Embed-RL: Reinforcement Learning for Reasoning-Driven Multimodal Embeddings

Haonan Jiang[1,2*], Yuji Wang[1,2*], Yongjie Zhu[2†], Xin Lu[2], Wenyu Qin[2], Meng Wang[2], Pengfei Wan[2], and Yansong Tang[1‡]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University, China
[2]Kling Team, Kuaishou Technology, China
{jiang-hn24@mails, yuji-wan24@mails, tang.yansong@sz}.tsinghua.edu.cn
{zhuyongjie, luxin09, qinwenyu, wangmeng46, wanpengfei}@kuaishou.com

Abstract. Leveraging Multimodal Large Language Models (MLLMs) has become pivotal for advancing Universal Multimodal Embeddings (UME) in addressing diverse cross-modal tasks. Recent studies demonstrate that incorporating generative Chain-of-Thought (CoT) reasoning can substantially enhance task-specific representations compared to discriminative methods. However, the generated reasoning CoTs of existing generative embedding methods are limited to the textual analysis of queries and are irrelevant to the retrieval of the targets. To address these limitations, we propose a reasoning-driven UME framework that integrates Embedder-Guided Reinforcement Learning (EG-RL) to optimize the Reasoner to produce evidential Traceability CoT (T-CoT). Our key contributions are threefold: (1) We design an EG-RL framework where the Embedder provides explicit supervision to the Reasoner, ensuring the generated CoT traces are aligned with embedding tasks. (2) We introduce T-CoT, which extracts critical multimodal cues to focus on retrieval-relevant elements and provides multimodal inputs for the Embedder. (3) With limited computational resources, our framework outperforms the pioneering embedding model on both MMEB-V2 and UVRB benchmarks. The integration of multimodal evidence in structured reasoning, paired with retrieval-oriented alignment, effectively strengthens cross-modal semantic consistency and boosts the model's fine-grained matching capability as well as its generalization across complex scenarios. Our work demonstrates that targeted reasoning optimization can significantly improve multimodal embedding quality, providing a practical and efficient solution for reasoning-driven UME development. Project page.

Keywords: Multimodal Embedding · Generative Reasoning · Reinforcement Learning

1 Introduction

Multimodal embedding, as a fundamental and core supporting technology for cross-modal tasks, has been widely applied to numerous important and practical

∗: Equal Contribution. †: Project Leader. ‡: Corresponding Author.

application directions such as image-text retrieval, video moment localization, and visual document understanding [18, 28]. Traditional multimodal embedding methods typically adopt dual-encoder-based architectures, such as CLIP [33], BLIP [26], and SigLIP [50]. These methods demonstrate weaker inherent capability in bridging the gap between different modalities compared with advanced Multimodal Large Language Models (MLLMs) [1,23,24,39]. Additionally, MLLMs benefit from their remarkably strong multimodal understanding and instruction-following capabilities, enabling them to adapt to diverse, complex, and real-world task requirements. Therefore, an increasing body of literature [11, 18, 31, 36, 47] proves that MLLMs can be used to learn Universal Multimodal Embedding (UME) that captures general-purpose content similarity. Meanwhile, evaluation benchmarks such as the Multimodal Embedding Benchmark (MMEB) [18] and its upgraded version MMEB-V2 [31] have addressed this academic demand for UME research, covering 78 instruction-aware tasks across three modalities.

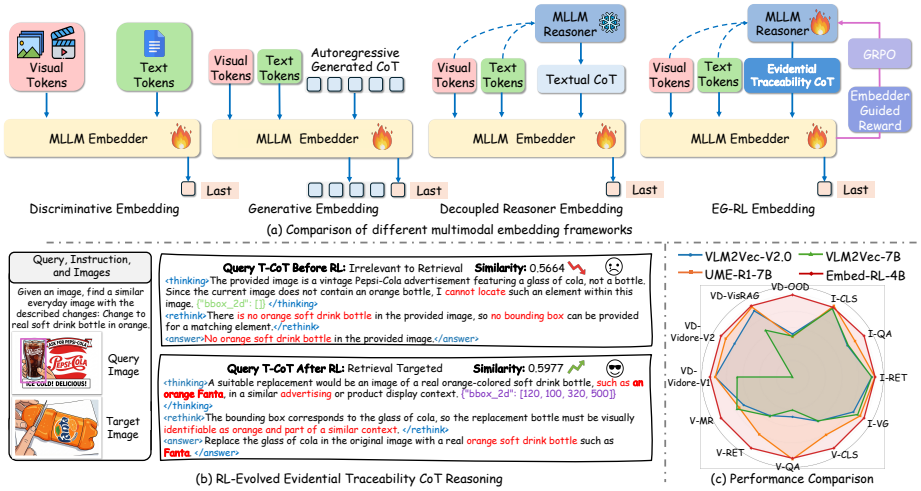

Fig. 1: Multimodal embedding optimization via Embedder-Guided Reinforcement Learning (EG-RL). (a) Frameworks evolution. (b) Reasoning enhancement with RL-optimized evidential Traceability CoT (T-CoT). (c) Comparison of multi-task performance.

Currently, the majority of existing MLLM-powered embedding methods belong to discriminative embedding models [21,30,53]. These models typically extract raw embedding features directly from the final hidden states of input tokens, which fails to fully leverage the generative capabilities and reasoning potential inherent but untapped in MLLMs. Consequently, recent studies have begun to actively explore the integration of generative reasoning into UME tasks, as shown in Figure 1. For instance, approaches like UME-R1 [22] effectively unify discriminative and generative embeddings through textual Chain-of-Thoughts (CoTs) generated by the MLLM Embedder. However, the simultaneous optimization of contrastive

loss and next-token prediction objectives can result in severe conflicting gradients, leading to suboptimal performance [3, 22]. In contrast, the decoupled Reasoner-Embedder paradigm proposed by TTE [6], which relies on massive computational resources and data distillation from large models, aims to alleviate this problem by decoupling the two processes, using pre-trained MLLMs to generate offline CoT reasoning to enhance embedding quality with only the Embedder trained. Nevertheless, the CoTs generated by the Reasoner of TTE are not specifically designed for embedding, as they are not trained together with the Embedder. This misalignment may introduce noise and even lead to hallucinations. Moreover, relying only on textual reasoning fails to leverage the MLLM Embedder's potential to process multimodal signals, whose representations could significantly enhance retrieval performance. This multimodal cue insufficiency leads to embedding alignment bias, where critical visual-spatial cues and video-temporal signals are poorly captured in retrieval, resulting in less effective cross-modal matching and restricted generalization on real-world multimodal tasks.

To address the above critical issues, this paper proposes an effective reasoning-driven decoupled UME framework. This framework specifically leverages the Embedder-Guided Reinforcement Learning (EG-RL) algorithm to optimize the CoTs generated by the Reasoner, using our novel process reward oriented to the accurate alignment between query and target and verifiable outcome reward for efficient retrieval. Firstly, we carefully construct a comprehensive dataset that initially trains the Embedder to generate high-quality embeddings conditioned on the sequence of preceding input and CoT tokens. The trained Embedder acts as a reward model and provides stable and reliable reward signals. Secondly, inspired by the region-aware paradigm [8,35,37,57] that makes the model precisely focus on the region of interest (RoI), we propose the evidential Traceability CoT (T-CoT) that explicitly guides the Embedder to focus on task-related information, effectively filter out redundant visual elements, and integrate modality-specific critical cues to adapt to long-text retrieval, coarse-grained semantic matching, and fine-grained alignment for enhanced robust performance across heterogeneous tasks. The main contributions of this paper are summarized as follows:

1. **Embedder-Guided Reinforcement Learning**. We propose a novel decoupled RL framework where the Embedder guides the Reasoner to optimize CoT trajectories for specific embedding tasks. This approach resolves key conflict between generative and embedding objectives, ensures the Reasoner's output greatly improves retrieval quality, and addresses core challenge of adapting general CoTs to embedding tasks.
2. **Evidential Traceability CoT for Embedding**. We further extend CoT reasoning to complex multimodal scenarios, integrating explicit visual localization information, video keyframes, and text keywords into detailed inference trajectories. This design enables the model to focus on core retrieval-related information and effectively mitigate the negative impact of redundant multimodal and text data on overall embedding alignment performance.
3. **Efficient Performance Improvement Across Multiple Benchmarks.** Under computationally constrained real-world settings, the framework pro-

posed in this paper effectively outperforms state-of-the-art generative embedding models on both MMEB-V2 [31] and Universal Video Retrieval Benchmark (UVRB) [13] two key benchmark datasets, and achieves exceptional retrieval performance across various combinatorial application scenarios.

2 Related Work

2.1 Universal Multimodal Embedding

Constructing robust multimodal representations is a fundamental core challenge in multimodal learning. Pioneering models such as CLIP [33] and ALIGN [17] adopt a dual-encoder architecture and learn effective representations through contrastive learning on rich large-scale image-text paired data. However, they struggle to handle interleaved image-text inputs, and their text encoders lack sufficient capacity to understand truly complex textual content. To address this issue, researchers leverage Multimodal Large Language Models (MLLMs) to build embedding models [18, 28, 30, 31, 53, 58], capitalizing on their strong multimodal comprehension capabilities to enhance overall learning performance.

Existing works focus on different aspects: VLM2Vec [18] transforms MLLMs into embedding models via contrastive learning and achieves outstanding performance on unconventional retrieval tasks such as visual question answering and localization; MM-Embed [28] explores using off-the-shelf MLLMs as zero-shot rerankers to optimize retrieval results; LamRA [30] unifies the multimodal retrieval paradigm through two-stage retrieval training and joint reranking. MegaPairs [58] and GME [54] address the modality imbalance problem with automated pipelines; LLaVE [21] and Unite [19] focus on hard negative sample mining.

Recent studies focus on instruction-aware representations: MMEB [18] and MMEB-V2 [31] construct a comprehensive evaluation benchmark covering 78 tasks. UME-R1 [22] first introduces reasoning mechanisms, yet simultaneous optimization of dual components via Reinforcement Learning (RL) leads to conflicts, and redundant Chain-of-Thought (CoT) trajectories dilute representations. TTE [6] adopts a decoupled yet computationally expensive architecture, and its Reasoner is misaligned with retrieval tasks, resulting in task-irrelevant outputs. In this paper, we propose a decoupled RL framework enabling separate optimization of dual components and generating retrieval-relevant reasoning trajectories through a dual reward mechanism, addressing the aforementioned challenges.

2.2 Multimodal Reasoning with Reinforcement Learning

MLLMs [1,23,29,39] extend and enrich the capabilities of Large Language Models (LLMs) to the multimodal domain, achieving promising results across diverse tasks including visual question answering [10, 29, 38], visual grounding [5, 20, 25], and keyframe extraction [27, 40, 51, 59]. Early works [16, 41, 44, 46, 55] mostly completed reasoning tasks using standardized CoT prompts. Since DeepSeek-R1 [12] proposed the Group Relative Policy Optimization (GRPO) RL algorithm,

numerous recent works have optimized advanced RL algorithms [48, 56] and enhanced the reasoning capabilities of MLLMs [2, 7, 34, 43, 45].

GRIT [8] interleaves bounding box coordinates with textual reasoning chains and designs an RL scheme based on the GRPO algorithm, enabling efficient training with dual robust rewards and no additional annotated data. Ground-R1 [2] proposes an RL framework to achieve grounded visual reasoning without extra annotations, guiding response generation through dual rewards to improve reasoning reliability and interpretability. BRPO [4] uses Intersection over Union (IoU)-based rewards to guide models to autonomously generate visual-text reflections, combined with a visual token mechanism to mitigate the problems of visual attention dilution and hallucinations. DeepEyes [57] adopts end-to-end RL to induce models to develop the ability of "thinking with images", improving performance on various reasoning tasks. TreeVGR [37] proposes the TreeBench benchmark and the TreeVGR training paradigm, enhancing visual grounding reasoning capabilities by jointly supervising localization and reasoning via RL. Inspired by grounding reasoning, this paper further proposes evidential Traceability CoT (T-CoT), which constructs structured multimodal reasoning chains by extracting bounding boxes of images, keyframes of videos, and keywords of text. This method enables the model to focus on the core regions of retrieval tasks, thereby improving embedding quality.

3 Methodology

3.1 Preliminaries

We focus on the universal multimodal retrieval task, where given a query q (text, image, or interleaved text-image modalities) and a candidate set $\Omega = \{c_n\}_{n=1}^{N}$, the goal is to retrieve the most relevant candidate from Ω.

To learn discriminative multimodal embeddings, we adopt contrastive learning with the InfoNCE loss [32], as shown in Figure 2(a). For a query q_i, its positive target t_i^+, and its negative target set $\mathcal{T}^- = \{t_j^-\}_{j\neq i}$, the InfoNCE loss optimizes the model to maximize the similarity between q_i and t_i^+ while minimizing the similarities to all $t_j^- \in \mathcal{T}^-$. The loss is defined as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\Big(\cos(\boldsymbol{h}_{q_i},\boldsymbol{h}_{t_i^+})/\tau\Big)}{\exp\Big(\cos(\boldsymbol{h}_{q_i},\boldsymbol{h}_{t_i^+})/\tau\Big)+\sum_{t^-\in\mathcal{T}^-}\exp\big(\cos(\boldsymbol{h}_{q_i},\boldsymbol{h}_{t^-})/\tau\big)}\ , \quad (1)$$

where $\boldsymbol{h}_{q_i}$ and $\boldsymbol{h}_t$ are embeddings of q_i and target t (extracted as the last-layer hidden states of the last token from a vision-language model), $\cos(\cdot,\cdot)$ denotes cosine similarity, and τ is the temperature hyperparameter.

3.2 Data Construction

To support the training of reasoning-driven universal multimodal embeddings, we construct a high-quality multimodal dataset following a "sampling-annotation-filtering-splitting" pipeline, as illustrated in Figure 2(a). The dataset integrates

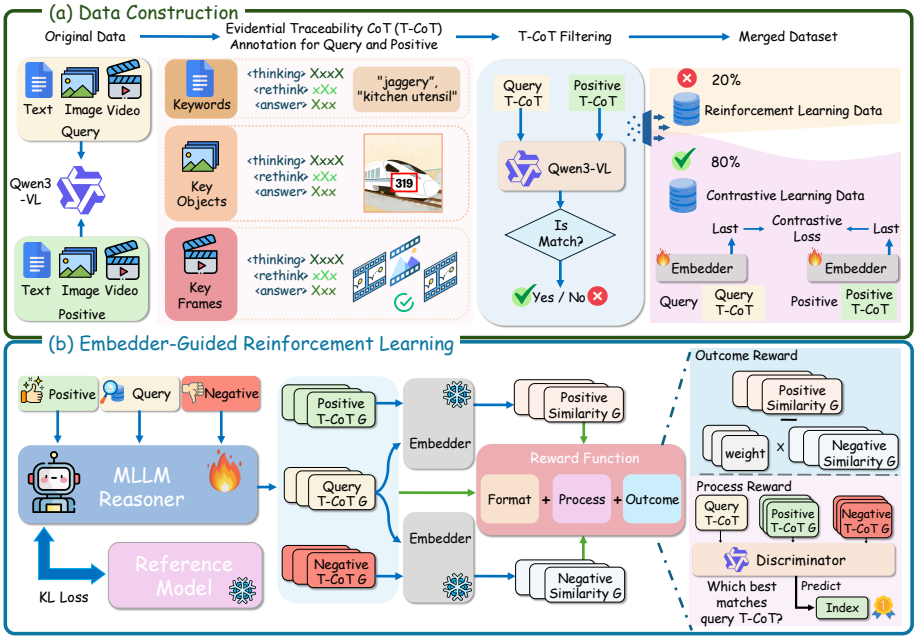

Fig. 2: Overview of the proposed data synthesis and EG-RL framework. (a) Data Construction generates T-CoT annotations for query-positive pairs, filters and splits the dataset to enable contrastive and reinforcement learning, laying the groundwork for reasoning-aware embedding. (b) Embedder-Guided Reinforcement Learning finetunes the MLLM with a process-outcome reward function, encouraging T-CoT trajectories that yield more discriminative and beneficial generative embeddings.

diverse modalities (image, video, visual document) and ensures alignment between reasoning trajectories and retrieval objectives through strict quality control.

We first curate the initial data pool by adopting a stratified sampling strategy across three core sources, and referencing the data paradigm of VLM2Vec-V2 [31]: (1) Image-centric tasks from MMEB-train [18], covering image classification, Question Answering, retrieval, and grounding; (2) Video-language instruction data from LLaVA-Hound [52], including video captioning, QA, and retrieval; (3) Visual document retrieval data from ViDoRe [9] and VisRAG [49].

Next, we perform evidential Traceability Chain-of-Thought (T-CoT) annotation for all query-positive pairs. Each T-CoT follows a structured three-part format: (1) `<thinking>` extracts modality-specific cues (text keywords via `text_keywords`, image spatial locations via `bbox_2d` (two-dimensional bounding box), video critical moments via `key_frames`); (2) `<rethink>` refines reasoning logic to focus on key retrieval-relevant aspects; (3) the final answer summarizes core retrieval-relevant information. We design task-specific prompts for annotation, ensuring T-CoT aligns with diverse multimodal retrieval scenarios.

Following annotation, we perform a strict CoT-guided relevance filtering to eliminate noisy samples. With a custom-designed judgment prompt, we assess whether the T-CoTs of queries and their positive samples are clearly irrelevant or contradictory to the task description. We retain only samples labeled "No", meaning those that are relevant and not contradictory, for contrastive learning. This filtering step effectively mitigates noise interference in contrastive learning. The initial dataset contains 2.22 million samples, and 1.83 million are preserved after filtering, yielding a retention rate of approximately 80%. Approximately 20% of the filtered-out samples are uniformly sampled and used in the reinforcement learning stage, as these hard examples are valuable for model exploration in reinforcement learning. In addition, we assign training weights to different datasets based on their task importance and data quality.

The constructed dataset exhibits three features: (1) modal diversity, covering three modalities of text, image, and video; (2) reasoning alignment, where T-CoT explicitly integrates multimodal cues and retrieval-related logic, avoiding informational redundancy; (3) quality assurance, as rigorous filtering and weighted sampling ensure that the dataset is free of significant noise and balanced across tasks. This dataset lays a solid foundation for the two-stage training, enabling Embedder to learn reasoning-aware representations and the Reasoner to optimize the generation of retrieval-centric T-CoTs through reinforcement learning.

3.3 Embedder-Guided Reinforcement Learning

To address the misalignment between generative reasoning and embedding objectives, we propose a decoupled reinforcement learning framework in which the pre-trained Embedder provides supervision to the Reasoner. This framework optimizes the generation of T-CoT to prioritize retrieval-relevant multimodal cues, leveraging a dual-guidance reward mechanism and Group Relative Policy Optimization (GRPO) [12]. Figure 2(b) illustrates the workflow of this stage.

EG-RL Framework Design. First, we fully train an Embedder using the InfoNCE loss to equip it with robust embedding capabilities. Our RL framework maintains strict separation between two components: the Reasoner, which is responsible for generating T-CoT, and the Embedder, which is frozen after contrastive training. This decoupling ensures three key benefits: (1) targeted optimization of reasoning without disrupting the Embedder's learned discriminative capabilities; (2) stable reward signals from the frozen Embedder that consistently evaluate T-CoT quality based on embedding alignment; and (3) flexible integration of multi-source rewards to internalize both retrieval and reranking knowledge. The Reasoner takes multimodal queries as input and outputs structured T-CoT, which integrates three critical cues: text keywords, image bounding boxes, and video keyframes. Additionally, we recrop the content within the bounding boxes and keyframes based on T-CoT to achieve multimodal reasoning-aware embeddings. This structured reasoning is then concatenated with the original input to form the Embedder's input, denoted as $\mathcal{I}$:

$$\mathcal{I} = [x_{\text{text}},\, x_{\text{img}},\, x_{\text{vid}},\, \text{T-CoT}(x),\, <\text{emb}>]\,. \tag{2}$$

In this equation, <emb> is a special token whose hidden state is extracted as the final embedding, and the evaluation of this embedding by the Embedder directly guides the Reasoner's policy update.

Reward Function with Process and Outcome Guidance. We design a three-component reward function to align T-CoT generation with embedding quality, combining format compliance, outcome-level retrieval effectiveness, and process-level T-CoT alignment:

Format Reward ($\mathcal{R}_{format}$): This reward ensures T-CoT strictly follows the predefined template (`<thinking>` → `<rethink>` → `<answer>`) and includes all required multimodal cues. Reward 1 for full compliance, 0 otherwise, guaranteeing T-CoT output interpretability and compatibility with the Embedder module.

Embedder-Guided Outcome Reward ($\mathcal{R}_{outcome}$): This reward measures how T-CoT improves embedding alignment by jointly assessing the ranking accuracy of positive samples and the similarity margin between positive and hard negative samples. The margin is a softmax-weighted average of negative similarities scaled by a temperature parameter. For a query q_i with positive target t_i^+ and in-batch negatives $\{t_j^-\}_{j\neq i}$, embeddings are $e_{q_i} = \pi_e(q_i, o_i^q)$ and $e_{t_j} = \pi_e(t_j, o_j^t)$, where o_i^q and o_j^t are T-CoT outputs for query and target. The reward for o_i^q is defined as:

$$\mathcal{R}_{\text{outcome}}(o_i^q) = \text{Acc}_k(e_{q_i}, t_i^+) \cdot \Big(\text{sim}(e_{q_i}, e_{t_i^+}) - \mathbb{E}_\tau\big[\text{sim}(e_{q_i}, e_{t_j^-})\big]\Big), \tag{3}$$

where $\text{Acc}_k(e_{q_i}, t_i^+)$ denotes the top-k retrieval accuracy, which measures whether t_i^+ is among the top-k ranked targets when sorted by cosine similarity to e_{q_i}; $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity between normalized embeddings; and $\mathbb{E}_\tau[\cdot]$ stands for the softmax-weighted average of cosine similarities between e_{q_i} and embeddings of in-batch negative targets.

Additionally, we compute $\mathcal{R}_{\text{outcome}}$ symmetrically for positive targets: taking t_i^+ as the anchor, q_i as its positive query, and other in-batch queries as negatives to calculate $\mathcal{R}_{\text{outcome}}$ for o_i^t. This symmetric computation enforces consistent embedding alignment in both query-to-target and target-to-query directions. This reward optimizes T-CoT with embedding learning as the core objective, enhancing its discriminative ability across samples.

T-CoT Process Reward ($\mathcal{R}_{process}$): We employ an independent pretrained Vision-Language Model Discriminator $\mathcal{D}$ for listwise comparison to align T-CoT outputs of queries and targets. Let q_{cot} be the query's T-CoT output and $\{c_{\text{cot}}^j\}_{j=1}^m$ the T-CoT outputs of m in-batch candidate targets, comprising positive samples from multiple rollouts of the query's data pair and negative samples. After shuffling, the index set of ground-truth positives is denoted $\mathcal{P}$. To mitigate position bias, we feed q_{cot} and shuffled $\{c_{\text{cot}}^j\}_{j=1}^m$ to $\mathcal{D}$ as pairwise inputs. The reward quantifies alignment via $\mathcal{D}$'s selection correctness, formally defined as:

$$\mathcal{R}_{\text{process}}(o_i) = \begin{cases} 1, & \text{if } \mathcal{D}\big(q_{\text{cot}}, \{c_{\text{cot}}^j\}_{j=1}^m\big) \in \mathcal{P}, \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where o_i denotes the T-CoT generation outcome of the i-th sample, and $\mathcal{D}(\cdot,\cdot)$ outputs the index of the candidate T-CoT most aligned with q_{cot} in the shuffled

candidate set. A reward of 1 indicates that $\mathcal{D}$ correctly selects a positive T-CoT from the ground-truth set, signifying well-aligned query-target T-CoT pairs; a reward of 0 means $\mathcal{D}$ fails to select any positive T-CoT, indicating misalignment between query and target T-CoT outputs. We compute $\mathcal{R}_{\text{process}}$ symmetrically in the reverse direction, moving from positive targets to their corresponding queries, to ensure consistent embedding alignment across both directions.

This reward guides Reasoner to align query and target T-CoT outputs. Since T-CoT is the intermediate process for embedding generation, T-CoT alignment quantified by $\mathcal{D}$ selection correctness directly improves embedding quality.

The total reward is a weighted combination of these three components:

$$\mathcal{R}_{\text{total}} = \alpha\mathcal{R}_{\text{format}} + \beta\mathcal{R}_{\text{process}} + \gamma\mathcal{R}_{\text{outcome}}, \tag{5}$$

where α, β, γ are non-negative weighting coefficients that balance the relative contributions of the three reward components in the total reward optimization.

Policy Optimization with GRPO. We adopt GRPO to optimize the Reasoner's policy, and use group-based rewards to stabilize the training process. For each query-target pair $q \sim \mathcal{S}$, where $\mathcal{S}$ denotes the training sample set of query-target pairs. We sample $G = 8$ candidate T-CoT sequences $\{o_i\}_{i=1}^{G}$ according to the old policy $\pi_{\theta_{\text{old}}}$. The optimization objective is defined as:

$$\begin{aligned}\mathcal{L}_{\text{grpo}} = \mathbb{E}_{\substack{q\sim\mathcal{S}, \\ \{o_i\}\sim\pi_{\theta_{\text{old}}}}} \Bigg[\frac{1}{G}\sum_{i=1}^{G} \Big(& \min\left(r_\theta(o_i)A_i, \text{clip}\left(r_\theta(o_i), 1-\epsilon, 1+\epsilon\right) A_i\right) \\ & - \beta\mathbb{D}_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right) \Big) \Bigg],\end{aligned} \tag{6}$$

where $r_\theta(o_i) = \pi_\theta(o_i|q)/\pi_{\theta_{\text{old}}}(o_i|q)$ denotes importance ratio, ϵ is the clipping threshold of importance ratio, β is a hyperparameter weighting the Kullback-Leibler divergence term, π_{ref} denotes the reference policy model before optimization, and $A_i = (r_i - \mu_r)/\sigma_r$ represents advantage, where $\mu_r = \text{mean}(\{r_1, ..., r_G\})$ and $\sigma_r = \text{std}(\{r_1, ..., r_G\})$ are the mean and std of group rewards.

4 Experiments

4.1 Implementation Details

We train Qwen3-VL-2B [1] and Qwen3-VL-4B [1] as Embedders with the DeepSpeed Zero2 optimization framework, and adopt a sub-batch strategy following VLM2Vec [18]. The models are trained for 2 epochs with a learning rate of 1e-4 and weight decay of 0.01; the batch size is computationally light: 512 for the 2B model and 256 for the 4B model, and we use Low-Rank Adaptation (LoRA) [14] for fine-tuning. For reinforcement learning, Qwen3-VL-8B [1] is trained as Reasoner via the GRPO algorithm [12] for 1 epoch, with a computationally light batch size of 256, a learning rate of 3e-6, and standard GRPO hyperparameters.

4.2 Baselines and Datasets

We compare with representative multimodal embedding models with diverse architectures, modalities and scales. These baselines cover image, video and visual document retrieval, ensuring thorough and fair evaluation. We evaluate against GME [54], ColPali [9], VLM2Vec [18], LamRA [30], CAFe [47], VLM2Vec-V2 [31], TTE [6],UME-R1 [22], InternVideo2 [42], Unite [11], and GVE [13].

For the training phase, we followed VLM2Vec-V2 [31] and constructed a comprehensive dataset from three key sources: video-language instruction data from LLaVA-Hound [52], visual document retrieval data from ViDoRe [9] and VisRAG [49], and image-based vision-language task data from MMEB-train [18]. Detailed training procedures, hyperparameters, and dataset construction are in the supplementary material. We evaluate on two comprehensive benchmarks:

MMEB-V2 (Massive Multimodal Embedding Benchmark) [31]: It is a comprehensive and robust benchmark consisting of 78 diverse tasks across three core visual modalities (image, video, and visual document). MMEB-V2 extends the original MMEB [18] by introducing five additional meta-tasks focused specifically on video and visual document understanding, bringing the total to nine meta-tasks. We adopted Hit@1 as the evaluation metric for image and video tasks, and Normalized Discounted Cumulative Gain (NDCG@5) [15] for visual documents.

UVRB (Universal Video Retrieval Benchmark) [13]: It is a suite of 16 datasets to identify capability gaps in video retrieval across tasks and domains. UVRB measures multi-dimensional generalization over textual, composite, and visual retrieval tasks, including across coarse-grained, fine-grained, and long-context scenarios. We report mean Average Precision (mAP) for all UVRB tasks.

4.3 Main Results

Table 2 compares the comprehensive performance of our proposed Embed-RL models with various baseline approaches on the MMEB-V2 benchmark. Our Embed-RL models consistently achieve superior performance directly compared to all baseline models. Specifically, Embed-RL-4B attains the best overall score of 68.1, outperforming the strong next top baseline UME-R1-7B by 3.6 points. Embed-RL-2B follows closely with an overall score of 66.8, also surpassing all baseline variants. Across different modalities, our models show clear and notable advantages. In the Image modality, Embed-RL-4B impressively achieves the best grounding (GD) performance of 91.4, with Embed-RL-2B ranking second at 90.4. For the Video modality, both Embed-RL-2B and Embed-RL-4B outperform all baselines in

Table 1: Video retrieval performance on UVRB. Domain dimensions: Coarse-grained (CG), Fine-grained (FG), Long-context (LC). The best and second-best results are marked in **bold** and underline.

Model	CG	FG	LC
InternVideo2-6B [42]	50.4	41.7	42.3
VLM2Vec-V2 [31]	49.8	50.2	76.2
GME-7B [54]	51.8	50.7	78.8
Unite-7B [19]	54.1	53.9	74.6
GVE-3B [13]	55.2	54.1	76.4
Embed-RL-2B	59.1	54.6	**86.9**
Embed-RL-4B	**60.7**	**55.6**	86.1

Table 2: Comparison of performance between baselines and our method on MMEB-V2. CLS: classification, QA: question answering, RET: retrieval, GD: grounding, MRET: moment retrieval, VDR: ViDoRe, VR: VisRAG, OOD: out-of-domain. The highest and second-highest values highlighted in **bold** and underline.

Model	Image					Video					VisDoc					All
	CLS	QA	RET	GD	Overall	CLS	QA	RET	MRET	Overall	VDRv1	VDRv2	VR	OOD	Overall	
# of Datasets	10	10	12	4	36	5	5	5	3	18	10	4	6	4	24	78
[HTML]EDEDED *Baseline Models*																
ColPali-V1.3-3B [9]	40.3	11.5	48.1	40.3	34.9	26.7	37.8	21.6	25.5	28.2	83.6	52.0	81.1	43.1	71.0	44.4
GME-2B [54]	54.4	29.9	66.9	55.5	51.9	34.9	42.0	25.6	32.4	33.9	86.1	54.0	82.5	43.1	72.7	54.1
GME-7B [54]	57.7	34.7	71.2	59.3	56.0	37.4	50.4	28.4	38.2	38.6	**89.4**	**55.6**	**85.0**	44.4	**75.2**	57.8
LamRA-2-7B [30]	59.2	26.5	70.0	62.7	54.1	39.3	42.6	24.3	34.6	35.2	22.0	11.5	37.4	21.0	23.9	40.4
LamRA-2.5-7B [30]	51.7	34.1	66.9	56.7	52.4	32.9	42.6	23.2	37.6	33.7	56.3	33.3	58.2	40.1	50.2	47.4
VLM2Vec-2B [18]	58.7	49.3	65.0	72.9	59.7	33.4	30.5	20.6	33.0	29.0	49.8	13.5	51.8	33.5	41.6	47.0
VLM2Vec-7B [18]	62.7	56.9	69.4	82.2	65.5	39.1	30.0	29.0	40.6	34.0	56.9	9.4	59.1	38.1	46.4	52.3
VLM2Vec-V2-2B [31]	62.9	56.3	69.5	77.3	64.9	39.3	34.3	28.8	38.5	34.9	75.5	44.9	79.4	39.4	65.4	58.0
VLM2Vec-V2-7B [31]	65.7	61.5	70.0	85.2	68.1	45.9	33.9	27.6	39.3	36.4	78.8	52.6	82.7	42.1	69.3	61.2
CAFe-7B [47]	63.6	61.7	69.1	87.6	67.6	35.8	58.7	34.4	39.5	42.4	70.7	49.6	79.5	38.1	63.9	60.6
TTE_s-2B [6]	**67.9**	66.6	70.2	84.1	70.1	47.3	49.1	34.4	33.2	32.1	77.5	53.2	83.2	41.1	68.8	63.1
UME-R1-2B [22]	64.8	62.8	67.6	77.2	66.6	44.3	51.2	32.9	39.7	42.2	72.4	46.2	79.2	37.2	63.9	60.1
UME-R1-7B [22]	67.1	69.2	**71.9**	84.9	**71.3**	48.6	**60.7**	38.2	39.3	47.5	75.7	50.5	83.7	37.6	67.1	64.5
[HTML]FFE5E5 *Ours*																
Embed-RL-2B	62.8	67.9	68.6	90.4	69.2	57.0	55.9	**45.1**	49.4	52.1	79.9	52.0	84.6	65.7	74.1	66.8
Embed-RL-4B	63.7	**70.5**	71.3	**91.4**	70.1	**57.6**	58.4	**45.1**	**49.5**	**53.0**	80.2	53.4	84.9	**67.1**	74.7	**68.1**

overall score, with Embed-RL-2B scoring 52.1 and Embed-RL-4B scoring 53.0, and they achieve a video retrieval (RET) score of 45.1. In the visual document modality, we observe significant improvements in out-of-domain (OOD) performance, where Embed-RL-4B reaches 67.1 and Embed-RL-2B 65.7, far exceeding prior baseline results. These results demonstrate the effectiveness of our proposed approach across diverse visual modalities and task types.

In the broader field of video retrieval, benefiting from the effectiveness of our T-CoT to accurately locate keywords or keyframes, our model exhibits significant advantages in Coarse-grained (CG), Fine-grained (FG), and Long-context (LC) retrieval tasks. As shown in Table 1, detailing video retrieval performance of different models on the UVRB dataset across the three domains, our Embed-RL models consistently excel: 4B tops CG at 60.7 and FG at 55.6, while 2B leads LC at 86.9, both outperforming all existing baselines by a clear margin.

Additionally, Figure 3 presents the visualization of our T-CoT on text, image and video; we crop bbox and keyframe to achieve multi-modal CoT input, and our T-CoT accurately locates retrieval needs to improve retrieval performance. More scores and visualizations are provided in the supplementary material.

4.4 Ablation Study

To dissect the contribution of each component in our framework, we conduct ablation experiments on MMEB-V2, where Embed-RL-2B serves as the model.

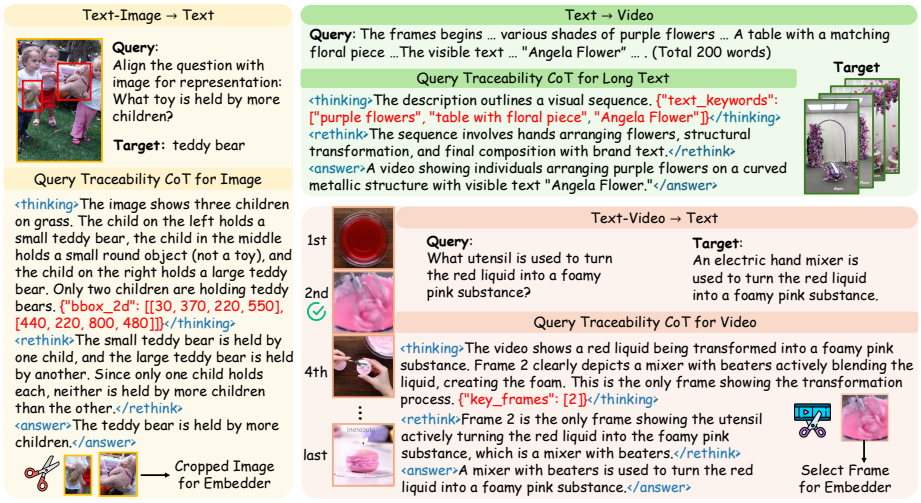

Fig. 3: Example visualization of our reasoning-driven embedding framework on multimodal retrieval tasks. The figure shows the evidential Traceability CoT reasoning process for video and visual document retrieval.

Analysis of the reward model in the RL Stage. As shown in Table 3, removing the RL stage leads to an overall performance decline of 1.5 points, dropping the score from 66.8 to 65.3. This result verifies that RL fine-tuning is indispensable for embedding alignment. Omitting weighted negative sampling, a core module of our contrastive reward mechanism, brings a performance reduction of 0.3 points from 66.8 to 66.5. This outcome emphasizes the component's key function in prioritizing hard negative examples to strengthen discriminative embedding learning.The process reward is formulated to reward logical reasoning steps and align query T-CoT with target T-CoT. It contributes 0.8 points to the overall performance, as its exclusion lowers the score from 66.8 to 66.0. This component shows the most notable influence on video tasks, where performance falls from 52.1 to 51.3. This trend reveals that video understanding strongly depends on step-by-step reasoning and further reflects the critical role of the process reward in T-CoT alignment.Additionally, the outcome reward is built to reward final predictions. It contributes 1.0 point to the overall performance, and its removal reduces the total score from 66.8 to 65.8. This reward ensures that the reasoning process remains consistent with the objective of the target task.

Impact of Reasoning Components on T-CoT. As shown in Table 4, removing the reasoning process while retaining the answer part leads to a performance decrease of 1.3 points from 66.8 to 65.5. Image grounding and video moment retrieval see notable drops, falling from 69.2 to 67.9 and 52.1 to 50.5, which highlights the importance of multimodal evidence tracking for fine-grained alignment tasks. Removing multimodal cues leads to a performance reduction of 1.0 points from 66.8 to 65.8, validating the necessity of extracting multimodal

cues through bounding boxes and keyframes to enhance alignment between multimodal representations and retrieval objectives. Most critically, using only raw input without T-CoT causes a catastrophic decrease in overall performance of 6.6 points, from 66.8 to 60.2. The greatest impact appears on video tasks, where performance falls from 52.1 to 43.7. This dramatic decline demonstrates the necessity of high-quality evidential Traceability CoT for retrieval accuracy, as it enables the model to decompose complex retrieval and understanding tasks into manageable steps and improve cross-modal embedding quality.

Table 3: Ablation Study on Reward Components in EG-RL stage.

Model	Image	Video	VisDoc	All
Embed-RL-2B	**69.2**	**52.1**	**74.1**	**66.8**
w/o EG-RL	68.0	50.1	72.7	65.3
w/o weighted negative	68.9	51.7	73.9	66.5
w/o process reward	68.3	51.3	73.5	66.0
w/o outcome reward	68.1	51.2	73.1	65.8

Table 4: Ablation Study on Reasoning Components in T-CoT.

Model	Image	Video	VisDoc	All
Embed-RL-2B	**69.2**	**52.1**	**74.1**	**66.8**
w/o reasoning	67.9	50.5	73.1	65.5
w/o multimodal cues	68.1	51.4	73.3	65.8
w/ raw input	60.4	43.7	72.4	60.2

Ablation Study on Model's Discriminative Ability for Candidates. We define top-ranked candidates with the highest similarity (excluding positive samples) as highly similar samples. On this basis, we study how optimizing the reasoner with EG-RL improves the model's ability to distinguish between similar candidates. Specifically, we calculate the difference between the query's similarity to the most and second-most similar candidates on each dataset, both before and after RL. This difference measures whether the model assigns significantly higher similarity to positive samples than to other highly similar candidates. As shown in Figure 4, we observe that on different datasets across three modalities, the radar chart obtained after RL prominently encloses the one obtained before RL. This indicates that the computed similarity difference becomes larger after RL, widening the gap between the query's similarity to the top-ranked candidate and the second-ranked candidate. It demonstrates that the model's ability to discriminate between similar candidates is effectively enhanced. Meanwhile, the bar chart shows that the model achieves consistent overall improvement on three-modality datasets. This verifies that optimizing T-CoT with RL strengthens the model's general ability to distinguish between different candidates.

Ablation on traceable evidence count and retrieval metrics. We also systematically analyze the relationship between the number of traceable evidence pieces and core retrieval metrics across all datasets before and after reinforcement learning. For images and visual document data, we count the change in the number of bounding boxes. For video data, we similarly count the change in the number of keyframes. We observe that after reinforcement learning, the T-CoT generated by the Reasoner tends to produce more bounding boxes. For the video modality, the model tends to focus on fewer keyframes. The retrieval metrics yield consistent and significant improvements, with the curves after reinforcement

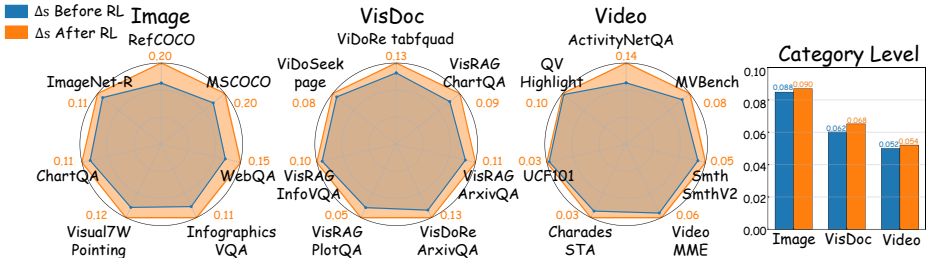

Fig. 4: Similarity difference $\Delta s = \text{sim}(\text{query}, \text{top1}) - \text{sim}(\text{query}, \text{top2})$ before and after EG-RL. Here, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity of normalized embeddings, top1 is the most similar positive candidate and top2 the second-most similar. This metric quantifies the model's discriminative ability over similar candidates on multimodal datasets.

learning lying entirely above those before. For the image modality, the model captures more visual evidence to boost reasoning accuracy and recall. For the video modality, it concentrates on critical frames and conducts precise keyframe extraction and temporal localization to identify key content. These changes are particularly pronounced on complex samples involving multi-object localization and multi-person relationship reasoning.

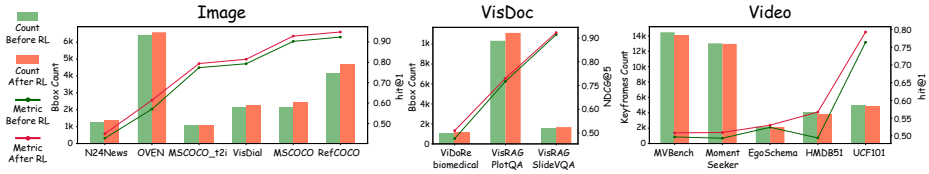

Fig. 5: Relationship between traceable evidence counts and retrieval metrics across datasets. Hit@1 is employed for Image and Video; NDCG@5 is used for VisDoc. Bounding box counts are shown for Image and VisDoc, while keyframe counts for Video.

5 Conclusion

This work addresses key limitations of generative universal multimodal embedding (UME) methods: chain-of-thought (CoT) remains text-only, resulting in poor retrieval relevance, while joint optimization of generative and embedding objectives gives rise to gradient conflicts that impede cross-modal matching. We propose Embed-RL, a reasoning-driven UME model built upon Embedder-Guided RL (EG-RL), which serves as a decoupled reinforcement learning framework that integrates multimodal evidential Traceability CoT (T-CoT) and a retrieval-oriented dual-reward mechanism to enable precise reasoning-embedding alignment. Extensive experiments on the MMEB-V2 and UVRB benchmarks demonstrate

that Embed-RL outperforms state-of-the-art counterparts within computational constraints, achieving significant improvements in cross-modal retrieval and out-of-domain generalization tasks. This work shows that targeted reasoning optimization can substantially enhance multimodal embeddings, providing an efficient solution for reasoning-driven UME and valuable insights into the integration of generative reasoning with multimodal representation learning.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62572270, and in part by the Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2025B1515020012).

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Cao, M., Zhao, H., Zhang, C., Chang, X., Reid, I., Liang, X.: Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning. arXiv preprint arXiv:2505.20272 (2025)
3. Chen, H., Liu, H., Luo, Y., Wang, L., Yang, N., Wei, F., Dou, Z.: Moca: Modality-aware continual pre-training makes better bidirectional multimodal embeddings. arXiv preprint arXiv:2506.23115 (2025)
4. Chu, X., Chen, X., Wang, G., Tan, Z., Huang, K., Lv, W., Mo, T., Li, W.: Qwen look again: Guiding vision-language reasoning models to re-attention visual information. arXiv preprint arXiv:2505.23558 (2025)
5. Chung, J., Kim, J., Kim, S., Lee, J., Kim, M.S., Yu, Y.: Don't look only once: Towards multimodal interactive reasoning with selective visual revisitation. arXiv preprint arXiv:2505.18842 (2025)
6. Cui, X., Cheng, J., Chen, H.y., Shukla, S.N., Awasthi, A., Pan, X., Ahuja, C., Mishra, S.K., Yang, Y., Xiao, J., et al.: Think then embed: Generative context improves multimodal embedding. arXiv preprint arXiv:2510.05014 (2025)
7. Duan, C., Fang, R., Wang, Y., Wang, K., Huang, L., Zeng, X., Li, H., Liu, X.: Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning. arXiv preprint arXiv:2505.17022 (2025)
8. Fan, Y., He, X., Yang, D., Zheng, K., Kuo, C.C., Zheng, Y., Narayanaraju, S.J., Guan, X., Wang, X.E.: Grit: Teaching mllms to think with images. arXiv preprint arXiv:2505.15879 (2025)
9. Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., Colombo, P.: Colpali: Efficient document retrieval with vision language models. arXiv preprint arXiv:2407.01449 (2024)

10. Geng, X., Xia, P., Zhang, Z., Wang, X., Wang, Q., Ding, R., Wang, C., Wu, J., Zhao, Y., Li, K., et al.: Webwatcher: Breaking new frontier of vision-language deep research agent. arXiv preprint arXiv:2508.05748 (2025)
11. 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)
12. 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)
13. Guo, Z., Li, M., Zhang, Y., Long, D., Xie, P., Chu, X.: Towards universal video retrieval: Generalizing video embedding via synthesized multimodal pyramid curriculum. arXiv preprint arXiv:2510.27571 (2025)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
15. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems (TOIS) **20**(4), 422–446 (2002)
16. Ji, D., Zhu, L., Gao, S., Xu, P., Lu, H., Ye, J., Zhao, F.: Tree-of-table: Unleashing the power of llms for enhanced large-scale table understanding. arXiv preprint arXiv:2411.08516 (2024)
17. 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)
18. Jiang, Z., Meng, R., Yang, X., Yavuz, S., Zhou, Y., Chen, W.: Vlm2vec: Training vision-language models for massive multimodal embedding tasks. arXiv preprint arXiv:2410.05160 (2024)
19. 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)
20. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9579–9589 (2024)
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)
22. Lan, Z., Niu, L., Meng, F., Zhou, J., Su, J.: Ume-r1: Exploring reasoning-driven generative multimodal embeddings. arXiv preprint arXiv:2511.00405 (2025)
23. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
24. 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)
25. Li, G., Xu, J., Zhao, Y., Peng, Y.: Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9098–9108 (2025)
26. 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)

27. Liao, Z., Xie, Q., Zhang, Y., Kong, Z., Lu, H., Yang, Z., Deng, Z.: Improved visual-spatial reasoning via r1-zero-like training. arXiv preprint arXiv:2504.00883 (2025)
28. 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)
29. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
30. 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)
31. 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)
32. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
33. 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)
34. Su, Z., Li, L., Song, M., Hao, Y., Yang, Z., Zhang, J., Chen, G., Gu, J., Li, J., Qu, X., et al.: Openthinkimg: Learning to think with images via visual tool reinforcement learning. arXiv preprint arXiv:2505.08617 (2025)
35. Su, Z., Xia, P., Guo, H., Liu, Z., Ma, Y., Qu, X., Liu, J., Li, Y., Zeng, K., Yang, Z., et al.: Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. arXiv preprint arXiv:2506.23918 (2025)
36. Thirukovalluru, R., Meng, R., Liu, Y., Su, M., Nie, P., Yavuz, S., Zhou, Y., Chen, W., Dhingra, B., et al.: Breaking the batch barrier (b3) of contrastive learning via smart batch mining. arXiv preprint arXiv:2505.11293 (2025)
37. Wang, H., Li, X., Huang, Z., Wang, A., Wang, J., Zhang, T., Zheng, J., Bai, S., Kang, Z., Feng, J., et al.: Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. arXiv preprint arXiv:2507.07999 (2025)
38. Wang, P., Ling, H.: Svqa-r1: Reinforcing spatial reasoning in mllms via view-consistent reward optimization. arXiv preprint arXiv:2506.01371 (2025)
39. 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)
40. Wang, Q., Liu, J., Liang, J., Jiang, Y., Zhang, Y., Chen, J., Zheng, Y., Wang, X., Wan, P., Yue, X., et al.: Vr-thinker: Boosting video reward models through thinking-with-image reasoning. arXiv preprint arXiv:2510.10518 (2025)
41. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
42. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: European Conference on Computer Vision. pp. 396–416. Springer (2024)
43. Wang, Y., Liu, W., Niu, J., Zhang, H., Tang, Y.: Vg-refiner: Towards tool-refined referring grounded reasoning via agentic reinforcement learning. arXiv preprint arXiv:2512.06373 (2025)

44. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
45. Wu, M., Yang, J., Jiang, J., Li, M., Yan, K., Yu, H., Zhang, M., Zhai, C., Nahrstedt, K.: Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. arXiv preprint arXiv:2505.19255 (2025)
46. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2087–2098 (2025)
47. Yu, H., Zhao, Z., Yan, S., Korycki, L., Wang, J., He, B., Liu, J., Zhang, L., Fan, X., Yu, H.: Cafe: Unifying representation and generation with contrastive-autoregressive finetuning. arXiv preprint arXiv:2503.19900 (2025)
48. 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)
49. Yu, S., Tang, C., Xu, B., Cui, J., Ran, J., Yan, Y., Liu, Z., Wang, S., Han, X., Liu, Z., et al.: Visrag: Vision-based retrieval-augmented generation on multi-modality documents. arXiv preprint arXiv:2410.10594 (2024)
50. 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)
51. Zhang, H., Gu, X., Li, J., Ma, C., Bai, S., Zhang, C., Zhang, B., Zhou, Z., He, D., Tang, Y.: Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. arXiv preprint arXiv:2508.04416 (2025)
52. Zhang, R., Gui, L., Sun, Z., Feng, Y., Xu, K., Zhang, Y., Fu, D., Li, C., Hauptmann, A.G., Bisk, Y., et al.: Direct preference optimization of video large multimodal models from language model reward. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 694–717 (2025)
53. Zhang, X., Zhang, Y., Xie, W., Li, M., Dai, Z., Long, D., Xie, P., Zhang, M., Li, W., Zhang, M.: Gme: Improving universal multimodal retrieval by multimodal llms. arXiv preprint arXiv:2412.16855 (2024)
54. Zhang, X., Zhang, Y., Xie, W., Li, M., Dai, Z., Long, D., Xie, P., Zhang, M., Li, W., Zhang, M.: Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9274–9285 (2025)
55. Zhang, X., Du, C., Pang, T., Liu, Q., Gao, W., Lin, M.: Chain of preference optimization: Improving chain-of-thought reasoning in llms. Advances in Neural Information Processing Systems **37**, 333–356 (2024)
56. Zheng, C., Liu, S., Li, M., Chen, X.H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., et al.: Group sequence policy optimization. arXiv preprint arXiv:2507.18071 (2025)
57. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)
58. 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)
59. Zhu, L., Chen, Q., Shen, X., Cun, X.: Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning. arXiv preprint arXiv:2505.23504 (2025)