LamRA: Large Multimodal Model as Your Advanced Retrieval Assistant

Yikun Liu[1,2*], Yajie Zhang[3], Jiayin Cai[3], Xiaolong Jiang[3], Yao Hu[3],
Jiangchao Yao[2], Yanfeng Wang[1†], Weidi Xie[1†]
[1]School of Artificial Intelligence, Shanghai Jiao Tong University, China
[2]CMIC, Shanghai Jiao Tong University, China [3]Xiaohongshu Inc., China

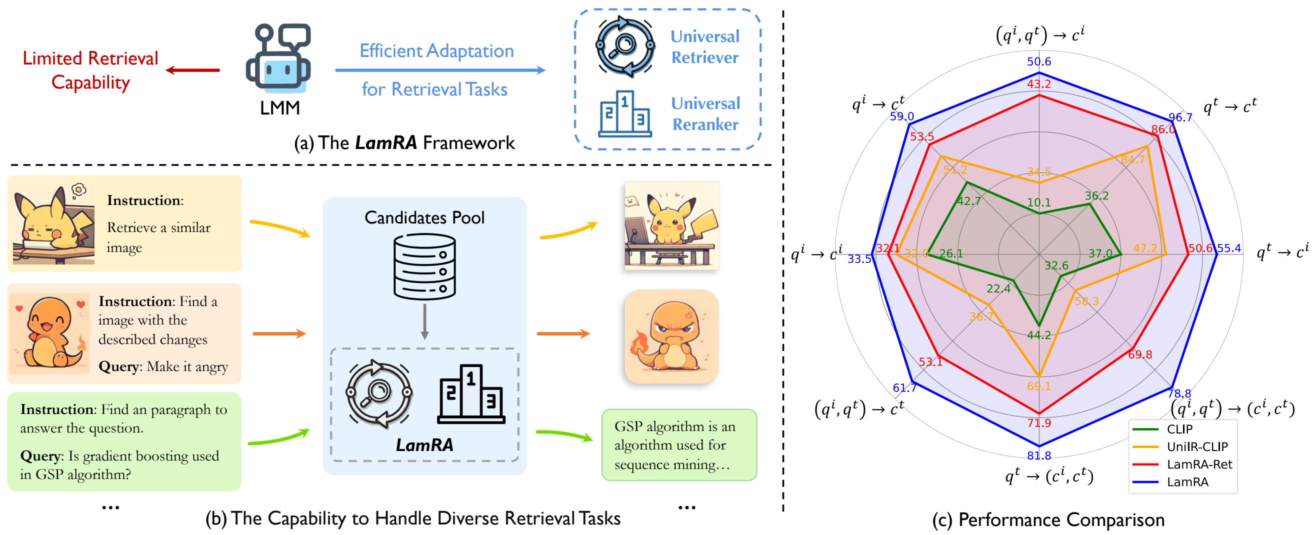

Figure 1. The **LamRA** framework empowers Large Multimodal Models with advanced retrieval and reranking capabilities. (a) LamRA enhances LMMs with universal retrieval and reranking capabilities by inserting lightweight LoRA modules into the LMMs. (b) Examples of varied retrieval tasks demonstrate LamRA's capability to handle diverse retrieval tasks. (c) Performance comparison on the M-BEIR test set shows LamRA's superior performance across a wide range of retrieval tasks. For instance, $q^t \rightarrow c^i$ represents text-to-image retrieval.

Abstract

*With the rapid advancement of multimodal information retrieval, increasingly complex retrieval tasks have emerged. Existing methods predominately rely on task-specific fine-tuning of vision-language models, often those trained with image-text contrastive learning. In this paper, we explore the possibility of re-purposing generative Large Multimodal Models (LMMs) for retrieval. This approach enables unifying all retrieval tasks under the same formulation and, more importantly, allows for extrapolation towards unseen retrieval tasks without additional training. Our contributions can be summarised in the following aspects: (i) We introduce **LamRA**, a versatile framework designed to empower LMMs with sophisticated retrieval and reranking capabilities. (ii) For retrieval, we adopt a two-stage training strategy comprising language-only pre-training and multimodal instruction tuning to progressively enhance LMM's retrieval performance. (iii) For reranking, we employ joint training of both pointwise and listwise reranking, offering two distinct ways to further boost the retrieval performance. (iv) Extensive experiments underscore the efficacy of our method in handling more than ten retrieval tasks, demonstrating robust performance in both supervised and zero-shot settings, including scenarios involving previously unseen retrieval tasks. Project page: https://code-kunkun.github.io/LamRA/.*

1. Introduction

Recently, the field of multimodal information retrieval has been significantly advanced by the remarkable success of Vision-Language Models (VLMs). For instance, CLIP [35]

*Work was done during internship in Xiaohongshu.
†Corresponding author.

and ALIGN [13], which are trained through contrastive learning [32] on large-scale image-text pairs, have demonstrated surprisingly impressive generalizability on cross-modal retrieval. However, as the landscape of information retrieval evolves, more challenging tasks have emerged, including composed image retrieval [29, 46], long-text image retrieval [52], and image/question to multimodal document retrieval [6, 30]. Existing methods [28, 48, 49] employ task-specific fine-tuning to adapt VLMs to these specialized tasks, which remains cumbersome and laborious.

In parallel, recent advancements in Large Multimodal Models (LMMs) [20, 21, 40] have spurred a growing trend toward leveraging LMMs for a diverse range of vision-language tasks. For instance, LISA [19] uses LMMs for segmentation tasks [55], while LLaVA [27] applies them to general Visual Question Answering (VQA). The success of LMMs in these domains can be attributed to two key factors: (i) the use of language as an interface between machines and humans, which naturally facilitates interaction across multiple tasks; and (ii) the training of Large Language Models (LLMs) on extensive text corpora, which has substantially improved their comprehension of natural language and, crucially, enriched their understanding of real-world knowledge. In this paper, we aim to extend this success by addressing a broad range of retrieval and reranking challenges within the scope of LMMs. This approach enables the unification of all retrieval tasks under the same formulation and allows for extrapolation to unseen retrieval tasks without the need for additional training.

We introduce a framework that enhances LMMs with universal retrieval and reranking capabilities, by simply inserting lightweight LoRA modules into LMMs, termed as **LamRA**. For quick retrieval, we adopt a two-stage training strategy to progressively improve the model's retrieval performance. Specifically, the first stage involves text-only pre-training, to enable the LMM to output embeddings in response to summarization prompts. The second stage involves instruction-tuning, where we fine-tune the model on diverse datasets across various retrieval tasks. For reranking, we train an additional LoRA module on the LMM, to handle multiple images or lengthy texts. Using hard sample mining from the quick retrieval model, we conduct joint training for both pointwise and listwise reranking, enabling flexible support for single or multiple candidate inputs, further boosting the overall performance significantly.

To validate the effectiveness of our method, we conduct extensive experiments across more than 10 distinct retrieval scenarios, including text-to-image, text-image-to-image, and text-image-to-text-image retrieval, *etc*. The robust performance indicates that our method can handle both simple and complex retrieval tasks effectively. Furthermore, we examine the model's task-level generalization by training on a subset of retrieval tasks and evaluating on previously unseen tasks. Remarkably, our method demonstrates strong generalization abilities, effectively performing on held-out tasks without prior exposure, underscoring its potential for transferability to unseen retrieval challenges.

To summarise, in this paper, we make the following contribution: (i) We introduce **LamRA**, a versatile framework designed to empower LMMs with sophisticated retrieval and reranking capabilities. (ii) To enable efficient retrieval, we propose LamRA-Ret, which adopts a two-stage training strategy comprising language-only pre-training and multimodal instruction tuning to progressively enhance the model's retrieval capability. (iii) For reranking, we present LamRA-Rank, which supports both pointwise and listwise reranking to further boost the retrieval performance. (iv) Through extensive experiments, we demonstrate the effectiveness of our method in both supervised and zero-shot settings, achieving robust performance even on previously unseen retrieval tasks. Consequently, our method performs on par with or significantly surpasses existing state-of-the-art (SOTA) models across more than ten retrieval tasks.

2. Related Work

Multimodal Information Retrieval. Traditional multimodal information retrieval has often been limited to cross-modal retrieval settings, with evaluation benchmarks typically restricted to datasets like MSCOCO [26] and Flickr30K [34]. As the field progresses, more complex retrieval tasks have emerged, including composed image retrieval [2, 29, 39, 46], long-text-to-image retrieval [52], and image/question-to-multimodal document retrieval [6, 11], among others. As discussed in [6, 30, 52], these tasks present substantial challenges for VLMs. Current approaches [1, 3, 31, 48] often address this by fine-tuning models on individual tasks, yet developing universal multimodal embeddings capable of handling these various tasks simultaneously remains a significant challenge.

Multimodal Representation Learning. The challenge of developing robust multimodal representation remains a foundational question in multimodal learning. Pioneering models such as CLIP [35] and ALIGN [13] have explored this by employing a dual-encoder architecture to learn effective representations through contrastive learning [32] on large-scale image-text pairs. However, given that dual-encoder networks like CLIP aim to enhance the alignment between the image and text, they encounter challenges when dealing with interleaved image-text inputs. Moreover, the text encoder learned in this way shows a limited understanding of complex text [10], suggesting the traditional dual-encoder models may require further refinement.

Numerous studies have pursued further exploration to achieve more universal multimodal representation. For instance, UniIR [45] introduces a benchmark comprising eight distinct retrieval tasks, demonstrating that CLIP, when

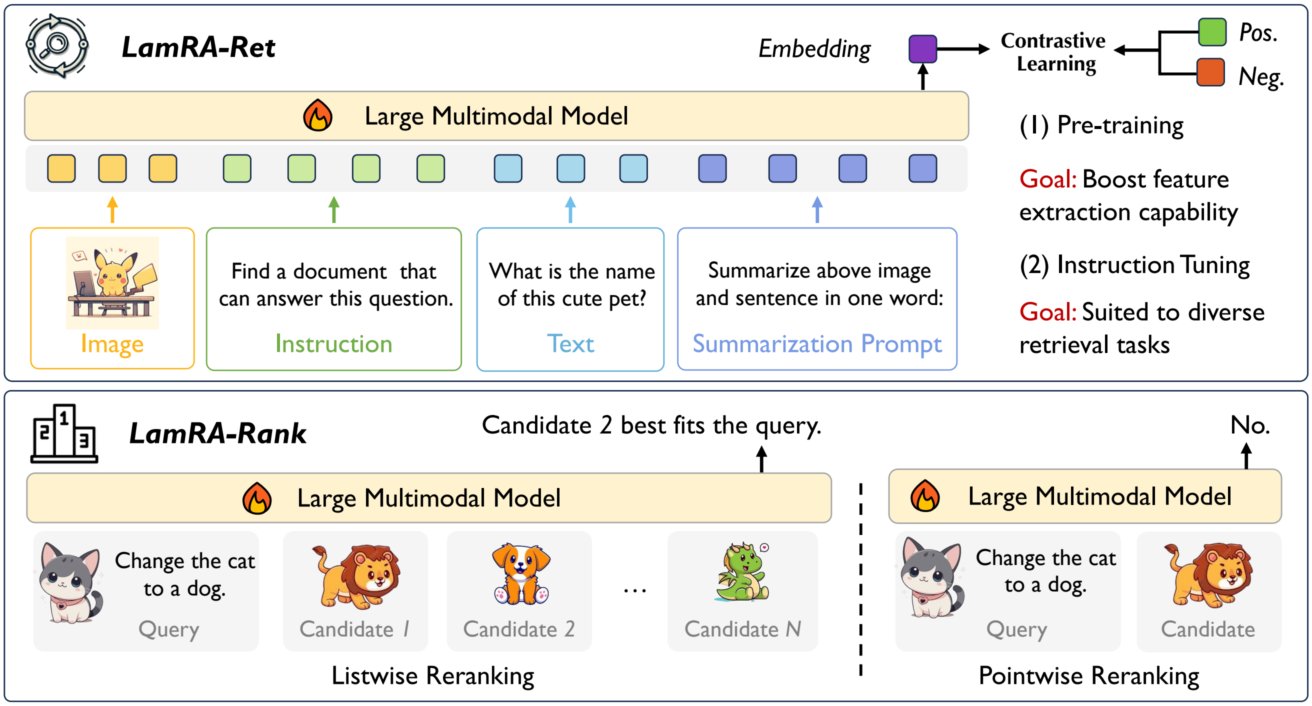

Figure 2. **Overview of the proposed LamRA framework.** LamRA consists of two components: LamRA-Ret and LamRA-Rank. The top section illustrates LamRA-Ret, encompassing both the pre-training and instruction-tuning stages, where contrastive learning is employed to enhance the retrieval capability of LMMs. The pre-training stage aims to improve the feature extraction capabilities through text-to-text retrieval, while the instruction tuning stage adapts the LMMs to various retrieval tasks by fine-tuning on diverse tasks with task-specific instructions. The bottom section depicts the joint training process of LamRA-Rank, which integrates both pointwise and listwise reranking.

trained on this benchmark, achieves enhanced generality and adaptability across various retrieval tasks. E5-V [16] leverages Large Multimodal Models (LMMs), using carefully designed prompts to map images and text into a shared language hidden space. Through fine-tuning on text pairs [9], E5-V demonstrates remarkable zero-shot retrieval capabilities. However, these methods demonstrate limited effectiveness on certain complex retrieval tasks. In this paper, we explore the potential of LMM as a universal retriever and reranker to address these challenges.

Large Multimodal Model. With recent advancements in Large Language Models (LLMs) [4, 8, 14], recent research have focused on Large Multimodal Models (LMMs) to align visual and textual modalities through visual instruction tuning. In particular, there is a growing trend toward using LMMs to tackle a wide range of vision-language tasks [36, 53, 55]. For instance, LISA [19] applies LMMs to segmentation tasks, DetGPT [33] utilizes LMMs for object detection, and VisionLLM [41] addresses various vision-centric tasks using LMMs. Nevertheless, relatively few studies have explored the potential of LMMs in universal retrieval tasks, which is our focus in this paper, where we propose a framework LamRA, that adapts LMMs with lightweight LoRAs for advanced retrieval and reranking, significantly enhancing their performance in various retrieval tasks. We notice a few concurrent studies [17, 25] also explore the use of LMMs for universal retrieval, and we hope that further research will continue to advance this field.

3. Method

This section starts by formulating our considered problem in Section 3.1; then we elaborate on the architecture and feature extraction process in Section 3.2; and describe the training details to equip the LMMs with retrieval capability in Section 3.3; then, we present the approach to endow the LMMs with universal reranking abilities in Section 3.4, aiming to further boost retrieval performance; lastly, we detail the inference pipeline of our method in Section 3.5.

3.1. Problem Formulation

We address the challenge of universal retrieval and reranking with a generative model. Specifically, for one specific query (q), which can be an image, text, or an interleaved image-text format, and a retrieval set of N candidates $\Omega = \{c_1, c_2, \cdots, c_N\}$ where each c_i comprises images, text, or interleaved formats, we first extract the embeddings for both the query and all candidates with LMMs, and rank all candidates in Ω based on the relevance, *i.e.*, cosine similarity between the query and candidate embeddings. This initial retrieval process yields the top-K candidates, represented as $\mathcal{C}_1 = \Phi_{\text{ret}}(q, \Omega)$. Subsequently, we refine this subset through a reranking process to produce the final ranked output, denoted as $\mathcal{C}_2 = \Phi_{\text{rerank}}(q, \mathcal{C}_1)$. Here, $\mathcal{C}_2$ represents the final reordered set of candidates.

In the following section, we provide details on extracting features for queries and candidates of arbitrary format.

3.2. Architecture & Feature Extraction

Generally speaking, the architecture of Large Multimodal Models (LMMs) typically consists of three main components: a vision encoder, a vision projector, and a language model. To compute the embedding with a generative model, we adopt a similar approach as [15, 16], using an Explicit One-word Limitation (EOL). Specifically, we use the following prompts: (i) for image-only input, we use: <image> Summarize above image in one word: <emb>; (ii) for text-only input, we use: <text> Summarize above sentence in one word: <emb>; (iii) for mixed image-text input, we utilize: <image$_1$><text$_1$>...<image$_i$><text$_j$> Summarize above image and sentence in one word: <emb>. where <image> and <text> denote placeholders for the input image and sentence, respectively. For any combination of images and text inputs, the LMM uses the vision encoder and vision projector to map images into language space, which are then processed alongside the text by the large language model. We use the last hidden state immediately preceding the <emb> token as the representation of the input.

3.3. Training for Retrieval

In this section, we describe a two-stage training scheme, to advance the retrieval capabilities of LMMs. The first stage involves language-only pre-training to output improved embeddings. The second stage, instruction tuning, further adapts the LMMs to diverse retrieval tasks.

Stage-I: Adapting LMMs for Retrieval Tasks. LMMs were trained for generative tasks, *e.g.*, next-token prediction, which limits their ability for retrieval, as evidenced by the performance metrics in Table 2, *i.e.*, directly employing Qwen2-VL-7B [40] for retrieval tasks results in inferior performance. To address this issue, we start by adapting LMMs for text-to-text retrieval by training LoRA modules on the Natural Language Inference (NLI) dataset [9], consisting pairs of text samples that are ideal to enhance LMMs for retrieval. For further discussion on the selection of pre-training datasets, please refer to Appendix A.

Stage-II: Instruction Tuning for Universal Retrieval. To further enhance the retrieval capabilities of the LMMs across various multimodal retrieval tasks, we train on M-BEIR, which comprises 8 retrieval tasks with 10 datasets, as curated by [45], including image-to-image retrieval, composed image retrieval, and image/question-to-multimodal-document retrieval, among others. For different retrieval tasks, we incorporate task-specific instructions tailored for each retrieval type. For example, to complete an image-to-image retrieval task, the instruction can be "*retrieval a similar image.*" For further details regarding the instructions and M-BEIR datasets, please refer to Appendix C.

Training Objective. We employ contrastive learning with the InfoNCE loss [32] for both language-only pre-training and instruction tuning stages. Specifically, given a batch size of B, the embeddings of n-th query q_n should be positioned close to the embeddings of its positive target c_n and far away from other negative instances, formulated as:

$$\mathcal{L}_{\text{ret}} = -\frac{1}{B}\sum_{n=1}^{B} \log\left[\frac{\exp\left[\kappa\left(\texttt{LMM}(q_n), \texttt{LMM}(c_n)\right)/\tau\right]}{\sum_{m=1}^{B}\exp\left[\kappa\left(\texttt{LMM}(q_n), \texttt{LMM}(c_m)\right)/\tau\right]}\right]$$

where τ refers to the temperature parameter, and $\kappa(\cdot,\cdot)$ denotes the cosine similarity. Additionally, $\texttt{LMM}(\cdot)$ represents the feature extraction process with LMMs, as mentioned in Section 3.2. The model after both pre-training and instruction tuning is referred as **LamRA-Ret**.

3.4. Training for Reranking

Here, we exploit the flexibility of LMMs to process multiple images or long texts and train a lightweight LoRA module to adapt them for reranking. As a result, we propose **LamRA-Rank**, which supports inputs of single or multiple candidates from the outputs of LamRA-Ret, to further improve the retrieval performance.

Collecting Training Data for Reranking. Here, we employ the LamRA-Ret model trained in Section 3.3 as the initial retriever, and train the reranking model on its top 100 retrieved candidates and use them as hard negatives. This approach enables LamRA-Rank to acquire a more nuanced understanding and improve its ranking performance.

Joint Training for Pointwise and Listwise Reranking. Given that the LMMs can accept flexible inputs, we explore joint training for both pointwise and listwise reranking.

In the case of pointwise reranking, for any query (q), we randomly select a negative candidate (c_{neg}) from the top 100 candidates and instruct the LamRA-Rank to output YES for the ground truth (c_{pos}) and NO for the negative candidate. Cross-entropy loss is employed as the loss function, expressed as $\mathcal{L}_{\text{point}} = \mathcal{L}_{\text{ce}}(\texttt{YES}, \texttt{Reranker}(q, c_{\text{pos}})) + \mathcal{L}_{\text{ce}}(\texttt{NO}, \texttt{Reranker}(q, c_{\text{neg}}))$, where $\texttt{Reranker}(\cdot,\cdot)$ represents the autoregressive output process of LamRA-Rank.

In contrast, for listwise reranking, we randomly select M negative samples $(c_1, c_2, \cdots, c_M)$ from the top 100 candidates, where M is a randomly determined integer between 2 and 5. We then randomly insert the ground truth (c_{pos}) at any position and prompt the LamRA-Rank to directly output the position number of the ground truth. This process can be formulated as $\mathcal{L}_{\text{list}} = \mathcal{L}_{\text{ce}}(\texttt{GT-POSITION}, \texttt{Reranker}(q, c_{\text{pos}}, c_1, c_2, \cdots, c_M))$.

The final loss is a weighted sum: $\mathcal{L}_{\text{rank}} = \mathcal{L}_{\text{point}} + \mathcal{L}_{\text{list}}$.

3.5. Inference Pipeline

At the inference stage, we employ LamRA-Ret for the initial retrieval phase (Φ_{ret}) and LamRA-Rank for the subsequent reranking (Φ_{rank}). Given a query (q) and a candidate set (Ω), LamRA-Ret is first utilized to compute an embedding similarity score (S_{ret}) for each candidate based on cosine similarity. Subsequently, the candidates are sorted to obtain the top-K candidate set ($\mathcal{C}_1$). For the reranking process, we offer two options: pointwise reranking and listwise reranking. In the case of pointwise reranking, the query and each of the top-K candidates are sequentially input into the LMM, performing K inference operations and assigning a reranking score (S_{rank}) to each candidate. This reranking score comes from the probability of the LMMs outputting `YES` for each candidate. Conversely, for listwise reranking, the query and the top-K candidates are simultaneously fed into LamRA-Rank, which directly output the serial number of the most relevant candidate. More discussion on these two methods can be found in Section 4.3. In the final step, the embedding similarity score from LamRA-Ret and the reranking score from LamRA-Rank are aggregated into a weighted sum to get the final score: $S = \alpha \times S_{\text{ret}} + (1 - \alpha) \times S_{\text{rank}}$, where α is a weight hyperparameter. This final score is then used to generate the reordered candidate set ($\mathcal{C}_2$). We refer to the combined framework of LamRA-Ret and LamRA-Rank as **LamRA**.

4. Experiments

4.1. Experimental Setup

Datasets and Metrics. We utilize the NLI dataset [9] for pre-training and the M-BEIR [45] dataset for instruction tuning. The M-BEIR dataset encompasses eight retrieval tasks across 10 retrieval datasets, comprising a total of 1.1M training samples. As shown in Table 1, to evaluate the versatility of LamRA across various retrieval tasks, we conduct assessments on the M-BEIR test set. Furthermore, we investigate LamRA's generalization ability on other previously unseen datasets, including ShareGPT4V [52], CIRCO [2], and Visual Dialog [7], among others. We adhere to the standard evaluation metrics established for each dataset. For the retrieval tasks, we utilize Recall@K as the evaluation metric, and for the image-text matching task, we employ Accuracy as the evaluation metric.

Experiment Settings & Baselines. We establish three distinct experiment settings: (i) To validate the versatility of our method across a range of retrieval tasks, we train it on all 8 tasks in the M-BEIR benchmark and evaluate its performance on the test sets. For the baseline in this setting, we follow UniIR [45], comparing with several strong VLMs, such as CLIP, SigLIP, and BLIP2, to assess their zero-shot performance. In the supervised setting, we compare our method with fully fine-tuned versions of UniIR-BLIP and UniIR-CLIP. (ii) To evaluate the generalization ability on previously **unseen retrieval datasets**, we perform zero-shot experiments on 10 datasets not encountered during training. In this case, the baseline includes a selection of universal retrievers, such as E5-V, MagicLens, and EVA-CLIP-18B. (iii) To investigate the generalization capacity on **unseen retrieval tasks**, we intentionally exclude data from three retrieval tasks: image-to-image retrieval, text-image-to-text retrieval, and text-image-to-text-image retrieval. Training is then conducted on the remaining five tasks with the evaluation of these excluded tasks.

Benchmark	Zero-shot	# Queries	# Candidates
M-BEIR [45]	✘	190K	5.6M
ShareGPT4V [52]	✔	1K	1K
Urban-1K [52]	✔	1K	1K
Flickr30K [34]	✔	1K	5K
CIRCO [2]	✔	800	120K
GeneCIS [39]	✔	8K	$10 \sim 15$
Visual Storytelling [12]	✔	5K	8K
Visual Dialog [7]	✔	2K	2K
Multi-round FashionIQ [50]	✔	2.4K	6.2K
CC-Neg [37]	✔	40K	2
Sugar-Crepe [10]	✔	7.5K	2

Table 1. **Summary of the evaluation benchmarks.** # Queries represents the number of test queries, and # Candidates denotes the number of test candidates per query.

Implementation Details. Our framework is implemented in Pytorch and leverages the Qwen2-VL-7B [40] by default. In the pretraining stage for retrieval, we conduct experiments on 8 A100 GPUs with a batch size of 576 and a learning rate of 4×10^{-4}, training for two epochs. During the instruction tuning stage, we train on 16 A100 GPUs with a batch size of 960 and a learning rate of 1×10^{-4} for one epoch. For the rerank training stage, we train for one epoch on 16 A100 GPUs, employing a batch size of 64 and a learning rate of 2×10^{-5}. Throughout all experiments, the parameters on the vision side remain fixed, while the LLM is fine-tuned using LoRA. LamRA-Rank adopts the pointwise reranking by default. During the M-BEIR evaluation, experiments are conducted in the local pool, with LamRA-Rank reranking the top-50 results. For experiments on unseen datasets, reranking is applied to the top-10 results. The weight hyperparameter α is set to a default value of 0.5.

4.2. Experimental Results

Versatility Across Various Retrieval Tasks. As demonstrated in Table 2, our model is capable of handling retrieval tasks across various input formats. We can draw the following observations: (i) Our approach significantly

Methods	$q^t \to c^i$			$q^t \to c^t$	$q^t \to (c^i, c^t)$		$q^i \to c^t$			$q^i \to c^i$	$(q^i, q^t) \to c^t$		$(q^i, q^t) \to c^i$		$(q^i, q^t) \to (c^i, c^t)$		Avg.
	VN	COCO	F200K	WebQA	EDIS	WebQA	VN	COCO	F200K	NIGHTS	OVEN	InfoS	FIQ	CIRR	OVEN	InfoS	
	R@5	R@5	R@10	R@5	R@5	R@5	R@5	R@5	R@10	R@5	R@5	R@5	R@10	R@5	R@5	R@5	
Zero-shot																	
CLIP-L [35]	**43.3**	61.1	6.6	36.2	43.3	45.1	41.3	79.0	7.7	26.1	24.2	20.5	7.0	13.2	38.8	26.4	32.5
SigLIP [51]	30.1	75.7	**36.5**	39.8	27.0	43.5	30.8	88.2	**34.2**	28.9	29.7	25.1	14.4	22.7	41.7	27.4	37.2
BLIP [22]	16.4	74.4	15.9	44.9	26.8	20.3	17.2	83.2	19.9	27.4	16.1	10.2	2.3	10.6	27.4	16.6	26.8
BLIP2 [23]	16.7	63.8	14.0	38.6	26.9	24.5	15.0	80.0	14.2	25.4	12.2	5.5	4.4	11.8	27.3	15.8	24.8
Qwen2-VL-7B [40]	9.3	55.1	5.0	42.0	26.2	9.4	5.4	46.6	4.0	21.3	21.4	22.5	4.3	16.3	43.6	36.2	23.0
Supervised - Dual Encoder																	
UniIR-BLIP$_{FF}$ [45]	23.4	79.7	26.1	80.0	50.9	79.8	22.8	89.9	28.9	**33.0**	41.0	22.4	29.2	52.2	55.8	33.0	46.8
UniIR-CLIP$_{SF}$ [45]	42.6	81.1	18.0	84.7	59.4	78.7	**43.1**	**92.3**	18.3	32.0	45.5	27.9	24.4	44.6	67.6	48.9	50.6
Supervised - LMMs																	
LamRA-Ret	41.6	**81.5**	28.7	**86.0**	**62.6**	**81.2**	39.6	90.6	30.4	32.1	**54.1**	**52.1**	**33.2**	**53.1**	**76.2**	**63.3**	**56.6**
LamRA	**48.0**	**85.2**	**32.9**	**96.7**	**75.8**	**87.7**	**48.6**	**92.3**	**36.1**	**33.5**	**59.2**	**64.1**	**37.8**	**63.3**	**79.2**	**78.3**	**63.7**

Table 2. **Comparison with up-to-date state-of-the-arts on M-BEIR test set.** The first row indicates the retrieval task type: q^t for text queries, q^i for image queries, c^t for text candidates, and c^i for image candidates. Abbreviations used include VN for VisualNews, F200K for Fashion200K, InfoS for InfoSeek, and FIQ for FashionIQ. Evaluation standards follow UniIR, with FashionIQ and Fashion200K using Recall@10, while all other evaluations employ Recall@5. The best (resp. second-best) numbers are in red (resp. blue).

outperforms the dual-encoder paradigm method, UniIR-CLIP, in both multimodal and pure text retrieval tasks. For instance, in text-image-to-text retrieval on the InfoSeek dataset, our method LamRA-Ret, achieves an improvement of 24.2 points over UniIR-CLIP in Recall@5. Similarly, in composed image retrieval on the CIRR dataset, our method exceeds UniIR-CLIP by 8.5 points in Recall@5. (ii) In simpler retrieval tasks, such as cross-modal retrieval, the performance of LamRA-Ret is either comparable to or slightly higher than that of UniIR-CLIP. (iii) Furthermore, LamRA (with reranking) achieves additional performance gains across all retrieval tasks, with an average improvement of 7.1 points over 16 M-BEIR retrieval tasks. (iv) Notably, Qwen2-VL-7B achieves an average zero-shot score of only 23.0 on the M-BEIR benchmark, which increases to 63.7 after applying our framework. This highlights the effectiveness of our proposed LMMs framework in enhancing general retrieval and reranking performance.

Advanced Generalization Capability on Unseen Dataset. To validate the strong generalization capability of our method, we conduct extensive experiments across 10 unseen datasets. As illustrated in Table 4, we can make the following observations: (i) Our method demonstrates excellent generalization across various unseen datasets, with performance either significantly exceeding or comparable to other strong baseline methods. (ii) Notably, our method performs exceptionally well in scenarios where dual-encoder approaches tend to underperform. For instance, in the image-to-long-text retrieval task on the Urban-1K dataset, our LamRA-Ret achieves over a 10-point improvement compared to other methods. Similarly, in tasks such as dialog-to-image retrieval and multi-round composed image retrieval, our method also yields substantial performance gains. (iii) In Image-Text Matching tasks, LMMs-based methods outperform those using the dual-encoder paradigm. We attribute this improvement primarily to the LMM's exceptional ability to comprehend complex text.

Methods	$q^i \to c^i$	$(q^i, q^t) \to c^t$		$(q^i, q^t) \to (c^i, c^t)$		Avg.
	NIGHTS R@5	OVEN R@5	InfoS R@5	OVEN R@5	InfoS R@5	
Supervised						
UniIR-BLIP$_{FF}$ [45]	33.0	41.0	22.4	55.8	33.0	37.0
UniIR-CLIP$_{SF}$ [45]	32.0	45.5	27.9	67.6	48.9	44.4
Zero-shot						
LamRA-Ret*	27.2	44.7	44.0	62.8	49.5	45.6
LamRA*	29.2	46.9	54.2	65.1	59.1	50.9

Table 3. **Held-out task generalization experiments on M-BEIR.** * means that the training is carried out on the remaining five tasks, with no exposure to these three held-out tasks.

Exceptional Generalization on Unseen Retrieval Task. As outlined in Section 4.1, we train our model on five tasks and evaluate it on excluded tasks to assess its generalization on these previously unseen tasks. As demonstrated in Table 3, our approach shows impressive performance on unseen retrieval tasks. Notably, when compared to supervised methods such as UniIR-CLIP, our method achieves competitive or even superior zero-shot results in more complex tasks, such as text-image-to-text retrieval task on the InfoSeek benchmark, our method exceeds UniIR-CLIP by 26.3 points in Recall@5. This remarkable generalization ability suggests that our method can extrapolate to unseen retrieval tasks without additional training and may hold significant practical applicability across a broader range of scenarios.

4.3. Ablation Study & Analysis

In this section, we conduct experiments to further investigate the effect of our proposal, for example, the two-stage

Methods	$q^t \to c^i$			$q^i \to c^t$			$(q^i, q^t) \to c^i$		$q^{\text{dialog}} \to c^i$	$(q^i \oplus q^t) \to c^i$		ITM	
	Share4V	Urban*	Flickr	Share4V	Urban*	Flickr	CIRCO*	GeneCIS*	VisD*	VIST	MT-FIQ*	CC-Neg	Sugar-Crepe*
	R@1	R@1	R@1	R@1	R@1	R@1	MAP@5	R@1	R@1	R@1	R@5	Acc.	Acc.
CLIP-L [35]	84.0	52.8	67.3	81.8	68.7	87.2	4.0	13.3	23.7	0.6	17.7	66.7	73.0
Long-CLIP-L [52]	**95.6**	86.1	76.1	**95.8**	82.7	89.3	5.7	16.3	37.9	1.1	18.5	76.3	80.9
UniIR-CLIP [45]	85.8	75.0	78.7	84.1	78.4	94.2	12.5	16.8	26.8	0.6	39.4	79.9	80.3
E5-V [16]	86.7	84.0	79.5	84.0	82.4	88.2	24.8	18.5	54.6	10.0	19.2	**83.2**	84.7
MagicLens-L [54]	85.5	59.3	72.5	60.9	24.2	84.6	29.6	16.3	28.0	3.3	22.6	62.7	75.9
EVA-CLIP-8B [38]	91.2	77.8	80.8	93.1	80.4	95.6	6.0	13.1	23.2	1.2	22.1	59.4	81.7
EVA-CLIP-18B [38]	92.1	81.7	**83.3**	94.0	83.3	**96.7**	6.1	13.6	24.7	1.0	21.9	63.8	83.1
LamRA-Ret	93.3	**95.1**	82.8	88.1	**94.3**	92.7	**33.2**	**18.9**	**62.8**	**23.1**	**60.9**	79.6	**85.8**
LamRA	**97.9**	**98.8**	**88.1**	**96.5**	**98.0**	**97.6**	**42.8**	**24.8**	**70.9**	**28.6**	**63.9**	**85.9**	**93.5**

Table 4. **Comparison with up-to-date state-of-the-arts on unseen dataset.** The first row indicates the retrieval task type: q^t for text queries, q^i for image queries, q^{dialog} for dialog queries, $(q^i \oplus q^t)$ for multi-interleaved image-text queries, c^t for text candidates, c^i for image candidates, and ITM for the Image-Text Matching task. Abbreviations used include Share4V for ShareGPT4V, Urban for Urban-1k, VIST for Visual Storytelling, VisD for Visual Dialog, and MT-FIQ for Multi-round FashionIQ. $*$ indicates that the images in these datasets are sourced from COCO or FashionIQ. However, due to significant differences in their captions or query formats, we still classify these datasets as unseen. We adhere to the standard evaluation metrics specific to each dataset. The best (resp. second-best) numbers are in red (resp. blue). For more details about these datasets, please refer to Appendix D.

training for LamRA-Ret, the scaling trends of LamRA, *etc*.

Effectiveness of Two-stage Training. According to the results presented in Table 5, both the pre-training and the instruction tuning stages significantly impact retrieval performance. Specifically, omitting the pre-training stage results in an average performance decline of three points on the M-BEIR benchmark. Similarly, the absence of the instruction tuning stage yields suboptimal results, as the model struggles with some complex multimodal retrieval tasks. Collectively, these stages together enable a progressive enhancement in the retrieval performance of the LMMs.

Pre-training	Instruction tuning	Avg.
✘	✘	23.0 (-33.6)
✔	✘	36.2 (-20.4)
✘	✔	53.6 (-3.0)
✔	✔	56.6

Table 5. **Effect of the two-stage training.** Avg. refers to the average recall performance across the M-BEIR test set.

Scaling Trends of LamRA. As shown in Table 6, to investigate the scaling effect associated with LamRA, we conduct experiments on LMMs of different sizes, specifically Qwen2-VL-2B and Qwen2-VL-7B. The results indicate that as the model size scales, performance improves correspondingly. Currently, due to the computational limitations, we are able to conduct experiments solely on the 2B and 7B models; however, our findings suggest that larger and more powerful LMMs potentially yield even superior results within our framework, thereby laying the groundwork for further exploration in this area.

Discussion on Pointwise and Listwise Rerank Methods. As outlined in Section 3.4, LamRA-Rank is designed to perform both pointwise and listwise reranking. We investigate

LMMs	LamRA-Ret	LamRA-Rank	Avg.
Qwen2-VL-2B [40]	✔	✘	51.6
	✔	✔	58.3
Qwen2-VL-7B [40]	✔	✘	56.6
	✔	✔	63.7

Table 6. **Scaling Trends of LamRA.** Experimental results on the M-BEIR test set across various model sizes.

Task	LamRA-Ret	LamRA-Rank(P)		LamRA-Rank(L)	
	R@1	R@1	Time	R@1	Time
$q^t \to c^t$	58.2	75.9	0.020s	75.9	0.010s
$(q^i, q^t) \to c^i$	18.5	24.5	0.071s	24.3	0.067s
$(q^i, q^t) \to c^t$	30.1	37.3	0.047s	36.6	0.017s
$(q^i, q^t) \to (c^i, c^t)$	33.4	39.9	0.084s	39.5	0.085s

Table 7. **Comparison of Recall@1 performance and inference costs between pointwise (LamRA-Rank(P)) and listwise (LamRA-Rank(L)) reranking on M-BEIR.** Reranking is applied to the top-5 results. Time denotes the per-query inference time cost measured on eight A100 GPUs with a batch size of 32.

these two distinct reranking methods on the M-BEIR test set, with the results presented in Table 7. Both methods demonstrate a significant improvement in retrieval performance. Although pointwise reranking generally incurs a higher computational overhead than listwise reranking, the latter is constrained by the LLM's context length, which limits the number of candidates it can process. Consequently, each reranking method offers distinct advantages and can be flexibly applied in different contexts based on specific application requirements.

Extending to Video Retrieval. As shown in Table 8, we evaluate our method on the MSR-VTT [47] and MSVD [5] datasets in a zero-shot setting for the text-to-video retrieval

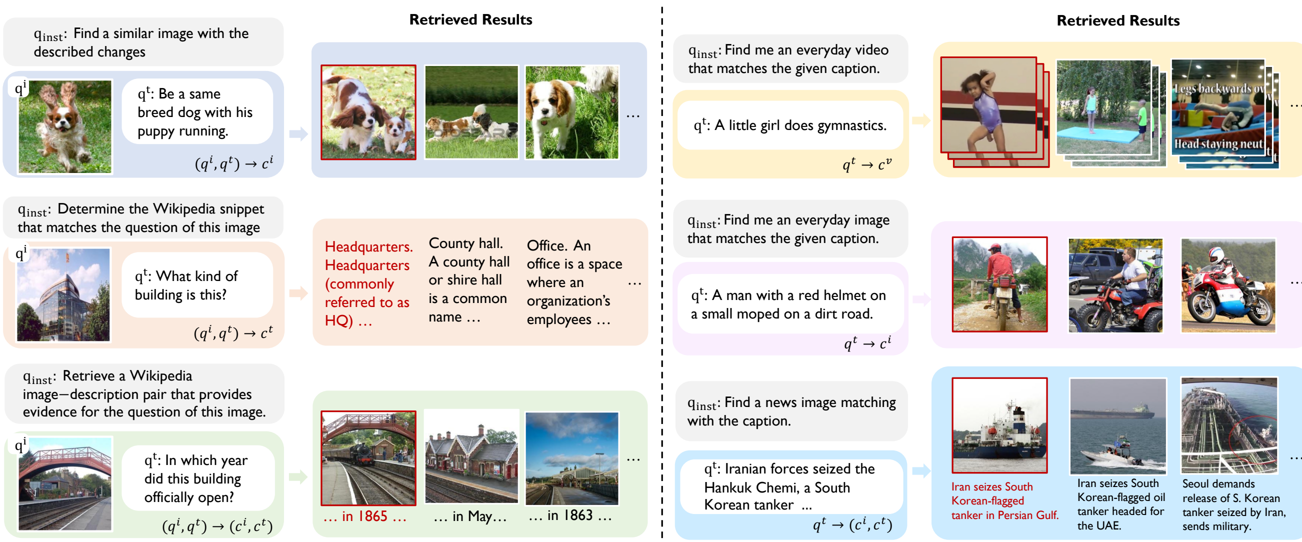

Figure 3. **Qualitative examples.** We show the results of our method across six different retrieval tasks, with the ground truth indicated by a red box and red text. Here, q^t for text queries, q^i for image queries, c^t for text candidates, c^i for image candidates, and c^v for video candidates. For more qualitative examples, please refer to Appendix H.

task. The results demonstrate that our approach exhibits competitive performance. For instance, on the MSR-VTT dataset, our method outperforms InternVideo [42] by 4.7 points in Recall@1. Similarly, on the MSVD dataset, we achieve an improvement of 2.5 points over UMT-L [24] in Recall@1. Notably, our method has not been exposed to any video data during the fine-tuning stage, indicating that it preserves Qwen2-VL's inherent video understanding capability. Although our approach currently falls short of the performance achieved by the state-of-the-art InternVideo2 [44], we intend to address this gap by exploring the integration of video data in future research.

Method	MSR-VTT			MSVD		
	R@1	R@5	R@10	R@1	R@5	R@10
InternVideo [42]	40.0	65.3	74.1	43.4	69.9	79.1
ViCLIP [43]	42.4	-	-	49.1	-	-
UMT-L [24]	42.6	64.4	73.1	49.9	77.7	85.3
InternVideo2$_{s2}$-6B [44]	55.9	78.3	85.1	59.3	84.4	89.6
LamRA	44.7	68.6	78.6	52.4	79.8	87.0

Table 8. **Zero-shot text-to-video retrieval performance.**

Discussion on Inference Cost. One potential drawback of applying LMMs to retrieval and reranking tasks is the high inference costs. Nevertheless, various strategies can be employed in practical applications to mitigate this issue. For instance, pre-extracting features for the candidate pool in advance or deploying our method in scenarios with low query-per-second requirements can significantly reduce inference costs. Moreover, several recent studies [18] focus on improving the inference speed of LMMs. We believe that leveraging LMMs for retrieval and reranking holds promising potential for future applications.

4.4. Qualitative Results.

In Figure 3, we present the qualitative results across six retrieval tasks. In the top left panel, we demonstrate our method's effectiveness in handling the composed image retrieval task, where the query requires retrieving an image of a dog of the same breed as shown in the reference image, along with its puppy running. Our method captures this complex intent, successfully retrieving the target image. In the third row of the first column, we showcase our method's performance on the text-image-to-text-image retrieval task, where it not only retrieves a Wikipedia snippet for the query but also identifies an image similar to the query image. For more qualitative examples, please refer to Appendix H.

5. Conclusion

In this paper, we have explored a novel paradigm for applying LMMs to address a broad range of retrieval tasks. By simply inserting lightweight LoRA modules, Our proposed universal framework, LamRA, endows LMMs with robust retrieval and reranking capabilities. Specifically, for efficient retrieval, we introduced LamRA-Ret, which adopts a two-stage training strategy to progressively enhance retrieval performance. For reranking, we presented LamRA-Rank, which supports both pointwise and listwise reranking to further optimize results. Comprehensive quantitative and qualitative comparisons across more than ten diverse retrieval tasks underscore the versatility and effectiveness of our method. We anticipate that our framework will provide valuable insights into the multimodal information retrieval field and inspire further research in this area.

References

[1] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[2] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the International Conference on Computer Vision*, 2023. 2, 5

[3] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. In *Proceedings of the European Conference on Computer Vision*, 2020. 2

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020. 3

[5] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Association for Computational Linguistics*, 2011. 7

[6] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the Conference on Empirical Methods in Natural Language Processinng*, 2023. 2

[7] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3

[9] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processinng*, 2021. 3, 4, 5

[10] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in Neural Information Processing Systems*, 2023. 2, 5

[11] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the International Conference on Computer Vision*, 2023. 2

[12] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2016. 5

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*, 2021. 2

[14] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 3

[15] Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models. *arXiv preprint arXiv:2307.16645*, 2023. 4

[16] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024. 3, 4, 7

[17] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024. 3

[18] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS Symposium on Operating Systems Principles*, 2023. 8

[19] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3

[20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2

[21] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 2

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, 2022. 6

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, 2023. 6

[24] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the International Conference on Computer Vision*, 2023. 8

[25] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*, 2024. 3

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 2014. 2

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. 2

[28] Yikun Liu, Jiangchao Yao, Ya Zhang, Yanfeng Wang, and Weidi Xie. Zero-shot composed text-image retrieval. In *Proceedings of the British Machine Vision Conference*, 2023. 2

[29] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the International Conference on Computer Vision*, 2021. 2

[30] Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the International Conference on Computer Vision*, 2023. 2

[31] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 4

[33] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, et al. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167*, 2023. 3

[34] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the International Conference on Computer Vision*, 2015. 2, 5

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021. 1, 2, 6, 7

[36] Jiayuan Rao, Haoning Wu, Chang Liu, Yanfeng Wang, and Weidi Xie. Matchtime: Towards automatic soccer game commentary generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processinng*, 2024. 3

[37] Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. Learn" no" to say" yes" better: Improving vision-language models via negations. In *Winter Conference on Applications of Computer Vision*, 2025. 5

[38] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024. 7

[39] Sagar Vaze, Nicolas Carion, and Ishan Misra. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 5

[40] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 4, 5, 6, 7

[41] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *Advances in Neural Information Processing Systems*, 2023. 3

[42] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 8

[43] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *Proceedings of the International Conference on Learning Representations*, 2024. 8

[44] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. In *Proceedings of the European Conference on Computer Vision*, 2024. 8

[45] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *Proceedings of the European Conference on Computer Vision*, 2024. 2, 4, 5, 6, 7

[46] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[47] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 7

[48] Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, Chun-Mei Feng, et al. Sentence-level prompts benefit composed image retrieval. In

Proceedings of the International Conference on Learning Representations, 2024. 2

[49] Yibin Yan and Weidi Xie. EchoSight: Advancing visual-language models with Wiki knowledge. In *Findings of the Association for Computational Linguistics: EMNLP*, 2024. 2

[50] Yifei Yuan and Wai Lam. Conversational fashion image retrieval via multiturn natural language feedback. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021. 5

[51] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the International Conference on Computer Vision*, 2023. 6

[52] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, 2024. 2, 5, 7

[53] Fei Zhang, Tianfei Zhou, Boyang Li, Hao He, Chaofan Ma, Tianjiao Zhang, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Uncovering prototypical knowledge for weakly open-vocabulary semantic segmentation. In *Advances in Neural Information Processing Systems*, 2023. 3

[54] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. In *Proceedings of the International Conference on Machine Learning*, 2024. 7

[55] Tianfei Zhou, Wang Xia, Fei Zhang, Boyu Chang, Wenguan Wang, Ye Yuan, Ender Konukoglu, and Daniel Cremers. Image segmentation in foundation model era: A survey. *arXiv preprint arXiv:2408.12957*, 2024. 2, 3