MIRe: Enhancing Multimodal Queries Representation via Fusion-Free Modality Interaction for Multimodal Retrieval

Yeong-Joon Ju[1], **Ho-Joong Kim**[1], **Seong-Whan Lee**[1]
[1]Department of Artificial Intelligence, Korea University
{yj_ju, hojoong_kim, sw.lee}@korea.ac.kr

Abstract

Recent multimodal retrieval methods have endowed text-based retrievers with multimodal capabilities by utilizing pre-training strategies for visual-text alignment. They often directly fuse the two modalities for cross-reference during the alignment to understand multimodal queries. However, existing methods often overlook crucial visual information due to a text-dominant issue, which overly depends on text-driven signals. In this paper, we introduce MIRe, a retrieval framework that achieves modality interaction without fusing textual features during the alignment. Our method allows the textual query to attend to visual embeddings while not feeding text-driven signals back into the visual representations. Additionally, we construct a pre-training dataset for multimodal query retrieval by transforming concise question-answer pairs into extended passages. Our experiments demonstrate that our pre-training strategy significantly enhances the understanding of multimodal queries, resulting in strong performance across four multimodal retrieval benchmarks under zero-shot settings. Moreover, our ablation studies and analyses explicitly verify the effectiveness of our framework in mitigating the text-dominant issue. Our code is publicly available: https://github.com/yeongjoonJu/MIRe

1 Introduction

Information retrieval aims to fetch relevant information from a large collection given a user query, underpinning numerous NLP tasks such as search engines, open-domain question answering (Chen, 2017; Zhu et al., 2021), and fact-checking (Thorne et al., 2018). Beyond conventional methods based on lexical similarities (e.g., TF-IDF and BM25 (Robertson et al., 2009)), embedding-based retrieval methods (Lee et al., 2019; Karpukhin et al., 2020; Izacard et al., 2022; Chen et al., 2024) have achieved rich semantic matching by learning high-dimensional representations of queries and passages via large-scale pre-training. However, they focus on textual queries, struggling to address multimodal queries that encompass both textual and visual information.

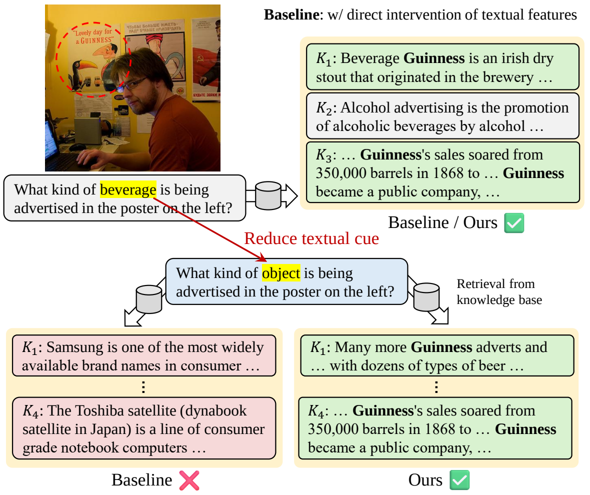

Baseline: w/ direct intervention of textual features
What kind of beverage is being advertised in the poster on the left?
K_1: Beverage Guinness is an irish dry stout that originated in the brewery …
K_2: Alcohol advertising is the promotion of alcoholic beverages by alcohol …
K_3: … Guinness's sales soared from 350,000 barrels in 1868 to … Guinness became a public company, …
Baseline / Ours ✓
Reduce textual cue
What kind of object is being advertised in the poster on the left?
Retrieval from knowledge base
K_1: Samsung is one of the most widely available brand names in consumer …
⋮
K_4: The Toshiba satellite (dynabook satellite in Japan) is a line of consumer grade notebook computers …
Baseline ✗
K_1: Many more Guinness adverts and … with dozens of types of beer …
⋮
K_4: … Guinness's sales soared from 350,000 barrels in 1868 to … Guinness became a public company, …
Ours ✓

Figure 1: Effect of the text-dominant issue in multimodal query retrieval.

In real-world scenarios, users often include visual references in their queries (e.g., complex objects or named entities depicted in an image), which are difficult to represent by text alone fully (Liu et al., 2023). Recent multimodal retrieval methods (Lin et al., 2023; Luo et al., 2023; Lin et al., 2024; Zhou et al., 2024a,b) have endowed text-based retrievers with multimodal capabilities by utilizing pre-training strategies for visual-text alignment. Most existing methods directly fuse the two modalities for cross-reference during visual-text alignment to enhance the understanding of multimodal queries. For instance, Luo et al. (2023) and Zhou et al. (2024a) facilitate modality interaction through early token fusion, where visual representations are prepended before passing through self-attention layers in the query encoder. Similarly, Lin et al. (2024) integrate modalities within the multimodal query using a cross-attention mechanism, where the textual query embeddings function as keys and values.

However, these methods often overlook crucial visual information due to a text-dominant issue induced by excessive reliance on text-driven signals

Findings of the Association for Computational Linguistics: ACL 2025, pages 5350–5363
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

during the alignment stage. In this stage, the retriever over-relies on textual similarities, thereby hindering proper visual alignment. Consequently, the retriever assigns high scores to irrelevant passages when textual cues are ambiguous. Fig. 1 shows the effect of the text-dominant issue. The baseline, which is trained with a direct fusion of textual features, fails to retrieve the desired passages due to its excessive reliance on text when the textual query becomes partially ambiguous (e.g., replacing a specific term like 'beverage' with a more generic word like 'object'). This text-dominant issue is further amplified through pre-training datasets constructed such that pseudo-queries are extracted from passages (Luo et al., 2023). Datasets obtained via this approach contain text-based queries that alone are sufficient to match relevant passages. This hinders visual-text alignment by relying on the high contextual similarity between textual queries and passages, even in the absence of visual information. This issue highlights the need for a retrieval framework that leverages multimodal queries by mapping both visual and textual cues into a linguistic space, capturing complementary interactions between these modalities without over-relying on textual features alone.

To address these issues, we introduce MIRe, a retrieval framework that achieves modality interaction without fusing textual features during the alignment stage. Instead of directly merging both modalities, MIRe allows the textual query to attend to patch-level visual embeddings without feeding text-driven signals back into the visual representations. We then fuse the two modalities during the relevance scoring stage based on a late-interaction mechanism (Khattab and Zaharia, 2020). This design alleviates the dependency on text-driven signals in the context of knowledge retrieval using a multimodal query. Furthermore, we construct a pre-training dataset by transforming multimodal query-response pairs into extensive passages via our response-to-passage conversion process that utilizes solely a text retrieval model. The constructed dataset requires the integration of both modalities to match a desired passage during training, enabling the model to link image understanding with complex textual queries. Our experiments demonstrate that our pre-training strategy significantly enhances multimodal query understanding for knowledge retrieval, resulting in strong performance across four multimodal retrieval benchmarks under zero-shot settings.

2 Related Work

Traditional methods such as TF-IDF and BM25 (Robertson et al., 2009) rely on keyword matching to retrieve relevant content but often fail to capture the deeper semantics underlying queries and documents. Beyond the surface-level lexical similarities, dense retrieval methods (Lee et al., 2019; Karpukhin et al., 2020; Izacard et al., 2022; Chen et al., 2024; Ni et al., 2022) leverage high-dimensional embedding models for richer semantic matching.

The transition from traditional text queries to multimodal queries has marked a significant evolution in information retrieval (Luo et al., 2021a). Early methods focused on converting images into textual representations, such as captions (Qu et al., 2021; Gao et al., 2022) and object tags (Gui et al., 2022; Yang et al., 2022). EnFoRe (Wu and Mooney, 2022) and DEDR (Salemi et al., 2023) improve image-query representations derived from a multimodal encoder with generated entities and captions, respectively. The OVEN dataset (Hu et al., 2023) has also provided insights into multimodal entity recognition. Whereas most of these approaches utilize DPR (Karpukhin et al., 2020) based on a single embedding for retrieval, FLMR (Lin et al., 2023) refines multimodal queries by incorporating RoIs and generated captions with the late-interaction mechanism. ReViz (Luo et al., 2023) represents an end-to-end multimodal retrieval system that removes the dependency on intermediate modules by pre-training on the VL-ICT, which automatically constructs a pre-training dataset by applying the Inverse Cloze Task (ICT) (Lee et al., 2019) to a multimodal knowledge base. UniIR (Wei et al., 2024) proposes an instruction-guided multimodal retriever along with its benchmark. They design two variants of the model architecture for modality interaction: score-level fusion and feature-level fusion based on CLIP and BLIP (Li et al., 2022). VISTA (Zhou et al., 2024a) introduces an in-depth fusion strategy by prepending visual tokens to the input of a text retrieval model to enhance multimodal understanding. PreFLMR (Lin et al., 2024) extends FLMR to investigate the scalability of multimodal retrievers under the late-interaction mechanism. In contrast to previous methods that rely heavily on text information within multimodal queries, we address the text-dominant issue in multimodal query representations caused by the direct intervention of textual features. We also adopt the late-interaction mechanism to fuse modalities during the scoring stage.

3 Method

In this section, we first define the problem of knowledge retrieval with multimodal queries. Next, we describe the architecture of our retrieval model and our data construction method.

3.1 Problem Definition

Given a multimodal query $Q = (I, T)$, the primary objective of our retriever $\mathcal{R}$ is to retrieve a set of relevant passages $K = \{D_1, D_2, \dots, D_n\}$ from a knowledge base U, where I and T denote an image and a textual query, respectively. Each D_i corresponds to a passage of text. To achieve this goal, $\mathcal{R}$ should encode the multi-modal query Q, integrating both the image and text modalities.

3.2 Background: Late Interaction in Retrieval

Late interaction (Khattab and Zaharia, 2020) is a retrieval strategy that preserves token-level embeddings for both queries and passages, enabling more fine-grained matching compared to single-vector retrieval. This mechanism defers the aggregation of embeddings to the scoring phase, retaining token-level signals. The retrieval model generates a set of low-dimensional embeddings $E = \{e_1, ..., e_l\}$ for tokens in both the query and the passage. Then, the final relevance score between query embeddings E_Q and document embeddings E_D is computed via the following MaxSim operation:

$$r_{Q,D} = \sum_{i=1}^{l_Q} \max_{j=1}^{l_D} \left(E_Q \cdot E_D^T \right), \tag{1}$$

where l_Q and l_D denote the number of tokens in the query and the document, respectively. Each query token is matched with its most relevant document token. In our MIRe framework, we extend this mechanism to handle retrieval with multimodal queries. Our rationale for this adoption is to mitigate the overemphasis on textual features during alignment by maintaining distinct representations for each modality.

3.3 Model Architecture

We detail our model architecture, focusing on how it integrates visual and textual features for multimodal query retrieval.

Textual Embeddings. We employ a pre-trained text retriever $\mathcal{R}_T$ to encode the input textual query T and passage D, utilizing multi-vector representations under the late-interaction mechanism. The text encoder generates token-level embeddings $E_t \in \mathbb{R}^{l_t \times d_t}$, where l_t denotes the number of tokens in T and d_t represents the embedding dimension.

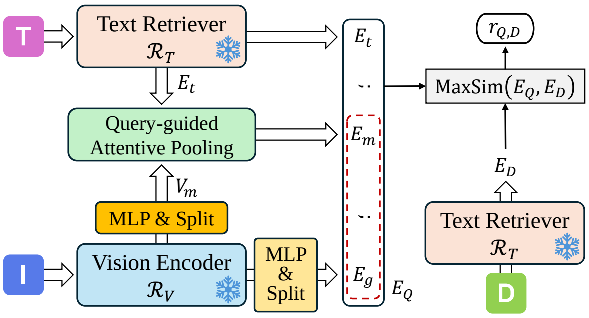


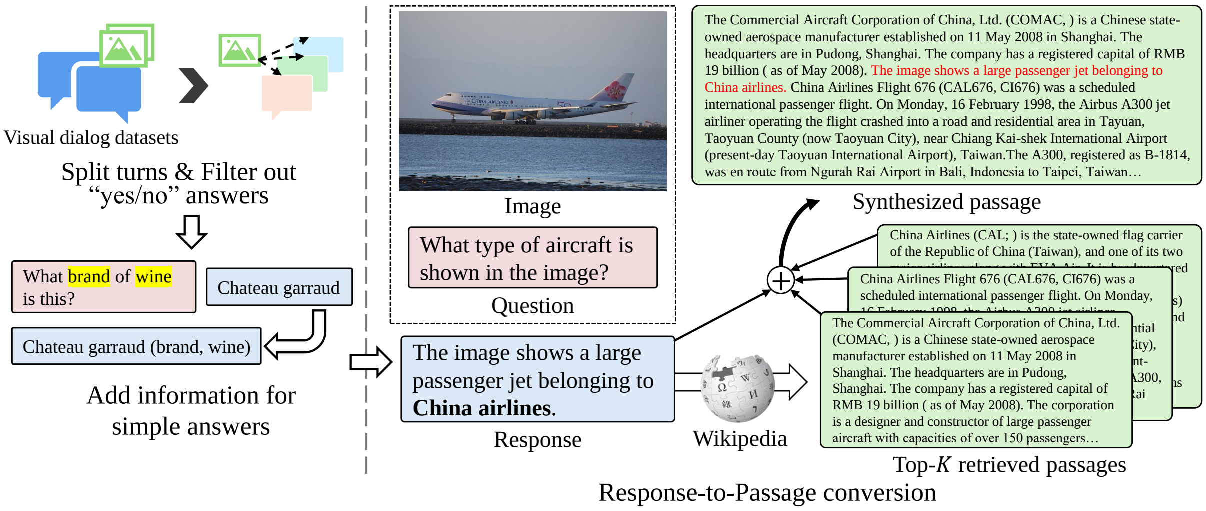
Figure 2: **Overview of the MIRe architecture.** This figure illustrates the interaction between the text encoder $\mathcal{R}_T$ and the vision encoder $\mathcal{R}_V$.

Visual Embeddings. We use ViT (Dosovitskiy et al., 2021) to encode image I. We adopt two kinds of visual embeddings: (1) global embeddings V_g derived from the CLS token, representing the overall content of the image, and (2) token-level embeddings V_m extracted from the penultimate layer of ViT, representing individual patches of the image. The global embedding $V_g \in \mathbb{R}^{d_v}$ is directly projected into the latent space of the text retriever $\mathcal{R}_T$ via a two-layer perception, producing embedding with dimension of $\mathbb{R}^{l_g \cdot d_t}$, where l_g is the pre-defined number of tokens. Subsequently, the projected V_g is reshaped into token-level embeddings $E_g \in \mathbb{R}^{l_g \times d_t}$.

Query-guided Attentive Pooling. Our architecture aims to achieve modality interaction without directly incorporating textual features in the pretraining stage for multimodal alignment, thereby mitigating the text dominance issue. To this end, we introduce a query-guided attentive pooling module employing an attention layer. This module retrieves visual information required by the textual query T from $V_m \in \mathbb{R}^{l_v \times d_v}$ and then aggregates the visual information based on its relevance to tokens within the textual query, where l_v denotes the number of image patches. Employing E_t as query vectors, attention scores $\mathcal{A} \in \mathbb{R}^{h \times l_t \times l_v}$ are calculated as follows:

$$\mathcal{A} = \text{Softmax} \left(\frac{E_t \cdot \mathcal{K}_m^\top}{\sqrt{d_t}} \right), \tag{2}$$

where $\mathcal{K}_m \in \mathbb{R}^{h \times l_v \times d_t}$ denotes key vectors of V_m projected by a linear layer and split into h tokens


Visual dialog datasets
Split turns & Filter out "yes/no" answers
What brand of wine is this?
Chateau garraud
Chateau garraud (brand, wine)
Add information for simple answers
Image
What type of aircraft is shown in the image?
Question
The image shows a large passenger jet belonging to **China airlines**.
Response
Wikipedia
Synthesized passage
Top-K retrieved passages
Response-to-Passage conversion

The Commercial Aircraft Corporation of China, Ltd. (COMAC,) is a Chinese state-owned aerospace manufacturer established on 11 May 2008 in Shanghai. The headquarters are in Pudong, Shanghai. The company has a registered capital of RMB 19 billion (as of May 2008). The image shows a large passenger jet belonging to China airlines. China Airlines Flight 676 (CAL676, CI676) was a scheduled international passenger flight. On Monday, 16 February 1998, the Airbus A300 jet airliner operating the flight crashed into a road and residential area in Tayuan, Taoyuan County (now Taoyuan City), near Chiang Kai-shek International Airport (present-day Taoyuan International Airport), Taiwan.The A300, registered as B-1814, was en route from Ngurah Rai Airport in Bali, Indonesia to Taipei, Taiwan…

China Airlines (CAL;) is the state-owned flag carrier of the Republic of China (Taiwan), and one of its two

China Airlines Flight 676 (CAL676, CI676) was a scheduled international passenger flight. On Monday,

The Commercial Aircraft Corporation of China, Ltd. (COMAC,) is a Chinese state-owned aerospace manufacturer established on 11 May 2008 in Shanghai. The headquarters are in Pudong, Shanghai. The company has a registered capital of RMB 19 billion (as of May 2008). The corporation is a designer and constructor of large passenger aircraft with capacities of over 150 passengers…

Figure 3: **Our data construction process.** Starting with visual dialogue datasets, our process involves two steps to convert the dialogue tasks to knowledge retrieval tasks. After preprocessing, we transform responses into a passage format by unifying the response and relevant passages retrieved from Wikipedia.

for each embedding within V_m. Then, the attended visual output $E_m \in \mathbb{R}^{h \times d_t}$ is calculated with value vectors $\mathcal{V}_m \in \mathbb{R}^{h \times l_v \times d_t}$ of V_m as follows:

$$E_m = \text{Linear}\left(\frac{1}{l_t}\sum_{i}^{l_t}(\mathcal{A} \cdot \mathcal{V}_m)\right), \tag{3}$$

where $\mathcal{V}_m$ is produced via operations identical with $\mathcal{K}_m$. Unlike the standard cross-attention mechanism, we apply mean-pooling along the sequence dimension without a residual connection, yielding h visual embeddings. In this modality interaction process, we only leverage textual embeddings E_t to calculate $\mathcal{A}$ as relevance scores for T without direct fusion. Our module alleviates the text dominance issue by explicitly limiting textual information during visual representation alignment, as demonstrated in our empirical analyses (see Sec. 5).

3.4 Dataset Construction

We aim to train our model to comprehend images based on textual queries, thereby enabling effective multimodal query understanding. To achieve this goal, we leverage existing multimodal question-response datasets, such as visual instruction-following data and VQA data. These datasets consist of query-response pairs where each pair is associated with a single image. In each pair, the response provides a concise and image-specific answer that directly addresses the textual query. Thus, the response serves as a clear bridge between the visual content and the query, explicitly linking image understanding to the language of the query. However, despite the explicit information provided by these responses, the datasets are not directly suitable for training the retriever $\mathcal{R}$ because of the inherent difference between concise responses and more expansive passages. In practice, responses can be matched with queries without ambiguity, whereas real-world retrieval tasks demand the identification of relevant information embedded within broader documents that often contain noisy content. To bridge this gap, we transform query-response pairs into a format suitable for multimodal retrieval tasks via response-to-passage conversion, as illustrated in Fig. 3.

Response-to-Passage Conversion. Let a multimodal query-response pair S as follows:

$$S = \{(I, T), R\}, \tag{4}$$

where R represents the response. We first attain multiple QA pairs for a single image from samples with several turns in source datasets. We divide the response R into two types: (1) detailed responses and (2) simple responses with a single word or a phrase. The simple responses often lack sufficient context to facilitate effective knowledge retrieval. Thus, we compensate simple responses with nouns extracted from the textual query T. Note that we filter out pairs of responses that do not contribute to knowledge-based retrieval, such as simple affirmations and negations (e.g., *"yes"* or *"no"*).

The nature of the data S guarantees a high correlation between (I, T) and R since the responses contain information conditioned on the given multimodal query while the textual queries have restrictive information. Thus, we utilize the response R to transform the response into an informative passage. From an arbitrary knowledge base U, we retrieve relevant passages using the response R as the query. Specifically, we obtain the top-k passages:

$$\{D_1, D_2, \ldots, D_k\} = \text{Retrieve}_{\mathcal{R}_T}(R, U, k), \quad (5)$$

where $\text{Retrieve}_{\mathcal{R}_T}$ denotes the retrieval function that returns the top k relevant passages from the knowledge base U based on the query R using the text retriever $\mathcal{R}_T$. To maintain contextual relevance with the multimodal query (I, T), we then augment the response R by combining it with the retrieved passages:

$$R' = [D_1; R; D_2; \ldots; D_k]. \quad (6)$$

This conversion strategy yields training data that more closely mimic the complexity and noise of real-world documents. Consequently, the retriever is exposed to more challenging and realistic scenarios during training, enabling it to effectively integrate visual cues and ultimately achieve more robust retrieval performance.

3.5 Training and Inference

We deal with passages including the golden answers to a given question Q as relevant passages K. To train our model, we employ in-batch negative sampling, which treats all passages in a training batch except for a passage D belonging to K as negative passages $\bar{K}$ for Q. We optimize our model by minimizing the following contrastive loss $\mathcal{L}_{CL}$ over the dataset $\mathcal{D}$:

$$\mathcal{L}_{CL} = -\sum_{\mathcal{D}} \log \frac{\exp(r_{Q,D}/\tau)}{exp(r_{Q,D}/\tau) + \sum_{\bar{D} \in \bar{K}} exp(r_{Q,\bar{D}}/\tau)}, \quad (7)$$

where τ is the temperature parameter that regulates the influence of penalties on negative samples. During the alignment stage, all parameters of $\mathcal{R}_T$ and $\mathcal{R}_V$ are frozen, preserving the established text retrieval performance. To focus on visual alignment, we exclude the textual embeddings E_t from the final query embedding E_Q, using only the visual features $[E_g; E_m]$ as E_Q. We also integrate a subset of a multimodal knowledge base, WiT (Srinivasan et al., 2021), into our dataset to enrich the world knowledge learned during alignment. Note that this addition does not affect multimodal query understanding because the dataset consists solely of pairs of an image and a passage (i.e., it does not include a textual query). For such data, we simply assign dummy prompts for multimodal queries (e.g., What is the core object or subject shown here?). We discuss this integration in Sec. 5 in detail.

After the alignment stage, we add textual embeddings E_t to E_Q when training on downstream tasks and the inference stage. For efficient retrieval, all passages within knowledge base U are pre-indexed using PLAID (Santhanam et al., 2022a), identical to ColBERTv2 (Santhanam et al., 2022b).

4 Experiments

4.1 Setup

Benchmarks. We employ four benchmarks for knowledge retrieval with multimodal queries: two variants of OK-VQA (Marino et al., 2019), ReMuQ (Luo et al., 2023), and E-VQA (Mensink et al., 2023). For OK-VQA, we use two versions based on different knowledge bases: OKVQA-GS, a corpus collected using Google search API as introduced in Luo et al. (2021a), and OKVQA-WK11M, a corpus containing 11 million Wikipedia passages compiled by Qu et al. (2020).

Metrics. We evaluate retrieval performance using Mean Reciprocal Rank at 5 (MRR@5), Recall@k (R@k), and Pseudo Recall@k (PR@k) across four benchmarks. MRR@5 measures the ranking quality of the first relevant passage. For OKVQA-GS and E-VQA, which do not provide explicit ground-truth passages, we compute PR@5 by checking whether retrieved documents contain the correct answer. For OKVQA-WK11M and ReMuQ, we evaluate R@k by verifying whether the target passages appear in the top-k results.

Implementation Details. Our pre-training dataset is synthesized from three visual instruction datasets (Zhang et al., 2023; Wang et al., 2023; Liu et al., 2024) and two VQA datasets (Singh et al., 2019; Biten et al., 2019), resulting in 1.35 million QA pairs, each paired with an image after preprocessing. We sampled to have no more than 12 question-response pairs per image. For the response-to-passage conversion, we utilize 6 million Wikipedia articles released by Chen et al. (2023) as our data pool. We retrieve three candidate passages for each response using ColBERTv2, trained with the MS MARCO Passage Ranking task (Nguyen et al., 2016). Each passage is truncated to three sentences, and the response is inserted between the first and second passages to ensure contextual consistency. We also added 0.5 million pairs randomly sampled from WiT.

For our base model, we adopt CLIP ViT-

Model	OKVQA-GS			OKVQA-WK11M			ReMuQ			E-VQA		
	MRR@5	PR@5	PR@10	MRR@5	R@5	R@10	MRR@5	R@5	R@10	MRR@5	PR@5	PR@10
CLIP (Radford et al., 2021)	19.08	34.54	50.48	16.45	29.81	43.0	0.34	0.78	1.36	-	-	-
FLMR (Lin et al., 2023)	38.15	57.25	69.42	32.56	50.61	62.58	66.67	72.10	74.95	29.97	42.0	50.75
ReViz (Luo et al., 2023)	45.77	64.05	75.39	44.03	62.43	73.44	23.61	39.43	46.77	-	-	-
UniIR (Wei et al., 2024)	53.27	73.94	84.19	-	-	-	79.15	84.34	86.62	31.59	44.21	55.31
VISTA (Zhou et al., 2024a)	55.33	72.83	81.61	-	-	-	78.32	84.21	87.03	33.9	47.73	56.72
PreFLMR† (Lin et al., 2024)	59.38	76.83	84.34	45.68	63.85	73.64	52.27	54.31	55.06	30.92	41.71	49.44
MIRe	**63.03**	**80.48**	**88.15**	**51.15**	**70.71**	**81.25**	**83.06**	**86.84**	**88.56**	**41.88**	**54.24**	**61.01**
w/ ViT-large	63.17	81.13	88.72	50.64	69.92	80.18	82.56	86.48	88.17	44.92	57.65	64.40

Table 1: **Zero-shot performance of MIRe and comparison methods.** Note that FLMR was only pre-trained on the WiT dataset. PreFLMR† were trained using our dataset and experimental settings. Bold indicates the highest performance.

base (Radford et al., 2021) as a vision encoder and ColBERTv2 as a text encoder based on BERT-base (Devlin et al., 2019). The number of tokens for visual embeddings E_g and E_m are set to 16 and 12, respectively. The value for E_m is determined by the number of heads h in the interaction module. The dimension of the final embeddings d_t is set to 128, consistent with the text encoder. Our base model has 211M parameters. Further implementation details are provided in Appendix A.

Comparison Methods. We benchmark our MIRe model against a diverse set of baseline models that employ pre-training stages for visual-text alignment: CLIP (Radford et al., 2021), FLMR (Lin et al., 2023), ReViz (Luo et al., 2023), PreFLMR (Lin et al., 2024), and VISTA (Zhou et al., 2024a). **UniIR** (Wei et al., 2024), which requires an explicit instruction input, is also included; we follow their protocol and use the instruction "Retrieve a passage that answers the given query about the image" during evaluation. Both FLMR and PreFLMR utilize the same vision and text encoders as our model, where FLMR was pre-trained with a subset of the WiT dataset. For direct comparison, PreFLMR was trained using the same pre-training procedure as our model, thereby highlighting the distinct advantages of our model architecture. For zero-shot (ZS) evaluation, we exclude baselines requiring external supervision for fairness, while for few-shot (FS) evaluation, models utilizing supervision datasets are included.

4.2 Main Results

Zero-shot Retrieval Performance. Tab. 1 shows that our method achieves superior zero-shot retrieval performance across all four benchmarks, significantly outperforming the comparison models. Despite employing a two-stage training strategy and directly optimizing the vision encoder for retrieval, VISTA still underperforms relative to our approach. Even though PreFLMR was trained under the same settings as our model, it exhibits a significant performance gap compared to our model. These results validate the effectiveness of our modality interaction approach. Our method also benefits from increased model capacity. The variant employing a larger vision encoder (ViT-large) shows similar performance to the standard model, but it further outperforms the standard model in E-VQA.

Fine-tuning on Downstream Tasks We further demonstrate the adaptability of our model and the effectiveness of our pre-training task by fine-tuning models on downstream tasks. Tab. 2 demonstrates remarkable adaptability when fine-tuned on downstream tasks. On the OKVQA-GS dataset, our model substantially outperforms all state-of-the-art models. On the ReMuQ dataset, our model still delivers strong performance, showing its competitive results. It is important to note that our method achieved higher performance on ReMuQ than VISTA in the zero-shot setting, which suggests that our pre-training and modality interaction approach endow our model with strong generalization capabilities. Notably, the variant without pre-training clearly lags behind the pre-trained model, highlighting the crucial role of our pre-training task. Furthermore, employing a larger vision encoder (ViT-large) yields additional improvements on OKVQA-GS, demonstrating the scalability of our approach. Overall, these results confirm that our model not only excels in zero-shot settings but also adapts effectively to fine-tuning on downstream tasks.

4.3 Ablation Studies

Our ablation studies, summarized in Tab. 3, reveal that each component in our framework plays a sig-

Model	OKVQA-GS		ReMuQ	
	PR@5	PR@10	R@5	R@10
FLMR (Lin et al., 2023)	70.63	81.23	62.76	74.67
VRR (Luo et al., 2021b)	71.5	81.5	-	-
ReViz (Luo et al., 2023)	73.35	83.17	23.61	39.43
GeMKR (Long et al., 2024)	78.6	86.2	90.3	92.7
VISTA (Zhou et al., 2024a)	82.06	90.11	**96.3**	**97.3**
Ours w/o Pre-training	74.26	84.07	92.44	94.38
Ours	**83.59**	**90.59**	94.40	96.20
w/ ViT-large	84.66	91.30	94.38	96.18

Table 2: **Fine-tuning performance on two tasks.**

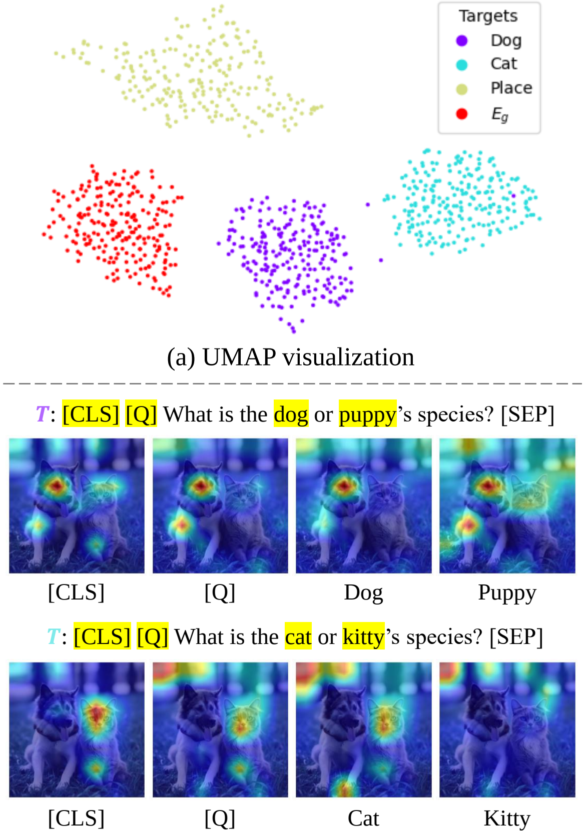
Method		OK-GS	OK-WK	ReMuQ	E-VQA	Avg.
Base		**63.03**	**51.15**	83.06	41.88	**59.78**
PT	*w/o WiT*	62.54	50.53	82.63	40.88	59.15
	w/o R2P	60.43	42.93	81.87	38.13	55.84
	w/ Single T	59.72	49.09	79.27	29.29	54.34
	w/ Residual	61.65	47.95	80.47	**43.06**	58.28
	w/o E_m	60.19	47.23	81.70	39.01	57.03
	w/ E_t	51.38	42.13	71.69	32.80	49.50
IF	*w/o* E_m	60.43	44.13	85.10	42.4	58.02
	w/o E_g	58.4	44.61	**85.91**	40.24	57.29
	w/o E_g&E_m	52.46	36.0	71.69	42.48	50.66
	w/o E_t	36.99	36.68	2.73	11.39	21.95

Table 3: **Ablation Studies.** Retrieval performance (MRR@5) in zero-shot settings across four datasets. "PT" and "IF" indicate ablations performed at the pre-training and inference stages, respectively.

nificant role in achieving robust zero-shot retrieval performance. We examine the contributions of our design from three perspectives: the dataset, model architecture during alignment, and the embeddings used at inference.

Dataset. In the pre-training stage (PT), omitting external knowledge from the WiT dataset causes only a slight performance drop, underscoring its supportive role (see Sec. 5). In contrast, training the model on original responses without applying the response-to-passage conversion (R2P) results in a substantially larger decline. These observations indicate that the R2P mechanism is essential for enhancing visual-text alignment and overall knowledge retrieval. We also investigate the effect of multiple QA pairs per image. As shown in Tab. 3, although sampling a single QA pair per image keeps the total number of images, this variant (w/ Single T) significantly degrades retrieval performance, suggesting the presence of hard-negative effects beyond simple visual-image alignment.

Model. We further examine how directly fusing text features during the alignment process affects performance. When we add a residual connection to our model architecture before sequential-wise pooling (*w/ Residual*), we observe a performance drop, indicating a slight exacerbation of the text-dominant issue. Moreover, when text features are allowed an even more direct influence, by setting $E_Q = [E_g, E_m, E_t]$ during alignment, the performance degrades considerably.



(a) UMAP visualization

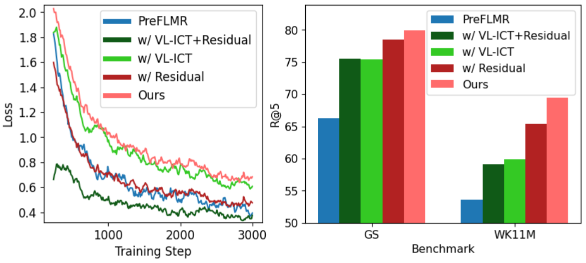
(b) Attention visualization for each token (averaged across heads)

Figure 4: **Visualization of multimodal query processing**, illustrating the alignment between textual and visual modalities.

Embeddings E_Q. At the inference stage (IF), our analysis shows that each embedding type plays a unique and complementary role. Removing either the modality-specific embedding (*w/o* E_m) or the general embedding (*w/o* E_g) leads to a moderate decline in performance, suggesting that both capture distinct yet essential aspects of the data. However, removing these components simultaneously causes a sharper performance drop. Notably, omitting the text embedding (*w/o* E_t) results in severe degradation of retrieval accuracy, indicating that E_t is indispensable for maintaining semantic coherence. This clear hierarchy in the impact of each embedding underscores their distinct functions and the need for their balanced integration.

Dataset		OKVQA-GS	ReMuQ	E-VQA	Infoseek
FLMR (Lin et al., 2023)		57.25	72.10	42.0	42.93
	w/o WiT	**81.11**	**87.45**	51.95	37.15
Ours	*w/ WiT (0.5M)*	80.48	86.84	**54.24**	42.61
	w/ WiT (1.0M)	79.63	86.59	54.05	**44.01**

Table 4: Zero-shot retrieval performance (R@5) under knowledge integration settings using WiT data.



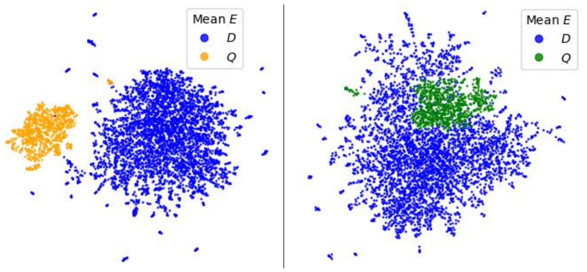
(a) Training loss curves (b) R@5 on OK-VQA

Figure 5: **Training convergence and retrieval performance.** All models were trained for only one epoch under the same settings.

5 Discussion

Effect of Query-guided Attentive Pooling To demonstrate the effectiveness of MIRe in capturing modality interactions, we visualize the embeddings and attention maps of multimodal queries on a controlled dataset. We synthesized 224 images with the prompt '*A dog and a cat in an image*' using Diffusion-XL (Podell et al., 2024), and conditioned the embeddings E_m on three distinct textual prompts: (1) *What is the dog or puppy's species?*, (2) *What is the cat or kitty's species?*, and (3) *Where is the place in the image?* In Fig. 4(a), the UMAP clustering (McInnes et al., 2018) of E_g and E_m illustrates MIRe effectively separates visual embeddings based on the query's intent. Additionally, Fig. 4(b) visualizes attention patterns of our pooling module, revealing how the model attends to specific visual patches relevant to each query. These results demonstrate that MIRe enhances interactions between textual and visual modalities.

Effect of Knowledge Integration. We further assess MIRe's capacity for external knowledge integration by incorporating the WiT dataset and analyzing its effect on retrieval performance, particularly on the Infoseek dataset (Chen et al., 2023). As shown in Tab. 4, *Ours w/o WiT* falls short on Infoseek relative to FLMR while competitively performing on other benchmarks. Notably, FLMR was learned with a subset of WiT without modality interaction. When we integrate external knowledge using 0.5 million WiT data, our model's performance on Infoseek is substantially improved to 42.61, bringing it on par with FLMR. Moreover, further increasing the WiT data to 1.0 million boosts the R@5 on Infoseek to 44.01. These findings, however, reveal that the performance gains observed on Infoseek are largely driven by its heavy reliance on external knowledge, which raises concerns about the generality of evaluation protocols that depend on such background information.



(a) *w/* E_t during alignment (b) Ours

Figure 6: **Comparison of Embedding Distribution.** (a) with E_t during alignment, where query embeddings (Q, orange) remain distinct from passage embeddings (D, blue); (b) our method, where Q (green) is better integrated into the textual space.

Text-dominant Issue. We analyze how certain pretraining strategies and model architectures exacerbate reliance on textual features during multimodal alignment. In Fig. 5(a), both PreFLMR and *w/ Residual* exhibit faster loss convergence compared to our model, suggesting that directly leveraging text features accelerates optimization. However, as shown in Fig. 5(b), the accelerated convergence does not translate to improved performance, with PreFLMR and *w/ Residual* underperforming relative to our model. The text-dominant issue is further exacerbated when using VL-ICT, introduced in ReViz, which constructs pseudo-queries from passages. Such behavior reveals the text-dominant issue, where excessive dependence on text features during alignment hinders the model's ability to fully leverage multimodal information. Fig. 6 illustrates this effect by visualizing the alignment of multimodal query embeddings Q with passage embeddings D. In (a), when E_t is explicitly used during alignment Q embeddings (orange) remain largely separated from the passage space, indicating poor alignment. In contrast, (b) demonstrates that our method effectively incorporates Q embeddings (green) into the linguistic space, improving alignment. These results suggest that excessive reliance on text features inhibits the multimodal query embeddings from adapting properly to the passage space.

6 Conclusion

We introduced MIRe, a novel retrieval framework designed for multimodal query retrieval without fusing textual features during the alignment stage. Our query-guided attentive pooling module allows textual embeddings to attend to visual patches while preventing text-driven signals from dominating the visual representations. We also constructed a pre-training dataset by converting concise question-answer pairs into extended passages, thereby exposing the model to more realistic retrieval tasks. Our extensive experiments demonstrate that MIRe consistently outperforms existing methods under both zero-shot and fine-tuned settings. Ablation studies further validate that each component of MIRe is crucial for achieving robust multimodal query retrieval.

7 Limitations

Despite the promising results, our work has several limitations that point to potential directions for future research. First, while our approach demonstrates strong performance across general-domain benchmarks, it remains untested in specialized domains (e.g., medical or legal documents), where multimodal content may exhibit more complex and domain-specific features. Second, we have not explored synergy with retrieval-augmented generative (RAG) frameworks, which typically prepend retrieved passages to a language model for downstream generation tasks. Although we believe our retrieval improvements would benefit RAG-based methods, in line with findings from Kim et al. (2024) showing that stronger retrievers enhance downstream generation, fully validating our approach in a RAG pipeline is left for future work. Finally, our current data construction method focuses on retrieval from large yet homogeneous corpora; adapting the framework to more diverse or dynamically changing knowledge sources may require additional techniques to handle domain shifts or continuously updated information.

8 Acknowledgment

This work was supported by the Agency For Defense Development by the Korean Government(UI247035TF)

References

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

D Chen. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics (ACL)*, pages 2318–2335.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 14948–14968.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. A thousand words are worth more than a picture: Natural language-centric outside-knowledge visual question answering. *arXiv preprint arXiv:2201.05299*.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. KAT: A knowledge augmented transformer for vision-and-language. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 956–968.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of Wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12065–12075.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research (TMLR)*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*, pages 39–48.

Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. 2024. Sure: Summarizing retrievals using answer candidates for open-domain QA of LLMs. In *The International Conference on Learning Representations (ICLR)*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6086–6096.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. 2024. PreFLMR: Scaling up fine-grained late-interaction multi-modal retrievers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5294–5316.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Siqi Liu, Weixi Feng, Tsu-Jui Fu, Wenhu Chen, and William Wang. 2023. EDIS: Entity-driven image search over multimodal web content. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4877–4894.

Xinwei Long, Jiali Zeng, Fandong Meng, Zhiyuan Ma, Kaiyan Zhang, Bowen Zhou, and Jie Zhou. 2024. Generative multi-modal knowledge retrieval with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 18733–18741.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2023. End-to-end knowledge retrieval with multi-modal queries. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8573–8589.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021a. Weakly-supervised visual-retriever-reader for knowledge-based question answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6417—-6431.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021b. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6417–6431.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3204.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, page 861.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3113–3124.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human-generated machine reading comprehension dataset.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9844–9855.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2024. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The International Conference on Learning Representations (ICLR)*.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the International ACM SIGIR conference on research and development in Information Retrieval (SIGIR)*, pages 539–548.

Chen Qu, Hamed Zamani, Liu Yang, W Bruce Croft, and Erik Learned-Miller. 2021. Passage retrieval for outside-knowledge visual question answering. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 1753–1757.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, pages 333–389.

Alireza Salemi, Juan Altmayer Pizzorno, and Hamed Zamani. 2023. A symmetric dual encoding dense retrieval framework for knowledge-intensive visual question answering. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 110–120.

Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. 2022a. Plaid: an efficient engine for late interaction retrieval. In *Proceedings of the ACM International Conference on Information & Knowledge Management (CIKM)*, pages 1747–1756.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022b. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 3715–3734.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8317–8326.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, page 2443–2449.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 809–819.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023. To see is to believe: Prompting GPT-4V for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision (ECCV)*, pages 387–404.

Jialin Wu and Raymond Mooney. 2022. Entity-focused dense passage retrieval for outside-knowledge visual question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8061–8072.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*, volume 36, pages 3081–3089.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. LLaVAR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024a. VISTA: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3185–3200.

Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. 2024b. MARVEL: Unlocking the multi-modal capability of dense retrieval via visual module plugin. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14608–14624.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

A Appendix

A.1 Training and Inference Details

In all experiments, we train models using the AdamW optimizer (Loshchilov and Hutter, 2019) with warm-up steps on a machine with 4 RTX A6000 GPUs. We chose model checkpoints based

Dataset	Hyperparameter					
	LR	# Epochs	# Batch per GPU	# Global Batch	# Warm-up	τ
Pre-training (R2P)	1e-4	5	128	512	300	0.3
OKVQA-GS	5e-5	15	128	512	10	0.8
ReMuQ	5e-5	5	128	512	10	0.8

Table 5: **Summary of hyperparameters utilized for training.** The LR denotes the learning rate.

Dataset	Size		
	#Train	#Test	KB U
OKVQA-GS	8,958	5,046	166,390
OKVQA-WK11M	-	2,523	11,000,000
ReMuQ	8,418	3,609	195,387
E-VQA	-	3,750	51,462

Table 6: **Summary of dataset statistics for evaluation.** This table presents the distribution of training and testing instances alongside the size of the knowledge bases for each dataset employed in our study. GS and KB denote the corpus collected from the Google Search API and used knowledge base, respectively.

Statistic	**Counts**
# Total data	1,356,536
# Images	264,262
# Max. queries per image	12
# Avg. queries per image	8.32
# Queries requiring description	230,877 (17.02%)
# Other types of queries	1,125,659 (82.98%)

Table 7: **Statistics of our constructed dataset.**

on the validation loss. We set hyperparameters for each dataset as shown in Tab. 5.

Pre-training. We used $E_Q = [E_g; E_m]$ without E_t to align visual embeddings with the linguistic space during the pre-training stage. In this stage, we only tuned the mapping network, such as a MLP layer for E_g and the query-guided attentive pooling module. PreFLMR and MIRe were set with the same hyperparameters.

Fine-tuning. For fine-tuning our model on downstream tasks, we tuned all parameters of our model except for the vision model in all experiments. Since the parameters of the vision model are not updated during training, we cached the outputs of the vision model before training. In our setting, training one epoch for our dataset took about 20 minutes on 4 RTX A6000 GPUs, where one epoch encompasses 3625 steps. We detail statistics of benchmark datasets in Tab. 6.

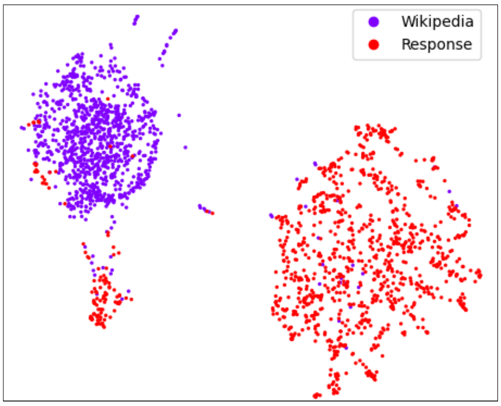


Figure 7: **UMAP visualization of embeddings** extracted using the Contriever model (Izacard et al., 2022), comparing Wikipedia documents (purple) and LLaVA responses (red). The separation between clusters highlights the structural and semantic differences.

Inference. Passages within the knowledge base were pre-indexed, following the method established by the previous work (Santhanam et al., 2022b). The indexing process consists of three critical steps: centroid selection, passage encoding, and index inversion. To enhance storage efficiency, embeddings were compressed to 2 bits per dimension. In the OK-VQA dataset using a corpus collected from Google search API, the retrieval time of MIRe and ColBERTv2 is approximately 0.085 seconds and 0.081 seconds per query on one RTX A6000 GPU, respectively. Thus, MIRe spends slightly more time retrieving relevant passages with multimodal queries, compared to the base text retriever.

A.2 Details for Our Dataset

To construct our dataset, we employ three visual instruction datasets (Zhang et al., 2023; Wang et al., 2023; Liu et al., 2024) and two VQA datasets (Singh et al., 2019; Biten et al., 2019). Initially, samples were split into individual turns. We removed turns with responses shorter than 30 characters only for detailed responses. Subsequently, we edited responses containing simple affirmations (“yes", “no") and excluded samples for tasks irrelevant to retrieval tasks (e.g., location and count),

Source Dataset	# Data	# Images	# Avg. S per I	# Max. S per I
ST-VQA (Biten et al., 2019)	25,154	18,518	1.36	7
TextVQA (Singh et al., 2019)	26,406	18,913	1.40	2
LLaVAR (Zhang et al., 2023)	42,690	19,787	2.16	7
Instruct4V (Wang et al., 2023)	222,711	26,663	8.35	12
LLaVA-1.5 (Liu et al., 2024)	1,017,622	158,429	6.42	12
Subset of WiT (Srinivasan et al., 2021)	500,000	500,000	1	1

Table 8: **Statistics of each source dataset within our dataset.** S per I denotes the number of queries per image.

Source Dataset	OKVQA-GS		OKVQA-WK11M		ReMuQ		E-VQA	
	PR@5	PR@10	R@5	R@10	R@5	R@10	PR@5	PR@10
ST-VQA (Biten et al., 2019)	72.29	81.45	57.35	67.97	85.79	87.81	51.01	58.37
TextVQA (Singh et al., 2019)	72.18	81.75	57.83	69.20	86.03	87.97	51.41	58.80
LLaVAR (Zhang et al., 2023)	73.11	82.62	60.88	71.90	86.34	88.39	51.89	58.75
Instruct4V (Wang et al., 2023)	78.72	86.88	65.83	75.90	86.01	87.81	52.0	59.25
LLaVA-1.5 (Liu et al., 2024)	79.41	87.77	68.05	78.32	86.56	88.31	**52.56**	**59.92**
Total	**81.11**	**88.84**	**70.55**	**82.20**	**87.45**	**88.45**	51.95	59.12

Table 9: **Zero-shot performance by each source dataset.** We apply our response-to-passage conversion process to each source dataset. Note that we did not add WiT data in this experiment.

where we automatically filtered out based on specific phrases.

Fig. 7 illustrates there exists a clear distinction between the concise responses and more expansive passages, supporting our perspective. After the pre-processing, we refined the data through a response-to-passage conversion using ColBERTv2, a text retriever trained on the MS MARCO Passage Ranking task (Nguyen et al., 2016). Responses were converted into passages using a pool of 6 million Wikipedia documents (Chen et al., 2023), with textual queries limited to 128 tokens. As shown in Fig. 8, our constructed dataset is featured by pairs of multimodal queries and passages including responses to different queries about the same image, advancing the capability to retrieve relevant information from multimodal queries. This process yielded a total of 1.36 million QA pairs; further data statistics are provided in Tab. 7 and Tab. 8.

Table 9 summarizes the zero-shot retrieval performance for each source dataset. The results demonstrate that our conversion process effectively leverages complementary strengths from various datasets, underscoring the robustness of our approach. Additionally, when unifying WiT data, we assigned textual queries by randomly sampling from the following prompts: "What is the main object?", "Identify the subject of this image.", "Who or what is the subject in this picture?", "Identify the main entity.", and "What is the core object or subject shown here?".

A.3 Architectural Differences: MIRe vs. PreFLMR

Our architecture differs from PreFLMR in several important aspects. In PreFLMR, visual tokens serve as queries, and the hidden states of the text encoder act as keys and values within a cross-attention mechanism. Attention scores $\mathcal{A} \in \mathbb{R}^{H \times Q \times K}$ are computed using these components, and value vectors $\mathcal{V}$ are derived from the truncated text encoder hidden states, which may cause information loss. The resulting outputs have dimensions $[H, Q, D]$ and are reshaped to $[Q, H \cdot D]$ before undergoing further reduction by the ColBERT head.

In contrast, MIRe uses textual embeddings E_t from the ColBERT head as queries, with visual tokens split into multiple heads for keys and values. Here, Q corresponds to the number of textual tokens and K to the number of visual tokens, reversing the modality roles compared to PreFLMR. Value vectors are obtained directly from the visual embeddings without truncation. Output aggregation is performed by mean-pooling across the sequence dimension, producing representations of shape $[H, D]$, which are passed through a linear layer without further dimensionality reduction.

Image *I*	Question *T*	Passage *D*
	What feature can be seen on the back of the bus?	In bus advertising, buses and their related infrastructure is a medium commonly used by advertisers to reach the public with their message. Usually, this takes the form of promoting commercial brands, but can also be used for public campaign messages. The back of the bus features an advertisement. School bus advertising is a form of advertising in the United States in which advertising space is sold on the sides of school buses. …
	What are the colors of the bus in the image?	Red & White Services was a bus company operating in south east Wales and Gloucestershire, England between 1929 and 1978. Red & White evolved into Red & White United Transport Ltd, formed in 1937, which owned bus and road freight companies in the United Kingdom and Southern Africa. The bus is white and red. Red Jammers are the vintage White Motor Company/Bender Body Company Model 706 buses used at Glacier National Park in the United States …
 Image *I*	What type of surface are the birds standing on?	The Sandy Island and Low Rock Important Bird Area comprises two small islands with a collective area of 9 lying at the western end of the Gulf of Carpentaria in the Northern Territory of Australia. They lie south-west of Groote Eylandt off the coast of Arnhem Land, with the nearest settlement there the Numbulwar community. The birds are standing on a sandy beach. Sandy Island is a 20 (ha) island lying about 5 (km) off the coast of Windy Harbour in south-west Western Australia, and near Point D' …
	Are there any specific objects near the birds in the image?	In agriculture, poultry litter or broiler litter is a mixture of poultry excreta, spilled feed, feathers, and material used as bedding in poultry operations. This term is also used to refer to unused bedding materials. There are straws lying on the sand near the birds. Bat-fowling is an archaic method of catching birds at night, while they are at roost. The process involves lighting straw or torches near their roost. After awakening them from their roost, the birds fly toward the flames, …
 Image *I*	What is the man doing in the image?	A passing shot is a forceful shot, as in tennis or team handball, that travels to one side out of the reach of one\'s opponent. In tennis, this shot is generally a groundstroke and is used when one\'s opponent is running to the net or if they are at the net already. … The man is playing tennis near the net and getting ready to hit a ball. he might have just made a play, and he is attempting to return the ball to continue the tennis match. Gamesmanship is the use of dubious (although not technically illegal) methods to win or gain a serious advantage in a game or sport. …
	Is there any official or umpire present in the image?	A challenge is a request made to the holder of a competitive title for a match between champion and challenger, the winner of which will acquire or retain the title. … There is an official looking on indicating that the tennis match is likely a formal or competitive one. The tennis scoring system is a standard widespread method for scoring tennis matches, including pick-up games. Some tennis matches are played as part of a tournament, which may have various categories, such as singles and doubles. The great majority are organised as a single-elimination tournament, with competitors being eliminated after a single loss, and the overall winner being the last competitor without a loss. A tournament is a competition involving at least three competitors, all participating in a sport or game. More specifically, the term may be used in …

Figure 8: **Examples for our dataset.** The figure illustrates samples in the dataset, where the red-colored text denotes inserted responses.