ELVA: Exploring Ranking-Driven Universal Multimodal Retrieval

Yuhan Liu[1†‡], Pei Fu[2†], Hang Li[2], Yukun Qi[2], Chao Jiang[2], Jingwen Fu[3§], Zhen Liu[1], Bin Qin[2], Zhenbo Luo[2], Jian Luan[2], and Jingmin Xin[1§]

[1] National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2] MiLM Plus, Xiaomi Inc
[3] Zhongguancun Academy, Beijing, China

Abstract. Leveraging Multimodal Large Language Models (MLLMs) via contrastive learning has become a mainstream paradigm for improving the performance of Universal Multimodal Retrieval (UMR). However, previous works have ignored the **grain blindness** when adapting the contrastive paradigm into retrieval tasks. Grain blindness refers to the tendency of the model to overlook grain-level information contained in the query, which is crucial for effectively handling complex queries. This stems from contrastive learning treating samples as a binary classification (positive/negative), while ignoring the different information carried by each negative sample. To address this, we argue that negatives should be treated differently according to their similarity to the positive sample, enabling the model to learn distinct grain information from each negative. In this paper, we introduce a simple but effective framework, called ELVA, a novel rule-based RL framework that mitigates grain blindness through ranking-driven MLLMs. 1) Instead of relying on reward models, we extend Reinforcement Learning with Verifiable Rewards (RLVR) to retrieval tasks, allowing the model to explore new ranking behaviors without explicit ranking labels. 2) By utilizing rule-based rewards, our approach jointly optimizes the ranking of negative samples while enlarging the similarity gap between positive and negative. To more precisely measure grain blindness, we further introduce MRBench, a new benchmark specifically designed for multi-grain query scenarios. ELVA achieves state-of-the-art results across standard retrieval benchmarks, and its notable 13.1% improvement on MRBench further demonstrates its effectiveness in alleviating grain blindness. Our code is available at `https://github.com/SeerRay-Lab/ELVA`.

1 Introduction

Universal Multimodal Retrieval (UMR) refers to a general retrieval paradigm that unifies diverse retrieval tasks within a single framework and enables generalization to unseen retrieval tasks [20, 35, 38, 79]. This marks a substantial shift

[†] Equal contribution.
[‡] Work done during the internship at Xiaomi.
[§] Corresponding author.

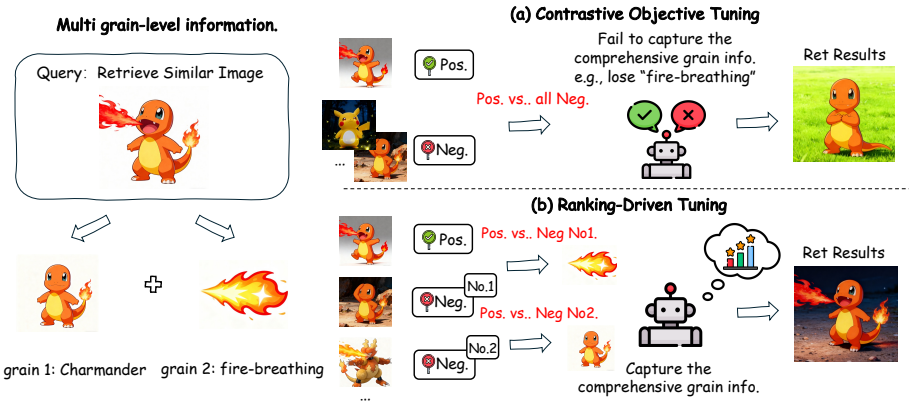

Fig. 1: The main idea of our proposed ELVA. Previous works [31, 35] fail on multi-grain queries due to grain blindness that emerges during contrastive training, as illustrated in (a). Build on its basis, our ELVA leverages the ranking-driven tuning with verifiable rewards to capture the comprehensive information, accurately retrieve the precise candidates shown in (b).

from prior efforts, which primarily focused on modality-specific retrieval tasks, including text-to-text [39,75], text-to-image [6,72], and image-to-image [3,47] retrieval. Recently, researchers have begun exploring Multimodal Large Language Models (MLLMs) [5, 23, 52, 63] for UMR, leveraging their extensive pretrained knowledge and strong generalization ability. Since MLLMs are originally trained for generative objectives (e.g., next-token prediction), recent works have adapted them to retrieval tasks via contrastive learning, effectively transferring their generative abilities to retrieval tasks [8, 30, 40, 74].

Although existing methods achieve impressive performance, our study reveals that they still suffer from *grain blindness* when adapting the contrastive paradigm to retrieval tasks. Grain blindness refers to the tendency of the model to overlook grain-level information contained in a query. Consequently, these methods struggle with multi-grain queries, as shown in Figure 1. In such a case, the query contains multiple levels of grain information, such as the action “fire-breathing" and the entity “Charmander", which place high demands on the model’s ability to comprehensively capture multi-grain information. To tackle the grain blindness, our study highlights two key challenges that need to be confronted.

Challenge I: *Which properties that the training paradigm have to equip in order to reduce the grain blindness?* Previous works [31, 35, 38] leverage contrastive objectives, learning embeddings by distinguishing positive and negative samples. However, as illustrated in Figure 1 (a), this training paradigm is suboptimal for retrieval tasks, as it fails to comprehensively capture the multiple levels of grain information contained in a query. Intuitively, the positive sample needs to be contrasted against diverse negative samples in order to learn distinct grain information. Yet, the contrastive paradigm treats all negatives equally [16,80], ignore the differential information carried by each negative, which

is crucial for accurate retrieval. To address this issue, we argue that the model should treat negative samples differently based on their similarity to the positive, which means negatives with higher similarity should be positioned closer to the positive. Toward this goal, we propose a ranking-driven approach that ranks candidates according to their relevance to the positive sample, enabling the model to capture comprehensive grain-level information, as shown in Figure 1 (b).

Challenge II: *How to incentivize the ranking ability without the ranking labels?* Previous retrieval models typically rely on supervised ranking objectives (e.g., listwise loss) to learn candidate ordering [9, 14, 28]. However, in UMR scenarios, obtaining ranking labels such as precisely ranking all negative samples by their exact relevance is extremely difficult and expensive. Meanwhile, forcing models to fit static labels severely restricts their capacity to explore subtle, grain-level hierarchical differences. To overcome this, we introduce an exploration-driven RL framework that operates without explicit ranking targets, utilizes the reward function as dynamic evaluators. Driven by the GRPO algorithm [48], the model autonomously discovers optimal inter-negative hierarchies by comparing the relative ranking quality of its variant generated candidate lists.

In this paper, we propose ***ELVA***: **E**xp**L**oring Ranking-Driven Uni**V**ersal Multimodal Retriev**A**l, a novel rule-based RL framework designed to overcome the grain blindness through ranking-driven MLLMs. Different from previous works that rely solely on contrastive objectives [35,38], ELVA adopts a ranking-driven tuning to capture richer grain-level information while simultaneously overcome the absence of ranking labels. Specifically, we propose two verifiable reward function to optimize the policy model: **1)** Ranking Reward, which encourages the model to rank candidates based on their relevance while rewarding the model for placing positive samples at higher ranks inspired by the NDCG [15]. Our Ranking Reward is a **continuous reformulation for RL**, ensuring continuous reward signal while optimizing negatives hierarchies; **2)** Margin Reward, which enforces explicit similarity-gap constraints, ensuring that positive samples remain closer to the query than negative ones. Additionally, we introduce a balanced negative sampling strategy to construct a ranking customized dataset for RL, filtering out excessively difficult negatives to ensure stable optimization.

For a more comprehensive evaluation, we construct a new benchmark, MRBench, derived from the M-BEIR dataset [56]. MRBench is specifically designed for multi-grain retrieval, where each query contains two or more grain-level attributes (e.g., an entity and an action), making it particularly challenging to preserve multi-grained information. Our method achieves a substantial 13.1% improvement in retrieval accuracy on this benchmark, demonstrating its effectiveness in mitigating grain blindness.

To summarize, we make the following contributions:

- We identify the issue of grain blindness when adapting the contrastive learning paradigm to retrieval tasks, and highlight two key challenges that need to be confronted.

- We propose ELVA to enable comprehensive multi-grain information acquisition by jointly optimizing ranking order and enforcing similarity-gap constraints.
- We construct a new dataset to evaluate model performance in complex multi-grain scenarios, and ELVA achieves state-of-the-art performance across diverse benchmarks including MRBench.

2 Related Work

Universal Multimodal Retrieval. Multimodal retrieval serves as the core task in information retrieval [2, 16, 45, 51], focusing on retrieving related content across diverse data modalities [22, 26]. As the landscape of information retrieval expands, more recent studies have shifted attention toward universal multimodal retrieval (UMR) [20, 35, 38, 79], where a unified model is capable of handling heterogeneous modalities and diverse retrieval tasks simultaneously. While earlier work in this domain often relied on small foundation models such as CLIP [57], recent advances [31, 35, 77] have demonstrated the promise of employing Multimodal Large Language Models (MLLMs) [24, 33, 43, 52, 59, 66, 78] to further enhance retrieval performance. As MLLMs are primarily trained with generative objectives, recent researches [21, 31, 35] adapt them for retrieval tasks via contrastive learning, utilize embeddings extracted from MLLMs performing similarity-based retrieval, leveraging their strong cross-modal representation capabilities. However, it remains a significant challenge to capture comprehensive grain-level information in order to retrieve complex queries with high precision. We propose a simple yet effective approach to incentivize the model's ranking ability, thus improving the UMR performance.

RL for Ranking Learning. Reinforcement learning (RL) has become a promising approach that enables models to adjust their behavior during training based on continuous feedback signals [37, 49, 68]. ReasonRank [34] optimizes discrete metric-based rewards (e.g., NDCG, Recall, RBO) defined over the entire ranked list. MM-R5 [65] introduces a position-weighted ranking reward, where each retrieved item receives a score according to its ranking position. These discrete rewards which means it only jumps when the ranking order changes, the optimization signal is discontinuous and high-variance, leading to unstable learning and poor convergence [62, 76]. In contrast, our reward enforces continuous feedback offering smoother gradients than purely discrete metric-based rewards. Search-R3 [10] incorporates continuous similarity scores, however the feedback easily saturates once the top-ranked positions converge, providing weak supervision for representation learning. In this paper, our ELVA not only maintains non-saturating learning signals via margin reward, but also models the intra-negative ranking structure via continuous ranking reward. Moreover, the ranking reward encourages negatives with high relevance ranked nearer to the positive, thereby enriching the grain-level information within the embedding space.

3 Method

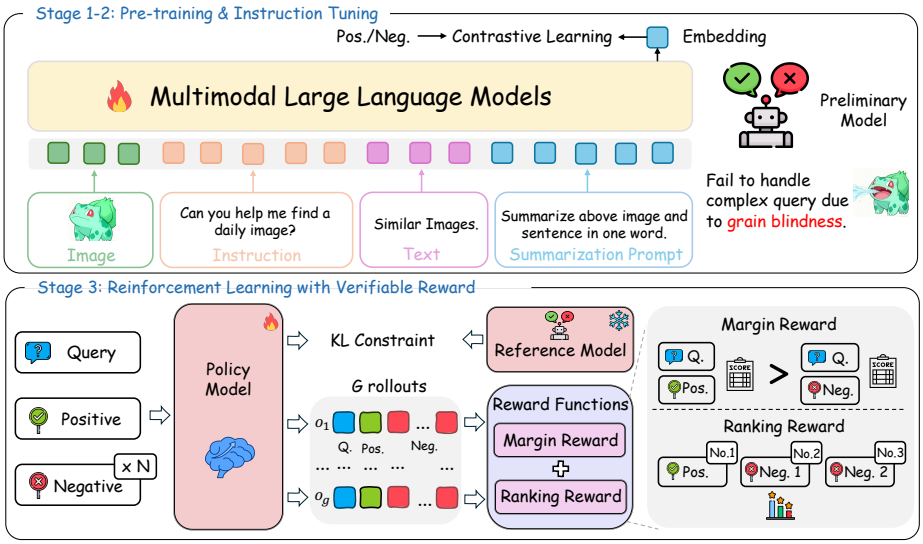

Fig. 2: Overview of the proposed ELVA framework. ELVA leverage the three-stage training framework for UMR tasks. Stage 1-2 as the pre-training and instruction tuning stage following [35], obtained the preliminary model struggle with the grain blindness. Stage 3 employ the RL tuning to incentivizing the ranking ability to address the issue. Given the input query q and candidates including pos. and N x neg., we first perform G rollouts to output G independent sets of embeddings from policy model. Then we compute the reward r_i for each output o_i using our proposed reward functions, detailed in Section 3.4. Finally, we optimize the policy model with GRPO [11] while ensuring that the model remains close to the reference policy model, via KL divergence.

3.1 Preliminary

This paper proposes a novel framework ELVA in Figure 2, to tackle the grain blindness in UMR. In this task, given a query q originating from any modality (text, image, or interleaved formats), the objective is to retrieve the most relevant sample from a set $\Omega = \{c_n\}_{n=1}^{N}$ with N candidates. To perform retrieval, we use a unified embedding extractor to encode both the query and candidates. Then we compute the cosine similarity s_i between the query and each candidate, and rank all candidates based on their similarity scores. The final retrieval output corresponds to the top-k candidates:

$$\mathcal{C} = \Phi_{\text{ret}}(q, \Omega) = \text{Top-}k\big(\{s_1, s_2, \ldots, s_N\}\big).$$

3.2 Formulation

To theoretically ground the concept of grain blindness, we formally define the mathematical formulation and this representational collapse to theoretically ground our framework.

Definition 1 (Grain and Query Composition). *A grain g is defined as the minimal, atomic semantic unit (e.g., an entity, an attribute, or an action) within a multimodal query. A query q is formulated as a set of these interdependent units, $q = \{g_1, g_2, \ldots, g_K\}$. Grains possess inherent structural dependencies; for instance, in the query "red dress," the attribute grain red is necessarily coupled with the entity grain dress to form a coherent search intent.*

Definition 2 (Grain Blindness). ***Grain blindness*** *is formalized as a representational collapse where the distance between the full query embedding and its de-grained version (where a specific grain g_k is removed) falls below a discriminative threshold δ:*

$$d(f_\theta(q), f_\theta(q \setminus \{g_k\})) < \delta, \tag{1}$$

where $f_\theta(\cdot)$ denotes the embedding function and $d(\cdot, \cdot)$ is a distance metric. This inequality indicates that the model fails to preserve the discriminative features of the specific grain g_k in the embedding space.

Theoretically, the emergence of grain blindness during contrastive learning can be attributed to Gradient Starvation [42, 46]. If certain dominant grains within q (e.g., a primary entity) provide a sufficient similarity margin to distinguish between positive and negative pairs, the contrastive loss drops rapidly. Consequently, other salient but secondary grains lose their necessary gradient signals for optimization. This premature convergence forces the model to ignore the suppressed grains, directly leading to the representational collapse described in Definition 2.

3.3 Pre-training & Instruction Tuning

To effectively leverage the contrastive learning paradigm in transforming the generative capability of MLLMs into discriminative representations, we first employ a two-stage training framework following [35, 38]. Since MLLMs are primarily trained for generative objectives such as next-token prediction, their inherent retrieval capability remains limited. The first stage conducts language-only pretraining on NLI datasets [7], enabling the generative model to produce more effective embeddings. The second stage, instruction tuning, further aligns the MLLMs with various retrieval tasks for better adaptability. These tasks include image-to-image retrieval, composed image retrieval, and image/question-to-multimodal-document retrieval, among others. Further details on the instruction templates and the M-BEIR datasets [56] are provided in the Suppl.

Training Objective. We adopt contrastive learning with the InfoNCE loss [41] as the optimization objective for both the language-only pretraining and instruction-tuning phases. During training, we input query q and instruction i into the model to obtain the query representation e_q. Similarly, each candidate c is fed into the model to derive its representation e_c. The training objective maximizes the similarity of positive samples and minimizes the similarity of negative samples, formulated as:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\log\left[\frac{\exp\left[\cos\left(e_q, e_c^+\right)/\tau\right]}{\sum_{m=1}^{N}\exp\left[\cos\left(e_q, e_{c_m}\right)/\tau\right]}\right]$$

where τ denotes the temperature parameter and N is the batch size. This approach enables the model to learns to discriminate between relevant and irrelevant information across various modalities.

3.4 ELVA

After the above training stages, we obtain a preliminary retrieval model. However, the model lacks sufficiently grain-level information, making it less effective for complex queries. To address this, we propose ELVA, a reinforcement learning based framework incentivizes the ranking ability of MLLMs for UMR. In contrast to conventional contrastive learning, ELVA optimizes model behavior through rule-based reward signals. However, a critical challenge in applying Policy Optimization to UMR tasks is that traditional EOL extraction [35, 38] (i.e., directly extracting the hidden state of a fixed prompt token [17]) yields no representational variance, rendering RL exploration impossible. To confront this, we formulate the feature extraction process as a **generative paradigm**. The model is mandated to first autoregressively generate a textual synthesis of the input, followed by a designated special token `[RET]`, which serves as the information bottleneck. We utilize the following prompt template:

Template for Generative Embeddings

USER: $\{I^{image}\}$ <Instruction> $\{I^{query}\}$.
Analyze and summarize the key information of the above input. Finally, append the special token `[RET]` to represent the entire input.
ASSISTANT: {`Generated Summary`} `[RET]`

Here, I^{image} denotes the image input, while I^{query} refers to the query text input. These modalities can be flexibly combined, and the corresponding instructions are adjusted accordingly. `[RET]` is a special token registered in the LLM, and we take the hidden state at this token's position as the retrieval embedding. In our RL framework, an action consists of generating embeddings for the query, positive, and all negatives by the policy model shown in the bottom of Figure 2. To facilitate policy optimization, the model generates G independent rollouts for a given query q and candidate set Ω through GRPO. This results in G groups

of embeddings, denoted as $\{\mathbf{e}_q^{(g)}, \mathbf{e}_{pos}^{(g)}, \{\mathbf{e}_{n,i}^{(g)}\}_{i=1}^{N}\}_{g=1}^{G}$, which serve as the basis for computing relative rewards.

Verifiable Reward Design. The reward model serves as a key component in reinforcement learning (RL), guiding the model's behavior to align with predefined correctness objectives. While conventional RL paradigms typically depend on human preference-based reward modeling [19, 32], recent advances such as DeepSeek-R1 [11] have shown that verifiable reward functions can substantially enhance reasoning capability. Building upon this insight, we extend Reinforcement Learning with Verifiable Rewards (RLVR) to the multimodal retrieval domain by developing a rule-driven, multi-criteria reward function that jointly evaluates ranking quality and distance between candidates. This design not only produces accurate retrieval results but also encourages the reasonable order among negative samples, thereby improving both robustness and interpretability. Our framework evaluates model output in terms of two complementary dimensions: **ranking orders and the distance between positive and negative**.

Margin Reward. The margin reward is designed to encourage the model to maintain a sufficient similarity gap between positive and negative samples, inspired by the triplet loss [12,36]. To achieve this, we compute the similarity scores between the query and all negative candidates, and select the hardest negative that has the highest similarity to the query. The reward then encourages the model to increase the similarity gap between the positive pair and the hardest negative. Given the input of query q and candidates $\Omega = \{c_p, c_n^0, \cdots, c_n^k\}$, where p and n represent positive samples and negative samples, the margin reward is defined as:

$$R_{\text{Margin}} = \max(0, \cos(q, c_p) - \cos(q, c_n)_{\max} - \delta),$$

where δ is a predefined hyperparameter that specifies the minimum required similarity gap. This formulation effectively drives the positive sample to rank at the top of the candidate list while imposing explicit similarity gap constraints. This reward enhances both retrieval precision and the discriminative structure of the learned representations.

Ranking Reward. We propose the ranking reward to explicitly encourage the model to rank the positive sample at the top of the candidate list while simultaneously promoting the ordering among negative samples to capture sufficient grained information. In contrast to the margin reward, which enforces a fixed similarity gap between positive and negative pairs, the ranking reward introduces rank-dependent weighting to optimize both rank precision and inter-negative structure.

Given a query and a candidate set, to translate these candidates into a ranked list, we compute cosine similarity scores $s_i = \cos(q, c_i)$ between the query and candidates within each specific rollout. These scores are then sorted to form the ranking used to calculate the reward. The reward is defined as:

$$R_{\text{Rank}} = s_{(r)} \cdot \frac{1}{1 + \log r} \; - \; \gamma \sum_{k \neq r} s_{(k)} \cdot (\log k - 1),$$

where $s_{(r)}$ denotes the similarity of the positive sample ranked at position r, k denotes the rank of negatives and γ controls the penalty strength for high-scoring negatives. The first term encourages the model to assign a higher similarity score to the positive sample and place it near the top of the ranking. The second term encourages more similar negative samples to be ranked closer to the top, thereby receiving smaller penalties, while still constraining their similarity to the query.

Moreover, R_{Rank} is a continuous reformulation designed for RL, ensuring smoother reward landscape while optimizing negatives hierarchies. This rank-aware formulation enhances retrieval quality by encouraging precise rank positioning and structured ranking among negatives, thereby improving the model's ability to learn adequate grained, discriminative representations.

Final Reward Function. The total reward combines the above rewards to optimize both ranking quality and similarity gap:

$$R_{\text{total}} = \alpha R_{\text{Margin}} + \varepsilon R_{\text{Ranking}},$$

where hyperparameters α and ε balance reward contributions and avoid reward hacking. This joint optimization enables the model to produce more precise retrieval outcomes by effectively capturing rich grained information, particularly in the context of complex or compositional queries.

Negative Sampling Strategy We introduce a balanced negative sampling strategy tailored for the RL stage. Prior work [35] directly sampled the top-100 candidates for SFT, which we find problematic for RL: when candidates are overly similar to the query, the reward distribution becomes too narrow, yielding weak or vanishing gradients [67] and hindering effective policy learning. To address this, we construct each query's candidate set using a balanced mix of negatives: (1) 50 filtered hard negatives, obtained by removing candidates above a similarity threshold and selecting the top 50 remaining ones [8, 29]; and (2) 50 randomly sampled negatives from the full pool. This combination increases reward variance and provides richer learning signals: the filtered subset offers controlled difficulty to enhance ranking capability, while the random subset adds distributional diversity and prevents overfitting. Overall, this mixed sampling strategy yields a more stable and informative signal for RL optimization.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We employ the NLI dataset [7] for pre-training and the M-BEIR dataset [56] for instruction tuning, following [35,38]. M-BEIR spans eight retrieval tasks across ten datasets, containing approximately 1.1M training instances. For the RL stage, we apply our negative sampling strategy to construct a training set of 11k instances from M-BEIR by sampling 1% of data from each dataset. We evaluate ELVA on the M-BEIR test set to assess its versatility

Table 1: Comparison with recent state-of-the-art methods on the M-BEIR test set. The first row denotes the retrieval task configuration, where q^t and q^i represent text and image queries, respectively, and c^t and c^i denote text and image candidates. Dataset abbreviations include VN for VisualNews, F200K for Fashion200K, InfoS for InfoSeek, and FIQ for FashionIQ. Following the UniIR [56] evaluation protocol, Recall@10 is reported for FashionIQ and Fashion200K, while Recall@5 is used for all other datasets. The best results are highlighted.

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 [45]	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 [69]	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 [25]	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 [24]	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 [52]	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
Qwen2.5-VL-7B [1]	40.2	71.9	20.3	71.9	49.4	64.5	29.3	84.6	19.4	25.5	42.4	32.1	25.0	55.1	60.8	54.9	46.7
Supervised - Dual Encoder																	
UniIR-BLIP$_{FF}$ [57]	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$_{SF}$ [57]	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 - MLLMs																	
Vision-R1-7B [13]	41.9	75.0	22.0	70.6	51.3	69.1	35.4	85.1	22.4	25.9	48.8	44.0	29.2	57.7	66.2	59.0	50.2
VLM-R1-7B [49]	40.5	77.2	22.5	72.3	50.0	67.9	36.2	86.3	20.9	26.4	48.8	37.5	29.9	57.4	64.0	62.3	50.0
MM-Embed-7B [31]	41.0	71.3	17.1	**95.9**	**68.8**	85.0	41.3	90.1	18.4	32.4	42.1	42.3	25.7	50.0	64.1	57.7	52.7
PUMA-3B [38]	35.7	79.5	25.8	86.2	58.2	78.4	35.2	90.1	29.0	31.4	52.7	48.3	30.6	49.9	74.0	65.2	54.4
LamRA-Ret-2B [35]	30.8	78.8	23.1	82.5	54.3	77.8	31.2	88.5	27.1	28.7	51.1	44.2	28.9	47.7	72.3	60.8	51.6
LamRA-Ret-7B [35]	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
ELVA-2B (Ours)	35.6	80.3	25.0	88.0	56.1	80.5	33.4	90.2	25.9	29.3	52.0	47.4	30.9	50.0	72.8	61.3	53.8$_{+4.3\%}$
ELVA-7B (Ours)	**43.5**	**83.0**	**29.2**	91.0	63.5	**83.1**	**41.7**	**92.2**	**32.1**	**32.8**	**56.0**	**55.5**	**34.6**	**55.4**	**77.5**	**67.1**	**58.7**$_{+3.9\%}$

across diverse retrieval scenarios. To further examine its generalization ability, we also evaluate ELVA on several unseen datasets [2, 70]. For multi-grain scenarios, we introduce a new benchmark, MRBench (Multi-gRain Benchmark), derived from M-BEIR. We first employ Qwen2.5-VL-7B [1] to automatically identify and filter queries containing at least two grain-level attributes, followed by human sampling verification to ensure data quality. Finally, we sample an equal number of instances from each task, resulting in a benchmark of 1k queries across 3 datasets and 4 retrieval tasks. We follow standard evaluation protocols for all datasets, using Recall@K as the primary metric for retrieval tasks.

Implementation Details. Our framework is implemented in PyTorch, by default, built upon Qwen2-VL-7B [52]. During the retrieval pretraining stage, experiments are conducted on 8 × H 20 GPUs with a batch size of 576, a learning rate of 4×10^{-5}, and trained for two epochs (3h completed). In the instruction tuning stage, we use 16 × H 20 GPUs with a batch size of 960 and a learning rate of 1×10^{-4} for one epoch following [35] (48h completed). For the RL stage, training is performed for one epoch on 8 × H 20 GPUs with a learning rate of 1×10^{-6}, using 8 rollouts and $\beta = 0.2$ (16h completed). Across all stages, the vision encoder remains frozen, while the language model is fine-tuned using LoRA. During M-BEIR evaluation, we conduct experiments in a local retrieval pool. The weight hyperparameter set to $\alpha = 0.4$ and $\varepsilon = 0.6$.

Table 2: Experimental results on unseen datasets. The first row denotes the type of retrieval task: q^t represents text queries, q^i image queries, q^{dialog} dialog-based queries, and $(q^i \oplus q^t)$ denotes interleaved image-text queries; c^t and c^i correspond to text and image candidates, respectively, while ITM refers to the Image-Text Matching task. Dataset abbreviations include Share4V for ShareGPT4V, Urban for Urban-1k, VisD for Visual Dialog, and MT-FIQ for Multi-round FashionIQ. The * symbol indicates that images in these datasets originate from COCO or FashionIQ; however, due to notable differences in captions and query structures, they are still treated as unseen datasets. We follow the standard evaluation metrics defined for each dataset, and the best-performing results are highlighted.

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*	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@5	Acc.	Acc.
CLIP-L [45]	84.0	52.8	67.3	81.8	68.7	87.2	4.0	13.3	23.7	17.7	66.7	73.0
Long-CLIP-L [70]	95.6	86.1	76.1	**95.8**	82.7	89.3	5.7	16.3	37.9	18.5	76.3	80.9
UniIR-CLIP [56]	85.8	75.0	78.7	84.1	78.4	94.2	12.5	16.8	26.8	39.4	79.9	80.3
E5-V [18]	86.7	84.0	79.5	84.0	82.4	88.2	24.8	18.5	54.6	19.2	83.2	84.7
MagicLens-L [71]	85.5	59.3	72.5	60.9	24.2	84.6	29.6	16.3	28.0	22.6	62.7	75.9
EVA-CLIP-8B [50]	91.2	77.8	80.8	93.1	80.4	95.6	6.0	13.1	23.2	22.1	59.4	81.7
EVA-CLIP-18B [50]	92.1	81.7	83.3	94.0	83.3	**96.7**	6.1	13.6	24.7	21.9	63.8	83.1
LamRA-Ret-7B [35]	93.3	95.1	82.8	88.1	94.3	92.7	33.2	18.9	62.8	60.9	79.6	85.8
ELVA-7B (Ours)	**96.6**	**96.1**	**84.4**	92.0	**95.5**	95.2	**34.5**	**20.2**	**65.3**	**61.2**	**87.3**	**91.1**

4.2 Experimental Results

Comparison of Effectiveness. We begin by evaluating the effectiveness of ELVA on the M-BEIR test set. Table 1 reports results in terms of Recall@K, covering 16 sub-tasks across 8 combinations of query and candidate modalities. To examine scalability, we present results for both ELVA-2B and ELVA-7B. We compare against three categories of methods: 1) Zero-shot general-purpose MLLMs, including BLIP-2 [24] and Qwen-VL [52]; 2) prior RL-based MLLMs, such as Vision-R1 [13] and VLM-R1 [49]; and 3) Retrieval-specialized MLLMs, including MM-Embed [31] and LamRA [35]. As shown in Table 1, ELVA consistently achieves state-of-the-art (SOTA) results across most settings. Notably, 1) even the 2B variant surpasses larger models such as MM-Embed-7B on most sub-tasks, and 2) on particularly challenging configurations like $(q^i, q^t) \to (c^i, c^t)$ on the InfoS dataset with 6.0% improvement, ELVA attains a substantial performance lead over previous methods. These results demonstrate the robustness and universality of ELVA, highlighting its strong retrieval capability across diverse multimodal inputs. We further apply LamRA-Rank in the reranking stage to boost accuracy shown in Suppl.

Comparison of Generalization on Unseen Dataset. To assess the generalization capability of our approach, we conduct extensive experiments on multiple ***unseen retrieval datasets***. As shown in Table 2, ELVA consistently delivers strong performance across all evaluation settings, demonstrating robust generalization to diverse data modalities and task types. In ITM tasks, our ELVA achieves over a 9.7% improvement to other methods. Similarly, in fixed-modal retrieval tasks such as text-to-image, our method also achieves substantial improvements in performance. These findings underscore the strong adaptability and scalability of ELVA, highlighting its promise as a unified framework for broader multimodal applications.

Table 3: Experimental results on held-out tasks. * indicates training on other tasks without exposure to the three held-out tasks.

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$_{FF}$	**33.0**	41.0	22.4	55.8	33.0	37.0
UniIR-CLIP$_{SF}$	32.0	45.5	27.9	**67.6**	48.9	44.4
Zero-shot						
Qwen2.5-VL	20.3	38.5	40.4	53.6	44.9	39.5
Vision-R1	22.9	39.8	42.9	57.4	46.5	41.9
LamRA-Ret*	27.2	44.7	44.0	62.8	49.5	45.6
ELVA-7B*	28.2	**46.5**	**49.2**	64.4	**53.0**	**48.3**

Table 4: Comparison of Reward Weighting.

Method	VN	COCO	F200K	Avg.
$\alpha = 0.6$, $\varepsilon = 0.4$	42.9	82.2	28.8	58.2
$\alpha = 0.5$, $\varepsilon = 0.5$	43.2	82.5	29.1	58.4
$\alpha = 0.4$, $\varepsilon = 0.6$ (ELVA)	**43.5**	**83.0**	**29.2**	**58.7**

Table 5: Generalizability of our ranking-driven RL framework.

Method	M-BEIR	MRBench
PUMA [38]	54.4	35.1
PUMA+ELVA	**56.3**	**37.0**
MM-Embed [31]	52.7	34.6
MM-Embed+ELVA	**54.9**	**36.2**

Table 6: Experimental results on MRBench datasets. The * symbol indicates that dataset are filtered for the multi-grain scene.

Methods	$q^i \to c^i$	$q^i \to c^t$	$q^i \to c^i$	$(q^i, q^t) \to c^i$	Avg.
	COCO* R@5	COCO* R@5	NIGHTS* R@5	CIRR* R@5	
Qwen2-VL-7B [52]	39.2	32.0	15.6	10.8	34.3
LamRA-Ret-7B [35]	50.4	54.0	22.4	25.7	38.1
ELVA-7B (Ours)	**55.5**	**60.6**	**24.1**	**32.5**	**43.2**

Table 7: Zero-shot text-to-video retrieval performance.

Method	MSR-VTT			MSVD		
	R@1	R@5	R@10	R@1	R@5	R@10
InternVideo [55]	40.0	65.3	74.1	43.4	69.9	79.1
ViCLIP [53]	42.4	-	-	49.1	-	-
UMT-L [27]	42.6	64.4	73.1	49.9	77.7	85.3
InternVideo2$_{s2}$-6B [54]	55.9	78.3	85.1	59.3	84.4	89.6
LamRA-7B [35]	44.7	68.6	78.6	52.4	79.8	87.0
ELVA-7B (Ours)	**46.4**	**70.5**	**78.9**	**53.9**	**80.7**	**87.9**

Comparison of Generalization on Unseen Task. Evaluate ELVA on ***unseen retrieval tasks*** by excluding specific tasks during training and testing the retrained model on these omitted tasks. As shown in Table 3, our method exhibits strong performance on unseen retrieval tasks. At the same parameter scale, our ELVA achieves 5.9% improvement over previous SOTA method. This strong generalization capability indicates that our method can effectively extend to unseen retrieval tasks without further training, showing great potential for broader real-world applications.

Comparison of Multi-Grain scene on MRBench. Table 6 reports the results on the MRBench benchmark. ELVA achieves superior performance compared to previous methods SOTA LamRA with 13.1% improvements and zero-shot models, highlighting the effectiveness of our approach in accurately retrieving queries with multi-grain information. As illustrated in Figure 3, when a query includes multiple grain-level attributes, such as "snow-capped mountains" and "trees", existing methods frequently fail to retrieve the correct result due to their inability to capture all grain components adequately. The qualitative comparisons show that our method captures the complex intent of the query and successfully retrieves the desired target. These results further confirm the effectiveness of our proposed approach in substantially alleviating the grain blindness problem. For more qualitative examples, please refer to Suppl.

4.3 Ablation Study

Ablating Reward Functions. To evaluate the effect of reward functions, we ablate ranking rewards and margin rewards, analyzing their effect across M-

Table 8: Ablation study. Avg. refers to the average recall performance across the M-BEIR test set. We select the image-to-text retrieval tasks as example.

#	Method	VN	COCO	F200K	Avg.
1	w/o Ranking Rewards	41.7 ↓1.8	82.0 ↓1.0	28.4 ↓0.8	57.2 ↓1.5
2	w/o Margin Rewards	42.3 ↓1.2	82.2 ↓0.8	28.5 ↓0.7	58.1 ↓0.6
3	w/o Negative Ranking	42.8 ↓0.7	82.0 ↓1.0	28.5 ↓0.7	58.1 ↓0.6
4	w/o Negative Sampling Strategy	43.1 ↓0.4	82.2 ↓0.8	28.5 ↓0.7	58.2 ↓0.5
5	w/o Randomly Sampled Negatives	43.3 ↓0.2	82.6 ↓0.4	28.9 ↓0.3	58.4 ↓0.3
6	ELVA-7B (Ours)	**43.5**	**83.0**	**29.2**	**58.7**

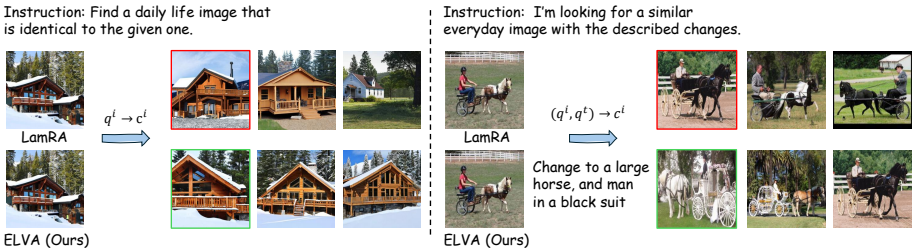

Fig. 3: Qualitative examples. We show the results of our method across different retrieval tasks, with the correct result indicated by the green box. Here, q^t for text queries, q^i for image queries, c^i for image candidates.

BEIR test set, as shown in Table 8. Excluding the ranking rewards leads to noticeable performance drops, highlighting its importance for enhancing the ranking quality, in order to learn sufficient grain information. Removing the margin rewards also reduces performance, indicating its role in ensuring the similarity gap for precise retrieval. In addition, we conduct another ablation in ranking reward, removing the second term designed for negative rank in row 3. The performance indicates that optimizing the negative ranking leads to more accurate retrieval. We conduct more ablation studies about hyperparameters in Suppl.

Ablating Negative Sampling Strategy. To evaluate the impact of the negative sampling strategy, we analyze different sampling strategies. As shown in Table 8, removing the sampling strategy leads to performance degradation, suggesting its importance in maintaining training data moderate difficulty in row 4. Introducing a random subset further enhances distributional diversity and prevents the model from overfitting in row 5. The results indicate that a properly training dataset is essential for model convergence, ensuring robust performance across datasets. More comparison results are shown in Suppl.

Ablating Reward Weighting. Additionally, we examine how different weighting schemes between the margin-based reward and the ranking-based reward affect the final performance in Table 4. This observation indicates that while the margin reward provides effective local pairwise guidance, it is the ranking reward that captures more holistic ordering signals and aligns more strongly with the final evaluation metric (Recall/K). Therefore, assigning a slightly higher weight to the ranking reward allows the model to better optimize global ranking behavior without losing the corrective constraints introduced by the margin term.

4.4 Deep Analysis

Post-Training Extension. Beyond the complete three stages pipeline, our proposed RL paradigm (Stage 3) functions as a highly adaptable, modular enhancement for existing multimodal retrievers. Future works can bypass the supervised fine-tuning stages and directly apply our ranking-driven RL to off-the-shelf models to achieve consistent performance gains. As shown in Table 5 integrating S3 as a post-training step significantly improves the average performance of PUMA [38] and MM-Embed [31]. This demonstrates our RL framework to be a universal "plug-and-play" booster that seamlessly scales to various multimodal retrieval architectures.

Extending to Video Retrieval. As presented in Table 7, we evaluate our method on the MSR-VTT [64] and MSVD [4] datasets under a zero-shot text-to-video retrieval setting. The results show that our approach achieves strong performance. For example, on MSR-VTT, our model achieves 16.5% improvements over InternVideo [55] , while on MSVD, it outperforms UMT-L [27] by 8.4%. It is noteworthy that our model has not been exposed to any video data during fine-tuning, yet it still retains Qwen2-VL's inherent video understanding ability. Although the current performance remains below the state-of-the-art InternVideo2 [54], we plan to incorporate video data in future work to further narrow this gap [44, 58].

Empirical and Qualitative Analysis of the Embedding Space. To validate the mitigation of grain blindness (Def. 2), we measure the representation distance $d(f_\theta(q), f_\theta(q \setminus \{g_k\}))$ on 100 sampled MRBench queries. By systematically masking a single phrase-level grain (e.g., dropping 'standing" from 'standing dog"), we compute the average cosine distance between full and masked query embeddings. The baseline yields a collapsed distance of 0.07, indicating the dropped grain was largely ignored. Conversely, ELVA significantly widens this gap to 0.15, quantitatively confirming its preservation of grain-level semantics. Furthermore, t-SNE visualization in Figure 4 of the query 'Standing Dog", shows that while the baseline entangles positives with hard negatives (e.g., 'Lying Dog" or "Standing Cat"), ELVA clearly separates them. Together, these results demonstrate ELVA's ability to prevent granularity loss and construct a highly discriminative embedding space.

Fig. 4: Distribution of embeddings from Baseline (left) and ELVA (right).

5 Limitations and Future Work

The limitation of MLLM-based retrieval is high inference costs. This overhead can be further mitigated via feature precomputation, layer pruning [38], efficient MLLM designs [60, 61, 73], or deploying the lightweight ELVA-2B. Future

works includes exploring joint retrieval-reranking frameworks to reduce pipeline complexity and scaling to larger MLLMs.

6 Conclusion

In this paper, we present ELVA, a novel framework designed to address the grain blindness when adapting the MLLMs via contrastive paradigm to Universal Multimodal Retrieval (UMR). We identify two challenges for addressing the grain blindness: 1) training paradigm's properties impact on retrieval; 2) how to incentivize the new ability without the ranking labels. We also introduce a new benchmark specifically designed to evaluate the grain blindness mitigation. Extensive experiments demonstrate that ELVA achieves significant gains and effectively mitigates grain blindness. We expect that our framework will offer meaningful guidance for advancing multimodal information retrieval and encourage continued exploration in this domain.

Acknowledgement

This work is supported by the Zhongguancun Academy (Grant No. XTS0070).

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 10
2. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15338–15347 (2023) 4, 10
3. Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A.: Effective conditioned and composed image retrieval combining clip-based features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21466–21474 (2022) 2
4. Chen, D., Dolan, W.B.: Collecting highly parallel data for paraphrase evaluation (2011) 14
5. Dong, H., Kang, Z., Yin, W., Liang, X., Feng, C., Ran, J.: Scalable vision language model training via high quality data curation. arXiv preprint arXiv:2501.05952 (2025) 2
6. Fu, Z., Zhang, L., Xia, H., Mao, Z.: Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26307–26316 (2024) 2
7. Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings (2021) 6, 9
8. Gu, T., Yang, K., Feng, Z., Wang, X., Zhang, Y., Long, D., Chen, Y., Cai, W., Deng, J.: Breaking the modality barrier: Universal embedding learning with multimodal llms. arXiv preprint arXiv:2504.17432 (2025) 2, 9

9. Gu, T., Yang, K., Zhang, K., An, X., Feng, Z., Zhang, Y., Cai, W., Deng, J., Bing, L.: Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning. arXiv preprint arXiv:2510.13515 (2025) 3
10. Gui, Y., Cheng, J.: Search-r3: Unifying reasoning and embedding generation in large language models. arXiv preprint arXiv:2510.07048 (2025) 4
11. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025) 5, 8
12. Hong, M., Lu, Y., Ye, N., Lin, C., Zhao, Q., Liu, S.: Unsupervised homography estimation with coplanarity-aware gan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17663–17672 (2022) 8
13. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025) 10, 11
14. Huang, X., Peng, H., Zou, D., Liu, Z., Li, J., Liu, K., Wu, J., Su, J., Yu, P.S.: Cosent: Consistent sentence embedding via similarity ranking. IEEE/ACM Transactions on Audio, Speech, and Language Processing **32**, 2800–2813 (2024) 3
15. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems (TOIS) **20**(4), 422–446 (2002) 3
16. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021) 2, 4
17. Jiang, T., Huang, S., Luan, Z., Wang, D., Zhuang, F.: Scaling sentence embeddings with large language models. arXiv preprint arXiv:2307.16645 (2023) 7
18. Jiang, T., Song, M., Zhang, Z., Huang, H., Deng, W., Sun, F., Zhang, Q., Wang, D., Zhuang, F.: E5-v: Universal embeddings with multimodal large language models. arXiv preprint arXiv:2407.12580 (2024) 11
19. Kaufmann, T., Weng, P., Bengs, V., Hüllermeier, E.: A survey of reinforcement learning from human feedback (2024) 8
20. Kong, F., Zhang, J., Liu, Y., Zhang, H., Feng, S., Yang, X., Wang, D., Tian, Y., Zhang, F., Zhou, G., et al.: Modality curation: Building universal embeddings for advanced multimodal information retrieval. arXiv preprint arXiv:2505.19650 (2025) 1, 4
21. Lan, Z., Niu, L., Meng, F., Zhou, J., Su, J.: Llave: Large language and vision embedding models with hardness-weighted contrastive learning. arXiv preprint arXiv:2503.04812 (2025) 4
22. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European conference on computer vision (ECCV). pp. 201–216 (2018) 4
23. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024) 2
24. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023) 4, 10, 11
25. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022) 10

26. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021) 4
27. Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L., Qiao, Y.: Unmasked teacher: Towards training-efficient video foundation models. In: ICCV (2023) 12, 14
28. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., Zhou, J., Lin, J.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv (2026) 3
29. Li, X., Li, C., Chen, S.Z., Chen, X.: U-marvel: Unveiling key factors for universal multimodal retrieval via embedding learning with mllms. arXiv preprint arXiv:2507.14902 (2025) 9
30. Lin, L., Long, J., Wan, Z., Wang, Y., Yang, D., Yang, S., Yao, Y., Chen, X., Guo, Z., Li, S., et al.: Sail-embedding technical report: Omni-modal embedding foundation model. arXiv preprint arXiv:2510.12709 (2025) 2
31. Lin, S.C., Lee, C., Shoeybi, M., Lin, J., Catanzaro, B., Ping, W.: Mm-embed: Universal multimodal retrieval with multimodal llms. arXiv preprint arXiv:2411.02571 (2024) 2, 4, 10, 11, 12, 14
32. Liu, C.Y., Zeng, L., Liu, J., Yan, R., He, J., Wang, C., Yan, S., Liu, Y., Zhou, Y.: Skywork-reward: Bag of tricks for reward modeling in llms. arXiv preprint arXiv:2410.18451 (2024) 8
33. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023) 4
34. Liu, W., Ma, X., Sun, W., Zhu, Y., Li, Y., Yin, D., Dou, Z.: Reasonrank: Empowering passage ranking with strong reasoning ability. arXiv preprint arXiv:2508.07050 (2025) 4
35. Liu, Y., Zhang, Y., Cai, J., Jiang, X., Hu, Y., Yao, J., Wang, Y., Xie, W.: Lamra: Large multimodal model as your advanced retrieval assistant. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4015–4025 (2025) 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12
36. Liu, Y., Huang, Q., Hui, S., Fu, J., Zhou, S., Wu, K., Li, P., Wang, J.: Semantic-aware representation learning for homography estimation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 2506–2514 (2024) 8
37. Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785 (2025) 4
38. Lyu, Y., Shao, R., Chen, G., Zhu, Y., Guan, W., Nie, L.: Puma: Layer-pruned language model for efficient unified multimodal retrieval with modality-adaptive learning. arXiv preprint arXiv:2507.08064 (2025) 1, 2, 3, 4, 6, 7, 9, 10, 12, 14
39. Ma, X., Wang, L., Yang, N., Wei, F., Lin, J.: Fine-tuning llama for multi-stage text retrieval. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2421–2425 (2024) 2
40. Meng, R., Jiang, Z., Liu, Y., Su, M., Yang, X., Fu, Y., Qin, C., Chen, Z., Xu, R., Xiong, C., et al.: Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. arXiv preprint arXiv:2507.04590 (2025) 2
41. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 7
42. Pezeshki, M., Kaba, O., Bengio, Y., Courville, A.C., Precup, D., Lajoie, G.: Gradient starvation: A learning proclivity in neural networks. Advances in Neural Information Processing Systems **34**, 1256–1272 (2021) 6
43. Qi, Y., Fu, P., Li, H., Liu, Y., Jiang, C., Qin, B., Luo, Z., Luan, J.: Patchcue: Enhancing vision-language model reasoning with patch-based visual cues. arXiv preprint arXiv:2603.05869 (2026) 4

44. Qi, Y., Zhao, Y., Zeng, Y., Bao, X., Huang, W., Chen, L., Chen, Z., Zhao, J., Qi, Z., Zhao, F.: Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. arXiv preprint arXiv:2504.07956 (2025) 14
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 4, 10, 11
46. Robinson, J., Sun, L., Yu, K., Batmanghelich, K., Jegelka, S., Sra, S.: Can contrastive learning avoid shortcut solutions? Advances in neural information processing systems **34**, 4974–4986 (2021) 6
47. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19305–19314 (2023) 2
48. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024) 3
49. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025) 4, 10, 11
50. Sun, Q., Wang, J., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, X.: Eva-clip-18b: Scaling clip to 18 billion parameters. arXiv preprint arXiv:2402.04252 (2024) 11
51. Tang, Y., Yu, J., Gai, K., Zhuang, J., Xiong, G., Gou, G., Wu, Q.: Missing target-relevant information prediction with world model for accurate zero-shot composed image retrieval. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24785–24795 (2025) 4
52. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 2, 4, 10, 11, 12
53. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: Internvid: A large-scale video-text dataset for multimodal understanding and generation. In: ICLR (2024) 12
54. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Xu, J., Wang, Z., et al.: Internvideo2: Scaling video foundation models for multimodal video understanding. In: ECCV (2024) 12, 14
55. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022) 12, 14
56. Wei, C., Chen, Y., Chen, H., Hu, H., Zhang, G., Fu, J., Ritter, A., Chen, W.: Uniir: Training and benchmarking universal multimodal information retrievers. In: ECCV (2024) 3, 6, 9, 10, 11
57. Wei, C., Chen, Y., Chen, H., Hu, H., Zhang, G., Fu, J., Ritter, A., Chen, W.: Uniir: Training and benchmarking universal multimodal information retrievers. In: European Conference on Computer Vision. pp. 387–404. Springer (2024) 4, 10
58. Wu, K., Li, P., Fu, J., Li, Y., Wu, Y., Liu, Y., Wang, J., Zhou, S.: Event-equalized dense video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8417–8427 (June 2025) 14
59. Wu, K., Li, P., Lyu, K., Lin, X., Zhao, L., He, Q., Wang, J., Liu, J.: Dual-anchoring: Addressing state drift in vision-language navigation. arXiv preprint arXiv:2604.17473 (2026) 4

60. Wu, Y., Deng, Y., Hui, S., Liu, Y., Wu, K., Huang, W., Wang, J.: Hierarchical frequency adaptation for all-in-one image restoration. Knowledge-Based Systems p. 116049 (2026) 14
61. Wu, Y., Deng, Y., Zhou, S., Liu, Y., Huang, W., Wang, J.: Cr-former: Single-image cloud removal with focused taylor attention. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024) 14
62. Xiao, T., Wang, S.: Towards off-policy learning for ranking policies with logged feedback. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 8700–8707 (2022) 4
63. Xiaomi, L., Xia, B., Shen, B., Zhu, D., Zhang, D., Wang, G., Zhang, H., Liu, H., Xiao, J., Dong, J., et al.: Mimo: Unlocking the reasoning potential of language model–from pretraining to posttraining. arXiv preprint arXiv:2505.07608 (2025) 2
64. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: CVPR (2016) 14
65. Xu, M., Dong, J., Hou, J., Wang, Z., Li, S., Gao, Z., Zhong, R., Cai, H.: Mm-r5: Multimodal reasoning-enhanced reranker via reinforcement learning for document retrieval. arXiv preprint arXiv:2506.12364 (2025) 4
66. Yang, Z., Liu, Y., Fu, J., Sugiyama, M., Zheng, N., et al.: Shaping schema via language representation as the next frontier for llm intelligence expanding. arXiv preprint arXiv:2605.09271 (2026) 4
67. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al.: Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476 (2025) 9
68. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.T., Sun, M., et al.: Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13807–13816 (2024) 4
69. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023) 10
70. Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J.: Long-clip: Unlocking the long-text capability of clip. In: European Conference on Computer Vision (2024) 10, 11
71. Zhang, K., Luan, Y., Hu, H., Lee, K., Qiao, S., Chen, W., Su, Y., Chang, M.W.: Magiclens: Self-supervised image retrieval with open-ended instructions (2024) 11
72. Zhang, Q., Lei, Z., Zhang, Z., Li, S.Z.: Context-aware attention network for image-text retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3536–3545 (2020) 2
73. Zhang, S., Fang, Q., Yang, Z., Feng, Y.: Llava-mini: Efficient image and video large multimodal models with one vision token. arXiv preprint arXiv:2501.03895 (2025) 14
74. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., et al.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176 (2025) 2
75. Zhao, W.X., Liu, J., Ren, R., Wen, J.R.: Dense text retrieval based on pretrained language models: A survey. ACM Transactions on Information Systems **42**(4), 1–60 (2024) 2
76. Zhou, J., Wang, X., Yu, J.: Optimizing preference alignment with differentiable ndcg ranking. arXiv preprint arXiv:2410.18127 (2024) 4

77. Zhou, J., Xiong, Y., Liu, Z., Liu, Z., Xiao, S., Wang, Y., Zhao, B., Zhang, C.J., Lian, D.: Megapairs: Massive data synthesis for universal multimodal retrieval. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 19076–19095 (2025) 4
78. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) 4
79. Zhu, L., Ji, D., Chen, T., Wu, H., Wang, S.: Retrv-r1: A reasoning-driven mllm framework for universal and efficient multimodal retrieval. arXiv preprint arXiv:2510.02745 (2025) 1, 4
80. Zhu, T., Jung, M.C., Clark, J.: Generalized contrastive learning for multi-modal retrieval and ranking. In: Companion Proceedings of the ACM on Web Conference 2025. pp. 661–670 (2025) 2