MATE: MATCHING-AWARE LATENT REASONING FOR UNIVERSAL MULTIMODAL EMBEDDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Universal multimodal embedding (UME) enables scalable retrieval by encoding queries and candidates independently in a shared vector space. Complex retrieval can nevertheless benefit from integrating evidence within each input and directly comparing queries with semantically similar candidates. Explicit reasoning and multi-step latent rollout increase encoding cost, while pairwise matching adds per-candidate computation at inference. We propose *MATE*, a one-step bi-encoder that learns from both forms of computation. Latent Reasoning Distillation (LRD) uses answer-guided weighting to summarize a teacher's multi-step latent trajectory and progressively distills this target into a single latent step. Latent-Aware Matching Distillation (LAMD) trains a matcher on paired latent states and mean-pooled embeddings, then selectively transfers its listwise preferences to the bi-encoder. Matching distillation focuses on queries with a sufficient matcher margin but an insufficient bi-encoder margin. At inference, *MATE* encodes queries and candidates independently and ranks candidates by normalized vector similarity. On the 78-task MMEB-V2 benchmark, the 2B *MATE* model achieves an overall score of 65.4 and delivers 62× the throughput of the explicit-reasoning model UME-R1 under matched inference settings. Code will be publicly available upon publication.

1 INTRODUCTION

Universal multimodal embedding (UME) maps text, images, videos, and visual documents into a shared space for cross-modal retrieval. By encoding queries and candidates independently, a bi-encoder allows corpus representations to be precomputed and indexed for efficient vector search. This design supports applications such as product search, document retrieval, and video question answering (Radford et al., 2021; Jiang et al., 2024b). Recent embedders built on multimodal large language models (MLLMs), including VLM2Vec, GME, and MoCa, have further improved retrieval quality (Zhang et al., 2025; Chen et al., 2026). Independent encoding, however, places substantial demands on each representation: it must integrate evidence within its own input while retaining the distinctions needed to separate relevant candidates from semantically close alternatives.

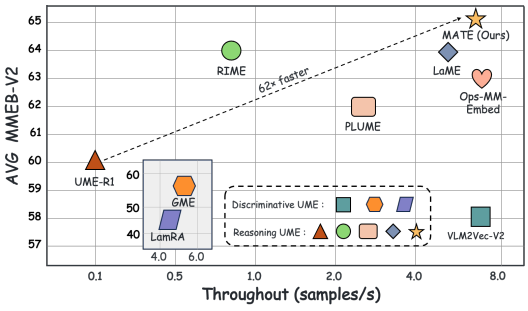

Figure 1: Retrieval quality and inference throughput on MMEB-V2. Throughput is measured on a single GPU, and quality is averaged over 78 tasks.

Meeting these demands can require two complementary forms of computation. Within an input, evidence may need to be integrated over multiple reasoning steps, for example to resolve compositional spatial relations or connect text with document layout. Across inputs, a positive candidate and a hard negative may share much of their content, yet differ in the evidence that determines relevance to a particular query. Direct query–candidate comparison can expose such differences. Standard contrastive learning supervises similarities between independently encoded representations, without explicitly supervising intermediate reasoning or pair-conditioned representations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Existing approaches obtain these capabilities at different costs. Explicit-reasoning embedders generate chain-of-thought (CoT) before forming representations (Hao et al., 2026), making large-scale encoding expensive. Latent reasoning avoids textual decoding, but methods that perform multi-step rollout at inference still incur additional encoding cost (He et al., 2026). Cross-encoders compare queries and candidates directly, although per-pair scoring typically confines them to post-retrieval reranking. Other work transfers matching information into a dual encoder: CPRD distills selected hard-negative ordering constraints from a cross-encoder (Chen et al., 2024), while ReMatch incorporates matching supervision into representation learning (Liu et al., 2026). These developments motivate using reasoning and matching together as training signals while retaining independent encoding at retrieval time.

We introduce *MATE*, which learns from a multi-step latent teacher and a training-only matcher, then deploys a one-step bi-encoder. Latent Reasoning Distillation (LRD) uses answer-guided weighting to extract a target from the teacher's latent trajectory and progressively transfers it to a single latent step. Latent-Aware Matching Distillation (LAMD) uses a matcher that jointly reads the latent states and mean-pooled embeddings of query–candidate pairs. When the matcher establishes a sufficient positive–negative margin but the bi-encoder's margin remains small, *MATE* distills the matcher's listwise preferences into the embedding space. Neither the multi-step teacher nor the matcher is used at inference: retrieval requires independent encoding and normalized vector similarity. Figure 1 illustrates the resulting quality–throughput trade-off.

Our main contributions are as follows.

- We develop MATE, a one-step bi-encoder trained with complementary signals from within-input latent reasoning and query–candidate matching.
- We introduce answer-guided, progressive latent distillation and selective listwise matching distillation to transfer these signals into independently computed embeddings.
- On the 78-task MMEB-V2 benchmark, the 2B scale, *MATE* achieves an overall score of 65.4 and delivers 62× the throughput of the explicit-reasoning model UME-R1 under matched inference settings.

2 Related Work

2.1 Universal Multimodal Embedding

Universal multimodal embedding (UME) uses a single model to map heterogeneous inputs into a shared retrieval space. Early dual encoders such as CLIP, ALIGN, and SigLIP primarily addressed image–text alignment (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023). Unified retrievers extended this setting to broader tasks and open-ended instructions (Wei et al., 2024; Zhang et al., 2024; Xu et al., 2025b). E5-V and MM-Embed adapted MLLMs as universal encoders (Jiang et al., 2024a; Lin et al., 2025), while the VLM2Vec family broadened modality coverage to include videos and visual documents (Jiang et al., 2024b; Meng et al., 2025). CAFe jointly optimizes representation and generation, while Ops-MM-Embed provides another MLLM embedding model (Yu et al., 2025; OpenSearch-AI, 2025). Other work has expanded training data with synthetic multimodal pairs or examined the factors that shape MLLM embedding quality (Lan et al., 2025; Chen et al., 2025; Zhou et al., 2025; Li et al., 2026b). Multi-token and late-interaction methods preserve finer-grained information for scoring (Xiao et al., 2026; Faysse et al., 2025). These studies advance modality coverage, training data, and representation granularity. *MATE* focuses on the reasoning and matching supervision used to train a single-vector encoder.

2.2 Reasoning-Enhanced Multimodal Embedding

Chain-of-thought (CoT) supports complex language and vision-language reasoning (Wei et al., 2022; Guo et al., 2025; Xu et al., 2025a). In retrieval, UME-R1 learns reasoning-driven embeddings, whereas Think-Then-Embed uses a separate reasoner (Lan et al., 2026; Cui et al., 2026). RIME uses rewriting to connect reasoning and multimodal embedding, RGE converts MLLM reasoning into retrieval representations, and Think-When-Needed adaptively routes reasoning for multimodal embeddings (Wu et al., 2026a; Liu et al., 2025a; Zhang et al., 2026). Embed-RL encourages reasoning grounded in visual evidence, and TRACE adapts reasoning to retrieval tasks (Jiang et al., 2026; Hao

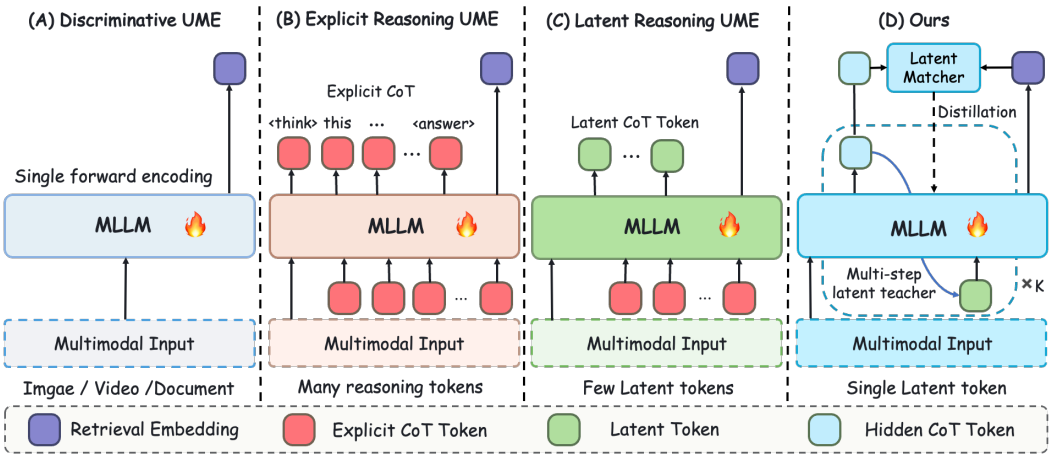

Figure 2: Approaches to universal multimodal embedding. (A) A discriminative encoder forms embeddings without a separate reasoning stage. (B) Explicit reasoning generates a textual rationale before forming an embedding. (C) Multi-step latent reasoning iterates over hidden states without decoding text. (D) *MATE* learns from a multi-step latent teacher and a training-only matcher, then retrieves with a one-step bi-encoder.

et al., 2026). Generating rationales for queries and corpus items, however, adds to encoding cost. Broader work on language models explores self-generated rationales and implicit CoT (Zelikman et al., 2024; Deng et al., 2023). Hao et al., CODI, and Think Silently study continuous latent reasoning (Hao et al., 2024; Shen et al., 2025; Tan et al., 2025), while pause tokens allocate additional computation and TokenSkip compresses textual reasoning sequences (Goyal et al., 2024; Xia et al., 2025). For dense text retrieval, LaSER internalizes explicit reasoning into latent space (Jin et al., 2026). PLUME performs autoregressive latent rollout at inference (He et al., 2026), whereas LaME uses a fixed budget of latent tokens in a single forward pass (Wu et al., 2026b). *MATE* distills an answer-guided summary of a multi-step latent teacher into a one-step encoder and complements this supervision with latent-aware matching.

2.3 Matching-Guided Retrieval Embeddings

Matching models evaluate queries and candidates jointly, allowing them to capture interactions that pooled vector similarity may miss. ALBEF and BLIP combine contrastive learning with image–text matching heads (Li et al., 2021; 2022). Retrieval systems also use cross-encoders to rerank candidates selected by a faster dual encoder (Miech et al., 2021; Geigle et al., 2022). CPRD transfers relative ordering among selected hard negatives from a cross-encoder to a dual encoder (Chen et al., 2024). ReMatch trains the embedding MLLM with a generative matching objective over raw inputs and their embeddings (Liu et al., 2026). Other MLLM-based systems use matching for reranking, external judging, or joint embedding and reranking (Liu et al., 2025b; Gu et al., 2026; Li et al., 2026a; Xu et al., 2025c). *MATE* trains a matcher on paired latent states and mean-pooled embeddings, then selectively distills its listwise preferences into the bi-encoder. At inference, candidates are scored by vector similarity.

3 Methodology

3.1 Overview

MATE trains a one-step encoder for universal multimodal retrieval with complementary supervision from latent reasoning and query–candidate matching. Latent Reasoning Distillation (LRD) forms an answer-guided summary of a multi-step teacher's latent trajectory and progressively distills it into a single latent step. Latent-Aware Matching Distillation (LAMD) uses a training-only matcher to jointly examine the latent states and mean-pooled embeddings of query–candidate pairs, then transfers selected listwise preferences into the bi-encoder's similarity space. At inference, only the

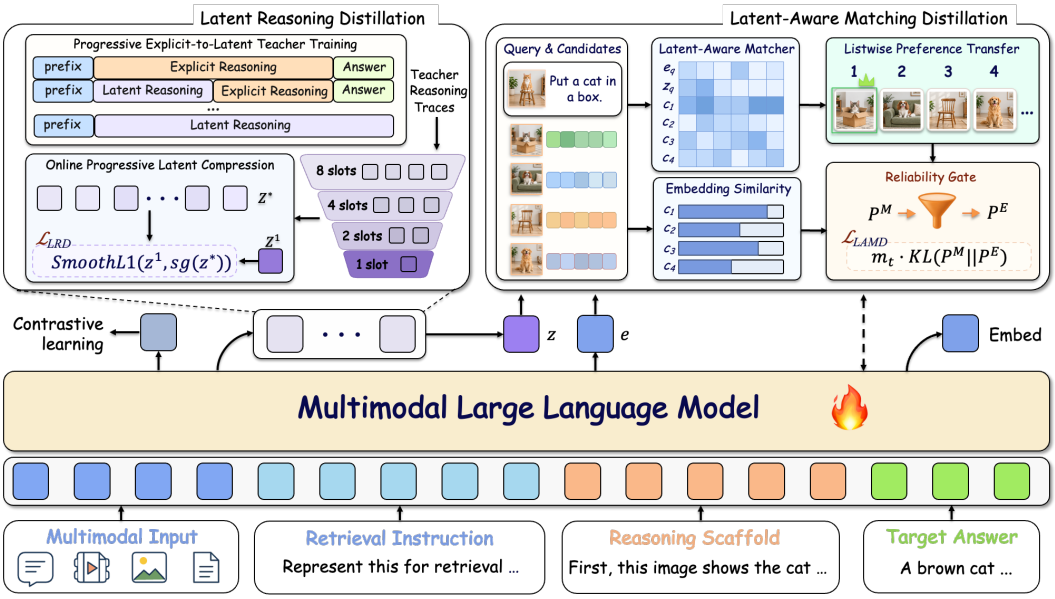

Figure 3: Overview of *MATE*. The bottom panel shows the shared MLLM backbone and the structured input used during training. Latent Reasoning Distillation (top left) trains a multi-step latent teacher with explicit rationales, then progressively distills its trajectory into a single latent state. Latent-Aware Matching Distillation (top right) uses a training-only matcher to examine the latent states and mean-pooled embeddings of queries and candidates. Selected listwise preferences from the matcher are transferred to the bi-encoder.

one-step encoder is retained: queries and corpus items are encoded independently, and candidates are ranked by normalized vector similarity. Figure 3 summarizes the method.

3.2 Problem Formulation

We consider universal multimodal retrieval, where a shared encoder maps text, images, videos, visual documents, and their combinations into a common embedding space. Given an input x, the encoder produces M embedding states $\mathbf{E}_x \in \mathbb{R}^{M \times D}$. We mean-pool these states and normalize the result to obtain the retrieval embedding

$$\mathbf{e}_x = \mathrm{Norm}(\mathrm{MeanPool}(\mathbf{E}_x)) . \tag{1}$$

Given a mini-batch of B query-candidate pairs $\{(q_i, c_i^+)\}_{i=1}^B$, the query-to-candidate contrastive objective is

$$\mathcal{L}_{\mathrm{con}}^{q \to c} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(\mathbf{e}_{q_i}^\top \mathbf{e}_{c_i^+} / \tau\right)}{\sum_{j \in \mathcal{V}_i} \exp\left(\mathbf{e}_{q_i}^\top \mathbf{e}_{c_j^+} / \tau\right)}, \tag{2}$$

where τ is the temperature and $\mathcal{V}_i$ indexes the positive candidate and valid in-batch negatives for query q_i. We define the candidate-to-query objective analogously and use their sum, $\mathcal{L}_{\mathrm{con}} = \mathcal{L}_{\mathrm{con}}^{q \to c} + \mathcal{L}_{\mathrm{con}}^{c \to q}$.

3.3 Latent Reasoning Distillation

Learning multi-step latent reasoning. We train the teacher on multimodal examples with explicit rationales and target answers. Training begins with the full textual rationale and gradually replaces its reasoning steps with continuous latent states. Answer supervision guides this transition, while the contrastive objective continues to shape the representations for retrieval.

Let $\mathcal{F}_{\theta_T}$ denote the resulting teacher. Given an input x, it produces N latent states:

$$\widehat{\mathbf{h}}_{x,j} = \mathcal{F}_{\theta_T}\left(x; \widehat{\mathbf{h}}_{x,<j}\right), \qquad j = 1, \ldots, N. \tag{3}$$

At each step, the current hidden state is fed back as the input embedding for the next step rather than decoded into text. The trajectory $\widehat{\mathbf{L}}_x = [\widehat{\mathbf{h}}_{x,1}; \ldots; \widehat{\mathbf{h}}_{x,N}] \in \mathbb{R}^{N \times D}$ collects the intermediate representations produced over these steps.

Answer-guided latent target. The teacher trajectory contains states from different stages of reasoning, which may differ in their alignment with the supervised answer. Rather than asking a one-step student to match every intermediate state, we use the answer representation to form a single target from the trajectory. During training, the retrieval states $\mathbf{E}_x$ precede the supervised answer span. Let $\widehat{\mathbf{h}}_{x,\mathrm{ans}}$ be the mean teacher hidden state over this span. Its similarity to each latent state determines a weight:

$$\alpha_{x,j} = \mathrm{softmax}_j \left(\frac{\widehat{\mathbf{h}}_{x,\mathrm{ans}}^{\top} \widehat{\mathbf{h}}_{x,j}}{\sqrt{D}} \right). \tag{4}$$

The answer-guided target is

$$\mathbf{z}_x^* = \sum_{j=1}^{N} \alpha_{x,j} \widehat{\mathbf{h}}_{x,j}. \tag{5}$$

States more similar to the answer representation contribute more strongly to $\mathbf{z}_x^*$, giving the student a summary of the trajectory without requiring it to reproduce each reasoning step.

Progressive one-step distillation. Directly compressing N latent steps into one can be difficult because the student has less scope for sequential computation than the teacher. We reduce this gap by training students with successively fewer latent steps, following $N \to \cdots \to 1$. Let $\mathbf{z}_x^{(k)}$ be the mean of the k latent states produced by a k-step student for input x. At each compression level, the student matches the same answer-guided target:

$$\mathcal{L}_{\mathrm{LRD}}^{(k)} = \mathrm{SmoothL1}\Big(\mathbf{z}_x^{(k)}, \mathrm{sg}(\mathbf{z}_x^*) \Big), \tag{6}$$

where $\mathrm{sg}(\cdot)$ stops gradients through the teacher target. We retain the contrastive objective throughout compression to preserve retrieval supervision. The final student uses a single latent step, with $\boldsymbol{\ell}_x = \mathbf{z}_x^{(1)}$. Retrieval uses the normalized embedding $\mathbf{e}_x$ defined in equation 1. The answer representation and multi-step teacher are used only during training.

3.4 Latent-Aware Matching Distillation

Latent Reasoning Distillation strengthens the representation of each input, but queries and candidates are still encoded independently. Their similarity is efficient for retrieval, yet reduces the relation between them to a single dot product. A pairwise matcher can examine their representations jointly. We use such a matcher during training and transfer its preferences to the bi-encoder, avoiding per-pair scoring at retrieval time.

Latent-aware pair matching. For an input x, let $\boldsymbol{\ell}_x$ be its one-step latent state and $\bar{\mathbf{e}}_x = \mathrm{MeanPool}(\mathbf{E}_x)$ its mean-pooled embedding before normalization. Given a query q and candidate c, the matcher scores their concatenated representations:

$$s_{\mathcal{M}}(q, c) = \mathcal{M}_\phi([\boldsymbol{\ell}_q; \bar{\mathbf{e}}_q; \boldsymbol{\ell}_c; \bar{\mathbf{e}}_c]). \tag{7}$$

We train the matcher with a positive candidate c_i^+ and a mined hard negative c_i^- for each query q_i, using the binary objective

$$\mathcal{L}_{\mathrm{match}} = -\frac{1}{B} \sum_{i=1}^{B} \left[\log \sigma\big(s_{\mathcal{M}}(q_i, c_i^+)\big) + \log\big(1 - \sigma\big(s_{\mathcal{M}}(q_i, c_i^-)\big)\big) \right], \tag{8}$$

where σ is the sigmoid function. Once trained on individual pairs, the matcher scores candidate sets to provide ranking supervision for the bi-encoder.

Preference-guided selection. For each query q_i, we construct a candidate set $\mathcal{C}_i$ containing one positive and K hard negatives. Let c_i^{hard} be the negative in $\mathcal{C}_i$ with the highest similarity to q_i under the current bi-encoder. We compare the margin between the positive and this same negative under the matcher and the bi-encoder:

$$\Delta_i^{\mathcal{M}} = s_{\mathcal{M}}(q_i, c_i^+) - s_{\mathcal{M}}(q_i, c_i^{\text{hard}}), \tag{9}$$

$$\Delta_i^{\mathcal{E}} = \mathbf{e}_{q_i}^\top \mathbf{e}_{c_i^+} - \mathbf{e}_{q_i}^\top \mathbf{e}_{c_i^{\text{hard}}}. \tag{10}$$

We apply matching supervision when the matcher margin exceeds $\delta_{\mathcal{M}}$ and the bi-encoder margin remains below $\delta_{\mathcal{E}}$:

$$m_i = \mathbb{I}\big[\Delta_i^{\mathcal{M}} > \delta_{\mathcal{M}}\big]\, \mathbb{I}\big[\Delta_i^{\mathcal{E}} < \delta_{\mathcal{E}}\big]\,. \tag{11}$$

This selects queries for which the matcher establishes a sufficient positive–negative margin that the bi-encoder has yet to achieve.

Listwise matching distillation. The matcher and bi-encoder convert their scores over $\mathcal{C}_i$ into distributions, using temperatures $\tau_{\mathcal{M}}$ and $\tau_{\mathcal{E}}$:

$$p_i^{\mathcal{M}} = \text{softmax}_{c\in\mathcal{C}_i}(s_{\mathcal{M}}(q_i, c)/\tau_{\mathcal{M}})\,, \tag{12}$$

$$p_i^{\mathcal{E}} = \text{softmax}_{c\in\mathcal{C}_i}\big(\mathbf{e}_{q_i}^\top \mathbf{e}_c/\tau_{\mathcal{E}}\big)\,. \tag{13}$$

Although the gate uses one hard negative, the distillation loss compares preferences over all candidates in the set:

$$\mathcal{L}_{\text{LAMD}} = \frac{1}{B}\sum_{i=1}^{B} m_i\, \text{KL}\big(p_i^{\mathcal{M}} \,\|\, p_i^{\mathcal{E}}\big)\,. \tag{14}$$

Only selected sets contribute to this loss, encouraging the independently encoded vectors to reflect the matcher's relative candidate scores.

3.5 Unified Learning and Retrieval

The bi-encoder retains contrastive retrieval supervision throughout both distillation stages. During Latent Reasoning Distillation, a student with k latent steps is trained with

$$\mathcal{L}^{(k)} = \mathcal{L}_{\text{con}} + \lambda_{\text{LRD}}\mathcal{L}_{\text{LRD}}^{(k)}. \tag{15}$$

The matcher is trained separately with $\mathcal{L}_{\text{match}}$. Once the student has been compressed to a single latent step, its embeddings are refined using the matcher's selected listwise preferences:

$$\mathcal{L}^{\text{final}} = \mathcal{L}_{\text{con}} + \lambda_{\text{LAMD}}\mathcal{L}_{\text{LAMD}}. \tag{16}$$

Here, λ_{LRD} and λ_{LAMD} weight the two distillation terms.

At inference, the multi-step teacher and matcher are removed. The one-step encoder independently maps queries and candidates to normalized embeddings, and candidates are ranked by $\mathbf{e}_q^\top \mathbf{e}_c$.

4 Experiments

4.1 Experimental Setup

Benchmarks and metrics. We evaluate *MATE* on MMEB-V2 and MMEB-V1. MMEB-V2 contains 78 tasks covering images, videos, and visual documents. Following the official protocol, we report Hit@1 for image and video tasks and NDCG@5 for visual-document tasks. The MMEB-V2 overall score is the unweighted average across all 78 tasks. The 36 image tasks are also reported under the official MMEB-V1 task-group convention.

Baselines. We compare *MATE* with discriminative, explicit-reasoning, and latent-reasoning UME models. These include VLM2Vec, VLM2Vec-V2, GME, LamRA, CAFe, Ops-MM-Embed, ReMatch, UME-R1, PLUME, and LaME, together with other recent reasoning-based embedders reported on MMEB-V1. For MMEB-V1, we also report published results from RIME, Embed-RL, RGE, Think-When-Needed, and Think-Then-Embed. Their reported backbone configurations include Qwen3-VL for Embed-RL and Think-When-Needed, and Qwen2.5-VL for RGE (Bai et al., 2025a;b).

Table 1: Results on the 78-task MMEB-V2 benchmark, grouped by approximate backbone size. The Image, Video, and VisDoc blocks contain 36, 18, and 24 tasks, respectively. CLS, QA, RET, GD, and MRET denote classification, question answering, retrieval, grounding, and moment retrieval. VDRv1, VDRv2, VR, and OOD refer to ViDoRe v1, ViDoRe v2, VisRAG, and out-of-distribution tasks. Avg. and All are unweighted averages over their constituent tasks. Bold marks the best score in each column within a size group, including ties.

Model	**Image**					**Video**					**VisDoc**					**All**
	CLS	QA	RET	GD	Avg.	CLS	QA	RET	MRET	Avg.	VDRv1	VDRv2	VR	OOD	Avg.	
# Datasets	10	10	12	4	36	5	5	5	3	18	10	4	6	4	24	78
∼2B Model Size																
VLM2Vec-V2	62.9	56.4	69.6	77.1	64.9	39.2	34.7	28.4	37.6	34.7	74.4	45.1	79.3	44.0	65.7	58.2
GME	54.6	29.8	66.8	55.6	51.9	34.8	41.8	25.4	31.8	33.6	**87.1**	53.9	82.4	43.1	**73.1**	54.2
LamRA	59.3	26.5	70.1	62.6	54.1	39.4	42.7	24.5	33.6	35.2	33.8	11.5	37.6	21.0	28.9	42.0
UME-R1	64.8	62.8	67.6	77.2	66.6	44.3	50.9	32.9	39.7	42.2	72.4	46.2	79.2	37.2	63.9	60.1
Ops-MM-Embed	**68.1**	65.1	69.2	80.9	69.0	**53.6**	**55.7**	**41.8**	33.7	**47.6**	76.4	53.2	77.6	32.6	65.5	63.0
PLUME	66.1	59.5	67.6	79.2	66.3	45.0	52.3	33.5	46.7	44.1	71.9	49.7	78.0	**57.4**	67.3	61.4
LaME	67.6	**66.2**	**70.5**	81.2	69.3	45.6	52.3	36.3	43.5	44.5	78.2	48.4	82.8	55.4	72.1	64.4
MATE (Ours)	64.9	65.8	69.7	**91.5**	**69.7**	47.8	52.7	39.2	**50.9**	47.3	81.3	**61.8**	**84.0**	44.7	72.6	**65.4**
∼7B Model Size																
GME	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	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	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	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	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
Ops-MM-Embed	69.7	69.6	**73.1**	87.2	72.7	**59.7**	62.2	**45.7**	43.2	**53.8**	80.1	59.6	79.3	67.8	74.4	68.9
LaME	**70.0**	**71.0**	**73.1**	85.8	**73.0**	51.0	**64.8**	40.3	44.6	50.8	81.5	58.2	85.0	**68.2**	75.9	68.8
MATE (Ours)	67.7	70.4	72.8	**93.1**	**73.0**	51.7	64.3	43.1	**57.3**	53.7	84.6	**67.0**	**85.8**	52.4	**76.6**	**69.7**

Implementation details. Unless otherwise stated, controlled experiments use Qwen2-VL-2B as the shared encoder (Wang et al., 2024). The training-only matcher is initialized from Qwen3-0.6B (Yang et al., 2025). We fully fine-tune the encoder with a global batch size of 1,024, an InfoNCE temperature of 0.02, and a learning rate of 5×10^{-5}. Neither the multi-step teacher nor the matcher is used at inference. Further implementation details are provided in the supplementary material.

4.2 Main Results

Results on MMEB-V2. Table 1 compares *MATE* with prior universal multimodal embedders at comparable backbone scales. *MATE* achieves overall scores of 65.4 at the 2B scale and 69.7 at the 7B scale, exceeding the strongest compared model in each size group by 1.0 and 0.8 points, respectively. Its clearest task-level advantages at both scales are in image grounding and video moment retrieval.

Results on MMEB-V1. Table 2 reports the same 36 image tasks using the MMEB-V1 meta-task and distribution-group breakdowns. *MATE*'s Overall score of 69.7 therefore matches its Image average in Table 1. It achieves the highest Overall score among the latent-reasoning models and the highest Grounding score in the table, although some explicit-reasoning models score higher overall on this image subset.

4.3 Ablation Studies

We examine the incremental contributions of LRD and LAMD. Unless otherwise stated, all variants use the same Qwen2-VL-2B backbone and evaluation protocol.

Table 3 summarizes the incremental contributions of LRD and LAMD. The multi-step teacher achieves 64.7 overall, compared with 59.5 for the one-step baseline. LRD raises the one-step student's score to 64.0, narrowing the gap to the teacher to 0.7 points. Adding LAMD provides a further gain of 1.4 points, bringing the final score to 65.4 and improving the averages across all three modality groups.

Table 3: Incremental contribution of LRD and LAMD on MMEB-V2. All is the unweighted average over 78 tasks.

Variant	Img.	Vid.	Doc.	All
One-step baseline	66.3	36.3	66.8	59.5
Multi-step teacher	69.1	45.6	72.3	64.7
LRD (one step)	68.4	44.9	71.6	64.0
LRD + LAMD	**69.7**	**47.3**	**72.6**	**65.4**

Table 2: Results on the 36 MMEB-V1 image tasks. Scores are averaged within each reported task group; IND and OOD denote in-distribution and out-of-distribution groups. Overall is the unweighted average across all 36 tasks. Bold marks the best score in each column within each method category.

Models	Backbone	Per Meta-Task Score				Average Score		
		Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# Datasets		10	10	12	4	20	16	36
Discriminative UME								
CLIP	CLIP-L/0.4B	42.8	9.1	53.0	51.8	37.1	38.7	37.8
GME	Qwen2-VL-2B	54.6	29.8	66.8	55.6	49.2	55.2	51.9
VLM2Vec	Qwen2-VL-2B	58.7	49.3	65.0	72.9	64.8	53.4	59.7
VLM2Vec-V2	Qwen2-VL-2B	62.9	56.4	69.6	77.1	68.8	60.1	64.9
Ops-MM-Embed	Qwen2-VL-2B	**68.1**	65.1	69.2	80.9	71.8	**65.6**	69.0
ReMatch	Qwen2-VL-2B	65.8	**65.9**	**70.1**	**83.3**	**72.8**	64.7	**69.2**
Explicit Reasoning UME								
UME-R1	Qwen2-VL-2B	64.8	62.8	67.6	77.2	71.5	60.4	66.6
RIME	Qwen2-VL-2B	67.9	64.4	69.8	82.1	64.5	**72.9**	69.2
Embed-RL	Qwen3-VL-2B	62.8	67.9	68.6	**90.4**	71.9	65.9	69.2
RGE	Qwen2.5-VL-3B	64.4	67.8	70.2	90.1	75.3	63.7	70.1
Think-When-Needed	Qwen3-VL-4B	68.6	71.7	68.0	87.0	75.4	66.2	71.3
Think-Then-Embed	Qwen2-VL-2B	**72.6**	**74.3**	**72.6**	85.2	**80.5**	67.0	**74.5**
Latent Reasoning UME								
PLUME	Qwen2-VL-2B	66.1	59.5	67.6	79.2	69.6	62.2	66.3
LaME	Qwen2-VL-2B	**67.6**	**66.2**	**70.5**	81.2	**73.1**	64.4	69.3
MATE (Ours)	Qwen2-VL-2B	64.9	65.8	69.7	**91.5**	71.6	**67.2**	**69.7**

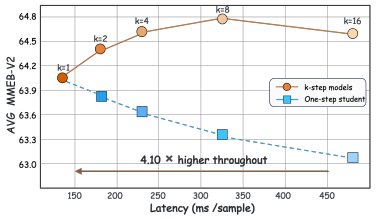

Figure 4: Retrieval quality versus encoding latency on MMEB-V2.

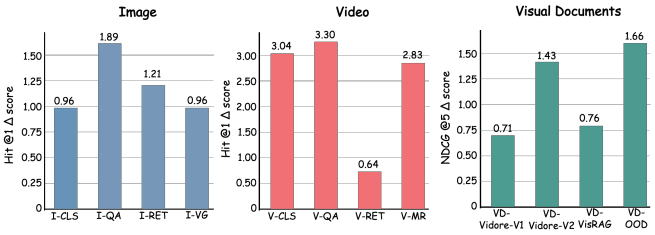

Figure 5: Absolute score gains across MMEB-V2 task groups, computed as LRD+LAMD minus LRD.

Quality and efficiency of latent reasoning. Figure 4 compares increasing sequential latent steps with increasing the inference-time latent capacity of a one-step student. For the multi-step models, retrieval performance improves up to $k = 8$, while latency rises; $k = 16$ adds cost without a further gain. Increasing the one-step student's latent capacity also raises latency without improving performance in the tested range.

Effects of latent-aware matching. Figure 5 shows improvements from LAMD across all reported task groups. Starting from the same LRD checkpoint and using the same additional training budget, LAMD achieves 65.4 overall, compared with 63.9 for continued InfoNCE training (Table 4). The input ablations further support combining latent states and pooled embeddings: neither input alone improves on the LRD baseline, whereas their combination provides effective matching supervision.

Design choices in Latent Reasoning Distillation. Table 5 compares three supervision targets for the same one-step student. The answer-guided target reaches 64.0, outperforming final-state distillation and uniform trajectory averaging by 2.6 and 1.8 points, respectively. The gain over uniform averaging highlights the value of answer-guided weighting when summarizing a multi-step trajectory for one-step distillation.

Table 4: LAMD ablations on MMEB-V2. The first block reports the answer-guided LRD checkpoint (†) and its InfoNCE continuation (‡), which uses the same training budget as LAMD. The remaining blocks examine matcher inputs and the number of hard negatives. Checkmarks indicate which representations are supplied to the matcher and whether listwise distillation is used.

Embed States	Latent States	LAMD	# Neg.	Image	Video	VisDoc	All
Without LAMD							
$\times^{\dagger}$	×	×	–	68.4	44.9	71.6	64.0
$\times^{\ddagger}$	×	×	–	68.2	45.1	71.5	63.9
Matcher inputs							
✓	×	✓	4	68.6	43.8	71.7	63.8
×	✓	✓	4	67.5	44.4	72.3	63.6
✓	✓	✓	4	**69.7**	**47.3**	**72.6**	**65.4**
Number of hard negatives							
✓	✓	✓	1	69.5	46.9	72.0	65.1
✓	✓	✓	2	69.6	47.2	72.3	65.3
✓	✓	✓	4	**69.7**	**47.3**	**72.6**	**65.4**
✓	✓	✓	8	69.2	46.8	71.9	64.9

Table 5: LRD supervision targets on MMEB-V2. The multi-step teacher is shown for reference; all distilled variants use one latent step at inference. Bold indicates the best distilled variant.

Variant	Supervision target	Image	Video	VisDoc	All
Multi-step teacher	–	69.1	45.6	72.3	64.7
Final-state distillation	Final state	67.6	41.4	67.2	61.4
Uniform trajectory distillation	Mean trajectory	68.1	42.2	68.3	62.2
Answer-guided LRD	Answer-guided target	**68.4**	**44.9**	**71.6**	**64.0**

4.4 Analysis of Distillation

Where trajectory summarization matters. The modality-level results in Table 5 reveal that the choice of distillation target has different effects across input types. Compared with final-state distillation, the answer-guided target improves the image, video, and visual-document averages by 0.8, 3.5, and 4.4 points, respectively. This pattern suggests that summarizing the full trajectory is particularly useful for recovering teacher performance on these tasks.

Where matching supervision helps. Comparing LRD with LRD+LAMD in Table 3, matching distillation improves the image, video, and visual-document averages by 1.3, 2.4, and 1.0 points, respectively. Figure 5 further shows positive gains across all 12 reported task groups, indicating that the improvement extends beyond a single modality or task group. Together with the matched-budget control in Table 4, these results show that matching supervision provides additional retrieval gains after the latent trajectory has been compressed.

5 Conclusion

We introduced *MATE*, which uses multi-step latent reasoning and query–candidate matching to train a one-step bi-encoder for universal multimodal retrieval. Latent Reasoning Distillation compresses an answer-guided summary of the teacher's trajectory, while Latent-Aware Matching Distillation transfers selected listwise preferences from a training-only matcher. At inference, *MATE* encodes queries and candidates independently and ranks candidates by normalized vector similarity, without the multi-step teacher or matcher. On MMEB-V2, *MATE* achieves the highest overall score among the compared models at both the 2B and 7B scales. These results show that reasoning and matching during training can strengthen universal embeddings while preserving scalable vector retrieval.

AI DISCLOSURE

Generative AI tools were used only to assist with language editing of the manuscript. The authors conducted the research, experiments, analysis, and interpretation of results, and take full responsibility for the final content.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

H. Chen, L. Wang, N. Yang, et al. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8254–8275, 2025.

H. Chen, H. Liu, Y. Luo, et al. Moca: Modality-aware continual pre-training makes better bidirectional multimodal embeddings. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 810–823, 2026.

Y. Chen, Z. Ma, Z. Zhang, et al. How to make cross encoder a good teacher for efficient image-text retrieval? In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26984–26993. IEEE, 2024.

X. Cui, J. Cheng, H. Chen, et al. Think then embed: Generative context improves multimodal embedding. In *International Conference on Learning Representations*, pp. 2690–2709, 2026.

Y. Deng, K. Prasad, R. Fernandez, et al. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023.

M. Faysse, H. Sibille, T. Wu, et al. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, pp. 61424–61449, 2025.

Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulić, and Iryna Gurevych. Retrieve fast, rerank smart: Cooperative and joint approaches for improved cross-modal retrieval. *Transactions of the Association for Computational Linguistics*, 10:503–521, 2022.

S. Goyal, Z. Ji, A. S. Rawat, et al. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations*, pp. 27896–27923, 2024.

T. Gu, K. Yang, K. Zhang, et al. Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21378–21386, 2026.

D. Guo, D. Yang, H. Zhang, et al. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

S. Hao, S. Sukhbaatar, D. J. Su, et al. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

X. Hao, S. Wang, T. Yang, et al. Trace: Task-adaptive reasoning and representation learning for universal multimodal retrieval. *arXiv preprint arXiv:2603.02929*, 2026.

Chenwei He, Xiangzhao Hao, Tianyu Yang, Yuxiang Ma, Yuheng Jia, Lingxiang Wu, Chaoyang Zhao, Haiyun Guo, and Jinqiao Wang. Plume: Latent reasoning based universal multimodal embedding. *arXiv preprint arXiv:2604.02073*, 2026.

C. Jia, Y. Yang, Y. Xia, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

H. Jiang, Y. Wang, Y. Zhu, et al. Embed-rl: Reinforcement learning for reasoning-driven multimodal embeddings. In *European Conference on Computer Vision*, pp. 39–56. Springer Nature Switzerland, 2026.

T. Jiang, M. Song, Z. Zhang, et al. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024a.

Z. Jiang, R. Meng, X. Yang, et al. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024b.

J. Jin, Y. Zhang, M. Li, et al. Laser: Internalizing explicit reasoning into latent space for dense retrieval. *arXiv preprint arXiv:2603.01425*, 2026.

Z. Lan, L. Niu, F. Meng, et al. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13721–13735, 2025.

Z. Lan, L. Niu, F. Meng, et al. Ume-r1: Exploring reasoning-driven generative multimodal embeddings. In *International Conference on Learning Representations*, pp. 115779–115806, 2026.

J. Li, R. Selvaraju, A. Gotmare, et al. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021.

J. Li, D. Li, C. Xiong, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026a.

X. Li, C. Li, S. Z. Chen, et al. U-marvel: Unveiling key factors for universal multimodal retrieval via embedding learning with MLLMs. In *International Conference on Learning Representations*, pp. 121298–121319, 2026b.

S. C. Lin, C. Lee, M. Shoeybi, et al. Mm-embed: Universal multimodal retrieval with multimodal llms. In *International Conference on Learning Representations*, pp. 44215–44234, 2025.

Chunxu Liu, Jiyuan Yang, Ruopeng Gao, et al. Reasoning guided embeddings: Leveraging MLLM reasoning for improved multimodal retrieval. *arXiv preprint arXiv:2511.16150*, 2025a.

Q. Liu, X. Liang, Z. Zhang, et al. Rematch: Boosting representation through matching for multimodal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16823–16833, 2026.

Y. Liu, Y. Zhang, J. Cai, et al. Lamra: Large multimodal model as your advanced retrieval assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4015–4025. IEEE, 2025b.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

A. Miech, J. B. Alayrac, I. Laptev, et al. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. *arXiv preprint arXiv:2103.16553*, 2021.

OpenSearch-AI. Ops-MM-Embedding-v1-7B. Hugging Face model card, 2025. URL `https://huggingface.co/OpenSearch-AI/Ops-MM-embedding-v1-7B`.

A. Radford, J. W. Kim, C. Hallacy, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Z. Shen, H. Yan, L. Zhang, et al. Codi: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693, 2025.

W. Tan, J. Li, J. Ju, et al. Think silently, think fast: Dynamic latent compression of LLM reasoning chains. In *Advances in Neural Information Processing Systems*, volume 38, pp. 4646–4668, 2025.

Peng Wang, Shuai Bai, Sinan Tan, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

C. Wei, Y. Chen, H. Chen, et al. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, pp. 387–404. Springer Nature Switzerland, 2024.

J. Wei, X. Wang, D. Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

Peixi Wu, Ke Mei, Feipeng Ma, et al. Beyond chain-of-thought: Rewrite as a universal interface for generative multimodal embeddings. *arXiv preprint arXiv:2604.22280*, 2026a.

Peixi Wu, Biao Yang, Feipeng Ma, Bosong Chai, Bo Lin, Wei Yuan, Fan Yang, Tingting Gao, Hebei Li, and Xiaoyan Sun. Lame: Learning to think in latent space for multimodal embedding via information bottleneck. *arXiv preprint arXiv:2606.13061*, 2026b.

H. Xia, C. T. Leong, W. Wang, et al. TokenSkip: Controllable chain-of-thought compression in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3351–3363, 2025.

Z. Xiao, Q. Ma, M. Gu, et al. Metaembed: Scaling multimodal retrieval at test-time with flexible late interaction. In *International Conference on Learning Representations*, pp. 23619–23638, 2026.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-CoT: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025a.

M. Xu, W. Zhou, Y. Babakhin, et al. Omni-embed-nemotron: A unified multimodal retrieval model for text, image, audio, and video. *arXiv preprint arXiv:2510.03458*, 2025b.

Mingjun Xu, Jinhan Dong, Jue Hou, Zehui Wang, Sihang Li, Zhifeng Gao, Renxin Zhong, and Hengxing Cai. MM-R5: Multimodal reasoning-enhanced reranker via reinforcement learning for document retrieval. *arXiv preprint arXiv:2506.12364*, 2025c.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

H. Yu, Z. Zhao, S. Yan, et al. Cafe: Unifying representation and generation with contrastive-autoregressive finetuning. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 6345–6356. IEEE, 2025.

Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah Goodman. Quiet-STaR: Language models can teach themselves to think before speaking. *arXiv preprint arXiv:2403.09629*, 2024.

X. Zhai, B. Mustafa, A. Kolesnikov, et al. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. MagicLens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*, 2024.

Longxiang Zhang, Weilong Dai, Guanghao Zhang, Hao Jiang, and Pipei Huang. Think when needed: Adaptive reasoning-driven multimodal embeddings with a dual-lora architecture. *arXiv preprint arXiv:2605.14448*, 2026.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9274–9285, 2025.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. MegaPairs: Massive data synthesis for universal multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 19076–19095, 2025.