LaME: Learning to Think in Latent Space for Multimodal Embedding via Information Bottleneck

Anonymous ACL submission

Abstract

Reasoning-driven universal multimodal embedding has advanced rapidly by introducing Chain-of-Thought (CoT) reasoning into the embedding pipeline. Despite the strong performance across both general and complex tasks, this paradigm suffers from two core limitations: (i) autoregressive CoT reasoning incurs high computational cost, making it impractical for low-latency retrieval; and (ii) embedding performance is heavily coupled with CoT annotation quality, making large-scale training unreliable. These raise fundamental questions: *Is textual CoT the optimal form of reasoning for embedding, and can effective embedding reasoning be accomplished in latent space?* To this end, we propose **LaME** (**La**tent Reasoning **M**ultimodal **E**mbedding), which formulates embedding-oriented latent reasoning as a weakly supervised information bottleneck. LaME employs K learnable reason tokens as a fixed-capacity bottleneck, completing all reasoning within a single forward pass. The two weak supervision signals structurally decouple contrastive from autoregressive objectives and eliminate dependence on CoT annotations, while a two-stage training pipeline ensures stable convergence. Experiments on MMEB-v2 and MRMR show that LaME achieves competitive performance, surpassing some explicit CoT-based models, while delivering $60\times$ faster inference than explicit CoT methods and $2\times$ faster than latent baselines with throughput comparable to discriminative embedding models. *Code will be released.*

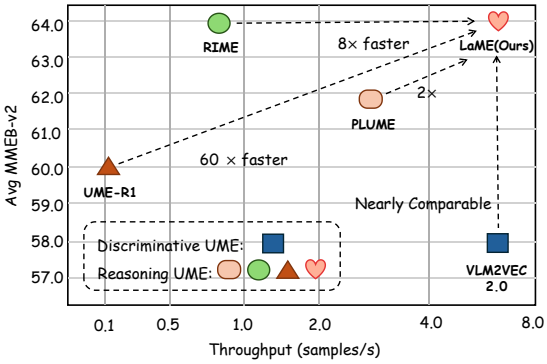

Figure 1: Performance vs. throughput on MMEB-V2 on a single H20 GPU. LaME achieves competitive performance with reasoning-driven methods while being $60\times$ faster than explicit CoT approaches and $2\times$ faster than iterative latent approaches, thanks to single-pass inference with only 8 reason tokens.

1 Introduction

The rapid development of Multimodal Large Language Models (MLLMs) (Liu et al., 2024; Chen et al., 2024; Wang et al., 2024) has fundamentally transformed multimodal embedding learning. Traditional dual encoders such as CLIP (Radford et al., 2021) and EVA-CLIP (Sun et al., 2023) align cross-modal representations via contrastive pre-training, yet struggle with complex multimodal inputs. Recent works including VLM2Vec (Jiang et al., 2025), E5-V (Jiang et al., 2024), GME (Zhang et al., 2025d) and so on (Liu et al., 2025c; Lin et al., 2024) demonstrate that MLLMs can be fine-tuned into powerful embedding backbones by leveraging intrinsic vision-language fusion and instruction-following capabilities, marking a paradigm shift from modality-specific encoding to unified multimodal representation learning.

Based on this, recent works further exploit the generative capacity of MLLMs by introducing explicit Chain-of-Thought (CoT) reasoning (Wei et al., 2022; Guo et al., 2025) to enrich the information available for generative embedding. UME-R1 (Lan et al., 2025), RIME (Wu et al., 2026), RGE (Liu et al., 2025b), and Think-Then-Embed (Chen et al., 2025) demonstrate that generative embeddings empowered by internal or external reasoning outperform purely discriminative ones. However, this paradigm suffers from two core limitations that constrain its practical feasibility. (1) The CoT-based reasoning incurs high com-

putational costs at the inference stage. Even with various acceleration strategies, it remains clearly impractical for large-scale and low-latency retrieval scenarios. (2) There exists a strong coupling between embedding generation and cold-start CoT annotation. Defective oracle reasoning trajectories readily degrade learned embeddings, making them highly susceptible to explicit textual annotations. These limitations raise a deeper question: *Is textual CoT the optimal form of reasoning for embedding?*

Motivated by Coconut (Hao et al., 2025), PLUME (He et al., 2026) integrates latent and explicit reasoning paradigms, iteratively feeding hidden states back as input embeddings to explore latent reasoning while retaining partial explicit CoT. Its training still follows a progressive learning scheme, replacing textual CoT with latent transitions (Ma et al., 2025; Deng et al., 2024; Zhang et al., 2025e,a). This inherently limits flexible latent reasoning, as the model is guided to imitate text-based reasoning patterns instead of learning thinking optimal for retrieval tasks. To this end, the information bottleneck (IB) (Tishby et al., 2000; Alemi et al., 2017) can be introduced to spark unconstrained exploration in latent space. By constraining *what* the latent thinking must preserve through a fixed-capacity bottleneck, IB leaves *how* the model organizes its internal representations entirely unconstrained, enabling it to spontaneously discover latent structures without adhering to any prescribed reasoning format.

Building on this perspective, we propose **LaME** (**La**tent Reasoning **M**ultimodal **E**mbedding) which instantiates the IB principle through K learnable reason tokens appended to the MLLM input as a fixed-capacity bottleneck. The two weak supervision signals shape the bottleneck: (1) a lightweight decoder head that reconstructs retrieval targets, predefined answers and keywords from the first K_r latent tokens, serving as a reasoning probe that requires no CoT process at all and thus decouples training from the quality of CoT annotations; (2) an aggregated embedding head that merges the remaining K_e latent tokens into one embedding and optimizes it against the retrieval objective, structurally separating contrastive from generative supervision and aligning latent reasoning with the retrieval objectives. Together, these two heads ensure latent tokens preserve adequate information without explicit supervision over intermediate latent reasoning.

However, end-to-end joint training is unstable because randomly initialized reasoning tokens lack semantic structure, making IB supervision ineffective in early training. To address this, we adopt a two-stage training pipeline. In the Bottleneck Warm-up stage, the MLLM backbone is frozen, and only the reasoning tokens and supervision heads are optimized to stabilize the bottleneck initialization. In the Joint Optimization stage, all parameters except the vision encoder are unfrozen, and an additional embedding token is appended to the end of the input for target contrastive supervision, enabling effective gradient propagation for embedding-oriented latent reasoning. In summary, our contributions are summarized as follows:

- We propose **LaME**, a framework that casts embedding-oriented latent reasoning as a weakly supervised information bottleneck, completely decoupling internal thinking from explicit CoT formats and exploring latent reasoning within a single forward pass.
- We design dual-head bottleneck supervision that structurally separates contrastive from generative supervision, together with a two-stage training pipeline that ensures stable bottleneck convergence.
- By confining latent reasoning to prefill tokens within a single forward pass, LaME achieves competitive performance on MMEB-v2 (Meng et al., 2025) and MRMR (Zhang et al., 2025c), with $60\times$ faster inference than explicit CoT methods and $2\times$ faster inference than iterative latent rollout baselines.

2 Related Works

2.1 Multimodal Embedding Models

Early multimodal embeddings rely on dual encoders with contrastive objectives, e.g., CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023). With MLLMs, VLM2Vec (Jiang et al., 2025), E5-V (Jiang et al., 2024), MM-Embed (Lin et al., 2024), GME (Zhang et al., 2025d), UniME (Gu et al., 2025a), UniME-V2 (Gu et al., 2025b), and RzenEmbed (Jian et al., 2025) show that MLLMs serve as strong embedding backbones via contrastive fine-tuning; MetaEmbed (Xiao et al., 2026) explores test-time scaling for multimodal retrieval. Recently, reasoning-enhanced methods have emerged: UME-R1 (Lan et al., 2025) and Think-Then-Embed (Chen et al., 2025) demonstrate that explicit CoT before embedding yields

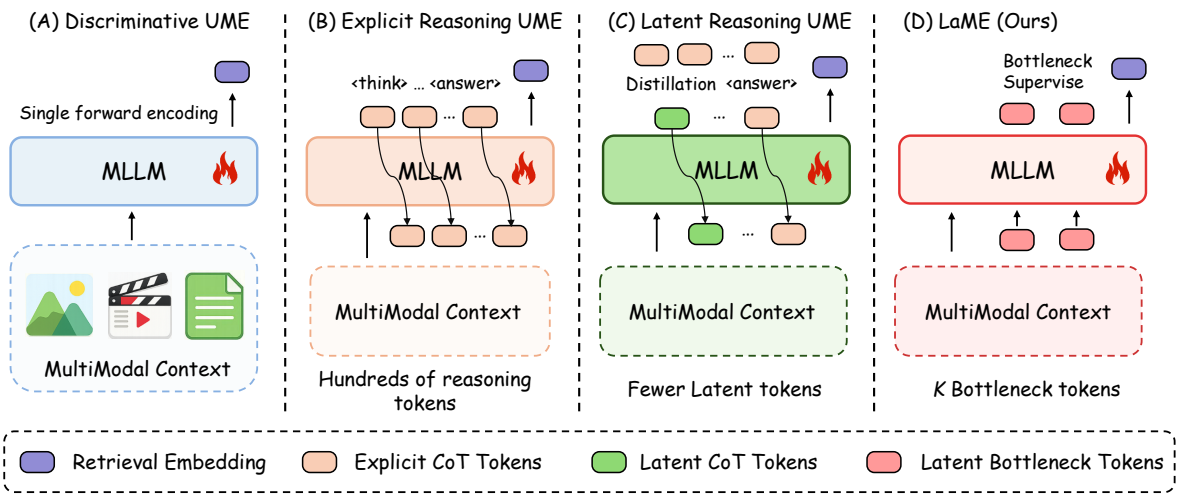

Figure 2: Comparison of universal multimodal embedding paradigms. (A) **Discriminative**: direct encoding without reasoning. (B) **Explicit reasoning**: autoregressive CoT generation before embedding. (C) **Latent reasoning**: iterative hidden-state rollout. (D) **LaME**: all reasoning is confined to bottleneck tokens in a single forward pass.

substantial gains, ReMatch (Liu et al., 2025a) leverages generative matching, RIME (Wu et al., 2026) introduces retrieval-friendly rewriting, while PLUME (He et al., 2026) explores latent reasoning via iterative hidden-state rollouts. However, explicit CoT methods entangle retrieval learning with autoregressive generation, , incurring heavy inference costs and CoT annotation dependency.

2.2 Latent Space Reasoning

Chain-of-thought prompting (Wei et al., 2022) enables complex reasoning but incurs decoding overhead. Coconut (Hao et al., 2025) and Pause tokens (Goyal et al., 2024) pioneer latent reasoning by feeding hidden states back as inputs or adding extra computation slots without text generation. Subsequent works compress CoT into latent spaces: COT-Valve (Ma et al., 2025), Implicit CoT (Deng et al., 2024), CODI (Shen and Yan, 2025), and CoLaR (Lu et al., 2025) achieve compression via variable-length sequences, explicit-to-implicit training, self-distillation, or dynamic single-pass schemes. TokenSkip (Xing et al., 2025) supports step-level controllable compression. Soft-CoT (Xu et al., 2025), LT-Tuning (Liu et al., 2026), SIM-CoT (Zhang et al., 2025e), and Light-Thinker (Zhang et al., 2025a) replace discrete CoT with instance-specific soft tokens. In retrieval, LaSER (Jin et al., 2026) and PLUME (He et al., 2026) extend latent reasoning to dense retrievers and multimodal embeddings, representing preliminary explorations for retrieval.

2.3 Information Bottleneck for Representation Learning

The IB principle (Tishby et al., 2000) learns representations that compress input X while preserving task-relevant information about Y, optimizing $\min I(Z, X) - \beta I(Z, Y)$. Deep Variational IB (Alemi et al., 2017) makes this tractable for neural networks; related ideas underpin CPC (van den Oord et al., 2019) and Deep InfoMax (Hjelm et al., 2019). Recent work shows that projection heads implicitly act as bottlenecks (Zhuo et al., 2025), and explicit IB objectives improve contrastive and vision-language pretraining (Li et al., 2025; Ji et al., 2025; Almudévar et al., 2025). OMIB (Wu et al., 2025) extends IB to multimodal settings with optimality guarantees. BToks (Sun et al., 2026) uses learnable bottleneck tokens as pooling mechanisms with generative condensation supervision. Our work differs in that bottleneck tokens serve as a medium for latent reasoning rather than mere pooling, supervised by an IB objective.

3 Our Framework

This section begins with preliminary on the background of universal multimodal embeddings (Section 3.1). We then introduce reason tokens as an information bottleneck that confines latent reasoning to a single forward pass (Section 3.2). Next, we present dual-head supervision that steers the bottleneck without relying on CoT annotations (Section 3.3). Finally, we introduce the two-stage training pipeline for stable bottleneck convergence (Sec-

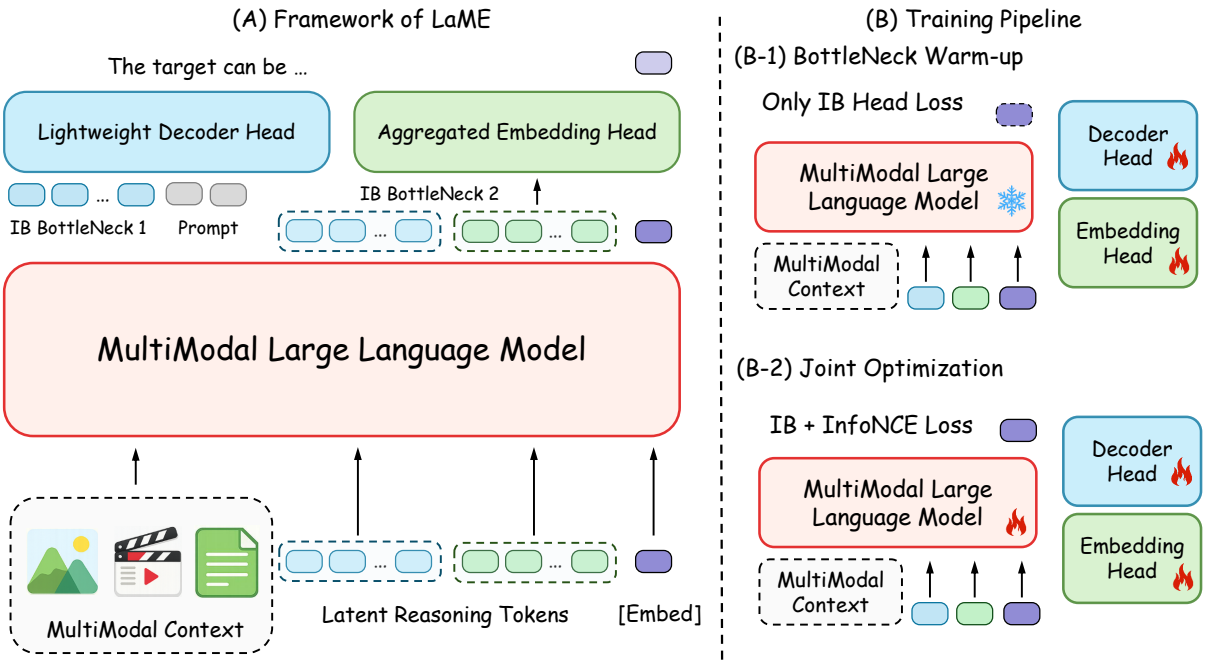

Figure 3: Overview of the LaME framework. K reason tokens form an information bottleneck. A Decoder Head and an Embedding Head supervise them with two-stage training: Bottleneck Warm-up and Joint Optimization.

tion 3.4).

3.1 Preliminary

Given a query q and a candidate set, the goal of Universal Embedding Model is to retrieve the most relevant target t^+ via similarity in a shared embedding space. Existing methods differ in the thinking process injected before embedding extraction. **Discriminative** methods directly encode the input into a normalized embedding $f_\theta(x)$ without any thinking process, and optimize with InfoNCE:

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(f_\theta(q_i)\cdot f_\theta(t_i^+)/\tau)}{\sum_{j=1}^{N}\exp(f_\theta(q_i)\cdot f_\theta(t_j)/\tau)} \tag{1}$$

where τ is the temperature. It is efficient but lacks intermediate reasoning for complex queries. **Explicit reasoning** methods address this by first generating a textual chain-of-thought $o^x = g_\phi(x)$, then embedding the concatenation $f_\theta(x, o^x)$. While effective, the autoregressive decoding of o^x is prohibitively slow at inference time. **Latent reasoning** methods replace partial textual thinking with hidden state iteration, but retain inference overhead and rely on retrieval-friendly CoT annotations for progressive imitation training. Our method takes a different route: it confines all latent reasoning to K learnable tokens within a single forward pass, governed by the IB objective.

3.2 Reason Tokens as Information Bottleneck

The core idea of LaME is to append K learnable reason tokens to the input and supervise them with an information bottleneck (IB) objective, thereby guiding the reasoning process indirectly. Concretely, given an input $\mathbf{x}$ which may contain interleaved text and visual tokens, we append K reason tokens $\mathbf{r} = [r_1, \ldots, r_K]$ to form the extended input $[\mathbf{x}; \mathbf{r}]$. A single forward pass through the MLLM produces hidden states for all positions:

$$[\mathbf{h_x}; \mathbf{h_r}] = \text{MLLM}([\mathbf{x}; \mathbf{r}]) \tag{2}$$

Under causal attention, the reason tokens $\mathbf{h_r}$ attend to all preceding context in $\mathbf{x}$ and aggregate task-relevant information. However, the final embedding is derived from the full hidden states, allowing it to spontaneously attend to either the reason tokens or the original input.

The bottleneck lies in the supervision of $\mathbf{h_r}$: we constrain the reason tokens via dedicated heads. Under this constraint, the reason tokens must think and abstract task-relevant cues from the input, while the finite capacity of K tokens forces them to retain only critical information and discard noise.

From an information-theoretic perspective, this instantiates the IB principle (Tishby et al., 2000). Let Z denote the representation carried by $\mathbf{h_r}$, X the input, and Y the target defined by the supervision heads. The supervision heads minimize

$I(Z;X)$ via the fixed capacity of K tokens while maximizing $I(Z;Y)$ via retrieval-oriented losses. Unlike variational IB, LaME requires no explicit rate regularization; the finite token count itself is the hard capacity constraint.

3.3 Dual-Head Supervision

We supervise the reason tokens with two weak IB supervision heads that address the two core limitations of explicit reasoning methods: dependency on explicit CoT annotation and misalignment between the reasoning and the retrieval objective.

Decoder Head (Reasoning Probe). We attach a lightweight autoregressive decoder that decodes target-relevant content from $\mathbf{h}_r$. Specifically, the hidden states of the first K_r reason tokens are used to reconstruct retrieval targets, predefined answers or keywords. In practice, we reuse the answer fields from existing research datasets (Jiang et al., 2026) as supervision targets:

$$\mathcal{L}_{\text{Dec}} = -\sum_{t=1}^{|y|} \log p_\psi(y_t \mid y_{<t}, \mathbf{h}_r^{1:K_r}, s) \quad (3)$$

where ψ denotes the decoder parameters, s is a decode prompt, and y is the reconstruction target. This head acts as a reasoning probe: if the latent thinking fails to distill useful information, decoding quality degrades, automatically providing gradient signals to guide reasoning optimization. Crucially, the supervision uses only the target as a weak signal without supervising intermediate thinking steps, decoupling training from CoT annotation.

Embedding Head (Retrieval Alignment). The remaining K_e reason token hidden states where $K_r + K_e = K$ are aggregated via mean pooling and normalized to produce the reason embedding:

$$\mathbf{e} = \text{Norm}\left(\frac{1}{K_e}\sum_{k=K_r+1}^{K} \mathbf{h}_r^k\right) \quad (4)$$

This embedding is trained with InfoNCE:

$$\mathcal{L}_{\text{Emb}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\mathbf{e}_{q_i} \cdot \mathbf{e}_{t_i^+}/\tau)}{\sum_{j=1}^{N} \exp(\mathbf{e}_{q_i} \cdot \mathbf{e}_{t_j}/\tau)} \quad (5)$$

Since this head bypasses the autoregressive generation path entirely, it structurally separates contrastive from generative optimization, ensuring that the latent reasoning process remains oriented toward the retrieval objective.

Joint Objective. In addition to the two IB heads, we append a learnable embed token after the reason tokens. Its hidden state is projected and normalized as the final retrieval embedding, trained with InfoNCE. The overall loss is:

$$\mathcal{L} = \mathcal{L}_{\text{NCE}} + \lambda_{\text{Dec}}\mathcal{L}_{\text{Dec}} + \lambda_{\text{Emb}}\mathcal{L}_{\text{Emb}} + \lambda_{\text{Div}}\mathcal{L}_{\text{Div}} \quad (6)$$

where λ_{Dec}, λ_{Emb}, λ_{Div} are balancing hyperparameters, and $\mathcal{L}_{\text{Div}}$ is a diversity regularizer penalizing pairwise cosine similarity among reason tokens:

$$\mathcal{L}_{\text{Div}} = \frac{1}{K(K-1)}\sum_{i \neq j} \cos(\mathbf{h}_r^i, \mathbf{h}_r^j) \quad (7)$$

which penalizes representational redundancy and encourages each reason token to attend to distinct facets of the input during latent reasoning.

3.4 Two-Stage Training

Directly optimizing $\mathcal{L}$ end-to-end is unstable. Randomly initialized $\mathbf{r}$ carries no semantic information, making the IB supervision prone to failure: $\mathcal{L}_{\text{Dec}}$ struggles to decode meaningful content from uninformative $\mathbf{h}_r$, and $\mathcal{L}_{\text{Emb}}$ lacks a reliable anchor for contrastive alignment, hindering bottleneck convergence. We therefore adopt a two-stage training pipeline to progressively aligns these latents as shown in Figure 3.

Stage 1: Bottleneck Warm-up. The MLLM backbone is frozen; only the reason tokens $\mathbf{r}$ and both IB heads are optimized. $\mathcal{L}_{\text{Dec}}$ establishes a mapping from $\mathbf{h}_r$ to target semantics, $\mathcal{L}_{\text{Emb}}$ provides initial retrieval-oriented structure, and $\mathcal{L}_{\text{Div}}$ prevents early collapse. After this stage, $\mathbf{h}_r$ already encodes meaningful information, ready to supply stable gradients to the backbone.

Stage 2: Joint Optimization. All parameters except the vision encoder are unfrozen for optimization. The contrastive loss $\mathcal{L}_{\text{NCE}}$ associated with the embed token is introduced. The two IB heads propagate high-quality gradients to the backbone network through the reasoning hidden states $\mathbf{h}_r$. Meanwhile, $\mathcal{L}_{\text{NCE}}$ directly optimizes the final retrieval embeddings within the output space of the backbone, facilitating collaborative adaptation of the entire model.

This progressive scheme ensures that by the time the backbone begins updating, the bottleneck already carries informative signals, preventing degenerate solutions and accelerating convergence.

Table 1: Results on the MMEB-V1 benchmark (Jiang et al., 2025). The scores are averaged per meta-task. **Bold** denotes the best performance within each category. Overall denotes the average score across 36 subtasks.

Models	Backbone	Per Meta-Task Score				Average Score		
		Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# of Datasets →		10	10	12	4	20	16	36
Discriminative UME								
CLIP	CLIP-L-0.4B	42.8	9.1	53.0	51.8	37.1	38.7	37.8
GME	Qwen2VL-2B	54.4	29.9	66.9	55.5	49.2	55.2	51.9
VLM2Vec	Qwen2VL-2B	58.7	49.3	65.0	72.9	64.8	53.4	59.7
VLM2Vec-V2	Qwen2VL-2B	62.9	56.3	69.5	77.3	68.8	60.1	64.9
Ops-MM-Embed	Qwen2VL-2B	**68.1**	65.1	69.2	80.9	71.8	**65.6**	69.0
ReMatch	Qwen2VL-2B	65.8	**65.9**	**70.1**	**83.3**	**72.8**	64.7	**69.2**
Explicit Reasoning UME								
UME-R1	Qwen2VL-2B	64.8	62.8	67.6	77.2	71.5	60.4	66.6
RIME	Qwen2VL-2B	67.9	64.4	69.8	82.1	64.5	72.9	69.2
Embed-RL	Qwen2.5VL-3B	62.8	67.9	68.6	90.4	65.9	71.9	69.2
RGE	Qwen3VL-2B	64.4	67.8	70.2	90.1	75.3	63.7	70.1
Think-When-Needed	Qwen3VL-4B	68.6	71.7	68.0	87.0	75.4	66.2	71.3
Think-Then-Embed	Qwen2VL-2B	**69.7**	**72.4**	**74.0**	**90.6**	**80.5**	**67.0**	**74.5**
Latent Reasoning UME								
PLUME	Qwen2VL-2B	66.1	59.5	67.6	79.9	69.6	62.2	66.3
LaME (Ours)	Qwen2VL-2B	**67.6**	**66.2**	**70.5**	**81.2**	**73.1**	**64.4**	**69.3**

Table 2: Results on the full MMEB-V2 benchmark (Meng et al., 2025). CLS: classification, QA: question answering, RET: retrieval, GD: grounding, MRET: moment retrieval, VDR: ViDoRe, VR: VisRAG, OOD: out-of-distribution. Best performance per model size in **bold**. All denotes the average score across 78 subtasks.

Model	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																
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	62.0	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
UME-R1	64.8	62.8	67.6	77.2	66.6	44.3	51.0	32.9	39.7	42.2	72.4	46.2	79.2	28.9	62.5	59.7
Ops-MM-Embed	**68.1**	65.1	69.2	80.9	69.0	**53.6**	**55.7**	**41.8**	33.7	**47.6**	76.4	53.2	77.6	32.6	65.5	63.0
PLUME	66.5	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
LaME (Ours)	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**
∼ 7B Model Size																
ColPali-V1.3	40.5	11.5	48.2	39.8	34.9	26.6	37.8	21.5	26.2	28.2	84.6	54.8	81.0	33.1	70.1	44.2
GME	57.7	34.7	71.2	59.3	56.0	37.4	50.4	28.4	38.2	38.6	**89.4**	55.6	85.0	44.4	75.2	57.8
LamRA	51.7	34.1	66.9	56.7	52.4	32.9	42.6	23.2	37.6	33.7	56.3	33.3	58.2	40.1	50.2	47.4
VLM2Vec	62.7	56.9	69.4	82.2	65.5	39.1	30.0	29.0	40.6	34.0	56.9	9.4	59.1	38.1	46.4	52.3
CAFe	63.6	61.7	69.1	**87.6**	67.6	35.8	58.7	34.4	39.5	42.4	70.7	49.6	79.5	38.1	63.9	60.6
UME-R1	67.1	69.2	71.9	84.9	71.3	48.6	60.7	38.2	39.3	47.5	75.7	50.5	83.7	37.6	67.1	64.5
Ops-MM-Embed	69.7	69.6	73.1	87.2	72.7	**59.7**	62.2	**45.7**	43.2	**53.8**	80.1	**59.6**	79.3	67.8	74.4	**68.9**
LaME (Ours)	**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

4 Experiments

4.1 Implementation Details

Following VLM2Vec-V2, we build LaME on Qwen2-VL-2B/7B-Instruct, with Qwen3-0.6B as the default decoder head and mean-pooling as the embedding head. We train on 1.5 million samples comparable to baselines, where the decoder head reuses answer fields without any CoT annotation. We only use K=8 reason tokens, with K_r=4 for decoder head and K_e=4 for embedding head. In Stage 1, the backbone is frozen and only the reason tokens and IB heads are warmed up for about 2000 steps. In Stage 2, all parameters are unfrozen and jointly trained for about 3200 steps. The loss weights are λ_{Dec}=1.0, λ_{Emb}=1.0, λ_{Div}=0.05, and the temperature τ=0.02. The batch size is 512 achieved through gradient accumulation. We use the AdamW optimizer with a learning rate of 1e−5.

Table 3: Results on the MRMR benchmark (Zhang et al., 2025c) including Art, Medicine (Med.), Science (Sci.), Humanities (Hum.), Math, Physics (Phy.), Engineering (Eng.), Business (Bus.), Negation (Neg.), Design (Des.), and Traffic (Tra.). All denotes the average score across 11 subtasks.

Model	**Backbone**	**Size**	**Knowledge**				**Theorem**				**Contradiction**			**All**
			Art	Med.	Sci.	Hum.	Math	Phy.	Eng.	Bus.	Neg.	Des.	Tra.	
Explicit Reasoning UME														
UME-R1	Qwen2-VL	7B	77.8	55.7	72.9	64.1	27.2	39.2	32.2	47.8	7.5	61.9	41.7	48.0
RIME	Qwen2-VL	7B	76.8	58.5	73.6	71.4	29.2	43.0	35.6	51.8	8.5	64.1	39.5	50.2
Discriminative UME & Latent Reasoning UME														
EVA-CLIP	EVA-ViT	0.4B	10.2	13.5	26.1	12.9	6.2	10.5	9.3	11.7	8.5	4.4	5.4	10.8
OpenCLIP	ViT-G/14	1B	56.0	17.9	33.2	22.0	5.7	5.0	7.0	9.7	13.0	8.1	12.4	17.3
VISTA	Qwen2-VL	2B	21.3	27.8	32.6	17.0	18.8	17.1	17.3	28.6	20.0	20.2	9.4	20.9
E5-V	LLaVA-Next	8B	25.1	11.7	16.6	10.8	2.1	3.4	2.5	5.2	11.5	3.7	2.1	8.6
VLM2Vec	Qwen2-VL	7B	53.5	22.4	36.7	24.0	2.1	2.8	2.8	2.9	11.5	5.6	18.3	18.1
ColPali	PaliGemma	3B	36.1	29.9	42.7	29.2	5.7	14.8	12.0	24.6	**28.5**	19.4	18.2	23.7
GME	Qwen2-VL	7B	54.3	40.1	46.8	45.6	28.8	36.0	30.2	45.1	15.0	26.3	29.6	36.2
MM-Embed	NV-Embed	8B	65.6	53.0	63.5	62.8	23.6	30.8	27.4	44.9	7.0	23.8	34.9	39.8
Ops-MM-Embed	Qwen2-VL	7B	**79.3**	52.5	70.0	**67.8**	27.7	39.5	30.1	**52.3**	8.0	55.9	**45.8**	48.1
LaME (Ours)	Qwen2-VL	7B	73.4	**58.2**	**73.8**	65.6	**29.5**	**44.4**	**36.4**	52.2	8.5	**64.9**	40.9	**49.8**

4.2 Datasets and Metrics

To evaluate the retrieval capability of LaME, we adopt MMEB-V2, a recently proposed benchmark for multimodal universal retrieval. As shown in Table 2, MMEB-V2 contains 36 and 18 sub-datasets for image and video scenarios respectively, covering classification, VQA, retrieval, visual grounding, moment retrieval, and so on with Hit@1 as the evaluation metric. For the Visual Document scenario which involves document-as-image tasks, we evaluate on 24 datasets from ViDoRe and VisRAG, with NDCG@5 as the evaluation metric.

To further validate the generalization of LaME in reasoning-intensive scenarios, we adopt MRMR, a benchmark specifically designed to assess retrieval capability in expert-level, reasoning-intensive tasks. Following their standard settings, we use nDCG@10 as the main metric.

4.3 Main Results

Results on MMEB-V1. As shown in Table 1, LaME achieves the best overall score of 69.3 among latent reasoning methods, surpassing PLUME by 3.0 points. It also outperforms competitive discriminative UME baselines such as ReMatch, demonstrating that latent reasoning can advance beyond discriminative models without incurring extra reasoning cost. Compared to explicit reasoning methods, LaME matches RIME and Embed-RL at 69.2 and trails Think-Then-Embed by only 5.2 points, despite the latter leveraging an external 72B reasoner to generate explicit CoT. LaME improves over PLUME on all four meta-tasks, indicating that the information bottleneck yields consistent gains across diverse task types.

Table 4: Ablation on core components on MMEB-V2. All crosses (✗✗✗) denote discriminative baseline.

Dec. Head	**Emb. Head**	**Two-Stage**	**Image**	**Video**	**VisDoc**	**All**
✗	✗	✗	68.5	43.9	71.3	63.8
✓	✓	✗	68.0	42.9	70.9	63.3
✓	✗	✓	68.8	44.2	71.2	64.0
✗	✓	✓	69.0	44.0	71.7	64.1
✓	✓	✓	**69.3**	**44.5**	**72.1**	**64.4**

Table 5: Effect of the number of reason tokens K on MMEB-V2 and MRMR. Throughput is measured on a single H20 GPU.

K	**Image**	**Video**	**VisDoc**	**MRMR**	**Throughput(samples/s)**
0	68.5	43.9	71.3	48.0	6.5 ± 0.1
2	68.3	44.0	71.5	48.1	6.4 ± 0.2
4	68.9	44.1	71.8	49.5	6.3 ± 0.3
8	**69.3**	**44.5**	**72.1**	49.8	6.2 ± 0.3
16	68.7	44.1	71.7	**49.9**	6.1 ± 0.4

Results on MMEB-V2. Table 2 extends the evaluation to the full MMEB-V2 benchmark covering image, video, and visual document modalities. Among 2B models, LaME ranks first overall at 64.4, leading Ops-MM-Embed by 1.4 and PLUME by 2.8 points. At the 7B scale, LaME scores 68.8, trailing Ops-MM-Embed by only 0.1 points, while achieving the best image and visual document averages. It confirms that the latent reasoning can generalize across modalities and model scales.

Results on MRMR. Table 3 presents results on the MRMR benchmark. At the 7B scale, LaME

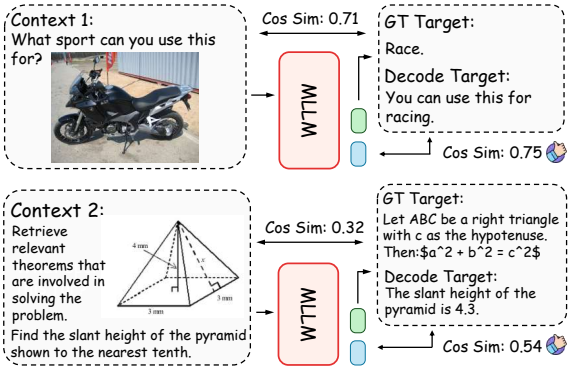

Figure 4: Qualitative examples of LaME. The green and blue block denotes the latent reason tokens for decoding and embedding.

achieves an overall score of 49.8. It ranks first on four theorem-reasoning subtasks: Science 73.8, Math 29.5, Physics 44.4, and Engineering 36.4. Its overall score is competitive with RIME at 50.2 and surpasses Ops-MM-Embed at 48.1. This shows that latent reasoning remains effective on expert-level and reasoning-intensive tasks, further validating its generality for complex queries.

4.4 Ablation Studies

Ablation on Core Components. As shown in Table 4, the discriminative baseline scores 63.8 overall. Adding both IB heads without two-stage training degrades performance to 63.3, showing that naive attachment without careful optimization can be harmful. With two-stage training, either the decoder head or the embedding head alone surpasses the baseline, yielding 64.0 and 64.1 respectively. The full configuration achieves the best result of 64.4 across all modalities, demonstrating that all components are jointly essential.

Ablation on Latent steps K. As shown in Table 5, increasing K from 0 to 8 steadily boosts performance across all MMEB-V2 modalities, with Image scores rising from 68.5 to 69.3, Video from 43.9 to 44.5, VisDoc from 71.3 to 72.1, and MRMR from 48.0 to 49.8. This validates that a moderate fixed-capacity bottleneck with learnable reasoning tokens enables effective retrieval without explicit CoT. Further increasing K to 16 yields only marginal gains in MRMR while degrading standard retrieval performance, indicating excessive capacity weakens the bottleneck constraint. Notably, the throughput merely drops slightly from 6.5 to 6.1 samples/s, verifying the inference efficiency of single-forward latent reasoning.

Table 6: Ablation on different head architectures.

Dec. Head	Emb. Head	Image	Video	VisDoc	All
Decoder Head variants (Emb. Head fixed)					
Qwen2.5-0.5B	Mean Pooling	69.0	44.2	71.8	64.3
Qwen3-0.6B	Mean Pooling	**69.3**	**44.5**	**72.1**	**64.4**
Qwen3-1.7B	Mean Pooling	69.0	44.2	71.8	64.1
Embedding Head variants (Dec. Head fixed)					
Qwen3-0.6B	Attn Pooling	68.7	43.8	71.0	63.8
Qwen3-0.6B	Mean Pooling	**69.3**	**44.5**	**72.1**	**64.4**

Ablation on Different IB Head. As shown in Table 6, Qwen3-0.6B decoder head reaches an overall score of 64.4, only 0.1 higher than Qwen2.5-0.5B, whereas Qwen3-1.7B drops to 64.1. This reveals decoder capacity does not monotonically benefit bottleneck supervision, with larger decoders incurring higher training costs and longer supervision chains. For embedding heads, mean pooling scores 64.4 and surpasses attention pooling at 63.8, proving simple aggregation more effectively compresses bottleneck information. The results verify the efficacy of our lightweight decoder and aggregated embedding head design.

4.5 Analysis on Latent Reasoning

Figure 4 shows two examples of how latent reasoning tokens enable deep thinking in a single forward pass. In Context 1, given a motorcycle image and the question, the decoder reconstructs "You can use this for racing," showing that latent reasoning activates world knowledge. In Context 2, a geometric query asks for the slant height of a pyramid. The decoder generates "The slant height of the pyramid is 4.3." indicating internalized mathematical reasoning such as the Pythagorean theorem. The similarity comparisons of reasoning latent embeddings also suggest the same conclusion.

5 Conclusion

We present LaME, which frames embedding-oriented latent reasoning as an information bottleneck problem. LaME achieves single-pass latent reasoning via learnable reason tokens, requiring no supervision on the structure or content of the latent thinking process. Only decode and contrastive losses guide the bottleneck. Experiments on MMEB-v2 demonstrate state-of-the-art performance with significant inference speedups over CoT and iterative methods. Future work will explore adaptive bottleneck capacity and extension to broader and more complex retrieval tasks.

Limitations

Although latent reasoning may be more likely to produce retrieval-friendly representations, the lack of explicit CoT traces reduces interpretability to some extent compared to CoT-based methods. Additionally, the two-stage training pipeline, while effective for stabilizing bottleneck optimization, introduces extra engineering complexity compared to single-stage approaches. Simplifying the training procedure without sacrificing performance remains an open problem.

Ethics Statement

This work studies multimodal retrieval and representation learning, and does not involve human subject recruitment, clinical data, or direct user interaction. As with other retrieval systems, our method may inherit biases, noise, and content imbalance from the source datasets. Although the rewrite-based design is intended to reduce redundant reasoning and semantic distortion, it can still produce imperfect or biased outputs; therefore, downstream use should include dataset curation and application-specific safety filtering.

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep variational information bottleneck. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Antonio Almudévar, José Miguel Hernández-Lobato, Sameer Khurana, Ricard Marxer, and Alfonso Ortega. 2025. Aligning multimodal representations through an information bottleneck. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Lei Chen, Rui Meng, Ziyan Jiang, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025. Think then embed: Generative context improves multimodal embedding. *arXiv preprint arXiv:2510.05014*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Yuntian Deng, Yejin Choi, and Samuel Bowman. 2024. Explicit CoT training for implicit CoT learning. *arXiv preprint arXiv:2311.01460*.

Manuel Faysse, Hugues Sibille, Tony Monet, Elia Music, Martin Music, Benjamin Musik, and Lucas Musik. 2025. ColPali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations (ICLR)*. OpenReview.net.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Sharan Narang. 2024. Think before you speak: Training language models with pause tokens. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. 2025a. Breaking the modality barrier: Universal embedding learning with multimodal LLMs. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM)*, Dublin, Ireland.

Tiancheng Gu, Kaicheng Yang, and 1 others. 2025b. UniME-V2: MLLM-as-a-judge for universal multimodal embedding. *arXiv preprint arXiv:2510.13515*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2025. Training large language models to reason in a continuous latent space. In *The Thirteenth International Conference on Learning Representations (ICLR)*. OpenReview.net.

Haoxiang He, Haoxiang Zhao, and 1 others. 2026. PLUME: Latent reasoning based universal multimodal embedding. *arXiv preprint arXiv:2604.02073*.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Greber, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Yingrui Ji and 1 others. 2025. CIBR: Cross-modal information bottleneck regularization for robust CLIP generalization. In *Artificial Neural Networks and Machine Learning – ICANN 2025*.

Wenhao Jian, Fan Zhang, and 1 others. 2025. RzenEmbed: Towards comprehensive multimodal retrieval. *arXiv preprint arXiv:2510.27350*.

Haonan Jiang, Yuji Wang, Yongjie Zhu, Xin Lu, Wenyu Qin, Meng Wang, Pengfei Wan, and Yansong Tang. 2026. Embed-RL: Reinforcement learning for reasoning-driven multimodal embeddings. *arXiv preprint arXiv:2602.13823*.

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

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2025. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations (ICLR)*. OpenReview.net.

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

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

Jin Li, Yaoming Wang, Xiaopeng Zhang, Dongsheng Jiang, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Qi Tian. 2025. Contrastive learning via variational information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7410–7427.

Sheng-Chieh Lin, Chankyu Mei, Barlas Oguz, Chenyan Xiong, Wen-tau Yih, and Xilun Xia. 2024. MM-Embed: Universal multimodal retrieval with multimodal LLMs. *arXiv preprint arXiv:2411.02571*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Weihao Liu, Dehai Min, and Lu Cheng. 2026. Latent thoughts tuning: Bridging context and reasoning with fused information in latent tokens. *arXiv preprint arXiv:2602.10229*.

Yifan Liu, Feng Liang, and 1 others. 2025a. ReMatch: Boosting representation through matching for multimodal retrieval. *arXiv preprint arXiv:2511.19278*.

Yifan Liu, Xinyi Yang, and 1 others. 2025b. Reasoning guided embeddings: Leveraging MLLM reasoning for multimodal embedding. *arXiv preprint arXiv:2511.16150*.

Yikun Liu and 1 others. 2025c. LamRA: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jinghui Lu and 1 others. 2025. Think silently, think fast: Dynamic latent compression of LLM reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xinyin Ma, Guangnian Jiang, Yifei Gong, and Zhiyuan Hu. 2025. COT-Valve: Length-compressible chain-of-thought tuning. *arXiv preprint arXiv:2502.09601*.

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. 2025. VLM2Vec-V2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Zhenyi Shen and others Yan. 2025. CODI: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: Improved training techniques for CLIP at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Siyu Sun, Jing Ren, Zhaohe Liao, Dongxiao Mao, Xiangyuan Ren, Yiyi Zhang, Haohua Zhao, Weixiong Lin, Jiang Shaohua, Liqing Zhang, and Yuchao Zheng. 2026. Bottleneck tokens for unified multimodal retrieval. *arXiv preprint arXiv:2604.11095*.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Cheng Chang, Jingren Yu, and Junyang Lin. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

Jinpeng Wu, Rui Mei, and 1 others. 2026. Beyond chain-of-thought: Rewrite as a universal interface for multimodal embedding. *arXiv preprint arXiv:2604.22280*.

Qilong Wu, Yiyang Shao, Jun Wang, and Xiaobo Sun. 2025. Learning optimal multimodal information bottleneck representations. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Zilin Xiao, Qi Ma, Mengting Gu, Chun-cheng Jason Chen, Xintao Chen, Vicente Ordonez, and Vijai Mohan. 2026. MetaEmbed: Scaling multimodal retrieval at test time. *The Thirteenth International Conference on Learning Representations (ICLR)*.

Heming Xing and 1 others. 2025. TokenSkip: Controllable chain-of-thought compression in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025. SoftCoT: Soft chain-of-thought for efficient reasoning with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations (ICLR)*. OpenReview.net.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

Jintian Zhang, Yulin Chen, Ningyu Wang, and Huajun Zhang. 2025a. LightThinker: Thinking step-by-step compression. *arXiv preprint arXiv:2502.15589*.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, and Yiming Yang. 2025b. Direct preference optimization of video large multimodal models from language model reward. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*.

Siyue Zhang, Yuan Gao, Xiao Zhou, Yilun Zhao, Tingyu Song, Arman Cohan, Anh Tuan Luu, and Chen Zhao. 2025c. MRMR: A realistic and expert-level multidisciplinary benchmark for reasoning-intensive multimodal retrieval. *arXiv preprint arXiv:2510.09510*.

Xin Zhang, Zehan Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2025d. GME: Improving universal multimodal retrieval by multimodal LLMs. *arXiv preprint arXiv:2412.16855*.

Yifei Zhang and 1 others. 2025e. SIM-CoT: Supervised implicit chain-of-thought. *arXiv preprint arXiv:2509.20317*.

Ouyang Zhuo, Hao Hu, and 1 others. 2025. Projection head is secretly an information bottleneck. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Prompt for Latent Decoder

```
<|im_start|>system
You are a helpful assistant.
<|im_end|>
<|im_start|>user
Please decode the following latent tokens back into the answer.
Now decode the latent representations below:
Latent representations:
<|im_end|>
<|im_start|>assistant
```

Figure 5: The prompt used by the latent decoder to reconstruct the target answer.

A More Training Details

A.1 Prompts for Latent Decoder

As shown in Figure 5, the decode prompt s follows a standard Qwen chat template: projected latent representations $\mathbf{h}_r^{1:K_r}$ are inserted after the reserved "Latent representations:" placeholder, and a lightweight decoder ψ then autoregressively decodes the target answer y. This prompt is used only during training.

A.2 Training Data Composition

The training data comprises approximately **1.55 million** training pairs drawn from the VLM2VEC-v2 (Meng et al., 2025) training corpus, spanning three modalities: images, videos, and visual documents. As shown in Table 7, we report the detailed per-dataset statistics. For image data, we adopt the image datasets in VLM2VEC-v2 including MSCOCO, ImageNet-1K, ChartQA, DocVQA, CIRR, and others across image-text matching, VQA, and retrieval scenarios ($\sim$677K samples after filtering). For video data, we use LLaVA-Hound (Zhang et al., 2025b) with video QA, retrieval, and caption subsets ($\sim$744K samples). For visual documents, we employ ViDoRe (Faysse et al., 2025) and VisRAG (Yu et al., 2025) covering complex charts, tables, and figures ($\sim$131K samples). Notably, the decoder head reuses the answer fields from existing cold-start CoT datasets (Jiang et al., 2026) as weak supervision targets, completely avoiding dependency on CoT annotations.

B More Experiments and Analysis

B.1 Ablation on Decoding Strategy

As shown in Table 8, Only-Answer achieves the best results (69.3 / 44.5 / 72.1), surpassing both CoT decoding and No Decoder (both 63.8 avg).

Table 7: Statistics of training data composition.

Dataset	Initial	Filtered	Ratio	Modality
Image-based (MMEB-train)				
A-OKVQA	50,000	34,750	69.50%	Text-Image Text
CIRR	50,000	31,950	63.90%	Text-Image Text-Image
ChartQA	50,000	35,900	71.80%	Text-Image Text
DocVQA	50,000	43,050	86.10%	Text-Image Text
HatefulMemes	25,500	15,150	59.41%	Text-Image Text
ImageNet-1K	50,000	40,200	80.40%	Text-Image Text
InfographicsVQA	50,000	36,850	73.70%	Text-Image Text
MSCOCO	50,000	23,800	47.60%	Text-Image Text-Image
MSCOCO-i2t	50,000	42,300	84.60%	Text-Image Text
MSCOCO-t2i	50,000	39,300	78.60%	Text Text-Image
N24News	50,000	27,700	55.40%	Text-Image Text
NIGHTS	47,823	39,300	82.17%	Text-Image Text-Image
OK-VQA	27,027	18,150	67.16%	Text-Image Text
SUN397	50,000	41,700	83.40%	Text-Image Text
VOC2007	23,532	18,600	79.05%	Text-Image Text
Visual7W	50,000	37,950	75.90%	Text-Image Text
VisDial	50,000	31,500	63.00%	Text Text-Image
VisualNews-i2t	50,000	31,300	62.60%	Text-Image Text
VisualNews-t2i	50,000	26,000	52.00%	Text Text-Image
WebQA	50,000	39,900	79.80%	Text Text-Image
Video-based (LLaVA-Hound)				
Caption Retrieval	300,000	258,200	86.07%	Video Text
Video QA	300,000	249,200	83.07%	Video-Text Text
Video Retrieval	300,000	236,900	78.97%	Text Video
Document-based				
ViDoRe	100,000	76,600	76.60%	Text-Image Text
VisRAG	100,000	54,850	54.85%	Text Image
Image-based	1,123,882	677,350	60.27%	Image-centric
Video-based	900,000	744,300	82.70%	Video-centric
Document-based	200,000	131,450	65.72%	Document-centric
Total	**2,223,882**	**1,553,100**	**69.84%**	Multimodal

Table 8: Ablation on decoding strategies on MMEB-V2.

Decoding Strategy	Image	Video	VisDoc	All
No Decoder (Emb. Only)	68.5	43.9	71.3	63.8
Only-Answer	**69.3**	**44.5**	**72.1**	**64.4**
Only-Thinking	68.4	43.5	71.4	63.7
Chain-of-Thought	68.4	44.1	71.4	63.8

This highlights a critical issue: in long CoT sequences, abundant easy-to-predict tokens (e.g., function words) rapidly saturate the LM loss and dominate gradients, thereby weakening IB supervision. The bottleneck collapses into a generic average representation suited for trivial prediction but lacking discriminative power. Only-Answer avoids this by restricting supervision to concise, informative answer tokens, preserving IB's role in encoding discriminative cues. No Decoder's degradation further confirms the decoder provides essential training-time reasoning supervision despite being discarded at inference.

B.2 Reason *vs.* Final Retrieval Embedding

LaME produces two types of embeddings: the reason embedding $\mathbf{e}$, derived from the reason tokens $\mathbf{h}_r^{K_r+1:K}$, and the final retrieval embedding from the [EMBED] token. We compare them across all modalities in Table 9. The reason embedding achieves a competitive 64.1 overall score, trailing the final retrieval embedding by merely 0.4 points. This small gap indicates that $\mathcal{L}_{\text{Emb}}$ effectively trains the reason tokens to encode retrieval-relevant infor-

Table 9: Comparison of reason embedding and final retrieval embedding on MMEB-V2.

Embedding Source	Image	Video	VisDoc	All
Reason	69.1	44.1	71.2	64.0
Final Retrieval	**69.3**	**44.5**	**72.1**	**64.4**

Table 10: Effect of the diversity regularizer weight λ_{Div} on MMEB-V2.

λ_{Div}	Image	Video	VisDoc	All
0.00	68.2	43.5	70.8	63.4
0.02	69.0	44.1	71.6	64.1
0.05	**69.3**	**44.5**	72.1	**64.4**
0.10	69.0	43.9	**72.5**	64.2

mation. Notably, the gap widens to 0.9 on VisDoc, compared to 0.2 on Image and 0.4 on Video, suggesting that fine-grained document understanding benefits more from the [EMBED] token's access to the complete hidden state sequence, whereas compressed reason tokens may lose granular details required for dense visual-text alignment.

B.3 Ablation on the Diversity Regularizer

Table 10 reports the results with varying diversity regularizer weight λ_{Div}. Removing the regularizer ($\lambda_{\text{Div}} = 0$) yields the lowest score of 63.4. Performance improves consistently as λ_{Div} increases from 0 to 0.05, reaching the best overall result of 64.4. These results confirm that the diversity regularizer is essential to prevent the reason tokens from *collapsing* into degenerate representations during training, thereby preserving the expressiveness of the latent reasoning process.

C Theoretical Analysis of Information Bottleneck

We provide a more formal treatment of the information bottleneck formulation in LaME. Let X denote the input (interleaved text and visual tokens), Z the representation carried by the reason token hidden states $\mathbf{h}_r$, and Y the supervision signal from both decoder and embedding heads. The IB objective is:

$$\min_{p(z|x)} \; I(Z;X) - \beta\, I(Z;Y), \tag{8}$$

where $\beta > 0$ trades off compression against predictive power.

In LaME, the capacity constraint is structural rather than variational: the representation Z is supported on at most K tokens of dimensionality d, imposing a hard information ceiling $I(Z;X) \leq K \cdot d \cdot \log(1 + \text{SNR})$, where SNR denotes the effective signal-to-noise ratio of the hidden states. This is distinct from variational IB (Alemi et al., 2017), which replaces $I(Z;X)$ with a KL-divergence upper bound requiring an auxiliary prior and Monte Carlo estimation; here, the token count K itself serves as the bottleneck radius.

The dual-head supervision provides the predictive term $I(Z;Y)$ via two complementary channels: the decoder head maximizes $I(Z_r; Y_{\text{dec}})$ through autoregressive reconstruction on the first K_r tokens, while the embedding head maximizes $I(Z_e; Y_{\text{emb}})$ through contrastive learning on the remaining K_e tokens. Since both heads share the same bottleneck, the finite capacity enforces a joint compression trade-off between reconstructive and discriminative information. This prevents degeneration into either a reconstructive autoencoder that ignores retrieval structure or a collapsed embedding that discards fine-grained semantics.

The two-stage training can be interpreted as progressively tightening the bottleneck: Stage 1 (frozen backbone) allows the reason tokens to discover a compact representation without altering the backbone, while Stage 2 (joint optimization) enables the backbone to co-adapt, routing task-relevant information through the reason tokens while preserving encoding capacity in the [EMBED] token path. This staged curriculum avoids the trivial solution where the backbone bypasses the bottleneck entirely, which occurs under end-to-end training as shown in Table 4.