DARE: DUAL-ANCHORED RECURRENT REASONING FOR UNIVERSAL MULTIMODAL EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

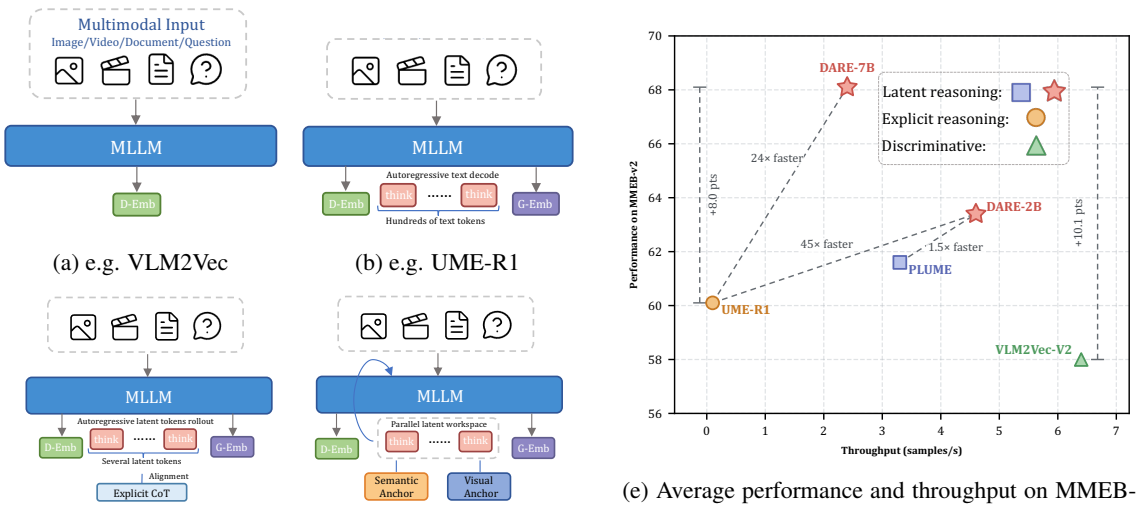
Reasoning-enhanced universal multimodal embeddings have achieved strong performance by incorporating chain-of-thought (CoT) reasoning into representation formation. However, explicit CoT introduces substantial autoregressive decoding latency, while existing latent alternatives largely inherit token-wise rollout and provide limited constraints on how intermediate states should evolve and interact with visual evidence. This raises a fundamental challenge: how to efficiently internalize multimodal reasoning while maintaining semantic structure and faithful visual alignment? To this end, we propose DARE (**D**ual-**A**nchored **R**ecurrent reasoning multimodal **E**mbeddings). DARE performs block-wise recurrent reasoning in a fixed-width latent workspace, updating all latent tokens in parallel within each step and progressively deepening the reasoning trajectory across recurrent iterations, regulated by a dual-anchored mechanism during training: (i) a semantic trajectory anchor that associates recurrent states with ordered reasoning stages to prevent semantic drift; and (ii) a recurrent visual evidence anchor that preserves input-conditioned perceptual evidence across successive refinements through prototype level structural alignment and instance level contrastive matching. Both anchors are used only during training and introduce no auxiliary modules or inputs at inference time. Extensive experiments on 78 MMEB-V2 subsets spanning image, video, and visual-document tasks demonstrate that DARE outperforms previous reasoning-enhanced embedding models, while achieving 45$\times$ and 1.5$\times$ speedups over prior explicit-CoT and latent-reasoning approaches.

1 INTRODUCTION

Multimodal embedding models play a pivotal role in downstream applications such as classification Jia et al. (2021); Girdhar et al. (2023), retrieval Li et al. (2022; 2021), and multimodal retrieval augmented generation Zhang et al. (2026c) by encoding and aligning diverse modalities into a unified vector space. CLIP-style models Radford et al. (2021); Sun et al. (2023); Zhai et al. (2023) employ vision and text transformers such as Dosovitskiy (2020); Devlin et al. (2019) to process visual and textual inputs independently, with subsequent works integrating fusion modules to accommodate interleaved image-text inputs Li et al. (2023). However, such paradigms inherently lack robust reasoning capabilities and world knowledge, leading to suboptimal performance in more demanding scenarios. Recently, embedding models built upon generative multimodal large language models (MLLMs) have demonstrated superior performance by tapping into the models' rich parametric knowledge and perceptual depth.

Existing MLLM-based embedding models can be categorized as illustrated in Figure 1. A typical paradigm Jiang et al. (2024); Meng et al. (2025) utilizes the hidden states of the last token from the final layer as semantic embeddings, aligning query and target via contrastive learning to ultimately yield discriminative embeddings. However, such methods merely exploit their well-aligned representations and initialized weights, leaving reasoning capabilities untapped. To address this, subsequent works employ explicit chain of thought

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.



(a) e.g. VLM2Vec (b) e.g. UME-R1

(c) e.g. PLUME (d) Ours

(e) Average performance and throughput on MMEB-v2. Inference throughput is tested on a single H20 GPU

Figure 1: We categorize current mainstream universal multimodal embedding models into three types in Fig.1a-Fig.1c. We propose our method as shown in Fig. 1d, which maintains a fixed-width latent workspace and performs parallel block-wise recurrent refinement, guided by semantic trajectory and visual evidence anchors during training. Experiments on MMEB-v2 in Fig. 1e demonstrate that we achieve improvements in both performance and throughput.

(CoT) via an external reasoner or the model itself, thereby improving multimodal embeddings by allowing the model to inspect, compare, and aggregate multimodal evidence before final representation generation Cui et al. (2025); Lan et al. (2025b). Nonetheless, sequential decoding of discrete text tokens inevitably introduces substantial computational latency.

Continuous latent reasoning Hao et al. (2024); Shen et al. (2025); Jin et al. (2026) provides a natural alternative by replacing discrete text tokens with hidden states. Nevertheless, prevailing multimodal latent reasoning embedding methods predominantly rely on token-wise autoregressive generation He et al. (2026), severely limiting reasoning depth and practical scalability. More fundamentally, existing methods primarily change the representational form of reasoning, from discrete tokens to continuous states, while largely inheriting the sequential execution, text-centric and loose visual alignment of autoregressive CoT. This leaves three critical dimensions of multimodal latent computation insufficiently explored: (1) how to parallelize latent computation without sacrificing multi-step reasoning; (2) how to effectively supervise intermediate latent states; (3) how to preserve and propagate input-conditioned visual evidence across successive latent refinements. To investigate these questions, we conduct systematic diagnostic probing experiments across perceptual tasks on Perception-Benchmark Lin et al. (2026) and reasoning tasks on MRMR-Benchmark Zhang et al. (2026b) as shown in Fig. 2. Our findings highlight the complementary roles of perceptual encoding and reasoning computation in multimodal latent reasoning. Visual representations contextualized by LLM retain fine-grained perceptual cues but are insufficient for reasoning-intensive tasks. Conversely, naively replacing explicit CoT with parallel latent computation substantially degrades performance on both perception and reasoning-oriented tasks. Inspired by Saunshi et al. (2025), increasing recurrent depth narrows this gap, and intermediate semantic supervision stabilizes latent-state progression. However, semantic guidance alone is insufficient to retain critical visual cues. Further grounding recurrent states in query-relevant visual evidence enables parallel latent reasoning to match or surpass explicit CoT across the evaluated tasks. These results

suggest that effective multimodal latent reasoning depends on both semantic guidance and the preservation of task-relevant visual cues.



Figure 2: Systematic diagnostic probing experiments across perceptual tasks on Perception-Benchmark and reasoning tasks on MRMR-Benchmark

Driven by these insights, we propose DARE, a dual-anchored recurrent reasoning framework for universal multimodal embeddings. Rather than relying on token-wise latent rollouts, DARE formalizes latent reasoning as iterative updates over a fixed-width latent workspace. At each reasoning step, all latent slots are refined concurrently, while across steps, a parameter-shared MLLM iteratively evolves this latent workspace.

To organize this recurrent trajectory and preserve its perceptual basis, DARE introduces two complementary training-only constraints. The semantic trajectory anchor aligns successive recurrent iterations with ordered blocks of the explicit reasoning trace, avoiding functional redundancy and enforcing step-wise progress. The recurrent visual evidence anchor maintains visual continuity throughout recurrent refinement, enabling successive latent states to preserve and build upon input-conditioned perceptual evidence by complementary prototype and instance alignment. In essence, the semantic anchor structures how the latent workspace progresses across reasoning stages, whereas the visual anchor ensures that this progression remains supported by consistent perceptual evidence. Both anchors are removed at inference, preserving an unencumbered latent embedding pipeline with zero auxiliary overhead. Our contributions are summarized as follows:

(1) We introduce **DARE**, a block-wise recurrent latent reasoning framework for multimodal embeddings that updates a fixed-width latent workspace in parallel across recurrent iterations, enabling deep deliberative reasoning without autoregressive token-wise rollout.

(2) We design a dual-anchored mechanism used exclusively during training: a semantic trajectory anchor promotes progressive reasoning and prevents semantic drift, alongside a recurrent visual evidence anchor that maintains visual evidence continuity across successive refinements via prototype and instance level constraints.

(3) Across 78 diverse subsets on MMEB-V2, DARE surpasses previous reasoning-enhanced embedding models while delivering massive $\mathbf{45}\times$ and $\mathbf{1.5}\times$ inference speedups over prior explicit-CoT and latent-reasoning methods, respectively, without requiring any auxiliary inference-time modules.

2 Related Work

Universal Multimodal Embedding. While CLIP-style models Radford et al. (2021); Sun et al. (2023); Zhai et al. (2023) adopt a dual-tower architecture to encode vision and language into vector and align them via contrastive learning, they struggle to effectively handle interleaved image-text inputs. Consequently, recent works have focused on leveraging MLLMs as backbones. Regarding training objectives, VLM2Vec Jiang et al. (2024) reformulates diverse tasks as retrieval, utilizing contrastive learning for representation training. CAFe Yu et al. (2025) and VLaDVA Ouali et al. (2025) integrate next-token prediction with contrastive

learning to preserve generative capabilities, while Moca Chen et al. (2026) introduces masked image modeling and masked language modeling before contrastive learning to enhance perceptual abilities. For training strategies, LLaVE Lan et al. (2025a) employs hard negative mining for a hardness-weighted curriculum, while Gu et al. (2025) bolsters understanding through offline textual discriminative distillation. For data, MegaPairs Zhou et al. (2025) synthesizes high-quality data based on visual similarity to sharpen embedding discriminability, VLM2Vec-V2 Meng et al. (2025) introduces MMEB-V2 dataset, incorporating video and visual documents to enrich evaluation dimensions. Recently, inspired by CoT's success in reasoning-heavy tasks, TTE Cui et al. (2025) utilizes a reasoner to generate CoT traces that guide embedding learning. UME-R1 Lan et al. (2025b) prefixes embedding extraction with CoT generation, enabling joint reasoning and embedding learning in a unified model, but imposing significant computational overhead.

Explicit and Latent Reasoning. CoT prompting Wei et al. (2022) and reasoning-specialized LLMs like DeepSeek-R1 Guo et al. (2025) have substantially pushed the performance upper bound of complex reasoning, inspiring subsequent extensions to multimodal reasoning and representation learning Hao et al. (2026); Jiang et al. (2026). To avoid the latency of generating hundreds of explicit tokens, Coconut Hao et al. (2024) embeds reasoning into latent vectors by curriculum learning Bengio et al. (2009) but scales poorly on Llama3-8B Grattafiori et al. (2024), while CODI's Shen et al. (2025) dual-path distillation incurs heavy training overhead. Furthermore, Xu et al. (2025); Wei et al. (2026); Zhang et al. (2025) substitute explicit CoT with instance-adaptive soft token. Inspired by DeepSeek-OCR Wei et al. (2025), VTI-CoT Zhang et al. (2026a) conducts reasoning within continuous visual embedding spaces. PLUME He et al. (2026) adapts Coconut's latent reasoning to embeddings, boosting performance while drastically reducing latency.

3 METHOD

In Section 3.1, we first provide the preliminaries of universal multimodal embedding and systematically analyze existing methods. In Section 3.2, we conduct diagnostic experiments, which establish the key motivation for our work. Finally, we detail the design of our proposed method, DARE.

3.1 PRELIMINARY

Let $x \in \mathcal{X}$ denote a multimodal input, a universal multimodal embedding model F_θ projects heterogeneous inputs into a shared embedding space as $\mathbf{e}_x = \mathrm{Norm}(F_\theta(x))$. Given a mini-batch of N positive query-target pairs $(q_i, t_i^+)_{i=1}^N$, the model is optimized using the InfoNCE objective:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\mathbf{e}_{q_i}^\top \mathbf{e}_{t_i^+}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{e}_{q_i}^\top \mathbf{e}_{t_j}/\tau)}, \tag{1}$$

where τ is the temperature parameter. The formulation above corresponds to direct embedding extraction without any intermediate reasoning. In reasoning-enhanced UME, the embedding is instead formed conditioned on an intermediate reasoning trajectory. Given an explicit reasoning trajectory $\mathbf{o}_x = [o_1, \ldots, o_L]$, the result embedding is defined as $\mathbf{e}_x^{\text{exp}} = \mathrm{Norm}\left(F_\theta(x, \mathbf{o}_x)\right)$, where the trajectory can be produced by either the embedding model itself or an external reasoner, the training objective is

$$\mathcal{L}_{\text{exp}} = \mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{CoT}} \sum_{\ell=1}^{L} \left(-\log p_\theta(o_\ell \mid x, o_{<\ell})\right). \tag{2}$$

where p_θ denotes the token distribution induced by its language-model head. Latent reasoning enables the embedding model to perform internal reasoning, replacing the explicit CoT with an autoregressive sequence

of continuous states $\mathbf{z}_x = [z_1, \ldots, z_K]$, and the corresponding embedding is $\mathbf{e}_x^{\text{imp}} = \text{Norm}\left(F_\theta(x, \mathbf{z}_x)\right)$, the training objective is

$$\mathcal{L}_{\text{lat}}^{(s)} = \mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{proc}} \left(- \sum_{\ell \in \mathcal{I}_s} \log p_\theta(o_\ell \mid x, \mathbf{z}_x, o_{<\ell}) \right). \tag{3}$$

where $\mathcal{I}_s \subseteq \{1, \ldots, L\}$ denotes the explicit CoT positions retained at curriculum stage s. As the curriculum proceeds, $\mathcal{I}_s$ progressively shrinks. In the final fully latent stage, $\mathcal{I}_s = \varnothing$, and therefore no direct process-level objective remains for organizing the intermediate latent states.

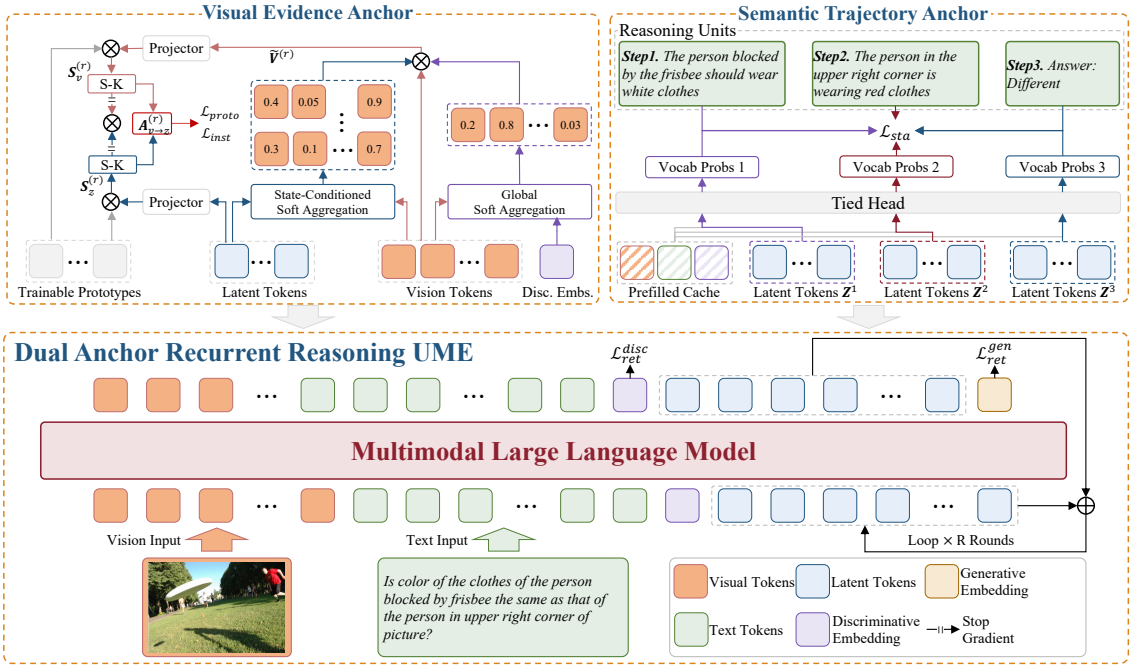


Figure 3: Overview of DARE. A fixed-width latent workspace undergoes block-wise recurrent refinement through a shared MLLM, guided by two training-time constraints: (a) semantic trajectory anchor for ordered reasoning progression, and (b) visual evidence anchor for preserving relevant perceptual information.

3.2 Motivation

To identify the core mechanisms underpinning effective latent multimodal reasoning, we conduct a diagnostic study on fine-grained perception and compositional reasoning tasks. As shown in Fig.2, contextualized visual states preserve fine-grained perceptual information but provide limited capacity for multi-step reasoning. Explicit CoT improves reasoning by sequentially transforming the available evidence, but slightly degrades performance on perception tasks, suggesting that language-centric reasoning may discard visual details. Naively compressing explicit CoT into parallel latent states causes a severe 5.6 points drop on average, demonstrating that one-shot parallelization alone cannot preserve reasoning capacity. Introducing recurrent refinement recovers 2.7 points by iterative computational depth, while aligning successive states with intermediate reasoning targets yields another 2.0 points gain by steering the latent process along a coherent semantic trajectory.

Nevertheless, semantic supervision alone cannot fully restore perceptual fidelity: it constrains what each recurrent state should represent, but not which instance-specific visual evidence should support that representation. We therefore inject query-relevant visual evidence into recurrent states and achieve a further 2.8 points, ultimately surpassing explicit CoT by 1.9 points overall. It shows that, with sufficient recurrent depth and complementary anchors, latent reasoning can remain competitive with explicit CoT while avoiding costly autoregressive generation.

3.3 DARE

Motivated by the above observations, we introduce **DARE** as illustrated in Fig.3. It replaces token-wise latent rollout with a fixed-width workspace whose states are updated in parallel and recurrently refined by a shared MLLM. The recurrent trajectory is structured by two complementary training anchors. The semantic trajectory anchor assigns an ordered reasoning role to each recurrent refinement, while the visual evidence anchor maintains visual evidence continuity across successive refinements.

Block-wise recurrent latent reasoning. Given a multimodal input x, the MLLM first encodes its visual and textual tokens and constructs a prefix key–value cache $\mathcal{C}_x$, which is shared by all subsequent recurrent computations. We decompose the reasoning process into R ordered blocks and associate the r-th block with the r-th recurrent refinement, where $r \in \{1, \dots, R\}$. DARE maintains a fixed-width workspace of N_{lat} latent tokens throughout these refinements, instead of allocating a separate latent sequence to each block. Let $\mathbf{E}_{\text{lat}} \in \mathbb{R}^{N_{\text{lat}} \times d}$ denote the initialized latent embeddings, and let $\text{Dec}_\theta(\mathbf{H}; \mathcal{C}_x)$ denote the hidden states returned by the shared MLLM decoder when an embedding sequence $\mathbf{H}$ is processed with the cached multimodal prefix. The workspace is recurrently updated as

$$\mathbf{Z}^{(1)} = \text{Dec}_\theta(\mathbf{E}_{\text{lat}}; \mathcal{C}_x), \quad \mathbf{Z}^{(r)} = \text{Dec}_\theta\Big(\mathbf{E}_{\text{lat}} + \mathbf{Z}^{(r-1)}; \mathcal{C}_x\Big), \quad r = 2, \dots, R. \tag{4}$$

At each recurrent pass, all N_{lat} latent states are updated concurrently in a single forward pass. The resulting states $\mathbf{Z}^{(r)}$ are regularized by the semantic trajectory anchor introduced below and reused for the subsequent refinement. This recurrent design maps the progression of reasoning directly onto recurrent depth while keeping the latent length fixed, thereby allowing the latent workspace width and the sequential refinement depth to be controlled independently.

After the final refinement, $\mathbf{Z}^{(R)}$ is committed to the context of the final MLLM forward. We extract the hidden state at the $\langle \text{gen_emb} \rangle$ token and apply ℓ_2 normalization to obtain the generative embedding $\mathbf{e}_x^{\text{gen}}$. The recurrent latent tokens therefore serve as an internal computation path for forming the embedding.

Semantic trajectory anchor. While recurrent refinement increases computational depth, terminal retrieval objectives alone provide no granular signals on what each intermediate pass should compute. We first segment the trace at explicit reasoning-step boundaries when available and otherwise use sentence boundaries. Each unit is tokenized using the MLLM backbone tokenizer and right-truncated to at most N_{lat} tokens, shorter units are padded only for batching, and padded positions are excluded from the loss. This produces the ordered units $\mathcal{O}_x = (\mathbf{o}_x^{(1)}, \dots, \mathbf{o}_x^{(R)})$. The r-th unit is denoted by $\mathbf{o}_x^{(r)} = (o_{x,1}^{(r)}, \dots, o_{x,L_r}^{(r)})$, where $L_r \le N_{\text{lat}}$ is its valid length after truncation and $L = \sum_{r=1}^{R} L_r$. We then align the tokens in $o_x^{(r)}$ and supervise them via the MLLM's tied language model head h_{LM} as follows

$$\mathcal{L}_{\text{STA}} = \frac{1}{L} \sum_{r=1}^{R} \sum_{j=1}^{L_r} \text{CE}\Big(h_{\text{LM}}(\mathbf{z}_j^{(r)}), o_{x,j}^{(r)}\Big), \tag{5}$$

Since $\mathbf{Z}^{(r)}$ is recurrently conditioned on $\mathbf{Z}^{(r-1)}$, each refinement is encouraged to extend the preceding computation toward the next reasoning phase, thereby preventing semantic drift and functionally redundant updates. The explicit reasoning trace and tied head supervision are used only during training, so DARE retains the efficiency of pure latent reasoning at inference.

Visual evidence anchor. Semantic anchor does not ensure that visual evidence survives successive recurrent updates. As the latent workspace advances from one reasoning block to the next, its states may remain semantically coherent while gradually weakening their connection to the specific visual content that supports the reasoning process. Motivated by our diagnostic findings, we extract $\mathbf{V}_x = (\mathbf{v}_1, \ldots, \mathbf{v}_P) \in \mathbb{R}^{P\times d}$ from the final MLLM decoder layer at the visual-token positions during prefix encoding as a visually rich reference. Before r-th recurrent refinement ($r > 1$), the discriminative embedding provides a stable global reference, while the preceding latent workspace $\mathbf{Z}^{(r-1)}$ identifies the visual content associated with the current reasoning state. We then produce a set of visual anchors by computing soft attention weights

$$\mathbf{W}^{(r)} = \mathrm{Softmax}_v\left(\frac{\mathrm{Norm}(\mathbf{G}^{(r-1)})\,\mathrm{Norm}(\mathbf{V}_x)^\top}{\tau_s}\right), \qquad \widetilde{\mathbf{V}}^{(r)} = \mathrm{sg}\Big(\mathbf{W}^{(r)}\mathbf{V}_x\Big). \tag{6}$$

where $\mathbf{G}^{(r-1)} = [\mathbf{e}_x^{\mathrm{disc}}; \mathbf{Z}^{(r-1)}]$, $[\,;\,]$ denotes sequence concatenation and τ_s controls the sharpness of visual aggregation. The resulting evidence $\widetilde{\mathbf{V}}^{(r)}$ records the visual basis associated with $\mathbf{Z}^{(r-1)}$ and defines the perceptual basis that the r-th refinement must carry forward and reason upon. Together with the block-wise semantic supervision, it allows the recurrent update to advance toward the next reasoning block without forgetting or attenuating evidence established by the preceding state.

We next align this retained visual basis $\widetilde{\mathbf{V}}^{(r)}$ with the updated latent workspace $\mathbf{Z}^{(r)}$ to enforce visual consistency. Since latent tokens dynamically shift their functional roles across recurrent reasoning iterations, imposing a rigid, token-wise or positional correspondence is overly restrictive. Therefore, we align $\widetilde{\mathbf{V}}^{(r)}$ and $\mathbf{Z}^{(r)}$ at the semantic level through a shared bank of N_{proto} learnable prototypes $\mathbf{P}_v \in \mathbb{R}^{N_{\mathrm{proto}}\times d_p}$. A shared projection head $g : \mathbb{R}^d \to \mathbb{R}^{d_p}$ first maps both state types into the prototype space:

$$\mathbf{S}_z^{(r)} = \mathrm{Norm}(g(\mathbf{Z}^{(r)}))\mathbf{P}_v^\top, \quad \mathbf{S}_v^{(r)} = \mathrm{Norm}(g(\widetilde{\mathbf{V}}^{(r)}))\mathbf{P}_v^\top, \quad \mathbf{Q}_a^{(r)} = \mathrm{sg}\Big[\mathrm{SK}(\mathbf{S}_a^{(r)}/\epsilon)\Big], \quad a \in \{z, v\}. \tag{7}$$

where Norm performs row-wise ℓ_2 normalization and SK denotes Sinkhorn-Knopp normalization Knight (2008). Sinkhorn balancing is computed over valid states collected across the batch to prevent prototype collapse, while visual latent correspondence is computed separately within each instance. Specifically, we obtain the row-normalized correspondence matrix

$$\mathbf{A}_{v\to z}^{(r)} = \mathrm{Softmax}_{\mathrm{row}}\Big(\mathbf{Q}_v^{(r)}\mathbf{Q}_z^{(r)\top}/\tau_m\Big), \tag{8}$$

where row-wise softmax is taken over the N_{lat} latent states and τ_m is matching temperature. Thus, each row of $\mathbf{A}_{v\to z}^{(r)}$ assigns one visual evidence state to a soft combination of latent states. Let $\mathbf{\Phi}_z^{(r)} = \mathrm{Softmax}(\mathbf{S}_z^{(r)}/\tau_p)$ denote the prototype predictions of refined latent states, where τ_p is the prototype temperature. We use the same correspondence to aggregate prototype predictions and continuous latent features:

$$\widehat{\mathbf{Q}}_v^{(r)} = \mathbf{A}_{v\to z}^{(r)}\mathbf{\Phi}_z^{(r)}, \qquad \widehat{\mathbf{U}}_v^{(r)} = \mathrm{Norm}\Big(\mathbf{A}_{v\to z}^{(r)}\,\mathrm{Norm}(g(\mathbf{Z}^{(r)}))\Big). \tag{9}$$

The visual evidence anchor jointly optimizes prototype-level structural coverage and instance-level fidelity:

$$\mathcal{L}_{\mathrm{vea}} = \frac{1}{R-1}\sum_{r=2}^{R}\Big[\,\mathrm{CE}\Big(\widehat{\mathbf{Q}}_v^{(r)}, \mathbf{Q}_v^{(r)}\Big) + \lambda_{\mathrm{inst}}\mathcal{L}_{\mathrm{InfoNCE}}\Big(\widehat{\mathbf{U}}_v^{(r)}, \mathrm{sg}\Big(\mathrm{Norm}(g(\widetilde{\mathbf{V}}^{(r)}))\Big)\Big)\Big]. \tag{10}$$

Specifically, the prototype term encourages the recurrent workspace to capture diverse semantic structures without imposing rigid, token-wise correspondences, while the balanced Sinkhorn assignment safeguards against prototype collapse. For the instance-level objective, each predicted visual anchor and its corresponding visual representation form a positive pair. Visual anchors from other inputs across devices serve as negatives, while anchors from the same input are excluded from the negative set. For variable-length visual sequences, padded states are masked before Sinkhorn normalization and contrastive matching.

Table 1: Results on the MMEB-V2 benchmark. CLS: classification, QA: question answer, RET: retrieval, GD: grounding, MRET: moment retrieval, VDR: ViDoRe, VR: VisRAG, OOD: out-of-distribution. Reported metrics adhere to the settings of VLM2Vec-V2. Detailed results can be found in Appendix.

Model	Backbone	Image					Video					VisDoc					All
		CLS	QA	RET	GD	Avg.	CLS	QA	RET	MRET	Avg.	VDRv1	VDRv2	VR	OOD	Avg.	
# of Datasets		10	10	12	4	36	5	5	5	3	18	10	4	6	4	24	78
~ 2B Model Size																	
ColPali-V1.3	PaliGemma	40.3	11.5	48.1	40.3	34.9	26.7	37.8	21.6	25.5	28.2	83.6	52.0	81.1	43.1	71.0	44.4
GME	Qwen2-VL	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
VLM2Vec	Qwen2-VL	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	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
LamRA	Qwen2-VL	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
UME-R1	Qwen2-VL	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
PLUME	Qwen2-VL	66.5	59.2	67.6	79.7	66.3	45.0	**52.3**	33.5	**46.7**	44.1	72.1	49.8	78.1	57.4	67.5	61.6
Ours	Qwen2-VL	**68.3**	62.6	68.6	**80.5**	**68.2**	**47.4**	51.1	**36.5**	43.6	**44.8**	74.7	50.2	81.4	**61.9**	70.2	**63.4**
~ 7B Model Size																	
GME	Qwen2-VL	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	Qwen2.5-VL	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	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
DUME	Qwen2-VL	64.2	57.0	70.8	81.8	66.4	32.9	47.4	8.6	28.0	29.4	67.1	35.2	82.6	34.9	60.3	55.9
CAFe	LLaVA-OV	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	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
Ours	Qwen2-VL	**70.8**	**70.7**	**72.4**	85.4	**72.9**	**51.1**	60.6	**39.5**	**46.2**	**49.7**	**79.8**	**56.3**	**84.0**	**66.1**	**74.6**	**68.1**

Training objective. The overall training objective integrates both discriminative and generative retrieval losses alongside the two training-time anchors

$$\mathcal{L} = \mathcal{L}_{\text{ret}}^{\text{disc}} + \mathcal{L}_{\text{ret}}^{\text{gen}} + \lambda_{\text{sta}}\mathcal{L}_{\text{sta}} + \lambda_{\text{vea}}\mathcal{L}_{\text{vea}}, \tag{11}$$

$\mathcal{L}_{\text{ret}}$ is the standard bidirectional InfoNCE loss applied to the query and target embeddings at $\langle \text{disc_emb} \rangle$ and $\langle \text{gen_emb} \rangle$ token. Unless otherwise specified, subsequent experiments are conducted using $\langle \text{gen_emb} \rangle$.

4 Experiments

4.1 Setup

Training. Following VLM2Vec-V2 and UME-R1, our model is built on Qwen2-VL-2B/7B-Instruct Wang et al. (2024) with the training corpus of UME-R1. We employ 75 latent tokens and set the number of recurrent refinement iterations to 3. The visual encoder remains frozen, while the remaining parameters are fully fine-tuned. We set all temperature parameters to 1 except $\tau = 0.02$, loss weight $\lambda_{vea} = 0.5$, $\lambda_{sta} = 1.0$. The global batch size is 1024, achieved through gradient accumulation. We optimize the model using AdamW with a learning rate of 2×10^{-5}.

Evaluation. To evaluate the performance of our approach, we conduct comprehensive experiments on the MMEB-V2 benchmark. Following standard protocols, we report Hit@1 for the image and video scenarios, and NDCG@5 for the visual document scenarios. As shown in Table 1, our compared methods include ColPali Faysse et al. (2025), LamRA Liu et al. (2025), etc.

4.2 Main Results

Table 1 summarizes the performance of DARE against representative discriminative, explicit-CoT, and latent-reasoning baselines on MMEB-V2. Overall, DARE achieves consistently superior performance across all three modalities: with Qwen2-VL-2B, it averages 63.4 across all modalities, outperforming explicit CoT reasoning model UME-R1 and latent reasoning model PLUME by 3.3 and 1.8 points while being 45× and

1.5× faster, respectively, as shown in Fig. 1e. It confirms that DARE constructs more visually-informed semantic representations via block-wise recurrent refinement, guided by the dual anchors. Furthermore, it exhibits favorable scaling properties: DARE-7B advances the average performance to 68.1, 8.0 points improvement with 24× speedup over UME-R1-2B.

4.3 Ablation Studies

We conduct ablation studies to verify the contribution of each core component and evaluate the impact of parameters related to the block-wise recurrent mechanism.

Core components. As shown in Table 2, all three components contribute to the final performance. Introducing block-wise recurrent reasoning improves the overall score from 59.2 to 60.9, demonstrating the benefit of increasing latent computational depth. The semantic trajectory anchor further improves the overall score by 1.2 points, confirming that assigning ordered semantic targets to recurrent states leads to more structured latent evolution. The visual evidence anchor yields larger improvements on image and video tasks, indicating that visual evidence is important for visual perception and reasoning. Combining the two anchors achieves the best performance, reaching 63.4 overall. Compared with discriminative embedding baseline, DARE improves 4.2 points on average. These results demonstrate that every proposed component contributes meaningfully.

Table 2: Ablation on core components on MMEB-V2. BWR: block-wise recurrent, STA: semantic trajectory anchor, VEA: visual evidence anchor. P: prototype term. I: instance term

BWR	STA	VEA-P	VEA-I	Image	Video	VisDoc	All
				65.3	37.7	66.2	59.2
✓				66.1	40.5	68.6	60.9
✓	✓			66.9	42.3	69.7	62.1
✓	✓	✓		67.7	42.6	69.3	62.4
✓	✓	✓	✓	**68.2**	**44.8**	**70.2**	**63.4**

Width and depth of latent reasoning. We study the effects of latent workspace width and recurrent depth as shown in Table 3. Expanding the workspace to $N_{lat} = 75$ yields a clear performance gain from 60.3 to 61.7 with negligible throughput degradation, benefiting from parallel latent updates. With 75 latent tokens, increasing the refinement depth to $R = 3$ further lifts the score by 1.7 points. Continuing to scale the depth to $R = 4$ still brings additional gains. Considering the trade-off between performance and throughput, we adopt $N_{lat} = 75$ and $R = 3$ as a favorable trade-off. These results demonstrate the scalability of our method, indicating that DARE can scale workspace capacity efficiently while extending reasoning depth to consistently boost performance.

Table 3: Ablation on width and depth of latent reasoning on MMEB-V2.

N_{lat}	R	Image	Video	VisDoc	All	Throu.
50	2	66.4	40.1	66.2	60.3	6.2
75	2	67.2	43.1	67.5	61.7	5.9
75	3	68.2	44.8	70.2	63.4	4.6
75	4	68.7	45.1	70.7	63.8	3.1

5 Conclusion

In this work, we presented DARE, a dual-anchored recurrent reasoning framework for universal multimodal embeddings. DARE performs block-wise refinement within a fixed-width latent workspace, where all latent states are updated in parallel at each recurrent pass, regulated by semantic trajectory anchor and visual evidence anchor. The former transfers ordered reasoning structure into successive latent states, while the latter prevents relevant perceptual information from being attenuated during semantic refinement. Across 78 MMEB-V2 subsets at both 2B and 7B scales, DARE achieves higher performance than prior reasoning-enhanced embedding models, with $45\times$ the throughput of explicit-CoT embedding and $1.5\times$ that of autoregressive latent reasoning. We will extend DARE with input-adaptive recurrent depth to allocate computation according to input complexity in future work.

AI USE STATEMENT

In this work, we used generative AI tools to polish the manuscript's language and to assist with refining initial code implementations. All foundational scientific components including motivation, methodology, experimental design were developed entirely independently by the authors. All AI-assisted text and code were rigorously vetted, verified, and edited across multiple authors to ensure technical correctness, factual accuracy, and proper citations. We take responsibility for the final content of this work, including all text and code produced or refined with the aid of generative AI.

REFERENCES

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Haonan Chen, Hong Liu, Yuping Luo, Liang Wang, Nan Yang, Furu Wei, and Zhicheng Dou. 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.

Xuanming Cui, Jianpeng Cheng, Hong-you Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Kumar Mishra, Yonghuan Yang, Jun Xiao, et al. Think then embed: Generative context improves multimodal embedding. *arXiv preprint arXiv:2510.05014*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 61424–61449, 2025.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. Breaking the modality barrier: Universal embedding learning with multimodal llms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2860–2869, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Xiangzhao Hao, Shijie Wang, Tianyu Yang, Tianyue Wang, Haiyun Guo, and Jinqiao Wang. 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.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Haonan Jiang, Yuji Wang, Yongjie Zhu, Xin Lu, Wenyu Qin, Meng Wang, Pengfei Wan, and Yansong Tang. Embed-rl: Reinforcement learning for reasoning-driven multimodal embeddings. In *European Conference on Computer Vision*, pp. 39–56. Springer, 2026.

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

Jiajie Jin, Yanzhao Zhang, Mingxin Li, Dingkun Long, Pengjun Xie, Yutao Zhu, and Zhicheng Dou. Laser: Internalizing explicit reasoning into latent space for dense retrieval. *arXiv preprint arXiv:2603.01425*, 2026.

Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. *arXiv preprint arXiv:2503.04812*, 2025a.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Ume-r1: Exploring reasoning-driven generative multimodal embeddings. *arXiv preprint arXiv:2511.00405*, 2025b.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Zichao Lin, Yifeng Xie, Bowen Qu, Haiming Wang, Jia Li, Haoning Wu, Yuhao Dong, Zuhao Yang, Jinguo Zhu, Haoyu Lu, et al. Perceptionbench: Evaluating atomic visual perception in multimodal large language models. *arXiv preprint arXiv:2607.24957*, 2026.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. 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, 2025.

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.

Yassine Ouali, Adrian Bulat, Alexandros Xenos, Anestis Zaganidis, Ioannis Maniadis Metaxas, Brais Martinez, and Georgios Tzimiropoulos. Vladva: Discriminative fine-tuning of lvlms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4101–4111, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J Reddi. Reasoning with latent thoughts: On the power of looped transformers. In *International Conference on Learning Representations*, volume 2025, pp. 14855–14881, 2025.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 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.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

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

Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Sim-cot: Supervised implicit chain-of-thought. In *International Conference on Learning Representations*, volume 2026, pp. 56721–56742, 2026.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. Softcot: Soft chain-of-thought for efficient reasoning with llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23336–23351, 2025.

Hao Yu, Zhuokai Zhao, Shen Yan, Lukasz Korycki, Jianyu Wang, Baosheng He, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, and Hanchao Yu. Cafe: Unifying representation and generation with contrastive-autoregressive finetuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6286–6297, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. Lightthinker: Thinking step-by-step compression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 13318–13339, 2025.

Shufan Zhang, Ziyue Lin, Bairun Wang, Lei Jin, Xuanding Ding, Xinzhu Ma, and Kunlin Yang. Vti-cot: Visual-textual interleaved chain of thought for video reasoning. *arXiv preprint arXiv:2606.05736*, 2026a.

Siyue Zhang, Yuan Gao, Xiao Zhou, Yilun Zhao, Tingyu Song, Arman Cohan, Anh Tuan Luu, and Chen Zhao. Mrmr: A realistic and expert-level multidisciplinary benchmark for reasoning-intensive multimodal retrieval. In *International Conference on Learning Representations*, volume 2026, pp. 100517–100555, 2026b.

Tao Zhang, Ziqi Zhang, Zongyang Ma, Yuxin Yang, Bing Li, Chunfeng Yuan, Kang Rong, Fengyun Rao, Jing Lyu, and Weiming Hu. Mmagent-r2: Learning to rerank and reject for agentic mrag. *arXiv preprint arXiv:2607.07383*, 2026c.

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 (Volume 1: Long Papers)*, pp. 19076–19095, 2025.

A APPENDIX

We report per-dataset results on all 78 subsets of MMEB-V2 as follows.

Table 4: Detailed results of baselines and DARE on MMEB-v2 benchmark. Video and VisDoc results are shown on the next table.

	ColPali	GME-7B	VLM2Vec-7B	VLM2Vec-V2-2B	CAFe-7B	UME-R1-2B	UME-R1-7B	PLUME-2B	DARE-2B	DARE-7B
Avg - All (78 tasks)	44.4	57.8	52.3	58.0	60.6	60.1	64.5	61.6	63.4	68.1
Avg - Image (36 tasks, Hit@1)	34.9	56.0	65.5	64.9	67.6	66.6	71.3	66.3	68.2	72.9
Avg - Video (18 tasks, Hit@1)	28.2	38.4	33.7	34.6	42.4	42.2	47.5	44.1	44.8	49.7
Avg - Visdoc (24 tasks, NDCG@5)	71.0	75.2	46.4	65.4	63.9	63.9	67.1	67.5	70.2	74.6
I-CLS (10)	40.3	57.7	62.7	62.9	63.6	64.8	67.1	66.5	68.3	70.8
I-QA (10)	11.5	34.7	56.9	56.3	61.7	62.8	69.2	59.2	62.6	70.7
I-RET (12)	48.1	71.2	69.4	69.5	69.1	67.6	71.9	67.6	68.6	72.4
I-VG (4)	40.3	59.3	82.2	77.3	87.6	77.2	84.9	79.7	80.5	85.4
V-CLS (5)	26.7	37.4	39.1	39.3	35.8	44.3	48.6	45.0	47.4	51.1
V-QA (5)	37.8	50.4	30.0	34.3	58.7	51.0	60.7	52.3	51.1	60.6
V-RET (5)	21.6	28.4	29.0	28.8	34.4	32.9	38.2	33.5	36.5	39.5
V-MR (3)	25.5	37.0	38.9	36.8	39.5	39.7	39.3	46.7	43.6	46.2
VD-Vidore-V1 (10)	83.6	89.4	56.9	75.7	70.7	72.4	75.7	72.1	74.7	79.8
VD-Vidore-V2 (4)	52.0	55.6	9.4	45.1	49.6	46.2	50.5	49.8	50.2	56.3
VD-VisRAG (6)	81.1	85.0	59.1	79.6	79.5	79.2	83.7	78.1	81.4	84.0
VD-OOD (4)	43.1	44.4	38.1	39.6	38.1	37.2	37.6	57.4	61.9	66.1
ImageNet-1K	42.4	64.6	80.1	80.8	77.3	75.3	80.4	74.1	78.9	83.9
N24News	25.5	50.5	79.7	72.9	83.2	81.1	82.3	81.1	80.6	81.9
HatefulMemes	50.6	53.6	69.7	56.3	78.7	75.2	79.0	75.5	77.5	79.0
VOC2007	69.8	80.3	80.7	85.0	89.8	80.0	90.8	86.1	89.9	91.9
SUN397	56.1	69.5	77.4	71.0	79.9	79.4	80.3	76.9	81.9	82.5
Place365	27.5	39.1	37.4	35.9	45.0	42.6	46.8	42.4	42.0	46.0
ImageNet-A	14.9	41.2	58.1	47.4	55.2	50.4	53.9	50.8	53.7	55.4
ImageNet-R	64.6	83.9	73.9	89.3	88.0	88.7	90.1	87.5	90.1	90.3
ObjectNet	45.6	69.0	40.1	65.2	22.5	52.0	42.3	61.5	65.9	70.6
Country211	6.0	24.8	29.8	25.2	16.7	23.4	25.0	25.0	22.8	26.9
OK-VQA	9.4	33.2	56.8	51.5	67.3	62.4	71.7	60.5	62.1	71.1
A-OKVQA	6.6	21.0	47.3	43.6	63.8	51.1	58.7	49.9	53.7	61.4
DocVQA	11.3	41.4	89.7	90.1	79.2	92.2	93.8	89.9	92.6	96.0
InfographicsVQA	5.0	20.3	60.0	58.8	53.3	67.7	79.2	59.6	68.4	78.2
ChartQA	5.7	17.8	56.9	47.4	48.8	64.9	75.1	49.8	55.2	71.1
Visual7W	6.1	22.2	52.7	52.9	52.5	54.1	55.2	47.6	53.5	59.3
ScienceQA	16.3	28.0	38.5	38.2	65.4	42.7	53.7	42.9	41.8	57.1
VizWiz	27.6	39.0	39.9	43.3	43.8	46.8	51.6	46.5	47.0	53.1
GQA	8.3	76.9	55.1	64.9	65.7	67.3	69.3	69.1	71.5	72.6
TextVQA	18.8	46.8	71.6	72.2	76.8	78.6	83.5	78.9	80.6	87.8
VisDial	41.2	60.8	81.9	82.7	82.7	76.6	80.7	72.6	75.3	83.2
CIRR	8.2	54.9	51.1	57.5	60.4	53.7	55.3	54.6	55.9	57.2
VisualNews_t2i	50.1	79.7	80.5	74.5	69.5	71.7	76.8	71.3	72.5	79.0
VisualNews_i2t	47.6	83.6	81.2	78.2	79.4	74.2	82.0	72.7	75.9	83.8
MSCOCO_t2i	59.2	71.2	77.2	75.3	75.4	75.1	78.3	74.1	76.6	79.3
MSCOCO_i2t	49.9	57.7	73.9	71.4	73.1	68.9	71.4	69.8	69.9	73.3
NIGHTS	65.5	67.6	67.6	68.6	66.7	67.2	68.1	68.0	69.0	69.2
WebQA	53.8	91.4	88.3	90.6	89.3	90.0	90.9	89.1	89.7	90.6
FashionIQ	5.9	37.8	17.1	19.5	39.0	17.1	23.4	20.3	19.9	20.4
Wiki-SS-NQ	80.5	78.2	62.3	66.9	61.2	62.0	72.5	68.6	67.3	73.6
OVEN	50.0	75.1	66.5	64.3	60.8	66.9	71.4	68.4	67.3	68.1
EDIS	64.7	96.0	85.7	84.1	71.3	88.0	92.0	81.8	83.5	91.6
MSCOCO	36.7	31.4	75.7	67.1	84.7	69.5	72.7	66.9	69.1	73.4
RefCOCO	64.5	60.9	87.6	87.1	89.4	83.3	91.4	86.5	86.1	92.2
RefCOCO-Matching	3.9	78.4	84.6	85.8	83.0	84.4	91.1	88.4	88.3	94.2
Visual7W-Pointing	56.1	66.5	81.0	69.2	93.2	71.5	84.2	74.9	78.5	81.9

Table 5: Detailed results of baselines and DARE on Video and Visual Doc of MMEB-v2 benchmark.

	ColPali	GME-7B	VLM2Vec-7B	VLM2Vec-V2-2B	CAFe-7B	UME-R1-2B	UME-R1-7B	PLUME-2B	DARE-2B	DARE-7B
K700	23.4	39.7	35.5	38.0	40.1	35.8	42.8	45.1	47.2	54.4
SmthSmthV2	25.1	30.6	32.1	42.8	35.8	44.1	50.4	44.8	51.1	53.8
HMDB51	24.8	47.9	42.2	40.9	46.9	54.4	58.3	51.2	56.0	58.3
UCF101	49.4	54.7	61.8	60.0	39.6	67.2	70.0	66.5	60.8	65.5
Breakfast	10.9	14.3	23.8	14.8	16.6	20.1	21.5	20.1	21.9	23.6
MVBench	33.7	46.6	28.5	33.7	48.9	49.9	58.2	47.4	45.6	57.7
Video-MME	30.6	39.2	27.8	30.7	46.0	41.7	47.3	40.0	38.9	48.3
NExTQA	35.2	53.6	20.3	20.9	62.4	59.9	69.6	57.3	56.5	69.5
EgoSchema	38.4	46.8	21.8	34.0	60.0	45.4	52.4	47.8	44.8	51.2
ActivityNetQA	51.3	65.6	51.4	52.3	76.0	57.8	76.0	69.2	69.9	76.4
DiDeMo	22.8	26.4	29.3	30.4	37.8	32.4	40.0	32.7	36.5	38.9
MSR-VTT	17.6	31.8	34.5	28.3	36.5	34.3	38.9	36.2	39.5	42.3
MSVD	45.4	49.7	46.7	48.1	56.4	55.4	60.8	56.1	59.1	62.5
VATEX	16.7	24.9	25.5	26.5	32.0	29.9	32.6	28.2	32.0	35.5
YouCook2	5.3	9.1	9.0	10.6	9.5	12.7	18.5	14.5	15.5	18.3
QVHighlight	19.9	59.5	57.7	49.4	58.4	57.5	54.9	57.1	60.5	65.7
Charades-STA	29.0	14.0	19.8	20.2	18.7	20.4	21.9	19.4	22.3	23.6
MomentSeeker	27.6	37.4	39.3	40.8	41.4	41.2	41.1	63.5	48.0	49.4
ViDoRe_arxivqa	81.7	86.9	60.2	80.6	73.3	73.9	73.6	72.6	77.7	80.3
ViDoRe_docvqa	56.6	57.5	34.7	44.9	38.3	37.9	41.1	36.2	42.3	50.6
ViDoRe_infovqa	84.9	91.6	70.4	83.7	80.6	76.2	80.8	79.0	79.5	83.1
ViDoRe_tabfquad	86.9	94.6	78.2	89.2	80.7	86.1	90.2	88.8	85.7	93.1
ViDoRe_tatdqa	70.9	74.1	27.6	43.8	37.8	40.6	46.7	36.6	43.9	49.8
ViDoRe_shiftproject	75.1	96.8	38.6	60.8	52.0	66.8	65.0	64.8	69.5	74.9
ViDoRe_artificial_intelligence	95.7	99.6	67.7	88.5	86.0	85.9	89.5	83.8	88.6	93.7
ViDoRe_energy	94.7	95.3	60.4	86.5	84.8	83.3	85.7	82.6	83.2	88.7
ViDoRe_government_reports	93.6	98.8	61.8	85.0	85.0	82.6	89.8	83.2	86.6	91.2
ViDoRe_healthcare_industry	95.9	99.3	69.9	92.2	88.4	90.8	94.3	91.1	89.7	92.7
ViDoRe_esg_reports_human_labeled_v2	51.3	63.4	6.8	45.6	50.7	50.2	50.4	52.1	56.9	58.6
ViDoRe_biomedical_lectures_v2_multilingual	54.7	49.5	5.1	44.3	50.9	46.2	50.7	48.2	47.5	54.1
ViDoRe_economics_reports_v2_multilingual	49.0	54.2	13.9	43.0	54.3	45.7	57.8	49.6	47.2	56.4
ViDoRe_esg_reports_v2_multilingual	52.9	55.4	11.9	46.6	42.3	42.7	43.2	49.0	49.2	56.1
VisRAG_ArxivQA	80.9	87.4	52.6	76.9	74.0	74.3	80.5	71.6	76.8	80.4
VisRAG_ChartQA	72.3	86.1	57.7	83.7	82.7	86.0	85.0	80.8	85.2	86.3
VisRAG_MP-DocVQA	82.0	89.7	60.6	88.1	75.1	75.6	83.4	74.9	79.8	82.7
VisRAG_SlideVQA	85.1	92.6	54.7	84.1	87.6	87.1	91.5	88.9	89.9	92.4
VisRAG_InfoVQA	83.5	88.6	66.0	82.3	87.9	84.4	89.2	85.7	86.1	88.5
VisRAG_PlotQA	79.3	76.5	62.7	75.9	69.4	68.0	72.7	66.2	70.4	73.7
ViDoSeek-page	38.1	32.6	16.3	29.1	22.5	21.2	21.3	80.6	85.4	88.4
ViDoSeek-doc	87.5	90.3	69.4	79.0	73.8	75.9	75.3	76.9	74.1	78.2
MMLongBench-page	27.1	36.9	0.4	15.8	13.3	11.9	12.3	39.9	45.9	50.5
MMLongBench-doc	80.4	85.2	28.8	63.0	42.6	39.7	41.3	32.0	42.4	47.3