Enhancing Multimodal Retrieval via Complementary Information Extraction and Alignment

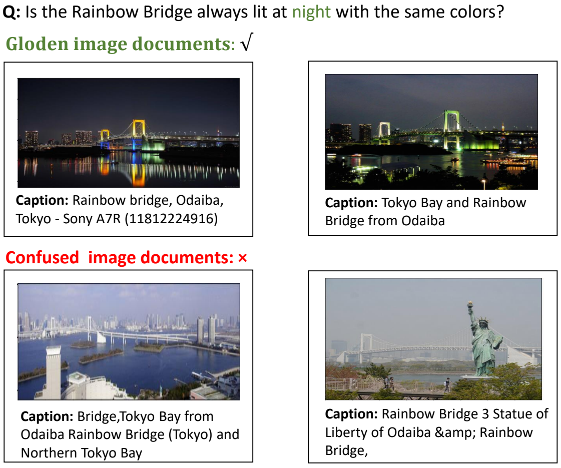
Delong Zeng[1,*] **Yuexiang Xie**[2] **Yaliang Li**[2] **Ying Shen**[1,3,†]
[1]School of Intelligent Systems Engineering, Sun Yat-sen University [2]Alibaba Group
[3]Guangdong Provincial Key Laboratory of Fire Science and Intelligent Emergency Technology
`zengdlong@mail2.sysu.edu.cn, {yuexiang.xyx, yaliang.li}@alibaba-inc.com`
`sheny76@mail.sysu.edu.cn`

Abstract

Multimodal retrieval has emerged as a promising yet challenging research direction in recent years. Most existing studies in multimodal retrieval focus on capturing information in multimodal data that is similar to their paired texts, but often ignores the complementary information contained in multimodal data. In this study, we propose CIEA, a novel multimodal retrieval approach that employs **C**omplementary **I**nformation **E**xtraction and **A**lignment, which transforms both text and images in documents into a unified latent space and features a complementary information extractor designed to identify and preserve differences in the image representations. We optimize CIEA using two complementary contrastive losses to ensure semantic integrity and effectively capture the complementary information contained in images. Extensive experiments demonstrate the effectiveness of CIEA, which achieves significant improvements over both divide-and-conquer models and universal dense retrieval models. We provide an ablation study, further discussions, and case studies to highlight the advancements achieved by CIEA. To promote further research in the community, we have released the source code at https://github.com/zengdlong/CIEA.

1 Introduction

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020b; Gao et al., 2023; Yi et al., 2024) has recently attracted widespread attention for its role in enhancing large language models (LLMs) (Touvron et al., 2023; Bai et al., 2023; Chung et al., 2024; Brown et al., 2020) by providing up-to-date information and alleviating hallucination issues. Most existing studies focus on retrieving textual information (Cheng et al., 2023; Shi et al., 2024; Wang et al., 2024a; Zhu et al., 2025), leveraging text corpora to provide factual support for model responses. As the scale of multimodal data continues to grow, effectively supporting multimodal retrieval to provide knowledge beyond text has emerged as a promising yet challenging research area (Zhao et al., 2023; Kuang et al., 2025).


Q: Is the Rainbow Bridge always lit at night with the same colors?

Gloden image documents: √

Caption: Rainbow bridge, Odaiba, Tokyo - Sony A7R (11812224916)

Caption: Tokyo Bay and Rainbow Bridge from Odaiba

Confused image documents: ×

Caption: Bridge,Tokyo Bay from Odaiba Rainbow Bridge (Tokyo) and Northern Tokyo Bay

Caption: Rainbow Bridge 3 Statue of Liberty of Odaiba &amp; Rainbow Bridge,

Figure 1: An example of multimodal retrieval. The images contain critical information that can assist in query responses, which might not be present in the text.

To promote the progress of multimodal retrieval, researchers have employed captioning models to convert multimodal data into text (Wu et al., 2023; Baldrati et al., 2023; Mahmoud et al., 2024), enabling the application of existing retrieval techniques developed for text corpora. However, these approaches can be heavily influenced by the effectiveness of the captioning models, which might result in the loss of crucial information during the conversion process (Che et al., 2023; Li et al., 2024b). Recent studies propose representation-based approaches, wherein textual and multimodal information are mapped into a unified embedding space for knowledge retrieval, after a separate encoding process (Radford et al., 2021; Zhang et al., 2021) or a project-based joint encoding process (Zhou et al., 2024; Li et al., 2024a).

[*]Work done as an intern at Alibaba Group.
[†]Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22092–22105
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

In this study, we focus on a novel representation-based approach that aims to transfer multimodal data into the latent vector space of language models. We identify a critical limitation of existing representation-based methods (Wang et al., 2024b; Zhou et al., 2024): they tend to focus on capturing the information in multimodal data that is similar to their textual counterparts (such as captions or associated text) while neglecting the complementary information in multimodal data. For example, as shown in Figure 1, a textual description paired with an image might reference the *Rainbow Bridge*. While the image can indeed provide similar visual details of *Rainbow Bridge*, supplementary information such as the nightscape and the color of the bridge can be also important. For the representation-based approach, extracting and preserving such complementary information in multimodal data can enhance the quality of the responses to queries that cannot be fully resolved using textual information alone.

Inspired and motivated by the above insights, we propose a novel multimodal retrieval approach that involves **C**omplementary **I**nformation **E**xtraction and **A**lignment, denoted as **CIEA**. Specifically, we adopt a language model to transform text into a latent space, while employing the CLIP model (Radford et al., 2021; Li et al., 2023) and a projector to map image information into the same unified latent space. Then, we design a complementary information extractor, which identifies the differences in representations between the images and the text in the documents. Based on these differences, we update the representations of the images to integrate complementary information. Furthermore, we introduce a novel optimization method tailored for CIEA by constructing two complementary contrastive loss functions: one that ensures the semantics integrity of the learned representations, and another that enhances the extraction of complementary information from the images.

Extensive experiments are conducted to demonstrate the effectiveness of CIEA in multimodal retrieval. The experimental results show that the proposed method achieves noticeable improvements compared to both divide-and-conquer models and universal dense retrieval models. An ablation study is carried out to highlight the contributions of the different components of CIEA. Besides, we provide further discussions on the effect of the language model and include some case studies for a better understanding of CIEA.

2 Related Work

Conventional retrieval models (Lewis et al., 2020b; Yu et al., 2023) encode queries and documents into vectors via language models (Chen et al., 2024), trained with contrastive learning and retrieved using KNN (Su et al., 2022). Multimodal retrieval, compared to text-only retrieval, incorporates multiple modalities of information, necessitating effective utilization strategies for information from different modalities (Yuan et al., 2021). One approach involves using caption models to convert information from other modalities into text, effectively converting multimodal retrieval tasks into text-only retrieval tasks (Baldrati et al., 2023). However, such an approach may lead to information loss (Che et al., 2023; Li et al., 2024b).

Another approach is to employ representation models, where visual and textual encoders encode information separately before fusion (Radford et al., 2021). Early studies adopt a divide-and-conquer strategy, encoding each modality separately and concatenating vectors to fuse information, which potentially causes modality competition. To tackle this, UniVL-DR (Liu et al., 2023b) proposes a universal multimodal retrieval framework by encoding queries and documents into a unified embedding space for retrieval, routing, and fusion. Some existing studies (Wang et al., 2022a,b) involve training large models to unify text and visuals. However, differing representations complicate the acquisition of sufficient data for effective semantic understanding (Lu et al., 2023).

Recent studies (Zhou et al., 2024; Wang et al., 2024b; Li et al., 2024a, 2023) propose a project-based framework, which leverages models like CLIP (Radford et al., 2021) to convert visual inputs into feature sequences and introduce projector layers to align these sequences with language model embeddings. Such a framework facilitates the comparison of visual and textual content at the embedding level and converts visual information into language model “tokens”, capitalizing on knowledge infused during the training of language models. For example, MARVEL (Zhou et al., 2024) utilizes the project-based framework for capturing multimodal information within the output space of the language model; MCL (Li et al., 2024a) adds a retrieval token for enhancing the model’s performance in representation learning.

Although remarkable progress has been made, visual information remains a novel and underex-

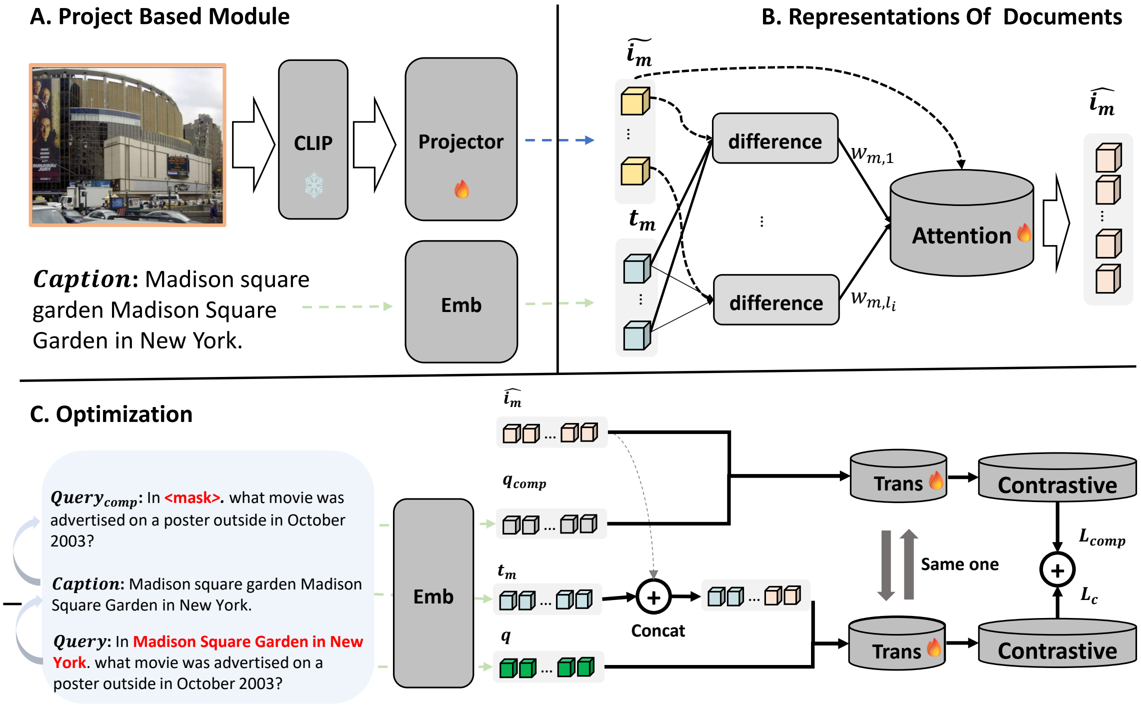

A. Project Based Module
CLIP
Projector
Caption: Madison square garden Madison Square Garden in New York.
Emb
B. Representations Of Documents
$\widetilde{i_m}$
t_m
difference
difference
$w_{m,1}$
w_{m,l_i}
Attention
$\widehat{i_m}$
C. Optimization
$\widehat{i_m}$
q_{comp}
t_m
q
Query$_{comp}$: In <mask>. what movie was advertised on a poster outside in October 2003?
Caption: Madison square garden Madison Square Garden in New York.
Query: In Madison Square Garden in New York. what movie was advertised on a poster outside in October 2003?
Emb
Concat
Trans
Trans
Same one
Contrastive
Contrastive
L_{comp}
L_c

Figure 2: The overall architecture of CIEA. The upper part illustrates how queries and multimodal documents are transformed into a unified latent space, while the lower part details the optimization process of CIEA.

plored component for language models, which motivates us to provide effective solutions for processing visual information to attain a comprehensive representation for enhancing multimodal retrieval.

3 Methodology

3.1 Preliminary

In this study, we focus on multimodal retrieval, which aims to find one or more relevant documents from a knowledge base in response to a given text query q. A knowledge base consisting of a total of N documents can be denoted as $D = \{d_1, d_2, \ldots, d_N\}$. Each document might contain text and images, i.e., $d_m = \{t_m, i_m\}$, where t_m and i_m represent the text and images in the document, respectively.

The main objective of multimodal retrieval is to align the representations of the query with the corresponding multimodal documents, ensuring that similar items are closely matched in a unified latent space. To achieve this, a multimodal encoder is employed to encode both the query and the documents, transforming them into dense representations, which can be given as:

$$\begin{aligned} \mathbf{q} &= \text{Encoder}(q), \\ \mathbf{d}_m &= \text{Encoder}(d_m), \forall d_m \in D. \end{aligned} \tag{1}$$

After that, the similarities between the query and documents are measured via cosine similarity:

$$\cos(\mathbf{q}, \mathbf{d}_m) = \frac{\mathbf{q} \cdot \mathbf{d}_m}{\|\mathbf{q}\| \|\mathbf{d}_m\|}. \tag{2}$$

In the following subsections, we provide a detailed introduction of the proposed **C**omplementary **I**nformation **E**xtraction **A**lignment, denoted as **CIEA**. The overall architecture of CIEA is illustrated in Figure 2. Specifically, we first encode the queries and multimodal documents via a language model and a CLIP model, as shown in Section 3.2. Then, in Section 3.3, we design a complementary information extractor to capture the complementary information contained in the images. The optimization designed for the proposed CIEA is introduced in Section 3.4.

3.2 Representations of Queries and Multimodal Documents

Firstly, as shown in part A of Figure 2, we transform both the queries and the multimodal documents into their dense representations. Inspired by previous studies (Zhou et al., 2024; Li et al., 2024a, 2023), we utilize a transformer-based language model as the backbone of the encoder to encode both text and images. A transformer-based language model typically consists of an embedding layer followed by several transformer blocks. We denote the embedding layer as $\text{Emb}(\cdot)$ and the transformer blocks as $\text{Trans}(\cdot)$.

The text query q can be transformed into dense representations via language models, which can be given as:

$$\mathbf{q} = \text{Trans}(\text{Emb}(q)), \tag{3}$$

where $\mathbf{q} \in \mathbb{R}^{l_q \times d}$, l_q denotes the number of tokens in the query, and d denotes the dimension of the dense representations.

For a document containing both text and images, i.e., $d_m = \{t_m, i_m\}$, the text t_m and image i_k are processed separately before being fused. Specifically, we begin by feeding the text within the document to the embedding layer:

$$\mathbf{t}_m = \text{Emb}(t_m), \tag{4}$$

where $\mathbf{t}_m = [\mathbf{t}_m^{(1)}, \mathbf{t}_m^{(2)}, \ldots, \mathbf{t}_m^{(l_t)}] \in \mathbb{R}^{l_t \times d}$ and l_t denotes the number of token in the text.

The images in the documents can be typically represented as RGB multi-channel matrices. We adopt a frozen CLIP (Radford et al., 2021) visual encoder to transform these matrices into a set of semantic vectors. Such a process involves dividing an image into multiple patches, with each patch representing a different region of the image, which can be formally given as:

$$\mathbf{i}_m = \text{CLIP}_{\text{Visual}}(i_m), \tag{5}$$

where $\mathbf{i}_m \in \mathbb{R}^{l_i \times d_{\text{clip}}}$, l_i denotes the number of patches, and d_{clip} denotes the hidden dimension of CLIP visual encoder.

Note that the hidden dimension of these image representations might be different from that of text representations, i.e., $d_{\text{clip}} \neq d$. Therefore we adopt a linear layer as the projector for further alignment:

$$\widetilde{\mathbf{i}_m} = \text{Proj}(\mathbf{i}_m), \tag{6}$$

where $\widetilde{\mathbf{i}_m} = [\widetilde{\mathbf{i}_m}^{(1)}, \widetilde{\mathbf{i}_m}^{(2)}, \ldots, \widetilde{\mathbf{i}_m}^{(l_i)}] \in \mathbb{R}^{l_i \times d}$.

3.3 Complementary Information Extractor

The proposed CIEA aims to capture the complementary information contained in images, i.e., the information not encompassed by the text in a document. To achieve this, we design a complementary information extractor that captures the differences between images and textual content. Specifically, we calculate the patch-level distances between text and image representations, identifying the maximum value to serve as a difference measurement, which is similar to those adopted in BLIP-2 (Li et al., 2023). Formally, the difference between the j-th patch of image and the text within document m, denoted as $r_{m,j}$, can be calculated as:

$$r_{m,j} = -\max_{c \in [l_t]} \cos(\widetilde{\mathbf{i}_m}^{(j)}, \mathbf{t}_m^{(c)}). \tag{7}$$

The obtained differences are normalized into a range of $[0, 1]$ for effectively weighing the image patches, as given by:

$$w_{m,j} = \frac{1 + r_{m,j}}{2}. \tag{8}$$

We adopt an attention layer for re-weighting the patches of the image based on the calculated $\mathbf{w}_m = w_{m,j}\ \forall j \in [l_i]$. The attention scores of embeddings are multiplied by the corresponding weights for normalization, which can be given as:

$$\mathbf{Q}_m = \widetilde{\mathbf{i}_m}\mathbf{W}_Q,\ \mathbf{K}_m = \widetilde{\mathbf{i}_m}\mathbf{W}_K,\ \mathbf{V}_m = \widetilde{\mathbf{i}_m}\mathbf{W}_V,$$

$$\widehat{\mathbf{i}_m} = \text{softmax}\left(\frac{\mathbf{Q}_m\mathbf{K}_m \cdot \mathbf{w}_m}{\sqrt{d}}\right)\mathbf{V}_m, \tag{9}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable matrices and "$\cdot$" here stands for the broadcasting method.

After that, as shown in part B of Figure 2, the embeddings of image and text are concatenated together and fed into transformer blocks for multimodal knowledge fusion. The obtained dense representation of the document d_m can be given as:

$$\mathbf{d}_m = \text{Trans}([\mathbf{e}_{\text{start}} \oplus \widehat{\mathbf{i}_m} \oplus \mathbf{e}_{\text{end}} \oplus \mathbf{t}_m]), \tag{10}$$

where $\oplus$ denotes the concatenate operation, and $\mathbf{e}_{\text{start}}$ and $\mathbf{e}_{\text{end}}$ denote the embeddings of special tokens <*start*> and <*end*>, respectively, for distinguishing the text and image embeddings explicitly.

Finally, for a given query, we rank the documents based on the similarities between their representations (refer to Eq.(10)) and the query representation (refer to Eq.(3)) to obtain the relevant documents.

3.4 Optimization

The trainable parameters of CIEA framework include the projector (as defined in Eq. (6)), the added attention weights $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ for re-weighting (as defined in Eq. (9)), and the transformer layers of the language models (as defined in Eq. (10)). In this section, we introduce how to optimize these trainable parameters in CIEA.

	mrr@10	**ndcg@10**	**mrr@20**	**ndcg@20**	**rec@20**	**rec@100**
BM25	22.11	22.92	22.80	25.41	46.27	62.82
CLIP-DPR	37.35	37.56	37.93	40.77	69.38	85.53
BM25 & CLIP-DPR	42.27	41.58	42.79	44.69	73.34	87.50
BM25 & CLIP (oracle)	61.05	58.18	61.37	60.45	80.82	90.83
VinVL-DPR	38.14	35.43	38.74	37.79	53.89	69.42
CLIP-DPR	48.83	46.32	49.34	49.11	69.84	86.43
UniVL-DR	62.40	59.32	62.96	61.22	80.37	89.42
T5-ANCE	64.13	62.03	64.40	64.02	83.81	92.07
MARVEL	65.43	63.07	65.67	65.00	84.32	92.27
CIEA (ours)	**66.16**	**63.89**	**66.41**	**65.85**	**85.43**	**92.75**

Table 1: Comparisons on the WebQA-Multi dataset. The best results are marked with bold.

Firstly, we identify the text segments from the queries in the training set that correspond to the text in the ground truth documents. For example, as shown in Figure 2, we mask out "*Madison Square Garden in New York*" that appears in the caption. We match the token IDs generated by the tokenizer of the language model for the query and the text in document for such identification.

After that, we replace these identified text segments with special tokens <*mask*> while preserving the remaining parts of the query. We denote the processed query as q_{comp}, whose representations can be obtained in a similar manner to the computation defined in Eq. (3):

$$\mathbf{q}_{\text{comp}} = \text{Trans}(\text{Emb}(q_{\text{comp}})). \tag{11}$$

We also construct negative samples for the queries to apply a contrastive loss for optimization. The documents annotated as relevant to the query are denoted as D^+, while N negative documents are sampled from the knowledge base and denoted as $D^- = \{d_1^-, d_2^-, \ldots, d_N^-\}$. The contrastive loss can be defined as:

$$\mathcal{L}_c = -\log \frac{e^{\cos(\mathbf{q},\mathbf{d}^+)/\tau}}{e^{\cos(\mathbf{q},\mathbf{d}^+)/\tau} + \sum_{D^-} e^{\cos(\mathbf{q},\mathbf{d}^-)/\tau}}, \tag{12}$$

where τ denotes the temperature, $\cos(\cdot)$ denotes the cosine similarity function, $\mathbf{q}$ denotes the representation of the query, $\mathbf{d}^+$ and $\mathbf{d}^-$ denote the representations of positive documents and negative documents, respectively.

Meanwhile, to enhance the extraction of complementary information from images within the documents, we construct another contrastive loss function based on $\mathbf{q}_{\text{comp}}$ and the representations of images, which can be given as:

$$\mathcal{L}_{\text{comp}} = -\log \frac{e^{\cos(\mathbf{q}_{\text{comp}},\mathbf{d}_{\text{img}}^+)/\tau}}{e^{\cos(\mathbf{q}_{\text{comp}},\mathbf{d}_{\text{img}}^+)/\tau} + \sum_{D^-} e^{\cos(\mathbf{q}_{\text{comp}},\mathbf{d}_{\text{img}}^-)/\tau}}, \tag{13}$$

where $\mathbf{d}_{\text{img}} = \text{Trans}(\widehat{\mathbf{i}_k})$ denotes the representations of the images in the documents. The intuition behind the above loss function is that we promote the representations of images to align with the information contained in the query that is not present in the text within documents.

Finally, the training objective of the proposed CIEA can be given as:

$$\mathcal{L} = \mathcal{L}_c + \lambda \cdot \mathcal{L}_{\text{comp}}, \tag{14}$$

where λ is a hyperparameter that balances the contributions of the two contrastive losses.

4 Experiments

4.1 Settings

Datasets We conduct extensive experiments on the WebQA (Chang et al., 2022) and EDIS (Liu et al., 2023a) dataset to compare the proposed method against baselines. WebQA is an open-domain multimodal question answering dataset in which each query is associated with one or more relevant documents. There are 787,697 documents containing only textual information, while 389,750 documents contain both text and images. For clarity, we refer to this dataset as **WebQA-Multi** in this study. Besides, to better show the effectiveness of multimodal retrieval, we extract all documents containing images and construct the **WebQA-Image**

	WebQa-Image				**EDIS**			
	mrr@10	**ndcg@10**	**mrr@20**	**ndcg@20**	**mrr@10**	**ndcg@10**	**mrr@20**	**ndcg@20**
CLIP-DPR	59.78	61.05	60.2	63.28	62.52	39.19	62.90	44.24
UniVL-DR	65.95	67.33	66.24	69.01	62.89	39.50	63.29	44.54
T5-ANCE	64.38	65.65	64.73	67.52	63.95	39.74	64.30	43.76
MARVEL-DPR	62.51	63.48	62.90	65.54	63.60	37.76	63.98	41.89
MARVEL	66.43	67.60	66.76	69.40	67.00	42.19	67.28	46.63
CIEA (ours)	**67.40**	**68.77**	**67.74**	**70.51**	**68.11**	**42.57**	**68.42**	**47.19**

Table 2: Comparisons on the WebQa-Image and EDIS. The best results are marked with bold.

dataset. **EDIS** retrieves images via textual queries, with 26,000 training, 3,200 development, and 3,200 test samples from 1M image corpus (each paired with a text caption). More statistics of datasets can be found in Appendix A.

Metrics To evaluate the effectiveness of the proposed method and baselines in multimodal retrieval, we adopt six widely-used metrics, including MRR@10, MRR@20, NDCG@10, NDCG@20, REC@20, and REC@100. The MRR (Bajaj et al., 2016) and NDCG (Taylor et al., 2008) metrics are calculated using the official code[1].

Implementation Details In the experiments, we employ T5-ANCE (Yu et al., 2023) as the foundational language model and CLIP as the visual understanding module to implement the proposed CIEA model to enhance the image understanding capabilities. The document representation capabilities of T5-ANCE are enhanced by using ANCE (Xiong et al., 2021) to sample negative documents based on T5 (Raffel et al., 2020). We truncate the text input length to 128 tokens. During training, we adopt the AdamW (Loshchilov and Hutter, 2019) optimizer with a maximum of 40 epochs, a batch size of 64, a learning rate of 5e-6, and set the temperature hyperparameter τ to 0.01. The hyperparameter λ is set to 0.0011 for WebQA-Multi, 0.019 for WebQA-Image, and 0.001 for EDIS, respectively. More implementation details can be found in Appendix B.

4.2 Baselines

We compare CIEA with two different types of baselines: divide-and-conquer models and universal dense retrieval models.

Divide-and-conquer models suggest retrieving image and text documents separately and then merging the retrieval results. We employ various single-modal retrievers to instantiate the divide-and-conquer models, including VinVL-DPR (Zhang et al., 2021), CLIP-DPR (Radford et al., 2021), and BM25 (Robertson et al., 2009). The results of multi-modality retrieval are combined based on their uni-modal rank reciprocals or oracle modality routing. The latter approach demonstrates the maximum potential performance of our divide-and-conquer models in retrieval tasks.

Regarding **universal dense retrieval models**, we adopt the following recent studies as baselines: (i) Pre-trained multimodal alignment models VinVL-DPR (Zhang et al., 2021) and CLIP-DPR (Radford et al., 2021), which are trained using the DPR framework; (ii) UniVL-DR (Liu et al., 2023b), which employs modality-balanced hard negatives to train text and image encoders, and uses image language adaptation techniques to bridge the modality gap between images and text; (iii) MARVEL (Zhou et al., 2024), a project-based approach that encodes image documents with CLIP and projects them into the embedding layer of a language model for fusion.

Comparisons The experimental results on WebQA and EDIS are shown in Table 1 and 2, respectively. From these results we can observe that, the proposed CIEA achieves noticeable improvements across various metrics compared to both divide-and-conquer models and universal dense retrieval models. CIEA achieves outperformance on both WebQA and EDIS, demonstrating its effectiveness in capturing crucial information from both text and images for enhancing multimodal retrieval. As shown in Tables 1, compared to universal dense retrieval models that employ projectors for language model grounding (e.g., MARVEL, which shows improvement over text-only methods with a 1.3 MRR@10 improvement over T5-ANCE), CIEA

[1] https://github.com/microsoft/MSMARCO-Passage-Ranking/blob/master/ms_marco_eval.py

Language model	setting	mrr@10	ndcg@10	mrr@20	ndcg@20	rec@20	rec@100
T5-ANCE	Text only	64.13	62.03	64.40	64.02	83.81	92.07
	Project-based	65.43	63.07	65.67	65.00	84.32	92.27
	CIEA	**66.16**	**63.89**	**66.41**	**65.85**	**85.43**	**92.75**
BERT	Text only	61.39	58.20	61.69	60.32	79.59	89.10
	Project-based	63.03	60.20	63.31	62.37	**81.80**	**90.78**
	CIEA	**63.63**	**60.46**	**63.90**	**62.49**	81.51	90.57
BART	Text only	60.14	56.82	60.49	59.14	78.41	89.25
	Project-based	62.67	59.70	62.98	62.04	81.77	90.77
	CIEA	**63.73**	**60.43**	**64.04**	**62.72**	**81.96**	**90.98**
GPT2	Text only	54.68	51.49	55.07	53.77	72.91	84.86
	Project-based	58.59	55.00	58.92	57.15	75.84	86.79
	CIEA	**59.25**	**55.35**	**59.59**	**57.62**	**76.47**	**86.92**
GPT2-LARGE	Text only	63.58	60.44	63.86	62.45	81.24	89.98
	Project-based	64.33	61.98	64.60	62.04	83.97	92.11
	CIEA	**65.38**	**63.00**	**65.62**	**64.96**	**84.96**	**92.68**

Table 3: Experiments results when using different language models. The *Text only* setting denotes we only use the document's captions for training without incorporating image information, while the *Project-based* setting is the most common projection method without visual alignment.

	mrr@10	ndcg@10	rec@20	rec@100
CIEA	**66.03**	**63.70**	**85.14**	**92.63**
w/o image query	65.66	63.44	84.74	92.54
w/o attention	65.90	63.39	84.95	92.60
Base	65.30	63.19	84.49	92.40

Table 4: Ablation study on WebQa-multi. Here, "w/o" stands for "without," and "Base" refers to the setting where both modules are removed.

achieves a further 0.74 MRR@10 improvement, with analogous enhancements replicated in Table 2. These CIEA-driven advancements underscore the critical benefits of complementary information alignment - a framework that enables superior visual information modeling and enhances the representational power of project-based models.

Ablation Study We conduct an ablation study to demonstrate the contributions of the image query and the attention layer. Specifically, we remove the attention layer designed for re-weighting (refer to Eq. (9)) and the utilization of image query (refer to Eq. (13)), respectively. We also perform *base* setting by removing these two components.

The experimental results are shown in Table 4. Although the removal of the image query does not lead to a decrease in rec@100, other metrics exhibit varying degrees of decline, particularly rec@20, which drops from 85.14 to 84.74. On the other hand, the removal of the attention mechanism results in larger significant decreases in ndcg@10 metrics, highlighting the importance of complementary information extraction. Further, we also compare the baseline configuration with the simultaneous removal of both modules. The results indicate that removing a single module yields improvements in various metrics compared to the baseline, demonstrating that both the image query and the attention layer contribute to enhancing retrieval accuracy in multimodal retrieval. More results can be found in Appendix C.

4.3 Further Discussions

The Effects of Language Models in CIEA To provide further discussions regarding the effects of language models used in the proposed CIEA, we conduct experiments with different language models as the backbone, including T5-ANCE (Yu et al., 2023), BART (Lewis et al., 2020a), BERT (Devlin et al., 2019), GPT-2, and GPT-2-LARGE (Radford et al., 2019). These language models cover three mainstream architectures, i.e., encoder-decoder, encoder-only, and decoder-only. More implementation details can be found in Appendix D.

The experimental results in Table 3 demonstrate the significant impacts of the backbone language model on multimodal retrieval effectiveness. For text-only retrieval, language models with larger parameter sizes, such as GPT-2 LARGE

Weights	MRR@10	NDCG@10	REC@20	REC@100
Dissimilar	**66.03±0.012**	**63.70 ± 0.030**	**85.14±0.052**	**92.63±0.008**
Similar	65.60±0.038	63.35±0.005	85.00±0.012	92.61±0.003

Table 5: Experimental results with different weights on WebQA-Multi.

and T5-ANCE, achieve better overall performance. Both project-based models and CIEA benefit from powerful backbones, showing performance gains through image-derived information. Notably, CIEA maintains competitive results across different backbones compared to project-based approaches, confirming its robustness.

Attention Weight Selection The application of maximum cosine similarity for computing patch-level similarities is designed to extract image regions exhibiting lower correspondence with textual descriptions. To determine the optimal extraction approach, we perform empirical experiments that retained original cosine values rather than their complements. As demonstrated in Table 5, explicit dissimilarity computation has proven more effective for capturing fine-grained patch-level image-text mappings.

Computational Efficiency Although the proposed dual losses might increase training complexity, the computation of the losses is relatively independent and would not lead to a multiplicative increase in overall complexity. The proposed method achieves similar computational efficiency compared to MARVEL, which is also the project-based method. For example, with the sample devices, the training on the WebQA-Multid dataset with 4,966 samples with MARVEL needs around 6.1 minutes while that of CIEA is around 6.5 minutes.

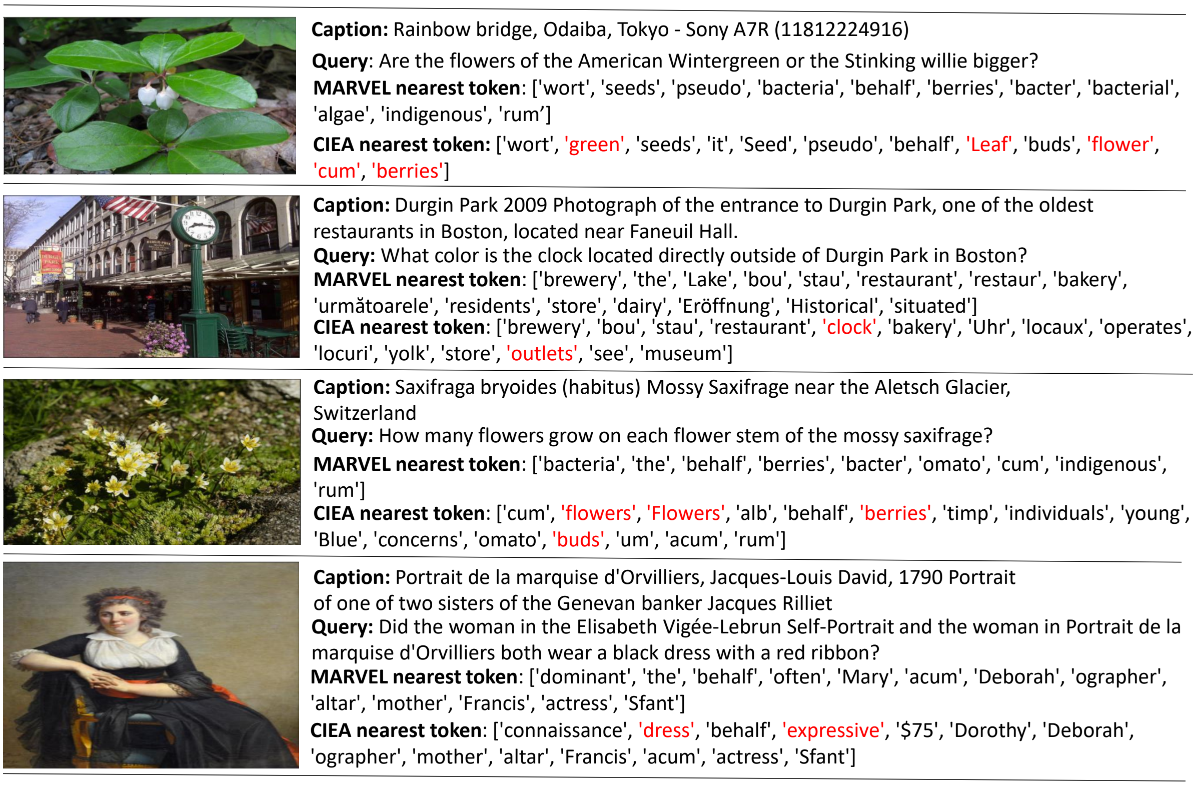
Case Study We conduct case studies on the WebQA dataset for better understanding. We apply cosine similarity to identify the words from the language model's vocabulary that are closest to the embeddings projected from images, showing what information is extracted from images. We compare the proposed CIEA and the strongest baseline MARVEL. The results are illustrated in Figure 3, which indicate that MARVEL focuses on image information that is close to the text, while CIEA captures more supplemental information from images. For example, CIEA captures some terms, such as *green*, *leaf*, and *clock*, that are semantically related to the image but are not mentioned in the caption. The term *clock* in the second case and *flowers* in the third case are relevant to the query, which can assist the model in efficiently locating and retrieving relevant content. These results further confirm the advances of CIEA for multimodal retrieval. Discussions regarding the failure cases are provided in Appendix E, which further illustrate the model's behavior in challenging scenarios and inspire future improvements.

5 Conclusion

In this paper, we propose CIEA to enhance the effectiveness of multimodal retrieval. The main idea of CIEA is to enhance the capturing of complementary information in multimodal data. Specifically, CIEA utilizes language models and CLIP to transform multimodal documents into a unified latent space. For complementary information extraction, we calculate the patch-level distances between text and images, which are then used to re-weight the image representations. Regarding the optimization of CIEA, besides applying a contrastive loss for learning the semantics of text, we also encourage the alignment of image representations with the complementary information in queries. We conduct a series of experiments to demonstrate the advantages of CIEA compared to two different types of multimodal retrieval approaches. Further discussions on the effect of language models show the robustness of CIEA, and several case studies are included for better understanding.

Limitations

Multimodal retrieval is currently in a phase of rapid development. In this paper, our exploration is limited to queries that contain only text, with a focus on extracting information from images. Expanding the proposed approach to include more types of multimodal data, such as audio and video, is a promising direction for future research. Besides, although we conduct experiments with various language models, we are constrained by computational resources and have not yet explored larger-sized models.



Caption: Rainbow bridge, Odaiba, Tokyo - Sony A7R (11812224916)

Query: Are the flowers of the American Wintergreen or the Stinking willie bigger?

MARVEL nearest token: ['wort', 'seeds', 'pseudo', 'bacteria', 'behalf', 'berries', 'bacter', 'bacterial', 'algae', 'indigenous', 'rum']

CIEA nearest token: ['wort', 'green', 'seeds', 'it', 'Seed', 'pseudo', 'behalf', 'Leaf', 'buds', 'flower', 'cum', 'berries']

Caption: Durgin Park 2009 Photograph of the entrance to Durgin Park, one of the oldest restaurants in Boston, located near Faneuil Hall.

Query: What color is the clock located directly outside of Durgin Park in Boston?

MARVEL nearest token: ['brewery', 'the', 'Lake', 'bou', 'stau', 'restaurant', 'restaur', 'bakery', 'următoarele', 'residents', 'store', 'dairy', 'Eröffnung', 'Historical', 'situated']

CIEA nearest token: ['brewery', 'bou', 'stau', 'restaurant', 'clock', 'bakery', 'Uhr', 'locaux', 'operates', 'locuri', 'yolk', 'store', 'outlets', 'see', 'museum']

Caption: Saxifraga bryoides (habitus) Mossy Saxifrage near the Aletsch Glacier, Switzerland

Query: How many flowers grow on each flower stem of the mossy saxifrage?

MARVEL nearest token: ['bacteria', 'the', 'behalf', 'berries', 'bacter', 'omato', 'cum', 'indigenous', 'rum']

CIEA nearest token: ['cum', 'flowers', 'Flowers', 'alb', 'behalf', 'berries', 'timp', 'individuals', 'young', 'Blue', 'concerns', 'omato', 'buds', 'um', 'acum', 'rum']

Caption: Portrait de la marquise d'Orvilliers, Jacques-Louis David, 1790 Portrait of one of two sisters of the Genevan banker Jacques Rilliet

Query: Did the woman in the Elisabeth Vigée-Lebrun Self-Portrait and the woman in Portrait de la marquise d'Orvilliers both wear a black dress with a red ribbon?

MARVEL nearest token: ['dominant', 'the', 'behalf', 'often', 'Mary', 'acum', 'Deborah', 'ographer', 'altar', 'mother', 'Francis', 'actress', 'Sfant']

CIEA nearest token: ['connaissance', 'dress', 'behalf', 'expressive', '$75', 'Dorothy', 'Deborah', 'ographer', 'mother', 'altar', 'Francis', 'acum', 'actress', 'Sfant']

Figure 3: Case studies. The nearest token represents the embeddings in the vocabulary that are closest to the visual embeddings, with duplicates removed. The words in red represent terms related to the image that are not found in MARVEL within CIEA.

Acknowledgement

This work is supported by Key-Area Research and Development Program of Guangdong Province (No.2024B1111100001).

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15338–15347.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16495–16504.

Chang Che, Qunwei Lin, Xinyu Zhao, Jiaxin Huang, and Liqiang Yu. 2023. Enhancing multimodal understanding with clip-based image-to-text transformation. In *Proceedings of the 2023 6th International Conference on Big Data Technologies*, pages 414–418.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2023. Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36:43780–43799.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al.

2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2025. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742.

Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan Kankanhalli. 2024a. Improving context understanding in multimodal large language models via multimodal composition learning. In *International Conference on Machine Learning*, pages 27732–27751.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024b. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773.

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

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2023b. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In *International Conference on Learning Representations*.

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

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *Advances in Neural Information Processing Systems*, 36:43447–43478.

Anas Mahmoud, Mostafa Elhoushi, Amro Abbas, Yu Yang, Newsha Ardalani, Hugh Leather, and Ari S Morcos. 2024. Sieve: Multimodal dataset pruning using image captioning models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22423–22432.

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*, pages 8748–8763.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 8364–8377.

Xi'ao Su, Ran Wang, and Xinyu Dai. 2022. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 672–679.

Michael Taylor, John Guiver, Stephen Robertson, and Tom Minka. 2008. Softrank: optimizing non-smooth rank metrics. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 77–86.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, et al. 2024a. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022b. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*.

Ziyang Wang, Heba Elfardy, Markus Dreyer, Kevin Small, and Mohit Bansal. 2024b. Unified embeddings for multimodal retrieval via frozen llms. In *Findings of the Association for Computational Linguistics*, pages 1537–1547.

Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. 2023. Cap4video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10704–10713.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*.

Shi Yu, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2023. Openmatch-v2: An all-in-one multimodality plm-based information retrieval toolkit. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3160–3164.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. 2023. Retrieving multimodal information for augmented generation: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4736–4756.

Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. 2024. Marvel: Unlocking the multi-modal capability of dense retrieval via visual module plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 14608–14624.

Xiangrong Zhu, Yuexiang Xie, Yi Liu, Yaliang Li, and Wei Hu. 2025. Knowledge graph-guided retrieval augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, pages 8912–8924.

A Datasets

Modality	Documents	Queries		
		Train	Dev	Test
Image	389,750	16,400	2,554	2,511
Text	787,697	15,366	2,446	2,455
Total	1,177,447	31,766	5,000	4,966

Table 6: Statistics of the adopted WebQA.

WebQA is an open-domain question answering dataset, where each query is associated with one or more multimodal documents to assist in generating responses. WebQA can be divided into three partitions: train, dev, and test, with data statistics as shown in Table 6. The retrieval corpus contains 389,750 image documents with visual information and captions, as well as 787,697 plain text documents. Our partitioned WebQA-Image focuses solely on image documents for testing the effectiveness of CIEA in multimodal modeling. In contrast, WebQA-Multi retrieves from a total of 1,177,447 documents to assess whether the model can maintain an accurate representation of text information.

WebQA is used by many baselines, such as MARVEL (Zhou et al., 2024). ClueWeb22-MM is also utilized in MARVEL. We have applied for this dataset from Carnegie Mellon University; unfortunately, our request was denied due to export control restrictions. Therefore, we also use EDIS to evaluate our methods. EDIS (Liu et al., 2023a) is a comprehensive dataset consisting of 1 million image-text pairs sourced from Google. The dataset includes a training set of 26,000 pairs, accompanied by validation and test sets, each containing 3,200 pairs. With a high entity count of 4.03, EDIS reflects a diverse range of semantic content, making it a valuable resource for research in image-text retrieval tasks.

B Implementation Details

In our experiment, we use T5-ANCE (Yu et al., 2023) as the base language model and utilize CLIP as the visual understanding module to implement our CIEA model, enhancing the image understanding capability of T5-ANCE. For easier comparison, we initialize our projector using the pre-trained projector from MARVEL. The visual encoder is initialized using the clip-vit-base-patch32 checkpoint from OpenAI[2]. We truncate the text input length to 128 tokens.

During the training process, we use the AdamW (Loshchilov and Hutter, 2019) optimizer, set the maximum training epochs to 40, with a batch size of 64, a learning rate of 5e-6, and the temperature hyperparameter τ set to 0.01. We follow the setup of UniVL-DR (Loshchilov and Hutter, 2019) by training with the ANCE sampling method. Starting from the CIEA-DPR model, fine-tuned with negative examples from within the batch, we continue to train CIEA-DPR to achieve a balanced modality with difficult negative examples. For negative sampling in evaluation, we shuffle the training set and select other samples within the same batch as negative examples in DPR, while in the ANCE approach, we utilize the top 100 samples with the highest similarity retrieval by CIEA-DPR as hard negatives. All models are evaluated every 500 steps, with early stopping set at 5 steps. For the parameter λ, we perform a grid search and select the parameter that yielded the lowest loss on the validation set, setting it to 0.0011 for WebQA-Multi, 0.019 for WebQA-Image, and 0.001 for EDIS. The results of various baselines for the WebQA-Multi dataset are provided in the MARVEL's paper(Zhou et al., 2024), while the experimental results for other baselines are reproduced using the open-source code from the original papers. All experiments are conducted on a single NVIDIA A100.

C Statistical significance

To avoid potential concerns that the results are justified, we repeat the experiments with different random seeds and report the average values and standard deviation values in Table 8. The results demonstrate the proposed CIEA achieves significant and consistent improvements compared to MARVEL. Besides, different modules make positive contributions to the overall performance of ICEA.

D The Effects of Language Models

The project-based approach relies on language models as the backbone, making the performance of the language model one of the key factors influencing retrieval effectiveness. To validate whether our method can function effectively with different language models as the backbone, we con-

[2] https://huggingface.co/sentence-transformers/clip-ViT-B-32

Language model	setting	mrr@10	ndcg@10	mrr@20	ndcg@20	rec@20	rec@100
T5-ANCE	Text only	52.67	45.69	53.08	48.48	67.70	82.07
	Project-based	56.91	54.26	57.31	56.88	77.20	89.26
	CIEA	**57.49**	**54.80**	**57.86**	**57.24**	**77.21**	**89.41**
BERT	Text only	39.50	36.23	40.06	38.69	55.37	72.72
	Project-based	52.65	49.41	53.21	51.97	71.16	85.40
	CIEA	**53.40**	**49.96**	**53.83**	**62.56**	**71.95**	**85.72**
BART	Text only	24.52	21.79	25.04	23.98	36.91	54.66
	Project-based	50.46	47.72	50.94	50.04	69.40	83.69
	CIEA	**50.97**	**48.06**	**51.43**	**50.72**	**70.44**	**84.81**
GPT2	Text only	20.38	18.62	20.86	20.19	29.97	44.33
	Project-based	**45.94**	42.45	46.46	45.08	63.16	78.34
	CIEA	45.82	**42.60**	**46.51**	**45.13**	**63.46**	**78.91**
GPT2-LARGE	Text only	37.25	34.52	37.67	37.11	54.37	73.08
	Project-based	51.44	48.98	51.89	51.64	72.06	86.61
	CIEA	**54.28**	**51.81**	**54.68**	**54.39**	**74.94**	**88.52**

Table 7: Results of different language models on DPR training framework.

Model	MRR@10	NDCG@10	REC@100
CIEA	**66.03±0.012**	**63.70±0.030**	**92.63±0.008**
w/o image query	65.66±0.008	63.44±0.072	92.54±0.003
w/o attention	65.90±0.098	63.39±0.003	92.60±0.001
MARVEL	65.30±0.017	63.19±0.005	92.40±0.012

Table 8: Standard Deviation Values on WebQa-multi.

duct experiments using T5-ANCE (Yu et al., 2023), BART (Lewis et al., 2020a), BERT (Devlin et al., 2019), GPT-2, and GPT-2-LARGE (Radford et al., 2019), encompassing encoder-decoder, encoder-only, and decoder-only architectures. For the encoder-decoder model, we follow MARVEL's approach(Zhou et al., 2024) and encode the multimodal document input into the encoder, while inputting a '\s' character into the decoder segment to use its final hidden layer representation as the document representation. For the encoder-only model, we use the representation of the first character, and for the decoder-only model, we use the final hidden layer representation of the last token, as each token only computes attention with preceding tokens. Due to the large number of parameters in GPT-2 LARGE, we set its batch size to 16 to avoid memory issues, while setting the batch size to 64 for the others. Besides the results trained with the ANCE negative sampling method presented in Table 3, we also include the results trained using the in-batch negative sampling method of DPR in Table 7. It is evident that while ANCE consistently outperforms DPR, CIEA surpasses both simple project-based and text-only methods under both negative sampling approaches, demonstrating that the proposed method aligns better with image information.

E Failure Case Analysis

We provide a failure case analysis in Table 9 for a better understanding of the proposed method and for promoting further research. Each case presents the ground-truth image-caption pair alongside the top three retrieved candidates from CIEA, with all instances classified as retrieval failures due to the absence of ground-truth matches among the top results. The first two cases illustrate the model's partial success in capturing color attributes (e.g., *"orange"* and *"green"*), while it might struggle to establish precise color-object mappings (e.g., *"ceiling"* and *"tiles"*), especially within multi-object environments. The third case reveals more fundamental limitations regarding action interpretation, as the retrieved results exhibit minimal relevance to the target action *"relax"*. These failures highlight persistent challenges in visual-semantic alignment, particularly in the contextual binding of object attributes and the dynamic interpretation of action representations.

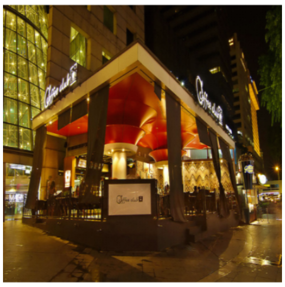
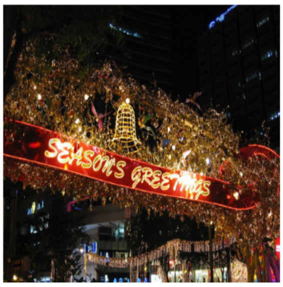
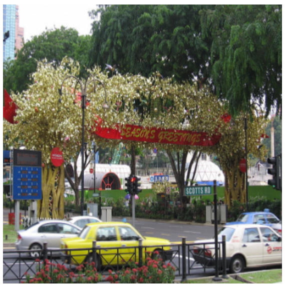
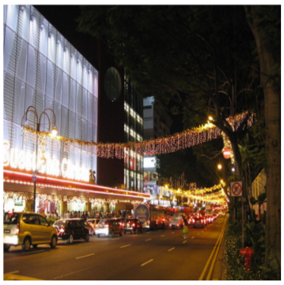
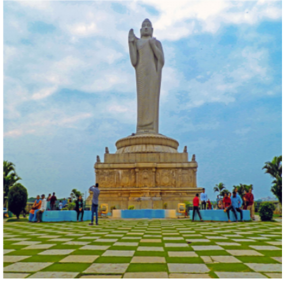
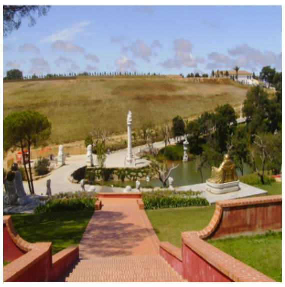
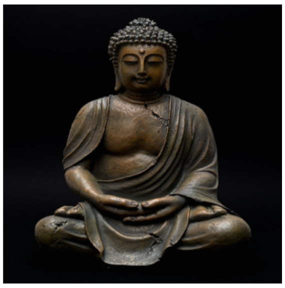
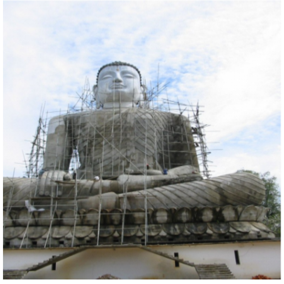
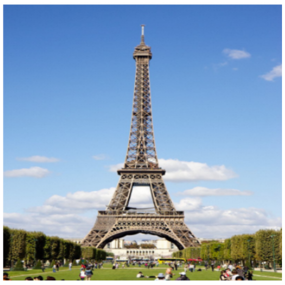
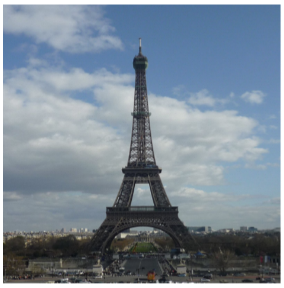
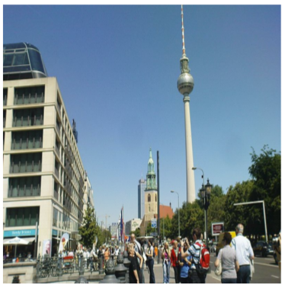
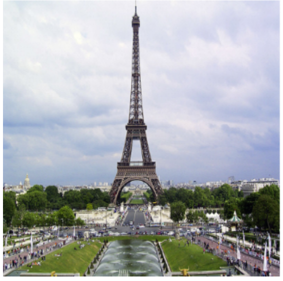
Question: Which restaurant on Orchard Road in Singapore has a tiered orange ceiling?			
Golden	**Retrieval (Top 3)**		
			
Coffee Club, Orchard Road (8171776629) Coffee Club at Wheelock Place, Singapore.	Orchard Road 12, Xmas, Dec 06 Orchard Road, Christmas Light-up 2006, Singapore.	Orchard Road, Xmas, Dec 06 Orchard Road, Christmas Light-up 2006, Singapore.	Orchard Road 15, Xmas, Dec 06 Orchard Road, Christmas Light-up 2006, Singapore.
Question: Which Buddha statue has green tiles in front of it?			
Golden	**Retrieval (Top 3)**		
			
Buddha statue no 1 Buddha statue at Hussain sagar	Buddha eden (34)	Buddha 1251876	Building the Buddha – panoramio
Question: Where can people relax in the grass next to the Eifel Tower?			
Golden	**Retrieval (Top 3)**		
			
Eiffel Tower, Paris 17 September 2010.	FW Eiffelturm.	Berlin , Mitte , Alexanderplatz , fernsehturm - panoramio.	Eiffelturm.

Table 9: Failure Case Analysis.