

MMEmb-R1: Reasoning-Enhanced Multimodal Embedding with Pair-Aware Selection and Adaptive Control

Yuchi Wang^{12*}, Dingkan Yang¹³, Haiyang Yu², Weikang Bian¹, Jiefeng Long²,
Xiao Liang², Chao Feng^{2†}, Hongsheng Li^{1†}

¹MMLab, The Chinese University of Hong Kong

²ByteDance Inc. ³Fudan University

wangyuchi369@gmail.com chaofeng.zz@bytedance.com hsli@ee.cuhk.edu.hk

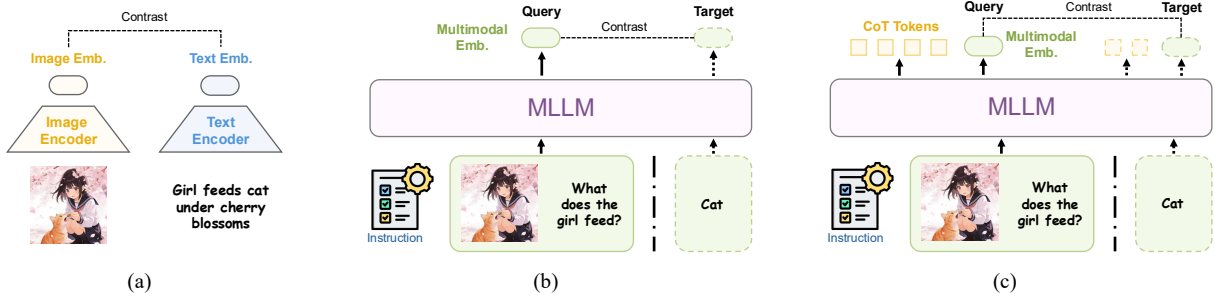


Figure 1: The evolution of multimodal embedding. (a) Early approaches employ modality-specific encoders to project different modalities into a shared semantic space. (b) MLLM-based methods process multimodal inputs with task instructions and are trained using semantically related query–target pairs. (c) Recent reasoning-enhanced embedding methods introduce chain-of-thought (CoT) reasoning prior to generating multimodal embeddings.

Abstract

MLLMs have been successfully applied to multimodal embedding tasks, yet their generative reasoning capabilities remain underutilized. Directly incorporating chain-of-thought reasoning into embedding learning introduces two fundamental challenges. First, structural misalignment between instance-level reasoning and pairwise contrastive supervision may lead to shortcut behavior, where the model merely learns the superficial format of reasoning. Second, reasoning is not universally beneficial for embedding tasks. Enforcing reasoning for all inputs may introduce unnecessary computation and latency, and can even obscure salient semantic signals for simple cases. To address these issues, we propose MMEmb-R1, an adaptive reasoning-based multimodal embedding framework. We formulate reasoning as a latent variable and introduce pair-aware reasoning selection to identify reasoning paths beneficial for query–target alignment. Furthermore, we adopt reinforcement learning to selectively invoke reasoning only when necessary. Experiments on the MMEB-V2 benchmark demonstrate that our model achieves a score of 71.2 with only 4B parameters, establishing a new state-of-the-art while significantly reducing reasoning overhead and inference latency.

* Work done during an internship at ByteDance.

† Chao Feng and Hongsheng Li are corresponding authors.

1 Introduction

Multimodal embedding models aim to project heterogeneous inputs, such as text, images, and interleaved image-text content, into a unified semantic space. They serve as a fundamental infrastructure for a wide range of applications, including recommendation systems (Lin et al., 2025; Zhang et al., 2025a), cross-modal retrieval (Faysse et al., 2025a; Wei et al., 2024), and retrieval-augmented generation (Yu et al., 2025b; Ren et al., 2025). Early work, exemplified by CLIP (Radford et al., 2021), leverages large-scale image-text pairs to align different modalities within a shared semantic space (Fig. 1(a)). More recently, multimodal large language models (MLLMs) have revolutionized this field (Meng et al., 2025; Zhang et al., 2025b; Jiang et al., 2025b) by providing rich world knowledge, compositional understanding, and strong instruction-following capabilities (Fig. 1(b)).

However, the current utilization of MLLMs in embedding models remains limited. Most existing approaches treat MLLMs primarily as static feature extractors, without fundamentally departing from the conventional paradigm. In contrast, the success of LLMs (Brown et al., 2020; Yang et al., 2025) and MLLMs (Comanici et al., 2025; OpenAI et al., 2024a) largely stems from their generative

capability: next-token prediction and the generative paradigm have substantially enhanced abstraction, reasoning, and structured understanding, giving rise to emergent abilities (Wei et al., 2022a). The embedding community has only marginally benefited from these strengths. This raises a fundamental question: *Can generative reasoning be effectively integrated into embedding learning, and if so, what is the appropriate formulation?*

By reexamining the paradigms of embedding and reasoning, as well as prior related works, we identify two key challenges. **(1) Structural misalignment between reasoning and representation learning may induce shortcut behavior.** Embedding models are trained under pairwise contrastive supervision, whereas reasoning is generated at the instance level. Existing pioneering reasoning-driven embedding models (Lan et al., 2025b; Cui et al., 2026), as illustrated in Fig. 1(c), typically require the model to learn or incorporate a single teacher-provided chain-of-thought (CoT) (Wei et al., 2022b) separately for the query and the target before generating the embedding. In this setup, reasoning quality is largely decoupled from the paired objective that governs contrastive representation learning. As shown in Fig. 2(a), embedding tokens in prior models such as UME-R1 (Lan et al., 2025b) attend heavily to the original input but minimally to CoT tokens, suggesting that reasoning is often treated as a deterministic procedural prefix rather than a latent variable subject to selection. Consequently, the model exhibits shortcut behavior: it mimics the surface format of reasoning without establishing a meaningful dependency between reasoning and the learned representation. **(2) Reasoning is not universally necessary for embedding tasks.** For simple or concise inputs, enforced autoregressive reasoning may induce “overthinking”, introducing unnecessary computation and latency (Kim et al., 2025). Moreover, excessive reasoning can obscure salient semantic signals and may even degrade performance by confusing the model, as shown in Fig. 2(b).

To harvest the ability of the generative paradigm and address above challenges, we propose MMEmb-R1, an adaptive Reasoning-based MultiModal Embedding framework. Instead of deterministically generating a single reasoning trajectory, we formulate the reasoning path as a latent variable and introduce a **pair-aware reasoning selection mechanism tailored to contrastive embedding**. Specifically, we employ multiple hetero-

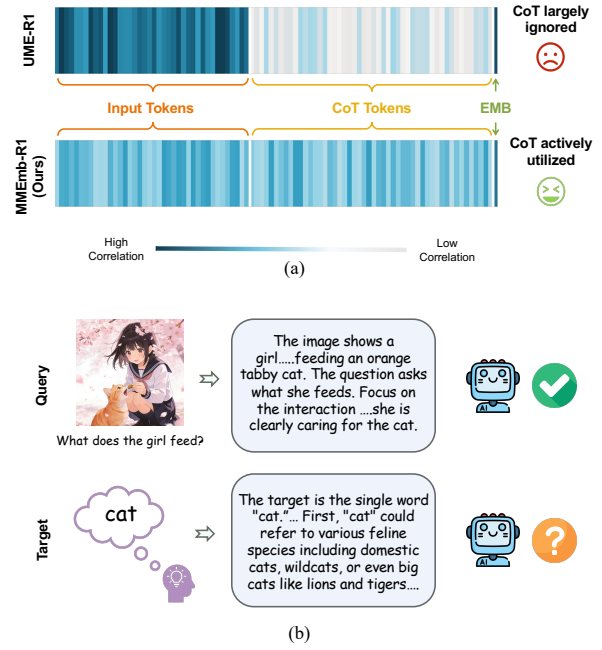


Figure 2: Two challenges of reasoning in embedding. (a) Shortcut behavior: UME-R1’s embedding token largely ignores CoT tokens, while MMEmb-R1 actively utilizes them. (b) Overthinking: reasoning helps the complex query (top) but introduces irrelevant noise for the simple target “cat” (bottom).

geneous worker MLLMs to generate diverse reasoning candidates, simulating a rich prior distribution over the latent reasoning space and mitigating single-teacher bias. We then design a pair-aware evaluator to score each reasoning path: by comparing the matching confidence with and without the rationale, we isolate its marginal contribution to query-target alignment, which subsequently guides model training. Furthermore, we develop an **adaptive reasoning mechanism** that explicitly models the utility of reasoning and mitigates unnecessary overthinking. We quantify the reasoning benefit by computing the similarity gap between reasoning-enhanced and direct embeddings. This continuous utility signal serves as a reward in reinforcement learning with GRPO (Guo et al., 2025), enabling the model to learn a policy that selectively invokes reasoning only when it provides substantial benefit. By integrating pair-aware selection with adaptive reasoning control, our framework achieves a principled balance between effectiveness and efficiency.

Extensive experiments on the MMEB-V2 benchmark (Meng et al., 2025) demonstrate the effectiveness of our approach. MMEmb-R1 achieves state-of-the-art performance across both small-size and medium-size settings, attaining 68.3 overall with

a Qwen3-VL-2B backbone and 71.2 with Qwen3-VL-4B, surpassing strong baselines such as Embed-RL (Jiang et al., 2026) (66.8) and RzenEmbed-v1 (Jian et al., 2025) (68.9) while using fewer parameters. The proposed adaptive mechanism increases inference throughput by $3.1\times$ compared to UME-R1 (Lan et al., 2025b), with improvement in retrieval accuracy. We hope this work offers a fresh perspective on reasoning-aware representation learning and opens new avenues for integrating generative paradigms into multimodal embedding.

2 Related Works

2.1 Multimodal Embedding Models

Multimodal embedding aims to learn compact, semantically meaningful representations for heterogeneous data. CLIP (Radford et al., 2021) established the dual-tower contrastive paradigm, training separate encoders via large-scale image-text alignment. Subsequent studies extended this paradigm to additional modalities like AudioCLIP (Guzhov et al., 2022) and CLIP4Clip (Luo et al., 2022). Other works further improve the contrastive learning paradigm by introducing novel training objectives or pre-training strategies, such as BLIP (Li et al., 2022) and SigLIP (Zhai et al., 2023). With the rise of MLLMs, the community has shifted toward MLLM-based embedding frameworks. Early representative works include VLM2Vec (Jiang et al., 2025b), GME (Zhang et al., 2025b), and ColPali (Faysse et al., 2025b). Building on this foundation, recent efforts have explored expanding modality coverage (Meng et al., 2025; Jian et al., 2025; Tzachor et al., 2026; Liu et al., 2025b), scaling data quality (Li et al., 2026a; Zhou et al., 2025; Gu et al., 2025b), and designing specialized architectures or training strategies (Chen et al., 2025a; Qin et al., 2025; Gu et al., 2026; Li et al., 2026b). More recently, several studies have explored incorporating generative reasoning into embedding learning. UME-R1 (Lan et al., 2025b) applies supervised fine-tuning to endow embedding models with reasoning capability; TTE (Cui et al., 2026) investigates diverse combinations of reasoners and embedders; and our concurrent work Embed-RL (Jiang et al., 2026) optimizes the reasoner to generate evidential chains of thought. While these pioneering efforts demonstrate the potential of reasoning for embedding, they largely overlook the structural misalignment between instance-level reasoning and pair-level contrastive supervision, which motivates

the design of MMEmb-R1.

2.2 Large Reasoning Models

Recent advances have shown that LLMs and MLLMs benefit substantially from enhanced reasoning capabilities (Guo et al., 2025; Comanici et al., 2025), as exemplified by OpenAI o1 (OpenAI et al., 2024b) and QwQ (Team, 2025). Early methods adopt chain-of-thought prompting (Kojima et al., 2022; Wang et al., 2023; Xu et al., 2025; Shao et al., 2024a) to elicit step-by-step rationales. Inspired by GRPO in DeepSeek-R1 (Guo et al., 2025), a growing body of work applies reinforcement learning to optimize reasoning trajectories across diverse domains, including visual understanding (Feng et al., 2025; Shen et al., 2025), text-to-image generation (Jiang et al., 2025a), mathematical reasoning (Lu et al., 2024; Zhang et al., 2024b), and domain-specific applications such as finance (Liu et al., 2026) and medicine (Lai et al., 2026). However, multimodal embedding and representation learning, an important subfield of multimodal learning, has yet to benefit from this paradigm much, a gap our work aims to bridge.

3 Methodology

We present our MMEmb-R1 framework, illustrated in Fig. 3, which consists of three stages: (1) constructing a pair-aware reasoning pool via diverse candidate generation and posterior selection (§ 3.2); (2) jointly training the model for reasoning generation and contrastive embedding (§ 3.3); and (3) adaptive reasoning control via utility-aware reinforcement learning (§ 3.4). We begin with preliminaries and an architectural overview in § 3.1.

3.1 Preliminaries and Framework Overview

Preliminaries of Multimodal Embedding.

Given a multimodal input $x = \{t, v\}$ consisting of text and visual (image or video) content, an embedding model $\mathcal{E}$ maps it to a d -dimensional representation $\mathbf{z} = \mathcal{E}(x) \in \mathbb{R}^d$. Training follows the contrastive paradigm: given a batch of N query-target pairs $\{(q_k, t_k^+)\}_{k=1}^N$, we compute embeddings $\mathbf{z}_{q_k} = \mathcal{E}(q_k)$ and $\mathbf{z}_{t_k} = \mathcal{E}(t_k^+)$. The objective pulls positive pairs closer while pushing in-batch negatives apart by optimizing the InfoNCE loss:

$$\mathcal{L}_{\text{con}} = -\frac{1}{N} \sum_{k=1}^N \log \frac{\exp(\text{sim}(\mathbf{z}_{q_k}, \mathbf{z}_{t_k})/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{z}_{q_k}, \mathbf{z}_{t_j})/\tau)}$$

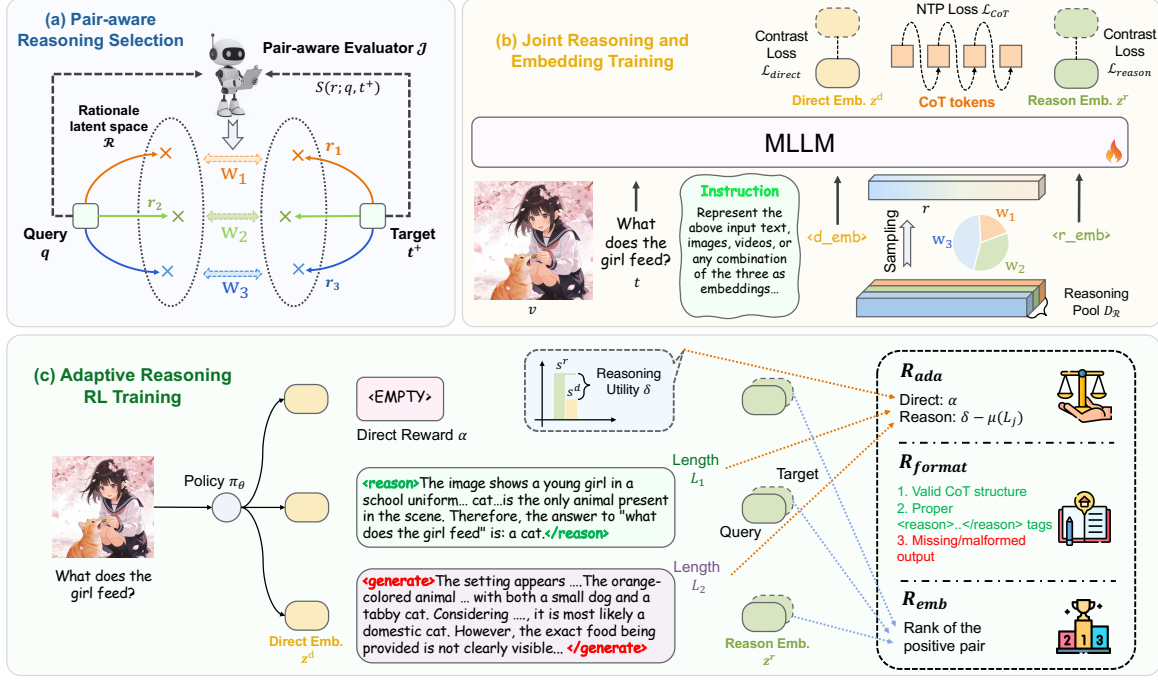


Figure 3: Overview of the MMEmb-R1 framework. **Upper Left:** Pair-aware reasoning selection—multiple heterogeneous workers generate diverse rationale candidates for the query and target, and an evaluator scores each candidate to produce selection weights w_1, w_2, w_3 . **Upper right:** Joint reasoning and embedding training—the MLLM is trained with a direct embedding path ($\mathcal{L}_{\text{direct}}$), a reasoning-enhanced embedding path ($\mathcal{L}_{\text{reason}}$), and a next-token prediction objective over CoT tokens ($\mathcal{L}_{\text{CoT}}$). **Lower:** Adaptive reasoning via GRPO—the policy π_θ decides whether to invoke reasoning or emit `<EMPTY>` for each query, guided by three reward signals: adaptive reward R_{ada} , format reward R_{format} , and embedding reward R_{emb} .

where τ is sampling temperature and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity.

Architecture Overview. MMEmb-R1 is built upon a multimodal large language model (MLLM). Visual inputs are first processed by a vision transformer (ViT) (Dosovitskiy et al., 2020) and projected into the language token space via a visual adapter, enabling unified sequence modeling across modalities. The model operates in two modes. In *direct mode*, the embedding is extracted from the hidden state of the final input special token: $\mathbf{z}^d = \mathcal{E}(x)$. In *reasoning mode*, the model first generates a reasoning path r conditioned on the input, and the embedding is derived from the final token after the reasoning trajectory: $\mathbf{z}^r = \mathcal{E}(x \oplus r)$, where $\oplus$ denotes sequence concatenation.

Reasoning as a Latent Variable. A central departure of MMEmb-R1 from prior work is the treatment of the reasoning path r as a *latent variable* rather than a deterministic output of a fixed teacher. Formally, we posit a latent reasoning space $\mathcal{R}$ with a prior distribution $\mathcal{P}(R)$, from which reasoning candidates are sampled: $r \sim \mathcal{P}(R)$. The

reasoning-enhanced embedding can then be written as a marginalization over this latent space: $\mathbf{z}^r = \mathbb{E}_{r \sim \mathcal{P}(R)} [\mathcal{E}(x \oplus r)]$. In practice, direct marginalization is intractable. Our framework addresses this by: (1) simulating $\mathcal{P}(R)$ through diverse multi-worker generation, (2) introducing a pair-aware scoring function $S(r; q, t^+)$ to perform structured posterior selection aligned with the contrastive objective, and (3) learning an adaptive policy that decides whether to sample from $\mathcal{P}(R)$ at all. We detail each component in the following sections.

3.2 Pair-Aware Reasoning Selection for Contrastive Embedding

As established in § 3.1, we model reasoning as a latent variable $r \sim \mathcal{P}(R)$ whose quality should be assessed under the joint query–target context.

3.2.1 Diverse Prior Simulation via Multi-Worker Generation

To approximate a rich prior $\mathcal{P}(R)$ and reduce single-teacher bias, we employ K heterogeneous worker MLLMs $\{M_k\}_{k=1}^K$ spanning complementary capabilities: (1) Instruct-based models (e.g., Qwen2-VL-Instruct (Wang et al., 2024)): pro-

duce concise, structured analyses of core semantics and retrieval-relevant keypoints. **(2)** Thinking models (e.g., GLM-4.1V-Thinking (Team et al., 2026)): generate exploratory, long-form reasoning chains that capture deeper analytical perspectives, though potentially with greater verbosity. **(3)** High-capacity proprietary models (e.g., Gemini 2.5 Pro (Comanici et al., 2025)): provide broad world knowledge and rich contextual coverage. As shown in Fig. 3(a), for each input x (either a query q or target t^+), each worker independently produces a candidate rationale $r_k = M_k(x)$, $k = 1, \dots, K$. Note that generation is still performed *single-sided* in this stage to avoid information leakage. The resulting candidates $\mathcal{R}_x = \{r_1, r_2, \dots, r_K\}$ collectively form empirical samples from the latent reasoning prior $\mathcal{P}(R)$. Detailed prompt and implementations can be found in Appendix. B.1.

3.2.2 Posterior Selection

Given samples from the prior, we perform *posterior selection*: identifying which reasoning paths are most useful for the pair (q_i, t_i^+) . Specifically, we employ an evaluator model $\mathcal{J}$ prompted to judge whether the query and target match, and extract the logit of the affirmative token [YES] as a confidence score. Inspired by causal intervention (Pearl, 2009), we isolate the contribution of reasoning by comparing the matching confidence computed without and with a rationale candidate: $c_0 = \text{Conf}_{\mathcal{J}}(q_i, t_i^+)$ and $c_r = \text{Conf}_{\mathcal{J}}(q_i, t_i^+, r)$. The *reasoning gain* is: $\Delta_r = c_r - c_0$, measuring how much rationale r improves recognizing query-target correspondence beyond raw input. Positive Δ_r indicates useful semantic bridging rather than mere rephrasing. This method can also serve as baseline calibration: subtracting the sample-wise “counterfactual baseline” c_0 yields a more normalized and robust scoring signal across different samples. We retain candidates with $\Delta_r > \epsilon$, forming $\mathcal{R}_i^+ = \{r \in \mathcal{R}_i \mid \Delta_r > \epsilon\}$, and normalize gains via softmax:

$$w_r = \frac{\exp(\Delta_r/\gamma)}{\sum_{r' \in \mathcal{R}_i^+} \exp(\Delta_{r'}/\gamma)}$$

where γ is a temperature controlling the sharpness of the selection distribution. This produces a weighted reasoning pool: $\mathcal{D}_R = \{(q_i, t_i^+, r_{i,j}, w_{i,j})\}_{i,j}$ where higher-gain reasoning paths contribute more strongly to subsequent training. More details can be found in Appendix B.2 and Appendix A.3.

3.3 Joint Reasoning and Embedding Training

With the curated reasoning pool $\mathcal{D}_R$ representing the selected posterior over latent reasoning, we fine-tune the MLLM to acquire: (1) contrastive alignment for embedding matching, and (2) coherent chain-of-thought generation internalizing the reasoning distribution. This is achieved through a multi-objective training scheme with two complementary embedding paths (Fig. 3(b)).

Reasoning-Enhanced Embedding Path. For each training pair (q_i, t_i^+) , we sample a reasoning path $r_{i,j}$ from $\mathcal{D}_R$ according to its posterior weight $w_{i,j}$. This path is optimized with the contrastive loss: $\mathcal{L}_{\text{reason}} = \mathcal{L}_{\text{con}}(\mathbf{z}_q^r, \mathbf{z}_t^r)$, where $\mathcal{L}_{\text{con}}$ follows the InfoNCE formulation defined above. To explicitly cultivate reasoning generation ability within the backbone, we additionally apply a next-token prediction loss over the chain-of-thought tokens:

$$\mathcal{L}_{\text{CoT}} = - \sum_{l=1}^{|r|} \log p_{\theta}(r_l \mid x, r_{<l}),$$

which trains the model to internalize the reasoning trajectories in $\mathcal{D}_R$ as generative knowledge.

Direct Embedding Path. To preserve embedding quality without reasoning overhead, we include a direct path encoding raw inputs as $\mathbf{z}^d = \mathcal{E}(x)$, optimized with: $\mathcal{L}_{\text{direct}} = \mathcal{L}_{\text{con}}(\mathbf{z}_q^d, \mathbf{z}_t^d)$.

Overall Objective. The complete training objective combines all three components:

$$\mathcal{L} = \mathcal{L}_{\text{reason}} + \lambda_{\text{CoT}} \mathcal{L}_{\text{CoT}} + \lambda_{\text{direct}} \mathcal{L}_{\text{direct}},$$

where λ_{CoT} and λ_{direct} are hyperparameters.

3.4 Adaptive Reasoning Control via Utility-Aware Optimization

While the joint training stage equips the model with reasoning capability, not all inputs benefit from explicit reasoning as discussed in § 1. We therefore introduce a reinforcement learning stage that trains the model to *selectively invoke reasoning only when it provides measurable benefit*.

3.4.1 Reasoning Utility Estimation

We estimate reasoning utility from the embedding geometry learned after the joint training stage. For each query q_i in the reinforcement learning dataset, we compute its similarity with the corresponding target using both normalized direct embeddings and reasoning-enhanced embeddings produced by

the jointly trained model, yielding s_i^d and s_i^r , respectively. We then define the *reasoning utility gap* as $\delta_i = s_i^r - s_i^d$. This continuous signal quantifies the marginal benefit of reasoning for each instance: $\delta_i > 0$ indicates that reasoning improves retrieval quality, whereas $\delta_i \leq 0$ suggests that direct embedding is sufficient or even preferable. Importantly, we treat δ_i as a continuous intrinsic signal rather than a binary supervision label, enabling more fine-grained and stable policy learning.

3.4.2 Policy Optimization with GRPO

We formulate adaptive reasoning as a sequential decision-making problem (Puterman, 1990; Chen et al., 2023). As shown in Fig. 3(c), for each query q_i , the model selects an action $a_i \in \{\text{DIRECT}, \text{REASON}\}$, indicating whether to generate the embedding directly or to invoke reasoning before embedding. If the REASON action is selected, the model first generates a rationale and then produces the embedding conditioned on it. We design a reward function that balances retrieval improvement and computational cost:

$$R_{\text{ada}} = \begin{cases} \alpha, & a_i = \text{DIRECT} \wedge (n \leq N) \\ \delta_i - \mu(L_i), & a_i = \text{REASON} \end{cases}$$

where α is a positive constant that encourages exploration of the DIRECT action during the early stage of training (i.e., when the training step $n \leq N$), mitigating the tendency to always generate rationales inherited from the supervised stage. For the REASON action, L_i denotes the length of the generated reasoning, and $\mu(\cdot)$ controls the trade-off between performance gain and computational overhead. In particular, we penalize excessively long rationales by applying an additional coefficient c when the reasoning length exceeds 512 tokens. Following Lan et al. (2025b), we also incorporate an embedding reward R_{emb} , which evaluates embedding quality based on the ranking position of the positive target among in-batch negatives, and a format reward R_{format} to ensure that the generated rationales follow the required output structure. To ensure symmetry, we additionally compute the reward in the reverse direction (target $\rightarrow$ query) and take the mean of the two scores. More details can be found in Appendix B.4. The overall objective is optimized using Group Relative Policy Optimization (GRPO) (Shao et al., 2024b), which maximizes the expected reward:

$$\max_{\theta} \mathbb{E}_{a_i \sim \pi_{\theta}(\cdot|q_i)} [R_{\text{ada}} + R_{\text{format}} + R_{\text{emb}}].$$

4 Experiments

4.1 Setup

4.1.1 Implementation Details

We build MMEmb-R1 on the Qwen-VL family. For diverse prior simulation, we use GLM-4.1V-Thinking (Team et al., 2026), InternVL3-14B-Instruct (Zhu et al., 2025), and Doubao-Seed-1.6-Vision (ByteDance, 2025). The pair-aware evaluator $\mathcal{J}$ is Qwen3-VL-32B-Instruct (Yang et al., 2025). For joint training, we use batch size 32 per GPU under DeepSpeed ZeRO-3, a cosine learning rate schedule with initial rate 5×10^{-5} , and train for 3 epochs. For adaptive reasoning, we use GRPO (Shao et al., 2024b) with $\alpha = 0.2$ and $\beta = 0.04$. All experiments run on $8 \times$ H20 90GB GPUs. See Appendix B for details.

4.1.2 Training Datasets and Benchmark

Following prior work, we use MMEB-Train (Meng et al., 2025) for training. After data filtering and pair-aware selection (§ 3.2.2), we obtain $\sim 1.2\text{M}$ samples for joint embedding and reasoning training and $\sim 10\text{K}$ for adaptive reasoning reinforcement learning. For evaluation, we use MMEB-V2 (Meng et al., 2025), a comprehensive benchmark covering 78 tasks across classification, VQA, retrieval, and visual grounding. Following the standard evaluation protocol, we report Hit@1 for image/video tasks and NDCG@5 for visual document tasks.

4.2 Main Results

Baselines. We compare MMEmb-R1 against a broad set of multimodal embedding models, including classical MLLM-based models such as GME (Zhang et al., 2025b), ColPali (Faysse et al., 2025b), VLM2Vec (Jiang et al., 2025b), and VLM2Vec-V2 (Meng et al., 2025); recent methods like LamRA (Liu et al., 2025c), CAFe (Yu et al., 2025a), and RzenEmbed (Jian et al., 2025); and reasoning-driven models including UME-R1 (Lan et al., 2025b), TTE (Cui et al., 2026), and Embed-RL (Jiang et al., 2026). All methods are evaluated on MMEB-V2 (Meng et al., 2025) across Image, Video, and VisDoc modalities (Tab. 1).

Analysis. We can see that MMEmb-R1 achieves state-of-the-art performance in both size categories. With a Qwen3-VL-2B backbone, MMEmb-R1 attains 68.3 overall, surpassing the strongest baselines Embed-RL and RzenEmbed-v1 by +1.5 and +3.9 points respectively. Scaling to Qwen3-VL-

Model	Backbone	Image					Video					VisDoc					All
		CLS	QA	RET	GD	Overall	CLS	QA	RET	MRET	Overall	VDRv1	VDRv2	VR	OOD	Overall	
Small-size Models																	
GME	Qwen2-VL-2B	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
ColPali-V1.3	PaliGemma-3B	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
VLM2Vec	Qwen2-VL-2B	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-V2	Qwen2-VL-2B	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
UME-R1	Qwen2-VL-2B	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
TTE _g [†]	Qwen3-VL-2B	67.9	66.6	70.2	84.1	<u>70.1</u>	47.3	49.1	34.4	33.2	32.1	77.5	53.2	83.2	41.1	68.8	63.1
RzenEmbed-v1	Qwen2-VL-2B	65.3	61.7	73.8	77.9	68.5	45.6	47.5	38.3	36.7	42.6	87.0	57.6	85.4	43.3	74.4	64.4
Embed-RL	Qwen3-VL-2B	62.8	67.9	68.6	90.4	69.2	57.0	55.9	45.1	49.4	<u>52.1</u>	79.9	52.0	84.6	65.7	<u>74.1</u>	<u>66.8</u>
MMEmb-R1 (Ours)	Qwen2-VL-2B	64.5	68.1	70.0	88.9	70.0	56.3	52.8	47.6	42.5	50.6	74.1	56.0	78.9	48.3	68.0	65.0
MMEmb-R1 (Ours)	Qwen3-VL-2B	63.5	73.7	70.2	89.8	71.5	59.8	60.3	50.6	49.2	55.6	82.0	55.7	80.7	56.7	73.2	68.3
Medium-size Models																	
GME	Qwen2-VL-7B	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	Qwen2-VL-7B	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	Qwen2.5-VL-7B	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	Qwen2-VL-7B	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
CAFe	LLaVA-OV-7B	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
UME-R1	Qwen2-VL-7B	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
Embed-RL	Qwen3-VL-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
TTE _g [†]	Qwen2-VL-7B	69.7	72.4	74.0	90.6	74.2	49.1	60.6	36.4	37.2	46.8	84.1	62.7	91.9	47.6	76.4	68.6
RzenEmbed-v1	Qwen2-VL-7B	69.8	68.7	76.8	85.7	73.6	52.8	56.2	41.9	41.8	48.9	89.5	60.8	87.9	44.4	76.8	68.9
MMEmb-R1 (Ours)	Qwen2-VL-7B	68.1	71.8	77.7	87.1	<u>74.4</u>	60.2	56.0	51.2	41.1	<u>53.3</u>	79.6	65.3	84.7	58.0	74.9	<u>69.7</u>
MMEmb-R1 (Ours)	Qwen3-VL-4B	67.7	74.2	74.5	94.9	74.8	60.7	61.1	51.6	50.6	56.6	83.0	60.6	83.7	66.6	<u>76.7</u>	71.2

Table 1: Comparison of baseline methods and MMEmb-R1 on MMEB-V2. Given the diversity of model backbones, we aggregate results by model size. Models with 2B–3B parameters are categorized as *small*, while those with 4B–7B parameters are categorized as *medium*. [†] indicates that for the TTE model, we adopt the student variant to ensure a fair comparison without relying on a large external teacher model. Metrics are abbreviated as follows: **CLS** (classification), **QA** (question answering), **RET** (retrieval), **GD** (grounding), **MRET** (moment retrieval), **VDR** (ViDoRe), **VR** (VisRAG), and **OOD** (out-of-domain).

4B yields 71.2, outperforming the best medium-size baseline RzenEmbed-v1-7B with nearly half the parameters. Notably, MMEmb-R1 at 2B already surpasses several 7B baselines, suggesting that high-quality reasoning can partially compensate for the capacity gap. The improvements are consistent across modality groups but particularly pronounced on Video, where MMEmb-R1 (Qwen3-VL-2B) achieves 55.6, outperforming Embed-RL by +3.5. This aligns with our expectation: video understanding demands compositional reasoning over temporal dynamics, precisely the setting where pair-aware latent reasoning provides the greatest benefit. Detailed results and results on MMEB-V1 can be found in Appendix A.5. We also provide qualitative analysis in Appendix A.1.

4.3 Further Analysis

4.3.1 Further Analysis of Adaptive Reasoning

Inference latency comparison. As discussed in § 3.4, our adaptive reasoning mechanism eliminates unnecessary reasoning paths, improving performance while reducing the inference latency of reasoning-driven embedding models. To verify this, we compare inference throughput and performance under different inference strategies. We

Strategy	Throughput	Accuracy
MMEmb-R1 (Direct)	22.70 queries/s	59.2
UME-R1	3.21 queries/s	60.1
MMEmb-R1 (Reason)	5.79 queries/s	63.6
MMEmb-R1 (Adaptive)	10.02 queries/s	65.0

Table 2: Inference throughput on a subset and overall accuracy on MMEB-V2.

measure end-to-end latency using a batch size of 24 on 8×H20 GPUs. For a fair comparison with UME-R1-2B, we use Qwen2VL-2B as the backbone. For the “always-reasoning” setting, we apply reinforcement learning but set the adaptive reward R_{ada} to zero. Results are summarized in Tab. 2. MMEmb-R1 Adaptive achieves 10.02 queries per second, a $1.7\times$ speedup over the always-reason variant and $3.1\times$ over UME-R1, while delivering the highest accuracy. The direct variant achieves the highest throughput but suffers a severe performance drop. The throughput gap between MMEmb-R1 Always-reason and UME-R1 reflects the efficiency of our pair-aware reasoning selection, which produces more concise rationales than UME-R1’s verbose single-teacher chains. The further improvement from always-reason to adaptive confirms that the learned policy effectively skips unnecessary rea-

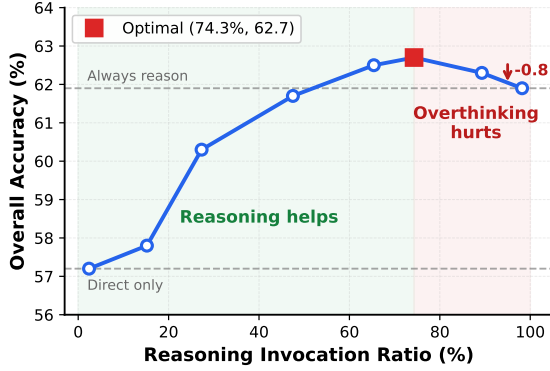


Figure 4: Reasoning invocation ratio vs. overall accuracy. The curve peaks at 74.3% reasoning ratio, after which accuracy declines.

soning for simple queries, yielding a model that is both faster and more accurate.

Pareto Frontier between Reasoning and Accuracy The coefficient c , which controls the trade-off between reasoning benefits and the length budget, provides an informative lens for analyzing how the adaptive policy allocates reasoning budget. In this experiment, we remove the 512-token limit and directly vary the cost coefficient c , tracing the resulting trade-off between reasoning invocation ratio and retrieval accuracy in Fig. 4. Each point corresponds to a policy trained with a different w and evaluated on a subset of MMEB-V2 for efficiency. As shown in the figure, the curve increases slowly at first and then rises steeply from around 57.2 to 62.7 at a reasoning ratio of 74.3%. Beyond this point, performance declines slightly to 61.9 under near-universal reasoning, representing a 0.8-point drop that empirically confirms the overthinking phenomenon discussed in § 1. These results indicate that the adaptive policy learns to prioritize the most reasoning-critical instances first: the earliest queries selected for reasoning yield the highest marginal returns, while those added later contribute diminishing or even negative gains.

4.3.2 Ablation Studies

We conduct ablation studies to validate our design choices (Tab. 3). (1) Replacing the diverse multi-worker prior with a single teacher causes the largest drop, confirming that diverse candidates cover the reasoning space more effectively. Uniform sampling without pair-aware scoring degrades performance by 2.2 points, and removing the “counterfactual baseline” c_0 causes an addi-

Variant	Score	Δ
MMEmb-R1 (Full)	65.0	–
<i>Pair-aware Reasoning Selection</i>		
Single-teacher rationale	61.2	–3.8
w/o pair-aware selection (uniform)	62.8	–2.2
w/o baseline calibration (use c_r only)	64.1	–0.9
<i>Training Objective</i>		
w/o $\mathcal{L}_{\text{reason}}$ (Direct only)	59.2	–5.8
<i>Adaptive Reasoning</i>		
Always reason	63.6	–1.4
Always direct	60.4	–4.6
Random (50%)	60.6	–4.4
Oracle	66.2	+1.2

Table 3: Ablation study on MMEB-V2. Each row removes or modifies one component from the full MMEmb-R1 framework.

tional 0.9-point drop, indicating both selection and baseline calibration filter genuinely useful rationales. (2) Removing the reasoning path entirely (Direct only) results in a 5.8-point drop, establishing reasoning-enhanced representations as the primary driver of our framework. (3) The always-reason variant achieves 63.6, 1.4 points lower than the full model, confirming that indiscriminate reasoning harms simple queries. The always-direct and random 50% strategies perform comparably, suggesting naive stochastic selection provides no advantage over skipping reasoning, whereas the learned policy captures meaningful structure. The oracle strategy (selecting the better of direct vs. reasoning-enhanced embedding) provides an upper bound of 66.2, indicating our learned policy recovers most achievable gain.

5 Conclusion

In this paper, we present MMEmb-R1, a framework integrating generative reasoning into multimodal embedding learning. To address the misalignment between instance-level reasoning and pair-level contrastive supervision, we introduce a pair-aware selection mechanism. To mitigate overhead from indiscriminate reasoning, we develop a utility-aware reinforcement learning stage for selective reasoning invocation. Experiments on MMEB-V2 demonstrate state-of-the-art performance with substantially reduced inference latency. We hope this work inspires further research in reasoning-driven multimodal representation learning.

Limitations

Our framework has several limitations that warrant future investigation. First, the pipeline nature of our approach—offline reasoning generation, pair-aware selection, and two-stage training—prevents joint optimization of these components. An end-to-end formulation that unifies reasoning generation, selection, and adaptive invocation within a single training loop could improve overall performance. Second, the reward design for the adaptive policy currently considers only binary decisions (whether to invoke reasoning or not), which may be suboptimal. Extending the policy to control reasoning depth or granularity (e.g., brief versus detailed reasoning chains) could enable more fine-grained allocation of computational resources. Third, reasoning-enhanced embedding inevitably incurs additional inference cost. While precomputing embeddings for the corpus side partially alleviates this in retrieval scenarios, fundamentally reducing the latency of reasoning-based models remains an open challenge.

References

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- ByteDance. 2025. [Doubao-seed-1.6-vision](#).
- Haonan Chen, Hong Liu, Yuping Luo, Liang Wang, Nan Yang, Furu Wei, and Zhicheng Dou. 2025a. [Moca: Modality-aware continual pre-training makes better bidirectional multimodal embeddings](#). *Preprint*, arXiv:2506.23115.
- Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. 2025b. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8254–8275.
- Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Tianyu Liu, and Baobao Chang. 2023. Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond. *arXiv preprint arXiv:2310.02071*.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.
- Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Xuanming Cui, Jianpeng Cheng, Hong you Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Kumar Mishra, Yonghuan Yang, Jun Xiao, Qi Guo, Ser-Nam Lim, Aashu Singh, and Xiangjun Fan. 2026. [Think then embed: Generative context improves multimodal embedding](#). *Preprint*, arXiv:2510.05014.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. [An image is worth 16x16 words: Transformers for image recognition at scale](#). *ArXiv*, abs/2010.11929.
- Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025a. [Colpali: Efficient document retrieval with vision language models](#). *Preprint*, arXiv:2407.01449.
- Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2025b. [Colpali: Efficient document retrieval with vision language models](#). *Preprint*, arXiv:2407.01449.
- Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. [Video-r1: Reinforcing video reasoning in mllms](#). *Preprint*, arXiv:2503.21776.
- Geonmo Gu, Byeongho Heo, Jaemyung Yu, Jaehui Hwang, Taekyung Kim, Sangmin Lee, HeeJae Jun, Yoohoon Kang, Sangdoo Yun, and Dongyoon Han. 2026. [Muco: Multi-turn contrastive learning for multimodal embedding model](#). *Preprint*, arXiv:2602.06393.
- Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. 2025a. Breaking the modality barrier: Universal embedding learning

- with multimodal llms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2860–2869.
- Tiancheng Gu, Kaicheng Yang, Kaichen Zhang, Xiang An, Ziyong Feng, Yueyi Zhang, Weidong Cai, Jiankang Deng, and Lidong Bing. 2025b. [Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning](#). *Preprint*, arXiv:2510.13515.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.
- Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. 2025. [Rzen-embed: Towards comprehensive multimodal retrieval](#). *Preprint*, arXiv:2510.27350.
- Dongzhi Jiang, Ziyu Guo, Renrui Zhang, Zhuofan Zong, Hao Li, Le Zhuo, Shilin Yan, Pheng-Ann Heng, and Hongsheng Li. 2025a. [T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot](#). *Preprint*, arXiv:2505.00703.
- Haonan Jiang, Yuji Wang, Yongjie Zhu, Xin Lu, Wenyu Qin, Meng Wang, Pengfei Wan, and Yansong Tang. 2026. [Embed-r1: Reinforcement learning for reasoning-driven multimodal embeddings](#). *Preprint*, arXiv:2602.13823.
- Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. 2025b. [Vlm2vec: Training vision-language models for massive multimodal embedding tasks](#). *Preprint*, arXiv:2410.05160.
- Wongyu Kim, Hochang Lee, Sanghak Lee, Yoonsung Kim, and Jaehyun Park. 2025. [Let multimodal embedders learn when to augment query via adaptive query augmentation](#). *Preprint*, arXiv:2511.02358.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. 2026. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE Transactions on Medical Imaging*.
- Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. 2025a. [Llave: Large language and vision embedding models with hardness-weighted contrastive learning](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. 2025b. Ume-r1: Exploring reasoning-driven generative multimodal embeddings. *arXiv preprint arXiv:2511.00405*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.
- Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibao Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2026a. [Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking](#). *Preprint*, arXiv:2601.04720.
- Qi Li, Yanzhe Zhao, Yongxin Zhou, Yameng Wang, Yandong Yang, Yuanjia Zhou, Jue Wang, Zuo-jian Wang, and Jinxiang Liu. 2026b. Magic-mm-embedding: Towards visual-token-efficient universal multimodal embedding with mllms. *arXiv preprint arXiv:2602.05275*.
- Lin Lin, Jiefeng Long, Zhihe Wan, Yuchi Wang, Dingkan Yang, Shuang Yang, Yueyang Yao, Xu Chen, Zirui Guo, Shengqiang Li, Weiran Li, Hanyu Li, Yaling Mou, Yan Qiu, Haiyang Yu, Xiao Liang, Hongsheng Li, and Chao Feng. 2025. [Sail-embedding technical report: Omni-modal embedding foundation model](#). *Preprint*, arXiv:2510.12709.
- Keliang Liu, Dingkan Yang, Ziyun Qian, Weijie Yin, Yuchi Wang, Hongsheng Li, Jun Liu, Peng Zhai, Yang Liu, and Lihua Zhang. 2025a. Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle. *arXiv preprint arXiv:2509.16679*.
- Qianying Liu, Xiao Liang, Zhiqiang Zhang, Zhongfei Qing, Fengfan Zhou, Yibo Chen, Xu Tang, Yao Hu, and Paul Henderson. 2025b. Rematch: Boosting representation through matching for multimodal retrieval. *arXiv preprint arXiv:2511.19278*.
- Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. 2025c. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4015–4025.

- Zhaowei Liu, Xin Guo, Zhi Yang, Fangqi Lou, Lingfeng Zeng, Mengping Li, Qi Qi, Zhiqiang Liu, Yiyang Han, Dongpo Cheng, Ronghao Chen, Huacan Wang, Xingdong Feng, Huixia Judy Wang, Chengchun Shi, and Liwen Zhang. 2026. [Fin-r1: A large language model for financial reasoning through reinforcement learning](#). *Preprint*, arXiv:2503.16252.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. [Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts](#). *Preprint*, arXiv:2310.02255.
- Huashao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.
- Rui Meng, Ziyang Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhui Chen, and Semih Yavuz. 2025. [Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents](#). *Preprint*, arXiv:2507.04590.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024b. [Openai o1 system card](#). *Preprint*, arXiv:2412.16720.
- Judea Pearl. 2009. *Causality*. Cambridge university press.
- Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.
- Martin L Puterman. 1990. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434.
- Jiajun Qin, Yuan Pu, Zhuolun He, Seungeun Kim, David Z. Pan, and Bei Yu. 2025. [Unimoco: Unified modality completion for robust multi-modal embeddings](#). *Preprint*, arXiv:2505.11815.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). *Preprint*, arXiv:2103.00020.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2025. [Videorag: Retrieval-augmented generation with extreme long-context videos](#). *Preprint*, arXiv:2502.01549.
- Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, et al. 2024. Ro-cov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1):688.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024a. Visual cot: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024b. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. 2025. [Vlm-r1: A stable and generalizable r1-style large vision-language model](#). *Preprint*, arXiv:2504.07615.
- Qwen Team. 2025. [Qwq-32b: Embracing the power of reinforcement learning](#).
- V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihang Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026. [Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning](#). *Preprint*, arXiv:2507.01006.
- Issar Tzachor, Dvir Samuel, and Rami Ben-Ari. 2026. Vidvec: Unlocking video mllm embeddings for video-text retrieval. *arXiv preprint arXiv:2602.08099*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei

- Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. [Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution](#). *Preprint*, arXiv:2409.12191.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). *Preprint*, arXiv:2203.11171.
- Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhui Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pages 387–404. Springer.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. [Emergent abilities of large language models](#). *Preprint*, arXiv:2206.07682.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098.
- Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, et al. 2023. Context de-confounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19005–19015.
- Dingkang Yang, Kun Yang, Haopeng Kuang, Zhaoyu Chen, Yuzheng Wang, and Lihua Zhang. 2024. Towards context-aware emotion recognition debiasing from a causal demystification perspective via de-confounded training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10663–10680.
- Hao Yu, Zhuokai Zhao, Shen Yan, Lukasz Korycki, Jianyu Wang, Baosheng He, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, and Hanchao Yu. 2025a. Cafe: Unifying representation and generation with contrastive-autoregressive finetuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6286–6297.
- Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025b. [Visrag: Vision-based retrieval-augmented generation on multi-modality documents](#). *Preprint*, arXiv:2410.10594.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.
- Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Xiangyu Zhao, Yan Gao, Yao Hu, and Enhong Chen. 2025a. Notellm-2: Multimodal large representation models for recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2815–2826.
- Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhui Chen, Yu Su, and Ming-Wei Chang. 2024a. [Magiciens: Self-supervised image retrieval with open-ended instructions](#). *Preprint*, arXiv:2403.19651.
- Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, Peng Gao, Chunyuan Li, and Hongsheng Li. 2024b. [Mavis: Mathematical visual instruction tuning with an automatic data engine](#). *Preprint*, arXiv:2407.08739.
- Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025b. [Gme: Improving universal multimodal retrieval by multimodal llms](#). *Preprint*, arXiv:2412.16855.
- Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. 2025. 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)*, pages 19076–19095.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. [Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models](#). *Preprint*, arXiv:2504.10479.

A Additional Experimental Results

A.1 Qualitative Analysis

We conduct qualitative analyses to demonstrate the capability of MMEmb-R1 and illustrate several design principles. For simplicity, we present only the main reasoning traces produced by the model.

Fig. 7 presents two retrieval cases highlighting the advantages of adaptive reasoning in MMEmb-R1. In the upper case, the query is a cartoon penguin, which is visually unambiguous. MMEmb-R1 adaptively skips reasoning and correctly retrieves “penguin,” whereas UME-R1’s enforced reasoning introduces spurious alternatives (“penguin, magpie, or puffin”) and ultimately retrieves the wrong target. In the lower case, the query is a cooking video that requires temporal inference. MMEmb-R1 invokes reasoning and correctly decomposes the cooking sequence, inferring that “the logical next step after stir-frying is to add seasoning.” In contrast, the non-reasoning baseline VLM2Vec-V2 appears to capture only the coarse semantic concept of cooking and retrieves a temporally preceding action instead. These examples demonstrate that MMEmb-R1 learns when reasoning is beneficial and when it is unnecessary.

Fig. 8 further illustrates why diverse workers combined with pair-aware selection outperform single-source reasoning approaches such as UME-R1 and TTE. Given a chart query asking “How common was it for people to feel depressed during the outbreak?” and a ground-truth target stating that “about one in four Americans (24%) reported feeling depressed some or a little of the time, while 9% felt depressed most or all of the time,” the three workers exhibit complementary strengths and weaknesses. The Instruct worker ($w = 0.28$) correctly extracts the relevant numbers (9%, 15%, 24%, 52%) but merely lists them without interpreting which frequency band each corresponds to, leaving a gap between the raw data and the natural-language phrasing of the target. The Thinking worker ($w = 0.17$) produces a detailed cross-category comparison—contrasting depression with anxiety and discussing clinical instruments—but this exhaustive analysis diverges from the specific query, burying the target-relevant information. The Proprietary worker ($w = 0.55$) receives the highest weight because it directly mirrors the target semantics: it rephrases “24%” as “about one in four respondents,” associates “9%” with the “most or all of the time” frequency band, and provides

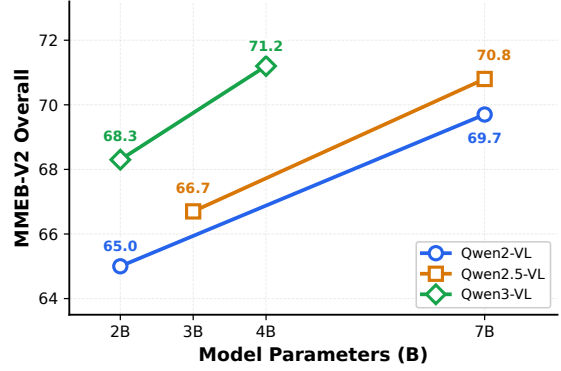


Figure 5: Scaling behavior of MMEmb-R1 across backbone families and parameter scales. Performance improves consistently within each family, and newer architectures achieve higher scores at comparable or smaller model sizes.

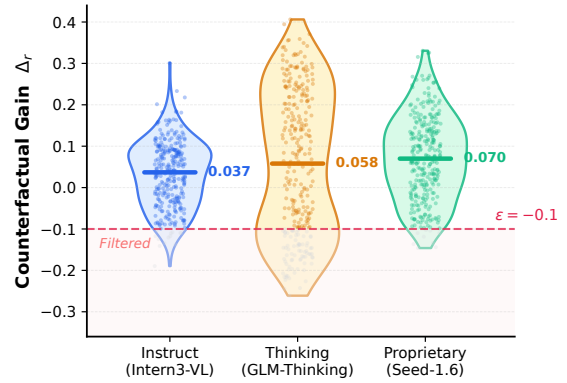


Figure 6: Distribution of reasoning gains Δ_r across worker types. Median values are shown as horizontal bars. The dashed line indicates the selection threshold $\epsilon = -0.1$; candidates below are filtered out.

the complementary statistic that the majority rarely felt depressed. These examples illustrate that the pair-aware evaluator identifies reasoning paths that effectively bridge the semantic gap between a specific query and its target, rather than favoring reasoning that is merely more elaborate or complex.

A.2 Scaling Behavior Across Backbones

To assess the generality and scalability of MMEmb-R1, we apply our framework to six backbone MLLMs spanning three model families and varying parameter scales: Qwen2-VL (2B, 7B), Qwen2.5-VL (3B, 7B), and Qwen3-VL (2B, 4B). Fig. 5 reports the overall MMEB-V2 performance for each configuration.

Two observations emerge. First, MMEmb-R1 exhibits consistent intra-family scaling: perfor-

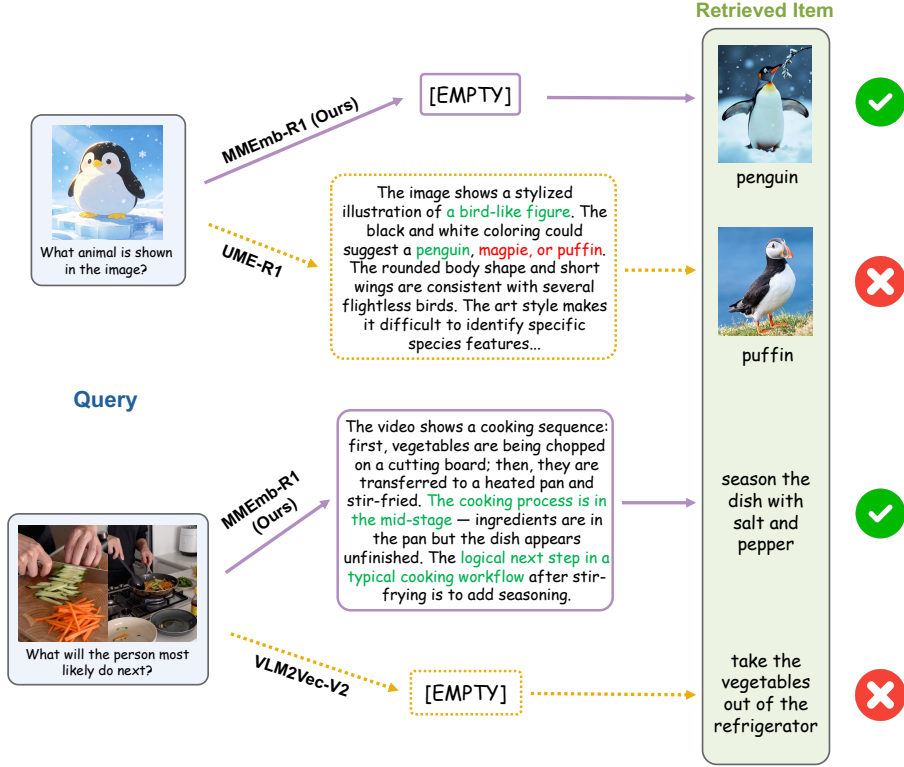


Figure 7: Adaptive reasoning: MMEmb-R1 skips reasoning for a simple visual query (top, avoiding overthinking) and invokes it for a complex temporal query (bottom), outperforming UME-R1 and VLM2Vec-V2 respectively.

mance improves monotonically with model size across all three families, indicating that our framework effectively leverages the additional capacity of larger backbones without saturating. Second, the gains from backbone architecture advancement are substantial and largely orthogonal to those from scaling. Qwen3-VL-2B surpasses Qwen2-VL-7B at less than one-third of the parameters, and Qwen3-VL-4B outperforms Qwen2.5-VL-7B at roughly half the size. This suggests that MMEmb-R1 benefits from both stronger representations and larger capacity, and that the pair-aware reasoning selection and adaptive invocation mechanisms transfer effectively across architectures without architecture-specific tuning.

A.3 Reasoning Gain Distribution

Fig. 6 presents the distribution of “counterfactual reasoning gains” Δ_r across the three worker types used in our diverse prior simulation (values are rescaled for better visualization). We can see that no single worker dominates the distribution. The Proprietary worker achieves the highest median gain and exhibits the most compact, positively skewed distribution, while the Instruct worker produces a tighter but lower-centered distribution. In

contrast, the Thinking worker shows the widest spread with a clearly bimodal shape—its upper mode reaches the highest individual Δ_r values among all workers, yet its lower mode extends well into negative territory. This observation supports our hypothesis that thinking models generate exploratory reasoning chains that are occasionally exceptional but frequently noisy, making them a high-variance complement to the more conservative Instruct and Proprietary workers. This complementarity further motivates our latent-variable formulation: diverse samples from heterogeneous workers collectively approximate a richer reasoning space than any single source, while the pair-aware scoring mechanism assigns higher weights to higher-quality samples.

We adopt a lenient threshold $\epsilon = -0.1$ across all worker types, allowing samples whose reasoning introduces only a small performance drop to be retained. As shown in Fig. 6, the Thinking worker produces the largest number of filtered samples, indicating that many of its reasoning chains substantially harm query–target alignment. Importantly, this does not imply that most samples are simply retained such that the selection mechanism becomes ineffective. Instead, the key aspect of our design

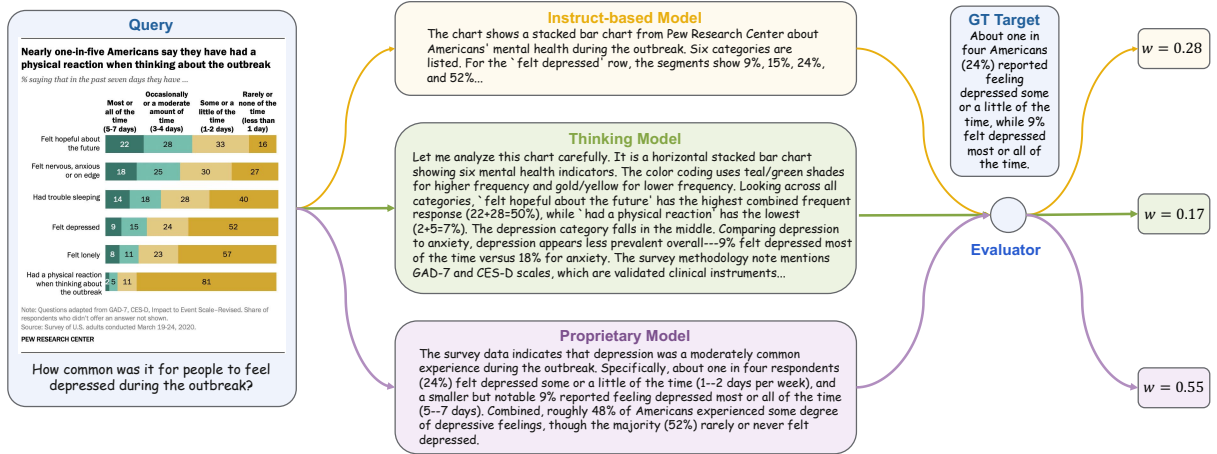


Figure 8: Pair-aware reasoning selection: three heterogeneous workers produce complementary rationales for the same query. The evaluator assigns the highest weight to the Proprietary worker ($w = 0.55$), which best bridges the query–target semantic gap.

lies in the relative weighting among the accepted samples, which is determined by the pair-aware alignment scores across different workers.

A.4 Distribution of Reasoning Utility

As described in § 3.4.1, we compute the reasoning utility $\delta_i = s_i^r - s_i^d$ for each training query by comparing the normalized similarity scores obtained from reasoning-enhanced and direct embeddings. Fig. 9 shows the distribution of δ_i across a subset. The distribution is unimodal and centered slightly above zero, with a longer right tail than left. Roughly 60% of instances exhibit positive utility, confirming that reasoning-enhanced embeddings are generally beneficial after pair-aware selection. However, a substantial 40% of instances show negative utility, meaning that reasoning introduces noise or obscures salient signals for these samples. This mixed landscape directly motivates our adaptive reasoning mechanism: a one-size-fits-all strategy—whether always reasoning or never reasoning—is inherently suboptimal. The continuous, instance-dependent nature of δ_i further justifies using reinforcement learning rather than a hard threshold to learn the decision boundary, as the optimal reasoning policy must account for fine-grained variations across inputs.

A.5 Detailed Results on MMEB-V2 and MMEB-V1

We provide per-dataset results on MMEB-V2 (78 datasets, three modality groups) in Tab. 12. For compatibility with prior work evaluated on the original image-only benchmark, we also report MMEB-

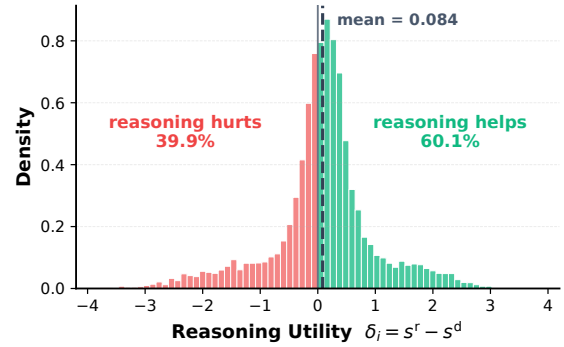


Figure 9: Distribution of reasoning utility δ_i over the training set. Green bars ($\delta_i \geq 0$) indicate instances where reasoning improves retrieval; red bars ($\delta_i < 0$) indicate the opposite.

V1 results (36 datasets) in Tab. 13. MMEB-R1 (Qwen3-VL-4B) achieves 74.8 overall on V1, outperforming all baselines including Embed-RL-4B and UME-R1-7B, confirming that the benefits of our framework are not specific to the video and document modalities newly introduced in V2.

To assess whether the performance improvements of our model are statistically significant across tasks, we conduct paired statistical tests based on the per-task scores reported in Table 12. We compare MMEB-R1 (Qwen3-VL-4B) against several strong baselines using the Wilcoxon signed-rank test as the primary non-parametric test, a paired t -test as a complementary analysis, and task-level bootstrap 95% confidence intervals with 10k resamples over the paired differences. As shown in Table 5, MMEB-R1 consistently outperforms the compared baselines across the 78 tasks. In partic-

== INSTRUCTION TO INSTRUCT-BASED AND HIGH-CAPACITY PROPRIETARY MODELS ==

The input above is a query or a candidate item for retrieval. Carefully analyze the input, which may contain text, images, videos, or any combination of modalities.

Identify the core semantic content, including the main topic, key entities, relevant relationships, contextual information, and any salient details that are essential for understanding its meaning.

Provide a concise, factual reasoning process that focuses on semantic understanding rather than speculation. Enclose this reasoning within `<reason>` and `</reason>` tags. Then, on a new line, produce a compact semantic summary that best represents the input for retrieval purposes. The summary should be either a single word or a short sentence, and must begin with `<sum>`.

Your response must strictly follow this format:

`<reason>`

...

`</reason>`

`<sum>`...

Any response that does not follow this format is invalid.

===== INSTRUCTION TO THINKING MODELS =====

The above input is a query/candidate for retrieval. Carefully examine and analyze the above input (which may include text, images, videos, or any combination). Identify and describe the key elements present in the input, such as the main topic, important entities, relationships, context, and any notable features or details that contribute to the overall meaning. Finally, synthesize your analysis and reflection into a single word or a concise sentence that best captures the essence of the input for retrieval purposes. If the input is a phrase or word, the summary is that word itself.

Table 4: Instructions to different workers MLLM for prior distribution simulation.

ular, the improvements over the closest 4B rival, Embed-RL-4B, are observed on 69 out of 78 tasks, with a mean improvement of 3.08 points and a 95% bootstrap confidence interval of [2.13, 4.02]. Both the Wilcoxon signed-rank test and the paired t -test yield highly significant p -values for all comparisons, while all corresponding confidence intervals remain strictly above zero. These results demonstrate that the improvements of MMEmb-R1 are statistically significant and robust across tasks.

A.6 Deep Analysis of Workers and Evaluator

We employ a worker-evaluator framework to approximate the latent reasoning space. To assess the robustness of different worker and evaluator configurations, as well as to gain deeper insight into the proposed pair-aware reasoning selection mechanism, we conduct the following experiments. We consider different types of workers, as listed in Tab. 6, and different evaluator configurations, as shown in Tab. 7.

We observe the following findings: **(1)** Basically, increasing the diversity and capability of

worker models consistently improves downstream performance, supporting our hypothesis that heterogeneous reasoning candidates better approximate the latent reasoning space. However, continuously adding more heterogeneous workers may not always be beneficial. When we further add three additional workers (Kimi/Claude/GPT series), performance decreases, likely because excessive diversity introduces additional challenges for CoT-based training and makes the optimization process more difficult. **(2)** We acknowledge that introducing multiple workers and a large evaluator increases engineering overhead. To reduce this cost, we replace the original 32B evaluator with a compact 2B reward model trained to mimic the original scoring process, denoted as *Evaluator-Flash*. Despite its significantly smaller size, this variant achieves a score of 63.2, outperforming UME-R1 (60.1) while providing a more compact and scalable pipeline. **(3)** The evaluation stage remains stable across different evaluator models and prompting strategies, provided that the evaluator is sufficiently capable. In contrast, worker quality is substantially more

Comparison (Ours-4B vs.)	Win/Loss/Tie	Mean Δ	95% CI of Δ	Wilcoxon p	Paired- t p
Embed-RL-4B (closest 4B rival)	69 / 9 / 0	+3.08	[+2.13, +4.02]	6.3×10^{-10}	1.5×10^{-8}
UME-R1-7B	63 / 13 / 2	+6.66	[+4.36, +9.09]	4.4×10^{-8}	4.3×10^{-7}
Embed-RL-2B	69 / 9 / 0	+4.39	[+3.27, +5.52]	3.4×10^{-11}	5.8×10^{-11}

Table 5: Statistical significance analysis of MMEB-R1 (Qwen3-VL-4B) against strong baselines over 78 tasks. Δ denotes the per-task performance difference between our model and the corresponding baseline.

Worker Configuration	MMEB-V2
InternVL3-2B (single)	54.8
GLM-4.1V-Thinking (single)	61.2
Qwen3VL-32B (single)	61.9
GLM-4.1V-Thinking + InternVL3-14B	63.0
MMEB-R1 (Full)	65.0
MMEB-R1 + Another 3 workers	62.1

Table 6: MMEB-V2 performance under different worker configurations.

Model/Evaluator	MMEB-V2
UME-R1	60.1
Evaluator-Flash	63.2
InternVL3-38B	63.7
Qwen3-VL-32B (Alt. Prompt)	64.7
Qwen3-VL-32B (Original)	65.0

Table 7: MMEB-V2 performance under different evaluator configurations. “Alt. Prompt” denotes the use of an alternative prompt template.

critical, since workers directly generate the reasoning data used for training. This observation further justifies our sophisticated pair-aware selection mechanism, which is specifically designed to filter and weight high-quality reasoning supervision. (4) We further conduct an experiment using Qwen3-VL-32B as the worker model for generating CoT data. As shown in Tab. 7, even with a substantially stronger teacher model, the resulting performance still falls short of our full framework. This suggests that the improvements mainly stem from the effectiveness of our proposed methodology, rather than simply distilling knowledge from more powerful models.

A.7 Results Generalized to Other Domains

To further demonstrate the generalizability of our framework, we conduct preliminary experiments on two additional benchmarks. First, we evaluate on MMStar (Chen et al., 2024), a widely used multimodal benchmark, to examine whether our pair-aware reasoning mechanism can transfer to generative multimodal tasks. Second, we evaluate on ROCov2 (Rückert et al., 2024), an out-of-domain med-

ical multimodal dataset. We compare model performance with and without the proposed pair-aware selection mechanism. As shown in Tab. 8, our method consistently improves performance across both settings, suggesting that the proposed framework generalizes effectively beyond the primary benchmark.

Benchmark	w/o Selection	w/ Pair-Aware Selection
MMStar	60.2	63.2
ROCov2	43.26	45.78

Table 8: Performance with and without the pair-aware selection mechanism on a generative task (MMStar) and an out-of-domain benchmark (ROCov2). ROCov2 is evaluated using CUI@5.

A.8 Query-Dependent Embeddings

Our embedding is designed to capture not only the semantics of the visual and textual content, but also the question or instruction associated with the query. To verify this property, we conduct controlled experiments on MMEB-V2 by perturbing the question text while keeping the target items unchanged. Specifically, we consider two settings: (1) *w/o question text*, where the question text is removed from every query, and (2) *random question text*, where the question text is randomly reassigned across queries. All other components of the queries and the target items remain unchanged. The results are shown in Tab. 9. Removing the question text causes a substantial performance drop from 65.0 to 52.8, indicating that the question text provides essential information for constructing discriminative query embeddings. Moreover, randomly replacing the question text further degrades the performance to 43.1, suggesting that the embedding is sensitive to the specific query intent rather than merely exploiting generic textual information. These results shows that our model produces query-dependent embeddings that adapt to the question or instruction associated with each query.

Setting	MMEB-V2
Original	65.0
w/o question text	52.8
Random question text	43.1

Table 9: MMEB-V2 performance under different question-text perturbations. The target items remain unchanged across all settings.

B More Implementation Details

B.1 Multi-Worker Reasoning Path Generation

As discussed in § 3.2.1, we leverage heterogeneous MLLMs to approximate the distribution of the reasoning latent space. The key principle is to maximize complementarity across workers: each type contributes distinct reasoning styles and knowledge coverage, collectively simulating a richer prior than any single model. Specifically, we employ the following models with carefully designed prompts, inspired by UME-R1 (Lan et al., 2025b) and TTE (Cui et al., 2026).

Instruct-based models. We use InternVL3-14B-Instruct (Zhu et al., 2025) with the prompt shown in Tab. 4 (top). The prompt enforces a structured `<reason>...<sum>` format, encouraging concise, factual semantic analysis. Empirically, this worker produces the most consistently formatted and retrieval-oriented rationales, serving as a stable baseline in the candidate pool.

Thinking models. We leverage GLM-4.1V-Thinking (Team et al., 2026), using the prompt adopted in UME-R1 (Lan et al., 2025b) (Tab. 4, bottom). Unlike the instruct-based prompt, we do not enforce a rigid output format, allowing the model to reason in its native chain-of-thought style. This produces longer, more exploratory chains with higher variance—occasionally yielding uniquely high-gain candidates, as shown in Appendix A.3.

High-capacity proprietary models. We further leverage the API of Doubao-Seed-1.6-Vision (ByteDance, 2025), using the same prompt template as the instruct-based models. Despite sharing the prompt, the proprietary model generates qualitatively richer rationales due to its broader world knowledge, resulting in the highest median reasoning gain among all worker types.

B.2 Pair-Aware Evaluator Implementation

For the pair-aware evaluator $\mathcal{J}$, we employ Qwen3-VL-32B-Instruct (Yang et al., 2025) and use vLLM¹ for efficient inference. We leverage this strong open-source model to obtain reliable logit scores for relevance estimation. The evaluator prompts are provided in Tab. 10. For each training pair (q_i, t_i^+) and each reasoning candidate $r \in \mathcal{R}_i$ generated by the worker models, we perform two inference passes through the evaluator. In the *baseline pass*, the evaluator receives only the raw query and target (with their associated images or videos) and is prompted to judge semantic relevance with a binary YES/NO response. In the *with-rationale pass*, the same query–target pair is augmented with the candidate rationale. We extract the logit of the first generated token for both [YES] and [NO], and compute a log-probability ratio $\text{diff} = \log p([\text{YES}]) - \log p([\text{NO}])$ for each pass. The “counterfactual reasoning gain” is then: $\Delta_r = \text{diff}_{\text{with}} - \text{diff}_{\text{baseline}}$. A positive Δ_r indicates that the rationale improves the evaluator’s confidence in the query–target match beyond what the raw inputs alone provide. After computing Δ_r for all three worker sources (Instruct, Thinking, Proprietary), we apply a softmax over the scores to obtain normalized selection weights: $w_k = \text{softmax}(\Delta_{r_k})$, $k \in \{\text{instruct, thinking, proprietary}\}$. These weights are stored alongside the rationales and used during training for weighted sampling.

Moreover, we notice that the prompts used in the baseline and rationale conditions are slightly different, raising a concern that the observed differences may be partially attributed to prompt-induced bias. To address this concern, we redesigned the evaluator prompts such that both conditions ask exactly the same relevance question (“Given the query and target, are they relevant for retrieval? YES/NO”), with the only difference being whether the rationale is included in the evaluator context. Using this matched-prompt setting, we recomputed all per-worker selection weights and compared them with the original ones. The resulting rankings show high agreement, with Kendall’s $\tau = 0.82$, indicating that the original weights are not primarily driven by prompt asymmetry. Moreover, the updated 2B model trained with the revised weights achieves 64.8 on MMEB-V2, comparable to the original setting. Therefore, we retain the original

¹<https://vllm.ai/>

===== BASELINE PROMPT (WITHOUT RATIONALE) ===== Above are two items for a retrieval task: Query Item: <image><video> Text: {query_text} Target Item: <image><video> Text: {target_text} Based on the query item and target item, are these two items semantically relevant for retrieval? Consider whether they match in terms of content, topic, and intent. Answer with YES or NO only.
===== WITH-RATIONALE PROMPT (FOR COUNTERFACTUAL EVALUATION) ===== Above are two items for a retrieval task: Query Item: <image><video> Text: {query_text} Rationale: {query_rationale} Target Item: <image><video> Text: {target_text} Rationale: {target_rationale} The rationales are reasoning paths generated to help understand the semantic content of each item. Based on the query item, target item, and their rationales, does adding the rationale improve the retrieval effectiveness? Consider whether the rationales capture essential semantic information that aids in matching these two items. Answer with YES or NO only.

Table 10: Prompts used for the pair-aware evaluator $\mathcal{J}$. The *baseline* prompt (top) evaluates query–target relevance without reasoning. The *with-rationale* prompt (bottom) includes generated rationales, enabling the counterfactual comparison $\Delta_r = c_r - c_0$.

prompt for simplicity.

B.3 Details of Joint Reasoning and Embedding Training

We adopt the instruction template shown in Tab. 11. Two special tokens are introduced: <d_emb>, appended at the beginning of the instruction to mark the direct embedding extraction point, and <r_emb>, generated after optional reasoning tokens to mark the reasoning-enhanced embedding extraction point. Although the adaptive reasoning policy is formally learned in the RL stage (§ 3.4), we find it beneficial to expose the model to a small fraction of direct-embedding samples during joint training,

preventing the model from becoming overly reliant on reasoning generation and easing the subsequent policy learning. Specifically, we select samples that are unlikely to benefit from reasoning based on two criteria: (1) samples with very low pair-aware selection weight w_r , indicating that no generated rationale meaningfully improves query–target alignment, and (2) samples with very short input text (fewer than 5 words), where reasoning would constitute overthinking. For these samples, we replace the rationale with an <empty> token with probability 0.1, training the model to directly produce embeddings without intermediate reasoning.

During training, the vision encoder is kept

<input_image><input_video><input_text><d_emb> Represent the above input text, images, videos, or any combination of the three as embeddings. You may output the thinking process in <reason> </reason> tags and then summarize the entire input in a word or sentence. Finally, use the <r_emb> tag to represent the entire input. If explicit reasoning is not necessary (e.g., the task is simple or the input is concise), you may directly produce the embeddings without generating intermediate thinking.

Table 11: Instruction template used during joint reasoning and embedding training (§ 3.3).

frozen, while both the multimodal projector and the LLM backbone are updated. We train for 3 epochs with a per-device batch size of 4 and gradient accumulation steps of 8, yielding an effective batch size of 256 across 8 GPUs. We use AdamW with a learning rate of 5×10^{-5} , cosine scheduling, and a warmup ratio of 0.03. The loss weights λ_{CoT} and λ_{direct} are both set to 1. The maximum sequence length is 12288 tokens, with image pixels clipped to [768, 2359296]. Training is conducted in bfloat16 precision with DeepSpeed ZeRO-3 and gradient checkpointing.

B.4 Details of Adaptive Reasoning Control

For adaptive reasoning control, we adopt the codebase of VLM-R1² (Shen et al., 2025). We sample 8 completions for each query with a maximum generation length of 1024 and temperature 1.0. The GRPO clipping coefficient is set to the range [0.8, 1.28], and the KL-divergence coefficient is set to 0.04. Training is performed with a batch size of 8 per device and 2 gradient accumulation steps. The learning rate is 1×10^{-6} , and the model is trained for 2 epochs. Additional details regarding GRPO are provided in Section 3.4.2.

For the adaptive reward design, we set $\alpha = 0.2$ and encourage the DIRECT action during the first 500 training steps. The reasoning cost coefficient is set to $c = 1 \times 10^{-3}$. For the format reward, any chain-of-thought (CoT) that deviates from our predefined format receives a reward of 0, while valid outputs receive a reward of 1. For the embedding reward, we adopt the design proposed in UME-R1 (Lan et al., 2025b), which measures how well the generated representations distinguish positive targets from negative ones. Specifically, the reward considers two criteria: (i) the ranking of positive targets among negative targets, and (ii) the similarity gap between positives and negatives. For

each query q with a positive target t^+ and a negative target t^- , we sample a group of responses $\{o_j^+\}_{j=1}^G$ associated with the positive target and another group $\{o_j^-\}_{j=1}^G$ associated with the negative target. Given the i -th sampled response o_i and embedding model $\mathcal{E}_\theta$, we compute its similarity scores with the positive targets as $S^+ = \{\mathcal{E}_\theta([q, o_i]) \cdot \mathcal{E}_\theta([t^+, o_j^+])\}_{j=1}^G$ and with the negative targets as $S^- = \{\pi_\theta(q, o_i) \cdot \pi_\theta(t^-, o_j^-)\}_{j=1}^G$. The embedding reward is defined as:

$$R_{\text{emb}}(o_i) = \frac{|S^+ \cap \text{top}_G(S^+ \cup S^-)|}{G} \times (\text{avg}(S^+) - \text{avg}(S^-)),$$

where $\text{top}_G(\cdot)$ selects the G largest elements from the input set. The first term measures whether positive similarities rank higher than negative ones, while the second term captures the magnitude of the similarity gap. Maximizing this reward encourages the model to produce reasoning trajectories that lead to more discriminative and informative embeddings. Moreover, we treat the query and the target symmetrically and compute the final reward as the mean of the rewards obtained from both directions.

B.5 Discussion on the Implementation of Other Baselines

We provide additional implementation details of existing reasoning-driven embedding methods to facilitate a clearer comparison with our framework. TTE employs Qwen2.5-VL-72B as the teacher model and designs dataset-specific prompts to generate CoT annotations. UME-R1 uses GLM-4.1V-Thinking for rationale generation followed by heuristic filtering, while EMB-RL relies on teacher-generated CoT data together with model-based filtering strategies. In terms of training pipelines, both UME-R1 and EMB-RL adopt a two-stage supervised fine-tuning (SFT) plus reinforcement

²<https://github.com/om-ai-lab/VLM-R1>

learning (RL) framework, whereas TTE separately trains the embedder and the reasoner. Therefore, the overall complexity of our framework is comparable to existing baselines rather than substantially higher. Moreover, as the data scale increases, our method may even become more efficient by avoiding dataset-specific manual prompt engineering and heuristic filtering procedures.

C Background and Preliminaries

C.1 Causal Inference and Causal Learning

Causal inference (Pearl, 2009) aims to identify cause-and-effect relationships beyond associational patterns. The structural causal model (SCM) framework represents data-generating processes as directed acyclic graphs, where Pearl’s do-operator $\text{do}(X = x)$ formalizes interventions by fixing a variable while severing its incoming causal edges. This distinguishes the interventional distribution $P(Y \mid \text{do}(X = x))$ from the observational conditional $P(Y \mid X = x)$, enabling isolation of the true causal effect from confounders. A key quantity is the average treatment effect: $\text{ATE} = \mathbb{E}[Y \mid \text{do}(X = 1)] - \mathbb{E}[Y \mid \text{do}(X = 0)]$. At a higher level, counterfactual reasoning (Pearl and Mackenzie, 2018) addresses “what if” questions—computing the outcome under an alternative intervention for the same instance. Causal perspectives have been increasingly adopted in the deep learning community (Yang et al., 2023, 2024). These works share a common principle: explicitly modeling causal pathways isolates target effects from confounders, yielding more robust systems. In our work, obtaining an ideal reasoning trace is challenging. From a Bayesian perspective, the diverse rationales generated by multiple MLLMs can be viewed as a noisy prior candidate set, and the evaluator-based reweighting process provides a more informative posterior distribution. During evaluation, the evaluator estimates whether introducing a reasoning path improves retrieval relevance and quantifies its contribution. The comparison between predictions with and without rationale is conceptually related to the intuition of comparing factual and counterfactual outcomes in causal analysis.

C.2 Group Relative Policy Optimization

Standard RLHF alignment via PPO (Schulman et al., 2017) requires a separate critic network to estimate per-token advantages, introducing substantial memory overhead when the policy is a

large language model. An alternative line of work replaces reinforcement learning with preference-based optimization. Direct Preference Optimization (DPO) (Rafailov et al., 2023) learns from paired preference data (x, y^+, y^-) by directly optimizing the likelihood difference between preferred and rejected responses, eliminating the need for an explicit reward model or policy gradient updates. However, DPO relies on curated preference pairs (Xu et al., 2024) and does not naturally accommodate reward signals derived from multiple sampled outputs. GRPO (Shao et al., 2024b) addresses these limitations by computing advantages at the group level. For each input x , G candidate outputs $\{o_1, \dots, o_G\}$ are sampled from π_θ and scored by a reward function $R(\cdot)$, with advantages normalized within the group: $\hat{A}_i = (R(o_i) - \text{mean}(\{R(o_j)\}))/\text{std}(\{R(o_j)\})$. The policy is updated via a clipped surrogate objective with KL regularization against a reference policy π_{ref} :

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E} \left[\min(r_i \hat{A}_i, \text{clip}(r_i, 1 - \epsilon, 1 + \epsilon) \hat{A}_i) \right] - \beta D_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}})$$

where $r_i = \pi_\theta(o_i \mid x)/\pi_{\text{old}}(o_i \mid x)$. GRPO has been widely adopted in reasoning model training, notably by DeepSeek-R1 (Guo et al., 2025). Its key advantages are: no critic network (lower memory), stable gradients via group normalization, and straightforward implementation atop standard LM training infrastructure (Liu et al., 2025a).

Table 12: Detailed results of baselines and MMEmb-R1 on full MMEB v2 benchmark. Rows are colored by modality: Image , Video , and VisDoc . Best results per row are in **bold**.

	ColPali v1.3	GME-7B	VLM2Vec-7B	VLM2Vec-V2-2B	CAFe-7B	UME-R1-2B	UME-R1-7B	Embed-RL-2B	Embed-RL-4B	Ours (Qwen2-VL-2B)	Ours (Qwen3-VL-4B)
ImageNet-1K	42.4	64.6	80.1	80.8	77.3	75.3	80.4	78.0	79.5	76.2	81.8
N24News	25.5	50.5	79.7	72.9	83.2	81.1	82.3	44.9	48.3	56.5	62.2
HatefulMemes	50.6	53.6	69.7	56.3	78.7	75.2	79.0	65.0	66.2	71.8	70.5
VOC2007	69.8	80.3	80.7	85.0	89.8	80.0	90.8	78.7	79.5	79.5	82.1
SUN397	56.1	69.5	77.4	71.0	79.9	79.4	80.3	75.4	79.2	76.2	79.7
Place365	27.5	39.1	37.4	35.9	45.0	42.6	46.8	43.9	43.1	48.5	46.8
ImageNet-A	14.9	41.2	58.1	47.4	55.2	50.4	53.9	59.2	58.1	53.8	62.3
ImageNet-R	64.6	83.9	73.9	89.3	88.0	88.7	90.1	88.5	88.2	89.2	91.5
ObjectNet	45.6	69.0	40.1	65.2	22.5	52.0	42.3	74.8	75.4	62.5	75.9
Country211	6.0	24.8	29.8	30.2	16.7	23.4	25.0	20.0	19.4	25.5	23.8
OK-VQA	9.4	33.2	56.8	51.5	67.3	62.4	71.7	61.4	67.3	66.8	68.9
A-OKVQA	6.6	21.0	47.3	43.6	63.8	51.1	58.7	54.7	59.3	57.2	64.1
DocVQA	11.3	41.4	89.7	90.1	79.2	92.2	93.8	92.4	94.3	92.8	95.2
InfographicsVQA	5.0	20.3	60.0	58.8	53.3	67.7	79.2	76.7	77.5	78.5	82.1
ChartQA	5.7	17.8	56.9	47.4	48.8	64.9	75.1	80.7	80.9	75.2	81.5
Visual7W	6.1	22.2	52.7	52.9	52.5	54.1	55.2	52.7	55.3	58.0	65.2
ScienceQA	16.3	28.0	38.5	38.2	65.4	42.7	53.7	57.3	61.6	49.8	64.8
VizWiz	27.6	39.0	39.9	43.3	43.8	46.8	51.6	54.5	56.2	54.5	64.5
GQA	8.3	76.9	55.1	64.9	65.7	67.3	69.3	64.9	68.5	64.2	69.8
TextVQA	18.8	46.8	71.6	72.2	76.8	78.6	83.5	83.8	84.3	83.5	85.9
VisDial	41.2	60.8	81.9	82.7	82.7	76.6	80.7	81.5	84.9	83.2	83.5
CIRR	8.2	54.9	51.1	57.5	60.4	53.7	55.3	47.6	61.2	52.8	69.7
VisualNews_i2i	50.1	79.7	80.5	74.5	69.5	71.7	76.8	71.9	73.7	63.5	75.1
VisualNews_i2t	47.6	83.6	81.2	78.2	79.4	74.2	82.0	73.6	73.9	74.8	77.3
MSCOCO_i2i	59.2	71.2	77.2	75.3	75.4	75.1	78.3	79.4	78.9	81.2	84.5
MSCOCO_i2t	49.9	57.7	73.9	71.4	73.1	68.9	71.4	75.3	76.3	73.5	76.2
NIGHTS	65.5	67.6	67.6	68.6	66.7	67.2	68.1	66.3	66.4	73.2	70.9
WebQA	53.8	91.4	88.3	90.6	89.3	90.0	90.9	89.3	90.5	87.8	91.2
FashionIQ	5.9	37.8	17.1	19.5	39.0	17.1	23.4	24.0	31.9	26.5	37.8
Wiki-SS-NQ	80.5	78.2	62.3	66.9	61.2	62.0	72.5	68.9	69.6	70.2	74.1
OVEN	50.0	75.1	66.5	64.3	60.8	66.9	71.4	61.4	60.7	66.8	61.5
EDIS	64.7	96.0	85.7	84.1	71.3	88.0	92.0	84.5	87.4	86.5	91.8
MSCOCO	36.7	31.4	75.7	67.1	84.7	69.5	72.7	92.9	93.6	87.2	94.2
RefCOCO	64.5	60.9	87.6	87.1	89.4	83.3	91.4	94.9	95.9	97.5	99.1
RefCOCO-Matching	3.9	78.4	84.6	85.8	83.0	84.4	91.1	85.8	88.0	83.8	92.5
Visual7W-Pointing	56.1	66.5	81.0	69.2	93.2	71.5	84.2	88.0	87.9	87.2	93.8
K700	23.4	39.7	35.5	38.0	40.1	35.8	42.8	55.8	56.8	57.2	57.3
SmithSmithV2	25.1	30.6	32.1	42.8	35.8	44.1	50.4	56.7	59.5	55.8	64.8
HMDB51	24.8	47.9	42.2	40.9	46.9	54.4	58.3	56.7	60.1	61.2	62.5
UCF101	49.4	54.7	61.8	60.0	39.6	67.2	70.0	79.3	78.5	74.5	81.2
Breakfast	10.9	14.3	23.8	14.8	16.6	20.1	21.5	36.7	33.0	32.8	37.5
MVBench	33.7	46.6	28.5	33.7	48.9	49.9	58.2	50.8	55.9	55.5	58.2
Video-MME	30.6	39.2	27.8	30.7	46.0	41.7	47.3	47.1	50.5	49.8	52.1
NExTQA	35.2	53.6	20.3	20.9	62.4	59.9	69.6	53.9	58.2	53.2	60.8
EgoSchema	38.4	46.8	21.8	34.0	60.0	45.4	52.4	53.0	52.8	37.5	57.2
ActivityNetQA	51.3	65.6	51.4	52.3	76.0	57.8	76.0	74.8	74.4	67.8	77.1
DiDeMo	22.8	26.4	29.3	30.4	37.8	32.4	40.0	45.3	46.8	48.5	55.3
MSR-VTT	17.6	31.8	34.5	28.3	36.5	34.3	38.9	45.7	46.2	47.2	51.8
MSVD	45.4	49.7	46.7	48.1	56.4	55.4	60.8	67.2	65.8	62.8	68.2
VATEX	16.7	24.9	25.5	26.5	32.0	29.9	32.6	43.6	43.4	44.2	52.9
YouCook2	5.3	9.1	9.0	10.6	9.5	12.7	18.5	23.5	23.3	35.5	29.7
QVHighlight	19.9	59.5	57.7	49.4	58.4	57.5	54.9	70.7	73.6	48.2	68.3
Charades-STA	29.0	14.0	19.8	20.2	18.7	20.4	21.9	26.4	25.0	34.8	31.5
MomentSeeker	27.6	37.4	39.3	40.8	41.4	41.2	41.1	50.9	49.9	44.5	52.1
ViDoRe_arxivqa	81.7	86.9	60.2	80.6	73.3	73.9	73.6	86.1	88.7	68.2	89.3
ViDoRe_docvqa	56.6	57.5	34.7	44.9	38.3	37.9	41.1	45.7	47.5	46.8	53.8
ViDoRe_infvqa	84.9	91.6	70.4	83.7	80.6	76.2	80.8	86.8	86.9	82.5	87.2
ViDoRe_tabfqquad	86.9	94.6	78.2	89.2	80.7	86.1	90.2	94.5	94.7	91.2	94.1
ViDoRe_tatdqa	70.9	74.1	27.6	43.8	37.8	40.6	46.7	54.6	54.8	43.5	72.3
ViDoRe_shiftproject	75.1	96.8	38.6	60.8	52.0	66.8	65.0	70.7	69.0	67.8	69.5
ViDoRe_artificial_intelligence	95.7	99.6	67.7	88.5	86.0	85.9	89.5	94.0	91.6	89.2	92.1
ViDoRe_energy	94.7	95.3	60.4	86.5	84.8	83.3	85.7	86.7	88.1	81.5	88.6
ViDoRe_government_reports	93.6	98.8	61.8	85.0	85.0	82.6	89.8	89.0	90.7	84.8	91.2
ViDoRe_healthcare_industry	95.9	99.3	69.9	92.2	88.4	90.8	94.3	91.1	90.4	85.8	91.8
ViDoRe_esg_reports_human_labeled_v2	51.3	63.4	6.8	45.6	50.7	50.2	50.4	56.9	59.8	56.2	67.5
ViDoRe_biomedical_lectures_v2_multilingual	54.7	49.5	5.1	44.3	50.9	46.2	50.7	51.0	50.1	47.5	55.7
ViDoRe_economics_reports_v2_multilingual	49.0	54.2	13.9	43.0	54.3	45.7	57.8	53.0	53.9	59.2	64.3
ViDoRe_esg_reports_v2_multilingual	52.9	55.4	11.9	46.6	42.3	42.7	43.2	46.9	49.7	61.2	54.9
VisRAG_ArxivQA	80.9	87.4	52.6	76.9	74.0	74.3	80.5	84.9	86.9	84.8	87.2
VisRAG_ChartQA	72.3	86.1	57.7	83.7	82.7	86.0	85.0	88.3	88.5	74.2	88.1
VisRAG_MP-DocVQA	82.0	89.7	60.6	88.1	75.1	75.6	83.4	79.1	79.3	71.5	84.8
VisRAG_SlideVQA	85.1	92.6	54.7	84.1	87.6	87.1	91.5	92.3	92.6	82.8	92.1
VisRAG_InfoVQA	83.5	88.6	66.0	82.3	87.9	84.4	89.2	90.0	89.6	91.8	90.8
VisRAG_PlotQA	79.3	76.5	62.7	75.9	69.4	68.0	72.7	73.0	72.4	68.5	58.9
ViDoSeek-page	38.1	32.6	16.3	29.1	22.5	21.2	21.3	82.0	84.4	32.2	74.8
ViDoSeek-doc	87.5	90.3	69.4	79.0	73.8	75.9	75.3	82.6	82.4	78.5	88.9
MMLongBench-page	27.1	36.9	0.4	15.8	13.3	11.9	12.3	47.7	51.0	32.4	47.5
MMLongBench-doc	80.4	85.2	28.8	63.0	42.6	39.7	41.3	50.3	50.7	49.9	55.2

Table 13: Results on the MMEB-V1 benchmark, which consists of 36 image embedding tasks. IND and OOD denote in-distribution and out-of-distribution datasets, respectively. Some results are adopted from Embed-RL (Jiang et al., 2026).

Model	Per Meta-Task Score				Average Score		
	Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# of Datasets	10	10	12	4	20	16	36
<i>Baseline Models</i>							
CLIP (Radford et al., 2021)	42.8	9.1	53.0	51.8	37.1	38.7	37.8
BLIP-2 (Li et al., 2023)	27.0	4.2	33.9	47.0	25.3	25.1	25.2
SigLIP (Zhai et al., 2023)	40.3	8.4	31.6	59.5	32.3	38.0	34.8
OpenCLIP (Cherti et al., 2023)	47.8	10.9	52.3	53.3	39.3	40.2	39.7
UniIR (BLIP _{FF}) (Wei et al., 2024)	42.1	15.0	60.1	62.2	44.7	40.4	42.8
UniIR (CLIP _{SE}) (Wei et al., 2024)	44.3	16.2	61.8	65.3	47.1	41.7	44.7
Magiclens (Zhang et al., 2024a)	38.8	8.3	35.4	26.0	31.0	23.7	27.8
<i>MLLM-based Baseline Models</i>							
VLM2Vec-2B (Jiang et al., 2025b)	59.0	49.4	65.4	73.4	66.0	52.6	60.1
VLM2Vec-7B (Jiang et al., 2025b)	62.6	57.8	69.9	81.7	72.2	57.8	65.8
VLM2Vec-V2 (Meng et al., 2025)	62.9	56.3	69.5	77.3	68.8	59.9	64.9
MMRet-7B (Zhou et al., 2025)	56.0	57.4	69.9	83.6	68.0	59.1	64.1
CAFe-V1-7B (Yu et al., 2025a)	65.2	65.6	70.0	91.2	75.8	62.4	69.8
CAFe-V2-7B (Yu et al., 2025a)	63.6	61.7	69.1	87.6	72.8	61.1	67.6
mmE5-11B (Chen et al., 2025b)	67.6	62.8	70.9	89.7	72.3	66.7	69.8
LLaVE-2B (Lan et al., 2025a)	62.1	60.2	65.2	84.9	69.4	59.8	65.2
LLaVE-7B (Lan et al., 2025a)	65.7	65.4	70.9	91.9	75.0	64.4	70.3
UniME-4B (Gu et al., 2025a)	54.8	55.9	64.5	81.8	68.2	52.7	64.2
UniME-7B (Gu et al., 2025a)	66.8	66.6	70.6	90.9	74.6	65.8	70.7
UME-R1-2B (Lan et al., 2025b)	64.8	62.8	67.6	77.2	71.5	60.4	66.6
UME-R1-7B (Lan et al., 2025b)	67.1	69.2	71.9	84.9	76.1	65.1	71.3
Embed-RL-2B (Jiang et al., 2026)	62.8	67.9	68.6	90.4	71.9	65.9	69.2
Embed-RL-4B (Jiang et al., 2026)	63.7	70.5	71.3	91.4	74.3	67.3	71.2
<i>Ours</i>							
MMEmb-R1 (Qwen2-VL-2B)	64.5	69.0	70.0	88.9	71.4	68.4	70.0
MMEmb-R1 (Qwen3-VL-4B)	67.7	74.2	74.5	94.9	75.4	74.0	74.8