MARVEL: Unlocking the Multi-Modal Capability of Dense Retrieval via Visual Module Plugin

Tianshuo Zhou[1*], Sen Mei[1*], Xinze Li[1], Zhenghao Liu[1†], Chenyan Xiong[2], Zhiyuan Liu[3], Yu Gu[1], Ge Yu[1]

[1]Department of Computer Science and Technology, Northeastern University, China
[2]Language Technologies Institute, Carnegie Mellon University, United States
[3]Department of Computer Science and Technology, Institute for AI, Tsinghua University, China
Beijing National Research Center for Information Science and Technology, China

Abstract

This paper proposes **M**ulti-mod**A**l **R**etrieval model via **V**isual modul**E** p**L**ugin (MARVEL), which learns an embedding space for queries and multi-modal documents to conduct retrieval. MARVEL encodes queries and multi-modal documents with a unified encoder model, which helps to alleviate the modality gap between images and texts. Specifically, we enable the image understanding ability of the well-trained dense retriever, T5-ANCE, by incorporating the visual module's encoded image features as its inputs. To facilitate the multi-modal retrieval tasks, we build the ClueWeb22-MM dataset based on the ClueWeb22 dataset, which regards anchor texts as queries, and extracts the related text and image documents from anchor-linked web pages. Our experiments show that MARVEL significantly outperforms the state-of-the-art methods on the multi-modal retrieval dataset WebQA and ClueWeb22-MM. MARVEL provides an opportunity to broaden the advantages of text retrieval to the multi-modal scenario. Besides, we also illustrate that the language model has the ability to extract image semantics and partly map the image features to the input word embedding space. All codes are available at https://github.com/OpenMatch/MARVEL.

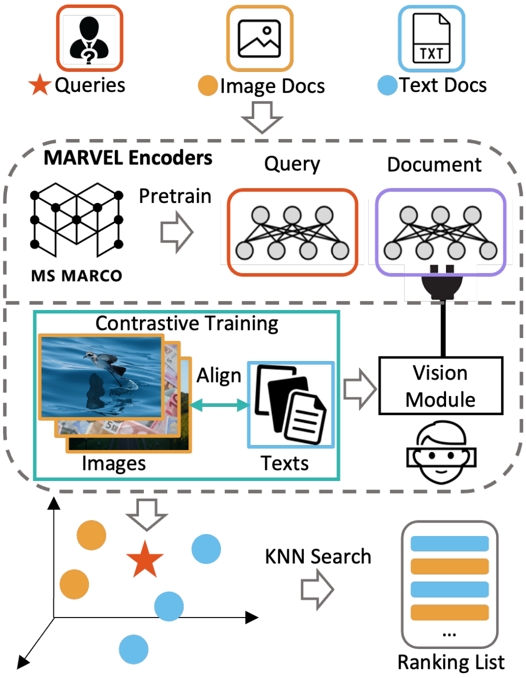


Figure 1: Retrieval Pipeline with Our MARVEL Model. MARVEL incorporates the visual module plugin, aiming to unlock the multi-modal capabilities of well trained dense retrieval model.

1 Introduction

With the growth of multimedia information on the Internet, search engines tend to return multi-modal retrieval results to better satisfy the user information need (Tautkute et al., 2019; Zhu et al., 2023). The media information provides more vivid retrieval results, such as texts, images, videos, and more, which improves users' experiences and even changes their browsing behaviors.

Multi-modal retrieval (Bain et al., 2021; Awad et al., 2021; Arni et al., 2008; Chang et al., 2022) aims to return fusion results of images and texts to answer user questions. The task can be modeled using a divide-and-conquer pipeline (Chang et al., 2022; Liu et al., 2023b) or universal dense retrieval (Liu et al., 2023b). UniVL-DR (Liu et al., 2023b) encodes queries and multi-modal documents into a universal embedding space for multi-modal retrieval. However, this work encodes image features and texts using different encoders from CLIP (Radford et al., 2021) and the separated text and image encoding leads to a modality gap in representing multi-modal documents. It makes UniVL-DR design an additional image verbalization method to alleviate the modality gap and also limits the text retrieval models (Karpukhin et al., 2020; Xiong et al., 2021a; Zhan et al., 2021; Li

分而治之

语言化

[*]indicates equal contribution.
[†]indicates corresponding author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14608–14624

et al., 2021b; Yu et al., 2021) to excel their advantages in multi-modal scenarios.

In this paper, we propose **M**ulti-mod**A**l **R**etrieval model via **V**isual modul**E** p**L**ugin (MARVEL). As shown in Figure 1, MARVEL is based on the text retriever T5-ANCE (Yu et al., 2023), regards the visual module as a plugin and pretrains the visual module with image-caption contrastive training for adaption. By incorporating a visual module into well-trained text retriever T5-ANCE, MARVEL seizes the opportunity to extend the benefits of unimodal learning to the multi-modal retrieval task.

通过加入一个 visual module，就可以将文本检索任务拓展到多模态检索当中

To facilitate the multi-modal retrieval task, we build a large-scale benchmark, ClueWeb22-MM, based on the web page dataset, ClueWeb22 (Overwijk et al., 2022). Following previous work in text retrieval (Zhang et al., 2020; Xie et al., 2023), we regard the anchor text as a query and assume that its linked web page is related to the query. Subsequently, we extract image and text documents from these anchor-linked web pages. After processing, the ClueWeb22-MM encompasses over 90k queries, maintaining a scale comparable to existing benchmark WebQA (Chang et al., 2022). Previous work (Xie et al., 2023) demonstrates that the high-quality training signals from anchor-document pairs contribute to developing a state-of-the-art dense retrieval model.

Our experiments show that MARVEL outperforms all baseline models, achieving improvements of over 2% and 7%, in the main metric MRR, on WebQA (Chang et al., 2022) and ClueWeb22-MM, respectively. The evaluation results indicate the effectiveness of MARVEL comes from the visual module plugin architecture, the visual module pretraining method, and the text matching knowledge learned by T5-ANCE. Our further analyses illustrate that the image representations encoded by the visual module can be easily captured by only finetuning the language model parameters. The training strategies guide the language model to assign more appropriate attention weights to image and text features, preventing the visual module from overfitting to the training signals. These encoded image representations not only inhabit the input embedding space for semantics alignment but also function as a kind of prompt.

2 Related Work

Existing dense retrieval models (Karpukhin et al., 2020; Xiong et al., 2021a; Ren et al., 2021; Xiong et al., 2021b; Gao and Callan, 2022; Luan et al., 2021; Khattab and Zaharia, 2020) usually focus on retrieving text documents and modeling the relevance between queries and documents. They usually employ pretrained language models to encode queries and text documents into an embedding space, followed by a KNN search to retrieve candidate documents.

Unlike the text retrieval task, the multi-modal retrieval task (Chang et al., 2022; Hannan et al., 2020; Singh et al., 2021; Talmor et al., 2021) aims to provide users with multi-modal documents that satisfy their information needs. Earlier work primarily focuses on building a divide-and-conquer pipeline for multi-modal retrieval (Chang et al., 2022; Liu et al., 2023b; Escalante et al., 2008; Grubinger et al., 2008). In these models, retrievers individually search candidates from the document collections of different modalities and then use a reranking model to fuse the retrieval results, such as vision-language models (Zhang et al., 2021). However, this approach usually struggles to fuse the retrieval results across different modalities (Liu et al., 2023b). UniVL-DR (Liu et al., 2023b) builds a universal multi-modal dense retrieval model. It encodes queries and multi-modal documents as embeddings and conducts retrieval, modality routing, and result fusion within a unified embedding space.

Representing images is also the core of multi-modal retrieval, aiming to alleviate the modality gap between images and texts. Existing work usually focuses on representing the images using captions and image features (Liu et al., 2023b) with different encoding methods. BERT-style pretrained visual-language models (Chen et al., 2019; Lu et al., 2019; Tan and Bansal, 2019; Su et al., 2020; Li et al., 2019, 2021a; Cho et al., 2021; Hu et al., 2020; Wang et al., 2022) provide an opportunity to model the captions and image features using the same model. However, these visual-language models typically aim to align the semantics between image features and captions instead of learning representations for image documents. Thus they show less effectiveness in learning an embedding space for multi-modal retrieval (Liu et al., 2023b).

学习的是图像和标题之间的关系，而不是图像的表征

思考，那么我们如何学习图像的特征呢? 回到最初的 Restnet 之类的模型

Another way to facilitate the image document representations is using the visual-language models that focus on representation learning, such as CLIP (Radford et al., 2021). It encodes image features and texts using different encoders. However, these approaches often only provide shallow interactions between texts and visual features. Thus, existing models (Liu et al., 2023b) pay more attention to alleviating the modality gap between texts and images by the image verbalization method, aiming to bridge the modality gap between images and texts in the raw text space.

Recent advancements in multi-modal large language models (Brown et al., 2020; Touvron et al., 2023) have introduced a novel approach to modeling multi-modality features. This approach incorporates a visual encoder module into large language models through a transformation layer (Li et al., 2023; Alayrac et al., 2022; Liu et al., 2023a). These models extract image features using the visual encoder module of CLIP and then optimize the prompt tokens and transformation layer to map the encoded image embeddings to the raw input space of large language models (Merullo et al., 2023; Lester et al., 2021). Such a visual encoder plugin method presents a unified modeling approach for handling image and text features. It not only enables the visual comprehension ability of large language models but also preserves their effectiveness by freezing their parameters.

3 Multi-Modal Retrieval Model via Visual Module Plugin (MARVEL)

In this section, we first describe the multi-modal retrieval (Sec. 3.1) and then introduce the model architecture of MARVEL (Sec. 3.2).

3.1 Preliminary of Multi-Modal Retrieval

Given a query q, the retrieval task requires the dense retrieval models to search relevant documents from the document collection $\mathcal{D}$ to meet the information needs of users.

Previous dense retrieval models (Karpukhin et al., 2020; Xiong et al., 2021a; Gao and Callan, 2021; Yu et al., 2023) usually focus on the text retrieval task, which aims to model the relevance between user query q and text documents $\mathcal{D} = \{d^1_{\text{Text}}, ..., d^m_{\text{Text}}\}$. They encode both query and the i-th document d^i_{Text} using language models, such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and T5 (Raffel et al., 2020):

$$\vec{q} = \text{TextEncoder}(q); \vec{d}^i_{\text{Text}} = \text{TextEncoder}(d^i_{\text{Text}}). \quad (1)$$

Different from text retrieval (Nguyen et al., 2016; Thakur et al., 2021), the multi-modal retrieval task (Chang et al., 2022) aims to return a fusion result of documents from the collection $\mathcal{D}$, which are from different modalities. The document collection $\mathcal{D}$ not only contains texts $\mathcal{T} = \{d^1_{\text{Text}}, ..., d^m_{\text{Text}}\}$, but also includes images $\mathcal{I} = \{d^1_{\text{Image}}, ..., d^n_{\text{Image}}\}$.

The multi-modal retrieval task requires retrievers to conduct relevance modeling, cross-modal matching, and modality fusion (Liu et al., 2023b). Previous work (Liu et al., 2023b) maps text and image documents in an embedding space for retrieval, encodes texts and images using different encoders, and tries to bridge the modality gap using image verbalization methods. However, this limits the capability of dense retrieval models, hindering the expansion of text matching knowledge for learning representations for multi-modal documents.

3.2 Universal Multi-Modal Encoding

We show the model architecture in Figure 2. Different from previous work (Liu et al., 2023b), we can universally encode query q and multi-modal documents $\mathcal{D} = \{d^1_{\text{Text}}, ..., d^m_{\text{Text}}, d^1_{\text{Image}}, ..., d^n_{\text{Image}}\}$ using one encoder, T5-ANCE-CLIP:

$$\begin{aligned} \vec{q} &= \text{T5-ANCE-CLIP}(q); \\ \vec{d}^i_{\text{Text}} &= \text{T5-ANCE-CLIP}(d^i_{\text{Text}}); \\ \vec{d}^i_{\text{Image}} &= \text{T5-ANCE-CLIP}(d^i_{\text{Image}}(I), d^i_{\text{Image}}(C)), \end{aligned} \quad (2)$$

where $d^i_{\text{Image}}(I)$ and $d^i_{\text{Image}}(C)$ are the image feature and caption of the i-th image document d^i_{Image}.

Then we calculate the relevance score $f(q, d^i)$ between query q and the i-th document d^i using cosine similarity:

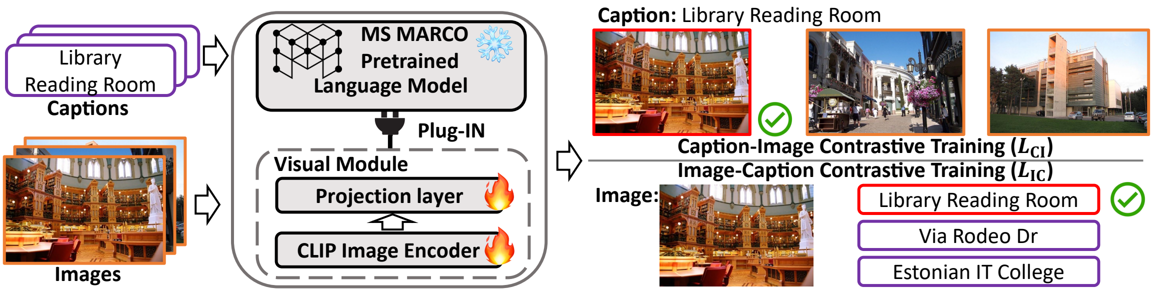
$$f(q, d^i) = \cos(\vec{q}, \vec{d}^i). \quad (3)$$

Following this, we conduct KNN search (Johnson et al., 2019) to retrieve multi-modal document candidates for the given query q.

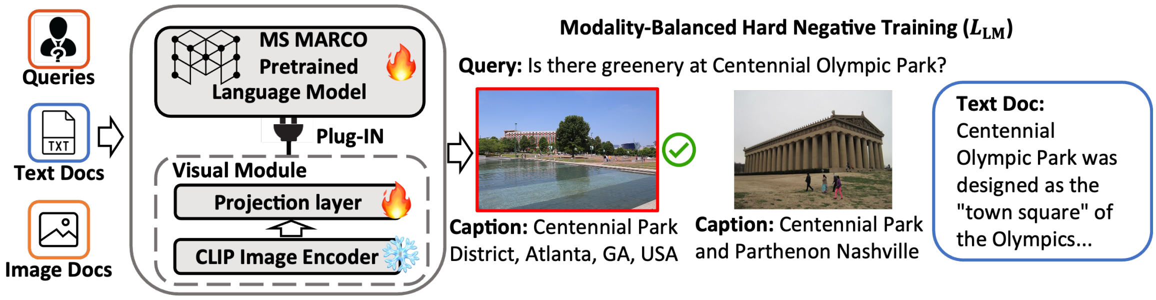
Subsequently, we first introduce the visual module plugin architecture of our MARVEL model (Sec. 3.2.1). Then we adapt the visual module to T5-ANCE by pretraining the visual understanding module (Sec. 3.2.2). Finally, we finetune the parameters of T5-ANCE to learn an embedding space for multi-modal retrieval (Sec. 3.2.3).

3.2.1 Dense Retrieval with Visual Plugin

MARVEL starts from the T5-ANCE model (Yu et al., 2023), which is a dense retrieval model that is well-trained using MS MARCO dataset (Nguyen et al., 2016). Then we enable T5-ANCE by incorporating the visual module from the vision-language model, CLIP (Radford et al., 2021), and conduct the T5-ANCE-CLIP model. We can use a universal



(a) Visual Module Adaption Pretraining.



(b) Modality-Balanced Language Model Finetuning. We follow previous work (Liu et al., 2023b) and sample one image document and one text document from corresponding negative document collections.

Figure 2: The Architecture of **M**ulti-mod**A**l **R**etrieval model via **V**isual modul**E** p**L**ugin (MARVEL). We first pretrain the visual modules using the image-caption alignment task (Figure 2(a)) and then finetune the language model to conduct multi-modal retrieval (Figure 2(b)).

encoder, T5-ANCE-CLIP, to encode texts, image features, and image documents.

Specifically, we encode the image feature I using the visual encoder of CLIP (Radford et al., 2021) and get its encoded visual representation $\vec{h}^I$:

$$\vec{h}^I = \text{CLIP}(I), \tag{4}$$

This representation is obtained from the grid features of the last layer of the visual encoder of CLIP, and $\vec{h}^I = \{\vec{h}^I_1, ..., \vec{h}^I_{49}\}$. Here 49 is the number of patches. Then we follow the previous visual-language model (Merullo et al., 2023) and use a linear transformation layer to adapt the visual features $\vec{h}^I_i$ into the embedding space of the inputs of dense retrieval model:

$$\vec{I}_i = \text{Linear}(\vec{h}^I_i). \tag{5}$$

Finally, we can feed these encoded image features $\vec{I} = \{\vec{I}_1, ..., \vec{I}_{49}\}$ as the ahead input embeddings $\vec{X}$ for T5-ANCE:

$$\vec{X} = \vec{e}(\text{<start>}); \vec{I}_1; ...; \vec{I}_{49}; \vec{e}(\text{<end>}); \vec{e}_1; ...; \vec{e}_k, \tag{6}$$

where ; is the concatenation operation. $\vec{e}(\text{<start>})$ and $\vec{e}(\text{<end>})$ are the embeddings of prompt tokens to denote the start and end of encoded image feature representations. $\{\vec{e}_1 ...; \vec{e}_k\}$ are the word embeddings of the text input sequence $T = \{T_1, ..., T_k\}$.

Different from these visual-language models (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023a; Tsimpoukelli et al., 2021), our MARVEL model aims to bring the advance of text retrieval-based pretraining to multi-modal retrieval tasks by using the visual model plugin to bridge the modality gap between images and texts.

3.2.2 Visual Module Adaption Pretraining

In MARVEL, we adapt the visual understanding module to T5-ANCE by only pretraining the parameters of the visual module (Eq. 4) and the projection layer (Eq. 5). We follow Radford et al. (2021) and leverage the image-caption contrastive training loss L_{VM} to pretrain the visual understanding module. The training loss utilizes the alignment between image features I and captions C:

$$L_{\text{VM}} = L_{\text{IC}} + L_{\text{CI}}, \tag{7}$$

where L_{IC} and L_{CI} are the dual direction training losses to regard image and caption as queries and then map them with corresponding caption and image, respectively:

$$L_{\text{IC}} = -\log \frac{e^{f(I,C^+)/\tau}}{e^{f(I,C^+)/\tau} + \sum_{C^- \in \mathcal{D}^-_C} e^{f(I,C^-)/\tau}}, \tag{8}$$

$$L_{\text{CI}} = -\log \frac{e^{f(C,I^+)/\tau}}{e^{f(C,I^+)/\tau} + \sum_{I^- \in \mathcal{D}^-_I} e^{f(C,I^-)/\tau}}, \tag{9}$$

Dataset	Modality	#Doc	#Query Train	Dev	Test
WebQA	Image	389,750	16,400	2,554	2,511
	Text	787,697	15,366	2,446	2,455
	Multi-Modal	1,177,447	31,766	5,000	4,966
ClueWeb22-MM	Image	368,710	35,873	5,041	5,030
	Text	363,508	36,155	4,959	4,970
	Multi-Modal	732,218	72,028	10,000	10,000

Table 1: Data Statistics.

where τ is the temperature used to scale the similarity score. $\mathcal{D}_C^-$ and $\mathcal{D}_I^-$ contain negative captions and negative images respectively, which are sampled from in-batch negatives.

3.2.3 Modality-Balanced Language Model Finetuning

During finetuning, we can freeze the parameters of the visual module (Eq. 4) and optimize other parameters of MARVEL. To enable the MARVEL model to learn a universal embedding space for both queries and multi-modal documents, we follow previous work (Liu et al., 2023b) and employ modality-balanced hard negative training to alleviate the modality discrimination of retrieval models:

$$L_{\text{LM}} = -\log \frac{e^{f(q,d^+)/\tau}}{e^{f(q,d^+)/\tau} + \sum_{d^- \in \mathcal{D}^-} e^{f(q,d^-)/\tau}}$$
$$\propto -\underbrace{f(q,d^+)/\tau}_{L_{\text{Align}}} + \log(\sum_{d^- \in \mathcal{D}^-} (\underbrace{e^{f(q,d^-_{\text{Image}})/\tau}}_{L_{\text{Image}}} + \underbrace{e^{f(q,d^-_{\text{Text}})/\tau}}_{L_{\text{Text}}})), \tag{10}$$

where $\mathcal{D}^-$ contains the same number of negative documents of image and text. L_{Align} teaches models to align the query with related documents. L_{Text} and L_{Image} guide retrievers to choose the modality and make the embedding space uniform (Liu et al., 2023b; Wang and Isola, 2020; Chen et al., 2020).

4 Experimental Methodology

This section describes datasets, evaluation metrics, baselines and implementation details.

Dataset. During pretraining, we collect the image-caption pairs from ClueWeb22 (Overwijk et al., 2022) to train the visual understanding module. More details of pretraining data are shown in Appendix A.2. Then two multi-modal retrieval datasets, WebQA and ClueWeb22-MM, are used for finetuning and evaluation. The data statistics are shown in Table 1.

WebQA is a multi-hop, multi-modal, open-domain question answering benchmark (Chang et al., 2022). The dataset contains images and passage snippets that are crawled from the general Web and Wikipedia. We follow previous work (Liu et al., 2023b) to keep the same experimental settings to preprocess the dataset. Besides, we build a new multi-modal retrieval dataset, ClueWeb22-MM, based on ClueWeb22 (Overwijk et al., 2022), which provides 10 billion web pages with rich information. We only retain web pages in English and build the ClueWeb22-MM dataset. We establish query-document relations by pairing anchors with their corresponding document (Xie et al., 2023; Zhang et al., 2020). And then we regard the anchor texts as queries and extract image documents and text documents from the linked documents. More details of building the ClueWeb22-MM dataset are shown in Appendix A.4.

Evaluation Metrics. We use NDCG@10, MRR@10 and Recall@100 as evaluation metrics. Following previous work (Liu et al., 2023b; Nguyen et al., 2016), we regard MRR@10 as our main evaluation. MRR and NDCG are computed using the official scripts[1]. Statistic significances are tested by permutation test with P<0.05.

Baselines. In our experiments, we follow previous work (Liu et al., 2023b) to conduct baseline models and divide them into three groups: single modality retrieval, divide-and-conquer, and universal dense retrieval models.

Single Modality Retrieval. In our experiments, we represent image documents using captions and use several text retrieval models as baselines. BM25 (Robertson et al., 2009) is widely used in text retrieval work, which conducts exact matches between queries and documents. DPR (Karpukhin et al., 2020) is trained using NQ dataset (Kwiatkowski et al., 2019) and uses a dual-encoder to encode queries and documents as dense vectors for retrieval. We start from vanilla BERT (Devlin et al., 2019) and DPR (Karpukhin et al., 2020) checkpoints and train the encoders using in-batch negatives to conduct BERT-DPR and NQ-DPR models. NQ-ANCE is also compared, which continuously trains NQ-DPR using hard negatives (Xiong et al., 2021a). Besides, T5-ANCE (Yu et al., 2023) and Anchor-DR (Xie et al., 2023) are dense retrieval models that are trained on MS MARCO and ClueWeb22, respectively.

Divide-and-Conquer. The divide-and-conquer models retrieve image documents and text documents individually and then fuse the retrieval re-

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

sults. Following previous work (Liu et al., 2023b), we use single modality retrievers, VinVL-DPR, CLIP-DPR and BM25, and fuse the retrieval results according to their unimodal rank reciprocals.

Universal Dense Retrieval. CLIP-DPR and VinVL-DPR employ the visual language models CLIP (Radford et al., 2021) and VinVL (Zhang et al., 2021) as image and text encoders and then are trained with in-batch negatives. UniVL-DR (Liu et al., 2023b) is our main baseline model, which further uses modality-balanced hard negative to train text and image encoders and also utilizes the image verbalization method to bridge the modality gap between images and texts.

Implementation Details. In our experiments, we use T5-ANCE (Yu et al., 2023) as our backbone language model, which is well-trained on the MS MARCO dataset (Nguyen et al., 2016). Then we implement our MARVEL model by utilizing CLIP as the visual understanding module to empower the image understanding capability of T5-ANCE. The visual encoder is initialized with the clip-vit-base-patch32 checkpoint from OpenAI[2]. For MARVEL, we truncate queries, text documents and image captions to 128 tokens and set the max number of visual tokens to 49.

During training, we use AdamW (Loshchilov and Hutter, 2019) optimizer and set maximum training epoch=20, batch size=64, learning rate=$5e-6$, and the temperature hyperparameter $\tau = 0.01$. We follow UniVL-DR (Liu et al., 2023b) and conduct MARVEL-ANCE by starting from in-batch negative finetuned MARVEL-DPR, and continuously training MARVEL-DPR with modality-balanced hard negatives. These hard negatives are randomly sampled from the top 100 retrieved negatives using MARVEL-DPR. All models are evaluated per 500 steps and the early stop step is set to 5.

5 Evaluation Result

In this section, we first evaluate the performance of MARVEL and conduct ablation studies. Then, we explore the effectiveness of different visual and language model fusion methods and analyze the role of visual module adaption pretraining in MARVEL. Some case studies are shown in Appendix A.8.

5.1 Overall Performance

The multi-modal retrieval performance of MARVEL and baseline models is shown in Table 2. Besides retrieval performance, we also compared retrieval efficiency in Appendix A.7.

Overall, MARVEL significantly outperforms baseline models on all datasets by achieving more than 2% improvements on both datasets, demonstrating its advantages in multi-modal retrieval tasks. Compared with text retrieval models, MARVEL improves their performance, showing that the image features are crucial in the multi-modal retrieval task. Furthermore, these universal multi-modal dense retrievers, UniVL-DR and MARVEL, outperform divide-and-conquer models by alleviating the modality fusion problem (Liu et al., 2023b). Compared with our main baseline UniVL-DR, MARVEL encodes queries and multi-modal documents using a universal encoder. Experimental results show that MARVEL significantly improves the retrieval effectiveness of UniVL-DR on both datasets, demonstrating its effectiveness in bridging the modality gap between images and texts.

5.2 Ablation Study

As shown in Table 3, we conduct ablation studies to explore the role of different modules of MARVEL in multi-modal retrieval. More ablation studies are shown in Appendix A.5.

In the comparison between MARVEL and MARVEL w/o CLIP Pretraining, pretraining the visual understanding module shows its effectiveness by improving the performance on single/multi-modal retrieval tasks. It shows that the image-caption alignment relations provide some opportunities to adapt the visual module to the language model via pretraining. Subsequently, MARVEL also outperforms MARVEL w/o MS MARCO Pretraining, especially on the text retrieval task. It demonstrates that MARVEL can broaden the advantage of text relevance modeling to the multi-modal retrieval task. To unify the multi-modal encoding, MARVEL follows previous work (Hannan et al., 2020) uses prompt tokens to indicate the start and end positions of encoded image features (Eq. 6), aiming to distinguish the image features from text token embeddings. These image prompt tokens bring light improvements, illustrating their roles in multi-modal document representation.

5.3 Retrieval Effectiveness of Different Visual-Language Fusion Methods

In this experiment, we show the retrieval effectiveness of MARVEL on the WebQA dataset by using different modality fusion and finetuning methods.

[2] https://github.com/openai/CLIP

Setting	Model	WebQA			ClueWeb22-MM		
		MRR@10	NDCG@10	Rec@100	MRR@10	NDCG@10	Rec@100
Single Modality (Text Only)	BM25	53.75	49.60	80.69	40.81	46.08	78.22
	DPR (Zero-Shot)	22.72	20.06	45.43	20.59	23.24	44.93
	CLIP-Text (Zero-Shot)	18.16	16.76	39.83	30.13	33.91	59.53
	Anchor-DR (Zero-Shot)	39.96	37.09	71.32	42.92	48.50	76.52
	T5-ANCE (Zero-Shot)	41.57	37.92	69.33	45.65	51.71	83.23
	BERT-DPR	42.16	39.57	77.10	38.56	44.41	80.38
	NQ-DPR	41.88	39.65	78.57	39.86	46.15	83.50
	NQ-ANCE	45.54	42.05	69.31	45.89	51.83	81.21
Divide-Conquer	VinVL-DPR	22.11	22.92	62.82	29.97	36.13	74.56
	CLIP-DPR	37.35	37.56	85.53	39.54	47.16	87.25
	BM25 & CLIP-DPR	42.27	41.58	87.50	41.58	48.67	83.50
UnivSearch	CLIP (Zero-Shot)	10.59	8.69	20.21	16.28	18.52	40.36
	VinVL-DPR	38.14	35.43	69.42	35.09	40.36	75.06
	CLIP-DPR	48.83	46.32	86.43	42.59	49.24	87.07
	UniVL-DR	$\underline{62.40}^{\dagger\S}$	$\underline{59.32}^{\dagger\S}$	$\underline{89.42}^{\dagger\S}$	$\underline{47.99}^{\dagger\S}$	$\underline{55.41}^{\dagger\S}$	$\underline{90.46}^{\dagger\S}$
	MARVEL-DPR	$55.71^{\dagger}$	$52.94^{\dagger}$	$88.23^{\dagger}$	$46.93^{\dagger}$	$53.76^{\dagger}$	$88.74^{\dagger}$
	MARVEL-ANCE	$\mathbf{65.15}^{\dagger\ddagger\S}$	$\mathbf{62.95}^{\dagger\ddagger\S}$	$\mathbf{92.40}^{\dagger\ddagger\S}$	$\mathbf{55.19}^{\dagger\ddagger\S}$	$\mathbf{62.83}^{\dagger\ddagger\S}$	$\mathbf{93.16}^{\dagger\ddagger\S}$

Table 2: Overall Performance. We keep the same experimental settings with previous work (Liu et al., 2023b). †, ‡ and § indicate statistically significant improvements over CLIP-DPR†, UniVL-DR‡ and MARVEL-DPR§.

Model	Modality	WebQA			ClueWeb22-MM		
		MRR@10	NDCG@10	Rec@100	MRR@10	NDCG@10	Rec@100
MARVEL-ANCE	Text	$64.72^{\ddagger}$	$58.88^{\ddagger\S}$	$90.26^{\ddagger\S}$	$71.73^{\dagger\ddagger\S}$	$75.40^{\dagger\ddagger\S}$	$92.29^{\ddagger\S}$
	Image	$66.12^{\dagger}$	$67.49^{\dagger\ddagger}$	$95.12^{\dagger\ddagger\S}$	$77.57^{\dagger\ddagger\S}$	$81.34^{\dagger\ddagger\S}$	$96.50^{\dagger\ddagger}$
	Multi	$65.15^{\ddagger}$	$62.95^{\dagger\ddagger}$	$92.40^{\ddagger\S}$	$55.19^{\ddagger\S}$	$62.83^{\ddagger\S}$	$93.16^{\ddagger\S}$
w/o CLIP Pretraining	Text	$64.63^{\ddagger}$	$58.79^{\ddagger}$	$90.21^{\ddagger\S}$	$70.92^{\S}$	$74.67^{\S}$	$92.13^{\ddagger\S}$
	Image	65.17	66.69	94.64	$76.99^{\ddagger\S}$	$80.83^{\ddagger\S}$	96.22
	Multi	64.66	$62.50^{\ddagger}$	$92.24^{\ddagger\S}$	$55.18^{\ddagger\S}$	$62.81^{\ddagger\S}$	$93.07^{\ddagger}$
w/o MS MARCO Pretraining	Text	63.37	56.93	88.54	$70.74^{\S}$	$74.35^{\S}$	91.27
	Image	65.73	66.91	94.66	76.26	80.11	96.08
	Multi	64.21	61.63	91.43	$54.61^{\S}$	$62.16^{\S}$	92.52
w/o Prompt	Text	63.86	$58.00^{\ddagger}$	$89.60^{\ddagger}$	69.99	73.82	91.65
	Image	$66.53^{\dagger\ddagger}$	$67.56^{\dagger\ddagger}$	94.42	76.07	80.14	$96.58^{\dagger\ddagger}$
	Multi	$64.92^{\ddagger}$	$62.50^{\ddagger}$	$91.81^{\ddagger}$	54.20	61.79	$92.93^{\ddagger}$

Table 3: Ablation Studies. †, ‡, and § indicate statistically significant improvements over MARVEL-ANCE w/o CLIP Pretraining†, MARVEL-ANCE w/o MS MARCO Pretraining‡ and MARVEL-ANCE w/o Prompt§.

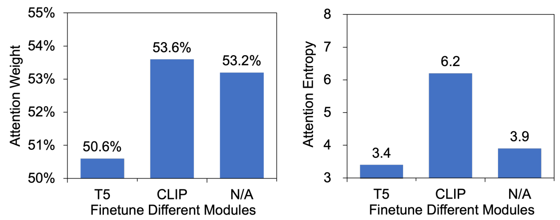
Modality Fusion. Three kinds of visual-language fusion methods are compared in our experiments, including Sum, Concat and Plugin. For Sum and Concat methods, we encode the captions and image features separately as embeddings, then sum or concatenate these embeddings, followed by joint training of T5-ANCE and CLIP models with in-batch negatives. We show the experimental results in Table 4. MARVEL's visual module plugin method outperforms other fusion methods. This highlights the effectiveness of utilizing pretrained attention heads of language models for extracting image semantics and fostering deeper interactions between image and text inputs. Our plugin method proves instrumental in mitigating the modality gap between texts and images, enabling MARVEL to better represent image documents by jointly modeling image captions and features.

Different from Liu et al. (2023b), we use T5-ANCE and CLIP as the text and image encoders, respectively. These models have different architectures and are pretrained on text retrieval and image-caption matching tasks. The multi-modal retrieval performance of CLIP-Sum decreases when we encode the image caption with a stronger retrieval model T5-ANCE (T5-CLIP-Sum) instead of CLIP. It demonstrates that incorporating an additional visual module into a well-trained dense retrieval model is still challenging for multi-modal retrieval. Notably, MARVEL provides a promising way to enable the image understanding ability of dense retrieval models by using the visual module plugin modeling method.

Finetuning Strategies. We then show the effectiveness of different finetuning strategies. In this experiment, we individually finetune the language model (T5) and visual module (CLIP) to show the changes of attention distributions and analyze the

Method	Modality	MRR@10	Rec@100
CLIP-Sum	Text	51.75	84.37
	Image	60.61	94.84
	Multi	48.83	86.43
T5-CLIP (Sum)	Text	51.84	85.06
	Image	58.09	93.13
	Multi	35.03	79.00
T5-CLIP (Concat)	Text	48.71	81.78
	Image	37.20	81.14
	Multi	25.19	62.77
T5-CLIP (Plugin)	Text	54.28	85.80
	Image	60.81	93.55
	Multi	55.58	88.50

Table 4: Retrieval Performance of the Models using Different Visual-Language Fusion Methods. T5-CLIP (Sum/Concat) is similar to previous work (Liu et al., 2023b), which only replace the image caption encoder with T5-ANCE. The CLIP-Sum model is the CLIP-DPR model from previous work (Liu et al., 2023b). All models are trained with in-batch negatives. MRR@10 is used to evaluate the retrieval effectiveness of all models.



(a) Attention Weights.

(b) Attention Entropy.

Figure 3: Attention Distribution of MARVEL-ANCE. The attention weights of image features are shown in Figure 3(a). And the attention weight entropy of image captions and features is shown in Figure 3(b).

behaviors of different finetuning strategies.

As shown in Figure 3. The attention scores are calculated by cross attentions from the decoder to the encoder module of T5. We first show the attention weight distribution of image features in Figure 3(a). When we only finetune the language model, the attention heads tend to allocate more balanced attention weights between image features and captions, helping to adapt the visual module in the language model. On the other hand, the image features win more attention weights when the CLIP module is finetuned. However, as shown in Figure 3(b), only finetuning the CLIP module shows a scattered attention weight mechanism than other models, which misleads the T5-ANCE to capture more important information from encoded representations of documents. All these phenomena demonstrate the necessity of the training strategies of MARVEL, which pretrain visual module for adaption and only finetune the language model for multi-modal retrieval. In addition, we show the retrieval effectiveness with different finetuning methods in Appendix A.3.

5.4 Effectiveness of Visual Module Adaption Pretraining

We show some cases in Table 5 to show the effectiveness of the visual understanding module by verbalizing encoded image features. Besides, more experiments about the encoded image features are shown in Appendix A.6.

We randomly select four image documents of different topics and represent the encoded image feature with some tokens to verbalize the image semantics. Specifically, we first use the visual plugin modules of MARVEL and MARVEL w/o CLIP Pretraining to encode the image features. Then, to show the semantics of the encoded image features, we utilize cosine similarity to find the k-NN tokens for each encoded image feature. Finally, the tokens with the highest score are used to represent the semantics of the encoded image features.

For the first two examples, MARVEL w/o CLIP Pretraing learns more similar representations for both image documents. The related word tokens of these image documents contain lots of same tokens, such as “7,000”, “Hi”, and “RAM”, which are unrelated to the semantics of the image documents. On the contrary, our MARVEL model can learn more similar semantics to both image documents. Specifically, MARVEL verbalizes the first image document using the words “brightness”, “resident” and “store”, which are related to the image description of “Bourbon Street”. Additionally, MARVEL captures the semantics of “animals”, “wildlife” and “creatures” in the second image document. The next two instances show the effectiveness of MARVEL in learning more fine-grained semantics of the image documents by verbalizing the image documents with more related words, such as “militari”, “vehicle”, “flag”, “legislatur”, and “government”. Overall, compared with the MARVEL w/o CLIP Pretraining model, MARVEL has the ability to learn more effective representations that are closer to the semantics of the images, demonstrating the important role of MARVEL’s visual module pretraining strategy in adapting the visual understanding module for dense retrievals.

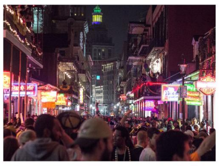
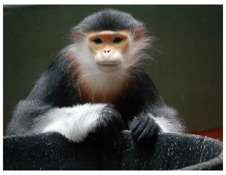
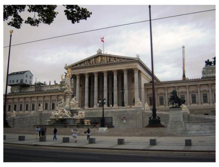
Figure	Text
	**Manual Caption:** Mardi Gras Bourbon Street 2015 Bourbon Street, New Orleans, during Mardi Gras Nearest Tokens: [“kehr”,“voted”, “brightness”, “event”, “city”, “venue”, “local”, “pub”, “bounce”, “island”, “ferry”, “keto”, “Ice”, “residents”, “lighting”, “store”, “lights”, “banks”, “Lake”, “impacted”, “lively”, “drinks”, “eye”] Nearest Tokens w/o CLIP Pretraining: [“OUG”, “7,000”, “ban”, “CU”, “edited”, “ition”, “Pop”, “imprisonment”, “O”, “militari”, “immuno”, “Ton”, “reset”, “États”, “concise”, “Arbeits”, “IN”, “Hi”, “RAM”, “Hello”, “stocked”, “charged”, “institu”]
	**Manual Caption:** Red-shanked Douc at the Philadelphia Zoo Nearest Tokens: [“Whale”, “endangered”, “horn”, “bird”, “goat”, “animals”, “wildlife”, “mammals”, “whale”, “Gib”, “Elephant”, “Savannah”, “dach”, “birds”, “creatures”, “Wildlife”, “lois”, “biomass”, “limb”, “Creatures”] Nearest Tokens w/o CLIP Pretraining: [“RAM”, “bilingual”, “MOD”, “native”, “recognizable”, “Graphic”, “charged”, “ordentlich”, “gray”, “suffisamment”, “colorful”, “clar”, “haunt”, “riad”, “CM”, “ammunition”, “ordre”, “thetic”, “Hi”, “auftrag”, “he”, “ban”, “sets”, “7,000”, “representation”]
	**Manual Caption:** Military parade in Baku on an Army Day28 Military parade in Baku on an Army Day Nearest tokens: [“vehicles”, “Fahrzeug”, “territories”, “flag”, “chemical”, “replies”, “migrants”, “parliament”, “bikes”, “militari”, “equipment”, “République”, “troops”, “clothing”, “gear”, “prisoners”, “machinery”, “tribe”, “vorgesehen”] Nearest Tokens w/o CLIP Pretraining: [“împreună”, “RAM”, “ened”, “troupe”, “Compet”, “sie”, “own”, “RGB”, “Ha”,“operation”, “arbeit”, “enforcement”, “Cor”, “EU”, “LCD”, “countries”, “SO”, “institu”, “grief”, “limbi”, “default”, “16”, “raum”, “haunt”, “unanimous”]
	**Manual Caption:**Parlament Wien Austria, Vienna, Austrian Parliament Building Nearest Tokens: [“Schloss”, “funds”, “Statut”, “furniture”, “Albany”, “structure”, “palace”, “Capitol”, “statute”, “locul”, “headquarters”, “occupie”, “structures”, “legislature”, “cinéma”, “legislation”, “governmental”, “Argentin”,“vederea”] Nearest Tokens w/o CLIP Pretraining: [“militari”, “reset”, “Ton”, “shrine”, “commands”, “hi”, “împreună”, “Achtung”, “genug”, “shake”, “RAM”, “iconic”, “committed”, “département”, “colo”, “Hi”, “Sammlung”, “pop”, “1951”, “ban”, “russia”, “Color”, “vivid”, “HM”, “arbeit”, “default”]

Table 5: The Nearest Tokens of Image Features. We randomly select five image documents, encode these image features using the visual module of MARVEL and MARVEL w/o CLIP Pretraining, and then show the nearest tokens of the encoded image features. The tokens related to the semantics of the image document are highlighted.

6 Conclusion

This paper proposes **M**ulti-mod**A**l **R**etrieval via **V**isual modul**E** p**L**ugin (MARVEL). MARVEL integrates a visual plugin module with a well-trained dense retriever and pretrains the visual module with image-caption contrastive training for adaption. Our MARVEL model achieves state-of-the-art on all benchmarks by unifying the multi-modal document encoding and alleviating the modality gap between images and texts.

Limitations

Even though MARVEL shows strong effectiveness in the multi-modal retrieval task, there are some limitations in our work. Existing multi-modal retrieval systems still highly depend on the semantics of image caption instead of the image understanding ability of the visual module. In this case, MARVEL pretrains the visual understanding module but achieves limited improvements. Building an effective visual understanding module is crucial for the multi-modal retrieval task.

Acknowledgments

This work is partly supported by the Natural Science Foundation of China under Grant (No. 62206042, No. U23B2019, No. 62137001, and No. 62272093), the Joint Funds of Natural Science Foundation of Liaoning Province (No. 2023-MSBA-081), and the Fundamental Research Funds for the Central Universities under Grant (No. N2416012).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of NeurIPS*, pages 23716–23736.

T Arni, M Sanderson, P Clough, and M Grubinger. 2008. Overview of the imageclef 2008 photographic retrieval task. *Working Notes of the CLEF*.

George Awad, Asad A Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, et al. 2021. Trecvid 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of ICCV*, pages 1728–1738.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of NeurIPS*.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the CVPR*, pages 16495–16504.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*, pages 1597–1607.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of ICML*, pages 1931–1942.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Hugo Jair Escalante, Carlos A Hérnadez, Luis Enrique Sucar, and Manuel Montes. 2008. Late fusion of heterogeneous methods for multimedia image retrieval. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 172–179.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. In *Proceedings of EMNLP*, pages 981–993.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of ACL*, pages 2843–2853.

M Grubinger, P Clough, A Hanbury, and H Müller. 2008. Overview of the imageclef 2008 photographic retrieval task. In *Working Notes of the 2008 CLEF Workshop. Aarhus, Denmark*.

Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. Manymodalqa: Modality disambiguation and QA over diverse inputs. In *Proceedings of AAAI*, pages 7879–7886.

Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. 2020. Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

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 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 SIGIR*, pages 39–48.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Proceedings of TACL*, pages 452–466.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of EMNLP*, pages 3045–3059.

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 *Proceedings of ICML*, pages 19730–19742.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021a. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of ACL*, pages 2592–2607.

Yizhi Li, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2021b. More robust dense retrieval with contrastive dual learning. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 287–296.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*.

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 *Proceedings of ICLR*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of ICLR*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of NeurIPS*, pages 13–23.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Proceedings of TACL*, pages 329–345.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. 2023. Linearly mapping from image to text space. In *Proceedings of ICLR*.

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. In *Proceesings of CoCo@NIPS*.

Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. Clueweb22: 10 billion web documents with rich information. In *Proceedings of SIGIR*, pages 3360–3362.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, pages 8748–8763.

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. *The Journal of Machine Learning Research*, pages 140:1–140:67.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of EMNLP*, pages 2825–2835.

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

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs.

Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. 2021. MIMOQA: Multimodal input multimodal output question answering. In *Proceedings of NAACL-HLT*, pages 5317–5332.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *Proceedings of ICLR*.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: complex question answering over text, tables and images. In *Proceedings of ICLR*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of EMNLP*, pages 5100–5111.

Ivona Tautkute, Tomasz Trzciński, Aleksander P Skorupa, Łukasz Brocki, and Krzysztof Marasek. 2019. Deepstyle: Multimodal search engine for fashion and interior design. *IEEE Access*, pages 84613–84628.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the 35th Conference on Neural Information Processing Systems, Datasets, and Benchmarks Track*.

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.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Proceedings of NeurIPS*, pages 200–212.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of ICML*, pages 9929–9939.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks.

Yiqing Xie, Xiao Liu, and Chenyan Xiong. 2023. Unsupervised dense retrieval training with web anchors. In *Proceedings of SIGIR*, pages 2476–2480.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021a. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of ICLR*.

Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick S. H. Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. 2021b. Answering complex open-domain questions with multi-hop dense retrieval. In *Proceedings of ICLR*.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. In *Proceedings of SIGIR*, pages 829–838.

Shi Yu, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2023. Openmatch-v2: An all-in-one multi-modality plm-based information retrieval toolkit. In *Proceedings of SIGIR*, pages 3160–3164.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of SIGIR*, pages 1503–1512.

Kaitao Zhang, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2020. Selective weak supervision for neural information retrieval. In *Proceedings of WWW*, pages 474–485.

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 CVPR*, pages 5579–5588.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey.

Finetune	Modality	MRR@10	NDCG@10	Rec@100
CLIP & T5	Text	64.89	58.71	89.98
	Image	64.36	65.41	94.19
	Multi	64.37	61.77	91.78
T5	Text	64.72	58.88	90.26
	Image	66.12	67.49	95.12
	Multi	65.15	62.95	92.40
CLIP	Text	48.38	41.63	75.11
	Image	56.28	56.17	87.67
	Multi	49.22	45.80	80.42
N/A	Text	48.38	41.39	74.57
	Image	55.09	54.99	87.26
	Multi	48.12	44.75	79.69

Table 6: The Retrieval Performance with Different Training Strategies. We freeze each module of MARVEL-ANCE to explore the benefits of training between different modules.

Data Type	Median	Average	Max	Min
Queries	8.0	9.9	245.0	1.0
Text Documents	52.0	127.8	1121183.0	1.0
Image Captions	6.0	8.1	998.0	1.0

Table 7: Length Statistics of Queries, Text Documents and Image Captions in ClueWeb22-MM Dataset.

Range of Image Sizes	Number
Height or Width $\geq$ 1024	23.8k
Height and Width $\geq$ 1024	7.4k
Height or Width $\geq$ 512	81.9k
Height and Width $\geq$ 512	43.9k
Height or Width $\geq$ 256	234.6k
Height and Width $\geq$ 256	170.2k

Table 8: Image Size Distribution of ClueWeb22-MM.

A Appendix

A.1 License

We show the licenses of the datasets that we use. WebQA uses CC0-1.0 license, while ClueWeb22 shows its terms of use at website[3]. All these licenses and agreements permit the use of their data for academic purposes.

A.2 Experimental Details of MARVEL Pretraining Data

In this subsection, we introduce the experimental details to process the pretraining data.

To pretrain the visual module in MARVEL, we collect the image-caption pairs from the ClueWeb22 dataset. We retain the English pages, extract the content within the image tag and use the image and alt-text to construct the image-caption pair. To ensure the quality of the pretraining dataset, following LAION-400M (Schuhmann et al., 2021), we use CLIP to calculate the embeddings of images and captions and compute the cosine similarity between the two embeddings. Subsequently, we discard all samples with a cosine similarity lower than 0.3. The pretraining dataset contains 1.6M image-caption pairs, and we randomly select 10,000 pieces of data as the development set and use the rest for the pretraining visual module.

A.3 Retrieval Effectiveness of Different Finetuning Strategies

In this experiment, we show the performance of single/cross and multi-modal retrieval tasks with different finetuning strategies.

[3] https://lemurproject.org/clueweb22/

As shown in Table 6, finetuning the CLIP module indeed improves the retrieval performance of the whole frozen model, especially in the image retrieval task. This observation shows that multi-modal training signals are effective to benefit the capability of visual modules. When we only tune the parameters of T5, MARVEL-ANCE achieves significant improvements over the frozen model, showing the language model's strong ability to adapt the visual module to the dense retriever. Nevertheless, the fully finetuned model decreases the retrieval performance of MARVEL-ANCE that only finetunes T5. It shows the necessity of the pretraining-and-then-finetuning strategy of MARVEL, which pretrains the visual understanding module for adaption and finetunes the language model for multi-modal retrieval.

A.4 More Details of ClueWeb22-MM

To show the details of our ClueWeb22-MM dataset, we show the data collection, data processing, and data statistics in this subsection.

Data Collection. Following previous work in text retrieval (Zhang et al., 2020; Xie et al., 2023), we regard the anchor text as a query and assume that its linked web page is related to the query. Then we extract image documents and text documents from these anchor-linked web pages. To obtain image documents, we parse HTML to extract the content within the image tag, then use alt-text as image caption, and crawl the image features from the image URL.

Data Processing. Ensuring the quality and meaningfulness of the ClueWeb22-MM dataset, we conduct additional processing on the data to filter out noise data according to the quality of images and alt-texts. Concerning images, we retain data

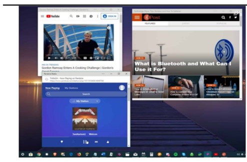
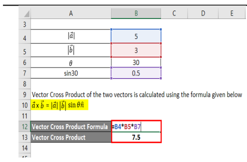
Queries with the Text Document as Label	
Query: Chinese Dragons — Facts, Culture, Origins, and Art **Text Document:** Live updates on China travel restrictions for 2022. Learn more Home Chinese Culture Traditional Chinese Clothes Chinese Dragons — Facts, Culture, Origins, and Art Written by Mike Ho Updated Dec. 14, 2021 Chinese dragons are powerful and benevolent symbols in Chinese culture, with supposed control over watery phenomenon, e.g. summoning rain during a drought. Dragons are everywhere in China — in legends, festivals, astrology, art, names, and idioms.	
Query: How to manage partitions with the Disk Management tool, in Windows I Digital Citizen **Text Document:** Disk Management A new window should pop up, listing the drive letter of the partition. Click or tap Change and, in the next window, select the new drive letter you wish to assign to it. Then, click or tap OK.	
Query: here's a small-batch peanut butter oatmeal cookie recipe for you **Text Document:** You are here: Home / Recipes / Small-batch Peanut Butter Oatmeal Cookies Small-batch Peanut Butter Oatmeal Cookies 02/21/19 I Cookies , Desserts , Recipes , Small-batch Dessert These Small-batch Peanut Butter Oatmeal Cookies are the perfect cookie hybrid. They're rich and peanut buttery, bendy and chewy, and the best of both worlds. A few weeks ago, I posted these (AMAZING) Peanut Butter Oatmeal Cookies . It was a big-batch recipe meant for sharing and freezing, so I promised that I'd add a small-batch version ASAP for those of you who are here for small-batch desserts. So here we go. Let's make a cute little batch of Peanut Butter Oatmeal Cookies and share with no one.	
Query: What foods increase uric acid **Text Document:** Vegetables and legumes that increase uric acid Legumes such as lentils, chickpeas or beans are known for their purine content, so their intake should be limited to only once or twice a week if you have high uric acid. Other vegetables that should be eaten in moderation are asparagus, mushrooms, cauliflower, spinach, radishes and leeks... Other foods that increase uric acid Other foods that increases uric acid and should be avoided are: All kinds of alcoholic beverages , especially beer and wine. Carbonated beverages, sugar-laden soft drinks and packaged juices. Avoid cooking with brewer's yeast...	
Queries with the Image Document as Label	
	**Query:** Use Web apps With the New Chromium Edge on Windows 10 **Image Caption:** Web Apps Running Chromium Edge
	**Query:** What are Runestones In Witcher 3? **Image Caption:** Witcher 3 best runewords
	**Query:** Everything We Know About Mindy Kaling and BJ Novak's Relationship—Including Sweet Details from Her Book **Image Caption:** mindy-kaling-bj-novak-removebg
	**Query:** Vector Cross Product Formula Excel Template **Image Caption:** Vector Cross Product Formula-1.2

Table 9: Examples of ClueWeb22-MM. We give practical examples of queries, image documents, and text documents.

with image file extensions such as jpg/png/jpeg and discard samples with image URLs containing keywords, *e.g.* "logo", "button", "icon", "plugin", or "widget". Besides, we exclude the example, which has empty alt-text, has "no alt attribute" and contains an alt-text that is shorter than 5.

To further guarantee the quality of the dataset, we use T5-ANCE (Yu et al., 2023) to estimate the relevance between the anchor and its corresponding image document. We encode all captions of image documents using T5-ANCE, use the anchor texts as queries to retrieve the images and reserve the anchors that are ranked in the top 10. Finally, we respectively sample 10,000 queries to build the development and test set. The rest data are used for finetuning models, which contain 72,028 queries.

Data Statistics. We provide length statistics on queries, text documents, and image captions in Table 7 and present the image size distribution in Table 8. Subsequently, as shown in Table 9, we show eight examples to illuminate the ClueWeb22-MM dataset. These examples show that the anchor-

Model	Modality	WebQA			ClueWeb22-MM		
		MRR@10	NDCG@10	Rec@100	MRR@10	NDCG@10	Rec@100
MARVEL-ANCE	Text	64.72	58.88	90.26	71.73	75.40	92.29
	Image	66.12	67.49	95.12	77.57	81.34	96.50
	Multi	65.15	62.95	92.40	55.19	62.83	93.16
w/o Image Caption	Text	64.67	58.30	89.49	69.75	73.32	90.60
	Image	3.85	4.32	24.81	18.26	20.65	45.07
	Multi	33.70	30.83	56.45	37.29	40.74	64.26
w/o Image Feature	Text	63.42	57.95	90.27	71.17	74.78	91.57
	Image	64.32	65.42	94.15	76.83	80.60	95.88
	Multi	63.60	61.43	91.99	54.98	62.64	92.60

Table 10: Additional Ablation Study Results on MARVEL-ANCE.

Model	MRR@10	NDCG@10	Rec@100
MARVEL-DPR	55.71	52.94	88.23
w/ 1-NN Token	38.80	35.89	73.59
w/ 5-NN Tokens	42.39	39.27	75.04
w/ Random Token	37.73	35.34	71.92
MARVEL-ANCE	65.15	62.95	92.40
w/ 1-NN Token	51.37	48.27	80.47
w/ 5-NN Tokens	52.22	49.35	81.88
w/ Random Token	44.22	41.55	71.23

Table 11: Multi-Modal Retrieval Performance of Different Image Feature Replacement Strategies. We conduct experiments on MARVAL-DPR and MARVAL-ANCE models by replacing the image features with the average of k-NN (k Nearest Neighbour) word embeddings. The k is set to 1 and 5.

document pairs are of high quality. Thus we can use them to build an evaluation benchmark for multi-modal retrieval.

A.5 Additional Ablation Studies on MARVEL

We conduct additional ablation studies to explore the role of image captions and image features in the multi-modal retrieval task.

As shown in Table 10, the relevance modeling between queries and image documents heavily depends on the image caption, which is also observed in previous work (Liu et al., 2023b). The image features contribute to approximately 1% improvements in the image retrieval task, demonstrating the effectiveness of image features in helping the model better understand the image documents.

A.6 Learned Semantics of Image Features

In this experiment, we explore the semantic information of image features encoded by the visual module on the WebQA dataset. During training MARVEL model, we map the encoded image features into the input space of T5-ANCE's word embeddings. We conduct several experiments by replacing the encoded image features with the embeddings of the nearest or random tokens.

As shown in Table 11, replacing encoded image features with k-NN token embeddings generally outperforms the retrieval model using randomly selected token embeddings. It demonstrates that the visual plugin module effectively maps image semantics in the input space of the language model, and the ability to keep growing with more token embeddings (5-NN). However, the retrieval performance significantly drops when employing k-NN token embeddings to replace the image features, compared to the MARVEL model. It demonstrates the role of encoded image features beyond the semantic representations of word embeddings. The encoded image features may act as a kind of prompt, encouraging language models to capture image semantics (Merullo et al., 2023).

A.7 Retrieval Efficiency of MARVEL

In this section, we compare the retrieval efficiency of MARVEL with other baselines on the same device, as shown in Table 12.

MARVEL follows the general dense retrieval framework (Karpukhin et al., 2020) for efficient document retrieval. It encodes the entire corpus offline and constructs the document index using FAISS (Johnson et al., 2019) for online searches. While the offline encoding time for queries and image/text documents in MARVEL is longer than that in UniVL and other baseline models, this encoding process does not impact retrieval efficiency.

When comparing the retrieval latency of these models in retrieving the top 100 relevant documents for each query, MARVEL's retrieval time is comparable to its base model, T5-ANCE, which has a retrieval time of 48.3 ms. Furthermore, the retrieval time for document encoding using the CLIP model is less than that of models such as T5 and BERT. This demonstrates that the retrieval time is influenced by the dimensionality of the embeddings, and MARVEL's architecture does not introduce any additional retrieval latency.

Setting	Model	Encoding Time (ms)			Retrieval Time (ms)
		Query	Img Doc	Text Doc	Query
Single Modality (Text Only)	BM25	-	0.1	0.1	41.9
	DPR (Zero-Shot)	1.8	2.7	4.7	48.4
	CLIP-Text (Zero-Shot)	0.4	0.2	0.3	32.7
	Anchor-DR (Zero-Shot)	2.2	2.0	2.0	47.7
	T5-ANCE (Zero-Shot)	2.1	2.0	2.0	48.3
	BERT-DPR	1.1	2.7	3.1	47.6
	NQ-DPR	1.3	2.8	4.9	47.8
	NQ-ANCE	1.3	2.7	4.7	48.1
Divide-Conquer	VinVL-DPR	1.4	5.2	4.7	48.4
	CLIP-DPR	0.4	0.8	0.3	32.8
	BM25 & CLIP-DPR	0.4	0.8	0.1	37.3
UnivSearch	CLIP	0.4	0.8	0.3	32.8
	VinVL-DPR	1.3	5.3	4.7	47.6
	UniVL-DR	0.4	0.9	0.3	32.9
	MARVEL	2.1	3.7	2.0	48.0

Table 12: Retrieval Efficiency. We compare the encoding and retrieval times of different architectural models on the same device. These models encode the multi-model documents and queries offline and construct the FlatL2 index using FAISS (Johnson et al., 2019) for online retrieval.

A.8 Case Studies

In Figure 4, we show two cases from WebQA and ClueWeb22-MM to study the multi-modal retrieval effectiveness of MARVEL. The top 5 documents retrieved by UniVL-DR, MARVEL-DPR, and MARVEL-ANCE are presented.

For the first case, UniVL-DR conducts shallow keyword matching and returns text documents that are related to “animal” and “Peace” mentioned in the query, which are unrelated to the query. In contrast, MARVEL can better understand that “Peace and Plenty” is a famous painting and retrieve more related images and text documents for users. In the second case, UniVL-DR, MARVEL-DPR, and MARVEL-ANCE all return documents related to “promotion ideas”. Notable, MARVEL can better understand the user’s query and return the expected modality. MARVEL-ANCE introduces a variety of sales promotion strategies rather than matching on “promotion” keywords. It shows the effectiveness of MARVEL in better fusing the retrieval results from different modalities, which thrives on universal multi-modal document encoding.

Query: What are two animals that can be found in "Peace and Plenty"?

UniVL-DR Retrieval Top5 Documents:

1st: A forest is a piece of land with many trees. Many animals need forests to live and survive. Forests are very important and grow in many places around the world. They are an ecosystem which includes many plants and animals. Temperature and rainfall are the two most important things for forests. Many places are too cold..

2nd: Animals include mice, rats that eat mostly seeds. Snakes, hawks and eagles eat mostly small animals. Big animals, mostly grazers, include zebras, ostriches, giraffes, rhinos, elephants, horses, and antelopes . ↑ Chapman, Kim Alan, Adelheide Fischer, and Mary Ziegenhagen. 1998.

3rd: The roughly 270 land-based even-toed ungulate species include pigs, peccaries, hippopotamuses, antelopes, deer, giraffes, camels, llamas, alpacas, sheep, goats, and cattle. Many of these are of great dietary, economic, and cultural importance to humans.

4th: Owls hunt mostly small mammals, insects, and other birds, although a few species specialize in hunting fish. They are found in all regions of the Earth except the polar ice caps and some remote islands. Owls are divided into two families: the true (or typical) owl family, Strigidae, and the barn-owl family, Tytonidae.

5th: A two-headed snake, with one head being where its tail would normally be. It is venomous and, if chopped in half, its two parts can reunite. A kangaroo -like animal with a flat, human-like face and a very long tail. A small, flat animal with pure white fur and bright red claws and teeth. Its head is feline...

MARVEL-DPR Retrieval Top5 Documents:

1st: Tolstoy's War and Peace and Chekhov's Peasants both feature scenes in which wolves are hunted with hounds and Borzois. The musical Peter and the Wolf involves a wolf being captured for eating a duck, but is spared and sent to a zoo.

2nd: 

3rd: Both paintings show humans and animals interacting together... In the case of \"Peaceable Kingdom\", there are settlers in the background, signing a treaty with the Native Americans.

4th: Images of exploitations are shown during the film. According to the primatologist Jane Goodall, \"Peaceable Kingdom is a piece of art\"The documentary's producers, have created a website called

5th: But you must not change one thing, one pebble, one grain of sand, until you know what good and evil will follow on that act. The world is in balance, in Equilibrium.

MARVEL-ANCE Retrieval Top5 Documents:

1st: 

2nd: #Argent (Silver or white) – Peace and sincerity Azure (Blue) – Truth and loyalty Furs (Ermine, ermines, erminois, vair, counter vair, pean, potent, counter potent) – Dignity Gules (Red) – Warrior or martyr...

3rd: 

4th: 

5th: Spot: The protagonist of the series. A yellow puppy with a brown spot on each side of his body and a brown tip on his tail, Spot is full of curiosity...

(a) Top5 Multi-modal Documents Retrieved from WebQA.

Query: 15 Insanely Effective Sales Promotion Ideas to Win More Customers

UniVL-DR Retrieval Top5 Documents:

1st: 

2nd: Iconic Sales Booster helps you implement a variety of sales techniques to boost the revenue of your ecommerce store. Some of the tactics that it gives you access to are: A "Frequently Bought Together" box on the single product page (like Amazon)…

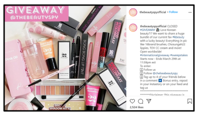
3rd: Another great example of social proof is a testimonial from your current user base - this is a genuine way to share real product experiences with your potential customers. When visitors are unsure about your brand or product, seeing that others similar to them have put their confidence and…

4th: Source: Optinmonster Promotional pricing attracts budget-conscious customers and helps to drive more revenue into your business within a short time. The rationale behind this pricing strategy is that any losses incurred would be recovered due to increased sales and new loyal customers…

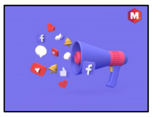
5th: Let's say you're in the middle of a product demo and there's been some head nodding so far, but not too many questions from your prospects. You're starting to feel that there's a little uncertainty about whether or not this is right for them, but you're not exactly sure why. Instead of pushing through your presentation…

MARVEL-DPR Retrieval Top5 Documents:

1st: Promotional products DESIGN YOUR INVITATION START YOUR PROMO ORDER

2nd: 

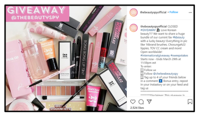
3rd: Clearly Offer a Discount or Lead Magnet Who doesn't want a great deal or something for free? Nobody! That's also why 92% of consumers are looking for…

4th: 

5th: Allow people to sell themselves As an educational marketer , you'll be able to educate people into a YES…

MARVEL-ANCE Retrieval Top5 Documents:

1st: Red Robin offered discounts to patrons who showed up for a burger with their Wolverine ticket stub in hand. While they may not have received a discount for the actual movie, the film benefitted from attracting people who were more excited about the burger… 12. Social media contests and giveaways A contest or giveaway on Facebook , Instagram, or your target audience's social platform of choice is a great way to get new customers interested in your business… 13. Shopping sprees One of the most exciting in-store sales promotions you can do is an "enter to win" contest that gives the winner a shopping spree…

2nd: 

3rd: Clearly Offer a Discount or Lead Magnet Who doesn't want a great deal or something for free? Nobody! That's also why 92% of consumers are looking for a deal when they shop and offering a discount will help convert them…

4th: It's likely that you've seen or used an exit intent pop up. These popups try to keep shoppers on the site by offering a discount. Where many brands get this wrong is by prompting it after an individual has just landed on the webpages…

5th: Create a segment of 'power users': users who purchase often or engage more frequently with your emails. Let them know they're appreciated, and encourage them to spread the word about your brand…

(b) Top5 Multi-modal Documents Retrieved from Clueweb22-MM.

Figure 4: Case Studies. We present two cases from WebQA and ClueWeb22-MM and show the top5 retrieved multi-modal documents. The ground-truth documents and related content are highlighted in red and blue respectively.