VISTA: Visualized Text Embedding For Universal Multi-Modal Retrieval

Junjie Zhou[1]*, **Zheng Liu**[2,3]†, **Shitao Xiao**[2], **Bo Zhao**[2], **Yongping Xiong**[1]
[1] State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
[2] Beijing Academy of Artificial Intelligence
[3] The Hong Kong Polytechnic University
zhoujunjie@bupt.edu.cn zhengliu1026@gmail.com stxiao@baai.ac.cn

数据合成

使用大量的弱标签数据

Abstract

Multi-modal retrieval becomes increasingly popular in practice. However, the existing retrievers are mostly text-oriented, which lack the capability to process visual information. Despite the presence of vision-language models like CLIP, the current methods are severely limited in representing the text-only and image-only data. In this work, we present a new embedding model **VISTA** for universal multi-modal retrieval. Our work brings forth threefold technical contributions. Firstly, we introduce a flexible architecture which extends a powerful text encoder with the image understanding capability by introducing visual token embeddings. Secondly, we develop two data generation strategies, which bring high-quality composed image-text to facilitate the training of the embedding model. Thirdly, we introduce a multi-stage training algorithm, which first aligns the visual token embedding with the text encoder using massive weakly labeled data, and then develops multi-modal representation capability using the generated composed image-text data. In our experiments, VISTA achieves superior performances across a variety of multi-modal retrieval tasks in both zero-shot and supervised settings. Our model, data, and source code are available at https://github.com/FlagOpen/FlagEmbedding.

1 Introduction

Information retrieval (IR) is a critical task in many real-world scenarios, e.g., search engines, open-domain question answering, and retrieval augmented generation (Karpukhin et al., 2020; Lewis et al., 2020a). It aims to find relevant data from a large database such that the downstream problems can be faithfully solved on top of proper knowledge. One important IR paradigm is dense retrieval, where the query and candidates, i.e. document, are represented as embeddings, and their semantic relationship can be reflected by the embedding similarity (Yates et al., 2021). With the continual progress on pre-trained model and training algorithm, increasingly powerful embedding models have been developed, such as DPR (Karpukhin et al., 2020), Contriever (Izacard et al., 2022), GTR (Ni et al., 2022), E5 (Wang et al., 2022), BGE (Xiao et al., 2023), etc., which substantially improves the quality and universality of dense retrieval.

Most of the existing dense retrieval models are text-oriented, which can only deal with the data presented in human language. However, a large portion of the world knowledge naturally contains both text and image, e.g., web articles with visual illustration (Chang et al., 2022); meanwhile, people's queries can also be flexibly expressed with multiple data modalities, e.g., search queries with exemplar images (Liu et al., 2021; Wu et al., 2021). Despite the development of visual-language representation models (VLM), like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), the above problem is still challenging in many perspectives. On one hand, the existing VLMs are severely limited in text representation capability, whose retrieval performance is far behind the recent text-only embedding models, like E5 and BGE. On the other hand, the existing VLMs focus more on the independent encoding of text and image; nevertheless, the joint representation of image-text data (e.g., documents with illustrations) is largely unexplored.

In this work, we propose VISualized Text embedding for universal multi-modal retrievAl, namely **VISTA**. It takes the best of the existing text encoder and image encoder where high-quality multi-modality embedding can be generated from it. In particular, our work presents the following three technical contributions.

First of all, we come up with a flexible model architecture to facilitate the generation of multi-modal embedding. It is built upon a powerful and

*Work done during Junjie's internship at BAAI.
†Corresponding author.

well-trained text encoder, which exhibits proficient text retrieval capability. Meanwhile, it makes the incorporation of visual tokens generated by an expressive image encoder, thereby augmenting the capability of image processing. Such an architecture brings forth two important advantages. 1) It establishes the *in-depth fusion* of text and image data, which substantially contributes to the quality of multi-modal embedding. 2) It also enables the *preservation of the original performance* of text embedding, as the text encoder is fully fixed while the visual tokens are incorporated.

深度融合和图像特征到文本嵌入模型中，同时保留了文本嵌入模型的能力

Secondly, we propose two innovative pipelines for the automatic generation of *Image-Text Composed datasets*, thereby securing large-scale, high-quality data for the training of multi-modal embedding models. These pipelines are designed to cater to scenarios where either the query or the candidate comprises image-text pairs, thereby facilitating the model to adapt to a diverse range of multi-modal retrieval situations.

迎合，满足需求

Thirdly, we design a two-stage training algorithm to learn the multi-modal embedding model. Initially, we perform the basic text-to-image matching task with massive weakly-labeled cross-modal data (Schuhmann et al., 2022), which aligns the visual token embedding with the text encoder. Subsequently, we perform composed text&image matching with our generated composed image-text datasets, which establishes the multi-modal representation capability for the embedding model.

visual toekn embedding 到 文本嵌入模型的对齐

VISTA is empirically verified by comprehensive experiments. Particularly, it achieves superior performance across various multi-modal retrieval tasks in both zero-shot and supervised settings. Without any task-specific optimization, VISTA is able to outperform or match the leading approach in every downstream evaluation scenario. Besides, VISTA's performance can also be substantially improved if it is continually fine-tuned for corresponding tasks.

2 Related Work

2.1 General Text Embedding

General text embedding plays an important role in various applications such as web search, question answering (Karpukhin et al., 2020), and retrieval augmented generation for large language models (Lewis et al., 2020b; Borgeaud et al., 2022; Shi et al., 2023). In recent, numerous effective general text embedding models have been developed, including Contriever (Izacard et al., 2022), Sentence-Transformer (Reimers and Gurevych, 2019), OpenAI text embedding (Neelakantan et al., 2022), BGE (Xiao et al., 2023), and M3 (Chen et al., 2024), etc. These models have demonstrated impressive generalizability and robust performance in the realm of text retrieval. However, they exhibit limitations when it comes to handling multi-modal data. This becomes particularly salient with the rising popularity of multi-modal retrieval (Chang et al., 2022; Vo et al., 2019; Luo et al., 2023) and multi-modal retrieval-augmented generation (Chen et al., 2022; Yasunaga et al., 2023).

2.2 General Multi-Modal Embedding

Multi-modal retrieval, characterized by queries and/or candidates composed of image-text data, is gaining increasing popularity in practice (Vo et al., 2019; Chang et al., 2022; Luo et al., 2023). Different from cross-modality retrieval models (Radford et al., 2021) which independently process image and text modalities, multi-modal retrieval necessitates models to have an in-depth understanding of the composed image-text data. Most existing models for multi-modal embedding primarily rely on the pre-trained CLIP (Radford et al., 2021) or BLIP (Li et al., 2022). For instance, models such as UniVL-DR (Liu et al., 2022), Clip4Cir (Baldrati et al., 2023), and UniIR (Wei et al., 2023) initially encode image and text separately using the corresponding encoders from CLIP or BLIP. These models then employ a fusion strategy, such as score fusion, to integrate features from both modalities.

However, these models lack in-depth image-text fusion mechanisms (Wei et al., 2023; Liu et al., 2022) or are designed for specific tasks (Liu et al., 2022; Saito et al., 2023; Baldrati et al., 2023), rather than for a broad spectrum of multi-modal embedding applications. Furthermore, the text embedding capabilities of CLIP and BLIP are not on par with recent general text embedding models, which can potentially compromise their performance in tasks that involve processing text-heavy multi-modal documents (Chang et al., 2022; Luo et al., 2023). A concurrent work, Marvel (Zhou et al., 2023), leverages pre-trained text embedding models as a foundation for encoding composed image-text documents, facilitated by a visual plugin. However, Marvel is a task-specific model trained for multi-modal document retrieval (Chang et al., 2022; Liu et al., 2022), and it cannot be utilized as a general multi-modal embedding model to handle other tasks, such as composed image retrieval.

3 VISTA Model

3.1 Model Architecture

The core idea of our VISTA is the use of the ViT encoder as an image tokenizer for the text encoder. This enables VISTA to encode a variety of data types, including images, text, and composed image-text data. As shown in Figure 1, we treat the Vision Transformer (ViT) (Dosovitskiy et al., 2021) as an image tokenizer of the text encoder, which allows the pre-trained text model to recognize image tokens while remaining frozen. The benefit of this approach is that it facilitates an *in-depth fusion* of text and image data, while the text encoder retains its robust text embedding capabilities.

Specifically, VISTA encodes text data directly using the pre-trained text encoder, as illustrated by the following formula:

$$\mathbf{e}_t = Bert(\{t_0, ..., t_m\}) \tag{1}$$

Here, $Bert$ represents the text encoder (Devlin et al., 2018) and is initialized with a pre-trained general text embedding model. $\{t_0, ..., t_m\}$ and $\mathbf{e}_t$ denote the text sequence and its corresponding text embedding, respectively. Notably, we utilize the normalized hidden state of Bert's special token, `[CLS]`, as the output of the embedding. For image data, the encoding process is defined as follows:

$$\begin{aligned} \{\epsilon_0, ..., \epsilon_n\} &= ViT(\{i_0, ..., i_n\}) \\ \mathbf{e}_i &= Bert(\{\epsilon_0, ..., \epsilon_n\}) \end{aligned} \tag{2}$$

where ViT is a vision transformer serving as an image tokenizer, $\{i_0, ..., i_n\}$ is the token sequence of the input image patches, while $\{\epsilon_0, ..., \epsilon_n\}$ corresponds to the sequence of hidden states for image tokens, as produced by ViT. The image token sequence $\{\epsilon_0, ..., \epsilon_n\}$ is then encoded by $Bert$ to derive the corresponding image embedding $\mathbf{e}_i$. For the composed image-text data, we encode it as:

$$\begin{aligned} \{\epsilon_0, ..., \epsilon_n\} &= ViT(\{i_0, ..., i_n\}) \\ \mathbf{e}_h &= Bert(\{\epsilon_0, ..., \epsilon_n\}; \{t_0, ..., t_m\}) \end{aligned} \tag{3}$$

We concatenate the sequence $\{\epsilon_0, ..., \epsilon_n\}$ and $\{t_0, ..., t_m\}$ together, forming an interleaved sequence of image and text tokens. This interleaved sequence is then encoded by $Bert$ to yield the hybrid multi-modal data representation $\mathbf{e}_h$.

We exclusively trained ViT during the training procedure while maintaining the text encoder $Bert$ in a frozen state. This strategy is adopted to preserve the powerful text embedding capabilities of the pre-trained text general embedding model.

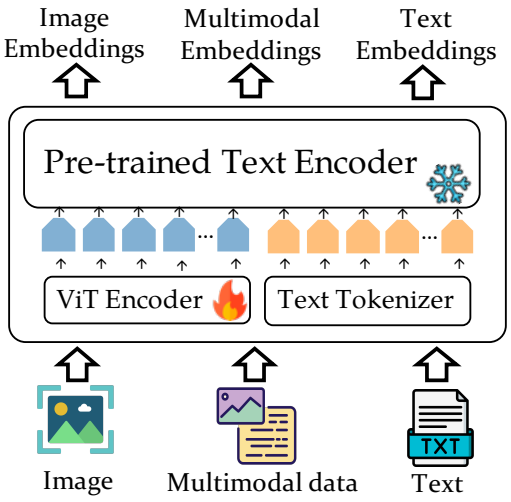

Figure 1: The model architecture of our VISTA model. We use the pre-trained language model as the foundation, making the ViT encoder transfer the Image to recognized tokens of the text encoder.

3.2 Data Construction

Existing hybrid multi-modal datasets predominantly require human annotation, such as writing queries for multi-modal document retrieval (Chang et al., 2022), annotating semantic relations for composed image retrieval (Liu et al., 2021; Wu et al., 2021), and creating questions and answers for knowledge retrieval (Luo et al., 2023). These costly human annotations limit the scale of hybrid multi-modal datasets, posing challenges for training multi-modal embedding models. To address these challenges, we have designed two pipelines to generate hybrid multi-modal data. These pipelines, based on the scenarios where either the ***query*** or the ***candidate*** is composed of image and text, provide a versatile training dataset that can accommodate diverse multi-modal retrieval situations. Our pipelines facilitate the production of two large-scale multi-modal embedding training datasets. The statistical information of our generated dataset is presented in Table 1.

3.2.1 Image&Text To Image (IT2I) Dataset

Inspired by InstructPix2Pix (Brooks et al., 2023), which devises a synthetic image-editing dataset for image editing models, we establish a pipeline for creating a dataset that is characterized by composed image-text queries. As shown in Figure 2, we feed the caption of the source image $\mathcal{C}_s$ into GPT-3.5 (Ouyang et al., 2022), prompting it to generate multiple distinct editing instructions $\{\mathcal{T}^1, ..., \mathcal{T}^m\}$

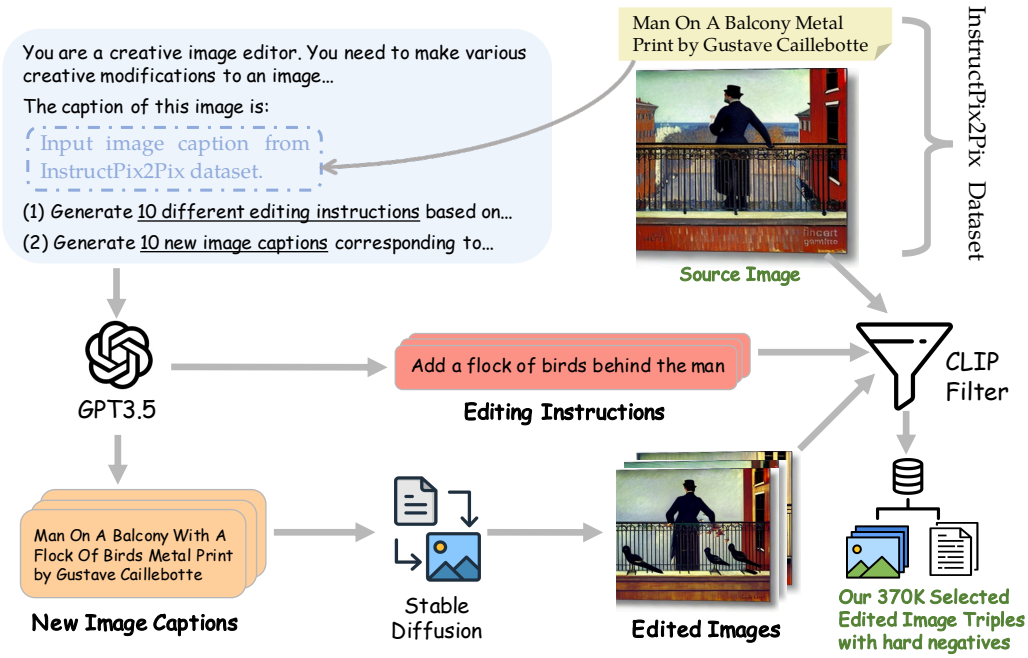

Figure 2: The construction pipeline of Image&Text To Image (IT2T) dataset.

Dataset	Queries	H. Annot.
CIRR (Liu et al., 2021)	36K	✓
FashionIQ (Wu et al., 2021)	30K	✓
Our IT2I Data	307K	×
WebQA (Chang et al., 2022)	21K*	✓
Our T2IT Data	213K	×

Table 1: Comparison of our generated datasets with existing datasets. Queries refers to the number of query-candidate pairs. H. Annot. denotes the necessity of human annotation. * The number of queries corresponding to multi-modal documents in WebQA.

along with their corresponding image captions $\{\mathcal{C}_t^1, ..., \mathcal{C}_t^m\}$, which are then fed into the stable diffusion model (Rombach et al., 2022) to generate the edited images $\{\mathcal{I}_t^1, ..., \mathcal{I}_t^m\}$. We designate different edited images $\mathcal{I}_t^i$ originating from the same source image $\mathcal{I}_s$ as hard negatives for each other. Consequently, we obtain multiple triples $(\mathcal{I}_s, \mathcal{T}^i, \mathcal{I}_t^i)$, where $\mathcal{I}_s$ and $\mathcal{T}^i$ are the composed image-text query, and $\mathcal{I}_t^i$ is the target image. We further employ CLIP (Radford et al., 2021) to filter these triples, resulting in 307K query-candidate pairs with hard negatives.

A major distinction is that our approach generates multiple editing instructions for each source image, while InstructPix2Pix provides only a single editing instruction per source image. Different edited images can work as hard negatives with each other. Therefore, it prevents the training task from collapsing into a naive image-to-image matching task, which enables the model to jointly understand the image and text data.

3.2.2 Text To Image&Text (T2IT) Dataset

We establish another pipeline to construct a pseudo multi-modal document retrieval dataset, in which the candidates are composed of both images and text. Our pipeline operates on a highly descriptive image captioning dataset ShareGPT4V (Chen et al., 2023). ShareGPT4V is characterized by the detailed textual image description that includes multi-granular information, encompassing world knowledge, properties of objects, spatial relationships, etc.

Specifically, for each image $\mathcal{I}$ accompanied by a descriptive caption $\mathcal{C}$, we first input $\mathcal{C}$ into GPT-3.5 and prompt it to generate an article $\mathcal{T}$ that is related to a subtopic of the image. Consequently, we obtain a multi-modal document candidate, denoted as $D = (\mathcal{I}, \mathcal{T})$. We then prompt GPT-3.5 to generate a query Q for the generated multi-modal document D. Through this process, we obtain over 213K triples $(Q, \mathcal{I}, \mathcal{T})$, where Q represents the query and $(\mathcal{I}, \mathcal{T})$ forms the multi-modal document candidate. We demonstrate that the data generated by this simple yet effective pipeline exhibits superior generalization capabilities compared to the manually annotated WebQA (Chang et al., 2022)

when used to train multi-modal embedding models, as detailed in Section 4.3.1. More details of the data generation process are shown in Appendix A.

3.3 Two-Stage Training

We develop a two-stage training strategy to facilitate the text encoder's ability to encode both image and hybrid multi-modal data into a unified embedding space. We initialize the text encoder with a general embedding model BGE-Base-v1.5 (Xiao et al., 2023) and initialize the ViT Encoder with EVA-CLIP-02-Base (Sun et al., 2023).

Stage 1: Cross-Modal Training. In the first training stage, we conduct contrastive language-image pre-training (Radford et al., 2021) to our VISTA. All training data are uni-modal in this stage, and we utilize the Laion-2B (Schuhmann et al., 2022) for in-depth alignment training, thereby transforming the ViT encoder into a high-quality image tokenizer for the general text embedding model. The training objectives are as follows:

$$\min_{\{\theta_I\}} \mathcal{L}_{s1} = \mathcal{L}_{con}(\mathbf{e}_t, \mathbf{e}_i) + \mathcal{L}_{con}(\mathbf{e}_i, \mathbf{e}_t) \quad (4)$$

where θ_I is the parameters of ViT. $\mathcal{L}_{con}(\mathbf{e}_t, \mathbf{e}_i)$ and $\mathcal{L}_{con}(\mathbf{e}_i, \mathbf{e}_t)$ are bidirectional cross-modal contrastive learning losses, and $\mathcal{L}_{con}(\mathbf{u}, \mathbf{v})$ can be formulated as:

$$\mathcal{L}_{con}(\mathbf{u}, \mathbf{v}) = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{exp(\mathbf{u}_i^T \mathbf{v}_i / \tau)}{\sum_{j \in \mathcal{B}} exp(\mathbf{u}_i^T \mathbf{v}_j / \tau)} \quad (5)$$

where $\mathcal{B}$ represents the set of in-batch samples, and τ is the temperature parameter that controls the strength of penalties on negative samples. Following the first stage of training, the image tokenizer develops the ability to encode image tokens in a format that the text encoder can interpret.

Stage 2: Multi-Modal Training. After the first stage of training, the text encoder has gained the ability to independently process image and text modalities, and align them into a unified embedding space. This has laid the groundwork for encoding interleaved sequences of text and image. Building upon this foundation, we further train VISTA to enhance its multi-modal encoding capabilities.

Specifically, we utilize our generated IT2I and T2IT datasets, as constructed in Section 3.2, for multi-modal training. The training objective of both the two tasks can be formulated as:

$$\min_{\{\theta_I\}} \mathcal{L}_{s2} = \mathcal{L}_{con}(\mathbf{q}, \mathbf{c}) \quad (6)$$

where $\mathbf{q}$ and $\mathbf{c}$ represent the embeddings of the query and candidate of these two tasks, respectively. We discover that a 600-step training process on our generated multi-modal training dataset is sufficient to equip VISTA with robust multi-modal embedding capabilities. This not only underscores the effectiveness of our model architecture but also validates the utility of our generated composed image-text training datasets. For more details on training and the hyper-parameter settings used, please refer to Appendix B.

4 Experimental Results

We carry out both zero-shot evaluations and supervised fine-tuning across various benchmarks to substantiate the efficacy and versatility of our VISTA model. In addition, we perform comprehensive ablation studies to scrutinize both the design of the VISTA model and the effectiveness of our stage-2 training datasets.

4.1 Zero-Shot Retrieval Performance

Benchmarks. We collect five distinct datasets, encompassing four different multi-modal retrieval tasks. To construct a challenging zero-shot evaluation setup, we perform our evaluation on the entire corpus of each dataset. The overall statistical information is shown in Table 2, while the detailed information for each benchmark can be found in Appendix C.

Metrics. We uniformly employ Recall@5 as the evaluation metric across all datasets. We employ the dense retrieval approach to evaluate all models across various dataset benchmarks. Each model encodes the query and candidate items into corresponding embedding spaces, and retrieval is performed based on cosine similarity scores using FAISS (Johnson et al., 2019).

Baselines. We benchmark our VISTA model against three established baseline models: CLIP-B[1] (Radford et al., 2021), BLIP-B (Li et al., 2022), and Pic2Word (Saito et al., 2023). We utilize the strategies outlined in (Wei et al., 2023; Liu et al., 2022; Saito et al., 2023) to encode composed image-text data for these baseline models. Further details can be found in Appendix D. Furthermore, to validate the universality of our generated composed image-text dataset, we apply the multi-modal training (Section 3.3) to all baseline models. These baseline models $\mathcal{X}$ are denoted as $\mathcal{X}$-MM.

[1] https://huggingface.co/openai/clip-vit-base-patch16

Dataset	Task	Query Count	Corpus Size	Domain
WebQA (Chang et al., 2022)	$q_t \rightarrow c_t/c_{it}$	4,966	944,766	Wikipedia
CIRR (Vo et al., 2019)	$q_{it} \rightarrow c_i$	4,181	21,551	Open-Domain
FashionIQ (Wu et al., 2021)	$q_{it} \rightarrow c_i$	6,016	74,381	Fashion Products
ReMuQ (Luo et al., 2023)	$q_{it} \rightarrow c_t$	3,609	195,387	Wikipedia
OVEN-QS (Hu et al., 2023)	$q_{it} \rightarrow c_t/c_{it}$	3,291	6,084,491	Wikipedia

Table 2: Statistical information for the zero-shot multi-modal retrieval benchmark datasets. q and c represent query and candidate respectively, with the subscripts i, t, and it denoting image, text, and composed image-text data respectively. During the zero-shot evaluation, we utilize the queries from the validation or test set of each dataset to perform retrieval assessments within the entire corpus of the respective dataset.

Models	# Params	WebQA	CIRR	FashionIQ	OVEN-QS	ReMuQ	Average
CLIP	149M	10.54	13.37	3.56	1.06	65.05	18.72
CLIP-MM	149M	28.77	19.64	5.55	0.40	58.86	22.64
BLIP	224M	10.03	8.25	1.50	0.06	1.25	4.22
BLIP-MM	224M	30.11	10.31	1.23	0.27	55.66	19.52
Pic2Word	224M	12.72	23.42	**8.24**	0.97	68.99	22.87
Pic2Word-MM	224M	24.15	**26.09**	7.65	0.82	78.09	27.36
VISTA (Ours)	196M	**60.11**	22.51	7.51	**8.39**	**84.73**	**36.65**

Table 3: Zero-shot evaluation results with Recall@5 on various hybrid multi-modal retrieval benchmarks. The '-MM' notation indicates baseline models that have undergone multi-modal training on our generated data. For zero-shot evaluation, we utilize the entire corpus of each dataset, encompassing all data splits, as the candidate pool.

Overall Performance. The zero-shot performance of various models is presented in Table 3. Our VISTA model achieves state-of-the-art average performance across all tasks, with more than 9% improvement on Recall@5. On the WebQA, OVEN-QS, and ReMuQ datasets, our model significantly outperforms all baselines in zero-shot retrieval performance. While the performance of our model on the composed image retrieval task on CIRR and FashionIQ is slightly lower than the proprietary model, pic2word, it should be noted that pic2word is a model specifically designed for this task. These results affirm the versatility and efficacy of VISTA in hybrid multi-modal retrieval. In addition, through multi-modal training on our generated dataset, we have seen a significant improvement in the zero-shot performance of all baseline models across various tasks. This demonstrates the efficiency and universality of our generated dataset. In addition, the qualitative zero-shot retrieval results of VISTA can be found in Appendix F.

4.2 Supervised Fine-Tuning Performance

We fine-tune our VISTA model across a variety of hybrid multi-modal retrieval benchmarks, including WebQA, CIRR, and ReMuQ. During the supervised fine-tuning process, we train all parameters of VISTA. Importantly, we abstain from making any task-specific modifications to VISTA and do not utilize any additional training data. The experimental results demonstrate the robustness and exceptional adaptability of VISTA across various hybrid multi-modal retrieval tasks.

弃权，避免

4.2.1 Fine-Tuning Performance on WebQA

Details & Metrics. Following the approach of (Liu et al., 2022), we fine-tune our VISTA on the training set of WebQA (Chang et al., 2022). We employ hard negatives from (Liu et al., 2022) for training and set the count of hard negatives to 9. During fine-tuning, we set the batch size to 288, and the initial learning rate to 2e-5, and fine-tune for 700 steps. During testing, we use the validation query set to retrieve from the entire corpus. Recall@5/10/20 and MRR@10 serve as our evaluation metrics.

Results. The experimental results are presented in Table 4. VISTA achieves 70.8% in Recall@5 and 71.0% in Recall@10. VISTA outperforms the previous SOTA method (Liu et al., 2022) by over 6% and exceeded the concurrent work Marvel (Zhou

Methods	R@5	R@10	R@20	MRR@10
CLIP-DPR	49.6	60.1	70.2	50.6
UniVL-DR	64.5	72.9	78.8	66.8
Marvel-DPR	60.1	69.6	78.0	61.6
Marvel-ANCE	70.8	78.8	84.3	71.0
VISTA (Ours)	**74.9**	**83.7**	**89.4**	**75.3**

Table 4: Supervised fine-tuning results on the WebQA dataset. The baseline models CLIP-DPR and UniVL-DR are taken from (Liu et al., 2022), while Mravel-DPR and Marvel-ANCE are taken from (Zhou et al., 2023). All retrievals are performed on the deduplicated corpus.

Methods	R@5	R_{sub}@1	Avg.
CIRPLANT (Liu et al., 2021)	52.6	39.2	45.9
CompoDiff (Gu et al., 2023)	54.4	35.8	45.1
Combiner (Baldrati et al., 2022)	65.4	62.4	63.9
Blip4CIR+Bi (Liu et al., 2024)	73.1	72.1	72.6
CLIP4Cir (Baldrati et al., 2023)	77.0	73.2	75.1
CoVR (Ventura et al., 2023)	**78.6**	75.0	**76.8**
VISTA (Ours)	76.1	**75.7**	75.9

Table 5: Supervised fine-tuning results on the CIRR test set.

et al., 2023) by more than 4%.

4.2.2 Fine-Tuning Performance on CIRR

Details & Metrics. Following the common protocols for composed image retrieval, we evaluate the model performance on the test set of the CIRR (Liu et al., 2021) dataset. CIRR includes two benchmarks: a standard one where the target search space encompasses the entire test corpus, and a fine-grained subset where the search space is limited to a subgroup of six images similar to the query image. We report Recall@5 (R@5) for the former and Recall@1 (R_{sub}@1) for the latter, and calculate the average of these two recall measures. During fine-tuning, we set the initial learning rate to 2e-5 and the batch size to 720, treating the subgroup as hard negatives for training. The model is fine-tuned for a total of 900 steps.

Results. The experimental results are shown in Table 5. Our VISTA achieves 76.1% in Recall@5, 75.7% in R_{sub}@1, and an overall average performance of 75.9% on the test set. Without employing any task-specific module, our VISTA achieves performance on par with state-of-the-art models that have been pre-trained or specifically designed for composed image retrieval.

Methods	R@5	R@10
BM25 (Robertson et al., 2009)	8.8	10.8
DPR (Karpukhin et al., 2020)	43.4	48.8
SEAL (Bevilacqua et al., 2022)	66.4	74.1
ReViz (Luo et al., 2023)	62.4	71.6
ReViz-ICT (Luo et al., 2023)	76.2	83.3
GeMKR (Long et al., 2024)	90.3	92.7
VISTA (Ours)	**96.3**	**97.3**

Table 6: Supervised fine-tuning results on the ReMuQ test set.

4.2.3 Fine-Tuning Performance on ReMuQ

Details & Metrics. We fine-tune our model on the training set of ReMuQ and test it on its test set. The evaluation is conducted on the entire knowledge base of ReMuQ. We report Recall@5 and Recall@10, and compare our results with state-of-the-art methods. The initial learning rate is set to 2e-5, the batch size to 1,920, and the model is fine-tuned for 200 steps.

Results. As shown in Table 6, VISTA achieves as high as 96.3% in Recall@5, surpassing the latest state-of-the-art method (Long et al., 2024) by more than 5%. These results demonstrate the powerful capability of VISTA in multi-modal knowledge base retrieval and highlight its considerable potential for application in multi-modal retrieval augmented generation for LLMs.

4.3 Ablation Analysis

4.3.1 The Impact of Stage-2 Training Data

Table 7 investigates the effect of our generated composed image-text training data (stage-2 training data) on the zero-shot performance across a variety of tasks. VISTA_{S1} denotes the model post the first stage of image-text cross-modal alignment. Building on this model, we use different training data configurations to analyze the influence of our generated stage-2 training data on the multi-modal embedding capabilities of VISTA.

Compared to the VISTA_{S1} model, the use of our generated Image&Text To Image (IT2I) Dataset leads to significant performance enhancements on the CIRR, FashionIQ, and ReMuQ benchmarks. These benchmarks are characterized by their inherent need for a comprehensive understanding of multi-modal queries. In comparison to training on the InstructPix2Pix dataset (Dai et al., 2023) that lacks hard negatives, the model trained on our IT2I

Model	WebQA	CIRR	FashionIQ	OVEN-QS	ReMuQ	Avg.
VISTA$_{S1}$	35.70	9.59	1.33	3.82	21.53	14.39
w/ InstructPix2Pix	44.24	14.47	2.88	5.14	80.60	29.47
w/ Ours-IT2I	51.87	21.29	6.73	3.40	**89.06**	34.47
w/ WebQA	-	10.64	2.03	3.92	77.42	-
w/ Ours-T2IT	57.28	15.86	3.81	5.38	74.67	31.40
VISTA	**60.11**	**22.51**	**7.51**	**8.39**	84.73	**36.65**
VISTA-SF	59.46	15.93	5.27	1.19	83.04	32.98

Table 7: Ablation studies: The zero-shot performance of models that use different stage-2 training data or multi-modal fusion methods. IT2I and T2IT respectively represent our generated Image&Text To Image Dataset and Text To Image&Text Dataset. Underlined values indicate where the associated training dataset has significantly improved performance on the corresponding benchmarks. VISTA-SF is an ablation model that employs the score-fusion method to encode composed image-text data.

dataset demonstrates superior multi-modal retrieval performance. The hard negatives that we generate can foster the model's comprehension of the interplay between images and text, thereby preventing the model from excessively relying on image feature similarity rather than semantic correlation during the training process.

In contrast to the VISTA$_{S1}$ model, the employment of our generated Text To Image&Text (T2IT) dataset notably enhances the zero-shot performance on the WebQA and OVEN-QS benchmarks. These benchmarks require a composed understanding of multi-modal candidates, a demand directly met by the training scenarios presented in our T2IT dataset. Additionally, we evaluate the performance of the model using WebQA for second-stage training. Except for ReMuQ, which is sourced from WebQA, the model exhibits superior performance when trained with our T2IT dataset compared to when it is trained with WebQA. This suggests that our T2IT data offers enhanced generalization capabilities compared to manually annotated datasets when used to train multi-modal embedding models.

In the final Stage-2 training, we conduct dual-task training using both our IT2I and T2IT datasets on the basis of VISTA$_{S1}$, resulting in the development of VISTA. Among all different training data configurations, VISTA achieves the best performance in four out of the five benchmarks. This result confirms the synergistic effect of integrating both datasets in the training of multi-modal embedding models.

4.3.2 Multi-Modal Fusion Methods

Table 7 also examines the benefits of the VISTA model architecture in encoding composed image-text data in comparison to VISTA-SF (last line). For VISTA-SF, a score-fusion approach is employed during the encoding of multi-modal data. This involves independently encoding images and text using the VISTA model, followed by an element-wise addition of the embeddings derived from both modalities. The experimental results demonstrate that VISTA achieves substantial improvement, significantly outperforming the VISTA-SF model. This can be attributed to the fact that score-fusion cannot deeply comprehend the integration of images and text, whereas VISTA is proficient in consistently encoding and interpreting composed image-text data. These findings underscore the advantages of the VISTA model in encoding interleaved text and visual sequences.

5 Conclusion

In this paper, we introduce **VISTA**, an VISualized Text embedding approach for universal multi-modal retrievAl. Our work makes three significant contributions. Firstly, we design a flexible model architecture that enables the in-depth fusion of text and image data, while maintaining the powerful performance of the general text embedding models. Secondly, we develop two data generation strategies for training multi-modal embedding models without the need for manual annotation. Lastly, we introduce a two-stage training algorithm that rapidly enhances the multi-modal representation capability of VISTA. Extensive experimental results

demonstrate the superior performance of VISTA in both zero-shot and supervised fine-tuning settings for various multi-modal retrieval tasks.

Limitations

As we reflect on the work conducted, we identify two areas of potential refinement in our approach. The first area concerns the diversity of image styles in our Image&Text To Image (IT2T) dataset. The images are generated using a stable diffusion model, which might have limited the range of styles in the dataset. The second area relates to the handling of image tokens. In our current approach, we feed all image tokens directly into the text encoder. This procedure might inadvertently heighten the computational load due to the uniform sequence length. A potential improvement could be the implementation of variable-length image token sequences, which could reduce sequence lengths and consequently lead to more efficient computation.

Ethics Statement

All training data used in our model have undergone rigorous screening to remove harmful content. This includes both the LAION-5B dataset constructed by (Schuhmann et al., 2022) and the multi-modal dataset we built ourselves. Despite our best efforts, we acknowledge that we cannot fully guarantee that these screenings were entirely comprehensive or without omissions. Furthermore, we strongly discourage the use of VISTA for encoding and retrieving sensitive content.

Acknowledgements

This research is supported by National Science and Technology Major Project (2023ZD0121504) and National Natural Science Foundation of China (NSFC-62306046, NSFC-62272054).

References

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2022. Effective conditioned and composed image retrieval combining clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21466–21474.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. 2023. Composed image retrieval using contrastive learning and task-oriented clip-based features. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–24.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. *Advances in Neural Information Processing Systems*, 35:31668–31683.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.

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.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5558–5570. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. 2023. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 12031–12041. IEEE.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

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

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. 2023. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400.

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2022. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In *The Eleventh International Conference on Learning Representations*.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134.

Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, and Stephen Gould. 2024. Bi-directional training for composed image retrieval via text prompt learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5753–5762.

Xinwei Long, Jiali Zeng, Fandong Meng, Zhiyuan Ma, Kaiyan Zhang, Bowen Zhou, and Jie Zhou. 2024. Generative multi-modal knowledge retrieval with large language models. *arXiv preprint arXiv:2401.08206*.

Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2023. End-to-end knowledge retrieval with multi-modal queries. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8573–8589. Association for Computational Linguistics.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and code embeddings by contrastive pre-training. *CoRR*, abs/2201.10005.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei

Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9844–9855. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: retrieval-augmented black-box language models. *CoRR*, abs/2301.12652.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. 2023. Covr: Learning composed video retrieval from web video captions. *arXiv preprint arXiv:2308.14746*.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2023. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39755–39769. PMLR.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: Bert and beyond. In *Proceedings of the 14th ACM International Conference on web search and data mining*, pages 1154–1156.

Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. 2023. Unlock multi-modal capability of dense retrieval via visual module plugin. *arXiv preprint arXiv:2310.14037*.

Appendix

A More Details of Data Construction

Image&Text To Image (IT2T) Dataset. The specific prompts utilized in creating the I2IT dataset are illustrated in Figure 3. We direct GPT-3.5 to generate a range of image editing instructions that significantly alter the semantic content of the images. This strategy diverges from how image editing datasets for image editing models are constructed. Datasets such as InstructPix2Pix (Brooks et al., 2023) prioritize dealing with the challenges of intricate image edits, whereas our focus with embedding models is more on understanding the relationships among image semantics.

Text to Image&Text (T2IT) Dataset. The steps and prompts used in the construction of the T2IT dataset are illustrated in Figure 4. As described in Section 3.2.2, the generation process of the T2IT dataset is divided into two parts: the first step involves generating an article about the image subtopic, and the second step involves generating a query for the multimodal document. It is worth noting that, in the second step of generating a query for the multimodal document, we still use descriptive captions from ShareGPT4V (Chen et al., 2023) to represent the images in the multimodal documents, as GPT-3.5 cannot directly process image data.

B More Training Details of VISTA

In the first training stage, we utilize the FLIP (Li et al., 2023) strategy to improve the time efficiency of image-text contrastive training. We randomly mask 50% of image tokens. This phase is trained for 116K steps. Subsequently, we conduct unmasked tuning (Li et al., 2023) for an extra 48K steps. The batch size is 16K throughout this stage. In the second training stage, the quantity of hard negative examples in the IT2I dataset is set to 3. This stage is trained for only 600 steps with a batch size of 1920. Across both stages, we set the temperature coefficient τ for contrastive learning at 0.02. We initiate the learning rate at $2e-5$ and apply a linear decay strategy for subsequent adjustments.

C Detailed Information of Benchmarks

To evaluate the effectiveness of our VISTA model in hybrid multi-modal retrieval tasks, we collected five distinct datasets, encompassing four different multi-modal retrieval tasks. Each dataset features hybrid data in either the query or candidate components, requiring a joint embedding of image and text data. Unless otherwise stated, for each dataset, all zero-shot evaluations are conducted using the dev split as the query set, with retrieval carried out across the entirety of its corpus.

> You are a creative image editor. You need to make various creative modifications to an image.
> The caption of this image is:
>
> Source Image caption from InstructPix2Pix dataset.
>
> You need to make different types of edits to this image, so you need to **write 10 edit instruction statements and the corresponding new captions based on the original caption**.
>
> Please unleash your creativity and devise imaginative image modifications, ensuring each one is unique to preserve the diversity of generated editing instruction statements. Generating editing instructions that can result in substantial and meaningful changes in the images' content is more preferred.
>
> Please return the generated edit statements and corresponding output captions in JSON format, with the format of ....

Figure 3: The specific prompts utilized during the generation of the Image&Text To Image (IT2T) dataset.

WebQA (Chang et al., 2022) is a multi-hop and multi-modal open-domain question answering dataset. The *Multi-modal Documents Retrieval* task on WebQA is proposed by (Liu et al., 2022). This task involves identifying suitable text or image-text pair candidates based on a query text. Notably, we de-duplicate the corpus of WebQA, following the same procedure as UniIR (Wei et al., 2023). The processed corpus encompasses 544,489 unique text documents, and 403,277 distinct image-caption pairs.

CIRR (Liu et al., 2021) is an open-domain dataset designed for the *Composed Image Retrieval (CIR)* task (Vo et al., 2019). In this task, each query comprises an image-text pair, which includes a reference image and instructive text that delineates the differences between the reference and target images.

FashionIQ (Wu et al., 2021) serves as another relevant dataset for the CIR task, specifically focusing on fashion products, including dresses, shirts, and top-tees.

ReMuQ (Luo et al., 2023) is a dataset curated for *Knowledge Retrieval with Multi-modal Queries*

Step1: Re-writing caption for image:

I will provide you with a detailed description of an image. Based on this description, please imagine the scene of this image and **write an article about a specific subtopic related to this image**.

The description of this image is:

Descriptive caption from ShareGPT4V dataset.

The topic of the article can be about an object, a character, or any content in the background of the image. You can imagine yourself as a sharer, a popular science writer, or any identity you like. You can write a description, reflection, or informative article about the image from any specific perspective. Use your imagination. However, please avoid generating generic image description text. The length of article is around...

Step2: Generating query sentence for image-caption pairs:

I will provide you with a detailed description of an image and its accompanying caption. Please imagine that you have already seen this image, and combine it with the corresponding caption to understand them as a multimodal document.

Here are the image description and caption:

Descriptive caption from ShareGPT4V dataset

Rewritten caption from Step 1

Please imagine yourself as a multimodal document search user and **write a suitable query statement that can locate this multimodal document**. The length of query sentence is around...

Figure 4: The specific prompts employed in the generation of the Text to Image&Text (T2IT) dataset, with the lengths of the articles and queries randomly assigned in each data generation iteration to ensure diversity. Typically, articles are approximately 50 words, and queries are within 20 words.

task. In the ReMuQ dataset, each query is composed of an image and an associated textual question, the objective being to search pertinent content from a textual knowledge base. Due to the absence of a dev split, we evaluate ReMuQ on the test split. **OVEN-QS** is the `Query Split` of the OVEN benchmark (Hu et al., 2023). OVEN-QS is proposed for the *Entity Retrieval with Visually-Situated Queries* task. Each query in OVEN-QS is an image-text pair, with the text component being visually situated and filtered out from VQA scenarios. This dataset necessitates identifying the correct entity from a diverse set of text or image-text candidates.

D Implementation details of Zero-Shot Baselines

For both CLIP and BLIP models, we adopt a score fusion approach on image-text pair data as outlined in (Wei et al., 2023; Liu et al., 2022). The score fusion process is represented as follows:

$$\mathbf{e}_h = \Phi_T(T) + \Phi_I(I) \tag{7}$$

where $\mathbf{e}_h$ denotes the embedding of composed image-text data, T and I represent the text and image data respectively, while Φ_T and Φ_I refer to the text and image encoders of the CLIP/BLIP models, correspondingly.

For Pic2Word model, a proprietary model dedicated to zero-shot composed image retrieval. We employ this model by mapping an image to a pseudo language token in the text encoder, adhering to their methodology for encoding composed image-text queries:

$$\begin{aligned} \texttt{[*]} &= Img2Txt(\Phi_I(I)) \\ \mathbf{e}_h &= \Phi_T(\texttt{"a photo of [*]"}; T) \end{aligned} \tag{8}$$

In this equation, $Img2Txt$ is the mapping network that translates the image feature into a pseudo language token, denoted as $[*]$. The image and text encoders, Φ_I and Φ_T are derived from the CLIP-L model (Radford et al., 2021). It derives the composed image-text features by integrating the pseudo language token with the text, subsequently processing this unified input through the text encoder.

E The Impact of Token Order

As outlined in Equation (3), we exclusively utilized the (image tokens, text tokens) order to process the interleaved image-text data in all previous experiments. To investigate the potential impact of

Table 8: Influence of token order on VISTA performance in processing interleaved image-text data.

Token Order	WebQA	CIRR	FashionIQ	OVEN-QS	ReMuQ	Avg.
(visual tokens, text tokens)	60.11	22.51	7.51	8.39	84.73	36.65
(text tokens, visual tokens)	59.49	22.10	7.36	9.54	84.68	36.63

token order on the performance of our VISTA, we conducted additional experiments using the (text tokens, image tokens) order. The evaluation results in zero-shot settings are presented in Table 8, with all reported results based on the Recall@5 metric. The experimental findings indicate that altering the order of image and text tokens does not significantly impact the model's performance.

F Zero-Shot Qualitative Results

Figures 5, 6, and 7 illustrate qualitative zero-shot examples from our VISTA model on the CIRR (Liu et al., 2021), Fashion IQ (Wu et al., 2021), and WebQA (Chang et al., 2022) datasets, respectively. In each figure, the query is presented on the extreme left, with the top@K retrieval results displayed to its right. The ground-truth candidate, as identified by the benchmark, is denoted by a green box. We conduct our retrieval process across the entire corpus for each dataset, which leads to identifying candidates that meet the retrieval requirements but are not tagged as ground truth. These results demonstrate the impressive zero-shot multi-modal retrieval performance of our VISTA model.

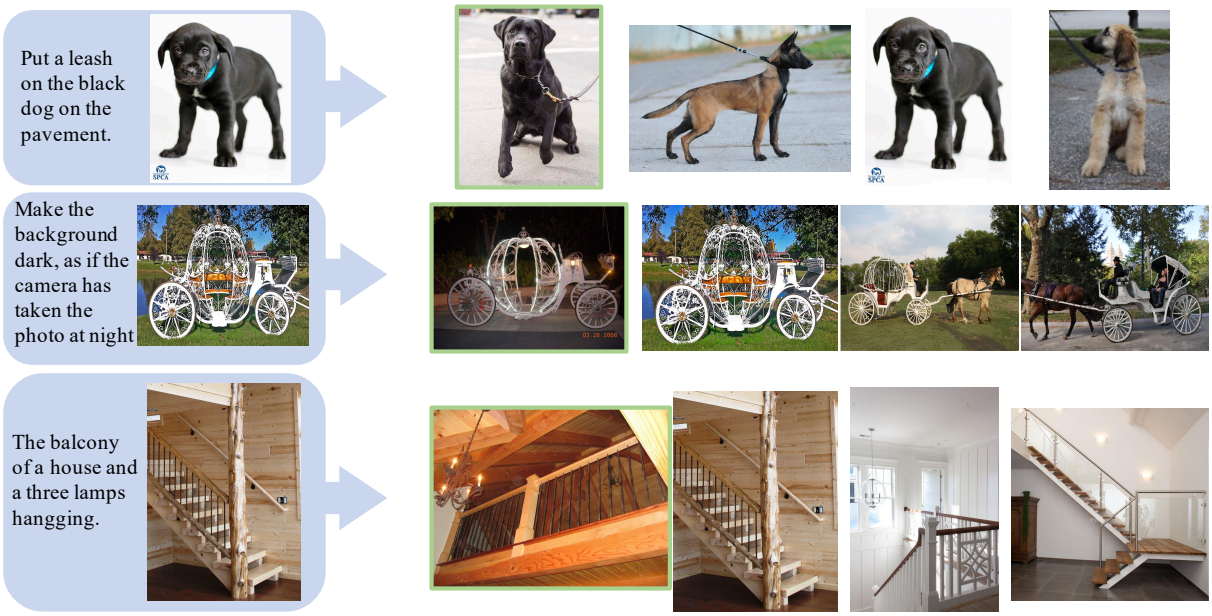

Figure 5: The Qualitative Examples of our VISTA Model on the CIRR Benchmark.

Figure 6: The Qualitative Examples of our VISTA Model on the FashionIQ Benchmark.

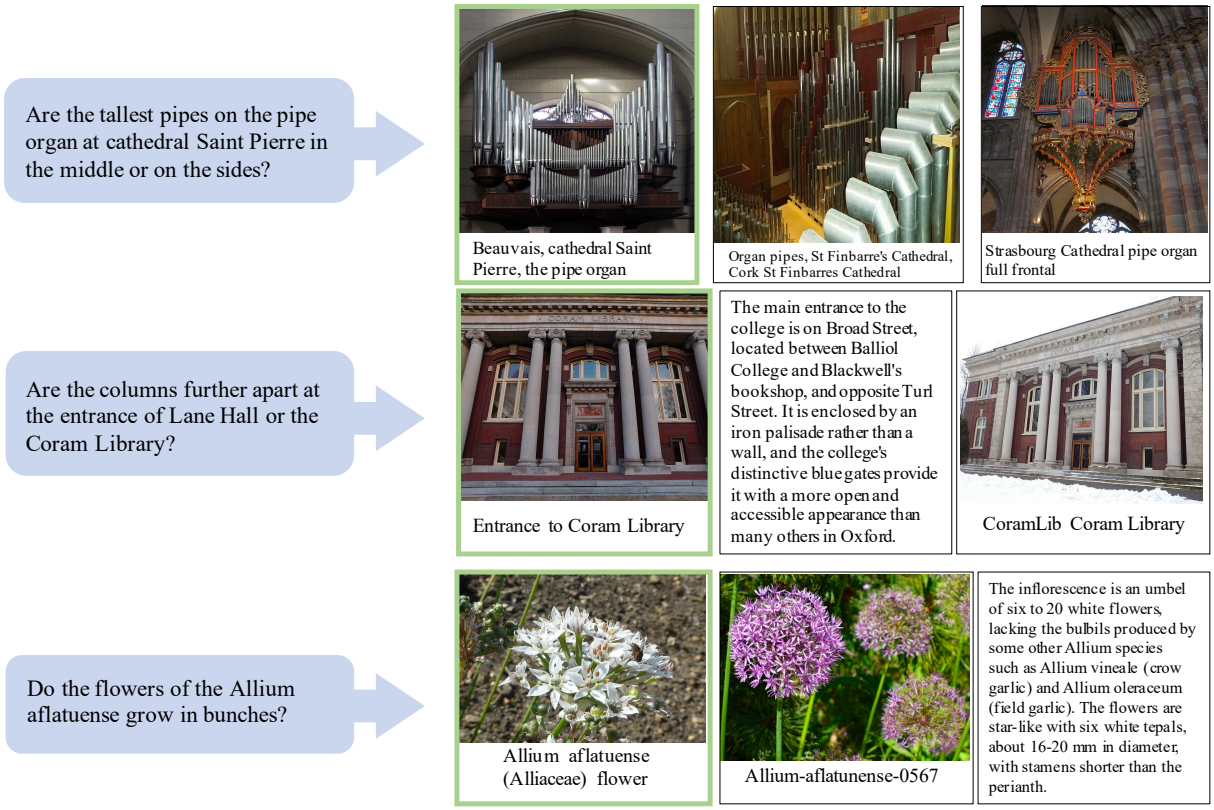

Figure 7: The Qualitative Examples of our VISTA Model on the WebQA Benchmark.