VisRet: Visualization Improves Knowledge-Intensive Text-to-Image Retrieval

Di Wu[1*], **Yixin Wan**[1*], **Kai-Wei Chang**[1]
[1]University of California, Los Angeles
{diwu,kwchang}@cs.ucla.edu

Abstract

Text-to-image retrieval (T2I retrieval) remains challenging because cross-modal embeddings often behave as bags of concepts, underrepresenting structured visual relationships such as pose and viewpoint. We propose **Visualize-then-Retrieve (VisRet)**, a retrieval paradigm that mitigates this limitation of cross-modal similarity alignment. VisRet first projects textual queries into the image modality via T2I generation, then performs retrieval within the image modality to bypass the weaknesses of cross-modal retrievers in recognizing subtle visual-spatial features. Across four benchmarks (Visual-RAG, INQUIRE-Rerank, Microsoft COCO, and our new Visual-RAG-ME featuring multi-entity comparisons), VisRet substantially outperforms cross-modal similarity matching and baselines that recast T2I retrieval as text-to-text similarity matching, improving nDCG@30 by 0.125 on average with CLIP as the retriever and by 0.121 with E5-V. For downstream question answering, VisRet increases accuracy on Visual-RAG and Visual-RAG-ME by 3.8% and 15.7% in top-1 retrieval, and by 3.9% and 11.1% in top-10 retrieval. Ablation studies show compatibility with different T2I instruction LLMs, T2I generation models, and downstream LLMs. VisRet provides a simple yet effective perspective for advancing in text-image retrieval. Our code and the new benchmark are publicly available at https://github.com/xiaowu0162/Visualize-then-Retrieve.

1 Introduction

Text-to-Image (T2I) retrieval selects the most relevant images from a visual corpus given a textual query. It plays a crucial role in knowledge-intensive applications that augment textual inputs with rich visual evidence (Chen et al., 2022; Wang et al., 2023; Sheynin et al., 2023; Braun et al., 2024).

A common approach embeds both the text query and image candidates into a shared representation space and ranks candidates by similarity (Frome et al., 2013; Kiros et al., 2014). However, producing rankings that reflect fine-grained alignments between text and image remains a long-standing challenge. Prior studies show that cross-modal embeddings often behave like "bags of concepts" for similarity matching and fail to capture structured relationships among visual elements (Yüksekgönül et al., 2023; Kamath et al., 2023). For example, Figure 1 presents a query that requires images of an entity (a Barnacle Goose) in a specific posture (wings unfolded). While the embedding model matches the entity type, it struggles to recognize subtler visual-spatial features such as the wing pose and the camera perspective (an upward-facing shot). To address these limitations, existing work improves embedding quality (Radford et al., 2021; Yu et al., 2022), performs query reformulation (Levy et al., 2023), or applies multi-stage reranking pipelines (Liu et al., 2024; Feng et al., 2025). Nevertheless, these strategies remain constrained by the intrinsic difficulty of cross-modal similarity alignment because they cannot bypass the text-to-image similarity search stage.

We propose **Visualize-then-Retrieve (VisRet)**, a retrieval paradigm that decomposes T2I retrieval into two stages: text-to-image *modality projection* followed by *within-modality retrieval*. VisRet first uses a text-to-image generation (T2I generation) model to visualize the textual query, with the goal of making key visual-spatial requirements explicit. The resulting visualization is then used as a query for direct image-to-image retrieval.

Compared to prior methods, VisRet offers two advantages. First, visualizations provide a more expressive and intuitive medium for representing compositional concepts such as entities, poses, and spatial relations. These requirements can be difficult to communicate through text alone,

[*] Equal contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25982–26001

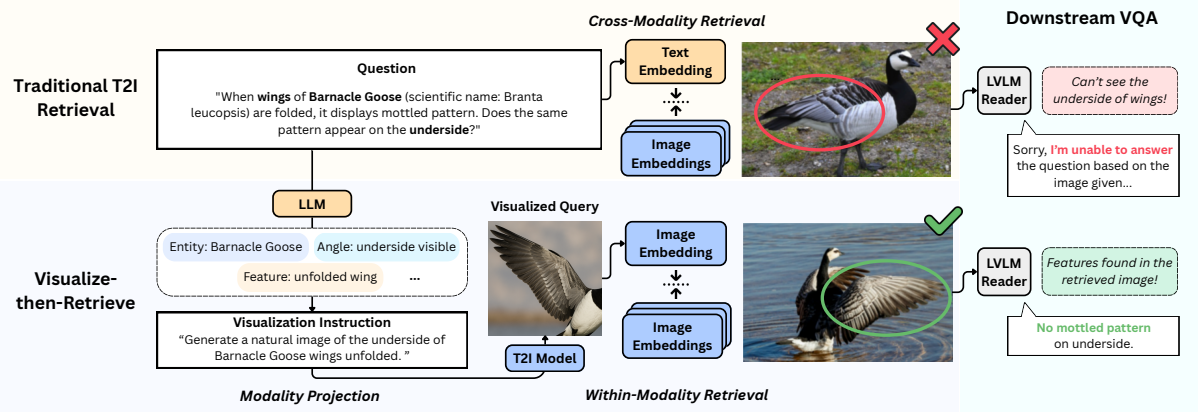


Figure 1: An overview of VisRet. Compared to the traditional T2I retrieval pipeline, VisRet first projects the text query into the image modality via T2I generation and then performs within-modality retrieval.

and encoding them exhaustively in a text query can hurt cross-modal matching due to limitations in embedding quality. In contrast, as shown in Figure 1, the visualized query can depict the desired entity, posture, and camera angle at the same time. Second, by operating entirely within the image modality during retrieval, VisRet avoids the weaknesses of cross-modal retrievers and leverages the stronger capacity of these retrievers in unimodal retrieval (Koishigarina et al., 2025).

To evaluate VisRet, we conduct experiments with two widely used cross-modal embedding models across four benchmarks: Visual-RAG (Wu et al., 2025b), INQUIRE-Rerank (Vendrow et al., 2024), Microsoft COCO (Chen et al., 2015), and Visual-RAG-ME, a new benchmark we introduce that features challenging visual feature comparison questions across multiple entities. VisRet substantially outperforms baseline cross-modal similarity matching and three strong baseline methods that cast T2I retrieval as text-to-text similarity matching (§5.2). When CLIP (Radford et al., 2021) is used as the retriever, VisRet improves nDCG@30 by 0.125 on average across the four benchmarks. With E5-V (Jiang et al., 2024) as the retriever, the improvement is 0.121. VisRet also improves downstream performance in retrieval-augmented generation (RAG) settings (§5.3), increasing T2I question answering accuracy on Visual-RAG and Visual-RAG-ME by 3.8% and 15.7% in top-1 retrieval and by 3.9% and 11.1% in top-10 retrieval. Ablation studies and analyses show that VisRet is robust to the choice of instruction-generation LLM, retriever, and downstream reader LVLM, while performance depends more strongly on the T2I generation model, with stronger generators yielding larger gains. We further analyze repeated runs and find that aggregating multiple visualizations is beneficial, while using a single visualization only slightly reduces overall performance. Finally, although visualized queries are effective for retrieval, they are generally insufficient as standalone knowledge to answer downstream questions. Our code and the new Visual-RAG-ME benchmark will be released publicly to facilitate future research.

2 Related Work

T2I Retrieval Benchmarks Early T2I retrieval benchmarks evaluate whether a system can retrieve an image given its paired human-written caption. These datasets span multiple domains and include widely used benchmarks such as COCO (Chen et al., 2015), Flickr8K (Hodosh et al., 2013), Flickr30K (Young et al., 2014), and Fashion200K (Han et al., 2017). As multimodal embedding models have matured, newer benchmarks have been introduced to assess retrieval in knowledge-intensive settings. Datasets such as WebQA (Chang et al., 2022), INQUIRE (Vendrow et al., 2024), Visual-RAG (Wu et al., 2025b), and MRAG-Bench (Hu et al., 2024) shift the focus from caption matching to retrieving images that provide the information needed to answer knowledge-intensive questions. These tasks evaluate T2I retrieval systems as components of retrieval-augmented generation (RAG) pipelines and emphasize support for downstream question answering. Extending this line of work, Visual-RAG-ME benchmarks T2I retrieval for reasoning about visual features shared across multiple entities.

T2I Retrieval Methods A large body of work improves T2I retrieval from several angles. First, many approaches aim to train stronger multimodal

embeddings through improved objectives and data mixtures (Faghri et al., 2018; Radford et al., 2021; Yu et al., 2022; Li et al., 2022). Other studies target specific stages of the pipeline, including textual query expansion (Levy et al., 2023; Lee et al., 2024) and reranking (Liu et al., 2024; Feng et al., 2025; Ding et al., 2025). Finally, a recent line of work explores generative image retrieval (Li et al., 2024; Qu et al., 2025), which trains a generative model to directly memorize an index of the image corpus. In contrast, VisRet expands query semantics by visualizing the query in the image space, which reduces the burden on cross-modal retrieval. It is also training-free and plug-and-play, accommodating off-the-shelf retrievers or pre-computed image embedding indices.

3 Approach

3.1 Problem Formulation

Given a textual query q and an image corpus $\mathcal{I}$, *Text-to-Image Retrieval* aims to retrieve $n \geq 1$ images $y_1, \ldots, y_n \in \mathcal{I}$ that best match the semantics of q. We further consider *Visual Question Answering* (VQA), where the query is a knowledge-seeking question with an expected answer a.

We consider a basic retrieval-augmented generation (RAG) pipeline for VQA. A multimodal retriever $\mathcal{R}$ retrieves k images from $\mathcal{I}$, denoted as $\{r_1, \ldots, r_k\} \equiv \mathcal{R}(q, \mathcal{I}) \subseteq \mathcal{I}$. A large vision-language model (LVLM) $\mathcal{M}$ then generates an answer conditioned on the question and retrieval results, i.e., $\mathcal{M}(q, \mathcal{R}(q, \mathcal{I}))$.

3.2 Visualize-then-Retrieve

We introduce *Visualize-then-Retrieve (VisRet)*, a two-stage T2I retrieval pipeline that bridges the modality gap through modality projection. Figure 1 illustrates the pipeline with an example.

Modality Projection In the first stage, VisRet uses a T2I generation system $\mathcal{T}$ to synthesize m visualizations $\{v_1, \ldots, v_m\} \equiv \mathcal{T}(q)$. Concretely, given the original query, an LLM first operates in the text space to draft a T2I instruction q' that describes images likely to satisfy the feature requirements implied by q. We then use an existing T2I generation method, such as Stable Diffusion (Esser et al., 2024), to project q' into a set of image queries $\{v_1, \ldots, v_m\}$, where each v_i is a synthesized image. To encourage diversity, randomness can be introduced either in the instruction drafting step or within the T2I generation process. In this work, we sample m times using the same q'.

Within-Modality Retrieval In the second stage, VisRet performs retrieval entirely within the image modality. Each synthesized image $v_i \in \{v_1, \ldots, v_m\}$ is independently used to retrieve a ranked list of images from the corpus:

$$\mathcal{R}(v_i, \mathcal{I}) = [r_1^{(i)}, \ldots, r_k^{(i)}].$$

We then apply Reciprocal Rank Fusion (RRF) (Cormack et al., 2009) to aggregate the m ranked lists. RRF assigns each candidate image r a fusion score based on its rank across the m lists:

$$\text{score}_{\text{RRF}}(r) = \sum_{i=1}^{m} \frac{1}{\lambda + \text{rank}_i(r)},$$

where $\text{rank}_i(r)$ is the rank position of image r in $\mathcal{R}(v_i, \mathcal{I})$, and λ is a hyperparameter that controls the influence of lower-ranked items. The final top-k results are obtained by selecting images with the highest $\text{score}_{\text{RRF}}(r)$.

4 Visual-RAG-ME Benchmark

To support the development of stronger T2I retrieval systems, we introduce **Visual-RAG-ME**, a multi-entity question answering and T2I retrieval dataset that extends Visual-RAG. Visual-RAG-ME constructs questions that compare visual features between the organism in a Visual-RAG query and another similar entity, and we manually annotate a new set of retrieval labels for the second entity. The Visual-RAG-ME annotation pipeline consists of the following three steps.

Second Entity Identification This step identifies potential entities that are biologically close to the original Visual-RAG entities, enabling plausible and challenging comparison questions. For each entity in Visual-RAG, we use BM25 (Robertson and Walker, 1994) to retrieve up to ten entities with the most similar scientific names.

Query Composition and Filtering We (the authors) manually review all 391 questions in Visual-RAG and attempt to construct a corresponding multi-entity question. We construct a question only when we can identify images for the second entity that clearly depict the same feature as the positive images for the original entity in Visual-RAG. The questions we compose typically use a comparison format that asks whether the two

	VR	VR-ME	IR-Hard	CC-Hard
# Queries	391	50	59	500
|Query| (word count)	18.5	25.1	6.0	12.2
# Images (per entity)	264	263	100	5000
# Positives (per entity)	14.3	20.8	12.5	1
CLIP Recall@1	0.210	0.220	0.000	0.000
E5-V Recall@1	0.240	0.340	0.000	0.000

Table 1: Dataset statistics and baseline performance. VR = Visual-RAG, IR = INQUIRE-Rerank, and CC = Microsoft COCO 2017.

organisms share a feature or which organism exhibits a more extreme stylistic feature (e.g., lighter coloration, smoother surface). This process yields 82 multi-entity questions. We then filter out (1) questions answerable by both GPT-4o and GPT-4o-mini without image information and (2) excessive questions covering the same topic. The final Visual-RAG-ME contains 50 high-quality multi-entity queries.

Retrieval Label Annotation and Balancing For each question, we collect images of the second entity from the iNaturalist 2021 dataset (Horn et al., 2021) and annotate retrieval labels. We assign a positive label only when the image clearly displays the feature required to answer the question. For some questions, many positive images exist in iNaturalist, so we apply a filtering step that keeps at most 50 positive images for each entity. Table 5 and Appendix Table 10 show example questions and their ground-truth images.

Table 1 reports the basic statistics of Visual-RAG-ME. While it contains slightly more positives per entity than Visual-RAG, it remains challenging. Both CLIP and E5-V achieve low Recall@1 due to the lengthy and knowledge-intensive queries.

5 Main Results

5.1 Experimental Setup

Benchmarks We evaluate on four benchmarks: (1) Visual-RAG (Wu et al., 2025b), a T2I retrieval and VQA benchmark featuring visual knowledge-intensive questions about natural species, where the relevant evidence is often not readily available in text corpora. (2) Visual-RAG-ME, our new benchmark that compares the same visual feature across multiple entities. (3) INQUIRE-Rerank (Vendrow et al., 2024), which requires accurate knowledge of species appearance and behavior for retrieval. (4) Microsoft COCO 2017 (Chen et al., 2015), a widely used image captioning dataset that we repurpose for caption-to-image retrieval evaluation. For (3) and (4), we further filter queries to retain the most difficult cases, yielding INQUIRE-Rerank-Hard and COCO-Hard. Table 1 reports dataset statistics, and we provide additional details in Appendix §A.

Metrics For all four retrieval benchmarks, we report Recall@k and nDCG@k with $k \in \{1, 10, 30\}$[1]. For Visual-RAG and Visual-RAG-ME, we also evaluate end-to-end VQA accuracy using an LLM judge with a prompt derived from Visual-RAG (Wu et al., 2025b).

Baselines We compare three families of T2I retrieval methods:

1. **Text-to-Image Matching:** Retrieve using the original textual query against the image index with off-the-shelf retrievers (**CLIP**, **E5-V**). We also include a query rewriting baseline, where an LLM rewrites the query to highlight desired features before performing standard T2I retrieval over the same embedding index.
2. **Text-to-Text Matching:** Convert images into textual surrogates and perform text-to-text retrieval avoiding cross-modal similarity matching. We compare with (1) captioning the entire corpus with BLIP, (2) top-k reranking with VISA, and (3) an ablated version of VISA, where we rerank top-k images using LLM-generated captions.
3. **Image-to-Image Matching:** At the time of writing (December 2025), VisRet is the only method in this category.

Implementation Details For all retrieval experiments, we use CLIP and E5-V as retrievers, `GPT-4o` (OpenAI, 2023b) as the T2I instruction generation LLM, and `gpt-image-1` (OpenAI, 2025) as the T2I model to generate $m = 3$ images. For downstream question answering on Visual-RAG and Visual-RAG-ME, we use `GPT-4o` as the reader LLM. Full implementation details for VisRet and baselines are provided in Appendix §B.

5.2 VisRet Improves T2I Retrieval

Table 2 summarizes retrieval results across four benchmarks and two retrievers. VisRet consistently outperforms both the original-query baseline and the LLM-based query rewriting baseline. With CLIP as the retriever, VisRet improves nDCG@10 by 0.109 (38%↑) over the original query and by 0.064 (19%↑) over LLM rephrasing, averaged over

[1] nDCG@1 is undefined and we use Recall@1 as its value.

Retrieval Method	Visual-RAG			Visual-RAG-ME			INQUIRE-Rerank-Hard			COCO-Hard		
	N@1	N@10	N@30	N@1	N@10	N@30	N@1	N@10	N@30	N@1	N@10	N@30
Retriever = CLIP												
Original Query	0.210	0.355	0.385	0.220	0.423	0.435	0.000	0.355	0.412	0.000	0.017	0.042
LLM Rewriting	0.238	0.360	0.395	0.410	0.575	0.572	0.136	0.349	0.407	0.014	0.047	0.093
Corpus Captioning (BLIP)	0.105	0.217	0.271	0.210	0.354	0.371	0.136	0.319	0.401	**0.054**	**0.122**	**0.153**
VISA Reranking (top-30)	0.223	0.378	0.388	0.240	0.443	0.457	0.000	0.000	0.000	0.000	0.000	0.000
Captioning Reranking (top-30)	0.200	0.336	0.367	0.160	0.362	0.390	0.000	0.000	0.000	0.000	0.000	0.000
VisRet	**0.251**	**0.431**	**0.438**	**0.460**	**0.632**	**0.605**	**0.170**	**0.452**	**0.455**	0.034	0.072	0.108
Retriever = E5-V												
Original Query	0.240	0.386	0.407	0.340	0.465	0.486	0.000	0.319	0.407	0.000	0.163	0.178
LLM Rewriting	0.223	0.368	0.391	0.460	0.569	0.566	0.170	0.367	0.412	0.038	0.133	0.182
Corpus Captioning (BLIP)	0.156	0.294	0.319	0.240	0.405	0.436	0.119	0.342	0.398	**0.064**	0.158	0.196
VISA Reranking (top-30)	0.279	0.413	0.419	0.350	0.495	0.500	0.022	0.002	0.002	0.000	0.000	0.000
Captioning Reranking (top-30)	0.248	0.369	0.396	0.400	0.429	0.454	0.000	0.000	0.005	0.000	0.000	0.000
VisRet	**0.299**	**0.452**	**0.461**	**0.560**	**0.643**	**0.622**	**0.220**	**0.377**	**0.425**	0.042	**0.173**	**0.205**

Table 2: nDCG (N@k) across four T2I retrieval benchmarks using different retrieval strategies and retrievers. Within each retriever group, the best results per column are bolded. We use different colors to denote similarity matching in **text-to-image**, **text-to-text**, or **image-to-image** style. VisRet achieves the state-of-the-art across most evaluation settings, especially excelling in handling knowledge-intensive queries from Visual-RAG and Visual-RAG-ME.

four benchmarks. Similar trends hold for E5-V, where VisRet yields gains of 0.078 (23%) and 0.052 (14%) in nDCG@10.

Text-to-text baselines behave less consistently across benchmarks. Corpus captioning with BLIP helps on COCO, which is closest to its training distribution, but does not improve over cross-modal retrieval on benchmarks with more knowledge-intensive queries. We hypothesize that this reflects the high information density in images and the corresponding loss of potentially useful details during captioning. In contrast, reranking approaches that operate on cross-modal retrieval results, including Captioning Reranking and VISA, outperform query-agnostic corpus captioning. However, they still do not outperform direct cross-modal retrieval by a large margin. These approaches are also constrained by cost, since they process only top-k candidates and their cost grows linearly with k. As a result, their performance is limited by the initial retrieval stage, leading to near-zero performance on INQUIRE-Rerank-Hard and COCO-Hard.

5.3 VisRet Improves Downstream QA

To assess VisRet in a RAG pipeline, we evaluate downstream VQA accuracy under three settings: (1) answering with only the model's internal knowledge, (2) RAG using retrieval from the original text query, and (3) RAG using VisRet. Figure 2 reports QA accuracy on Visual-RAG and Visual-RAG-ME with GPT-4o as the LVLM reader and CLIP as the retriever. The original query often

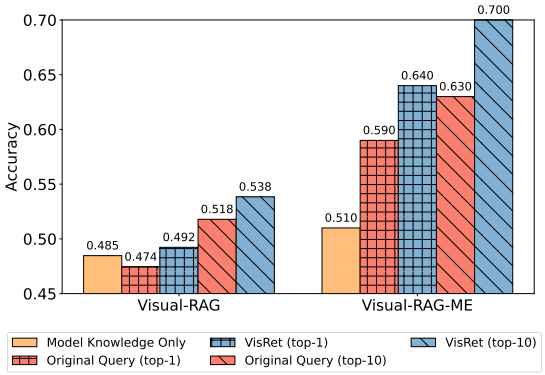


Figure 2: Downstream RAG-based VQA accuracy on Visual-RAG and Visual-RAG-ME with CLIP as the retriever and GPT-4o as the reader LVLM.

provides low-quality retrieval augmentation, and it slightly reduces performance on Visual-RAG in the top-1 retrieval setting relative to no retrieval, suggesting that inaccurate evidence can be more harmful than no evidence.

In contrast, VisRet improves QA accuracy in both top-1 and top-10 settings on both benchmarks, raising accuracy to 0.538 on Visual-RAG and 0.700 on Visual-RAG-ME. Notably, on Visual-RAG-ME, VisRet with top-1 retrieval outperforms the original-query setting even with top-10 retrieval, highlighting its ability to retrieve images that contain the required features with high precision. Overall, these results show that improvements in retrieval quality translate into tangible gains in downstream VQA performance.

6 Analyses

6.1 Robustness across modules

When VisRet is used in a RAG pipeline, four modules affect end-to-end performance: (1) visualization instruction generation, (2) query visualization, (3) retrieval, and (4) answer generation with an LVLM. We conduct analyses to probe the robustness of VisRet under variations of these modules. Overall, we find that changing the model for (1), (3), or (4) does not substantially alter VisRet's effectiveness, while performance is more sensitive to (2), which depends strongly on the capability of the T2I generation model. We first summarize the findings below:

1. Using both 8B and 70B Llama (Grattafiori et al., 2024) for instruction generation instead of GPT-4o, VisRet still outperforms using the LLM itself to rephrase the query (§C.1).
2. While most T2I models improve over cross-modality retrieval, strong proprietary T2I generation models yield substantially larger gains when used in VisRet (Table 3, §C.2). We provide further discussions below.
3. VisRet continues to deliver consistent gains when using a state-of-the-art embedding model (Table 9, §C.5).
4. Repeating the VQA experiments on Visual-RAG and Visual-RAG-ME with a stronger and a weaker reader LVLM, we find that VisRet's retrieval improvements consistently translate into downstream QA gains (§C.3).
5. Although the visualized queries improve retrieval, our preliminary study suggests they still cannot replace retrieved natural images in most cases (§C.4).

T2I Model Dependency Despite the retrieval performance can be improved with a range of T2I generation models, VisRet is limited by the fact that it generally benefits more from stronger T2I generation models. As shown in Table 3, the latest models from Google and OpenAI dominates the performance on Visual-RAG-ME by a large margin. Setting a lower quality for OpenAI Image-1 slightly harms the performance as well. What leads to the performance gap? In Figure 3, we present three representative cases where Image-1 outperforms Dall-E 3 (OpenAI, 2023a) on Visual-RAG-ME. We find three recurring failure modes:

- **Lack of focus** As shown in the top example, Dall-E 3 fails to zoom in and highlight the key feature wanted, i.e., a flower cluster.
- **Factuality issues** As shown in the middle example, Dall-E 3 fails to confidently render the bulb of the organism queries.
- **Weak instruction following** As shown in the bottom example, Dall-E 3 fails to comply with the instruction to show the abdomen of the organism, especially in the first visualization.

Method	N@1	N@10	N@30	Avg
`Retriever = CLIP`				
Original Query	0.220	0.423	0.435	0.359
VisRet Variants:				
Stable Diffusion 3.5	0.270	0.467	0.484	0.407
FLUX.1-dev	0.320	0.501	0.494	0.438
Qwen-Image	0.330	0.501	0.518	0.450
DALL-E 3	0.346	0.554	0.553	0.484
Gemini	0.410	0.579	0.583	0.524
Image-1 (low quality)	0.420	0.585	0.589	0.531
Image-1 (high quality)	**0.460**	**0.632**	**0.605**	**0.566**
`Retriever = E5-V`				
Original Query	0.340	0.465	0.486	0.430
VisRet Variants:				
Stable Diffusion 3.5	0.410	0.550	0.547	0.502
FLUX.1-dev	0.370	0.535	0.532	0.479
Qwen-Image	0.410	0.542	0.548	0.500
DALL-E 3	0.450	0.599	0.587	0.545
Gemini	**0.615**	0.585	**0.666**	0.589
Image-1 (low quality)	0.520	0.629	0.612	0.587
Image-1 (high quality)	0.560	**0.643**	0.622	**0.608**

Table 3: nDCG@k retrieval results on Visual-RAG-ME for different T2I generation models used in VisRet. We report N@1, N@10, N@30, and their average. Best numbers within each retriever are boldfaced. Gemini stands for the Gemini-2.5-Flash-Image-Preview model. The detailed experiment setup is presented in §C.2.

Among them, factuality reflects a core capability gap, while the other two issues can often be partially mitigated through prompting. For fairness, we use the same prompt across models in Table 3, but we expect that practical deployments can further benefit from model-specific prompt tuning.

Nevertheless, we argue that this limitation does not render VisRet inapplicable in real-world application. To begin with, VisRet is in fact efficient to serve because the image embedding index can be built once and reused across queries. In contrast, reranking baselines such as VISA incur much higher per-query cost because they rely on an LVLM to process top-k candidates, and their latency grows with k. In our experiments, VISA has roughly 5× higher latency than VisRet while

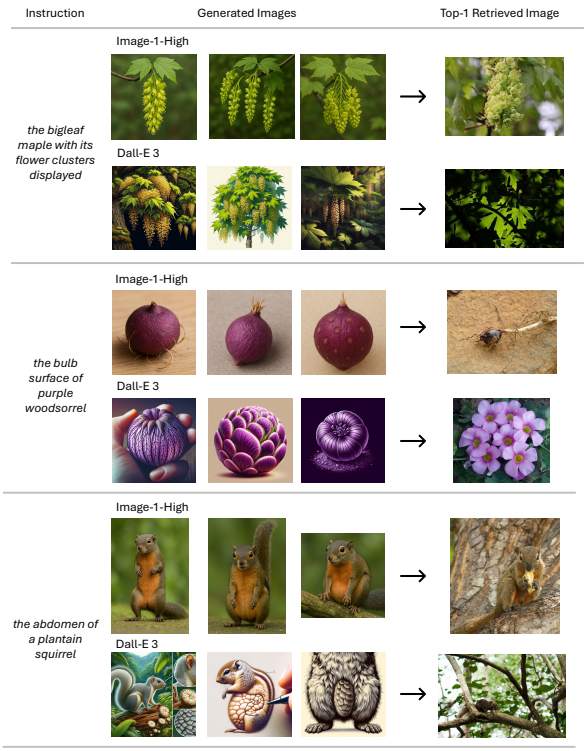


Figure 3: Examples of T2I generations from Image-1 and Dall-E 3 on Visual-RAG-ME. For these examples, VisRet with Image-1 achieves a nearly perfect retrieval score while VisRet with Dall-E 3 scores close to zero.

reranking only the top-30 candidates. In addition, in agentic settings, VisRet can be deployed as a client-side tool that is invoked on demand for difficult queries. This avoids changes to existing retrieval serving infrastructure, reduces client-side latency, and retains VisRet's performance gains.

6.2 Effectiveness of Repeated Runs

Is multi-image aggregation necessary for VisRet? As shown in Figure 4, for most examples, VisRet's N@30 outperforms at least two of its individual samples, highlighting the benefit of diversifying the query through multiple visualizations. At the same time, Table 4 shows that using only one sample only slightly degrades overall performance. This is because in many cases, such as example 2 and 3 in Figure 3, independently generated images share substantial information overlap. In such cases, a single visualization can be an efficient option.

6.3 Success and Failure Examples

We conclude the paper with a qualitative study. Table 5 presents examples where the visualization step captures subtle visual requirements implied by the original query. Additional examples are provided in §D. In many ground-truth images,

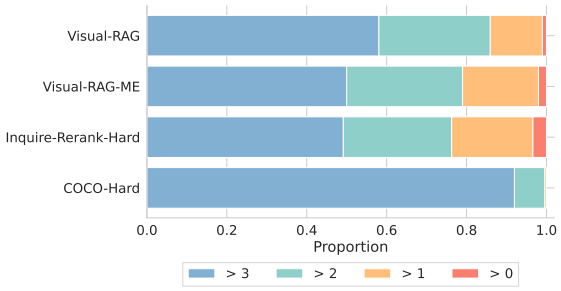


Figure 4: Distribution of counts where three-image VisRet outperforming its individual samples on N@30.

Benchmark	# Img	N@1	N@10	N@30	Avg
Retriever = CLIP					
Visual-RAG	3	0.251	0.431	0.438	**0.373**
	1	0.243	0.416	0.425	0.361
Visual-RAG-ME	3	0.460	0.632	0.605	**0.566**
	1	0.449	0.623	0.602	0.558
INQUIRE-Rerank-Hard	3	0.170	0.452	0.455	**0.359**
	1	0.166	0.432	0.457	0.352
COCO-Hard	3	0.034	0.072	0.108	**0.071**
	1	0.033	0.068	0.105	0.069
Retriever = E5-V					
Visual-RAG	3	0.299	0.452	0.461	**0.404**
	1	0.292	0.438	0.451	0.394
Visual-RAG-ME	3	0.560	0.643	0.622	**0.608**
	1	0.525	0.638	0.617	0.593
INQUIRE-Rerank-Hard	3	0.220	0.377	0.425	**0.341**
	1	0.205	0.292	0.426	0.308
COCO-Hard	3	0.042	0.173	0.205	**0.140**
	1	0.039	0.167	0.203	0.136

Table 4: Retrieval results (N@k) across benchmarks and image settings for two retrievers (CLIP and E5-V). Avg is the mean of N@1, N@10, and N@30.

the target entity appears in a specific posture that reveals the relevant evidence (Visual-RAG and Visual-RAG-ME), or it participates in complex relationships with other entities (INQUIRE and COCO). These constraints are challenging to be preserved in the embedding and surfaced during the cross-modal similarity matching. By contrast, VisRet incorporates them easily through query visualization, reducing the burden on the downstream embedding model.

In Figure 5, we show failure cases where VisRet does not outperform the LLM query rewriting baseline. We observe three salient error patterns:

- **Instruction Generation** As shown in the top row, retrieving a natural image that contains the answer requires two features: (1) the larva form and (2) the relevant body part being shown. The generated T2I instruction omits (1) and includes only (2). Addressing this issue likely requires prompt adjustments.

Dataset	Query	Ground Truth	Baseline Scores	Visualized Query	VisRet Scores
Visual-RAG	Does the Mountain Tree Frog (scientific name: Hyla eximia) have any distinctive pattern on the underside of its body?		Rank:49, nDCG@10: 0.00		Rank:4, nDCG@10: 0.39
	How many petals are on each of the Tower Mustard (scientific name: Turritis glabra)'s flowers?		Rank:143, nDCG@10: 0.00		Rank:2, nDCG@10: 0.76
Visual-RAG-ME	Which one has striped primary flight feathers, Willet (scientific name: *Tringa semipalmata*) or Grey-tailed tattler (scientific name: *Tringa brevipes*)?		Entity: Willet Rank:104, nDCG@10: 0.00		Entity: Willet Rank:1, nDCG@10: 0.72
			Entity: Grey-tailed tattler Rank:74, nDCG@10: 0.00		Entity: Grey-tailed tattler Rank:1, nDCG@10: 1.00
INQUIRE	A male and female cardinal sharing food		Rank:12, nDCG@10: 0.00		Rank:1, nDCG@10: 1.00
COCO	Plate with pizza, knife and fork laying on edge of plate with two glasses next to them.		Rank:39, nDCG@10: 0.00		Rank:3, nDCG@10: 0.3937

Table 5: Examples: VisRet improves retrieval by highlighting visual features implied by the textual query. "Baseline" refers to cross-modal retrieval with the original query. CLIP embedding scores are used for this visualization.

Query	Instruction	Visualization	Target
The larva of Columbia silk moth (scientific name: Hyalophora columbia) and that of tulip-tree silkmoth (scientific name: Callosamia angulifera), which one has more colorful sparse protuberance on its body?	*the tulip-tree silkmoth with its body shown*		
Mistle Thrush (scientific name: Turdus viscivorus) and Fieldfare (Turdus pilaris) which one has black dots on the underside of its tail feather?	*the Fieldfare with its tail feather underside shown*		
Does the underside of an adult Meso-American slider (scientific name: Trachemys venusta) shell have more blocks of scute than that of an adult northern map turtle (scientific name: Graptemys geographica)?	*the shell underside of northern map turtle*		

Figure 5: Examples of VisRet failing to outperform the LLM query rewriting baseline. Error patterns span T2I instruction generation, visualization, and retrieval failure.

- **Visualization** As shown in the second row, the instruction correctly identifies that the organism's *underside* must be shown, but the T2I model fails to generate a perspective that reliably highlights the feature.

- **Retrieval** Even when instruction generation and visualization succeed, retrieval can still fail when the embedding model is sensitive to variations such as background or object appearance. In the bottom-row example, both the background (table versus human hand and ground) and the object form (alive versus dead) differ, which leads to retrieval failure.

7 Conclusion

This work introduces VisRet, a novel framework that visualizes text queries to reframe cross-modality T2I retrieval as within-modality retrieval. By operating entirely in the visual domain, VisRet addresses key limitations of cross-modal embedding alignment. Our experiments confirm that visualized queries substantially improve both retrieval precision and downstream VQA accuracy across four benchmark datasets and two retrievers. The simplicity and modularity of VisRet open up promising directions for future knowledge-intensive multimodal systems.

Ethics Statement

In this section, we describe the ethical considerations related to this paper.

Potential Risks Although the goal of this paper is to introduce techniques to improve the text-to-image retrieval performance, the new approach could create new social risks. Specifically, in addition to the neural embedding model, our approach involves two neural models: an LLM and a T2I generation model. It is possible for these large models to bring in new social bias in generating the visualize query and thus bias the retrieval results. For instance, when depicting certain scenes of social activity, the models could reinforce stereotypical social roles. We urge practitioners to implement model debiasing and bias detection measure when deploying our proposed T2I retrieval method in real-world applications.

Artifact Release Our Visual-RAG-ME annotation is based on Visual-RAG, which is under CC BY-NC 4.0 license and the images shared by the iNaturalist 2021 dataset, which are under one of CC BY 4.0, CC BYNC 4.0, CC BY-NC-ND 4.0, CC BY-NC-SA 4.0, CC0 1.0, CC BY-ND 4.0, CC BY-SA 4.0. We adhere to the intended non-commercial research use of iNaturalist 2021 dataset and do not re-distribute the images. Following Visual-RAG, we will release our Visual-RAG-ME annotations under CC BY-NC 4.0 license.

Human Annotation Two authors, who are graduate students studying Natural Language Processing, are the only annotators involved in Visual-RAG-ME annotation. Both annotators are supported by the research stipend and the annotation work counted into the working hours. Consent was obtained from both annotators before benchmark curation. The entire benchmark creation process was automatically determined exempt by the institution's IRB policy. The annotators actively discussed whenever they encounter ambiguity during annotation and reached agreements before proceeding. After the benchmark annotation, we performed a round of human auditing to ensure no question may cause privacy or ethics concerns.

AI Assistant Use AI assistants, specifically ChatGPT, are used only for revising the paper draft, fixing grammar mistakes, and improving the outlook of the figures.

Acknowledgment

The research is supported in part by Taboola. We also thank Jia-Chen Gu and Hongwei Wang for their valuable feedback.

References

Black Forest Labs. 2024. Announcing black forest labs. https://bfl.ai/blog/24-08-01-bfl. News, published August 1, 2024.

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2024. DEFAME: dynamic evidence-based fact-checking with multimodal experts. *CoRR*, abs/2412.10510.

Yingshan Chang, Guihong Cao, Mridu Narang, Jianfeng Gao, Hisami Suzuki, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal QA. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16474–16483. IEEE.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5558–5570. Association for Computational Linguistics.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 758–759. ACM.

Guofeng Ding, Yiding Lu, Peng Hu, Mouxing Yang, Yijie Lin, and Xi Peng. 2025. Visual abstraction: A plug-and-play approach for text-visual retrieval. In *Forty-second International Conference on Machine Learning*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: improving visual-semantic embeddings with hard negatives. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 12. BMVA Press.

Chun-Mei Feng, Yang Bai, Tao Luo, Zhen Li, Salman H. Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Yong Liu. 2025. VQA4CIR: boosting composed image retrieval with visual question answering. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 2942–2950. AAAI Press.

Alisa Fortin, Guillaume Vernade, Kat Kampf, and Ammaar Reshi. 2025. Introducing gemini 2.5 flash image, our state-of-the-art image model. https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/. Google Developers Blog; posted Aug. 26, 2025.

Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomás Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2121–2129.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.*, 47:853–899.

Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge J. Belongie, and Oisin Mac Aodha. 2021. Benchmarking representation learning for natural world image collections. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12884–12893. Computer Vision Foundation / IEEE.

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. 2024. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. *CoRR*, abs/2410.08182.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024. E5-V: universal embeddings with multimodal large language models. *CoRR*, abs/2407.12580.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4933–4944. Association for Computational Linguistics.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. 2025. CLIP behaves like a bag-of-words model cross-modally but not uni-modally. *CoRR*, abs/2502.03566.

Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. 2024. Interactive text-to-image retrieval with large language models: A plug-and-play approach. In *Proceedings of the 62nd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 791–809. Association for Computational Linguistics.

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. 2023. Chatting makes perfect: Chat-based image retrieval. *Advances in Neural Information Processing Systems*, 36:61437–61449.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Yongqi Li, Wenjie Wang, Leigang Qu, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. 2024. Generative cross-modal retrieval: Memorizing images in multimodal language models for retrieval and beyond. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11851–11861. Association for Computational Linguistics.

Zheyuan Liu, Weixuan Sun, Damien Teney, and Stephen Gould. 2024. Candidate set re-ranking for composed image retrieval with dual multi-modal encoder. *Trans. Mach. Learn. Res.*, 2024.

OpenAI. 2023a. Dall·e 3. `https://openai.com/index/dall-e-3/`. Accessed: 2025-05-18.

OpenAI. 2023b. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

OpenAI. 2025. Gpt image 1. `https://platform.openai.com/docs/models/gpt-image-1`. Accessed: 2025-05-18.

Leigang Qu, Haochuan Li, Tan Wang, Wenjie Wang, Yongqi Li, Liqiang Nie, and Tat-Seng Chua. 2025. Tiger: Unifying text-to-image generation and retrieval with large multimodal models.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

S. E. Robertson and S. Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94*, pages 232–241, London. Springer London.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. 2023. knn-diffusion: Image generation via large-scale retrieval. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel J. Brostow, Kate E. Jones, Oisin Mac Aodha, Sara Beery, and Grant Van Horn. 2024. INQUIRE: A natural world text-to-image retrieval benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Xiaodan Wang, Chengyu Wang, Lei Li, Zhixu Li, Ben Chen, Linbo Jin, Jun Huang, Yanghua Xiao, and Ming Gao. 2023. Fashionklip: Enhancing e-commerce image-text retrieval with fashion multi-modal conceptual knowledge graph. In *Proceedings of the The 61st Annual Meeting of the Association for Computational Linguistics: Industry Track, ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 149–158. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. 2025a. Qwen-image technical report. *Preprint*, arXiv:2508.02324.

Yin Wu, Quanyu Long, Jing Li, Jianfei Yu, and Wenya Wang. 2025b. Visual-rag: Benchmarking text-to-image retrieval augmented generation for visual knowledge intensive queries. *CoRR*, abs/2502.16636.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics*, 2:67–78.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.*, 2022.

Mert Yüksekgönül, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *The*

Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

Supplementary Material: Appendices

A Benchmark Data Details

In this section, we present the details of data processing for Visual-RAG, INQUIRE-Rerank-Hard, and COCO-Hard.

A.1 Visual-RAG

Visual-RAG (Wu et al., 2025b) releases 391 queries with associated image names from iNaturalist 2021 (Horn et al., 2021) and corresponding retrieval labels[2]. To prepare the data, we download the original iNaturalist 2021 dataset and re-collect the images from the train and test sets. We were able to identify all annotated images in Visual-RAG except for a single image due to a likely path error.

A.2 INQUIRE-Rerank-Hard

To prepare INQUIRE-Rerank-Hard, we accessed the publicly released INQUIRE-Rerank (Vendrow et al., 2024) dataset[3]. The original test set contained 160 queries, each paired with 100 images retrieved by CLIP. In a pilot study, we tested the retrieval performance of off-the-shelf CLIP and E5-V models. Results showed that CLIP can achieve 0.438 Recall@1 while E5-V achieved 0.506 Recall@1. After manually inspecting the data, we found that for many instances, the negative images are not challenging enough and it is often very straightforward to identify the ground-truth images. To highlight the challenging questions, we therefore filtered out the questions for which either CLIP or E5-V achieves a perfect Recall@1. Overall, we observe that the remaining 59 questions require more nuanced image context understanding and a higher level of knowledge of the organisms themselves, with more challenging confounding negative images.

A.3 COCO-Hard

We derive COCO-Hard from the publicly released validation set of Microsoft COCO 2017 (Chen et al., 2015). The original validation set contained 5000 queries, and we directly use the ground-truth images to construct the image corpus. In a pilot study, we found that CLIP is able to achieve 0.424 Recall@1 and 0.756 Recall@10, while E5-V has 0.594 Recall@1 and 0.892 Recall@10. A manual analysis shows that the queries of most examples in the original COCO dataset are short phrases centered on a single entity. To increase the task's difficulty, we rank the examples in reverse order by nDCG@30 and retain the hardest 500 instances. We find that these hard queries often describe complex spatial relationships among multiple entities.

B Implementation Details

In this section, we present the implementation details of VisRet and baselines.

VisRet: T2I Generation To project the text query into the image space, we first instruct an LLM to analyze the query and highlight the key visual features in implies. The prompts for three benchmarking datasets are shown in Figure 6, Figure 7, and Figure 8, respectively. Then, we wrap the rephrased query in a prompt template "`Generate a small image of the {rephrased_query}`" to obtain the final instruction for T2I generation. We use the model `gpt-4o-2024-08-06` via OpenAI API with temperature = 0 for instruction generation and the `gpt-image-1` model for T2I generation with the quality flag set to "high". For generating multiple images for each query, we find that calling the `gpt-image-1` API to return multiple images given a single instruction already results in images with a high level of diversity. Therefore, we followed this setting in this paper and save further perturbing the instruction as future work.

VisRet: Retrieval After obtaining the generated visualizations, we encode both the visualized images and the image corpus via an off-the-shelf CLIP[4] or E5-V[5] encoder and perform a similarity search. Cosine similarity is used as the similarity metric. For RRF, we use $\lambda = 1$ to merge the rankings from multiple queries. All the retrieval experiments were performed on a local server with NVIDIA A100 GPU.

VQA Answer Generation We slightly modify the prompt in Visual-RAG to use chain-of-thought prompting (Wei et al., 2022). Concretely, the model is asked to always extract visual information,

[2] https://github.com/visual-rag/visual-rag

[3] https://huggingface.co/datasets/evendrow/INQUIRE-Rerank

[4] https://huggingface.co/openai/clip-vit-large-patch14-336

[5] https://huggingface.co/royokong/e5-v

perform reasoning, and conclude their reasoning with self-verification. We provide the full prompt in Figure 9 and Figure 10.

VQA Evaluation For Visual-RAG and Visual-RAG-ME, we use the same evaluation prompt released by the authors of Visual-RAG (Wu et al., 2025b), as shown in Figure 11. Since this prompt is already human-engineered for evaluating more complex references and long-form answers and the answers of Visual-RAG-ME are short and easy to evaluate, we did not perform additional prompt engineering. We use the same prompt and `gpt-4o-2024-08-06` as the LLM judge for all the VQA experiments in this paper.

Baselines For the **LLM rephrase** baseline, we use the same prompt for VisRet T2I instruction to highlight the most important features that the query is seeking. At an early stage of the project, we performed manual tuning on the prompt and found that the best-performing rephrase also serves as the best-performing T2I generation instruction. Therefore, we report the results using the same prompt for the final version for the paper. For the **Corpus Captioning** baseline, we use the `Salesforce/blip-image-captioning-large` model downloaded from huggingface. We follow the recommended parameters to run inference on local GPUs, except setting the minimum output length to 10 tokens to encourage a more detailed caption. For the **VISA Reranking** baseline, we followed the original paper's public implementation[6], using the sample prompt except changing the model to `Qwen2.5-VL-72B-Instruct` which we empirically find performs better. For the **Captioning Reranking** baseline, we simplify the VISA pipeline by directly prompting for a caption for each of the top retrieved images using the following prompt:

"Write a caption that describes the appearance and posture of the main subject. The caption should be concise but informative so that it can be used to retrieve the images of the same subject in different postures, states, or environments."

C Further Analyses

In this section, we provide more comprehensive analyses to investigate the effectiveness of VisRet from additional perspectives, including the choices of T2I instruction LLM, T2I generation model, and the downstream VQA reader LVLM. Finally, inspired by the generative retrieval literature, we conduct a pilot study of whether the generated images could be directly used as the knowledge context for downstream question answering.

C.1 Query Generation Model Choice

Does VisRet work well with other LLMs as the T2I instruction generator? In Table 6, we study two differently sized open-weight LLMs for rephrasing the query and generating the T2I instruction: Llama 3.1 8B Instruct and Llama 3.3 70B Instruct (Grattafiori et al., 2024). Overall, we observe promising results: For all three LLMs, using the LLMs themselves to generate T2I instructions for VisRet outperforms using the LLMs themselves to rephrase the query. While more expensive LLMs achieve higher performance, the small 8B Llama model can already achieve decent performance at a similar level as GPT-4o.

C.2 T2I Generation Model Choice

How strong does the T2I generation model need to be for VisRet to work well? We compare the default T2I generation model (Image-1 with high quality setting) with several other commonly used, strong T2I generation models:

- Stable Diffusion 3 (Esser et al., 2024).
- FLUX-1.dev (Black Forest Labs, 2024).
- Qwen-Image (Wu et al., 2025a).
- DALL-E 3 API (OpenAI, 2023a).
- Gemini-2.5-Flash-Image-Preview API (Fortin et al., 2025).
- Image-1 API with the low quality setting.

Table 3 shows the results with GPT-4o as the T2I instruction generation model and CLIP as the retriever model. We observe that while all the models can improve over cross-modal retrieval to some extent in the VisRet pipeline, their effectiveness increases with image quality. The best performance is achieved by the latest proprietary models Image-1 and Gemini-2.5-Flash-Image-Preview. Together, these results suggest that a good T2I generation model with strong instruction-following ability is necessary for VisRet. As further T2I generation

[6] https://github.com/XLearning-SCU/2025-ICML-VISA

Retrieval Strategy	LLM	R@1	R@10	R@30	N@1	N@10	N@30
Original Query	-	0.210	0.583	0.737	0.210	0.355	0.385
LLM Rewriting	Llama 3.1 8B Instruct	0.238	0.563	0.737	0.238	0.365	0.385
	Llama 3.3 70B Instruct	0.240	0.575	0.742	0.240	0.377	0.399
	GPT-4o	0.238	0.586	0.737	0.238	0.360	0.395
Visualize-then-Retrieve	Llama 3.1 8B Instruct	0.243	0.606	0.780	0.243	0.405	0.428
	Llama 3.3 70B Instruct	**0.256**	0.627	0.790	**0.256**	0.413	0.437
	GPT-4o	0.251	**0.645**	**0.793**	0.251	**0.431**	**0.438**

Table 6: Retrieval performance on Visual-RAG of with CLIP retriever, using different LLMs as T2I instruction generator for VisRet. R = Recall. N = nDCG. The best results are boldfaced.

Knowledge	Visual-RAG				Visual-RAG-ME			
	# images	GPT-4o-mini	GPT-4o	GPT-4.1	# images	GPT-4o-mini	GPT-4o	GPT-4.1
Model Knowledge Only	0	0.385	0.485	0.492	0	0.410	0.510	0.470
Direct T2I Retrieval	1	0.409	0.474	0.515	2	0.490	0.590	0.610
	10	0.460	0.518	0.571	10	0.480	0.630	0.650
Visualize-then-Retrieve	1	0.418	0.492	**0.572**	2	0.530	0.640	0.620
	10	**0.464**	**0.538**	0.567	10	**0.550**	**0.700**	**0.710**

Table 7: VQA performance comparison using different LVLMs as instruction generators for VisRet and query rephrase models. CLIP is used as the retriever. Boldfaced numbers indicate the best in each column.

methods improve, we anticipate that building more cost-efficient versions of VisRet is viable and promising.

C.3 Downstream VQA Model Choice

While we have shown the benefit of VisRet for VQA for GPT-4o, does the improvement hold across LVLMs with different capabilities? In Table 7, we repeat the VQA experiments with two additional LVLMs: GPT-4o-mini (version `gpt-4o-mini-2024-07-18`) and GPT-4.1 (version `gpt-4.1-2025-04-14`). Overall, we observe similar trends as those presented in Figure 2. Both direct T2I retrieval and VisRet outperform relying only on the model's knowledge, with VisRet substantially outperforming the former. These results form the foundation for VisRet as a general plug-and-play method to enhance RAG pipelines that rely on accurate T2I retrieval.

C.4 Image Queries as Knowledge

As demonstrated by previous results, a T2I generation model with strong ability to follow instructions and generate realistic images is crucial to the success of VisRet. A natural question is whether it is still necessary to perform retrieval rather than directly using the generated images as the knowledge? In Table 8, we compare the performance of using a single image as the context with VisRet. Overall, we observe mixed results. For Visual-RAG, GPT-4o-mini achieves slightly higher performance with the generated image than top-1 retrieval, but GPT-4o and GPT-4.1 exhibit the reverse pattern. For Visual-RAG-ME, both GPT-4o-mini and GPT-4.1 prefer the generated image to top-1 retrieval (and even top-10 retrieval). However, when provided with top-10 retrieved images, the models generally exhibit higher VQA performance than when using the generated image. Therefore, we conclude that retrieving natural images is still crucial for challenging VQA tasks like Visual-RAG and Visual-RAG-ME and cannot be fully replaced by pure T2I generation at this stage. An important direction for future work is to combine image generation and image retrieval to improve the quality of the retrieved knowledge.

C.5 Stronger Embedding Models

In Table 9, we provide additional results using RzenEmbed-v2-7B[7], which is one of the best models on MMEB at the time of writing (2025 December). VisRet still strongly outperforms most baselines in most settings except captioning reranking for Visual-RAG's N@30. This result highlights both weaker and stronger embedding models benefit from query visualization.

[7] https://huggingface.co/qihoo360/RzenEmbed

Knowledge	Visual-RAG				Visual-RAG-ME			
	# images	GPT-4o-mini	GPT-4o	GPT-4.1	# images	GPT-4o-mini	GPT-4o	GPT-4.1
Model Knowledge Only	0	0.385	0.485	0.492	0	0.410	0.510	0.470
Generated Image (Image-1)	1	0.431	0.425	0.444	2	**0.590**	0.580	**0.800**
Visualize-then-Retrieve	1	0.418	0.492	**0.572**	2	0.530	0.640	0.620
	10	**0.464**	**0.538**	0.567	10	0.550	**0.700**	0.710

Table 8: VQA performance comparison using different knowledge contexts on Visual-RAG and Visual-RAG-ME. CLIP is used as the retriever. Boldfaced numbers indicate the best in each column.

Retrieval Method	Visual-RAG			Visual-RAG-ME		
	N@1	N@10	N@30	N@1	N@10	N@30
Original Query	0.279	0.416	0.420	0.330	0.546	0.543
LLM Rewriting	0.302	0.414	0.420	0.390	0.529	0.560
Corpus Captioning (BLIP)	0.128	0.260	0.310	0.220	0.375	0.418
VISA Reranking (top-30)	0.288	0.431	0.442	0.380	0.547	0.541
Captioning Reranking (top-30)	0.280	0.465	**0.496**	0.140	0.250	0.257
VisRet	**0.335**	**0.476**	0.480	**0.550**	**0.667**	**0.642**

Table 9: Additional evaluation results for `Retriever = RzenEmbed-v2-7B`. nDCG (N@k) is reported. VisRet still strongly outperforms most baselines in most settings except captioning reranking for Visual-RAG's N@30.

D Further Qualitative Examples

In Table 10, we present additional qualitative examples on the four evaluation datasets. Consistent with quantitative findings, VisRet reduces the embedding model's workload by expressing the important details in a visualization that is closer to the retrieval target, resulting in substantially better performance in terms of ground-truth rank and nDCG@10.

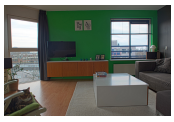
Dataset	Query	Ground Truth	Baseline Scores	Visualized Query	VisRet Scores
Visual-RAG	What is the color of the head of larva Urania Swallowtail Moth (scientific name: Urania fulgens)?		Rank:96, nDCG@10: 0.00		Rank:1, nDCG@10: 0.66
	What color are the ventral abdomen of Golden Buprestid Beetle (scientific name: Buprestis aurulenta)?		Rank:23, nDCG@10: 0.00		Rank:4, nDCG@10: 0.39
	Are there visually distinctive scales on feet of Azure Tit (scientific name: Cyanistes cyanus)?		Rank:26, nDCG@10: 0.40		Rank:2, nDCG@10: 0.76
INQUIRE	Mexican grass-carrying wasp visiting a purple flower		Rank:45, nDCG@10: 0.00		Rank:4, nDCG@10: 0.39
	great golden digger wasp carrying an orthopteron		Rank:12, nDCG@10: 0.45		Rank:2, nDCG@10: 0.83
Visual-RAG-ME	Which one has striped primary flight feathers, Willet (scientific name: Tringa semipalmata) or Grey-tailed tattler (scientific name: Tringa brevipes)?		Entity: Willet Rank:104, nDCG@10: 0.00		Entity: Willet Rank:1, nDCG@10: 0.72
			Entity: Grey-tailed tattler Rank:74, nDCG@10: 0.00		Entity: Grey-tailed tattler Rank:1, nDCG@10: 1.00
	Which one has less prominent color patterns, silvery checkerspot (scientific name: Chlosyne nycteis) caterpillar or theona checkerspot (scientific name: Chlosyne theona) caterpillar?		Entity: silvery checkerspot Rank:188, nDCG@10: 0.45		Entity: silvery checkerspot Rank:4, nDCG@10: 0.88
			Entity: theona checkerspot Rank:120, nDCG@10: 0.52		Entity: theona checkerspot Rank:3, nDCG@10: 0.85
COCO	A den with a large coffee table, couch and a television.		Rank:61, nDCG@10: 0.00		Rank:4, nDCG@10: 0.3937
	A computer workstation with a laptop and a desktop computer.		Rank:37, nDCG@10: 0.00		Rank:1, nDCG@10: 1.0

Table 10: Additional qualitative results on the three benchmarking datasets.

```
You are given a query, rephrase the query into a short descriptive phrase that highlights the key part
of the entity where the queried feature could be found. DO NOT include the asked feature (shape, color,
etc.) but instead include the part of the entity where the feature could be found. Output only the
rephrased query.

Examples:

Original query: What shape are the flowers of bush flax (scientific name: Astelia fragrans)?

Rephrased query: flowers of bush flax

Original query: Is the any specific color pattern on underside wings of tawny pipit (scientific name:
Anthus campestris) displayed during flight, or is it uniformly colored?

Rephrased query: tawny pipit with its underside wings shown

Original query: {question}

Rephrased query:
```

Figure 6: Prompt for instructing an LLM to generate the T2I generation instruction for Visual-RAG questions.

```
You are given a query about two entities, as well as an entity of interest. Rephrase the query into a
short descriptive phrase that highlights the key part of the entity of interest on which the queried
feature could be found. DO NOT include the asked feature (shape, color, etc.) but instead include the
entity name + part of the entity where the feature could be found. Output only the rephrased query.

Examples:

Original query: Are the tongues of grass snake (scientific name: Natrix helvetica) and Chicken Snake
(scientific name: Spilotes pullatus) the same color?

Entity of interest: Spilotes pullatus

Rephrased query: Chicken Snake with its tongue shown

Original query: Which one has a more slender matured legume, common milkpea (scientific name: Galega
officinalis) or narrowleaf lupin (scientific name: Lupinus angustifolius)?

Entity of interest: Galega officinalis

Rephrased query: the legume of common milkpea

Original query: {question}

Entity of interest: {entity}

Rephrased query:
```

Figure 7: Prompt for instructing an LLM to generate the T2I generation instruction for Visual-RAG-ME questions.

```
You are given an image retrieval query, rephrase the query to add in a bit detail (no longer than 30
words). The rephrased query should highlight the appearance, posture, actions of the main entity so
that it is easier to retrieve the desired image among (1) images of the same entity with different
posture and (2) images of different entities with the same posture.

Original query: {question}

Rephrased query:
```

Figure 8: Prompt for instructing an LLM to generate the T2I generation instruction for INQUIRE-Rerank-Hard and COCO-Hard questions.

```
Given a question from the user regarding a visual feature of an organism (animal, plant, etc.), answer
it using systematic reasoning. You will be provided with one or more images that may contain the key
information for answering the question. Your output should consist of two parts.

1. Reasoning:
- Look at the image carefully. Pick out the feature that can help you correctly answer the question.
- If no useful information can be inferred from the image, you should summarize your own knowledge
related to the question.
- If the image contradicts your own knowledge, you should trust the image.
- If the image is blurry, you should summarize your own knowledge related to the question.

2. Answer:
- Only your conclusion that directly answers the question.
- No need to repeat the reasoning.

Please always follow the answer format without bolding texts: "### Reasoning: {reasoning}\n### Answer:
{your_answer}"
```

Figure 9: Prompt for VQA on Visual-RAG.

```
You are a model that rigorously answers a question that compares a visual feature of two organisms
(animal, plant, etc.) using systematic reasoning. You will be provided with one or more images of both
organisms that may contain the key information for answering the question. Your output should consist
of two parts.

1. Reasoning:
- Look at the images carefully. Pick out the features that can help you correctly answer the question.
- If no useful information can be inferred from the image, you should summarize your own knowledge
related to the organism.
- If the image contradicts your own knowledge, you should trust the image.
- If the image is blurry, you should summarize your own knowledge related to the question.
- Then, compare the features of the two organisms and reason through the question step by step.
- Finally, conclude your reasoning with a verification step that confirms the correctness of your answer
based on the evidence you have gathered.

2. Answer:
- Only your conclusion that directly answers the question.
- No need to repeat the reasoning.

Please always follow the answer format without bolding texts: "### Reasoning: {reasoning}\n### Answer:
{your_answer}"
```

Figure 10: Prompt for VQA on Visual-RAG-ME.

```
Please evaluate the answer to a question, score from 0 to 1. The reference answer is provided, and
the reference is usually short phrases or a single keyword. If the student answer is containing the
keywords or similar expressions (including similar color), without any additional guessed information,
it is full correct. If the student answer have missed some important part in the reference answer,
please assign partial score. Usually, when there are 2 key features and only 1 is being answered,
assign 0.5 score; if there are more than 2 key features, adjust partial score by ratio of correctly
answered key feature. The reference answer can be in the form of a Python list, in this case, any one
of the list item is correct.

If student answer contain irrelevant information not related to question, mark it with "Redundant",
but it does not affect score if related part are correct. (e.g. Question: what shape is leave of
Sanguinaria canadensis, Student Answer: shape is xxx, color is yyy, this is Redundant answer)

If student answer contain features not listed in reference answer, mark it with "Likely Hallucination"
and deduct 0.5 score. (e.g., Reference Answer: black and white. Student Answer: black white, with
yellow dots, “yellow dots” is not mentioned in reference). The reference answer sometimes contains an
add-on enclosed by brackets (), to help verifying hallucinations (e.g.: “shape is xxx (color is yyy)”).
Not mentioning add-on information in answer is not considered wrong. Answering "I don’t know", "Not
enough information" is considered wrong.

Format Instructions: Separate the remarks with score using "|", that is, use the syntax of: "Score:
{score} | Likely Hallucination", "Score: {score}", "Score: {score} | Likely Hallucination | Redundant".
If any explanation on why giving the score is needed, do not start a new line and append after remark
with brackets, e.g. "Score: {score} | Redundant | (Explanation: abc)".

Following are few examples:

Question: Is there any specific color marking around the eyes of a semipalmated plover (scientific
name: Charadrius semipalmatus)?
Reference Answer: black eye-round feather, white stripe above eyes. (sometimes connected to the white
forehead)

Student Answer: Yes, the bird has a distinctive black line that runs through the eye, which is a key
identifying feature.
Score: 0 | Likely Hallucination

Student Answer: They have a black vertical band in front of the eye, a white band above the eye, and a
single black band that wraps partially around the eye, creating a partial "mask" appearance.
Score: 1

Student Answer: Yes, the semipalmated plover has a distinctive black/dark ring around its eye, surrounded
by a bright white ring or patch
Score: 0.5 | Likely Hallucination (Explanation: not white ring, but only a line above the eye)

Question: What is the typical color of the antennae of Harris’s checkerspot butterfly (scientific name:
Chlosyne harrisii)?
Reference Answer: alternating black and white band, with yellow on the tip

Student Answer: The antennae of Harris’s checkerspot butterfly are black with orange-tipped clubs.
Score: 0.5 (Explanation: not mentioning black and white)

Student Answer: The typical color of the antennae of Harris’s checkerspot butterfly is black with white
spots.
Score: 0.5 | Likely Hallucination (Explanation: not white spot but band. Not mentioning the tip)

Question: Are the leaves of burro-weed (scientific name: Ambrosia dumosa) usually covered in small
hairs?
Reference Answer: yes

Student Answer: Yes, the leaves of burro-weed (Ambrosia dumosa) are typically covered in small hairs,
giving them a grayish or whitish-green appearance.
Score: 1 | Redundant

Now, score the following question:

Question: {question}
Reference Answer: {reference_answer}

Student Answer: {model_answer}
```

Figure 11: Prompt for the LLM VQA judge used for Visual-RAG and Visual-RAG-ME.