QueryFormer: Multi-modal Query Expansion on

Zijian Hu[1],**Xuanang Chen**[2],**Ben He**[1, 2*]

[1]School of Computer Science and Technology, University of Chinese Academy of Sciences
[2]Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences
huzijian2025@iscas.ac.cn, xuanang2020@iscas.ac.cn, benhe@ucas.ac.cn

Abstract

As technology advances quickly, the amount and variety of data people create and use every day have skyrocketed. This includes different types like images, text, sound, and video. This diversity poses a great challenge for efficiently extracting the information users need from massive, heterogeneous data. Multimodal retrieval, as an emerging technology, aims to realise cross-modal information querying and matching through semantic associations between different data modalities (e.g., image and text). Inspired by the limitations of existing methods in terms of insufficient retrieval performance when dealing with fuzzy or incomplete user queries, this study proposes a multimodal dynamic query expansion optimisation pipeline, QueryFormer. The pipeline utilises the generative power of multimodal large language models (MLLMS) combined with a retrieval result feedback mechanism to iteratively optimise the query representations, thereby enhancing cross-modal retrieval performance. The main contributions include designing this pipeline, proposing an adaptive query tuning strategy.

Introduction

Multimodal retrieval (Multimodal Retrieval) refers to the task of achieving cross-modal information query and matching based on the semantic associations between multiple heterogeneous data modalities, such as images, text, audio, and video. The core objective is to jointly model and align multimodal data to construct a unified or mappable semantic space. This allows users to input a query in one modality (such as text) and efficiently retrieve semantically related but differently formatted cross-modal results (such as images or videos), or vice versa. This task emphasizes deep semantic understanding and consistency measurement across modalities and is widely applied in areas such as open-domain visual question answering (VQA), cross-modal knowledge retrieval, and intelligent recommendation systems.

The Contrastive Language-Image Pre-training (CLIP) model (Radford et al. 2021) pioneered the unification of images and text into the same feature space. By performing contrastive learning on large-scale image-text pair datasets, it achieved semantic alignment across modalities. Simultaneously, the rise of large language models (LLMs) such as GPT-4 (OpenAI et al. 2024), Qwen (Qwen et al. 2025), and LLaMA (Grattafiori et al. 2024) has introduced new paradigms and technical momentum into the field of multimodal retrieval. These models, through their extensive training on massive text data, exhibit strong semantic understanding, reasoning, and generation capabilities, gradually becoming core components in the construction of cross-modal systems.

By integrating visual and speech modality encoders, multimodal large language models (MLLMs) have further pushed the boundaries of multimodal retrieval systems. MLLMs, such as GPT-4V (OpenAI et al. 2024), Qwen-2.5-VL (Qwen et al. 2025), and mLlama (Grattafiori et al. 2024), not only inherit the powerful text processing capabilities of traditional language models but also, by incorporating encoders for images, videos, and audio, construct a more comprehensive cross-modal understanding framework (Zhang et al. 2024). These models achieve a leap from "multimodal perception" to "multimodal reasoning" through end-to-end joint training or modular adaptation strategies, serving as the foundational models for multimodal systems empowered by large models (Chen et al. 2024).

Despite the significant improvements in cross-modal semantic alignment brought by CLIP and MLLMs, user input queries in practical retrieval scenarios often suffer from incomplete information or semantic ambiguity, such as blurry images or brief text descriptions, leading to retrieval results that deviate from expectations (Anand, Anand, and Setty 2023). To address this, this paper proposes a multimodal dynamic query expansion and optimisation pipeline. By combining the generative capabilities of MLLMs with feedback from retrieval results, this pipeline iteratively optimises query representations to enhance retrieval performance (Long et al. 2024; Zhang et al. 2025). Additionally, a multimodal RAG system based on the Gradio user interface has been implemented to facilitate user interaction and demonstrate the effectiveness of the proposed method.

Related Work

Large Language Models and Fine-tuning

Large language models (LLMs) have shown potential in retrieval tasks, such as understanding queries (Jagerman et al. 2023) and assisting in re-ranking (Ferrara 2023). However,

*Corresponding author

fine-tuning LLMs for specific tasks poses significant computational challenges due to their size. Efficient fine-tuning techniques, including prompt tuning, prefix tuning (Li and Liang 2021), adapters, and low-rank (LoRA) methods (Hu et al. 2021), address these challenges by updating only a subset of parameters, reducing resource costs while maintaining task-specific performance.

Multimodal Large Language Models

Multimodal large models are a combination of natural language processing and computer vision (Zhang et al. 2024). Usually, a multimodal large model consists of an LLM, a visual coder, and a linear projection layer (Li et al. 2024). The role of the projection layer is to align the information processed by the visual coder with the word embedding dimensions processed by the LLM, and to transform the visual features into "pseudo-texts" that can be understood by the LLM. For example, LLAVA uses Clip-ViT-L/14 as a visual coder to generate a 256x1024 feature matrix, while the word embedding dimension of the Vicuna language model used in the LLAVA architecture is 4096, so the linear projection layer converts each patch feature to the 4096 dimension to get the 256x4096 visual token sequence, and stitches together the visual token sequence to get a 256x4096 visual token sequence (Liu et al. 2023). Visual token sequences are spliced to get a complete input, and the answer is generated by auto-regression. With such processing, the large model extends the capability of analysis and reasoning in multimodal context and is widely used in multimodal retrieval (Long et al. 2024).

Multi-Modal Knowledge Retrieval

Multi-modal tasks often require additional knowledge support to address complex problems, as the information provided by the context is usually insufficient. Current solutions typically integrate various retrieval tools, such as BM25 for text retrieval (Robertson and Zaragoza 2009), CLIP for matching images with text (Radford et al. 2021), and GENER for entity linking and retrieval (De Cao et al. 2021). However, these methods fall short in two key areas: they fail to capture the interactions between visual and textual queries and rely on complex pipelines with external tools, complicating usage.

Recent studies have addressed these limitations by proposing new datasets and models. For instance, Luo et al. (2021) developed a dataset based on OKVQA, combining questions and images as queries to retrieve relevant evidence from small and large knowledge bases (112K to 21M records). Luo et al. (2023) introduced a more demanding multi-modal knowledge retrieval dataset requiring advanced cross-modal understanding. Single-stream vision-language models trained on contrastive frameworks have been explored to jointly encode visual and textual queries. While such methods have improved cross-modal representation, multi-modal knowledge retrieval remains underexplored in terms of effectiveness and training efficiency.

Generative Retrieval

Recent advancements in generative retrieval simplify pipelines by generating document identifiers, such as titles or n-grams, instead of searching through large-scale corpora. For instance, the DSI framework (Tay et al. 2022) uses numeric IDs stored in Transformer memories, but this approach struggles with scalability and efficiency. Alternatively, methods like FM-Index-based retrieval (Bevilacqua et al. 2022) generate semantic n-grams, though they face challenges such as ambiguity when n-grams correspond to multiple documents, requiring additional re-ranking steps.

Methodology

Problem Formulation. Multi-modal retrieval seeks to establish a family of reversible semantic mapping functions across heterogeneous data modalities - images ($I \in R^{H\times W\times 3}$), text ($T \in \Sigma$) and audio ($A \in R^{t\times f}$). The optimisation objective can be formally expressed as $\mathcal{F} = \{f_\theta : X_m \times Y_n \to R \mid \theta \in \Theta, m, n \in \{I, T, A, \ldots\}\}$. The optimisation objective can be formally expressed as $\min_\theta E_{(x,y)\sim D}\left[\mathcal{L}(f_\theta(x), g_\theta(y))\right] + \lambda\Omega(\theta)$. where $\mathcal{L}$ denotes the cross-modal similarity metric (e.g., cosine similarity) $\Omega(\theta)$ represents the model complexity regularisation term. x, y correspond to the query and retrieved result, respectively.

Query Preprocessing Query preprocessing transforms raw multimodal inputs into representations that enhance retrieval accuracy. We employ three key techniques: multimodal query rewriting, multimodal query expansion, and adaptive term importance weighting.

Multimodal Query Rewriting. Raw queries often contain noise, such as irrelevant image backgrounds or ambiguous text, which can degrade retrieval performance. Multimodal query rewriting refines queries by resolving ambiguities and enhancing specificity. For example, given a query like "How many days does the gestation of this animal take?" paired with an image of a deer, the system extracts implicit keywords (e.g., "deer") from the image and reformulates the query as "How many days does the gestation period of a deer take?" This process leverages a vision-language transformer (e.g., ViLT (Kim, Son, and Kim 2021)) to encode image and text inputs jointly, ensuring that the rewritten query captures user intent accurately.

Multimodal Query Expansion. To improve retrieval precision, we apply query expansion to enrich queries with semantically related terms. This is particularly effective in multimodal corpora, where queries may need to bridge visual and textual content. For instance, a query containing the term "Starry Night" is expanded to include synonyms like "night sky" or "Starry Night painting" using keyword embedding similarity metrics. These expanded terms are appended to the reformulated query, enhancing the discriminative power of sparse retrieval models (e.g., BM25) and dense encoders.

Adaptive Term Importance Weighting. To prioritize relevant information, we introduce adaptive term importance weighting, which dynamically assigns weights to keywords and features from different modalities. For example, given

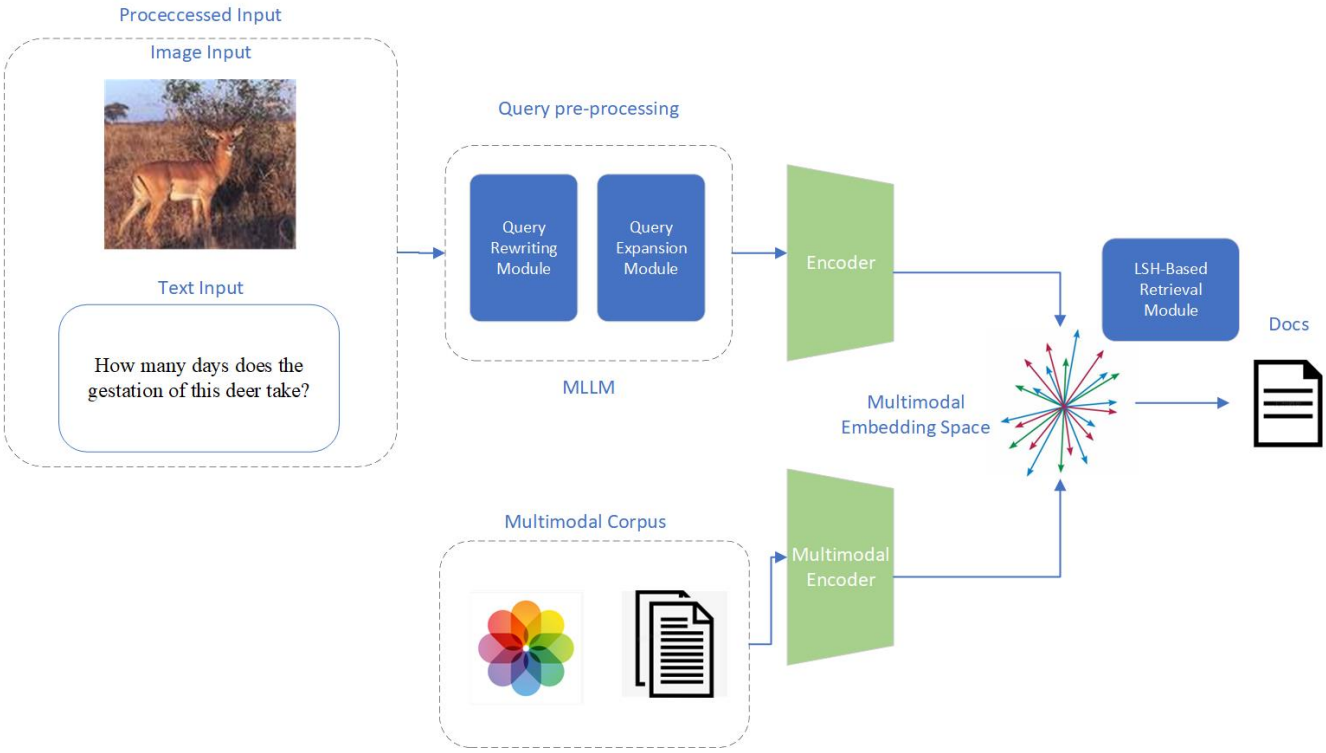

Figure 1: The framework of QueryFormer. The framework is structured to process raw multimodal queries (image and text) and retrieve relevant knowledge from a large corpus efficiently. It comprises two main phases: query preprocessing and retrieval. Query preprocessing refines raw inputs into structured representations, while the retrieval phase employs similarity search to identify relevant knowledge. The framework is designed to operate end-to-end, eliminating reliance on intermediate modules such as caption generators or object detectors.

a query "Where in the world could you ride on a boat such as this?" with an image of a canal boat, the system extracts explicit keywords (e.g., "boat," "world") from the text and implicit keywords (e.g., "waterway," "canal") from the image. A transformer-based model with an attention mechanism evaluates cross-modal interactions and assigns weights based on contextual relevance. This results in an augmented query, such as "Where in the world could you ride on a boat such as this in a waterway or canal?"

Retrieval Strategies The retrieval phase identifies the most relevant knowledge from a large corpus using similarity search in a multimodal embedding space. To address the computational challenges of searching large corpora, we employ Locality-Sensitive Hashing (LSH) (Datar et al. 2004) via the FAISS library for efficient approximate nearest neighbor search.

LSH-Based Retrieval. LSH maps similar data points to the same hash buckets, reducing the number of comparisons required during retrieval. The process consists of three steps:

- **Indexing Phase.** Each data point in the corpus is mapped to multiple hash tables using a family of hash functions (e.g., random hyperplane projections). These functions ensure that similar samples collide with high probability.
- **Query Phase.** For a given query embedding, the system computes its hash values and retrieves data points from the corresponding buckets across hash tables, forming a candidate set.
- **Refinement Phase.** The system computes precise similarity metrics (e.g., cosine similarity) between the query and each candidate, ranks them, and returns the top-K results.

This LSH-based approach achieves sublinear time complexity, enabling real-time retrieval from corpora with millions of entries while maintaining high recall rates.

Model Architecture Our model, inspired by ReViz (Luo et al. 2023), is an end-to-end vision-language retriever comprising a multimodal query encoder and a knowledge encoder. The architecture is designed to process raw image and text inputs directly, eliminating the need for intermediate cross-modal translators.

Multimodal Query Encoder. We use ViLT (Kim, Son, and Kim 2021), a vision-language transformer, to jointly encode the image I and text T. The image is partitioned into fixed-size patches, which are encoded as visual tokens via a linear projection layer . These visual tokens are concatenated with text tokens, summed with position embeddings, and processed through self-attention blocks. The final multimodal representation is obtained by applying a linear pro-

Algorithm 1: LSH Build Index

Require: Dataset D, hash tables L, hyperplanes k, dimension d
Ensure: Hyperplane collection $\mathcal{H}$, hash tables $\mathcal{T}$

```
function BUILDINDEX(D, L, k, d)
    H ← ∅
    T ← {InitializeHashTable()}_{i=1}^{L}
    for i = 1 to L do                                   ▷ Main loop
        H_i ← ∅
        for j = 1 to k do
            r_j ~ N(0, I_d)                             ▷ Generate Gaussian vector
            H_i ← H_i ∪ {r_j}
        end for
        H ← H ∪ {H_i}
    end for
    for all x ∈ D do
        x̂ ← x/‖x‖                                       ▷ Normalize vector
        for i = 1 to L do
            b ← COMPUTEHASH(x̂, H_i)
            T_i[b] ← T_i[b] ∪ {x̂}
        end for
    end for
    return H, T
end function
function COMPUTEHASH(v, H)
    s ← ""                                              ▷ Initialize empty string
    for all r ∈ H do
        if v^⊤ r ≥ 0 then
            s ← s ⊕ "1"                                 ▷ Bit concatenation
        else
            s ← s ⊕ "0"
        end if
    end for
    return s
end function
```

jection and hyperbolic tangent activation to the first token embedding:

$$\mathbf{Z}_q = \text{ViLT}(I, T)$$

Knowledge Encoder. The knowledge encoder employs BERT (Zhang et al. 2019) to encode textual knowledge K into a dense vector representation. The final representation is the vector corresponding to the [CLS] token:

$$\mathbf{Z}_k = \text{BERT}(K)$$

The relevance score between the query and knowledge is computed as the inner product of their embeddings:

$$\text{Score}(I, T, K) = \mathbf{Z}_k^\top \cdot \mathbf{Z}_q$$

Training

The training objective is to maximize the similarity between multimodal queries and relevant knowledge while minimizing similarity with irrelevant knowledge. We employ the InfoNCE(van den Oord, Li, and Vinyals 2019) loss, a contrastive learning objective that enhances cross-modal semantic alignment. The loss function is defined as:

$$\mathcal{L} = -\log \frac{\exp(\mathbf{z}_q \cdot \mathbf{z}_k)}{\exp(\mathbf{z}_q \cdot \mathbf{z}_k) + \sum_{\mathbf{k}' \in \mathbf{B_k}, \mathbf{k}' \neq \mathbf{k}} \exp(\mathbf{z}_q \cdot \mathbf{z}_{\mathbf{k}'})}$$

where $\mathbf{z}_q$ is the query embedding, $\mathbf{z}_k$ is the relevant knowledge embedding, and $\mathbf{z}_{\mathbf{k}'}$ represents negative knowledge embeddings from the batch $\mathbf{B_k}$.

Hard Negative Mining To further improve performance, we incorporate hard negative mining, which selects negative samples that are highly similar to the query but incorrect. The process involves:

1. **Preliminary Retrieval.** The model retrieves the top-K (e.g., $K = 100$) knowledge instances from the corpus for a given query.
2. **Hard Negative Selection.** Positive samples are excluded, and samples ranked between 50 and 70 (highly similar but incorrect) are selected as hard negatives.
3. **Hybrid Training.** Hard negatives are combined with in-batch negatives to compute the InfoNCE loss, enhancing the model's ability to handle challenging cases.

Pretraining Task We propose a pretraining task, VL-ICT (Vision-Language Inverse Cloze Task), adapted from ICT for multimodal scenarios. Using the dataset=, we construct triplets (I, T, K) where the image I and text T contain mutually exclusive information, and K is the relevant knowledge. For each WiT entry, the image is used as I, a sentence containing the title or caption is selected as T, and the remaining passage (after removing T) is used as K. Keywords in T overlapping with the caption are masked to enforce mutual exclusivity. This results in 3.2 million training triplets, enabling large-scale pretraining.

Experiments

Settings

Dataset. This study employs three parallel datasets to validate methodological efficacy. OKVQA serves as the benchmark dataset, an open-domain visual question answering task constructed upon COCO; OKVQA-GS112K offers 112,724 knowledge-enhanced corpora to verify medium-scale precision retrieval; OKVQA-WK21M incorporates 21 million Wikipedia knowledge units to assess ultra-large-scale retrieval efficiency. Concurrently, the Infoseek-VQA-WK6M multimodal dataset provides image-text alignment validation to refine cross-modal semantic mapping.The Table.1 below shows the corpus size of the datasets.

Evaluation Metrics. Consistent with established methodologies in prior research, our experiments employ Pseudo-Relevance Precision@K (P@K) and Recall@K (R@K) as core evaluation metrics. Recall performance is measured through R@5 and R@10 across all datasets, reflecting retrieval coverage at different granularities, while precision metrics are adapted to dataset characteristics: P@1 is prioritised for Infoseek-VQA to evaluate exact top-result accuracy

given its single-ground-truth design, whereas P@5 quantifies ranked-list precision for other benchmarks. Metric definitions rigorously follow the formal specifications outlined in Luo et al. (2021, 2023), ensuring methodological comparability.

Baselines. (1)Direct retrieval without QE: BM25 (Robertson and Zaragoza, 2009) is a standard term-matching sparse retriever, and DPR (Karpukhin et al., 2020) is a dense retriever utilizing dual BERT encoders. (2) Image-to-text retriever: CLIP (Radford et al., 2021) aligns visual and textual modalities through contrastive learning, projecting both images and text into a unified embedding space where cross-modal similarity is computed via cosine distance, serving as a foundational model for vision-language alignment tasks. (3) Task-specific multi-modal retrievers: ReViz (Luo et al., 2023) designs a dual-tower architecture with modality-specific encoders and a joint embedding space optimized via contrastive learning, while its enhanced variant ReViz+ICT incorporates interactive cross-modal transformers to dynamically fuse visual-textual features during retrieval. (4) GeMKR-based QE for Multimodal Retrieval: GeMKR (Zhang et al., 2024) leverages object-aware prefix tuning to align multi-grained vision-text features. The framework generates document-anchored knowledge clues (e.g., cross-modal implicit identifiers) through a generate-then-retrieve two-phase paradigm, efficiently decoupling neural reasoning from lightweight database operations. A knowledge-guided generation strategy is introduced during decoding steps to enforce semantic constraints, ensuring the uniqueness and discriminability of generated knowledge clues.

Implement details. This investigation employs the Llama-3.2-11B-Vision-Instruct architecture (Touvron et al., 2023) as the multimodal large language model core, which facilitates cross-modal reasoning through vision-language joint instruction fine-tuning (comparative experiments detailed in Appendix D.1). The retrieval module utilizes the classical BM25 sparse retrieval algorithm (Robertson & Zaragoza, 2009) with critical parameter configurations as follows: during generation phases, temperature parameters are set to 0.2 to maintain optimal diversity equilibrium, fixed generation of 12 query expansions, and construction of retrieval contexts based on Top-3 BM25-scoring documents; during optimization phases, dynamic temperature adjustment to 0.7 enhances semantic coverage with a re-ranking window of 8 expanded queries. Then we use the refined queries to retrieve the documents with BM25 to get the top-k documents. Experimental implementation was conducted on a 3 NVIDIA A800 80GB GPU infrastructure.

Main Results

The experimental outcomes presented in the table offer a comprehensive evaluation of various models across multiple metrics, specifically Precision@5 (P@5), Recall@5 (R@5), and Recall@10 (R@10), on the OKVQA-GS112K, OKVQA-WK21M, and Infoseek-vqa datasets. These results provide profound insights into the efficacy of different methodologies in the realm of multimodal retrieval-augmented generation (RAG) systems, highlighting the strengths and limitations of each approach while underscoring the superior performance of our proposed method.

We compare our model against several baseline models, including BM25 (Query+Caption), DPR (Query+Caption), CLIP, ReViz, ReViz+ICT, and GEMKR, across three datasets: OKVQA-GS112K, OKVQA-WK21M, and Infoseek-vqa. On OKVQA-GS112K, BM25, relying on keyword-based retrieval, struggles with semantic nuances, while DPR shows a slight improvement due to its use of dense embeddings. CLIP significantly advances performance, leveraging vision-language pre-training for better cross-modal alignment. ReViz+ICT outperforms ReViz by incorporating iterative context tuning, enhancing relevance in retrieval. GEMKR, with its robust multimodal integration, achieves strong results, closely competing with CLIP.

On OKVQA-WK21M, a similar trend persists. BM25 and DPR exhibit limited performance, while CLIP's results dip compared to OKVQA-GS112K, likely due to the dataset's increased complexity. ReViz+ICT again surpasses ReViz, showing the value of context tuning, and GEMKR maintains its competitive edge. On Infoseek-vqa, BM25 and DPR face challenges, whereas CLIP and ReViz+ICT demonstrate notable improvements. GEMKR stands out with superior performance, highlighting its effectiveness in handling diverse multimodal retrieval tasks.

Overall, models like CLIP, ReViz+ICT, and GEMKR consistently outperform traditional methods like BM25 and DPR, underscoring the importance of advanced cross-modal techniques in multimodal knowledge retrieval across varying dataset complexities.

Analysis

Ablation Study In this section, we perform hierarchical ablation studies from foundational to advanced components by sequentially removing key modules. Results are detailed in Table 3. First, we removed the images and text inputs from the dataset to test the accuracy in single-modality scenarios, excluding the possibility that the dataset could correctly retrieve the corresponding documents with only one modality input. After removing the image module (no images) and the text module, the R@5 metric decreased by 20.7% and 57.0%, respectively. This indicates that in this dataset, the model cannot accurately obtain answers using single-modal information. Generating image captions can increase query accuracy to some extent.Furthermore, we tested the impact of model fine-tuning on multimodal retrieval results. Converting the fine-tuned model Llama-3.2-11B-Vision-Instruct back to the original Llama-3.2-11B-Vision resulted in a significant 10% drop in P@5. This outcome demonstrates the essential role of the VLLM, which operates as a virtual knowledge base for generating precise knowledge clues to support query expansion. Fine-tuning can enhance the adaptability of multimodal large models for this task to some extent. Finally, we evaluated the effectiveness of the optimization module in our pipeline. After removing the Query Expansion (QE) module while keeping the model architecture unchanged, P@5, R@5, and R@10 all showed moderate decreases. This demonstrates the role of our QE module

Table 1: Cross-dataset performance comparison of retrieval methods

Method	OKVQA-GS112K			OKVQA-WK21M			Infoseek-vqa-WK6M		
	P@5	R@5	R@10	P@5	R@5	R@10	P@1	R@5	R@10
Baselines									
BM25 (Query+Caption)	32.0	54.9	64.3	32.6	57.6	68.4	9.8	16.0	18.5
DPR (Query+Caption)	28.2	57.2	68.5	27.8	59.0	70.4	8.7	14.2	16.2
CLIP	11.1	34.5	50.5	9.7	29.8	43.0	14.7	15.8	19.3
ReViz △	34.5	66.1	77.8	30.1	60.9	72.2	–	–	–
ReViz+ICT △	41.7	73.4	83.2	31.4	61.9	72.6	–	–	–
GEMKR △	49.1	78.6	86.2	46.0	70.8	79.1	–	–	–
Our method									
QueryFormer+BM25	55.3	81.7	87.6	56.8	80.3	86.7	7.8	14.2	16.5
QueryFormer+BGE-VL	**58.1**	**83.4**	**89.2**	**59.6**	**82.7**	**88.3**	9.2	16.8	18.9

Table 2: Ablation studies on the GS112K dataset

Delete Module	P@5	R@5	R@10
Full BM25 Pipeline	55.3	81.7	87.6
w/o picture	34.3	61.0	71.1
w/o query (with caption)	27.3	24.7	48.6
w/o caption	53.8	79.1	85.5
w/o Lora	44.5	75.3	84.2
w/o Query Expansion	52.5	78.5	85.1

in improving retrieval accuracy.

Effect of our strategies Our proposed method, QueryFormer, surpasses all baselines across the OKVQA-GS112K and OKVQA-WK21M datasets, achieving a P@5 of 55.3, R@5 of 81.7, and R@10 of 87.6 on OKVQA-GS112K, and a P@5 of 56.82, R@5 of 80.26, and R@10 of 86.72 on OKVQA-WK21M. The success on these two datasets demonstrates the commonality of our approach to both small and large query ranges. These results underscore the efficacy of our approach, which integrates query, image, and answer components with a refined BM25 framework, augmented by the LLAMA-3.2 architecture. The substantial improvement over GEMKR (the closest competitor) by approximately 6.2 in P@5, 3.1 in R@5, and 1.4 in R@10 on OKVQA-GS112K highlights the advantage of our multimodal fusion strategy. This superiority can be attributed to our meticulous negative sample construction, incorporating both In-batch Negatives and Hard Negative Mining, which enables the model to discern fine-grained semantic differences and enhances its robustness across varied datasets.

Reflection In evaluating QueryFormer's performance on the Infoseek-VQA dataset, it is evident that it underperforms compared to traditional methods, diverging significantly from its robust results on datasets like OKVQA-GS112K and OKVQA-WK21M. This performance gap is largely attributed to the limitations of the LLAMA-3.2-11B-Vision model, which forms the backbone of QueryFormer and exhibits shortcomings in fine-grained entity detection—a critical requirement for Infoseek-VQA. While the OKVQA datasets primarily demand general knowledge and language comprehension, Infoseek-VQA requires the precise identification of specific entities within images, such as distinguishing visually similar objects or detecting subtle attributes. Despite the LLAMA-3.2-11B-Vision model's proficiency in capturing broad semantic features, its inability to handle detailed visual tasks results in erroneous query modifications that fail to align with the image content, ultimately degrading retrieval performance. In contrast, traditional methods like CLIP excel on Infoseek-VQA due to their effective cross-modal alignment capabilities. CLIP, a standard vision-language pre-trained model, maps images and text into a shared embedding space, enabling direct and efficient matching between visual entities and textual descriptions. This design allows CLIP to accurately associate fine-grained visual details with corresponding text, making it particularly adept at tasks requiring precise entity recognition. Unlike QueryFormer, which struggles with the complexities of multimodal integration and query modification errors, CLIP bypasses these challenges by leveraging its pretrained alignment to directly address the entity-focused demands of Infoseek-VQA. To enhance QueryFormer's performance, future improvements could focus on incorporating a stronger visual encoder or optimizing query modification strategies to mitigate errors in visual recognition.

Conclusion

In conclusion, this work presents a novel end-to-end vision-language retriever for knowledge retrieval with multimodal queries, introducing the ReMuQ dataset and the ReViz model. By leveraging query preprocessing techniques, efficient LSH-based retrieval strategies, and a robust training paradigm with the VL-ICT pretraining task, our approach achieves superior performance on ReMuQ and OK-VQA datasets in both zero-shot and fine-tuned settings. The proposed framework eliminates reliance on intermediate modules, enhances retrieval accuracy, and demonstrates strong generalization across domains. These contributions lay a solid foundation for future advancements in multimodal information retrieval, with potential applications in question

answering and personal assistants.

Acknowledgments

We gratefully acknowledge the support of the CIPL Laboratory at the Chinese Academy of Sciences, where this work was completed. Their resources and collaborative environment were instrumental in the success of this project.

References

Anand, A.; Anand, A.; and Setty, V. 2023. Query Understanding in the Age of Large Language Models. *arXiv preprint arXiv:2306.16004*.

Bevilacqua, M.; Ottaviano, G.; Lewis, P. S. H.; Yih, S. W.-t.; Riedel, S.; and Petroni, F. 2022. Autoregressive Search Engines: Generating Substrings as Document Identifiers. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *arXiv preprint arXiv:2404.16821*.

Datar, M.; Immorlica, N.; Indyk, P.; and Mirrokni, V. S. 2004. Locality-Sensitive Hashing Scheme Based on p-Stable Distributions. In *Proceedings of the 20th Annual Symposium on Computational Geometry*, 253–262.

De Cao, N.; Izacard, G.; Riedel, S.; and Petroni, F. 2021. Autoregressive Entity Retrieval. In *9th International Conference on Learning Representations (ICLR 2021)*.

Ferrara, E. 2023. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, 6(1): 3.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Jagerman, R.; Zhuang, H.; Qin, Z.; Wang, X.; and Bendersky, M. 2023. Query Expansion by Prompting Large Language Models. *arXiv preprint arXiv:2305.03653*.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. arXiv:2102.03334.

Li, W.; Yuan, Y.; Liu, J.; Tang, D.; Wang, S.; Qin, J.; et al. 2024. TokenPacker: Efficient Visual Projector for Multimodal LLM. *arXiv preprint arXiv:2407.02392*.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL/IJCNLP 2021)*, 4582–4597.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*.

Long, X.; Zeng, J.; Meng, F.; Ma, Z.; Zhang, K.; Zhou, B.; and Zhou, J. 2024. Generative Multi-Modal Knowledge Retrieval with Large Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2024)*, 18733–18741.

Luo, M.; Fang, Z.; Gokhale, T.; Yang, Y.; and Baral, C. 2023. End-to-End Knowledge Retrieval with Multi-Modal Queries. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, 8573–8589.

Luo, M.; Zeng, Y.; Banerjee, P.; and Baral, C. 2021. Weakly-Supervised Visual-Retriever-Reader for Knowledge-Based Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, 6417–6431.

OpenAI; Achiam, J.; Adler, S. S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; et al. 2024. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Qwen; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; et al. 2025. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*.

Robertson, S. E.; and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4): 333–389.

Tay, Y.; Tran, V. A.; Dehghani, M.; Ni, J.; Bahri, D.; Mehta, H.; et al. 2022. Transformer Memory as a Differentiable Search Index. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*.

van den Oord, A.; Li, Y.; and Vinyals, O. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Zhang, D.; Yu, Y.; Dong, J.; Li, C.; Su, D.; Chu, C.; and Yu, D. 2024. MM-LLMs: Recent Advances in Multimodal Large Language Models. *arXiv preprint arXiv:2401.13601*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. BERTScore: Evaluating Text Generation with BERT. *arXiv preprint arXiv:1904.09675*.

Zhang, X.; Zhang, Y.; Xie, W.; Li, M.; Dai, Z.; Long, D.; et al. 2025. GME: Improving Universal Multimodal Retrieval by Multimodal LLMs. *arXiv preprint arXiv:2412.16855*.