adds image perturbation to the retriever and generator respectively. The hierarchical two-stage strategy we employed for generating retrieval visual attack is shown in the lower block. We optimize the added perturbation step-by-step by first breaking the modality alignment, then disrupting the semantic alignment.

10, 32], which have full access to the attacked model, aim to maximize the adversarial effect. Fast Gradient Sign Method (FGSM) [9] generates adversarial examples by adding a small perturbation in the direction of the gradient of the loss function with respect to the input. The Projected Gradient Descent (PGD) [25] is an iterative variant of FGSM that refines the perturbation through multiple steps. The Carlini and Wagner (CW) attack [4] employs optimization techniques to find adversarial examples that achieve best attack while being close to the original input. Gray-box [3, 15] and black-box [5] attacks have limited knowledge about the attacked model, and rely on transferability or surrogate models to craft effective perturbations. In this paper, we focus on white-box attack on images, assuming full access to the open-source retriever in MRAG pipeline.

3. Method

3.1. MRAG Preliminaries and Problem Definition

A multimodal RAG pipeline consists of a retriever \mathbf{R} , a generator \mathbf{G} and an external knowledge base \mathcal{KB} . When a user inputs an image I and a text query T , the retriever \mathbf{R} first retrieves top-k augmented knowledge from \mathcal{KB} . Then the generator \mathbf{G} generates an answer using the multimodal query along with k augmented knowledge. The retrieval process follows the DPR [16] structure, which comprises an image encoder E_I and a text encoder E_T . The encoders embed input and knowledge into a common semantic space. A similarity-based retrieval can be conducted by calculating and ranking the distance between the input and knowledge in this semantic space. For each query, the top-k knowledge embeddings with highest similarity are retrieved as results. The multimodal embedding of a user query is obtained by

summing the corresponding image and text embeddings, i.e., $E_I(I) + E_T(T)$. The knowledge embedding is depended on its modality. For text corpus, each passage K_t 's embedding is $E_T(K_t)$. For multimodal KB, each image-text pair (K_i, K_t) 's embedding is $E_I(K_i) + E_T(K_t)$.

In this paper, our goal is to attack the MRAG pipeline by solely modifying the input image. We achieve our attack by misaligning and disrupting the two inputs of the generator component in MRAG, which results in confusion within inputs along with misleading information. To be more specific, for the augmented knowledge, we optimize a small perturbation δ_r added at the input image I_q to mislead the retriever to return irrelevant knowledge. For the image input of the generator, we add another small perturbation δ_g that breaks down uni-modal semantic within the image to disrupt image understanding. Consequently, the generator produces incorrect answers based on the misaligned input query and augmented knowledge, with misleading irrelevant contents. The attack objective can be formulated as follows:

$$\begin{aligned} I_{adv_r} &= attack_r(I_q, \delta_r), \quad \|I_q - I_{adv_r}\|_\infty = \|\delta_r\|_\infty < \epsilon, \\ I_{adv_g} &= I_q + \delta_g, \quad \|I_q - I_{adv_g}\|_\infty = \|\delta_g\|_\infty < \epsilon, \\ Topk_{adv} &= \mathbf{R}(T, I_{adv_r}, \mathcal{KB}), \\ A_{adv} &= \mathbf{G}(T, I_{adv_g}, Topk_{adv}), \end{aligned}$$

where δ_r and δ_g are noises added to the image in retrieval and generation, $attack_r(\cdot)$ represents our hierarchical attack method of learning retrieval and generation perturbations respectively, ϵ is the bound of added perturbation to ensure that it is imperceptible to human eye.

3.2. Overall Hierarchical Attack Framework

The overall attack framework is illustrated in Figure 3. To produce confusion for the generator in MRAG, we aim to misalign the two inputs of it: the input query and augmented knowledge. Thus, we hierarchically apply two different perturbations to the image, resulting in misleading effect in two directions.

For the augmented knowledge, the added perturbation δ_r mainly takes effect by disrupting the retrieval process. We further adopt a hierarchical two-stage strategy to learn δ_r , targeting at modality alignment and semantic alignment within retrieval respectively. The detailed design of this hierarchical strategy is described in subsection 3.3. For the generator input query, the added perturbation δ_g breaks down the uni-modal semantic within the image input, so as to disrupt image understanding in generation.

3.3. Hierarchical Visual Attack on Augmented Knowledge

The visual attack on the augmented knowledge as inputs to the generator is achieved by disrupting the retrieval process.

We employ a hierarchical two-stage strategy to break down modality and semantic alignments respectively, which are key characteristics in success retrieval.

Stage 1: Modality Alignment Attack

Multimodal dual-encoder retrievers, such as CLIP [29], achieve strong alignment between text and image modalities. The corresponding text and image embeddings are close in the embedding space, enabling relevant knowledge to be retrieved through the ranking of embedding similarities during cross-modal retrieval. Thus, the first stage of visual attack on augmented knowledge is to break down the multimodal alignment between these modalities.

We achieve this by pushing the query image close to the least similar sample. Given a user query image, we search for a reference image from the database whose embedding has the smallest similarity with the query image's embedding. Then, given the retriever image encoder E_i , text encoder E_t and image captioning model C , we obtain the multimodal image representation $f_{multi_stage_1}$, the clean query caption's embedding f_{clean_cap} and the reference image caption's embedding f_{ref_cap} as follows:

$$\begin{aligned} f_{clean_cap} &= E_t(C(I_q)), \\ f_{ref_cap} &= E_t(C(I_r)), \\ f_{multi_stage_1} &= E_i(I_p) + E_t(C(I_p)), \end{aligned} \quad (1)$$

where I_q and I_r are the input user query image and its reference image, I_p is the query image with added perturbation.

To learn the best the perturbation, we conduct contrastive learning between $f_{multi_stage_1}$, f_{clean_cap} and f_{ref_cap} and design the loss function \mathcal{L}_{CM} based on hinge loss [8]:

$$\begin{aligned} \mathcal{L}_{CM} &= \max (\|clean_sim - \beta \cdot ref_sim\| + \gamma, 0), \\ clean_sim &= sim(f_{multi_stage_1}, f_{clean_cap}), \\ ref_sim &= sim(f_{multi_stage_1}, f_{ref_cap}). \end{aligned} \quad (2)$$

As shown in Equation 2, by minimizing \mathcal{L}_{CM} , we minimize the similarity between the query and clean image's caption, while maximize the similarity with the reference caption. The detailed optimization process can be found in Algorithm 1. In this stage, since we focus on cross-modal alignment, we consider the alignment between multimodal representation of the query image and the corresponding text embeddings. Through iterations, we minimize the similarity between the multimodal embedding and its text caption embedding, while maximizing the similarity with the reference text caption embedding.

Stage 2: Semantic Alignment Attack

The second stage of our attack aims to further break down the semantic relevance between cross-modal embeddings, which is crucial for retrieving useful knowledge for generation. The form of the loss function is similar to that in the first stage. However, the contrastive learning is now

Algorithm 1 Modality Alignment Attack (Stage 1)

Input: User image I_q , reference image I_r , retriever image encoder E_i , retriever text encoder E_t , image captioning model \mathbf{C} , generation steps s , step length α , perturbation bound ϵ , trade-off parameter β , margin parameter γ .

Output: Stage 1's perturbation δ_{r_1}

- 1: Get clean and reference image caption embedding:
 $f_{clean_cap} \leftarrow E_t(\mathbf{C}(I_q))$, $f_{ref_cap} \leftarrow E_t(\mathbf{C}(I_r))$
- 2: Initialize perturbation $\delta_{r_1} \leftarrow 0$
- 3: Initialize perturbed image $I_p \leftarrow I_q + \delta_{r_1}$
- 4: **for** step $\leftarrow 1$ to s **do**
- 5: $I_p \leftarrow I_q + \delta_{r_1}$
- 6: $f_{multi_stage1} \leftarrow E_i(I_p) + E_t(\mathbf{C}(I_p))$
- 7: clean_sim $\leftarrow \text{sim}(f_{multi_stage1}, f_{clean_cap})$
- 8: ref_sim $\leftarrow \text{sim}(f_{multi_stage1}, f_{ref_cap})$
- 9: $\mathcal{L}_{CM} \leftarrow \max(\|\text{clean_sim} - \beta \cdot \text{ref_sim}\| + \gamma, 0)$,
- 10: Optimize $\delta_{r_1} \leftarrow \delta_{r_1} - \alpha \cdot \text{sign}(\nabla_{\delta_{r_1}} \mathcal{L}_{CM})$
- 11: $\delta_{r_1} \leftarrow \text{Clip}(\delta_{r_1}, -\epsilon, \epsilon)$
- 12: **end for**
- 13: **Return** δ_{r_1}

conducted between the multimodal representation of user query, the query text embedding and reference passage embedding. We retrieve the positive text knowledge for the reference image and use its embedding as the positive embedding in this stage. The embeddings are obtained by:

$$\begin{aligned} f_{clean_query} &= E_t(T_q + \mathbf{C}(I_q)), \\ f_{ref_passage} &= E_t(T_p), \\ f_{multi_stage2} &= E_i(I_p) + E_t(T_q + \mathbf{C}(I_p)), \end{aligned} \quad (3)$$

where T_q and T_p are the user text query and retrieved reference text knowledge respectively.

The contrastive learning of the second stage is designed as shown in [Equation 4](#):

$$\begin{aligned} \mathcal{L}_R &\leftarrow \max(\|\text{clean_sim} - \beta \cdot \text{ref_sim}\| + \gamma, 0), \\ \text{clean_sim} &\leftarrow \text{sim}(f_{multi_stage2}, f_{clean_query}), \\ \text{ref_sim} &\leftarrow \text{sim}(f_{multi_stage2}, f_{ref_passage}). \end{aligned} \quad (4)$$

The detailed process of the second stage is shown in [Algorithm 2](#). Building on the perturbation generated in the first stage, the second stage iteratively refines and outputs the final perturbation.

With the two-stage hierarchical strategy, we obtain the final retrieval perturbation added to the query image and retrieve the adversarial augmented knowledge. The generator's disrupted query is obtained by adding noise generated by advanced attack method on LMMs to the image. The two inputs that misalign with each other achieve the hierarchical attack on MRAG pipeline.

Algorithm 2 Semantic Alignment Attack (Stage 2)

Input: Perturbation δ_{r_1} from first stage, user image I_q , User text query T_q , reference positive passage T_p , models E_i, E_t, \mathbf{C} and hyperparameters $s, \alpha, \epsilon, \beta, \gamma$ the same as stage 1.

Output: Final Perturbation δ_r

- 1: Get reference positive passage embedding:
 $f_{ref_passage} \leftarrow E_t(T_p)$
- 2: Get clean query text embedding: $f_{clean_query} \leftarrow E_t(T_q + \mathbf{C}(I_q))$
- 3: Initialize perturbation $\delta_r \leftarrow \delta_{r_1}$
- 4: Initialize perturbed image $I_p \leftarrow I_q + \delta_r$
- 5: **for** step $\leftarrow 1$ to s **do**
- 6: $I_p \leftarrow I_q + \delta_r$
- 7: $f_{multi_stage2} \leftarrow E_i(I_p) + E_t(T_q + \mathbf{C}(I_p))$
- 8: clean_sim $\leftarrow \text{sim}(f_{multi_stage2}, f_{clean_query})$
- 9: ref_sim $\leftarrow \text{sim}(f_{multi_stage2}, f_{ref_passage})$
- 10: $\mathcal{L}_R \leftarrow \max(\|\text{clean_sim} - \beta \cdot \text{ref_sim}\| + \gamma, 0)$,
- 11: Optimize $\delta_r \leftarrow \delta_r - \alpha \cdot \text{sign}(\nabla_{\delta_r} \mathcal{L}_R)$
- 12: $\delta_r \leftarrow \text{Clip}(\delta_r, -\epsilon, \epsilon)$
- 13: **end for**
- 14: **Return** δ_r

4. Experiments

4.1. Datasets and evaluation metrics

We conduct our experiments on two wide-used datasets:

- **OK-VQA** [26]: This is a large knowledge-based VQA dataset. Each data sample consists of a question, an image and 10 gold answers. The images are from the COCO dataset [19] and each question requires external knowledge to answer. We use the test set for attack evaluation, which contains 5046 samples. For OK-VQA, we follow RAVQA [20], FLMR [21] to use the Google Search corpus as the external knowledge base, which contains 169,306 passages in total.
- **Infoseek** [6]: We follow PoisonedEye [37] to attack 1000 samples from the visual question answering dataset Infoseek for evaluation. The knowledge database is the same subset of 2M image-text pairs from OVEN-Wiki [12] with PoisonedEye.

Evaluation metrics. Our attack method is evaluated from two perspectives: retrieval and generation. For retrieval, we first use $S(q, p)$ to identify the relation between a query q and a passage p , which is classified based on whether the passage contains a ground-truth answer to the query.

$$S(q, p) = \begin{cases} 1, & \text{if } p \text{ contains answer to } q, \\ 0, & \text{if } p \text{ does not contain answer to } q. \end{cases}$$

Then, we adopt the following two retrieval metrics.

Table 1. VQA Performance on OK-VQA. (ASR* is calculated as $(s_{clean} - s_{adv})/s_{clean}$ for each metric s , EM: Exact Match, “ \uparrow ” indicates that a higher value is better for this metric, while “ \downarrow ” indicates that a lower value is better. All VQA metrics are reported with RAG knowledge number $K = 5$. **Bold** and underline represent the best and second best results respectively.)

Retriever	Method	Retrieval(%)				BLIP-2 VQA(%)			
		R@5(\downarrow)	ASR*(\uparrow)	P@5(\downarrow)	ASR*(\uparrow)	VQA Score(\downarrow)	ASR*(\uparrow)	EM(\downarrow)	ASR*(\uparrow)
CLIP ViT-L/14	Clean	50.57	-	27.86	-	38.53	-	41.76	-
	AnyAttack [39]	50.12	0.89	27.21	2.33	37.94	0.59	41.14	0.62
	X-Transfer [14]	48.14	4.81	25.27	9.30	32.95	14.48	36.03	13.72
	LMM Attack [7]	12.76	74.77	4.48	83.92	<u>32.53</u>	<u>15.57</u>	<u>35.26</u>	<u>15.57</u>
CLIP ViT-L/14 <i>finetuned</i>	Ours	<u>14.00</u>	<u>72.32</u>	<u>4.86</u>	<u>82.56</u>	28.39	26.32	30.99	25.79
	Clean	74.63	-	44.03	-	41.25	-	44.55	-
	AnyAttack [39]	74.18	0.60	43.56	1.07	41.11	0.34	44.47	0.18
	X-Transfer [14]	72.69	2.60	40.99	6.90	35.54	13.84	38.55	13.47
	LMM Attack [7]	<u>56.92</u>	<u>23.73</u>	<u>27.34</u>	<u>37.91</u>	<u>33.38</u>	<u>19.08</u>	<u>36.29</u>	<u>18.54</u>
	Ours	31.81	57.38	12.40	71.84	28.80	30.18	31.25	29.85

Table 2. VQA Performance on InfoSeek. (ASR* is calculated as $(s_{clean} - s_{adv})/s_{clean}$ for each metric s , EM: Exact Match, “ \uparrow ” indicates that a higher value is better for this metric, while “ \downarrow ” indicates that a lower value is better. All VQA metrics are reported with RAG knowledge number $K = 5$. **Bold** and underline represent the best and second best results respectively.)

Retriever	Method	Retrieval(%)				BLIP-2 VQA(%)		LLava VQA(%)	
		R@5(\downarrow)	ASR*(\uparrow)	P@5(\downarrow)	ASR*(\uparrow)	EM	ASR*(\uparrow)	EM	ASR*(\uparrow)
Siglip-so400m	Clean	44.47	-	20.12	-	21.63	-	31.96	-
	X-Transfer [14]	24.42	45.09	11.01	45.28	<u>11.79</u>	45.49	19.93	37.64
	LMM Attack [7]	<u>5.35</u>	<u>87.97</u>	<u>1.68</u>	<u>91.65</u>	<u>11.91</u>	<u>44.94</u>	11.79	63.11
	Ours	4.37	90.17	1.56	92.25	5.47	74.71	<u>12.52</u>	60.83
CLIP ViT-H	Clean	43.26	-	18.78	-	20.41	-	29.77	-
	AnyAttack [39]	32.68	24.46	14.09	24.97	16.16	20.82	22.96	22.88
	X-Transfer [14]	13.49	68.82	4.45	76.30	<u>5.10</u>	<u>75.01</u>	11.91	59.99
	LMM Attack [7]	<u>4.86</u>	<u>88.77</u>	<u>1.43</u>	<u>92.39</u>	13.00	36.31	<u>9.48</u>	<u>68.16</u>
	Ours	2.31	94.66	0.51	97.28	3.28	83.93	10.45	64.90

- **Recall@K** evaluates how likely the retrieved K passages are to contain answers to the query, which is the proportion of queries that have positive passages in the retrieved results:

$$Recall@K = \min\left(\sum_{k=1}^K S(q, p_k), 1\right). \quad (5)$$

- **Precision@K** evaluates how much percent of the retrieved K passages contain answers to the query:

$$Precision@K = \frac{1}{K} \sum_{k=1}^K S(q, p_k). \quad (6)$$

For generation, we adopt the corresponding metric for each dataset. For OK-VQA, VQA score and Exact Match are used following RAVQA [20]. For InfoSeek, Exact

Match is calculated. For each metric, we calculate the non-target attack success rate proposed by X-transfer [14], which is $(s_{clean} - s_{adv})/s_{clean}$ for each metric s .

4.2. Implementation Details

We implement the proposed method based on the open-source PyTorch [28] framework. All the experiments are conducted on NVIDIA RTX3090. We adopt PGD-step 50, step size $\alpha = 1/255$, perturbation bound $\epsilon = 8/255$ for the two stages respectively in retrieval perturbation generation. The trade-off parameter β and margin parameter γ are set to 0.4 and 0.6 in all loss functions. We adopt the X-transfer noise as δ_g in our experiments.

For OK-VQA, we use the off-the-shell CLIP ViT-L/14 as well as a fine-tuned version of it as retrievers. For Infoseek, we use the off-the-shell Siglip-so400m and CLIP ViT-H as

retrievers following PoisonedEye [37]. For generators, we use the off-the-shell BLIP-2-flan-T5-xl and LLava-NEXT models.

4.3. Baseline Models

We compare our method with various state-of-the-art visual noise generation algorithms. AnyAttack [39] finetuned a decoder to generate noise for any image that transforms it to any target. We use the released AnyAttack decoder to generate noises using the reference image in our method for comparison. X-transfer [14] learned a transferrable noise that can be applied to any image. We employ the "xtransfer_large_linf_eps12_non_targeted" noise and set $\epsilon = 8$ to apply to all the images. LMM Attack [7] learned the noise using PGD algorithm, with the cross entropy loss to minimize similarity between the adversarial image and its image caption as the objective function. Since the code was not released, we re-implemented the algorithm and set PGD-step to 50.

4.4. Experimental Results

4.4.1. Main Attack Results

The main attack results of attacking the MRAG system on two datasets are shown in [Table 1](#) and [Table 2](#). The overall results show that our proposed method achieved the most severe attack influence on retrieval and generation process of MRAG. Several key conclusions can be made:

Our hierarchical attack works on both retrieval and generation. Among the datasets, our hierarchical attack method all achieve declines in both retrieval and generation metrics. As shown in the results, former baseline methods demonstrate attack advantage either in retrieval or generation. X-transfer [14]'s noise is more effective at attacking the generator (the retrieval metrics' decline are subtle compared to the VQA metrics). While LMM attack [7]'s noise is comparably more effective at attacking the retrieval process. For our attack, the hierarchical structure leads to a more balanced attack effectiveness on both retrieval and generation, and disrupts the whole MRAG chain by only modifying the image input, making the attack highly imperceptible to human vision while effective.

Our hierarchical attack works on the fine-tuned retriever. As shown in the tables, the off-the-shell models are highly vulnerable against adversarial noise in the retrieval task. However, previous attack methods can not achieve equivalent effectiveness towards both off-the-shell and fine-tuned models. For example, as shown in [Table 1](#), the non-hierarchical LMM attack can achieve 74.77% attack success rate on the R@5 metric of off-the-shell CLIP, while only getting 23.73% success rate on the fine-tuned version. While for our hierarchical attack method, we achieved high success rate (72.32% and 57.38%) on non-fine-tuned and fine-tuned retrievers. Note that though the attack perfor-

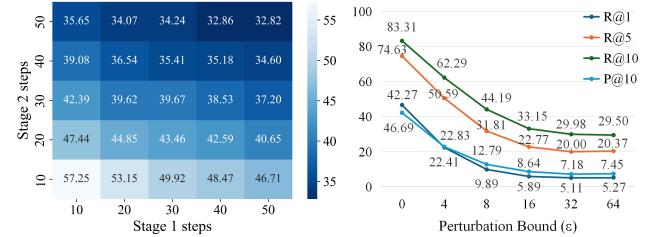


Figure 4. Performance of ablated models with different steps.

mance of LMM Attack and our method are compatible on the original CLIP, our method achieves over two times the success rate on the fine-tuned version.

Our hierarchical attack works on various black-box generators. We use BLIP-2 and LLava as black-box generators. VQA results of OK-VQA using LLava is shown in [Table 3](#). With using different black-box generators, our attack with adversarial augmented knowledge and query noise can all achieve damaging attack effect.

	VQA Score(\downarrow)	ASR*(\uparrow)	EM(\downarrow)	ASR*(\uparrow)
Clean	63.30	-	67.34	-
AnyAttack	61.31	3.14	65.46	2.79
X-transfer	57.17	9.68	60.98	9.44
LMM Attack	<u>56.72</u>	<u>10.39</u>	<u>60.42</u>	<u>10.23</u>
Ours	54.19	14.39	57.77	14.21

Table 3. VQA results on OK-VQA with LLava as generator. Top-5 retrieved documents are used for generation.

4.4.2. Ablation Studies

Analysis of Hierarchical Retrieval Attack. [Table 4](#) and [Table 5](#) show the detailed retrieval results on OK-VQA and Infoseek datasets. Both stages of the attack demonstrate individual effectiveness. Using stage 1 and stage 2 alone both degrades retrieval metrics. Stage 2's individual attack performance is comparably better than stage 1, since its attack goal is more directly targeted at retrieval. While across all datasets and retriever evaluated, the hierarchical two-stage attack consistently achieves optimal performance, demonstrating the attack effect of breaking both modality and semantic alignment.

Effect of PGD Steps. As shown in [Figure 4](#) (left), we report the Recall@5 metric on OK-VQA dataset with fine-tuned CLIP using 10 to 50 steps of stage 1 and 2. As the PGD algorithm increases optimization steps, the attack results become better, eventually coming to convergence. We finally choose PGD-step50 for each stage of optimization.

Effect of Perturbation Budget. We explore the effect of setting ϵ to different values. As shown in [Figure 4](#) (right),

Table 4. Retrieval Performance on OK-VQA. (**Bold** and underline represent the best and second best results respectively.)

CLIP ViT-L/14				
	R@5	R@10	P@5	P@10
Clean	50.57	62.25	27.86	27.48
AnyAttack [39]	50.12	61.06	27.21	26.86
X-Transfer [14]	48.14	59.02	25.27	24.73
LMM Attack [7]	12.76	19.26	4.48	4.65
Ours w/ Stage 1	21.94	31.85	8.52	8.74
Ours w/ Stage 2	19.70	28.22	7.56	7.77
Ours	<u>14.00</u>	<u>20.77</u>	<u>4.86</u>	<u>4.93</u>
CLIP ViT-L/14 finetuned				
	R@5	R@10	P@5	P@10
Clean	74.63	83.31	44.03	42.27
AnyAttack [39]	74.18	82.82	43.56	41.83
X-Transfer [14]	72.69	81.25	40.99	38.91
LMM Attack [7]	56.92	68.59	27.34	26.56
Ours w/ Stage 1	54.48	65.93	25.73	25.07
Ours w/ Stage 2	<u>37.65</u>	<u>49.68</u>	<u>15.56</u>	<u>15.78</u>
Ours	31.81	44.19	12.40	12.78

Table 5. Retrieval Performance on Infoseek. (**Bold** and underline represent the best and second best results respectively.)

Siglip-so400m				
	R@5	R@10	P@5	P@10
Clean	43.47	51.89	19.61	17.06
X-Transfer [14]	24.42	30.01	11.01	9.68
LMM Attack [7]	5.35	8.14	1.68	1.7
Ours w/ Stage 1	5.1	8.51	1.97	1.91
Ours w/ Stage 2	<u>4.74</u>	<u>7.9</u>	1.51	<u>1.51</u>
Ours	4.37	7.29	<u>1.56</u>	1.51
CLIP ViT-H				
	R@5	R@10	P@5	P@10
Clean	43.26	48.97	18.78	15.55
AnyAttack [39]	32.69	40.34	14.09	11.85
X-Transfer [14]	13.49	18.59	4.45	3.94
LMM Attack [7]	4.86	7.65	1.43	1.49
Ours w/ Stage 1	3.77	5.95	0.92	0.89
Ours w/ Stage 2	<u>2.61</u>	<u>4.39</u>	<u>0.69</u>	<u>0.69</u>
Ours	2.31	3.89	0.51	0.52

when ϵ gets larger, the attack effect is more obvious. Different colors refer to various retrieval metrics, when ϵ increases, all the metrics degrade. With a limited budget $\epsilon = 8$, attack effectiveness can already be achieved while

Question: Why might someone go to this place?

GT answers: business, nyc, shop



Clean docs:

1. macy's on state street: the marshall fields department **store** chicago illinois united states of america macy's on state ...
2. ... the busy streets surrounding the **shop**, known as the jordaan, are bustling with locals and tourists alike. ...
3. ... admittedly, it's a bit touristy and pricey, but you can't beat the location and browsing the unique mom-and-pop **shops**.
4. ... state street, macy's **department store** on right, formerly marshall field's. chicago famous marshall fields clock ...
5. ... broadway is synonymous with theatre, fifth avenue is automatically paired with **shopping**, madison avenue means the advertising industry, ...

Adversarial docs:

1. **winter sports**: sledding, skiing, snowboarding, skating (for kids) - nemours kidsshealth [skip to content] open search for parents parents site sitio para padres ...
2. ... **winter sports** are lots of fun — just ask any kid who's just scored the winning goal during an ice-hockey game or finished sledding ...
3. aug 7, 2019 · whether it is at a christmas market or end of year sale at zara, you should expect to find people scavenging for all sorts of things. ...
4. but you need the right equipment: flexible flyer baby pull sledbecause babies love **snow** tocheck price this is a great baby sled for pulling you ...
5. baby sled – this flexible flyer baby pull sled is the perfect way to tow babies on **packed snow**, groomed paths, or snow up to 4 inches deep secure support ...

LLaVA clean answer: 'Someone might go to this place to **shop**, as it is a busy shopping district with a large department store, Macy's, and other stores.'

LLaVA adv answer: 'Someone might go to this place to enjoy the **winter sports** activities offered, such as sledding, skiing, snowboarding, and skating.'

Figure 5. An example showing the original image, the adversarial images as inputs to retriever and generator, as well as the clean and adversarial augmented knowledge, along with the generated answer based on them. The example demonstrates the attack effect of our hierarchical method within imperceptible disruption.

being imperceptible.

4.4.3. Case studies

We provide a case study in Figure 5. This example shows the attack effect of our hierarchical method under the perturbation budget $\epsilon = 8$. As shown in the example, with the original query image without noise, the retriever returns augmented knowledge that are closely related to answering the question. The generator provides a correct answer accordingly. However, with added imperceptible noise, the retriever returns irrelevant augmented knowledge that are about winter sports. With the disruptions, the final prediction of the generator is wrong and unreasonable.

5. Conclusion

In this paper, we proposed a novel attack method against MRAG pipeline that focused solely on image inputs. With hierarchical optimization, we target MRAG's retriever and generator across different levels of abstraction, achieving

severe while stealth attack impact. Our research reveals that MRAG technologies, while widely adopted in practice, remain vulnerable to security threats posed by imperceptible adversarial visual noise. Future work will focus on uncovering deeper and more diverse potential threats brought by visual attacks on MRAG systems and developing robust defense mechanisms to balance effectiveness and security.

References

- [1] Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *arXiv preprint arXiv:2502.08826*, 2025. [1](#) [2](#)
- [2] Jay Barach. Cross-domain adversarial attacks and robust defense mechanisms for multimodal neural networks. In *International Conference on Advanced Network Technologies and Intelligent Computing*, pages 345–362. Springer, 2024. [2](#)
- [3] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017. [3](#)
- [4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. [3](#)
- [5] Huanran Chen, Yichi Zhang, Yinpeng Dong, Xiao Yang, Hang Su, and Jun Zhu. Rethinking model ensemble in transfer-based adversarial attacks. *arXiv preprint arXiv:2303.09105*, 2023. [3](#)
- [6] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*, 2023. [5](#)
- [7] Xuanming Cui, Alejandro Aparcero, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24625–24634, 2024. [2](#) [6](#) [7](#) [8](#)
- [8] Claudio Gentile and Manfred KK Warmuth. Linear hinge loss and average margin. *Advances in neural information processing systems*, 11, 1998. [4](#)
- [9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. [3](#)
- [10] Zhongliang Guo, Weiye Li, Yifei Qian, Ognjen Arandjelovic, and Lei Fang. A white-box false positive adversarial attack method on contrastive loss based offline handwritten signature verification models. In *International Conference on Artificial Intelligence and Statistics*, pages 901–909. PMLR, 2024. [3](#)
- [11] Hyeonjeong Ha, Qiusi Zhan, Jeonghwan Kim, Dimitrios Bralios, Saikrishna Sanniboina, Nanyun Peng, Kai-Wei Chang, Daniel Kang, and Heng Ji. Mm-poisonrag: Disrupting multimodal rag with local and global poisoning attacks. *arXiv preprint arXiv:2502.17832*, 2025. [1](#) [2](#)
- [12] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075, 2023. [5](#)
- [13] Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. *arXiv preprint arXiv:2410.08182*, 2024. [1](#)
- [14] Hanxun Huang, Sarah Erfani, Yige Li, Xingjun Ma, and James Bailey. X-transfer attacks: Towards super transferable adversarial attacks on clip. In *ICML*, 2025. [2](#) [6](#) [7](#) [8](#)
- [15] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pages 2137–2146. PMLR, 2018. [3](#)
- [16] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781, 2020. [3](#)
- [17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. [2](#)
- [18] 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. [1](#) [2](#)
- [19] 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. [5](#)
- [20] Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*, 2022. [5](#) [6](#)
- [21] Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36: 22820–22840, 2023. [5](#)
- [22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. [1](#) [2](#)
- [23] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 5:208–215, 2024. [1](#)
- [24] Yinuo Liu, Zenghui Yuan, Guiyao Tie, Jiawen Shi, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. Poisoned-mrag: Knowledge poisoning attacks to multimodal retrieval

- augmented generation. *arXiv preprint arXiv:2503.06254*, 2025. 1, 2
- [25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 3
- [26] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 5
- [27] Lang Mei, Siyu Mo, Zhihan Yang, and Chong Chen. A survey of multimodal retrieval-augmented generation. *arXiv preprint arXiv:2504.08748*, 2025. 1, 2
- [28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6
- [29] 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. 4
- [30] Christian Schlaremann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3677–3685, 2023. 1
- [31] Haodi Wang, Kai Dong, Zhilei Zhu, Haotong Qin, Aishan Liu, Xiaolin Fang, Jiakai Wang, and Xianglong Liu. Transferable multimodal attack on vision-language pre-training models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 1722–1740. IEEE, 2024. 1
- [32] Ruofan Wang, Xingjun Ma, Hanxu Zhou, Chuanjun Ji, Guangnan Ye, and Yu-Gang Jiang. White-box multimodal jailbreaks against large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6920–6928, 2024. 3
- [33] 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. *arXiv preprint arXiv:2410.13085*, 2024. 1
- [34] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1
- [35] Ziyi Yin, Muchao Ye, Tianrong Zhang, Jiaqi Wang, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vqattack: Transferable adversarial attacks on visual question answering via pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6755–6763, 2024. 1
- [36] Yixiao Zeng, Tianyu Cao, Danqing Wang, Xinran Zhao, Zimeng Qiu, Morteza Ziyadi, Tongshuang Wu, and Lei Li. Rare: Retrieval-aware robustness evaluation for retrieval-augmented generation systems. *arXiv preprint arXiv:2506.00789*, 2025. 2
- [37] Chenyang Zhang, Xiaoyu Zhang, Jian Lou, Kai Wu, Zilong Wang, and Xiaofeng Chen. Poisondeye: Knowledge poisoning attack on retrieval-augmented generation based large vision-language models. In *Forty-second International Conference on Machine Learning*. 2, 5, 7
- [38] Chiyu Zhang, Lu Zhou, Xiaogang Xu, Jiafei Wu, and Zhe Liu. Adversarial attacks of vision tasks in the past 10 years: A survey. *ACM Computing Surveys*, 58(2):1–42, 2025. 2
- [39] Jiaming Zhang, Junhong Ye, Xingjun Ma, Yige Li, Yunfan Yang, Yunhao Chen, Jitao Sang, and Dit-Yan Yeung. Anyat-attack: Towards large-scale self-supervised adversarial attacks on vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19900–19909, 2025. 2, 6, 7, 8
- [40] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. {PoisonedRAG}: Knowledge corruption attacks to {Retrieval-Augmented} generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3827–3844, 2025. 1