FROM EXPLICIT COT TO RELEVANCE GUIDANCE: DISCRIMINATIVE LATENT REASONING FOR UNIVERSAL MULTIMODAL EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning has shown promise for improving universal multimodal embeddings, through both explicit chain-of-thought (CoT) and CoT-supervised latent reasoning. However, explicit CoT may not reliably convey the discriminative information that embeddings rely on to preserve distinctions among inputs. We argue that such information can be acquired and internalized by reasoning without necessarily relying on explicit CoT. To realize this view, we introduce Discriminative Latent Reasoning for Universal Multimodal Embedding (DiLRE), which establishes a discriminative foundation for latent reasoning through contrastive supervision and internalizes complementary information through relevance guidance. We train a Discriminative Guidance Model to jointly assess query–target relevance, providing guidance that complements embedding similarity. At test time, this guidance is combined with the embedding's current retrieval distribution to form discriminative targets for latent refinement. An iterative loop of retrieval and latent refinement then updates the latent reasoning tokens to align the resulting embedding's retrieval distribution with these targets. Extensive experiments and ablation studies demonstrate the effectiveness of DiLRE across diverse multimodal embedding tasks and model scales.

1 INTRODUCTION

Universal multimodal embedding (UME) represents inputs across modalities and tasks in a shared vector space, providing a unified interface for multimodal information access. By comparing independently encoded queries and candidates, this interface supports diverse tasks over text, images, videos, and visual documents. The VLM2Vec family brings the multimodal understanding and instruction-following capabilities of multimodal large language models into embedding learning, allowing representations to respond to different task requirements (Jiang et al., 2024b; Meng et al., 2025). Building on this foundation, reasoning for embedding has emerged as a promising direction for extending the information that these representations can capture and use.

Recent studies demonstrate the potential of explicit reasoning to improve multimodal embeddings (Cui et al., 2025; Lan et al., 2025; Liu et al., 2025a). A central focus has been the design of explicit CoT, with recent work seeking to make the generated reasoning more informative for the resulting embedding (Jiang et al., 2026a; Wu et al., 2026a). However, how different CoT formulations affect embedding quality is not yet fully understood, and recent evidence suggests that indiscriminately applying CoT may even be detrimental (Wang et al., 2026b). Latent-reasoning methods for universal multimodal embedding replace long explicit CoT traces with a small number of latent tokens, yet still rely on explicit CoT supervision and explore which explicit formulations provide more useful information for the output embedding (He et al., 2026; Wu et al., 2026b). Taken together, the ongoing exploration of CoT formulations across explicit and latent reasoning points to a more fundamental question: *what information internalized by reasoning can benefit the resulting embedding?*

An embedding organizes multimodal semantics into a shared representation space that preserves distinctions, and reasoning offers a way to enrich the information supporting those distinctions. Our view is that *reasoning can benefit embeddings by internalizing discriminative information, whose*

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.

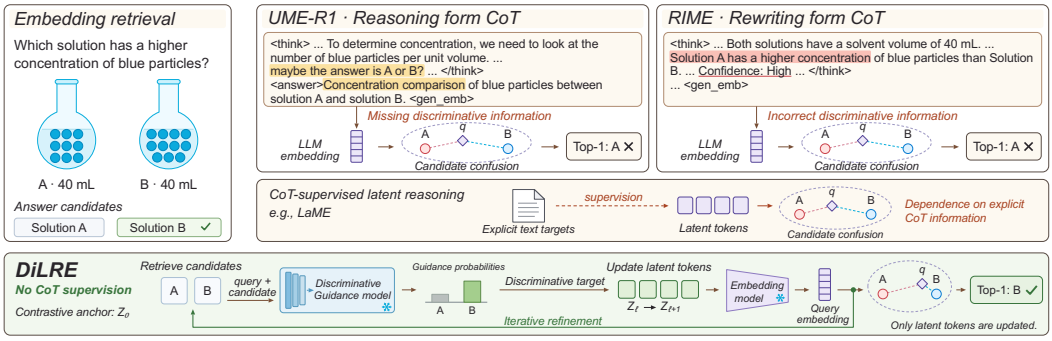


Figure 1: **Missing or incorrect discriminative information in explicit CoT.** The two examples show explicit CoT either leaving the decisive candidate distinction unresolved or stating it incorrectly. DiLRE instead uses relevance guidance to provide discriminative information directly to latent reasoning, refining the query embedding to retrieve the correct candidate.

acquisition does not necessarily depend on explicitly verbalized reasoning. As illustrated in Figure 1, explicit CoT may remain semantically relevant to a query yet leave the distinction between plausible answers unresolved, or even encode an incorrect distinction in a more definite rationale. These cases suggest that explicit CoT may have limitations as a means of reliably conveying the discriminative information needed by the embedding. This motivates us to explore an alternative in which discriminative information is supplied directly to the reasoning process, without being mediated by explicit CoT. A latent formulation offers this possibility: discriminative signals can be directly internalized within latent reasoning states to improve the resulting embedding.

To realize this view, we introduce *Discriminative Latent Reasoning for Universal Multimodal Embedding* (DiLRE). DiLRE implements latent reasoning with a small set of special tokens and organizes its design around two aspects: the provision of discriminative guidance and its internalization into latent reasoning. For guidance provision, contrastive supervision provides a discriminative foundation, while a Discriminative Guidance Model learns to provide complementary relevance guidance beyond embedding similarity. For internalization, contrastive supervision trains the latent reasoning tokens to establish a discriminative anchor during training. At test time, relevance guidance is combined with current retrieval preferences to guide iterative refinement of the latent reasoning tokens, incorporating additional discriminative information into latent reasoning to refine the resulting embedding. In this way, DiLRE enables reasoning to internalize discriminative information directly from contrastive supervision and relevance guidance, without relying on explicit CoT to convey that information.

Extensive experiments across diverse multimodal retrieval tasks and model scales demonstrate the effectiveness of DiLRE. Ablation studies substantiate the complementary contributions of discriminative supervision, relevance guidance, and latent refinement, while further analyses show that the method can balance retrieval quality and inference efficiency. Together, these results support the central premise that latent reasoning can improve universal multimodal embeddings by internalizing discriminative information.

Our contributions are threefold:

- We propose a discriminative perspective on reasoning for UME: reasoning can benefit embeddings by internalizing discriminative information, whose acquisition does not necessarily depend on explicitly verbalized reasoning.
- We introduce DiLRE, a framework that couples the provision of discriminative signals with their internalization into latent reasoning. Contrastive supervision establishes a discriminative anchor for latent reasoning tokens without explicit CoT supervision, while complementary relevance signals from the Discriminative Guidance Model guide their iterative refinement at test time.
- We demonstrate the effectiveness of DiLRE across diverse multimodal embedding tasks and model scales through extensive experiments and ablations.

2 Related Work

2.1 Multimodal Embedding Models

Early vision–language models learn shared embeddings through image–text alignment (Radford et al., 2021; Zhai et al., 2023). Multimodal large language models extend this paradigm to universal multimodal embedding (Jiang et al., 2024a;b; Lin et al., 2025; Zhang et al., 2024a; Meng et al., 2025). Building on CoT reasoning in language and multimodal models (Wei et al., 2022; Zhang et al., 2024b), explicit-reasoning approaches condition embeddings on generated rationales (Cui et al., 2025; Liu et al., 2025a), refine reasoning with retrieval-oriented objectives (Lan et al., 2025; Jiang et al., 2026a;b), or replace CoT with retrieval-friendly rewrites (Wu et al., 2026a). Adaptive methods further select when reasoning is useful (Wang et al., 2026b; Zhang et al., 2026). Latent reasoning reduces textual generation at inference, but several approaches still rely on language-based supervision. PLUME (He et al., 2026) progressively compresses explicit CoT into latent rollouts, while LaME (Wu et al., 2026b) uses auxiliary textual decoding alongside contrastive supervision. DiLRE instead supplies discriminative information directly to latent reasoning, without relying on the generation or compression of explicit CoT or on auxiliary textual decoding to convey that information.

2.2 Latent Space Reasoning

Latent space reasoning enables intermediate computation without verbalizing each reasoning step. Pause tokens provide additional computation before generation (Goyal et al., 2024), while Coconut feeds hidden states back as continuous thoughts (Hao et al., 2025). Knowledge distillation transfers explicit reasoning into hidden representations (Deng et al., 2023; Shen et al., 2025), and SoftCoT supplies soft thought tokens through an auxiliary model (Xu et al., 2025). Beyond training, LTPO optimizes latent thoughts at test time using confidence-based rewards for answer generation (Ye et al., 2025). DiLRE instead targets embedding quality, using discriminative guidance to align the retrieval distribution induced by latent reasoning with relevance judgments, rather than optimizing text generation.

3 Method

DiLRE is organized around the provision of discriminative signals and their internalization into latent reasoning (Figure 2). A Discriminative Guidance Model provides relevance judgments that complement embedding similarity. Contrastive pretraining establishes a discriminative anchor for the latent reasoning tokens, which are subsequently refined using relevance guidance and current retrieval preferences. This design allows discriminative information to enter the reasoning process directly, without being expressed as explicit CoT.

3.1 Preliminaries

Universal multimodal embedding. Given a query q and a candidate set $\mathcal{T}(q)$, the goal is to retrieve a matching target t^+ through similarity in a shared embedding space (Lan et al., 2025; Wu et al., 2026b). Queries and targets can contain text, images, videos, or visual documents, with the query instruction specifying the required matching relation. Let $f_\theta(x) \in \mathbb{R}^d$ denote a normalized embedding produced by a multimodal encoder. The similarity is $s(q,t) = f_\theta(q)^\top f_\theta(t)$. For a minibatch of N matched query–target pairs $\{(q_i, t_i^+)\}_{i=1}^N$, the standard InfoNCE objective is

$$\mathcal{L}_{\mathrm{NCE}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\big(f_\theta(q_i)^\top f_\theta(t_i^+)/\tau_{\mathrm{ret}}\big)}{\sum_{j=1}^{N} \exp\big(f_\theta(q_i)^\top f_\theta(t_j^+)/\tau_{\mathrm{ret}}\big)}, \tag{1}$$

where $\tau_{\mathrm{ret}} > 0$ is the temperature and the other targets in the minibatch serve as negatives.

Latent reasoning tokens. Following prior work on reasoning with special tokens (Goyal et al., 2024; Xu et al., 2025), we append K learnable special tokens to the multimodal input, followed by an embedding token:

$$[x; \texttt{[R}_1\texttt{]}; \ldots; \texttt{[R}_K\texttt{]}; \texttt{[EMBED]}]. \tag{2}$$

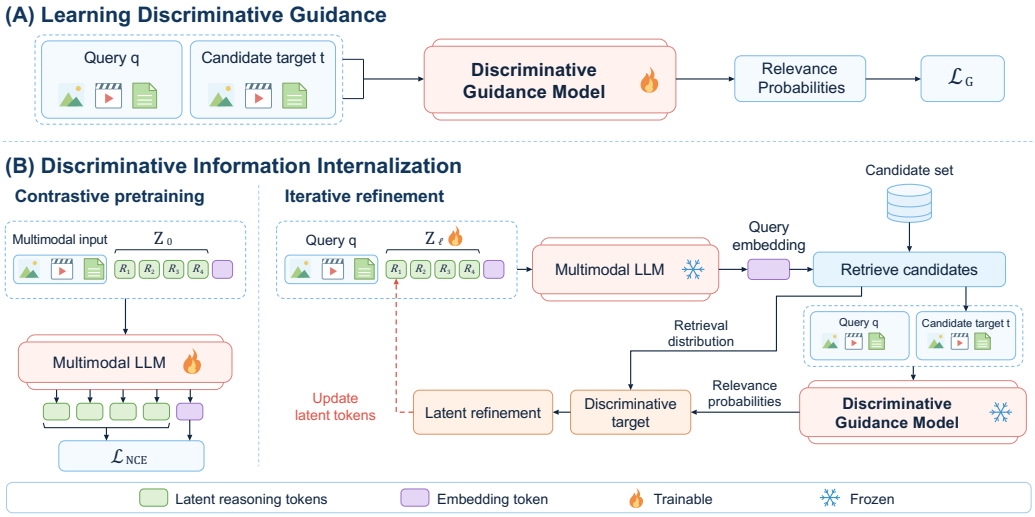


Figure 2: **Overview of DiLRE.** (A) The Discriminative Guidance Model learns to jointly assess query–target relevance. (B) Contrastive pretraining establishes a discriminative anchor for latent reasoning tokens. At test time, the retrieval distribution and the model's relevance probabilities form a discriminative target for iterative latent refinement. Only the input embeddings of the latent reasoning tokens are updated; all model parameters and candidate embeddings remain frozen. The refined latent reasoning produces a single retrieval embedding without explicit CoT.

Their hidden states support intermediate computation without textual decoding. In a single forward pass under causal attention, each reasoning token attends to the input and preceding reasoning tokens, while the embedding token attends to both. Let $Z \in \mathbb{R}^{K\times h}$ denote the input embeddings of the latent reasoning tokens, where h is the hidden dimension.

3.2 Learning Discriminative Guidance

To provide latent reasoning with discriminative information beyond embedding similarity, we learn to assess query–target relevance jointly. The *Discriminative Guidance Model* $g_\phi(q,t)$ takes two inputs: a query q with its task instruction and a candidate target t. It outputs a scalar relevance logit for the pair.

We train the model on positive and negative targets using two complementary objectives: binary cross-entropy for relevance prediction and a pairwise ranking loss for improving retrieval preferences:

$$\mathcal{L}_G = \mathcal{L}_{\text{BCE}} + \lambda_V \mathcal{L}_{\text{rank}}, \tag{3}$$

where λ_V balances the two objectives. For query–target pairs with binary relevance labels $y \in \{0,1\}$, the first objective is

$$\mathcal{L}_{\text{BCE}} = -\mathbb{E}_{(q,t,y)}[y \log \sigma(g_\phi(q,t)) + (1-y)\log(1-\sigma(g_\phi(q,t)))], \tag{4}$$

where σ is the sigmoid function, mapping the model's logits to relevance probabilities.

To learn guidance that complements embedding similarity, the ranking objective encourages positive targets t^+ to outrank negatives t^- under a combined score:

$$\mathcal{L}_{\text{rank}} = -\mathbb{E}_{(q,t^+,t^-)}\left[\log \sigma\big(S_\phi(q,t^+) - S_\phi(q,t^-)\big)\right], \tag{5}$$

where the combined score incorporates both retrieval similarity and the guidance model's relevance prediction:

$$S_\phi(q,t) = \frac{s(q,t)}{\tau_V} + \log \sigma(g_\phi(q,t)), \tag{6}$$

with $\tau_V > 0$ scaling the retrieval similarity. The ranking loss penalizes weak or incorrect combined preferences more strongly than confident, correct ones, encouraging the guidance model to provide complementary discriminative signals for latent refinement.

3.3 Discriminative Information Internalization

Our aim is to internalize this discriminative information directly within latent reasoning, without relying on explicit CoT to convey it. During training, contrastive supervision establishes a discriminative anchor for the latent reasoning tokens. At test time, we internalize feedback from the Discriminative Guidance Model through an iterative loop of retrieval and latent refinement to improve the resulting embedding.

Contrastive pretraining. We train the embedding backbone with the retrieval objective in Equation 1 on the final embedding, together with the same contrastive objective applied to the mean-pooled hidden states of the latent reasoning tokens. The pretrained input embeddings of the latent reasoning tokens, denoted by Z_0, serve as a discriminative anchor for subsequent refinement.

Constructing discriminative targets. To supervise latent refinement, we construct a target distribution that combines the current embedding's retrieval preferences with complementary judgments from the Discriminative Guidance Model. At round ℓ, we obtain the query embedding $e_\ell = f_\theta(q; Z_\ell)$ using Z_ℓ as the input embeddings of the latent reasoning tokens. It retrieves up to M candidates $C_\ell \subseteq \mathcal{T}(q)$. With fixed candidate embeddings $v_t = f_\theta(t; Z_0)$, the retrieval preferences are expressed as

$$\pi_\ell(t) = \frac{\exp(e_\ell^\top v_t/\tau)}{\sum_{u\in C_\ell} \exp(e_\ell^\top v_u/\tau)}, \qquad t \in C_\ell, \tag{7}$$

where $\tau > 0$ is the retrieval temperature. The relevance probability is obtained from the guidance model's logit as $r_\phi(q,t) = \sigma(g_\phi(q,t))$. These judgments adjust the current preferences to form the discriminative target:

$$p_\ell(t) = \frac{\pi_\ell(t)\, r_\phi(q,t)}{\sum_{u\in C_\ell} \pi_\ell(u)\, r_\phi(q,u)}, \qquad t \in C_\ell. \tag{8}$$

The target specifies relative preferences for the embedding to learn, without requiring a textual explanation of the underlying distinctions.

Internalization through latent refinement. To internalize discriminative guidance, we refine the latent reasoning tokens so that the resulting embedding's retrieval distribution approaches p_ℓ, the target combining the guidance model's judgments with current retrieval preferences. Let $\pi_\ell(t; Z)$ denote the retrieval probability in Equation 7 with e_ℓ replaced by $f_\theta(q; Z)$ and C_ℓ held fixed. All encoder and guidance model parameters, as well as candidate embeddings, remain frozen. We optimize only Z, the input embeddings of the latent reasoning tokens, by minimizing the cross-entropy with the fixed target:

$$\mathcal{L}_{\text{int}}^{(\ell)}(Z) = -\sum_{t\in C_\ell} p_\ell(t) \log \pi_\ell(t; Z). \tag{9}$$

We apply stop-gradient to p_ℓ, treating it as a constant during each update. We optimize Z to align the resulting embedding's retrieval distribution $\pi_\ell(\cdot; Z)$ with the discriminative target p_ℓ.

Iterative refinement. Starting from Z_0, we repeat target construction and latent refinement for L rounds. At round $\ell = 0, \ldots, L-1$, we take one Adam step from Z_ℓ on $\mathcal{L}_{\text{int}}^{(\ell)}$. The updated token embeddings are projected to a bounded neighborhood of Z_0, yielding $Z_{\ell+1}$. The updated embedding $e_{\ell+1} = f_\theta(q; Z_{\ell+1})$ is used to retrieve candidates and reconstruct the discriminative target for the next round. After L rounds, e_L ranks the full candidate set $\mathcal{T}(q)$ by similarity to the fixed candidate embeddings.

4 Experiments

We first describe the experimental setup and compare DiLRE with existing multimodal embedding methods across benchmarks and model scales. Ablation studies then examine whether discriminative information provides an effective alternative to explicit CoT supervision for latent reasoning, why guidance is combined with retrieval scores, and whether internalizing this guidance through latent reasoning improves retrieval. Finally, we analyze the effect of additional refinement rounds and the performance–efficiency trade-off.

Table 1: **Results on MMEB-V2.** DiLRE leads in overall performance at both model scales, with the highest image and visual-document averages.

Model	Image					Video					VisDoc					All
	CLS	QA	RET	GD	**Avg.**	CLS	QA	RET	MRET	**Avg.**	VDRv1	VDRv2	VR	OOD	**Avg.**	
# Tasks	10	10	12	4	36	5	5	5	3	18	10	4	6	4	24	78
∼ 2B Model Size																
VLM2Vec	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	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
GME	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
LamRA	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
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
UME-R1	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	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
ReLoop-UME	67.4	63.5	69.8	83.9	68.9	44.2	46.3	33.6	36.5	40.5	76.9	46.5	83.5	65.3	71.5	63.2
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
DiLRE (Ours)	**69.4**	**69.1**	**71.3**	**84.8**	**71.7**	**55.1**	**56.4**	41.5	40.7	**49.3**	81.1	53.3	**85.5**	**65.5**	**75.0**	**67.5**
∼ 7B Model Size																
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
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
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
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
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
ReLoop-UME	69.5	67.5	72.5	**87.8**	72.0	44.5	52.0	34.3	39.4	42.9	78.9	50.6	**85.7**	67.1	73.9	65.9
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
DiLRE (Ours)	**71.1**	**74.0**	**74.9**	**87.8**	**75.0**	54.0	61.3	43.8	**47.1**	52.0	83.0	**59.7**	**85.7**	**68.9**	**77.5**	**70.5**

4.1 Experimental Setup

Implementation Details. We use Qwen2-VL-2B and Qwen2-VL-7B backbones with $K = 4$ latent reasoning tokens. Backbone pretraining applies contrastive supervision to both the final embedding and mean-pooled latent reasoning states, using a learning rate of 10^{-5}, an effective batch size of 512, and $\tau_{\text{ret}} = 0.02$. The Discriminative Guidance Model matches the parameter scale of the embedding backbone and is trained for one epoch with rank-16 LoRA and a learning rate of 2×10^{-5}. For its training objective, we set $\lambda_V = 1$ and $\tau_V = 0.02$. At test time, each round retrieves up to $M = 32$ candidates and takes one Adam step on the input embeddings of the latent reasoning tokens with learning rate 10^{-3}. All model parameters and candidate embeddings remain fixed. We use $\tau = 0.02$ and a relative update radius of 0.2 around Z_0. We select the number of refinement rounds L on the validation set for each task and keep it fixed for all test queries within that task. Additional implementation details are provided in Appendix B.

Benchmarks and metrics. We evaluate DiLRE on MMEB-V1 (Jiang et al., 2024b) and MMEB-V2 (Meng et al., 2025). MMEB-V1 covers 36 image tasks spanning classification, visual question answering, retrieval, and visual grounding. MMEB-V2 adds 18 video and 24 visual-document tasks, yielding 78 tasks across nine meta-tasks. We report Hit@1 for image and video tasks and nDCG@5 (Järvelin & Kekäläinen, 2002) for visual-document tasks.

Baselines. We compare DiLRE with three categories of methods: direct embedding, explicit reasoning, and latent reasoning. Direct embedding baselines include CLIP (Radford et al., 2021), VLM2Vec (Jiang et al., 2024b), VLM2Vec-V2 (Meng et al., 2025), GME (Zhang et al., 2024a), Ops-MM-Embed (OpenSearch-AI, 2025), and ReMatch (Liu et al., 2025b). Explicit reasoning baselines include UME-R1 (Lan et al., 2025), RIME (Wu et al., 2026a), Embed-RL (Jiang et al., 2026a), RGE (Liu et al., 2025a), and Think-When-Needed (Zhang et al., 2026). Latent reasoning baselines include PLUME (He et al., 2026), ReLoop-UME (Wang et al., 2026a), and LaME (Wu et al., 2026b).

Table 2: **Results on MMEB-V1.** DiLRE surpasses the compared direct, explicit-reasoning, and latent-reasoning baselines overall without explicit CoT supervision.

Model	Backbone	Per Meta-Task				Average Score		
		CLS	VQA	RET	GD	IND	OOD	**All**
# Tasks		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	54.4	29.9	66.9	55.5	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	56.3	69.5	77.3	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	72.9	64.5	69.1
Embed-RL	Qwen3-VL-2B	62.8	67.9	68.6	**90.4**	71.9	65.9	69.2
RGE	Qwen3-VL-2B	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**
Latent Reasoning UME								
PLUME	Qwen2-VL-2B	66.5	59.2	67.6	79.7	69.6	62.2	66.3
LaME	Qwen2-VL-2B	67.6	66.2	70.5	81.2	73.1	64.4	69.3
DiLRE (Ours)	Qwen2-VL-2B	**69.4**	**69.1**	**71.3**	**84.8**	**74.0**	**68.9**	**71.7**

4.2 Main Results

Gains across modalities and scales. Table 1 shows that DiLRE achieves overall scores of 67.5 and 70.5 on MMEB-V2 at the 2B and 7B scales, exceeding the strongest baseline in each scale group by 3.1 and 1.6 points, respectively. It obtains the highest image and visual-document averages at both scales, alongside the highest video average at 2B. These results suggest that the benefits extend across different visual modalities and persist with a larger embedding model.

Strong performance without CoT supervision. Table 2 shows that DiLRE achieves the highest overall MMEB-V1 score among the compared methods, outperforming the strongest direct embedding, explicit reasoning, and latent reasoning baselines. Notably, DiLRE achieves higher overall performance with a 2B backbone than strong baselines based on the newer Qwen3-VL family, including a 4B model. These results demonstrate strong retrieval performance without using explicit CoT to convey the information for latent reasoning.

4.3 Ablation Studies

We study three questions concerning discriminative supervision, target construction, and information internalization, using the 2B backbone on MMEB-V2.

Table 3: **Supervision for latent reasoning.** Contrastive supervision outperforms the compared text-decoding objectives across all three modalities.

Variant	Image	Video	VisDoc	All
No latent tokens	67.8	45.4	61.8	60.8
No aux. loss	67.4	44.8	61.0	60.2
Text: UME-R1	67.0	42.2	69.5	62.0
Text: RIME	66.7	42.1	69.9	62.0
Text: Embed-RL	67.7	46.1	67.1	62.5
Contrastive (Ours)	**69.1**	**46.3**	**71.2**	**64.5**

Can discriminative information provide an effective alternative to explicit CoT for supervising latent reasoning? Table 3 compares supervision for latent reasoning before test-time refinement.

Four-token variants use no auxiliary supervision, text decoding toward UME-R1, RIME, or Embed-RL targets (Lan et al., 2025; Wu et al., 2026a; Jiang et al., 2026a), or contrastive supervision on mean-pooled latent reasoning states. All variants retain the final embedding objective. Without auxiliary supervision, latent tokens do not improve over the encoder without them (60.2 vs. 60.8). Contrastive supervision reaches 64.5, exceeding the strongest text-decoding variant at 62.5 and leading across all three modalities. This supports establishing a discriminative anchor through contrastive supervision, without prescribing textual reasoning targets.

Table 4: **Discriminative target construction.** Combining retrieval scores and guidance outperforms either source alone across all three modalities.

Setting	Image	Video	VisDoc	All
Retrieval only	69.1	46.2	71.3	64.5
Guidance only	58.5	38.9	69.2	57.3
Random guidance only	4.4	11.6	11.3	8.2
Retrieval + Guidance	**71.2**	**48.2**	**74.6**	**66.9**

Why combine discriminative guidance with retrieval scores? Table 4 evaluates candidate rankings induced by retrieval scores, relevance guidance, and their combination in the discriminative target p_ℓ, before latent refinement. Guidance alone underperforms retrieval (57.3 vs. 64.5), whereas their combination reaches 66.9 and improves all three modality averages. These results support combining the retrieval distribution with complementary relevance guidance when constructing targets for latent refinement, rather than relying on the guidance model alone.

Does internalizing discriminative information through latent reasoning improve retrieval? Table 5 compares using the discriminative target for static reranking, directly updating the final embedding, and refining the input embeddings of latent reasoning tokens. Latent refinement raises the overall score from 64.5 to 67.5, surpassing static reranking at 66.9 and direct embedding updates at 66.5 across all three modalities. The gains support internalizing guidance through latent reasoning to align the resulting embedding's retrieval distribution with the discriminative target. Removing contrastive supervision on latent reasoning states reduces the refined score to 64.8, further supporting the role of the pretrained discriminative anchor.

Table 5: **Internalizing discriminative information.** Refining latent reasoning tokens yields greater retrieval gains than reranking candidates or directly updating the final embedding.

Variant	Output	Image36	Video18	VisDoc24	All
Unrefined encoder	Embedding	69.1	46.2	71.3	64.5
Static reranking	Ranking	71.2	48.2	74.6	66.9
Direct embedding update	Embedding	70.9	48.1	73.7	66.5
Latent refinement (Ours)	Embedding	**71.7**	**49.3**	**75.0**	**67.5**
w/o latent contrastive supervision	Embedding	68.9	45.1	73.5	64.8

4.4 Further Analysis

Do additional refinement rounds improve retrieval performance? Figure 3 compares fixed refinement rounds on MMEB-V2. Each round refreshes the discriminative target and updates the latent reasoning tokens. Relative to no refinement, the overall gain increases from 1.2 points after one round to 1.6 points after two rounds. The second round improves all three modality averages. Image and video performance then declines, whereas visual-document gains largely plateau at three to four rounds. These trends show that additional refinement can help, but its benefits depend on the number of rounds and modality. The oracle curves provide retrospective upper bounds, suggesting that more precise selection of refinement rounds could further improve retrieval.

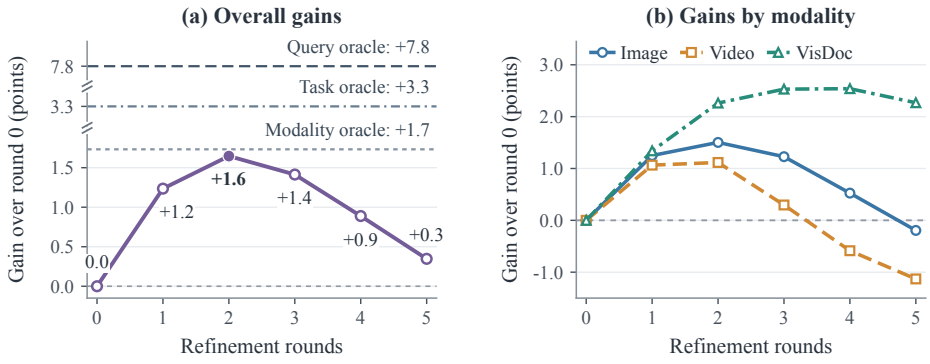


Figure 3: **Effect of refinement rounds on MMEB-V2 (2B).** Retrieval gains peak after a few rounds, with image and video tasks favoring fewer rounds than visual-document tasks.

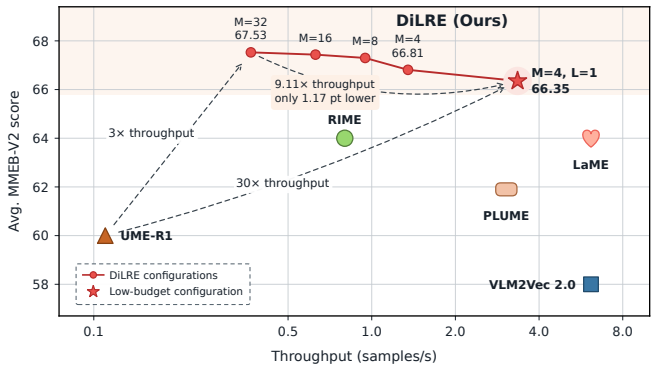


Figure 4: **Performance–efficiency trade-off.** DiLRE retains most of its retrieval gains with fewer guidance evaluations and higher refinement throughput.

What performance–efficiency trade-off does DiLRE offer? Figure 4 shows that fewer guidance evaluations retain most of DiLRE's retrieval gains while improving throughput. Reducing the number of candidates from $M = 32$ to $M = 4$ on a single H20 GPU with one process and batch size 1 yields 3.3× higher throughput, with a 0.7-point decrease in the overall MMEB-V2 score (67.5 to 66.8). At $M = 4$, latent refinement further improves over one refinement update (66.8 vs. 66.4), indicating that internalizing guidance remains beneficial even with fewer candidates. Overall, DiLRE attains state-of-the-art retrieval performance with throughput at the level of latent-reasoning methods, substantially outperforming explicit-reasoning methods in both effectiveness and efficiency, while the candidate budget M and refinement rounds L provide flexible control over the quality–cost trade-off.

5 Conclusion

We introduced DiLRE to bring discriminative information into latent reasoning for universal multimodal embedding without relying on explicit CoT. Contrastive pretraining establishes a discriminative anchor for the latent reasoning tokens, while a Discriminative Guidance Model provides relevance judgments that complement embedding similarity. At test time, DiLRE combines these judgments with the current retrieval preferences to construct a discriminative target and iteratively refines the query's latent reasoning token embeddings. The resulting single embedding ranks the full candidate set, with model parameters and candidate embeddings fixed throughout refinement. Experiments across image, video, and visual-document tasks at both 2B and 7B scales, together with ablations, support the value of complementary relevance guidance and its internalization through latent refinement.

AI USE STATEMENT

Generative AI tools were used to aid in polishing the writing of this paper and, under the authors' supervision, to assist with coding during implementation of the method. They were not used to develop theoretical models or conceptual frameworks; formulate mathematical claims; provide critical ingredients for proving mathematical claims; assist in writing mathematical proofs; propose or refine hypotheses; or design research methodology or experiments.

REFERENCES

Xuanming Cui, Jianpeng Cheng, Hong-you Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Kumar Mishra, Yonghuan Yang, Jun Xiao, Qi Guo, Ser-Nam Lim, Aashu Singh, and Xiangjun Fan. Think then embed: Generative context improves multimodal embedding. *arXiv preprint arXiv:2510.05014*, 2025. URL https://arxiv.org/abs/2510.05014v4.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023. URL https://arxiv.org/abs/2311.01460v1.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ph04CRkPdC.

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. In *Second Conference on Language Modeling*, 2025. URL https://arxiv.org/abs/2412.06769v4.

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. URL https://arxiv.org/abs/2604.02073v3.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418. URL https://doi.org/10.1145/582415.582418.

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. *arXiv preprint arXiv:2602.13823*, 2026a. URL https://arxiv.org/abs/2602.13823v3.

Mingzhou Jiang, Peixi Wu, Hang Cheng, Yunhao Zhou, Biao Yang, Wei Yuan, Yun Li, Fan Yang, Wenwu Ou, and Honghui He. Reason what matters: Retrieval-grounded reasoning for universal multimodal embeddings. *arXiv preprint arXiv:2609.15296*, 2026b. URL https://arxiv.org/abs/2609.15296.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-V: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*, 2024a. URL https://arxiv.org/abs/2407.12580v1.

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*, 2024b. URL https://arxiv.org/abs/2410.05160v3.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. UME-R1: Exploring reasoning-driven generative multimodal embeddings. *arXiv preprint arXiv:2511.00405*, 2025. URL https://arxiv.org/abs/2511.00405v2.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-Embed: Universal multimodal retrieval with multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.02571v2.

Chunxu Liu, Jiyuan Yang, Ruopeng Gao, Yuhan Zhu, Feng Zhu, Rui Zhao, and Limin Wang. Reasoning guided embeddings: Leveraging MLLM reasoning for improved multimodal retrieval. *arXiv preprint arXiv:2511.16150*, 2025a. URL https://arxiv.org/abs/2511.16150v1.

Qianying Liu, Xiao Liang, Zhiqiang Zhang, Zhongfei Qing, Fengfan Zhou, Yibo Chen, Xu Tang, Yao Hu, and Paul Henderson. ReMatch: Boosting representation through matching for multimodal retrieval. *arXiv preprint arXiv:2511.19278*, 2025b. URL https://arxiv.org/abs/2511.19278v2.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2Vec-V2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025. URL https://arxiv.org/abs/2507.04590v1.

OpenSearch-AI. Ops-MM-embedding-v1. Hugging Face model card, 2025. URL https://huggingface.co/OpenSearch-AI/Ops-MM-embedding-v1-2B.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.36. URL https://aclanthology.org/2025.emnlp-main.36/.

Shijie Wang, Xiangzhao Hao, Yueti Li, Guangyu Cao, Xinyu Tang, and Haiyun Guo. ReLoop-UME: Recurrent depth with learnable retrieval registers for universal multimodal embedding. *arXiv preprint arXiv:2607.28751*, 2026a. URL https://arxiv.org/abs/2607.28751v1.

Yuchi Wang, Dingkang Yang, Haiyang Yu, Weikang Bian, Jiefeng Long, Xiao Liang, Chao Feng, and Hongsheng Li. MMEmb-R1: Reasoning-enhanced multimodal embedding with pair-aware selection and adaptive control. *arXiv preprint arXiv:2604.06156*, 2026b. URL https://arxiv.org/abs/2604.06156v2.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1800. URL https://arxiv.org/abs/2201.11903.

Peixi Wu, Ke Mei, Feipeng Ma, Bosong Chai, Zhibin Lan, Chenxi Zhao, Shannan Yan, Jie Chen, Zhangchi Hu, Yansong Peng, Bo Lin, Junjie Zhou, Dacheng Yin, Tianyi Wang, Fengyun Rao, Jing Lyu, Hebei Li, and Xiaoyan Sun. Beyond chain-of-thought: Rewrite as a universal interface for generative multimodal embeddings. *arXiv preprint arXiv:2604.22280*, 2026a. URL https://arxiv.org/abs/2604.22280v1.

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. URL https://arxiv.org/abs/2606.13061v3.

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. Association

for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1137. URL https://aclanthology.org/2025.acl-long.1137/.

Wengao Ye, Yan Liang, and Lianlei Shan. Thinking on the fly: Test-time reasoning enhancement via latent thought policy optimization. *arXiv preprint arXiv:2510.04182*, 2025. URL https://arxiv.org/abs/2510.04182v4.

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. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

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. URL https://arxiv.org/abs/2605.14448v1.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. GME: Improving universal multimodal retrieval by multimodal LLMs. *arXiv preprint arXiv:2412.16855*, 2024a. URL https://arxiv.org/abs/2412.16855v2.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024b. URL https://openreview.net/forum?id=y1pPWFVfvR.

A Training Data

A.1 Embedding Model Training Data

The embedding model in DiLRE is contrastively pretrained on 1,684,211 query–positive pairs drawn from the multimodal training corpus described by Embed-RL (Jiang et al., 2026a). The corpus spans the image, video, and visual-document modalities covered by MMEB-V2 (Meng et al., 2025). Image tasks contribute 720,844 pairs across classification, visual question answering, and retrieval; LLaVA-Hound contributes 819,137 pairs across video caption retrieval, question answering, and retrieval; and ViDoRe and VisRAG contribute 144,230 visual-document pairs. We use the paired inputs for contrastive learning without supervising the model with textual CoT annotations. Table 6 reports the number of pairs and query–positive modalities for each dataset or task subset.

Table 6: **Training data composition.** Per-dataset counts and shares of the 1,684,211 query–positive pairs used by DiLRE.

Dataset or task subset	Pairs	Share	Query → positive
Image-based (MMEB-train)			
A-OKVQA	37,929	2.25%	Text+image → text
ChartQA	39,512	2.35%	Text+image → text
CIRR	35,085	2.08%	Text+image → text+image
DocVQA	47,401	2.81%	Text+image → text
HatefulMemes	16,572	0.98%	Text+image → text
ImageNet-1K	44,409	2.64%	Text+image → text
InfographicsVQA	40,746	2.42%	Text+image → text
MSCOCO	26,429	1.57%	Text+image → text+image
MSCOCO-i2t	46,596	2.77%	Text+image → text
MSCOCO-t2i	43,173	2.56%	Text → text+image
N24News	30,320	1.80%	Text+image → text
NIGHTS	43,167	2.56%	Text+image → text+image
OK-VQA	19,900	1.18%	Text+image → text
SUN397	45,864	2.72%	Text+image → text
VisDial	34,652	2.06%	Text → text+image
Visual7W	41,677	2.47%	Text+image → text
VisualNews-i2t	34,364	2.04%	Text+image → text
VisualNews-t2i	28,684	1.70%	Text → text+image
VOC2007	20,454	1.21%	Text+image → text
WebQA	43,910	2.61%	Text → text+image
Video-based (LLaVA-Hound)			
Caption Retrieval	284,721	16.91%	Text+video → text
Video QA	274,006	16.27%	Text+video → text
Video Retrieval	260,410	15.46%	Text → text+video
Visual-document			
ViDoRe	83,964	4.99%	Text+image → text
VisRAG	60,266	3.58%	Text → image
Image-based	720,844	42.80%	
Video-based	819,137	48.64%	
Visual-document	144,230	8.56%	
Total	**1,684,211**	**100.00%**	

A.2 GUIDANCE MODEL TRAINING DATA

We derive relevance supervision for the Discriminative Guidance Model from the same multimodal corpus as in Section A.1. Within each task, repeated query records are consolidated and their annotated positive targets are collected. An internal validation partition contains 167,294 task-specific query identities, leaving 1,367,682 for training. For each training comparison, an annotated positive is paired with a candidate from the corresponding task's candidate pool. Known positives and equivalent targets are excluded from the comparison candidates. Each comparison contributes one positive and one negative query–candidate relevance example; the model receives the task instruction, query, and a single candidate as input.

Across the 25 task subsets, this yields 844,480 comparisons, or 1,688,960 labeled query–candidate examples with equal numbers of positive and negative labels. Table 7 gives the per-task composition of the training comparisons; the internal validation partition is excluded from these counts. The supervision uses relevance labels rather than textual reasoning annotations.

Table 7: **Guidance model training data.** Distribution of 844,480 relevance comparisons across the 25 task subsets.

Dataset or task subset	**Comparisons**	**Share**
Image-based (MMEB-train)		
A-OKVQA	12,217	1.45%
ChartQA	8,214	0.97%
CIRR	29,916	3.54%
DocVQA	5,619	0.67%
HatefulMemes	806	0.10%
ImageNet-1K	10,513	1.24%
InfographicsVQA	7,808	0.92%
MSCOCO	18,663	2.21%
MSCOCO-i2t	41,936	4.97%
MSCOCO-t2i	38,854	4.60%
N24News	2,674	0.32%
NIGHTS	25,383	3.01%
OK-VQA	5,608	0.66%
SUN397	7,656	0.91%
VisDial	31,187	3.69%
Visual7W	37,429	4.43%
VisualNews-i2t	30,928	3.66%
VisualNews-t2i	24,038	2.85%
VOC2007	1,528	0.18%
WebQA	26,368	3.12%
Video-based (LLaVA-Hound)		
Caption Retrieval	129,064	15.28%
Video QA	154,982	18.35%
Video Retrieval	119,751	14.18%
Visual-document		
ViDoRe	24,452	2.90%
VisRAG	48,886	5.79%
Image-based	367,345	43.50%
Video-based	403,797	47.82%
Visual-document	73,338	8.68%
Total	**844,480**	**100.00%**

B Additional Method Details

We describe how DiLRE trains the representations used for retrieval and discriminative guidance, followed by the query-local refinement procedure in Section 3. Algorithm 1 summarizes inference in the notation of the main paper.

B.1 Embedding Backbone and Latent Reasoning

The embedding backbone is based on Qwen2-VL at the 2B or 7B scale. We append $K = 4$ latent reasoning tokens and a final `[EMBED]` token after the multimodal input, in the causal order of Equation 2. The four reasoning-token states can integrate information from the input and preceding reasoning tokens; the final token can additionally attend to all four states. Its normalized hidden state is the single embedding $f_\theta(x)$ used for retrieval. This path does not decode textual reasoning at inference.

Contrastive pretraining updates the embedding backbone, including the latent reasoning token embeddings. The final embedding follows the retrieval objective in Equation 1. We also mean-pool the four reasoning-token hidden states, normalize the pooled representation, and apply the same form of contrastive objective to it. Both objectives use the query–positive pairs in Appendix A.1; no textual CoT annotations are required. The learned token input embeddings Z_0 provide the starting point for query refinement. Candidate targets are encoded once with this pretrained state, giving fixed vectors $v_t = f_\theta(t; Z_0)$.

B.2 Discriminative Guidance

The Discriminative Guidance Model is a single-sequence multimodal relevance scorer. DiLRE-2B uses a 2B guidance backbone and DiLRE-7B uses a 7B guidance backbone, matching the scale of the respective Qwen2-VL embedding backbone. The guidance and embedding models are trained separately and do not share parameters. Its task-conditioned prompt presents the query q and one candidate t together, including their text and available visual inputs (Appendix C). The multimodal transformer jointly processes this pair rather than producing separate query and candidate embeddings. The hidden state at the last non-padding position is passed through LayerNorm, training-time dropout, and a scalar linear head to produce the relevance logit $g_\phi(q, t)$; no text is decoded.

The pretrained visual encoder and language backbone weights remain frozen at each scale. A rank-16 language LoRA adapter is trained on the query, key, value, and output attention projections and the gate, up, and down feed-forward projections, together with the normalization and scoring head. Within each scale, the same adapter and head serve all tasks. For each training comparison, positive and negative candidates are scored in separate forward passes through the corresponding guidance model; their labels and logits enter the binary relevance and pairwise ranking objectives in Equations 3–6. The supervision is detailed in Appendix A.2. During latent refinement, the fixed guidance model scores the current retrieved candidates.

B.3 Test-Time Latent Refinement

At round ℓ, the current query embedding e_ℓ retrieves the top $M = 32$ eligible targets from the full task-specific candidate set $\mathcal{T}(q)$. The resulting set C_ℓ is used only to construct the round's discriminative target and optimization objective; the final embedding ranks the full candidate set. Repeated query–target pairs can reuse their guidance scores across rounds. The retrieval distribution π_ℓ and relevance probabilities $r_\phi(q, t)$ are combined as in Equation 8 to obtain p_ℓ.

For a given query, write $Z_\ell = Z_0 + \Delta_\ell$, with $\Delta_0 = 0$. Each round takes one Adam step on $\mathcal{L}_{\text{int}}^{(\ell)}$ in Equation 9, holding both C_ℓ and the stop-gradient target p_ℓ fixed within that step. The updated input embeddings are projected onto $\{Z : \|Z - Z_0\|_F \le 0.2\|Z_0\|_F\}$. The optimizer state and the latent update are local to the query. After the step, we recompute $e_{\ell+1}$, retrieve candidates again from $\mathcal{T}(q)$, and construct the next target. No encoder parameter, guidance parameter, or candidate embedding is updated at test time.

Algorithm 1 DiLRE inference for a query q in a task with fixed round count L.

Require: Query q; eligible targets $\mathcal{T}(q)$; pretrained Z_0; fixed f_θ, g_ϕ, and $\{v_t\}_{t\in\mathcal{T}(q)}$; M, L, τ

Initialize the query-local Adam state at Z_0
for $\ell = 0, \ldots, L-1$ **do**
  $e_\ell \leftarrow f_\theta(q; Z_\ell)$
  Retrieve C_ℓ, the top M targets in $\mathcal{T}(q)$ by $e_\ell^\top v_t$
  Obtain $r_\phi(q,t)$ for $t \in C_\ell$ from guidance logits
  Form π_ℓ on C_ℓ by Equation 7
  Form p_ℓ by Equation 8 and stop its gradient
  $\widetilde{Z} \leftarrow \mathrm{AdamStep}\Big(Z_\ell, \nabla_Z \mathcal{L}_{\mathrm{int}}^{(\ell)}(Z)|_{Z=Z_\ell}\Big)$
  $Z_{\ell+1} \leftarrow \Pi_{\{Z:\|Z-Z_0\|_F \le 0.2\|Z_0\|_F\}}(\widetilde{Z})$
end for
$e_L \leftarrow f_\theta(q; Z_L)$
return targets in $\mathcal{T}(q)$ ranked by $e_L^\top v_t$

B.4 Selecting the Number of Refinement Rounds

As described in Section 4.1, the number of rounds L is chosen at the task level using training-side validation examples and is then held fixed for every test query in that task. We use a 10% subset of each available training task for this decision. Where a benchmark task lacks a corresponding training task, the decision is informed by a preassigned source group matched by modality, retrieval direction where available, and query–target structure. The resulting fixed value of L for each task is reported in Table 8. The round-specific analysis in Figure 3 evaluates fixed horizons and reports retrospective oracle references separately.

Table 8: **Task-level refinement rounds.** The validation-selected horizon L is fixed for all test queries in each MMEB-V2 task. Read down the left column, then down the right column; the average is the unweighted mean across all 78 tasks. ViDoRe-SynDoc abbreviates the SyntheticDocQA domains.

No.	Evaluation task	L
Image tasks (36)		
1	ImageNet-1K	1
2	N24News	1
3	HatefulMemes	1
4	VOC2007	1
5	SUN397	1
6	Place365	2
7	ImageNet-A	2
8	ImageNet-R	2
9	ObjectNet	5
10	Country211	5
11	OK-VQA	1
12	A-OKVQA	1
13	DocVQA	1
14	InfographicsVQA	1
15	ChartQA	1
16	Visual7W	2
17	ScienceQA	2
18	VizWiz	1
19	GQA	1
20	TextVQA	1
21	MSCOCO-i2t	2
22	VisualNews-i2t	1
23	VisDial	1
24	MSCOCO-t2i	1
25	VisualNews-t2i	1
26	WebQA	1
27	EDIS	1
28	Wiki-SS-NQ	2
29	CIRR	2
30	NIGHTS	2
31	OVEN	2
32	FashionIQ	2
33	MSCOCO	1
34	RefCOCO	1
35	RefCOCO-Matching	1
36	Visual7W-Pointing	3
Video tasks (18)		
37	SmthSmthV2	3
38	HMDB51	2
39	UCF101	1

No.	Evaluation task	L
Video tasks (continued)		
40	K700	2
41	Breakfast	4
42	MSR-VTT	1
43	MSVD	1
44	DiDeMo	1
45	YouCook2	3
46	VATEX	1
47	QVHighlight	1
48	Charades-STA	1
49	MomentSeeker	2
50	Video-MME	1
51	NExTQA	1
52	EgoSchema	1
53	MVBench	1
54	ActivityNetQA	1
Visual-document tasks (24)		
55	ViDoRe-arxivqa	1
56	ViDoRe-docvqa	4
57	ViDoRe-infovqa	1
58	ViDoRe-tabfquad	1
59	ViDoRe-tatdqa	1
60	ViDoRe-shiftproject	2
61	ViDoRe-SynDoc-artificial-intelligence	2
62	ViDoRe-SynDoc-energy	1
63	ViDoRe-SynDoc-government-reports	4
64	ViDoRe-SynDoc-healthcare-industry	1
65	ViDoRe-esg-reports-human-labeled-v2	1
66	ViDoRe-biomedical-lectures-v2-multilingual	5
67	ViDoRe-economics-reports-v2-multilingual	1
68	ViDoRe-esg-reports-v2-multilingual	1
69	VisRAG-ArxivQA	1
70	VisRAG-ChartQA	1
71	VisRAG-MP-DocVQA	5
72	VisRAG-SlideVQA	4
73	VisRAG-InfoVQA	1
74	VisRAG-PlotQA	1
75	ViDoSeek-page	1
76	ViDoSeek-doc	1
77	MMLongBench-doc	2
78	MMLongBench-page	1

Average across 78 tasks	**1.65**

C Prompt Design for Discriminative Guidance

The Discriminative Guidance Model evaluates one query–candidate pair at a time, producing the scalar relevance logit $g_\phi(q, t)$ defined in Section 3.2. Its prompt identifies the retrieval task, presents the query and one candidate in their original modalities, and states the relation that makes the candidate relevant. This common format lets the model judge different target types—an answer, image, video, or document page—without changing the meaning of its output. The resulting candidate judgments contribute to the discriminative target in Equation 8.

C.1 Shared Prompt

The template below shows the common instruction and the fields populated for each task. The *relevance criterion* specifies the required query–target relation; it directs the judgment toward the requested answer or visual evidence rather than general semantic association. Braces denote substituted task content, and bracketed media markers denote the corresponding image or video inputs.

Shared guidance prompt

Instruction. Determine whether the candidate is a correct retrieval target for the query under the stated task. Compare the query and candidate evidence. When the task requires an exact answer, a specified edit, or the same visual instance, give those conditions priority over generic association.

Task {task name}
Relevance criterion {task-specific matching rule}
Query {task instruction, text, and available media}
Candidate {candidate text or media}

Judgment. Estimate this candidate's relevance under the stated criterion. Prefer correctness and complete satisfaction of the query constraints over broad similarity.

Input construction. The task instruction accompanies the query; content-bearing text such as a question, edit request, or answer options remains visible to the guidance model. Images, sampled video frames, and document pages are provided as media inputs, rather than replaced with textual descriptions. The candidate field contains the object actually being judged: answer text for answer selection, an image or video for visual retrieval, and a page image for document retrieval. Candidate titles or descriptions are included only when they are part of the task's target. Evaluation labels and relevance judgments are never supplied in these prompts.

C.2 Representative Instantiations

The following abbreviated instances illustrate how the same prompt changes with the task. Each box supplies the query and candidate fields and the task-specific criterion; the shared instruction above is omitted for space.

ImageNet-1K — **Image classification**

Query [image] and *Represent the given image for classification.*
Candidate {class label}.
Relevance criterion The label identifies the visual class depicted in the image.

A-OKVQA — **Visual question answering**

Query [image] and {question}.
Candidate {answer text}.
Relevance criterion The answer resolves the question using the image and, when needed, the knowledge required by the question.

CIRR — **Composed image retrieval**

Query [source image] and {edit instruction}.
Candidate [candidate image].
Relevance criterion The candidate preserves the relevant source content while satisfying the specified change.

NIGHTS — **Image-to-image retrieval**

Query [reference image].
Candidate [candidate image].
Relevance criterion The candidate resembles the reference in the everyday visual details relevant to the task.

Video-MME — **Video question answering**

Query [video frames], {question}, and {answer options}.
Candidate {answer text}.
Relevance criterion The answer is supported by the video's visual and temporal evidence, including the order of events when the question depends on it.

MSR-VTT — **Text-to-video retrieval**

Query *Find a video that contains the following visual content:* {caption}.
Candidate [candidate video frames].
Relevance criterion The video depicts the visual content and events described by the caption.

VisRAG — **Visual-document retrieval**

Query *Find a document image that matches the given query.* {*query*}
Candidate [page image].
Relevance criterion The page provides evidence for the query in its text, table, chart, or visual layout.

These instances illustrate distinct kinds of evidence under one pairwise interface. Classification and question answering compare an image or video with a textual target; composed retrieval requires the candidate image to satisfy an explicit modification; image similarity can be judged directly from two visual inputs; and video or document retrieval judges the candidate media itself. Thus, the task criterion expresses the intended relation while the query and candidate fields retain the evidence needed to assess it.

For every instance, the guidance model returns a score for the pair, not a generated answer or explanation. The relevance criterion changes what evidence the model should attend to; the resulting logit retains the same role in the discriminative target of Equation 8.

D Additional Results

D.1 Detailed MMEB-V2 Results

Table 9 reports the full 78-task breakdown for DiLRE-2B and DiLRE-7B together with the published baselines in UME-R1 (Lan et al., 2025). Image and video tasks use Hit@1; visual-document tasks use linear-gain nDCG@5. The overall, modality, and task-family averages give equal weight to their constituent tasks, as in Table 1. Baseline values retain their published one-decimal precision, and both DiLRE columns are shown to one decimal.

Table 9: **Fine-grained MMEB-V2 results.** The nine baseline columns are drawn from Table 5 of UME-R1 (Lan et al., 2025); the shaded DiLRE columns report our results. Numbers in parentheses indicate task counts.

Task	ColP. 1.3	LamRA 7B	V2Vec 7B	V2Vec V2	CAFe 7B	DUME 2B	DUME 7B	UME 2B	UME 7B	DiLRE 2B	DiLRE 7B
Average: all (78)	44.4	47.4	52.3	58.0	60.6	52.7	55.9	60.1	64.5	67.5	70.5
Average: image (36)	34.9	52.4	65.5	64.9	67.6	62.5	66.4	66.6	71.3	71.7	75.0
Average: video (18)	28.2	33.6	33.7	34.6	42.4	33.2	29.4	42.2	47.5	49.3	52.0
Average: VisDoc (24)	71.0	50.2	46.4	65.4	63.9	52.8	60.3	63.9	67.1	75.0	77.5
Image classification (10)	40.3	51.7	62.7	62.9	63.6	59.3	64.2	64.8	67.1	69.4	71.1
Image QA (10)	11.5	34.1	56.9	56.3	61.7	54.9	57.0	62.8	69.2	69.1	74.0
Image retrieval (12)	48.1	66.9	69.4	69.5	69.1	66.3	70.8	67.6	71.9	71.3	74.9
Visual grounding (4)	40.3	56.7	82.2	77.3	87.6	78.0	81.8	77.2	84.9	84.8	87.8
Video classification (5)	26.7	32.9	39.1	39.3	35.8	37.7	32.9	44.3	48.6	55.1	54.0
Video QA (5)	37.8	42.6	30.0	34.3	58.7	46.6	47.4	51.0	60.7	56.4	61.3
Video retrieval (5)	21.6	23.2	29.0	28.8	34.4	17.1	8.6	32.9	38.2	41.5	43.8
Moment retrieval (3)	25.5	37.2	38.9	36.8	39.5	30.0	28.0	39.7	39.3	40.7	47.1
ViDoRe V1 (10)	83.6	56.3	56.9	75.7	70.7	67.6	67.1	72.4	75.7	81.1	83.0
ViDoRe V2 (4)	52.0	33.3	9.4	45.1	49.6	43.3	35.2	46.2	50.5	53.3	59.7
VisRAG (6)	81.1	58.2	59.1	79.6	79.5	47.1	82.6	79.2	83.7	85.5	85.7
VisDoc OOD (4)	43.1	40.1	38.1	39.6	38.1	33.8	34.9	37.2	37.6	65.5	68.9
Image tasks (36; Hit@1)											
ImageNet-1K	42.4	58.9	80.1	80.8	77.3	74.6	76.6	75.3	80.4	81.9	82.3
N24News	25.5	29.8	79.7	72.9	83.2	69.7	77.2	81.1	82.3	77.7	79.5
HatefulMemes	50.6	51.3	69.7	56.3	78.7	65.3	79.6	75.2	79.0	69.2	72.6
VOC2007	69.8	78.7	80.7	85.0	89.8	68.9	85.5	80.0	90.8	90.6	92.5
SUN397	56.1	66.5	77.4	71.0	79.9	71.4	74.6	79.4	80.3	82.6	82.7
Place365	27.5	37.4	37.4	35.9	45.0	41.0	41.9	42.6	46.8	45.6	47.4
ImageNet-A	14.9	36.3	58.1	47.4	55.2	41.3	48.6	50.4	53.9	54.5	61.6
ImageNet-R	64.6	77.0	73.9	89.3	88.0	90.7	88.8	88.7	90.1	91.5	92.1
ObjectNet	45.6	59.4	40.1	65.2	22.5	46.2	44.8	52.0	42.3	73.8	72.8
Country211	6.0	21.7	29.8	25.2	16.7	23.9	24.7	23.4	25.0	26.9	27.0
OK-VQA	9.4	39.9	56.8	51.5	67.3	56.8	61.6	62.4	71.7	68.1	75.8
A-OKVQA	6.6	34.1	47.3	43.6	63.8	46.9	51.4	51.1	58.7	57.9	63.7
DocVQA	11.3	37.1	89.7	90.1	79.2	86.0	86.3	92.2	93.8	94.7	96.0
InfographicsVQA	5.0	23.7	60.0	58.8	53.3	59.2	62.3	67.7	79.2	70.4	80.7
ChartQA	5.7	15.0	56.9	47.4	48.8	39.1	49.8	64.9	75.1	58.3	70.1
Visual7W	6.1	24.6	52.7	52.9	52.5	46.9	52.1	54.1	55.2	60.8	64.0
ScienceQA	16.3	31.3	38.5	38.2	65.4	38.7	45.5	42.7	53.7	54.3	62.5
VizWiz	27.6	32.0	39.9	43.3	43.8	42.0	44.3	46.8	51.6	57.9	59.9
GQA	8.3	57.4	55.1	64.9	65.7	60.2	46.9	67.3	69.3	81.5	79.1
TextVQA	18.8	46.1	71.6	72.2	76.8	73.9	69.9	78.6	83.5	86.7	88.2
VisDial	41.2	62.5	81.9	82.7	82.7	75.9	75.7	76.6	80.7	80.3	83.8
CIRR	8.2	44.7	51.1	57.5	60.4	52.0	51.6	53.7	55.3	60.9	66.3
VisualNews-t2i	50.1	70.1	80.5	74.5	69.5	71.2	76.9	71.7	76.8	72.1	75.9
VisualNews-i2t	47.6	74.2	81.2	78.2	79.4	72.5	82.3	74.2	82.0	78.2	83.1
MSCOCO-t2i	59.2	65.7	77.2	75.3	75.4	74.5	77.1	75.1	78.3	74.1	79.1
MSCOCO-i2t	49.9	71.1	73.9	71.4	73.1	68.3	71.2	68.9	71.4	73.7	75.8
NIGHTS	65.5	64.4	67.6	68.6	66.7	67.5	69.6	67.2	68.1	67.4	68.0
WebQA	53.8	85.7	88.3	90.6	89.3	90.2	90.3	90.0	90.9	88.4	91.5
FashionIQ	5.9	33.4	17.1	19.5	39.0	11.5	20.5	17.1	23.4	25.3	32.2
Wiki-SS-NQ	80.5	67.0	62.3	66.9	61.2	60.0	70.6	62.0	72.5	77.1	79.1
OVEN	50.0	84.8	66.5	64.3	60.8	65.2	70.5	66.9	71.4	65.0	69.6
EDIS	64.7	78.7	85.7	84.1	71.3	86.5	92.8	88.0	92.0	93.2	94.4
MSCOCO	36.7	36.0	75.7	67.1	84.7	68.1	72.3	69.5	72.7	71.1	74.5
RefCOCO	64.5	57.1	87.6	87.1	89.4	85.1	86.8	83.3	91.4	91.2	93.7
RefCOCO-Matching	3.9	82.6	84.6	85.8	83.0	89.3	85.1	84.4	91.1	88.7	92.3
Visual7W-Pointing	56.1	51.2	81.0	69.2	93.2	69.5	83.1	71.5	84.2	88.3	90.8

Continued on next page

Table 9 continued

Task	ColP. 1.3	LamRA 7B	V2Vec 7B	V2Vec V2	CAFe 7B	DUME 2B	DUME 7B	UME 2B	UME 7B	DiLRE 2B	DiLRE 7B
Video tasks (18; Hit@1)											
K700	23.4	32.1	35.5	38.0	40.1	22.7	27.3	35.8	42.8	53.8	55.2
SmthSmthV2	25.1	25.3	32.1	42.8	35.8	37.7	25.1	44.1	50.4	61.9	58.3
HMDB51	24.8	33.8	42.2	40.9	46.9	53.4	42.6	54.4	58.3	57.6	57.1
UCF101	49.4	53.0	61.8	60.0	39.6	55.7	48.8	67.2	70.0	76.7	77.7
Breakfast	10.9	20.1	23.8	14.8	16.6	18.9	20.8	20.1	21.5	25.4	21.5
MVBench	33.7	37.6	28.5	33.7	48.9	48.8	47.4	49.9	58.2	54.4	61.7
Video-MME	30.6	35.1	27.8	30.7	46.0	39.2	40.2	41.7	47.3	44.8	50.0
NExTQA	35.2	44.9	20.3	20.9	62.4	55.2	48.6	59.9	69.6	62.1	69.9
EgoSchema	38.4	47.0	21.8	34.0	60.0	23.2	50.4	45.4	52.4	48.0	52.0
ActivityNetQA	51.3	48.5	51.4	52.3	76.0	66.7	50.2	57.8	76.0	72.9	72.8
DiDeMo	22.8	22.8	29.3	30.4	37.8	16.9	0.10	32.4	40.0	41.5	45.1
MSR-VTT	17.6	25.0	34.5	28.3	36.5	16.2	0.10	34.3	38.9	44.0	46.0
MSVD	45.4	41.9	46.7	48.1	56.4	34.9	28.8	55.4	60.8	65.4	67.2
VATEX	16.7	18.7	25.5	26.5	32.0	11.1	13.8	29.9	32.6	35.8	38.2
YouCook2	5.3	7.5	9.0	10.6	9.5	0.06	0.00	12.7	18.5	21.0	22.6
QVHighlight	19.9	60.9	57.7	49.4	58.4	40.3	29.4	57.5	54.9	54.7	66.6
Charades-STA	29.0	18.8	19.8	20.2	18.7	16.1	15.8	20.4	21.9	21.2	24.2
MomentSeeker	27.6	31.8	39.3	40.8	41.4	33.7	38.8	41.2	41.1	46.1	50.4
Visual-document tasks (24; nDCG@5)											
ViDoRe-arxivqa	81.7	53.0	60.2	80.6	73.3	68.7	66.6	73.9	73.6	83.9	85.1
ViDoRe-docvqa	56.6	25.4	34.7	44.9	38.3	33.6	35.8	37.9	41.1	49.9	51.2
ViDoRe-infovqa	84.9	72.3	70.4	83.7	80.6	74.5	72.8	76.2	80.8	86.2	87.9
ViDoRe-tabfquad	86.9	66.1	78.2	89.2	80.7	78.3	89.2	86.1	90.2	93.7	95.6
ViDoRe-tatdqa	70.9	25.9	27.6	43.8	37.8	35.3	38.5	40.6	46.7	56.6	56.7
ViDoRe-shiftproject	75.1	27.3	38.6	60.8	52.0	61.8	61.9	66.8	65.0	71.4	81.5
ViDoRe-SynDoc-artificial-intelligence	95.7	72.0	67.7	88.5	86.0	74.3	69.3	85.9	89.5	94.8	94.3
ViDoRe-SynDoc-energy	94.7	65.2	60.4	86.5	84.8	78.4	68.4	83.3	85.7	87.6	90.8
ViDoRe-SynDoc-government-reports	93.6	72.2	61.8	85.0	85.0	83.0	83.1	82.6	89.8	93.2	92.8
ViDoRe-SynDoc-healthcare-industry	95.9	83.8	69.9	92.2	88.4	88.2	84.9	90.8	94.3	94.2	94.3
ViDoRe-esg-reports-human-labeled-v2	51.3	33.0	6.8	45.6	50.7	48.0	40.4	50.2	50.4	56.3	64.7
ViDoRe-biomedical-lectures-v2-multilingual	54.7	35.9	5.1	44.3	50.9	39.8	37.4	46.2	50.7	53.8	58.2
ViDoRe-economics-reports-v2-multilingual	49.0	31.9	13.9	43.0	54.3	44.1	29.6	45.7	57.8	50.9	58.3
ViDoRe-esg-reports-v2-multilingual	52.9	32.5	11.9	46.6	42.3	41.1	33.5	42.7	43.2	52.4	57.8
VisRAG-ArxivQA	80.9	37.7	52.6	76.9	74.0	35.8	77.3	74.3	80.5	82.9	84.3
VisRAG-ChartQA	72.3	68.2	57.7	83.7	82.7	47.2	83.4	86.0	85.0	88.9	86.8
VisRAG-MP-DocVQA	82.0	72.0	60.6	88.1	75.1	35.3	83.8	75.6	83.4	84.6	86.4
VisRAG-SlideVQA	85.1	71.1	54.7	84.1	87.6	61.3	91.5	87.1	91.5	95.0	95.3
VisRAG-InfoVQA	83.5	67.9	66.0	82.3	87.9	64.7	88.2	84.4	89.2	90.8	90.5
VisRAG-PlotQA	79.3	56.4	62.7	75.9	69.4	38.5	71.3	68.0	72.7	71.1	71.1
ViDoSeek-page	38.1	10.7	16.3	29.1	22.5	20.0	20.2	21.2	21.3	78.9	85.5
ViDoSeek-doc	87.5	63.9	69.4	79.0	73.8	69.5	73.2	75.9	75.3	84.0	83.7
MMLongBench-page	27.1	0.5	0.4	15.8	13.3	10.4	10.3	11.9	12.3	50.0	55.5
MMLongBench-doc	80.4	51.4	28.8	63.0	42.6	35.4	36.0	39.7	41.3	48.9	51.0

Column names. ColP. denotes ColPali v1.3; LamRA uses Qwen2.5-VL; V2Vec denotes VLM2Vec, with V2 indicating VLM2Vec-V2.0; UME denotes UME-R1. The four ViDoRe-SynDoc rows denote the corresponding SyntheticDocQA domains.

D.2 Number of Candidates for Discriminative Guidance

At refinement round ℓ, the current query embedding retrieves the top M targets C_ℓ from the task-specific candidate set $\mathcal{T}(q)$. The Discriminative Guidance Model evaluates these query–target pairs, and the resulting relevance probabilities are combined with the retrieval distribution to construct p_ℓ on C_ℓ (Section 3.3). Thus, M controls how many retrieved targets contribute guidance to each latent update; after refinement, e_L still ranks the full set $\mathcal{T}(q)$. We vary M for DiLRE-2B on MMEB-V2 using the same task evaluation protocol as the main results.

Table 10: **Effect of the number of candidates scored by the Guidance Model.** DiLRE-2B results on MMEB-V2 as the top-M retrieved targets are retained at each refinement round. Numbers in parentheses give task counts.

M	Image (36)	Video (18)	VisDoc (24)	All (78)
4	71.4	49.1	73.3	66.8
8	71.5	49.2	74.5	67.3
16	71.6	49.3	74.8	67.4
32	71.7	49.3	75.0	67.5

Table 10 expands the $M = 4$ and $M = 32$ comparison in Figure 4; $M = 32$ is the setting used for the main MMEB-V2 results. The overall score rises from 66.8 at $M = 4$ to 67.3 at $M = 8$ and reaches 67.5 at $M = 32$. The intermediate $M = 16$ setting reaches 67.4. The gains are most visible for visual-document tasks, while the video average changes little across these settings. Smaller M therefore retains much of the retrieval performance with fewer candidate pairs evaluated by the Guidance Model at each round.

The per-task results for all four values of M are reported in Table 11.

Table 11: **Per-task results for the top-M guidance ablation.** DiLRE-2B scores on MMEB-V2 for each number of retrieved targets scored by the Guidance Model at a refinement round. Image and video tasks use Hit@1; visual-document tasks use linear-gain nDCG@5. Scores are shown on a 0–100 scale. ViDoRe-SynDoc abbreviates SyntheticDocQA.

Task	$M = 4$	$M = 8$	$M = 16$	$M = 32$
Image tasks (36; Hit@1)				
ImageNet-1K	81.9	81.9	81.9	81.9
N24News	77.7	77.5	77.6	77.7
HatefulMemes	69.2	69.2	69.2	69.2
VOC2007	90.3	90.9	90.6	90.6
SUN397	82.6	82.5	82.6	82.6
Place365	45.9	46.0	45.6	45.6
ImageNet-A	53.9	54.2	54.4	54.5
ImageNet-R	91.4	91.6	91.5	91.5
ObjectNet	73.6	74.4	74.4	73.8
Country211	27.7	26.8	26.9	26.9
OK-VQA	67.3	68.1	68.1	68.1
A-OKVQA	57.9	57.9	57.9	57.9
DocVQA	94.7	94.7	94.7	94.7
InfographicsVQA	70.5	70.0	70.4	70.4
ChartQA	59.3	59.1	58.4	58.3
Visual7W	60.2	60.5	61.0	60.8
ScienceQA	51.6	52.2	53.7	54.3
VizWiz	57.8	57.9	57.8	57.9
GQA	79.6	81.2	81.7	81.5
TextVQA	86.6	86.6	87.0	86.7
MSCOCO-i2t	72.9	73.7	73.7	73.7
VisualNews-i2t	78.1	78.6	77.4	78.2
VisDial	80.2	80.3	80.3	80.3
MSCOCO-t2i	73.9	74.0	74.0	74.1
VisualNews-t2i	72.1	72.1	72.1	72.1
WebQA	88.6	88.4	88.4	88.4
EDIS	92.4	92.4	92.4	93.2
Wiki-SS-NQ	74.5	75.6	76.9	77.1

Continued on next page

Table 11 continued

Task	$M = 4$	$M = 8$	$M = 16$	$M = 32$
CIRR	60.6	60.9	61.1	60.9
NIGHTS	68.0	67.2	67.4	67.4
OVEN	64.4	64.5	65.1	65.0
FashionIQ	25.3	25.6	25.1	25.3
MSCOCO	70.9	71.2	71.1	71.1
RefCOCO	90.7	91.1	91.2	91.2
RefCOCO-Matching	88.7	88.7	88.7	88.7
Visual7W-Pointing	87.7	88.1	88.4	88.3
Video tasks (18; Hit@1)				
SmthSmthV2	60.1	61.2	61.6	61.9
HMDB51	58.4	58.7	58.5	57.6
UCF101	75.6	76.5	76.6	76.7
K700	53.5	54.0	53.6	53.8
Breakfast	26.8	25.4	25.4	25.4
MSR-VTT	43.8	44.0	44.0	44.0
MSVD	65.2	65.4	65.4	65.4
DiDeMo	40.0	39.8	41.6	41.5
YouCook2	20.0	20.7	20.9	21.0
VATEX	35.6	35.7	35.8	35.8
QVHighlight	54.6	54.7	54.7	54.7
Charades-STA	21.6	21.0	20.8	21.2
MomentSeeker	46.4	45.8	46.1	46.1
Video-MME	44.8	44.4	44.5	44.8
NExTQA	62.1	62.1	62.1	62.1
EgoSchema	48.0	48.0	48.0	48.0
MVBench	54.6	54.6	54.5	54.4
ActivityNetQA	72.9	72.9	72.9	72.9
Visual-document tasks (24; linear-gain nDCG@5)				
ViDoRe-arxivqa	82.6	83.4	83.9	83.9
ViDoRe-docvqa	45.9	48.0	48.4	49.8
ViDoRe-infovqa	85.3	86.1	86.2	86.2
ViDoRe-tabfquad	93.5	93.7	94.0	93.7
ViDoRe-tatdqa	54.4	56.1	56.5	56.6
ViDoRe-shiftproject	71.4	74.5	73.8	71.4
ViDoRe-SynDoc-artificial-intelligence	94.0	93.6	94.3	94.8
ViDoRe-SynDoc-energy	85.0	88.8	87.3	87.6
ViDoRe-SynDoc-government-reports	90.7	92.3	93.2	93.2
ViDoRe-SynDoc-healthcare-industry	93.3	92.0	91.9	94.2
ViDoRe-esg-reports-human-labeled-v2	54.2	56.0	56.3	56.3
ViDoRe-biomedical-lectures-v2-multilingual	51.1	52.9	54.1	53.8
ViDoRe-economics-reports-v2-multilingual	49.2	51.0	50.9	50.9
ViDoRe-esg-reports-v2-multilingual	47.6	51.6	52.3	52.4
VisRAG-ArxivQA	81.2	82.3	82.7	82.9
VisRAG-ChartQA	88.3	87.5	86.8	88.9
VisRAG-MP-DocVQA	81.1	83.1	84.0	84.6
VisRAG-SlideVQA	94.1	94.7	94.9	95.0
VisRAG-InfoVQA	90.2	90.8	90.9	90.8
VisRAG-PlotQA	70.4	71.0	71.0	71.1
ViDoSeek-page	77.6	78.6	78.9	78.9
ViDoSeek-doc	82.5	83.5	83.9	84.0
MMLongBench-doc	47.0	48.1	48.2	48.9
MMLongBench-page	47.8	49.2	49.7	50.0

D.3 Guidance Evaluation Cost Under Candidate Refresh

Each latent update changes the query embedding and can bring new targets into the top-M set C_ℓ. Since the Guidance Model is fixed and scores the same pair (q, t) identically across rounds, a score obtained in an earlier round can be reused when that target reappears. For a query refined for L rounds, the number of Guidance Model evaluations is thus $|\bigcup_{\ell=0}^{L-1} C_\ell|$, rather than the sum of the round-wise candidate counts. This lets the method refresh its retrieval evidence while avoiding repeated evaluation of overlapping candidates.

Table 12: **Guidance evaluation cost across candidate budgets.** Pairs/query counts distinct query–target pairs scored by the Guidance Model after reusing scores for repeated pairs. The last column gives the reduction relative to $M = 32$ on the same cost-audit queries. Overall scores are the full MMEB-V2 results from Table 10.

M	All (78)	Pairs/query	Pair reduction
4	66.8	5.561	84.56%
8	67.3	10.140	71.84%
16	67.4	18.946	47.39%
32	67.5	36.011	0.00%

At $M = 32$, the initial retrieved set contains 28.433 eligible targets per query on average in the cost audit. Candidate refresh raises the mean number of distinct pairs scored to 36.011, or 1.27 times that of a single pass over the initial set. The additional 7.578 pairs per query quantify the cost of admitting new candidates after latent updates; repeated candidates do not require another Guidance Model evaluation. Smaller budgets reduce this cost substantially: $M = 8$ and $M = 16$ use 10.140 and 18.946 pairs per query, respectively, representing 71.84% and 47.39% fewer evaluations than $M = 32$.

The quality gains remain meaningful at these lower pair budgets. On the full MMEB-V2 benchmark, $M = 8$ and $M = 16$ reach 67.3 and 67.4 overall, compared with 66.9 for the $M = 32$ one-pass static reranking control in Table 5. They therefore exceed that control by 0.4 and 0.5 points while evaluating fewer guidance pairs on the cost-audit workload. The full $M = 32$ latent method reaches 67.5, a further 0.6-point gain over static reranking at the 1.27-times pair count. Table 10 also shows diminishing returns as M increases: the overall score rises by 0.5 points from $M = 4$ to $M = 8$, then by only 0.1 point at each subsequent doubling. These results show that candidate refresh and latent internalization can make effective use of a limited Guidance Model evaluation budget. The pair counts isolate Guidance Model work; wall-clock throughput, which also includes latent updates, is reported in Figure 4.

D.4 Validation-Tuned Static Fusion

Equations 7 and 8 suggest a static alternative to latent refinement. If the candidates and guidance scores were fixed and each round exactly matched its target distribution, then $\pi_L(t) \propto \pi_0(t) r_\phi(q, t)^L$. DiLRE instead takes a bounded Adam step and refreshes the candidates after each round, so this identity is only a motivation for testing whether a stronger weight on the same guidance scores explains the gain. We therefore keep the initial query and candidate embeddings fixed and rank with

$$S_\alpha(q, t) = \frac{e_0^\top v_t}{\tau} + \alpha \log r_\phi(q, t). \tag{10}$$

For each task, we select α using its corresponding training-side validation examples and then hold it fixed on test queries. This control changes the static fusion weight without updating the query representation.

Table 13: **Validation-tuned static fusion on MMEB-V2.** DiLRE-2B test results with the same pretrained embedding and guidance models. The coefficient α is selected on training-side validation examples; the fixed-weight and latent-refinement rows are from Table 5.

Variant	Image (36)	Video (18)	VisDoc (24)	**All (78)**
Static reranking ($\alpha = 1$)	71.2	48.2	74.6	66.9
Validation-tuned static fusion	70.5	47.3	72.6	65.8
Latent refinement (DiLRE)	**71.7**	**49.3**	**75.0**	**67.5**

The validation-selected fusion weight reaches 65.8 overall on held-out test queries, below both fixed-weight static reranking (66.9) and latent refinement (67.5). In this setting, tuning the scalar guidance weight does not reproduce the benefit of refining the query representation. Together with the direct embedding update in Table 5, this comparison supports updating latent reasoning tokens rather than only changing the final scores or directly updating the output embedding.

D.5 Retrieval Under Candidate-Set Expansion

A retrieval index may grow after a query embedding has been computed. We examine whether the discriminative information acquired through latent refinement remains useful when the final embedding encounters targets that were not present during that query's refinement. In a separate panel of 384 queries from 12 source groups, each query is refined using only its original nine candidates for up to two rounds. We then hold the resulting query embedding e_L fixed and add further candidates from the same source group. The added targets are ranked by $e_L^\top v_t$ using fixed candidate embeddings, with no additional Guidance Model evaluation or latent update. For comparison, we rank the same expanded sets with the unrefined query embedding e_0.

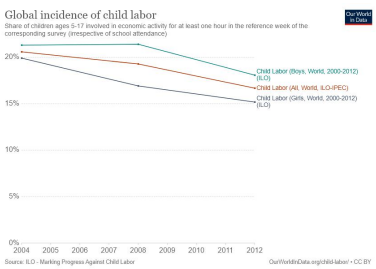
Table 14: **Retrieval with a fixed query embedding as the candidate set grows.** Hit@1 (%) measures recovery of each query's original positive target on the 384-query panel. Candidate counts are means across queries; additional targets were not used to refine the query embedding.

Mean candidates	Unrefined e_0	Refined e_L
41.58	68.23	72.14
74.75	68.23	72.14
141.17	68.23	72.14
273.75	68.23	72.14

As the mean search set grows to approximately 274 targets, more than 30 times the nine used during refinement, the refined embedding retains 72.14% Hit@1, compared with 68.23% for the unrefined embedding. This result shows that the query-local update can produce a reusable retrieval vector: its benefit persists when same-source targets are added without asking the Guidance Model to score those new targets. The test measures retention of the original positives under added candidates; it is separate from the full MMEB-V2 evaluation in Table 1.

E Qualitative Retrieval Examples

Figures 5–8 illustrate four retrieval decisions from MMEB-V2. Each comparison uses the same query and task candidate set. The *DiLRE-static rerank* row uses the combined retrieval and guidance scores for static reranking of retrieved candidates; its target rank is measured within that set. The remaining rows report top-1 targets and annotated-target ranks over the full task candidate set. We retain the opening and decisive statements of UME-R1's generated reasoning verbatim, marking omitted intervening text with ellipses, alongside the targets retrieved by the explicit-text methods and DiLRE. DiLRE's result comes from the final refined query embedding, rather than directly taking the Guidance Model's highest-scoring candidate. These examples complement the aggregate comparisons in Tables 1 and 5.



Query. "When does the gap between data about boys and girls reach become largest?"

Annotated target. 2008.

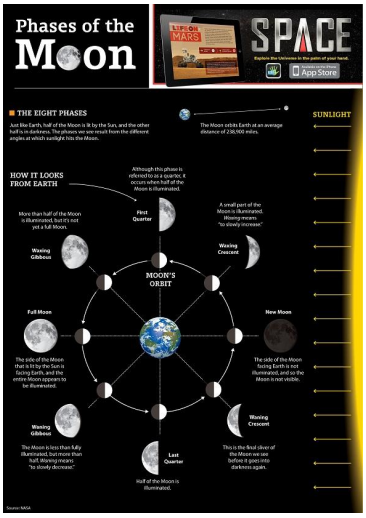
Incorrect discriminative information

UME-R1. "So, let's analyze the input. The question is asking about when the gap between data about boys and girls reaches its largest gap. … Looking at the graph, the gap between the two lines is largest around 2004, then decreases."

Method	Retrieved top-1	Target rank
LaME	2004	2
UME-R1	2004	2
RIME	2004	2
DiLRE	2008	1
DiLRE-static rerank	2004	2

The boys–girls gap is largest in 2008, not 2004. UME-R1's reasoning selects 2004, matching its retrieved top-1. Without latent refinement, static reranking still leaves 2008 second; DiLRE's final embedding retrieves the year with the largest inter-series gap.

Figure 5: **Comparing visual quantities in ChartQA.** The query requires comparing the vertical distance between two series across years. DiLRE retrieves the annotated year from its refined embedding.



Query. "what is the next phase after waning Crescent?"

Annotated target. New moon.

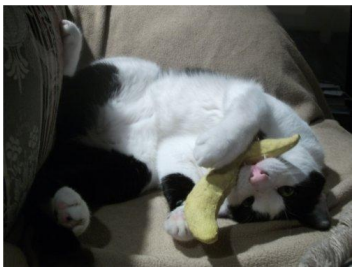
Incorrect discriminative information

UME-R1. "So, let's analyze the input. The question is asking for the next phase after Waning Crescent. … Synthesize: The next phase after Waning Crescent is Waning Gibbous."

Method	Retrieved top-1	Target rank
LaME	Waxing gibbous	3
UME-R1	Waning gibbous	3
RIME	Waning gibbous	2
DiLRE	New moon	1
DiLRE-static rerank	Waxing gibbous	3

The diagram places the waning crescent immediately before the new moon. Static reranking leaves the immediate successor third, whereas DiLRE's refined embedding ranks it first.

Figure 6: **Relational order in InfographicsVQA.** The retrieved target must follow the phase indicated by the query, rather than merely describe another phase of the same cycle.



Query. "Who is in the photo?"

Annotated target. No one.

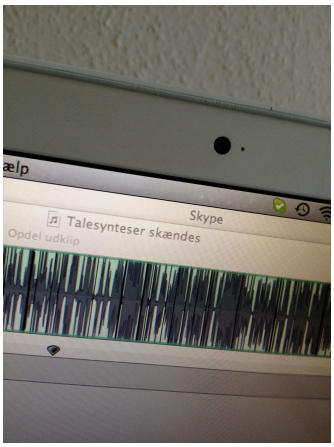
Incorrect discriminative information

UME-R1. "So, let's analyze the input. ... The key element is the cat, which is the main subject. ... Then, the synthesis would be the cat."

Method	Retrieved top-1	Target rank
LaME	The cat.	3
UME-R1	The cat.	2
RIME	The cat.	2
DiLRE	No one.	1
DiLRE-static rerank	The cat.	3

The image contains a cat but no person. UME-R1 names the salient object without resolving the question's entity type. The ablation ranks "No one" third; DiLRE's refined embedding places this answer first.

Figure 7: **Entity type in Visual7W.** The visual subject is a cat, whereas the question asks who is present. DiLRE retrieves the annotated answer, "No one," from its refined embedding.



Query. "What is it?"

Annotated target. Skype.

Missing discriminative information

UME-R1. "So, let's analyze the input. The image shows a laptop screen with the Skype application open, ... So the main thing is a laptop with Skype open showing a chat message." Its summary reads: "A laptop screen displaying the Skype application with a chat message."

Method	Retrieved top-1	Target rank
LaME	Computer screen	2
UME-R1	Computer	2
RIME	Computer screen	2
DiLRE	Skype	1
DiLRE-static rerank	Computer screen	2

The application name is visible in the image. UME-R1 mentions both Skype and the laptop screen, but its summary does not distinguish the requested application from its display. Static reranking leaves Skype second; DiLRE's refined embedding retrieves the application first.

Figure 8: **Application versus display in VizWiz.** A textual description may contain the right name while leaving a broad physical container salient. DiLRE retrieves the annotated application.

Across these examples, the discriminative distinction is a relation between the query and competing targets: a comparison across years, a successor in a visual sequence, the distinction between an object and a person, or a named application instead of its display. The within-case comparisons illustrate how the Guidance Model's candidate-level evidence becomes a full-set retrieval decision after DiLRE internalizes it in the query's latent reasoning tokens.

F Further Discussion

F.1 Provenance of Benchmark Results

The DiLRE entries in Tables 1 and 2 are from our evaluations. Baseline entries are literature-reported results under the MMEB benchmarks introduced by VLM2Vec and extended by VLM2Vec-V2 (Jiang et al., 2024b; Meng et al., 2025). The established MMEB-V2 baseline families are reported by VLM2Vec-V2 and used as comparison points in subsequent studies including UME-R1, PLUME, and LaME (Meng et al., 2025; Lan et al., 2025; He et al., 2026; Wu et al., 2026b); for newer methods, we use the results in their respective reports. For MMEB-V1 in-distribution and out-of-distribution averages not tabulated in the original method reports, VLM2Vec-V2 and PLUME follow LaME's published breakdown, while RIME follows its published task-level results (Wu et al., 2026b;a). These external results are cited comparisons, not independent reruns by us.

Table 9 has a single source for its detailed baselines: its nine prior-method columns reproduce the 78-task MMEB-V2 results in Appendix D, Table 5 of UME-R1 (Lan et al., 2025), at the reported one-decimal precision. The two DiLRE columns report our evaluations on the same benchmark.

F.2 Explicit Reasoning and CoT-Derived Supervision

In this paper, *explicit CoT* in reasoning for embeddings refers broadly to explicit reasoning expressed in generated text, rather than only to a stepwise trace or content enclosed by `<think>` tokens. That text may take the form of a rationale, a structured response, a teacher-produced answer or summary drawn from a reasoning response, or a task-directed rewrite; it may be produced by the embedding model or by a separate model. Some prior work reserves "CoT" for stepwise reasoning, but our broader terminology concerns the role of generated reasoning text in forming an embedding, without changing those works' own distinctions. We call training targets originating in such text *CoT-derived textual supervision*; *process-level CoT supervision* specifically reconstructs intermediate steps. Raw queries, targets, and pre-existing answer labels are not CoT-derived merely because they are text.

This broader framing reflects the variety of reasoning-for-embedding methods. TTE, UME-R1, and RGE use generated rationales to inform embeddings (Cui et al., 2025; Lan et al., 2025; Liu et al., 2025a); MMEmb-R1 and Think-When-Needed additionally learn when explicit reasoning is useful (Wang et al., 2026b; Zhang et al., 2026). Embed-RL and ReWAM organize textual reasoning around retrieval evidence (Jiang et al., 2026a;b). RIME instead learns retrieval-oriented rewrites and explicitly distinguishes them from CoT (Wu et al., 2026a); it belongs to the broader explicit-text family, not to stepwise CoT. Even a latent method can draw on this family: PLUME uses explicit CoT as a temporary training scaffold (He et al., 2026). Across these designs, the generated text may differ in structure and use, but reasoning for an embedding is not defined by a `<think>` span.

Embed-RL (Jiang et al., 2026a) illustrates why the distinction between text provenance and supervision granularity matters. Its Reasoner produces a structured response for each query–positive pair, ending with an answer that summarizes retrieval-relevant information. That answer is a component of a generated CoT response, although it is not itself the intermediate reasoning trajectory. LaME (Wu et al., 2026b) explicitly avoids supervising intermediate steps and trains aggregated latent states with InfoNCE, but also uses an auxiliary decoder for retrieval targets, answers, or keywords. In its released Embed-RL-format data path, the decoded answer is extracted from the generated query and positive reasoning responses, rather than copied from their raw inputs. Thus, LaME uses text targets originating in generated CoT and remains dependent on explicit CoT supervision in this training path. This illustrates that latent reasoning for embeddings can still rely on supervision supplied by an explicit reasoning process.

DiLRE uses neither generated reasoning targets nor auxiliary textual decoding to supervise its latent states. It learns from retrieval pairs and query–candidate relevance labels, then uses guidance from the current retrieved candidates to refine the query's latent token inputs at test time while keeping model parameters and candidate embeddings fixed. LaME also uses contrastive supervision on latent states, and the absence of stepwise CoT labels alone is not our contribution. The substantive difference is how discriminative evidence is acquired and internalized through retrieval-conditioned test-time refinement.

F.3 Why Internalize Guidance in Latent States?

The useful output of a universal multimodal embedding system is a query vector that can be compared with pre-encoded targets. Static reranking can improve the order of candidates scored by the guidance model, but leaves the query embedding e_0 unchanged: its feedback does not alter similarity to targets outside that scored set. DiLRE instead internalizes the feedback by refining the query's latent reasoning token inputs. The frozen encoder then produces a single updated embedding e_L, which ranks the full candidate set $\mathcal{T}(q)$ against unchanged target embeddings. This preserves the embedding interface while allowing relevance guidance to affect retrieval beyond the shortlist on which it was obtained.

The iterative procedure also changes which candidates can supply feedback. After each latent update, $e_{\ell+1}$ retrieves a new top-M set $C_{\ell+1}$; a target outside C_ℓ can enter this set and be assessed in a later round. This possibility is absent from static reranking over a fixed shortlist. Feedback is mediated by the pretrained latent reasoning tokens and the encoder's attention before the final embedding is formed, rather than being applied only as a fixed correction to retrieval scores. Contrastive supervision anchors these tokens, and each query-local update remains bounded around their pretrained inputs.

The controlled comparison in Table 5 supports this distinction: on MMEB-V2 with the 2B backbone, latent refinement reaches 67.5 overall, compared with 66.9 for static reranking and 66.5 for a direct update to the final embedding. Even selecting a static fusion weight on training-side validation examples yields 65.8 (Table 13), so adjusting that weight alone does not recover the observed gain. In the separate candidate-set expansion experiment, the refined embedding retains 72.14% Hit@1 versus 68.23% for the unrefined embedding as the mean candidate count grows to about 274, without further guidance evaluations (Table 14). Finally, the query oracle in Figure 3 reaches 72.3 overall by retrospectively selecting among evaluated rounds using test labels. It is an upper bound, not a deployable result, but shows that the sequence of refined embeddings contains further room for better selection.

G Limitations

DiLRE selects a task-level refinement horizon from training-side validation data and applies it uniformly to test queries. This policy provides a reproducible round budget, requires no test labels, and already yields strong retrieval performance. Its remaining scope for improvement lies in the granularity of round selection: different queries within a task may benefit from different numbers of updates. The retrospective query oracle in Figure 3 reaches 72.3 overall, compared with 67.5 under the validation-selected policy, indicating that the refinement trajectory contains substantial additional potential. Because the oracle uses ground-truth labels, this gap does not itself establish the gain of a deployable policy. It does, however, motivate learning a query-adaptive stopping rule from training-side data, using observable signals such as retrieval margins, changes in the candidate ranking, and agreement between embedding similarity and relevance guidance. Such a rule could capture more of DiLRE's refinement potential while preserving its embedding output and requiring no labels at inference.