Visualized Text-to-Image Retrieval

Anonymous ACL submission

Abstract

We propose Visualize-then-Retrieve (*VisRet*), a new paradigm for Text-to-Image (T2I) retrieval that mitigates the limitations of cross-modal similarity alignment of existing multi-modal embeddings. *VisRet* first projects textual queries into the image modality via T2I generation. Then, it performs retrieval within the image modality to bypass the weaknesses of cross-modal retrievers in recognizing subtle visual-spatial features. Experiments on three knowledge-intensive T2I retrieval benchmarks, including a newly introduced multi-entity benchmark, demonstrate that *VisRet* consistently improves T2I retrieval by 24.5% to 32.7% NDCG@10 across different embedding models. *VisRet* also significantly benefits downstream visual question answering accuracy when used in retrieval-augmented generation pipelines. The method is plug-and-play and compatible with off-the-shelf retrievers, making it an effective module for knowledge-intensive multi-modal systems.

1 Introduction

Text-to-Image (T2I) retrieval is the task of selecting the most relevant images from a visual corpus based on a textual query. It plays a crucial role in enabling knowledge-intensive applications that require supporting textual inputs with rich visual content (Chen et al., 2022; Wang et al., 2023; Sheynin et al., 2023; Braun et al., 2024).

A common approach to T2I retrieval is to embed both the query and candidate images into a shared representation space, where similarity scores are computed (Frome et al., 2013; Kiros et al., 2014). However, obtaining accurate similarity rankings that capture fine-grained semantics in both text and image remains a long-standing challenge. Prior studies have observed that cross-modal embeddings often behave like "bags-of-concepts", failing to model structured relationships among visual elements (Yüksekgönül et al., 2023; Kamath et al., 2023). For instance, Figure 1 presents a query that requires images of an entity (a Barnacle Goose) at specific postures (wings unfolded) to answer. While the embedding model succeeds at matching the entity type, it struggles to recognize subtler visual-spatial features such as the pose of the wing (unfolded) and the camera perspective (upshot). To address these limitations, existing work has explored improving the embedding quality (Radford et al., 2021; Yu et al., 2022), query reformulation (Levy et al., 2023), and multi-stage reranking pipelines (Liu et al., 2024; Feng et al., 2025). Yet, all these strategies are ultimately constrained by the intrinsic difficulty of cross-modal similarity alignment, as they cannot bypass the stage of text-to-image similarity search.

We propose *Visualize-then-Retrieve (VisRet)*, a novel retrieval paradigm that decomposes T2I retrieval into two stages: *text-to-image modality projection* followed by *within-modality retrieval*. Concretely, the textual query is first visualized as one or more images via a T2I generation model. Then, the visualized query, which better exhibits the desired visual-spatial features, is used to perform image-to-image retrieval.

Compared to prior methods, *VisRet* offers two key advantages. First, visualizations provide a more expressive and intuitive medium for encoding multiple compositional concepts such as entities, poses, and spatial relations, which are difficult to express via text alone. As shown in Figure 1, the visualized query is able to accurately 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 instead leverages the stronger capacity of these retrievers in uni-modal retrieval (Koishigarina et al., 2025).

We evaluate *VisRet* on three challenging T2I retrieval benchmarks: INQUIRE-Rerank (Vendrow et al., 2024), Visual-RAG (Wu et al., 2025), and

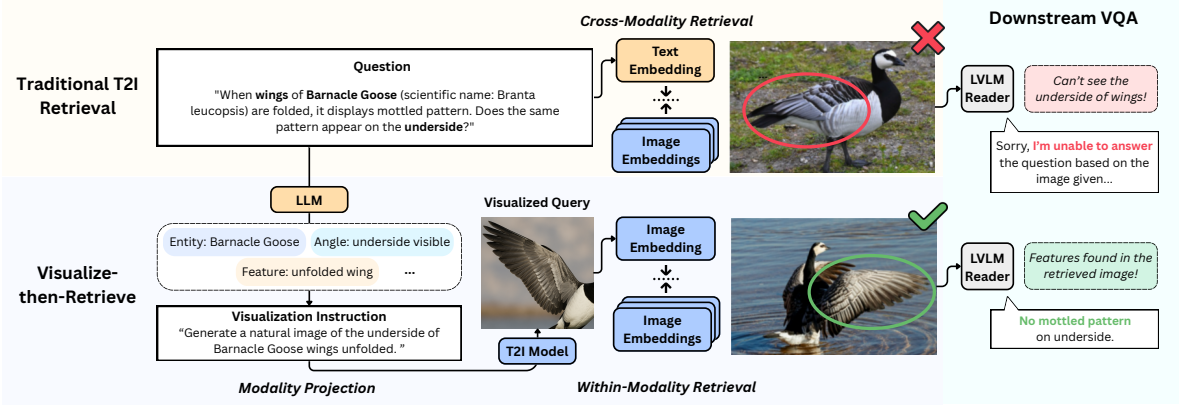


Figure 1: An overview of *VisRet*. Compared to the traditional T2I retrieval pipeline, *VisRet* first projects the text query into the image modality via T2I generation and then performs within-modality retrieval.

Visual-RAG-ME, a new benchmark we introduce that features feature comparison questions across *multiple entities*. Results show that *VisRet* substantially outperforms baseline T2I retrieval methods and a strong LLM-based query rewriting approach (§4.2). When CLIP (Radford et al., 2021) is used as the retriever, *VisRet* outperforms the two baselines by 32.7% and 15.6% higher NDCG@10 respectively, averaged over three benchmarks. With E5-V (Jiang et al., 2024) as the retriever, the performance gain becomes 24.5% and 12.4%. Moreover, *VisRet* enhances downstream performance in retrieval-augmented generation (RAG) settings (§4.3). It improves T2I question answering accuracy on Visual-RAG and Visual-RAG-ME by 3.8% and 15.7% in top-1 retrieval setting and 3.9% and 11.1% in top-10 retrieval setting. Our code and benchmark will be shared publicly to facilitate future research.

2 Related Work

T2I Retrieval Benchmarks Early T2I retrieval benchmarks evaluate the ability to identify images based on their paired human-written captions. These datasets span multiple domains and include widely used benchmarks such as Flickr8K (Hodosh et al., 2013), Flickr30K (Young et al., 2014), and Fashion200K (Han et al., 2017). As multi-modal embedding models have matured, more challenging benchmarks have been introduced to assess retrieval in knowledge-intensive settings. These newer datasets—such as WebQA (Chang et al., 2022), INQUIRE (Vendrow et al., 2024), Visual-RAG (Wu et al., 2025), and MRAG-Bench (Hu et al., 2024)—shift the focus from caption matching to retrieving images that contain the knowledge necessary to answer complex natural language questions. These tasks challenge retrieval systems to support downstream reasoning in retrieval-augmented generation (RAG) pipelines.

T2I Retrieval Methods There have been extensive research on improving text-to-image retrieval from different perspectives. First, a number of works aim to train better multi-modal embeddings by designing better training objectives and data mixtures (Faghri et al., 2018; Radford et al., 2021; Yu et al., 2022; Li et al., 2022). Other studies improve various stages in the retrieval pipeline, such as textual query expansion (Levy et al., 2023; Lee et al., 2024) and reranking (Liu et al., 2024; Feng et al., 2025). Finally, a recent line of work introduces 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. Different from these existing approaches, *VisRet* expands the query semantics by directly visualizing it in the image space, thereby alleviating the workload of cross-modal retrieval. In addition, *VisRet* is a training-free plug-and-play framework that can accommodate any off-the-shelf retriever.

3 Approach

3.1 Problem Formulation

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

In this paper, we consider a basic retrieval-augmented generation (RAG) VQA pipeline: A multi-modal retriever $\mathcal{R}$ retrieves k images from $\mathcal{I}$, denoted as $\{r_1, ..., r_k\} \equiv \mathcal{R}(q, \mathcal{I}) \subseteq \mathcal{I}$. Then, a large vision-language model (LVLM) $\mathcal{M}$ directly generates the answer based on the question and the retrieval results $\mathcal{M}(q, \mathcal{R}(q, \mathcal{I}))$.

Retrieval Method	Visual-RAG						Visual-RAG-ME			INQUIRE-Rerank-Hard		
	R@1	R@10	R@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	0.583	0.737	0.210	0.355	0.385	0.220	0.423	0.435	0.000	0.355	0.412
Query Expansion	0.238	0.586	0.737	0.238	0.360	0.395	0.410	0.575	0.572	0.136	0.349	0.407
Visualize-then-Retrieve	**0.251**	**0.645**	**0.793**	**0.251**	**0.431**	**0.438**	0.460	**0.632**	**0.605**	0.170	**0.452**	0.455
- *multi-image*	0.246	0.637	0.772	0.246	0.414	0.421	**0.480**	0.629	**0.605**	**0.237**	0.428	**0.469**
Retriever = E5-V												
Original Query	0.240	0.568	0.706	0.240	0.386	0.407	0.340	0.465	0.486	0.000	0.319	0.407
Query Expansion	0.223	0.560	0.719	0.223	0.368	0.391	0.460	0.569	0.566	0.170	0.367	0.412
Visualize-then-Retrieve	0.299	**0.673**	**0.801**	0.299	**0.452**	**0.461**	**0.560**	**0.643**	**0.622**	**0.220**	0.377	0.425
- *multi-image*	**0.307**	0.645	**0.772**	**0.307**	0.442	0.446	0.520	0.640	0.617	0.203	**0.384**	**0.445**

Table 1: Evaluation results across three T2I retrieval benchmarks using different retrieval strategies and retrievers. The best results in each column within each retriever group are boldfaced. R = Recall. N = NDCG.

3.2 Visualize-then-Retrieve

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

Modality Projection The first stage of *VisRet* leverages a T2I generation system $\mathcal{T}$ to directly generate m visualizations $\{v_1, ..., v_m\} \equiv \mathcal{T}(q)$. Empirically, we find it helpful to use an LLM within $\mathcal{T}$ to first rephrase q into a T2I instruction q', before feeding into existing T2I generation models such as Stable Diffusion (Esser et al., 2024). To generate diverse $\{v_1, ..., v_m\}$, randomness can be injected either into q' or into the T2I generation.

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

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

To aggregate the m separate retrieval results, we apply Reciprocal Rank Fusion (RRF) (Cormack et al., 2009). RRF assigns a fusion score to each candidate image r based on its rank across 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 list $\mathcal{R}(v_i, \mathcal{I})$, and λ is a hyperparameter that controls the influence of lower-ranked items. The final top-k retrieval result is formed by selecting the highest-scoring images according to $\text{score}_{\text{RRF}}(r)$.

4 Results and Analyses

4.1 Experimental Setup

We evaluate on three challenging benchmarks: (1) INQUIRE-Rerank (Vendrow et al., 2024), a T2I retrieval benchmark requiring accurate knowledge of species appearance and behavior. We perform additional filtering to remove overly simple queries and call the resulting dataset INQUIRE-Rerank-Hard. (2) Visual-RAG (Wu et al., 2025), a T2I retrieval and VQA benchmark featuring visual knowledge intensive questions on features of natural species that are not commonly documented in text corpus. (3) Visual-RAG-ME, a new benchmark we introduce featuring queries that compare the same visual feature across *multiple entities*. We present the benchmark details in Appendix A.

For all the three benchmarks, we evaluate T2I retrieval with Recall@k and NDCG@k with k = 1, 10, 30. For Visual-RAG and Visual-RAG-ME, we additionally use an LLM judge to evaluate the end-to-end VQA accuracy, following Wu et al. (2025).

For the experiments presented in the main text, we use CLIP and E5-V as the retriever, `GPT-4o` (OpenAI, 2023b) as the downstream reader, and `gpt-image-1` (OpenAI, 2025) as the T2I Model to generate $m = 3$ images. We analyze more model choices in Appendix B and present all the prompts and other hyperparameters in Appendix C.

4.2 Retrieval Performance

Table 1 summarizes retrieval performance across all benchmarks and retrievers. We compare four strategies: using the original textual query, applying query expansion via an LLM, our proposed *VisRet*, and *VisRet* with only a single generated image ("*-multi-image*"). Across all datasets, *VisRet* consistently outperforms both the original query

Dataset	Question	Ground Truth	Baseline	Generated Image	VisRet
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
INQUIRE	A male and female cardinal sharing food		Rank:12, NDCG@10: 0.00		Rank:1, NDCG@10: 1.00

Table 2: Examples: *VisRet* improves retrieval by highlighting visual features implied by the textual query.

and query expansion baselines by a large margin. When CLIP is used as the retriever, *VisRet* outperforms the original query and LLM-based rephrase by 32.7% and 15.6% relatively higher NDCG@10 over three benchmarks. Similar trends hold when E5-V is used as the retriever, exhibiting 24.5% and 12.4% performance gain in NDCG@10. Further, using only one generated image as the query only slightly harms the performance, indicating the flexibility of *VisRet*. Table 2 presents several examples to demonstrate how the visualization step successfully captures subtle visual semantics implied by the original text query. We present more analyses on the T2I model and the rephrase LLM in Appendix B.1 and Appendix B.2.

4.3 Downstream QA Performance

To assess the utility of *VisRet* in real-world applications, we evaluate its downstream VQA accuracy in a RAG pipeline. We compare three settings: (1) using only the model's internal knowledge, (2) RAG with original text query-based retrieval and (3) RAG with *VisRet*. Figure 2 shows the QA accuracy on Visual-RAG and Visual-RAG-ME using GPT-4o as the LVLM reader and CLIP as the retriever. The original query results in low-quality retrieval augmentation, even slightly harming the performance on Visual-RAG in top-1 retrieval setting compared to no retrieval. By contrast, *VisRet* significantly improves QA accuracy in both top-1 and top-10 settings on both benchmarks, boosting accuracy to 0.538 on Visual-RAG and 0.700 on Visual-RAG-ME. Remarkably, on Visual-RAG-ME, *VisRet* outperforms top-10 retrieval in the original query setting with only top-1 retrieval, highlighting its high accuracy in retrieving the images containing the required features. Overall, the results confirm that *VisRet* not only improves

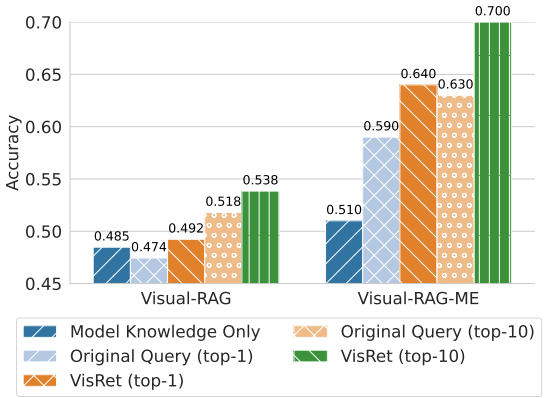


Figure 2: Downstream RAG-based VQA accuracy on Visual-RAG and Visual-RAG-ME with CLIP as the retriever and GPT-4o as the reader LVLM.

retrieval accuracy but also leads to tangible gains in downstream VQA performance. In Appendix B.3, we further demonstrate that *VisRet* can bring similar performance gains to other models as the VQA reader. In Appendix B.4, we analyze the performance of directly using the generated images as the context and find that while T2I generation improves retrieval, it still cannot replace the retrieved natural images in most cases.

5 Conclusion

This work introduces *VisRet*, a framework that visualizes text queries to enable more accurate T2I retrieval. By operating entirely in the visual domain during retrieval, *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 three benchmarking datasets and two retrievers. The simplicity and modularity of VisRet open up promising directions for future knowledge-intensive multi-modal systems.

Limitations

While this paper has proposed a novel framework and achieves strong empirical performance, our study has a few limitations as well. First, generating high quality images as queries can incur non-negligible latency costs. We would like to emphasize that retrieval is often a pipeline and the improved retrieval accuracy saves latency from a range of downstream operations such as reranking over a large number of candidates or iterative retrieval. Also, for applications that are accuracy-driven but not latency-sensitive, such as deep research, the latency of *VisRet* is often justifiable. Another limitation is that the paper only considers off-the-shelf T2I generation models and frozen embedding weights. Further work can consider using in-domain images to further fine-tune or condition the T2I generation model, producing visualizations that emphasize salient features while mitigating noise from domain shifts. It is also a promising direction to use T2I generation to synthesize more text-image alignment data to further improve the knowledge of the embedding model of fine-grained implied semantics.

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.

References

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.

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.

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.

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.

Yin Wu, Quanyu Long, Jing Li, Jianfei Yu, and Wenya Wang. 2025. 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 Visual-RAG data processing and Visual-RAG-ME annotation. Then, we provide the details of how we processed INQUIRE-Rerenk-Hard. Finally, we document the dataset statistics and the baseline T2I retrieval performance of CLIP (Radford et al., 2021) and E5-V (Jiang et al., 2024) in Table 3.

A.1 Visual-RAG and Visual-RAG-ME

Visual-RAG (Wu et al., 2025) releases 391 queries with associated image names from iNaturalist 2021 (Horn et al., 2021) and corresponding retrieval labels[1]. To prepare the data, we download the original iNaturalist 2021 dataset and re-collect the images from the train and test set. We were able to identify all annotated images in Visual-RAG except a single image due to a likely path error.

We annotate **Visual-RAG-ME** as an extension of Visual-RAG to visual question answering in the multi-entity setting. Concretely, Visual-RAG-ME reuses the visual features queried in Visual-RAG and constructs questions that compares those features between the organism covered in Visual-RAG with another similar entity, for which we manually annotate a new set of retrieval labels. The Visual-RAG-ME annotation pipeline consists of three steps: second entity identification; query composition and filtering; retrieval label annotation and balancing. We next describe the steps in detail.

Second Entity Identification The goal of this step is to identify entities that are biologically close to the original entities in Visual-RAG so that plausible and challenging questions could be constructed. For this purpose, for each entity covered in Visual-RAG, we use BM25 (Robertson and Walker, 1994) to retrieve ten entities that have the closest full scientific names.

Query Composition and Filtering In this step, we (the authors) manually traverse all the 391 questions in Visual-RAG and attempt to construct a corresponding multi-entity question. A question is constructed when we can identify images for the second entity that clearly depict the same feature as in the positive images for the original entity in Visual-RAG. The question we compose generally take a comparison style that asks whether the two organisms have the same feature or which organism feature a more extreme stylistic feature (e.g., lighter coloration, smoother surface etc.). We were able to construct 82 multi-entity questions after this step. Next, we perform a round of filtering to remove (1) the questions that both GPT-4o and GPT-4o-mini can answer correctly without image information and (2) questions can cover overly similar topics. After the filtering step, we ended up with 50 high quality multi-entity queries.

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

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

Table 3: Dataset statistics and baseline performance. VR denotes Visual-RAG and IR denotes INQUIRE-Rerank.

Retrieval Label Annotation and Balancing Finally, for each question, we collect images of the second entity from iNaturalist and annotate their retrieval label. A positive label is assigned only if the image clearly displays the feature required to answer the question. For some of the questions, we find that a large number of positive images exist in the iNaturalist database. We therefore implement a filtering step where at most 50 positive images are kept for each entity. Table 8 shows two examples with their questions and ground-truth images.

In Table 3, we present the basic statistics of Visual-RAG-ME. While it has slightly more positive per entity compared to Visual-RAG, our new benchmark is still challenging, with both CLIP and E5-V achieving a low Recall@1 due to its lengthy and knowledge-intensive queries.

A.2 INQUIRE-Rerank-Hard

To prepare INQUIRE-Rerank-Hard, we accessed the publicly released INQUIRE-Rerank (Vendrow et al., 2024) benchmark[2]. 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

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

Method	R@1	R@10	R@30	N@1	N@10	N@30
Baselines						
Original Query	0.210	0.583	0.737	0.210	0.355	0.385
LLM Rephrase	0.238	0.586	0.737	0.238	0.360	0.395
VisRet						
DALL-E 3	0.169	0.581	0.757	0.169	0.344	0.376
Stable Diffusion 3	0.166	0.517	0.691	0.166	0.319	0.349
Image-1 (low quality)	0.243	0.611	0.760	0.243	0.397	0.415
Image-1 (high quality)	**0.251**	**0.645**	**0.793**	**0.251**	**0.431**	**0.438**

Table 4: T2I Retrieval performance across different T2I generation models used for *VisRet*. We use GPT-4o to generate the T2I instruction and CLIP as the retriever. R = Recall. N = NDCG. The best results are boldfaced.

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 Rephrase	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 5: Retrieval performance on Visual-RAG of with CLIP retriever, using LLMs as T2I instruction generator for *VisRet*. R = Recall. N = NDCG. The best results are boldfaced.

can achieve 0.438 Recall@1 while E5-V achieved 0.506 Recall@1. After manually inspecting the data, we found that for a lot of 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 on which either CLIP and E5-V can achieve 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 organism themselves, with more challenging confounder negative images.

B *VisRet*: Further Analyses

In this section, we provide more comprehensive analyses to investigate the effectiveness of *VisRet* from more perspectives, including the choices of T2I generation Model, T2I Instruction LLM, the downstream VQA LVLM reader. 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.

B.1 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 three other models: DALL-E 3 (OpenAI, 2023a), Stable Diffusion 3 (Esser et al., 2024), and Image-1 with the low generation quality setting. Table 4 shows the results with GPT-4o as T2I instruction generation model and CLIP as the retriever model. Interestingly, compared to the cross-modality retrieval baselines, we find that DALL-E 3 and Stable Diffusion 3 do not provide significant performance improvements, while Image-1 low quality clearly and consistently improve the performance. The best performance is achieved by the newest and the most expensive Image-1 high quality setting. Together, these results suggest that a good T2I generation model with strong instruction following ability is necessary for *VisRet*. As further T2I generation methods improve, we anticipate that building more cost-efficient version of *VisRet* is a viable and promising further direction.

B.2 T2I Instruction LLM Choice

Does *VisRet* work well with other LLMs as the T2I instruction generator? In Table 5, 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

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	38.49	48.47	49.23	0	41.00	51.00	47.00
Direct T2I Retrieval	1	40.92	47.44	51.53	2	49.00	59.00	61.00
	10	46.04	51.79	57.06	10	48.00	63.00	65.00
Visualize-then-Retrieve	1	41.81	49.23	**57.16**	2	53.00	64.00	62.00
	10	**46.42**	**53.84**	56.65	10	**55.00**	**70.00**	**71.00**

Table 6: 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.

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	38.49	48.47	49.23	0	41.00	51.00	47.00
Generated Image (Image-1)	1	43.09	42.45	44.37	2	**59.00**	58.00	**80.00**
Visualize-then-Retrieve	1	41.81	49.23	**57.16**	2	53.00	64.00	62.00
	10	**46.42**	**53.84**	56.65	10	55.00	**70.00**	71.00

Table 7: 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.

promising results. For all the three LLMs, using them to generate T2I instructions for *VisRet* outperforms using the LLM themselves to rephrase the query. While more expensive LLMs achieve a high performance, the small 8B Llama model can already achieve decent performance at a similar level as GPT-4o.

B.3 Downstream VQA LVLM Choice

While we have shown the benefit of *VisRet* on VQA for GPT-4o, does the improvement hold across LVLMs with different capabilities? In Table 6, 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 only relying 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.

B.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*. It is a natural question then, that it is still necessary to perform retrieval instead of directly using the generated images as the knowledge? In Table 7, we compare the performance of using a single image as the context with *VisRet*. Overall, we observe a mixed result. 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 prefers the generated image over top-1 retrieval (and even top-10 retrieval). However, when provided with top-10 retrieved images, the models generally exhibit a higher VQA performance than using the generated image. Therefore, we conclude that retrieving natural images is still crucial for challenging VQA tasks like Visual-RAG and Visual-RAG-MR and cannot be fully replaced by pure T2I generation at this stage. It is an important future work to further combine image generation and image retrieval to improve the quality of the retrieved knowledge.

B.5 Further Qualitative Studies

In Table 8, we additionally provide some qualitative examples of how *VisRet* successfully improves T2I retrieval performance. As can be observed in top-ranked examples in text query-only retrieval, these retrieved images often fall short in correctly conveying visual semantics information such as angle, body part depicted, form of the subject depicted, visual distance of the subject, and so on. In contrast, *VisRet* is able to represent the inferred nuanced visual semantics in the image generation step first, then utilize this within-modality semantic-rich image query to obtain precise visual knowledge related to the VQA task.

C 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 it. The prompts for three benchmarking datasets are shown in Figure 3, Figure 4, and Figure 5, respectively. Then, we wrap the rephrased query with 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 signle instruction already results in images generated 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[3] or E5-V[4] encoder and perform a similarity search. Cosine similarity is used as the similarity metric. For RRF, we used $\lambda = 1$ 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, perform reasoning, and conclude its reasoning with self-verification. We show the detailed prompt in Figure 6 and Figure 7.

Baselines For the LLM rephrase baseline, we use the same prompt for *VisRet* T2I instruction to highlight the most important feature that the query is seeking. At the early stage of the prompt, 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 with the same prompt for the final version for the paper.

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., 2025), as shown in Figure 8. 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.

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

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

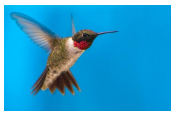
Dataset	Question	Ground Truth	Baseline	Generated Image	VisRet
Visual-RAG	What is the color of the head of larva Urania Swallowtail Moth (scientific name: Urania fulgens)?		Rank:96, NDCG@10: 0.00		Rank:1, NDCG@10: 0.66
	What color are the ventral abdomen of Golden Buprestid Beetle (scientific name: Buprestis 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	A male and female cardinal sharing food		Rank:12, NDCG@10: 0.00		Rank:1, NDCG@10: 1.00
	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
	male ruby-throated hummingbird in flight		Rank:60, NDCG@10: 0.00		Rank:7, NDCG@10: 0.39
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

Table 8: 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 3: 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 4: 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 5: Prompt for instructing an LLM to generate the T2I generation instruction for INQUIRE-Rerank 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 6: 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 7: 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 8: Prompt for the LLM VQA judge used for Visual-RAG and Visual-RAG-ME.