REASON WHAT MATTERS: RETRIEVAL-GROUNDED REASONING FOR UNIVERSAL MULTIMODAL EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Universal multimodal embedding (UME) maps multimodal inputs into a shared embedding space for diverse retrieval tasks. Recent methods improve embeddings through Chain-of-Thought (CoT) reasoning optimized with GRPO using retrieval rewards. However, existing methods overlook the mismatch between candidate-aware retrieval supervision and input-only CoT generation: (1) trajectory-level rewards convey retrieval outcomes without explicitly identifying the input-supported evidence that distinguishes the positive from hard negatives; (2) input-only generation cannot directly assess whether further reasoning improves retrieval, potentially producing redundant CoTs with substantial latency. To bridge this gap, we propose **Re**ason **W**h**a**t **M**atters (**ReWAM**), a retrieval-grounded framework that aligns candidate-aware supervision with input-only generation. Specifically, we introduce Retrieval-Aware Self-Distillation (RASD), which extracts privileged guidance from input-supported facts and evidence distinguishing the positive from hard negatives. Conditioned on this guidance, an on-policy self-teacher provides token-level feedback to refine credit assignment, directing policy updates toward retrieval-relevant reasoning grounded in the input. We further propose Retrieval-Adaptive Inference (RAI), which learns a retrieval-aware stopping criterion from prefix-level retrieval feedback. It stops redundant reasoning without candidate access and uses speculative decoding to further reduce CoT latency. Extensive experiments on MMEB-V2 and MRMR demonstrate that ReWAM achieves state-of-the-art retrieval performance while delivering up to $\mathbf{5}\times$ the inference throughput of competitive explicit-CoT UME methods. ReWAM thus enables high-quality retrieval through efficient input-only reasoning, making explicit CoT practical for corpus-scale multimodal retrieval. *The code will be publicly available.*

1 INTRODUCTION

Building on vision-language representation learning (Radford et al., 2021; Li et al., 2022; Zhai et al., 2023), universal multimodal embedding (UME) supports heterogeneous retrieval across modalities and tasks with a single model and a shared embedding space. Recent methods adapt multimodal large language models (MLLMs) into unified embedders through contrastive learning (Jiang et al., 2024b; Zhang et al., 2025; Meng et al., 2025). These models typically derive embeddings from final-layer hidden states in a fixed-depth forward pass. Although efficient and scalable, this design fails to fully leverage the reasoning capabilities of MLLMs, limiting performance on tasks that require compositional relations, fine-grained evidence, or multi-step visual understanding.

Recently, some studies have extended UME by incorporating Chain-of-Thought (CoT) reasoning into embedding generation, which we refer to as *Reasoning-enhanced embeddings*. As illustrated in Fig. 1, prior methods generally fall into two paradigms. Generative reasoning UME (Lan et al., 2026; Wu et al., 2026b) uses a shared MLLM backbone for both reasoning and embedding generation. However, this paradigm often yields suboptimal results due to potential gradient conflicts between the next-token prediction and contrastive learning objectives (Cui et al., 2026). Decoupled reasoning UME (Cui et al., 2026; Jiang et al., 2026; Zhang et al., 2026a) separates the reasoning and embedding components, conditioning the embedder on both the original input and the generated CoT. Despite the extra parameters and computation, decoupling mitigates conflicts between training objectives and often leads to better retrieval performance (Jiang et al., 2026). Recent methods in

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.

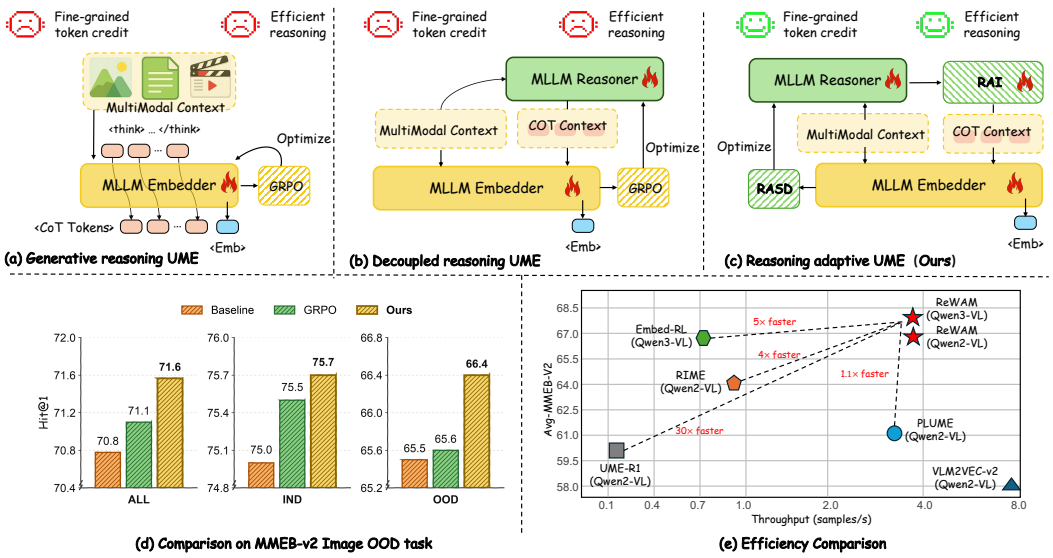


Figure 1: Overview of reasoning-enhanced UME paradigms and empirical comparisons. (a)–(c) show Generative reasoning UME, Decoupled reasoning UME, and ReWAM, respectively. (d) compares in-distribution and OOD retrieval, showing smaller gains from GRPO on OOD tasks. (e) compares retrieval performance and throughput across UME methods.

both paradigms further optimize CoT generation with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), using retrieval rewards to encourage retrieval-relevant reasoning.

Despite these advances, existing methods generate CoTs from the input alone to support independent encoding and reuse of corpus representations, but assess their retrieval value through comparisons with positive and negative candidates. Input-supported facts may not distinguish confusable candidates, while seemingly discriminative claims may lack input support. Existing methods overlook this mismatch between candidate-aware supervision and input-only reasoning: (1) Trajectory-level rewards reflect the retrieval outcome of a complete CoT without identifying which claims are both input-supported and discriminative. Applying the same advantage to all tokens can therefore reinforce incorrect or irrelevant claims alongside useful reasoning. As shown in Fig. 1(d), GRPO improves in-distribution retrieval but provides only marginal gains on out-of-distribution (OOD) tasks. (2) A shorter COT prefix may already contain sufficient retrieval evidence, but independent generation cannot directly compare candidates to assess the benefit of continuing. As a result, the model may continue after obtaining sufficient evidence, adding redundant computation and latency.

To address these limitations, we propose **Re**ason **Wh**a**t** **M**atters (**ReWAM**), a retrieval-grounded framework that aligns candidate-aware retrieval supervision with input-only reasoning, as illustrated in Fig. 1(c). Specifically, we introduce Retrieval-Aware Self-Distillation (RASD), which uses a multimodal analyzer to extract input-supported facts and distinctions between the positive and retrieved hard negatives. The analyzer uses this evidence to identify supported and conflicting CoT claims, constructing privileged guidance for an on-policy self-teacher. Conditioned on this guidance, the teacher provides token-specific feedback to modulate trajectory advantages and focus policy updates on input-grounded reasoning that distinguishes confusable candidates. We further introduce Retrieval-Adaptive Inference (RAI), which learns a retrieval-aware stopping criterion from prefix-level candidate comparisons for input-only generation. RAI estimates remaining retrieval utility from the current reasoning state alone to truncate redundant reasoning, while speculative decoding reduces generation cost without sacrificing the retrieval benefits of explicit CoT.

Extensive experiments on MMEB-V2 and MRMR demonstrate that ReWAM achieves state-of-the-art retrieval performance with substantially improved inference efficiency. As shown in Fig. 1(e), ReWAM outperforms the decoupled explicit-CoT method Embed-RL (Jiang et al., 2026) while delivering up to $5\times$ its throughput. It even surpasses the efficiency-oriented latent-reasoning method PLUME (He et al., 2026) in both retrieval quality and throughput while retaining explicit CoT.

These results show that aligning candidate-aware supervision with input-only reasoning enables high-quality retrieval and efficient encoding at corpus scale. Our main contributions are:

- We propose ReWAM to align candidate-aware retrieval supervision with input-only CoT generation. This guides reasoning toward input-supported evidence that distinguishes relevant targets, improving CoT-based retrieval while preserving independent query and target encoding.
- We introduce RASD for token-level credit assignment through retrieval evidence verification, and RAI for adaptive CoT truncation guided by prefix-level remaining retrieval utility.
- Extensive experiments on MMEB-V2 and MRMR demonstrate state-of-the-art retrieval performance with up to $5\times$ the throughput of prior decoupled explicit-CoT methods.

2 Related Work

Universal Multimodal Embedding. Contrastive vision–language pretraining learns to align representations across images and text (Radford et al., 2021; Li et al., 2022; Zhai et al., 2023). Building on this foundation, UME methods adapt MLLMs into general-purpose encoders and explore diverse architectures and training objectives for multimodal retrieval (Jiang et al., 2024b; Zhang et al., 2025; Meng et al., 2025; Faysse et al., 2025; Liu et al., 2025; Chen et al., 2026b; Yu et al., 2025). More recent studies integrate reasoning into representation learning to better capture compositional relations and fine-grained cross-modal cues (Lan et al., 2026; Wu et al., 2026b; Cui et al., 2026; Jiang et al., 2026; He et al., 2026; Tsai et al., 2026).

On-Policy Self-Distillation. Outcome-based RL provides verifiable feedback, yet sparse trajectory rewards do not reveal the contribution of intermediate decisions (Shao et al., 2024). On-policy self-distillation addresses this gap by conditioning the policy on privileged information to supervise its own rollouts at the token level (Zhao et al., 2026; Hübotter et al., 2026; Yang et al., 2026a; Song & Zheng, 2026). Later methods refine these signals through confidence weighting, cross-rollout reflection, skill guidance, and self-evolving supervision (Zheng et al., 2026a;b; Yang et al., 2026b; Lu et al., 2026; Wu et al., 2026a).

Inference Acceleration. Autoregressive decoding generates tokens sequentially, making long reasoning traces costly at inference time. Speculative decoding reduces this cost by verifying multiple draft tokens in parallel (Leviathan et al., 2023), while multi-token prediction and multi-head decoding predict several future tokens per step (Gloeckle et al., 2024; Cai et al., 2024). Block-diffusion and semi-autoregressive methods further increase decoding parallelism (Chen et al., 2026a; Cheng et al., 2026). Complementary methods reduce reasoning length through efficient input routing (Zhang et al., 2026a) or task-agnostic semantic redundancy detection (Sun et al., 2026).

3 Method

3.1 Preliminaries

Universal multimodal retrieval aims to identify relevant targets for queries across text, images, videos, and visual documents. To learn discriminative multimodal embeddings, we adopt contrastive learning with the InfoNCE loss. For a batch of B pairs, we optimize the model to maximize similarity between each query q_i and its positive target t_i^+ while minimizing similarities to all negative targets in $\mathcal{T}^- = \{t_j^- \mid j \neq i\}$:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\cos(\mathbf{e}_{q_i}, \mathbf{e}_{t_i^+})/\tau_E)}{\exp(\cos(\mathbf{e}_{q_i}, \mathbf{e}_{t_i^+})/\tau_E) + \sum_{t_j^- \in \mathcal{T}^-}\exp(\cos(\mathbf{e}_{q_i}, \mathbf{e}_{t_j^-})/\tau_E)}. \quad (1)$$

Here $\mathbf{e}_{q_i}$ and $\mathbf{e}_t$ denote ℓ_2-normalized query and target embeddings, extracted from the final-layer hidden states of the MLLM embedder conditioned on the multimodal input and its corresponding CoT. $\cos(\cdot,\cdot)$ denotes cosine similarity, and τ_E is the temperature. We compute the InfoNCE loss in both query-to-target and target-to-query directions and sum the two losses to obtain $\mathcal{L}_{\text{con}}$.

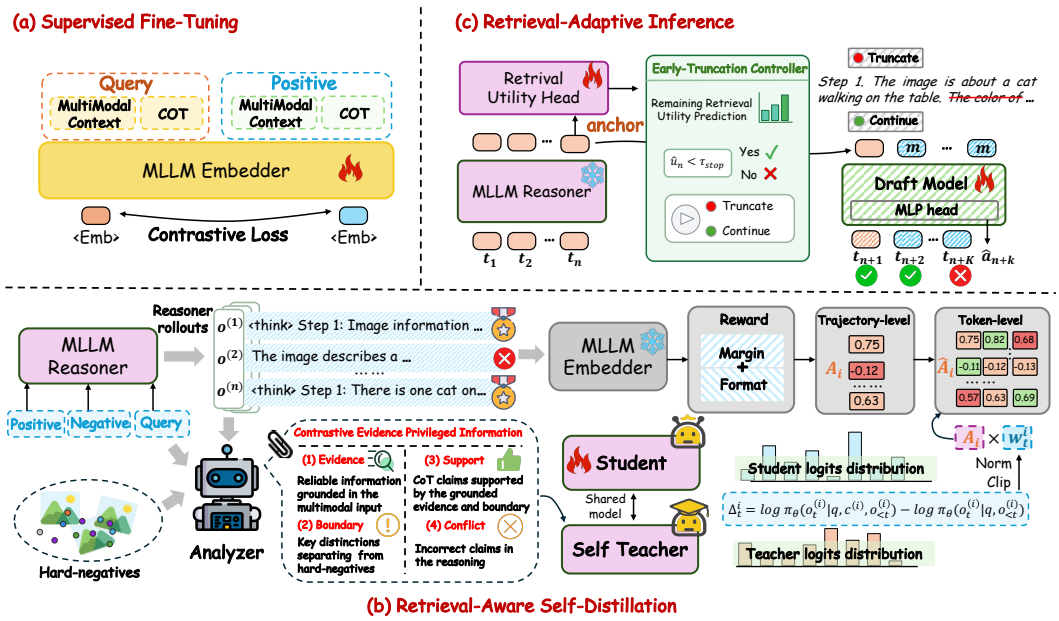


Figure 2: Overview of ReWAM. (a) Supervised fine-tuning trains a CoT-conditioned embedder. (b) RASD uses privileged same-token rescoring to derive token-wise weights that modulate trajectory-level retrieval advantages. (c) RAI stops a partial CoT when its predicted remaining retrieval utility is low and accelerates generation through speculative decoding.

3.2 Supervised Fine-Tuning

As shown in Fig. 2(a), the embedder E_ϕ encodes each query and target together with an offline CoT and is fine-tuned using the bidirectional contrastive objective $\mathcal{L}_{\text{con}}$. No language-modeling loss is applied, keeping representation learning separate from autoregressive generation. We then freeze E_ϕ to evaluate and optimize reasoner rollouts with a stable retrieval criterion.

3.3 Retrieval-Aware Self-Distillation

To address coarse credit assignment in retrieval-based RL, RASD constructs privileged guidance by contrasting the positive with hard negatives (Fig. 2(b)). An on-policy self-teacher uses this guidance to derive token-level feedback for reweighting trajectory advantages. The resulting supervision focuses policy updates on reasoning that distinguishes the target from confusable candidates.

Retrieval-based reward function. Specifically, for each query q, the reasoner policy π_θ samples a group of G CoTs $\{o^{(i)}\}_{i=1}^{G}$. We apply RASD symmetrically to targets, with paired queries as positives and other in-batch queries as negatives. Absolute retrieval performance alone is insufficient to assess the contribution of CoT, as embeddings obtained without CoT may already be highly discriminative. To quantify this contribution, we define a Margin Reward that estimates the margin improvement over a no-CoT embedding baseline on the same candidates. For the positive target t^+ and each in-batch negative $t^- \in \mathcal{T}^-$, the reasoner likewise samples G CoTs independently. The frozen embedder produces the ℓ_2-normalized query embedding $\mathbf{e}_q^i = E_\phi(q, o^{(i)})$ and the corresponding positive and negative target embeddings $\mathbf{e}_{t^+}^j$ and $\mathbf{e}_{t^-}^j$. The Margin Reward compares $M(o^{(i)})$ with $M(\emptyset)$, the margin obtained with the embedder conditioned on multimodal inputs alone:

$$M(o^{(i)}) = \mathbb{E}_j\left[\cos(\mathbf{e}_q^i, \mathbf{e}_{t^+}^j)\right] - \mathbb{E}_{t^-,j}\left[\cos(\mathbf{e}_q^i, \mathbf{e}_{t^-}^j)\right],$$

$$R_{\text{margin}}(o^{(i)}) = \begin{cases} +1, & M(o^{(i)}) - M(\emptyset) > \epsilon_v, \\ 0, & |M(o^{(i)}) - M(\emptyset)| \le \epsilon_v, \\ -1, & M(o^{(i)}) - M(\emptyset) < -\epsilon_v, \end{cases} \tag{2}$$

Here $\epsilon_v = 0.01$ suppresses minor embedding fluctuations. Using the no-CoT margin as an anchor, this reward encourages more discriminative reasoning. The Format reward $R_{\text{format}}(o^{(i)})$ is 1 for CoT enclosed in `<think>` and `</think>`, and 0 otherwise. The trajectory-level advantage is:

$$r_i = R_{\text{margin}}(o^{(i)}) + \lambda_f R_{\text{format}}(o^{(i)}), \qquad A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\})}. \tag{3}$$

Privileged guidance construction. The reasoner generates each CoT from the input alone, whereas A_i summarizes its retrieval feedback from candidate comparisons. This trajectory-level signal does not identify which claims are supported by the input and distinguish the positive from hard negatives. Applying A_i to all tokens can therefore reinforce unsupported or irrelevant claims alongside useful reasoning. To provide this missing guidance, RASD uses a multimodal analyzer $\mathcal{A}$ to construct Contrastive Evidence Privileged Information (CEPI):

$$c^{(i)} = \mathcal{A}(q, t^+, h^-, o^{(i)}) = \big(c^{\text{evi}}, c^{\text{bnd}}, c^{\text{sup}}, c^{\text{con}}\big), \tag{4}$$

where $h^- \subseteq \mathcal{T}^-$ contains the *top-3* in-batch hard negatives retrieved by the frozen embedder without CoT. The analyzer extracts input-grounded facts c^{evi} and positive–negative decision boundaries c^{bnd}, then checks $o^{(i)}$ against both to identify supported claims c^{sup} and conflicting claims c^{con}.

Token-level credit assignment. Conditioned on $c^{(i)}$, the same reasoner acts as an on-policy self-teacher and rescores $o^{(i)}$ without generating a new CoT. For the t-th token $o_t^{(i)}$ with prefix $o_{<t}^{(i)}$, we estimate teacher–student divergence using their log-probability difference:

$$\Delta_t^i = \log \pi_\theta(o_t^{(i)} \mid q, c^{(i)}, o_{<t}^{(i)}) - \log \pi_\theta(o_t^{(i)} \mid q, o_{<t}^{(i)}), \quad s_t^i = \exp\big(\text{sign}(A_i)\Delta_t^i\big). \tag{5}$$

The likelihood difference Δ_t^i measures how privileged evidence changes the policy's support for each sampled token, while s_t^i exponentiates this feedback after alignment with the advantage sign. Treating Δ_t^i as a stop-gradient signal for token-level credit assignment, RASD derives bounded weights w_t^i to modulate the trajectory advantage A_i:

$$w_t^i = 1 + \lambda_R \,\text{clip}\big(s_t^i - 1, -\epsilon_R, \epsilon_R\big), \qquad \hat{A}_t^i = A_i w_t^i, \tag{6}$$

where λ_R controls the strength of token-level modulation, and ϵ_R is the clipping threshold that bounds the adjustment. The modulated advantage $\hat{A}_t^i$ strengthens reinforcement of evidence-supported tokens when $A_i > 0$ and reduces their penalties when $A_i < 0$. The positive weights preserve the sign of each trajectory's advantage. The RASD objective is:

$$\mathcal{L}_{\text{RASD}} = -\mathbb{E}_{i,t}\Big[\min\Big(\rho_t^i \hat{A}_t^i, \text{clip}(\rho_t^i, 1-\epsilon_P, 1+\epsilon_P)\hat{A}_t^i\Big)\Big] + \beta\, D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \tag{7}$$

where $\rho_t^i = \pi_\theta(o_t^{(i)} \mid q, o_{<t}^{(i)}) / \pi_{\text{old}}(o_t^{(i)} \mid q, o_{<t}^{(i)})$ is the per-token importance sampling ratio. ϵ_P is the clipping threshold. The KL term D_{KL} penalizes deviation from the reference policy π_{ref}. This objective encourages reasoning grounded in input evidence that distinguishes confusable candidates.

3.4 Retrieval-Adaptive Inference

A CoT prefix may already contain sufficient retrieval evidence, yet input-only generation cannot directly compare candidates to assess the value of continuing. This can lead to redundant reasoning and increased latency. To reduce this overhead, we introduce Retrieval-Adaptive Inference (RAI), combining a Retrieval Utility Head trained on prefix-level retrieval feedback for early stopping with speculative decoding for faster generation (Fig. 2(c)).

Retrieval-aware early stopping. To learn a retrieval-grounded stopping criterion, we quantify the retrieval gains still available beyond each CoT prefix. For each query q, the frozen reasoner performs one greedy rollout to obtain a complete CoT o of length L. Each prefix $o_{\le n}$ ($n \le L$) contains the first n tokens of o and is encoded as $\mathbf{e}_q^n = E_\phi(q, o_{\le n})$. The candidate embeddings $\mathbf{e}_{t^+}$ and $\mathbf{e}_{t^-}$ remain fixed and are conditioned on their own full CoTs. The no-CoT margin m_0 uses $E_\phi(q)$ against these same candidates. The prefix margin and remaining retrieval utility are:

$$\begin{aligned} m_n &= \cos(\mathbf{e}_q^n, \mathbf{e}_{t^+}) - \mathbb{E}_{t^- \in \mathcal{T}^-}\big[\cos(\mathbf{e}_q^n, \mathbf{e}_{t^-})\big], \\ g_n &= \max\{m_n - m_0, 0\}, \qquad u_n = \frac{\max\{g_k \mid n \le k \le L\} - g_n}{\max\{g_k \mid 1 \le k \le L\}}. \end{aligned} \tag{8}$$

With k indexing evaluated prefix lengths, the numerator measures the largest additional retrieval gain available beyond the current prefix, while the denominator is the maximum gain over the no-CoT baseline across the rollout. Their ratio u_n quantifies the relative benefit of continued reasoning, with values near zero indicating little further gain. To predict this utility during generation, we use a Retrieval Utility Head H_ω, a two-hidden-layer MLP with sigmoid output, that maps the frozen reasoner's final-layer state h_n after the n-th token to $\hat{u}_n$. We train the head using the offline utility labels u_n with a SmoothL1 loss:

$$\hat{u}_n = H_\omega(h_n), \qquad \mathcal{L}_H = \mathbb{E}_{(h_n,u_n)}[\mathrm{SmoothL1}(\hat{u}_n, u_n)]\,. \tag{9}$$

At inference, generation stops when $\hat{u}_n < \tau_{\mathrm{stop}}$, using the current hidden state without candidate comparisons or embedder calls. See Appendix B.3 for prefix sampling and decoding details.

Speculative decoding. To accelerate CoT generation, we adopt speculative decoding with a semi-autoregressive Draft model (Chen et al., 2026a; Cheng et al., 2026), using the frozen reasoner as the target model to verify draft tokens. As shown in Fig. 2(c), conditioned on multimodal context features F_n extracted from the frozen reasoner, the Draft backbone D processes the last prefix token t_n and $K-1$ mask tokens m in parallel. A low-rank Markov head adds a correction based on the preceding token:

$$\begin{aligned}(z_{n+1},\dots,z_{n+K}) &= D([t_n, m, \dots, m]; F_n),\\ q_{n+k} &= \mathrm{softmax}(W_{\mathrm{LM}} z_{n+k} + W_2 W_1(t_{n+k-1}))\,, \quad k = 1,\dots,K.\end{aligned} \tag{10}$$

Here q_{n+k} is the Draft model's vocabulary distribution for t_{n+k}, and z_{n+k} is its prediction state, mapped to base logits by the frozen W_{LM}. Using the preceding token t_{n+k-1}, W_2W_1 provides a logit correction for predicting the next token t_{n+k}. The backbone computes all K hidden states in parallel, while the lightweight Markov head generates tokens sequentially.

To learn proposals consistent with the frozen reasoner, we train on offline CoTs using reference tokens $t^\star_{n+k}$ and target distributions p_{n+k} under the same prefix. Following Cheng et al. (2026), we combine next token prediction, distribution alignment, and acceptance estimation:

$$\mathcal{L}_D = \mathbb{E}_{n,k}\Big[\omega_k\Big(-\lambda_{\mathrm{ntp}} \log q_{n+k}(t^\star_{n+k}) + \lambda_{\mathrm{align}} \|q_{n+k} - p_{n+k}\|_1 + \lambda_{\mathrm{acc}} \ell^{\mathrm{acc}}_{n+k}\Big)\Big]. \tag{11}$$

The λ coefficients balance the losses, and unit-mean weights $\omega_k \propto e^{-k/\gamma}$ emphasize earlier proposals over valid training positions, with γ controlling the decay. The next-token prediction loss supervises reference continuations, while distribution alignment promotes overlap and hence acceptance under speculative sampling. The auxiliary acceptance loss is defined as:

$$a_{n+k} = 1 - \tfrac{1}{2}\|q_{n+k} - p_{n+k}\|_1, \quad \ell^{\mathrm{acc}}_{n+k} = \mathrm{BCE}(\hat{a}_{n+k}, a_{n+k}). \tag{12}$$

Here a_{n+k} is the Draft–target distribution overlap, equal to the expected acceptance probability under speculative sampling. An auxiliary sigmoid head predicts $\hat{a}_{n+k}$ from z_{n+k} and the preceding token. BCE denotes binary cross-entropy, which trains the head to match the soft target a_{n+k}.

4 Experiments

4.1 Experimental setup

Implementation details. We train CoT-conditioned embedders based on Qwen2-VL (2B and 7B) and Qwen3-VL (2B and 4B). Following Embed-RL (Jiang et al., 2026), we adopt Qwen3-VL-8B-Instruct as the reasoner. *For SFT,* we use the same training data and prompt templates as RIME (Wu et al., 2026b). The embedders are trained on 1.4M query–target pairs for one epoch with a learning rate of 5×10^{-5} and a batch size of 512. *For RASD,* we train on 20K query–target pairs with hard-negative mining, as described in Appendix A. We use a fixed Qwen3.5-122B-A10B API as the multimodal analyzer to construct privileged guidance. The reasoner is trained for one epoch with a learning rate of 1×10^{-6} and a batch size of 32 query–target pairs. We sample $G = 8$ CoTs per input and set $\epsilon_R = \epsilon_P = 0.2$, $\lambda_R = 0.5$, and KL weight $\beta = 0.01$. *For RAI,* the RASD-trained reasoner greedily generates 730K offline CoT rollouts from randomly sampled training inputs. We train the Retrieval Utility Head and Draft model for 10 epochs with a batch size of 128 and learning rates of 3×10^{-4} and 6×10^{-4}, respectively. Following Cheng et al. (2026), we set $\lambda_{\mathrm{ntp}} = 0.1$, $\lambda_{\mathrm{align}} = 0.9$, $\lambda_{\mathrm{acc}} = 1$, and $\gamma = 4$ in Eq. (11). We set the stopping threshold to $\tau_{\mathrm{stop}} = 0.01$.

Table 1: Results on the MMEB-V2 benchmark. Best and second-best scores are bolded and underlined, respectively. CLS: classification, QA: question answering, RET: retrieval, GD: grounding, MRET: moment retrieval, VDR: ViDoRe, VR: VisRAG, and OOD: out-of-distribution. We follow the evaluation protocol of VLM2Vec-V2.

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																
ColPali-V1.3 (PaliGemma-3B)	40.3	11.5	48.1	40.3	34.9	26.7	37.8	21.6	25.5	28.2	83.6	52.0	81.1	43.1	71.0	44.4
GME (Qwen2-VL-2B)	54.4	29.9	66.9	55.5	51.9	34.9	42.0	25.6	32.4	33.9	**86.1**	54.0	82.5	43.1	72.7	54.1
VLM2Vec (Qwen2-VL-2B)	58.7	49.3	65.0	72.9	59.7	33.4	30.5	20.6	33.0	29.0	49.8	13.5	51.8	33.5	41.6	47.0
VLM2Vec-V2 (Qwen2-VL-2B)	62.9	56.3	69.5	77.3	64.9	39.3	34.3	28.8	38.5	34.9	75.5	44.9	79.4	39.4	65.4	58.0
UME-R1 (Qwen2-VL-2B)	64.8	62.8	67.6	77.2	66.6	44.3	51.2	32.9	39.7	42.2	72.4	46.2	79.2	37.2	63.9	60.1
TTEs (Qwen2-VL-2B)	67.9	66.6	70.2	84.1	70.1	47.3	49.1	33.2	32.1	41.3	77.5	53.2	83.2	41.1	68.8	63.1
Embed-RL (Qwen3-VL-2B)	62.8	67.9	68.6	**90.4**	69.2	**57.0**	55.9	**45.1**	**49.4**	**52.1**	79.9	52.0	**84.6**	65.7	74.1	66.8
RIME (Qwen2-VL-2B)	67.9	64.4	69.8	82.1	69.1	48.0	52.1	33.6	39.2	43.7	76.4	51.4	81.7	63.9	71.4	64.1
PLUME (Qwen2-VL-2B)	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
UMER-R (Qwen2-VL-2B)	64.9	68.4	69.7	74.1	68.5	51.4	51.6	35.3	33.3	44.0	75.1	48.3	**84.6**	**67.9**	71.8	63.9
ReWAM (Qwen2-VL-2B)	68.8	70.3	**71.2**	82.7	71.6	53.0	61.3	36.2	35.4	47.7	77.4	51.8	84.2	65.2	72.8	66.5
ReWAM (Qwen3-VL-2B)	**68.9**	**70.7**	70.5	86.0	**71.9**	53.6	**63.3**	38.0	39.9	49.7	78.6	**55.4**	84.2	66.7	**74.2**	**67.4**
≥4B model size																
GME (Qwen2-VL-7B)	57.7	34.7	71.2	59.3	56.0	37.4	50.4	28.4	38.2	38.6	**89.4**	55.6	85.0	44.4	75.2	57.8
LamRA-2 (Qwen2-VL-7B)	59.2	26.5	70.0	62.7	54.1	39.3	42.6	24.3	34.6	35.2	22.0	11.5	37.4	21.0	23.9	40.4
LamRA-2.5 (Qwen2.5-VL-7B)	51.7	34.1	66.9	56.7	52.4	32.9	42.6	23.2	37.6	33.7	56.3	33.3	58.2	40.1	50.2	47.4
VLM2Vec (Qwen2-VL-7B)	62.7	56.9	69.4	82.2	65.5	39.1	30.0	29.0	40.6	34.0	56.9	9.4	59.1	38.1	46.4	52.3
VLM2Vec-V2 (Qwen2-VL-7B)	65.7	61.5	70.0	85.2	68.1	45.9	33.9	27.6	39.3	36.4	78.8	52.6	82.7	42.1	69.3	61.2
CAFe (LLaVA-OV-7B)	63.6	61.7	69.1	87.6	67.6	35.8	58.7	34.4	39.5	42.4	70.7	49.6	79.5	38.1	63.9	60.6
UME-R1 (Qwen2-VL-7B)	67.1	69.2	71.9	84.9	71.3	48.6	60.7	38.2	39.3	47.5	75.7	50.5	83.7	37.6	67.1	64.5
RIME (Qwen2-VL-7B)	**70.3**	71.7	73.2	86.3	73.4	52.6	62.0	38.4	41.6	49.4	80.9	55.6	**85.8**	66.9	75.6	68.6
Embed-RL (Qwen3-VL-4B)	63.7	70.5	71.3	**91.4**	70.1	**57.6**	58.4	**45.1**	**49.5**	**53.0**	80.2	53.4	84.9	67.1	74.7	68.1
ReWAM (Qwen3-VL-4B)	**70.3**	71.8	72.3	88.4	73.4	52.2	**63.2**	39.5	40.4	49.8	80.1	**58.2**	85.7	**68.1**	**75.8**	68.7
ReWAM (Qwen2-VL-7B)	**70.3**	**72.5**	**74.2**	87.8	**74.2**	53.6	62.6	38.6	38.6	49.5	80.9	56.1	85.7	67.3	75.7	**69.0**

Baselines. Our comparisons cover both standard multimodal embedders and reasoning-enhanced methods. Standard embedding baselines include EVA-CLIP (Sun et al., 2023), OpenCLIP (Cherti et al., 2023), VISTA (Zhou et al., 2024), E5-V (Jiang et al., 2024a), ColPali (Faysse et al., 2025), GME (Zhang et al., 2025), VLM2Vec and VLM2Vec-V2 (Jiang et al., 2024b; Meng et al., 2025), LamRA (Liu et al., 2025), and CAFe (Yu et al., 2025). Reasoning-enhanced baselines include UME-R1 (Lan et al., 2026), RIME (Wu et al., 2026b), TTEs (Cui et al., 2026), Embed-RL (Jiang et al., 2026), and UMER-R (Chen et al., 2026b). We also include PLUME (He et al., 2026) as a latent-reasoning baseline. On MRMR, we include MM-Embed (Lin et al., 2025), Ops-MM-Embed (OpenSearch-AI, 2025), and LaME (Wu et al., 2026c).

Evaluation. MMEB-V2 (Meng et al., 2025) covers 78 datasets for general-purpose multimodal retrieval, including 36 image, 18 video, and 24 visual-document (visdoc) datasets. MRMR (Zhang et al., 2026b) assesses reasoning-intensive retrieval across 11 subtasks in Knowledge, Theorem, and Contradiction. Following Wu et al. (2026b), we report Hit@1 for image and video tasks and NDCG@5 for visdoc tasks on MMEB-V2. For MRMR, we report NDCG@10 for all subtasks except Negation, which uses Hit@1. We measure inference speed on a single H800 GPU.

4.2 Main results

Universal multimodal retrieval. ReWAM achieves the best overall scores of 67.4 and 69.0 in the two model-size groups of Table 1, outperforming both standard and reasoning-enhanced multimodal embedding methods. With Qwen2-VL-2B, ReWAM surpasses VLM2Vec-V2 by 8.5 points overall. Using the same SFT data and prompt templates as RIME, it gains 2.4 points overall and 2.5, 4.0, and 1.4 points on image, video, and visdoc tasks, respectively. It also exceeds the latent-reasoning method PLUME by 4.9 points overall while retaining explicit CoT. ReWAM further outperforms Embed-RL by 0.6 points at both the 2B and 4B scales with matched embedding backbones. These comparisons suggest that retrieval-aware supervision improves the retrieval utility of explicit CoT.

Reasoning-intensive retrieval. ReWAM achieves the top two overall scores on MRMR (Table 2), with 53.3 and 51.4 for the 4B and 7B models, respectively. ReWAM-4B surpasses RIME, LaME, Embed-RL, and UME-R1 by 3.1, 3.5, 6.5, and 5.3 points, respectively. It leads on eight of the eleven subtasks. In Knowledge tasks, it achieves the best results in Art, Medicine, and Humanities. The gains are particularly pronounced in Theorem tasks, where it ranks first on all four subtasks,

Table 2: Results on the MRMR benchmark, covering Art, Medicine (Med.), Science (Sci.), Humanities (Hum.), Math, Physics (Phy.), Engineering (Eng.), Business (Bus.), Negation (Neg.), Design (Des.), and Traffic (Tra.). Best and second-best scores are bolded and underlined, respectively.

Model	Backbone	Size	Knowledge				Theorem				Contradiction			All
			Art	Med.	Sci.	Hum.	Math	Phy.	Eng.	Bus.	Neg.	Des.	Tra.	
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
Embed-RL	Qwen3-VL	4B	73.5	53.1	60.5	69.8	32.9	45.6	35.1	52.6	6.0	54.3	31.7	46.8
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
LaME	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
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
ReWAM	Qwen3-VL	4B	**79.9**	**60.0**	73.0	**74.8**	**40.2**	**52.2**	**43.1**	**56.0**	6.5	46.7	**53.5**	**53.3**
ReWAM	Qwen2-VL	7B	77.8	57.4	72.3	73.3	33.3	45.8	40.5	53.6	8.0	60.1	43.3	51.4

exceeding the strongest prior results by 7.3, 6.6, 6.7, and 3.4 points in Math, Physics, Engineering, and Business, respectively. In Contradiction tasks, it reaches 53.5 on Traffic, improving over the best prior result by 7.7 points. These gains support the value of guiding input-only reasoning toward input-grounded evidence that distinguishes relevant targets in reasoning-intensive retrieval.

Retrieval quality and inference efficiency. In Fig. 1(e), ReWAM achieves 3.75 *samples/s*, versus 0.75 for decoupled Embed-RL and 3.41 for latent-reasoning PLUME. ReWAM thus delivers approximately $5\times$ and $1.1\times$ their throughput, respectively, while retaining explicit CoT and outperforming PLUME on MMEB-V2 with the same embedding backbone.

4.3 Ablation studies and analysis

Ablation on Core Components. We compare ReWAM with three variants in Table 3. The *w/ raw input* variant trains an embedder on raw multimodal content alone. The *w/o CoT* variant retains the trained CoT-conditioned embedder but omits CoT during encoding. The *w/o RASD* variant omits Retrieval-Aware Self-Distillation. ReWAM outperforms all three variants across all modalities, with overall gains of 3.4, 4.2, and 0.8 points, respectively. Removing CoT from the trained embedder yields a lower overall score than raw-input training, indicating reliance on the reasoning context. RASD further improves retrieval with CoT, supporting the value of directing reasoning toward retrieval-relevant distinctions.

Table 3: Main ablation studies using the Qwen2-VL-2B embedder.

Model	Image	VisDoc	Video	ALL
ReWAM	**71.6**	**72.8**	**47.7**	**66.5**
w/o RASD	70.8	72.1	46.9	65.7
w/o CoT	67.0	70.1	42.6	62.3
w/ raw input	68.5	70.4	42.5	63.1

Table 4: RASD training ablation. Best results are bolded.

Model	Image	VisDoc	Video	ALL
Trajectory-level	71.1	72.4	47.3	66.0
$\lambda_R = 0.1$	71.4	72.4	47.5	66.2
$\lambda_R = 0.3$	71.4	72.6	47.6	66.3
$\lambda_R = 0.5$	**71.6**	**72.8**	**47.7**	**66.5**
$\lambda_R = 0.7$	71.4	72.7	47.7	66.3
$\lambda_R = 1.0$	71.6	72.7	47.5	66.4

Effectiveness of RASD. Despite using the same training data and configuration, trajectory-level GRPO achieves lower overall scores than RASD across all modulation strengths (Table 4). Even a small modulation strength of $\lambda_R = 0.1$ raises the overall score from 66.0 to 66.2. Evidence-guided weights differentiate token updates within a CoT and favor input-supported, retrieval-relevant reasoning. Even weak modulation therefore introduces guidance absent from a uniform trajectory-level signal. Performance peaks at 66.5 with $\lambda_R = 0.5$. Further increasing λ_R to 0.7 or 1.0 yields 66.3 and 66.4, respectively, both above the trajectory-level baseline. These results demonstrate the robustness of RASD to the choice of modulation strength. We further examine the contribution of Contrastive Evidence Privileged Information (CEPI) by progressively adding its components to trajectory-level supervision, as shown in Table 5. Grounded evidence supplies input-supported facts, while decision boundaries distinguish the positive from hard negatives. Claim verification checks the sampled CoT against this guidance to identify supported and conflicting statements, completing CEPI. The full configuration raises the overall score from 66.0 to 66.5, with gains

of 0.5, 0.4, and 0.4 points on image, visdoc, and video tasks, respectively. These gains highlight the benefit of ReWAM over trajectory-level supervision.

Effectiveness of RAI. To evaluate RAI's effectiveness in accelerating retrieval-grounded reasoning, we vary the number of draft tokens per block ($K = 4$ or 7) and ablate the Retrieval Utility Head (Table 6). Speculative decoding alone yields 2.0–2.4× the baseline throughput with nearly unchanged CoT lengths and retrieval scores. Adding the head reduces CoT length by 26–36% and further raises throughput by 26–64%. Increasing K from 4 to 7 raises full-RAI throughput from 4.1 to 4.5, 3.1 to 3.3, and 2.9 to 3.4 samples/s on image, visdoc, and video tasks, respectively. At $K = 7$, RAI averages a 3.3× speedup across modalities, with only a 0.3–0.4-point drop in retrieval scores. This speedup supports practical CoT-based retrieval at scale.

Table 5: Ablation of Contrastive Evidence Privileged Information (CEPI). Components are added cumulatively.

Variant	Image	VisDoc	Video	ALL
Trajectory-level	71.1	72.4	47.3	66.0
+ Grounded evidence	71.4	72.7	47.5	66.3
+ Decision boundary	71.5	72.8	47.6	66.4
+ Claim verification	**71.6**	**72.8**	**47.7**	**66.5**

Table 6: Ablation of RAI on MMEB-V2. We report Hit@1 for Image and Video, NDCG@5 for VisDoc, the average number of CoT tokens, and throughput (samples/s).

Variants	Retrieval Utility Head	Image			VisDoc			Video		
		Hit@1	Avg. tokens	Samples/s	NDCG@5	Avg. tokens	Samples/s	Hit@1	Avg. tokens	Samples/s
Baseline	×	71.6	136.3	1.2	72.8	131.4	1.1	47.7	143.5	1.1
$K = 4$	×	71.5	137.4	2.5	72.7	132.2	2.2	47.9	144.1	2.3
	✓	71.4	87.9	4.1	72.5	97.1	3.1	47.4	102.0	2.9
$K = 7$	×	71.4	137.3	2.9	72.8	132.1	2.3	47.7	144.5	2.5
	✓	71.3	87.9	4.5	72.5	97.3	3.3	47.3	101.8	3.4

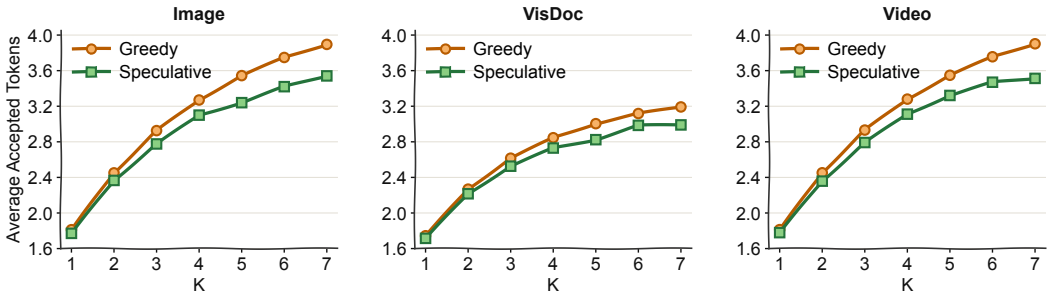


Figure 3: Ablation of Retrieval-Adaptive Inference (RAI). Average accepted tokens as draft length K varies under greedy and speculative acceptance on MMEB-V2. Counts include the bonus token.

Fig. 3 shows continued gains in accepted-token counts up to $K = 7$, especially under greedy verification. This suggests further speedup potential from scaling draft-training data and block length. We further compare greedy and speculative verification, with greedy verification accepting more tokens across all three modalities. This gap may reflect stronger draft–target top-1 agreement encouraged by greedy-generated training traces. Such agreement suffices for greedy verification, whereas speculative verification must preserve the target distribution and may therefore reject a matching token when the draft overestimates its probability (Hu et al., 2025).

5 Conclusion

We presented ReWAM for retrieval-grounded reasoning in universal multimodal embeddings. Retrieval-Aware Self-Distillation (RASD) conditions an on-policy self-teacher on input-supported contrastive evidence to refine token-level credit assignment. Retrieval-Adaptive Inference (RAI) predicts remaining retrieval utility to terminate unproductive reasoning and accelerates token generation through speculative decoding. ReWAM achieves state-of-the-art performance on MMEB-V2 and MRMR with up to 5× the throughput of competitive explicit-CoT methods. These advances make high-quality explicit reasoning more practical for corpus-scale multimodal retrieval.

AI USE STATEMENT

AI tools assisted with drafting and revising the manuscript. The authors carefully reviewed and revised every paragraph and take full responsibility for the final content.

ETHICS STATEMENT

We conduct experiments on existing multimodal datasets and do not collect new data from human participants. As with other retrieval systems, ReWAM may inherit biases from pretrained models and training data. Downstream applications should account for these biases and protect the privacy of indexed content.

REPRODUCIBILITY STATEMENT

Section 3 specifies the SFT, RASD, and RAI objectives and procedures. Appendix A describes the training data and its construction. Appendix B provides training settings, the RAI algorithm, and evaluation protocols. Additional results appear in Appendix C, and CEPI prompt templates in Appendix D. We will make the code publicly available to support reproducibility.

REFERENCES

Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D Lee, Deming Chen, and Tri Dao. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*, 2024.

Jian Chen, Yesheng Liang, and Zhijian Liu. Dflash: Block diffusion for flash speculative decoding. *arXiv preprint arXiv:2602.06036*, 2026a.

Libiao Chen, Xiyang Liu, Yanheng Wei, Tao Wang, and Zhenyu Tang. Umer: Unifying embedding and ranking via pair-aware discriminative reasoning for universal multimodal retrieval. *arXiv preprint arXiv:2608.18504*, 2026b.

Xin Cheng, Xingkai Yu, Chenze Shao, Jiashi Li, Yunfan Xiong, Yi Qian, Jiaqi Zhu, Shirong Ma, Xiaokang Zhang, Jiasheng Ye, et al. Dspark: Confidence-scheduled speculative decoding with semi-autoregressive generation. *arXiv preprint arXiv:2607.05147*, 2026.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829. IEEE, 2023.

Xuanming Cui, Jianpeng Cheng, Hong-you Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Mishra, Taipeng Tian, Qi Guo, et al. Think then embed: Generative context improves multimodal embedding. In *International Conference on Learning Representations*, volume 2026, pp. 2690–2709, 2026.

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

Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737*, 2024.

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.

Zhengmian Hu, Tong Zheng, Vignesh Viswanathan, Ziyi Chen, Ryan Rossi, Yihan Wu, Dinesh Manocha, and Heng Huang. Towards optimal multi-draft speculative decoding. In *International Conference on Learning Representations*, volume 2025, pp. 3181–3203, 2025.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

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*, 2026.

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.

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.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Ume-r1: Exploring reasoning-driven generative multimodal embeddings. In *International Conference on Learning Representations*, volume 2026, pp. 115779–115806, 2026.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International conference on machine learning*, pp. 19274–19286. PMLR, 2023.

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

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pp. 44215–44234, 2025.

Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4015–4025. IEEE, 2025.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, et al. Self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2605.15155*, 2026.

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

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

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

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

Renliang Sun, Wei Cheng, Dawei Li, Haifeng Chen, and Wei Wang. Stop when enough: Adaptive early-stopping for chain-of-thought reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27250–27268, 2026.

Yu-Che Tsai, Kuan-Yu Chen, Yuan-Chi Li, Yuan-Hao Chen, Ching-Yu Tsai, and Shou-De Lin. Let llms speak embedding languages: Generative text embeddings via iterative contrastive refinement. In *International Conference on Learning Representations*, volume 2026, pp. 19099–19119, 2026.

Jinyang Wu, Shuo Yang, Zhengxi Lu, Fan Zhang, Yuhao Shen, Lang Feng, Haoran Luo, Zheng Lian, Shuai Zhang, Zhengqi Wen, et al. Seed: Self-evolving on-policy distillation for agentic reinforcement learning. *arXiv preprint arXiv:2607.14777*, 2026a.

Peixi Wu, Ke Mei, Feipeng Ma, Bosong Chai, Zhibin Lan, Chenxi Zhao, Shannan Yan, Jie Chen, Zhangchi Hu, Yansong Peng, et al. Beyond chain-of-thought: Rewrite as a universal interface for generative multimodal embeddings. *arXiv preprint arXiv:2604.22280*, 2026b.

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*, 2026c.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026a.

Shuo Yang, Jinyang Wu, Zhengxi Lu, Yuhao Shen, Fan Zhang, Lang Feng, Shuai Zhang, Haoran Luo, Zheng Lian, Zhengqi Wen, et al. Opid: On-policy skill distillation for agentic reinforcement learning. *arXiv preprint arXiv:2606.26790*, 2026b.

Hao Yu, Zhuokai Zhao, Shen Yan, Lukasz Korycki, Jianyu Wang, Baosheng He, Jiayi Liu, Lizhu Zhang, Xiangjun Fan, and Hanchao Yu. Cafe: Unifying representation and generation with contrastive-autoregressive finetuning. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 6345–6356. IEEE, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

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*, 2026a.

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

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

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. Scope: Signal-calibrated on-policy distillation enhancement with dual-path adaptive weighting. *arXiv preprint arXiv:2604.10688*, 2026a.

Binbin Zheng, Zijun Xie, Guanqun Zhao, Enlei Gong, Xing Ma, Xiaoliang Fu, and Zeyu Chen. Group-reflective self-distillation for agentic reinforcement learning. *arXiv preprint arXiv:2607.28076*, 2026b.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. Vista: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3185–3200, 2024.

APPENDIX CONTENTS

A Data Construction **15**

A.1 SFT Data 15

A.2 RASD Data Construction 15

A.3 RAI Data Construction 16

B Implementation and Evaluation Details **18**

B.1 Training Details 18

B.2 Analysis of Retrieval-Aware Self-Distillation 18

B.3 Retrieval-Adaptive Inference Algorithm 19

B.4 Evaluation Protocol 20

C Experimental Results **21**

C.1 Additional Analysis 21

C.2 Detailed Results on MMEB-V2 21

C.3 Qualitative Case Studies 24

D CEPI Prompt Templates **35**

E More Discussions **36**

A Data Construction

A.1 SFT Data

We use 1.4M positive query–target pairs with the same SFT data and prompt templates as Wu et al. (2026b), covering image, video, and visual-document tasks. Each query and target is encoded with its original multimodal content and corresponding offline CoT. These paired representations provide the positive examples for bidirectional contrastive learning. The CoTs serve as conditioning context for the embedder.

A.2 RASD Data Construction

Our RASD data consist of positive query–target pairs and hard-negative relationships between them. We use the VLM2Vec-V2 training datasets (Meng et al., 2025) to identify difficult retrieval comparisons, filter them by reasoning benefit and semantic validity, and retain the original pairs needed for training. Table 7 summarizes the final composition.

(1) Hard-negative mining. To focus RASD on retrieval decisions that direct embeddings do not resolve reliably, we mine confusable candidates from the original multimodal inputs. We first remove exact duplicate pairs within each dataset, then use a frozen Qwen2-VL-2B embedder without CoT conditioning to retrieve the top 50 eligible candidates from the same dataset by cosine similarity, excluding known positives, duplicate inputs, and shared media. Mining is bidirectional except for classification datasets, which use only query-to-target retrieval. An input is selected if its hardest negative outranks all known positives or its smallest positive–negative similarity gap falls in the lowest 20% for its dataset and retrieval direction. We retain up to three negatives per input and direction and form a 200,000-pair pool.

(2) Retrieval-gain filtering. Retrieval difficulty alone does not show whether CoT improves discrimination. To select comparisons with a measurable reasoning benefit, we evaluate the same input, positive, and mined negative with and without CoT. The frozen Qwen3-VL-8B-Instruct reasoner greedily generates one CoT for each query and target independently, using the task- and modality-specific prompts of Wu et al. (2026b). A separate frozen Qwen2-VL-2B embedder trained with CoT conditioning encodes each item both with its own CoT and without CoT. The retrieval margin is the input's cosine similarity to the positive minus that to the negative. Let m_{CoT} and m_{noCoT} denote the CoT-trained embedder's margins with and without CoT, respectively. A comparison is retained only when both conditions hold:

$$m_{\mathrm{CoT}} > 10^{-6}, \qquad m_{\mathrm{CoT}} - m_{\mathrm{noCoT}} > 10^{-6}. \tag{13}$$

The first condition requires the positive to rank above the negative, while the second requires CoT to improve the margin over the same embedder's no-CoT baseline.

(3) Semantic verification. Mined negatives may include valid alternative matches, which would introduce incorrect supervision if treated as negatives. We therefore use the fixed Qwen3.5-122B-A10B model as an LLM judge to verify positive and negative labels. It receives the original multimodal content of the input item, its positive, and up to three retained hard negatives, with their retrieval roles explicitly labeled. The judge returns a validity decision for the positive and flags alternative positives, semantic duplicates, and uncertain judgments among the negative candidates. It also provides a self-reported confidence score for the positive decision and each candidate assessment. We retain a comparison only when the positive is judged valid, the negative has none of these flags, and both confidence scores are at least 0.8.

(4) Training-set assembly. To retain informative negative comparisons within training batches, we select positive pairs together with their verified hard-negative relationships. We collect both original pairs from each verified input–negative relationship and exclude the five classification datasets to avoid shared-label false negatives in symmetric in-batch training. We then select 20,000 pairs from the remaining 20 datasets under modality quotas, preserving verified hard-negative relationships among the selected pairs. These relationships guide batch construction so that difficult competitors can be evaluated together. RASD trains on the selected pairs with freshly sampled CoTs; the offline CoTs above are used for data selection.

A.3 RAI Data Construction

For training the Retrieval Utility Head and Draft model in RAI, we sample distinct query–target pairs from the SFT data. The RASD-trained Qwen3-VL-8B-Instruct reasoner greedily generates CoTs separately for query and positive inputs using the task- and modality-specific prompt templates of Wu et al. (2026b). Within each sampled pool, CoTs shared by identical reasoner inputs are counted once, while the original query–target associations are retained for retrieval-label construction. This yields 730,003 successful CoT sequences. We append a closing `</think>` tag to sequences that lack one. Table 8 summarizes the final dataset composition.

Table 7: Training data composition. Initial and SFT report the initial and filtered positive-pair counts reported by Wu et al. (2026b), respectively; RASD reports our selected pairs. Modality indicates the query-to-target direction, where T, I, and V denote text, image, and video; text includes task instructions.

Dataset	Initial	SFT	RASD	Modality
Image-based (MMEB-train)				
A-OKVQA	17,000	15,304	384	$T{+}I \rightarrow T$
CIRR	26,000	22,189	416	$T{+}I \rightarrow T{+}I$
ChartQA	28,000	25,479	1,472	$T{+}I \rightarrow T$
DocVQA	40,000	37,049	1,408	$T{+}I \rightarrow T$
HatefulMemes	8,000	7,352	0	$T{+}I \rightarrow T$
ImageNet-1K	50,000	43,607	0	$T{+}I \rightarrow T$
InfographicsVQA	24,000	21,059	1,280	$T{+}I \rightarrow T$
MSCOCO	50,000	23,417	224	$T{+}I \rightarrow T{+}I$
MSCOCO-i2t	50,000	44,630	512	$T{+}I \rightarrow T$
MSCOCO-t2i	50,000	43,229	384	$T \rightarrow T{+}I$
N24News	49,000	40,960	0	$T{+}I \rightarrow T$
NIGHTS	16,000	12,615	256	$T{+}I \rightarrow T{+}I$
OK-VQA	9,000	8,022	224	$T{+}I \rightarrow T$
SUN397	20,000	18,879	0	$T{+}I \rightarrow T$
VOC2007	8,000	6,456	0	$T{+}I \rightarrow T$
Visual7W	50,000	44,382	512	$T{+}I \rightarrow T$
VisDial	50,000	39,537	672	$T \rightarrow T{+}I$
VisualNews-i2t	50,000	41,177	640	$T{+}I \rightarrow T$
VisualNews-t2i	50,000	39,823	608	$T \rightarrow T{+}I$
WebQA	17,000	12,978	288	$T \rightarrow T{+}I$
Subtotal	662,000	548,144	9,280	Image-centric
Video-based (LLaVA-Hound)				
Caption Retrieval	300,000	232,044	3,296	$T{+}V \rightarrow T$
Video QA	300,000	207,829	1,472	$T{+}V \rightarrow T$
Video Retrieval	300,000	244,327	3,232	$T \rightarrow T{+}V$
Subtotal	900,000	684,200	8,000	Video-centric
Document-based				
ViDoRe	100,000	90,415	1,728	$T{+}I \rightarrow T$
VisRAG	100,000	81,086	992	$T \rightarrow I$
Subtotal	200,000	171,501	2,720	Document-centric
Total	**1,762,000**	**1,403,845**	**20,000**	Multimodal

Table 8: RAI data composition for Retrieval Utility Head and Draft model training. Counts report single-input CoT records after input deduplication within each sampled pool. Query and Positive denote each record's source role; percentages use all 730,003 sequences.

Dataset	Query	Positive	Total	Ratio
Image / VQA / classification				
A-OKVQA	6,208	2,997	9,205	1.26%
CIRR	8,448	6,045	14,493	1.99%
ChartQA	9,472	4,681	14,153	1.94%
DocVQA	11,648	7,984	19,632	2.69%
HatefulMemes	2,944	4	2,948	0.40%
ImageNet-1K	11,136	1,852	12,988	1.78%
InfographicsVQA	9,344	4,136	13,480	1.85%
MSCOCO	5,184	4,287	9,471	1.30%
MSCOCO (i2t)	11,648	11,482	23,130	3.17%
MSCOCO (t2i)	10,752	10,432	21,184	2.90%
N24News	7,104	47	7,151	0.98%
NIGHTS	7,040	5,491	12,531	1.72%
OK-VQA	3,456	1,921	5,377	0.74%
SUN397	9,088	751	9,839	1.35%
VOC2007	3,200	40	3,240	0.44%
VisDial	8,704	8,704	17,408	2.38%
Visual7W	10,429	6,588	17,017	2.33%
VisualNews (i2t)	8,576	8,576	17,152	2.35%
VisualNews (t2i)	7,168	7,168	14,336	1.96%
WebQA	7,296	4,182	11,478	1.57%
Subtotal	158,845	97,368	256,213	35.10%
Video / LLaVA-Hound				
LLaVA-Hound caption retrieval	71,296	71,296	142,592	19.53%
LLaVA-Hound QA	68,608	68,237	136,845	18.75%
LLaVA-Hound video retrieval	65,152	65,152	130,304	17.85%
Subtotal	205,056	204,685	409,741	56.13%
Document retrieval				
ViDoRe	20,992	16,336	37,328	5.11%
VisRAG	11,617	15,104	26,721	3.66%
Subtotal	32,609	31,440	64,049	8.77%
Total	**396,510**	**333,493**	**730,003**	**100.00%**

B Implementation and Evaluation Details

B.1 Training Details

SFT Training Details. The embedders are initialized from Qwen2-VL (2B and 7B) and Qwen3-VL (2B and 4B) and trained on 1.4M query–target pairs for one epoch with a learning rate of 5×10^{-5}. Following the sub-batch scheme of VLM2Vec (Jiang et al., 2024b), each contrastive sub-batch contains 128 pairs drawn from a single dataset. The contrastive loss in Eq. (1) is computed within each sub-batch with temperature $\tau_E = 0.02$, while the global training batch contains 512 pairs. We use a cosine learning-rate schedule with a warm-up ratio of 0.03. The maximum sequence length is set to 8,192 tokens. For images and visual-document pages, we set the minimum and maximum pixel counts to $16 \times 28 \times 28$ and $576 \times 28 \times 28$, respectively. For videos, the corresponding per-frame limits are $256 \times 28 \times 28$ and $1024 \times 28 \times 28$. We sample 4–8 frames per video. Images and frames outside these ranges are resized to satisfy the pixel constraints. RASD and RAI use the same visual preprocessing settings.

RASD Training Details. The SFT embedder is frozen, and the Qwen3-VL-8B-Instruct reasoner is optimized on 20K query–target pairs for one epoch with a learning rate of 1×10^{-6}. Each training batch contains 32 pairs from the same dataset, and each query or positive input produces $G = 8$ CoT rollouts with a maximum response length of 1,024 tokens. Rollouts use temperature 1.0, top-p 0.9, and top-k -1. The fixed Qwen3.5-122B-A10B analyzer uses the multimodal context of the input, its positive, and retrieved hard negatives to extract input-grounded evidence and discriminative boundaries. For each input, we retrieve the *top-3* in-batch hard negatives using the frozen embedder without CoT conditioning. The analyzer then checks each sampled CoT against the original multimodal context and extracted evidence, identifying supported claims and incorrect or borrowed details. The resulting evidence, decision boundaries, and claim-verification summaries constitute CEPI. The analyzer uses the same visual preprocessing as SFT and generates at most 512 tokens per response. We set the KL-penalty coefficient to 0.01, the format reward weight to $\lambda_f = 0.01$, the margin reward tolerance to $\epsilon_v = 0.01$, both clipping thresholds to $\epsilon_R = \epsilon_P = 0.2$, and the token-modulation strength to $\lambda_R = 0.5$. The CEPI prompt templates are provided in Appendix D.

RAI Training Details. The RASD-trained reasoner remains frozen while the Retrieval Utility Head and Draft model are trained on the CoT collection in Appendix A.3. Both modules are trained for 10 epochs with a batch size of 128 CoT sequences and learning rates of 3×10^{-4} and 6×10^{-4}, respectively. The Retrieval Utility Head is an MLP with hidden widths 64 and 128 and a sigmoid output, trained with the SmoothL1 objective in Eq. (9); its stopping threshold is $\tau_{\text{stop}} = 0.01$. The Draft model architecture follows Chen et al. (2026a) and Cheng et al. (2026). We train it with the objective in Eq. (11), setting $\lambda_{\text{ntp}} = 0.1$, $\lambda_{\text{align}} = 0.9$, $\lambda_{\text{acc}} = 1$, and $\gamma = 4$.

B.2 Analysis of Retrieval-Aware Self-Distillation

Trajectory-level retrieval feedback does not identify which reasoning claims are supported by the input and distinguish the positive from hard negatives. RASD uses CEPI to provide token-level guidance for input-only reasoning; we establish bounds on its modulation and its effect on the policy gradient.

Consider one policy update with fixed sampled CoTs $o^{(i)}$, privileged guidance $c^{(i)}$, finite advantages A_i, and old policy π_{old}. The two log-probabilities defining Δ_t^i use the same pre-update reasoner parameters and token prefix. As in Section 3.3, A_i, Δ_t^i, s_t^i, and w_t^i are held constant when differentiating the loss; the policy surrogate differentiates only the input-only likelihood in ρ_t^i.

Proposition B.1 (Bounded modulation and monotonicity)**.** *Suppose* $0 < \epsilon_R < 1$, $\lambda_R \geq 0$, *and* $\lambda_R \epsilon_R < 1$. *The weights and modulated advantages in Eq. (6) satisfy*

$$1 - \lambda_R \epsilon_R \leq w_t^i \leq 1 + \lambda_R \epsilon_R, \qquad |\hat{A}_t^i - A_i| \leq \lambda_R \epsilon_R |A_i|, \\ \operatorname{sign}(\hat{A}_t^i) = \operatorname{sign}(A_i). \tag{14}$$

For fixed A_i, *let* $F_i(d)$ *be the modulated advantage obtained by setting* $s_t^i = \exp(\operatorname{sign}(A_i) d)$ *in Eq. (6). Then* F_i *is nondecreasing and* $F_i(0) = A_i$.

Proof. The clipped term lies in $[-\epsilon_R, \epsilon_R]$, giving $|w_t^i - 1| \le \lambda_R \epsilon_R$ and hence $|\hat{A}_t^i - A_i| = |A_i||w_t^i - 1| \le \lambda_R \epsilon_R |A_i|$. Since $w_t^i \ge 1 - \lambda_R \epsilon_R > 0$, multiplying by w_t^i preserves the sign of A_i. For $A_i \neq 0$, within the unclipped interval,

$$F_i'(d) = \lambda_R |A_i| \exp(\mathrm{sign}(A_i) d) \ge 0. \tag{15}$$

Outside this interval, F_i is constant in each clipped region. Continuity at the clipping boundaries therefore establishes monotonicity. For $A_i = 0$, F_i is identically zero. At $d = 0$, $s_t^i = 1$ and $w_t^i = 1$, so $F_i(0) = A_i$. □

For $\lambda_R > 0$ and $A_i \neq 0$, positive teacher feedback ($\Delta_t^i > 0$) increases the advantage coefficient: it strengthens reinforcement for $A_i > 0$ and weakens penalties for $A_i < 0$, with the reverse effect for negative feedback. With $\epsilon_R = 0.2$ and $\lambda_R = 0.5$, weights lie in $[0.9, 1.1]$, so the advantage changes by at most 10% of $|A_i|$. If $A_i = 0$, the modulated advantage is also zero regardless of Δ_t^i.

To analyze the policy update, define the per-token clipped surrogate

$$S(\rho, A) = \min(\rho A, \mathrm{clip}(\rho, 1 - \epsilon_P, 1 + \epsilon_P) A). \tag{16}$$

Let $\ell_{i,t}^0 = -S(\rho_t^i, A_i)$ and $\ell_{i,t}^{\mathrm{RASD}} = -S(\rho_t^i, A_i w_t^i)$ use identical samples and rewards. Their averages $\mathcal{L}_{\mathrm{surr}}^0 = \mathbb{E}_{i,t}[\ell_{i,t}^0]$ and $\mathcal{L}_{\mathrm{surr}}^{\mathrm{RASD}} = \mathbb{E}_{i,t}[\ell_{i,t}^{\mathrm{RASD}}]$ use the fixed empirical averaging in Eq. (7) and exclude its KL regularizer.

Proposition B.2 (Gradient reweighting and perturbation bound)**.** *Under the conditions of Proposition B.1, with A_i and w_t^i detached and $0 < \epsilon_P < 1$,*

$$\ell_{i,t}^{\mathrm{RASD}} = w_t^i \ell_{i,t}^0, \qquad \nabla_\theta \ell_{i,t}^{\mathrm{RASD}} = w_t^i \nabla_\theta \ell_{i,t}^0, \tag{17}$$

where the gradient identity holds wherever the surrogate is differentiable. If all token terms are differentiable, then for any norm,

$$\left\| \nabla_\theta \mathcal{L}_{\mathrm{surr}}^{\mathrm{RASD}} - \nabla_\theta \mathcal{L}_{\mathrm{surr}}^0 \right\| \le \lambda_R \epsilon_R \, \mathbb{E}_{i,t}\left[\left\| \nabla_\theta \ell_{i,t}^0 \right\| \right]. \tag{18}$$

Proof. For $w > 0$, factoring w out of both arguments of the minimum gives $S(\rho, wA) = wS(\rho, A)$ and preserves the active clipping branch. Since w_t^i is detached, differentiating gives the gradient identity. The triangle inequality and $|w_t^i - 1| \le \lambda_R \epsilon_R$ imply

$$\begin{aligned} \left\| \mathbb{E}_{i,t}\left[(w_t^i - 1) \nabla_\theta \ell_{i,t}^0 \right] \right\| &\le \mathbb{E}_{i,t}\left[|w_t^i - 1| \left\| \nabla_\theta \ell_{i,t}^0 \right\| \right] \\ &\le \lambda_R \epsilon_R \, \mathbb{E}_{i,t}\left[\left\| \nabla_\theta \ell_{i,t}^0 \right\| \right]. \end{aligned}$$

□

RASD therefore rescales each token's surrogate gradient without changing its clipping branch; gradients suppressed by clipping remain zero. The perturbation bound is scaled by the mean of individual token-gradient norms. Adding identical KL terms preserves this bound.

B.3 Retrieval-Adaptive Inference Algorithm

Algorithm 1 expands the offline training and online decoding procedure in Section 3.4. We use its query-side notation $q, t^+, \mathcal{T}^-$; target-side processing is symmetric.

Offline rollouts and prefix labels. The frozen reasoner generates one complete greedy CoT o of length L per input. For query q, candidate embeddings are cached from the targets' full CoTs and held fixed. Let $\mathcal{N} \subseteq \{1, \ldots, L\}$ denote evaluated prefix lengths. Endpoints start at token 32 with stride 28, shifting to a sentence boundary within the next 12 tokens when available; at most seven are retained, including L. For each $n \in \mathcal{N}$, Eq. (8) gives m_n, g_n, and u_n from $\mathbf{e}_q^n = E_\phi(q, o_{\le n})$. The baseline m_0 is computed from the query embedding $E_\phi(q)$ without CoT and the cached candidate embeddings. Maxima range only over $\mathcal{N}$; all-zero gains give zero utility. The implementation adds 10^{-6} to the denominator for numerical stability.

Training the two modules. Pairs (h_n, u_n) train H_ω through Eq. (9), using the frozen reasoner's post-token states h_n. The Draft model D learns from offline CoT blocks via Eq. (11), with reference predecessor $t_{n+k-1}^\star$ and label $t_{n+k}^\star$. Its auxiliary prediction $\hat{a}_{n+k}$ uses Draft states z_{n+k} and the preceding token to estimate acceptance; $\hat{u}_n$ uses target states h_n to control CoT stopping.

Online draft verification and stopping. Starting from a target-generated anchor, D proposes up to K tokens using Eq. (10). The frozen reasoner verifies the block, retains the longest prefix matching its greedy predictions, and supplies a correction or bonus token. Verification returns this continuation and valid target states, discarding states of rejected tokens and respecting natural termination and the token budget. Corrected or bonus tokens are checked once their own target states become available. For each valid state in token order with $n \geq 32$, $H_\omega(h_n) < \tau_{\text{stop}}$ closes the CoT at n with `</think>`; otherwise decoding continues. Online stopping reuses target states without candidate comparisons or embedder calls.

Algorithm 1 RAI: training and inference

Require: Frozen reasoner π_θ and embedder E_ϕ; head H_ω; Draft model D
Require: Offline groups $\mathcal{G} = \{(q, o, t^+, \mathcal{T}^-)\}$ with cached full-CoT target embeddings; query $q_\star$; draft length K; stopping threshold τ_{stop}

Training

$\mathcal{D}_H \leftarrow \emptyset$ # Utility-head training set
for $(q, o, t^+, \mathcal{T}^-) \in \mathcal{G}$ **do**
  Select prefix endpoints $\mathcal{N}$ from o, including L. # Prefix sampling
  Compute m_0 from $E_\phi(q)$ and cached targets. # No-CoT margin
  Compute m_n and g_n using Eq. (8) for $n \in \mathcal{N}$. # Prefix retrieval gains
  Extract h_n at each selected endpoint. # Frozen reasoner's states
  for $n \in \mathcal{N}$ **do**
    $u_n \leftarrow \dfrac{\max\{g_k \mid k \in \mathcal{N},\ k \geq n\} - g_n}{\max\{g_k \mid k \in \mathcal{N}\} + 10^{-6}}$ # Remaining utility
    $\mathcal{D}_H \leftarrow \mathcal{D}_H \cup \{(h_n, u_n)\}$ # State–utility pair
  end for
end for
Train H_ω on $\mathcal{D}_H$ using Eq. (9). # Utility regression
Train D on CoT blocks using Eq. (11). # Draft model training

Inference

Initialize o and $\{h_n\}$ with $\pi_\theta(q_\star)$. # First greedy token
while true **do**
  Predict $\hat{u}_n = H_\omega(h_n)$ for new verified states with $n \geq 32$. # Stopping signal
  if any $\hat{u}_n < \tau_{\text{stop}}$ **then**
    return $o_{\leq n}$ at the first such n, closed with `</think>`. # Truncate
  else if CoT is closed or the token budget is exhausted **then**
    return o. # Generation complete
  else
    D proposes up to K tokens using Eq. (10). # Draft proposals
    Verify with π_θ and update o and $\{h_n\}$. # Verified continuation
  end if
end while

B.4 Evaluation Protocol

MMEB-V2 (Meng et al., 2025) contains 78 datasets: 36 image, 18 video, and 24 visual-document datasets. We report Hit@1 for image and video tasks and NDCG@5 for visual-document tasks. MRMR (Zhang et al., 2026b) contains 11 subtasks spanning Knowledge, Theorem, and Contradiction. We report Hit@1 on Negation and NDCG@10 on the remaining subtasks. Overall scores are averaged over individual datasets or subtasks. The main retrieval results and training ablations are evaluated without RAI. RAI is evaluated separately in the inference-efficiency experiments. In the ablation studies, throughput is averaged over three runs. All inference speed measurements use a single H800 GPU with a batch size of 1. Throughput covers the complete encoding pipeline, including CoT generation and embedding extraction. To accelerate evaluation, we use vLLM to schedule reasoner inference. We adopt greedy decoding with temperature 0.0, top-p 1.0, and top-k -1. The reasoner's context limit is 8,192 tokens, with at most 7,168 prompt tokens and 1,024 generated tokens. The embedder input length is capped at 8,192 tokens. Our repetition checks use pattern lengths of 4–64 tokens and a minimum repeat count of 4.

C EXPERIMENTAL RESULTS

C.1 ADDITIONAL ANALYSIS

Fixed-length Truncation versus RAI. RAI accelerates CoT generation through speculative decoding and uses predicted remaining retrieval utility to stop unproductive continuations. To evaluate adaptive truncation, we compare RAI with fixed-budget and full-CoT generation at $K = 4$ and $K = 7$, alongside an autoregressive baseline. All variants use the same Qwen2-VL-2B embedder and RASD-trained reasoner, with speculative-decoding settings held fixed within each K.

As shown in Table 9, short fixed budgets yield high throughput but sharply reduce retrieval quality. Increasing the budget gradually restores performance at the cost of speed, yet even the longest fixed budget leaves a 0.9-point overall-score gap to the baseline. RAI incurs the smallest overall-score loss among truncation strategies at both draft lengths. Mean throughput reaches 3.4 and 3.7 samples/s at $K = 4$ and $K = 7$, respectively, with only a 0.3–0.4-point score decrease. At $K = 7$, RAI is $1.6\times$ as fast as the longest fixed budget while scoring 0.5 points higher. It thus provides the highest throughput among configurations that preserve near-baseline retrieval quality, demonstrating that retrieval-aware stopping offers a better quality–efficiency trade-off than a uniform token budget.

Table 9: Fixed-length truncation versus RAI. L is the fixed CoT token budget and K is the draft block length. Baseline uses autoregressive decoding without truncation. $\uparrow$ and $\downarrow$ indicate absolute increases and decreases from Baseline; $\uparrow$ 0.0 denotes no change.

Stopping rule	K	Image		VisDoc		Video		ALL	
		Hit@1	Samples/s	NDCG@5	Samples/s	Hit@1	Samples/s	Score	Samples/s
Baseline	–	71.6	1.2	72.8	1.1	47.7	1.1	66.5	1.1
Full CoT	4	71.5 $\downarrow$0.1	2.5 $\uparrow$1.3	72.7 $\downarrow$0.1	2.2 $\uparrow$1.1	47.9 $\uparrow$0.2	2.3 $\uparrow$1.2	66.4 $\downarrow$0.1	2.3 $\uparrow$1.2
$L = 16$	4	67.2 $\downarrow$4.4	7.8 $\uparrow$6.6	70.3 $\downarrow$2.5	7.0 $\uparrow$5.9	42.4 $\downarrow$5.3	6.6 $\uparrow$5.5	62.5 $\downarrow$4.0	7.1 $\uparrow$6.0
$L = 32$	4	67.9 $\downarrow$3.7	7.1 $\uparrow$5.9	71.0 $\downarrow$1.8	5.6 $\uparrow$4.5	43.6 $\downarrow$4.1	5.1 $\uparrow$4.0	63.2 $\downarrow$3.3	5.9 $\uparrow$4.8
$L = 64$	4	68.6 $\downarrow$3.0	4.1 $\uparrow$2.9	72.0 $\downarrow$0.8	3.5 $\uparrow$2.4	45.0 $\downarrow$2.7	3.5 $\uparrow$2.4	64.2 $\downarrow$2.3	3.7 $\uparrow$2.6
$L = 128$	4	70.4 $\downarrow$1.2	2.5 $\uparrow$1.3	72.8 $\uparrow$0.0	2.0 $\uparrow$0.9	46.5 $\downarrow$1.2	1.7 $\uparrow$0.6	65.6 $\downarrow$0.9	2.1 $\uparrow$1.0
RAI	4	71.4 $\downarrow$0.2	4.1 $\uparrow$2.9	72.5 $\downarrow$0.3	3.1 $\uparrow$2.0	47.4 $\downarrow$0.3	2.9 $\uparrow$1.8	66.2 $\downarrow$0.3	3.4 $\uparrow$2.3
Full CoT	7	71.4 $\downarrow$0.2	2.9 $\uparrow$1.7	72.8 $\uparrow$0.0	2.3 $\uparrow$1.2	47.7 $\uparrow$0.0	2.5 $\uparrow$1.4	66.4 $\downarrow$0.1	2.6 $\uparrow$1.5
$L = 16$	7	67.2 $\downarrow$4.4	10.1 $\uparrow$8.9	70.4 $\downarrow$2.4	7.3 $\uparrow$6.2	42.5 $\downarrow$5.2	7.3 $\uparrow$6.2	62.5 $\downarrow$4.0	8.3 $\uparrow$7.2
$L = 32$	7	67.8 $\downarrow$3.8	8.9 $\uparrow$7.7	71.0 $\downarrow$1.8	5.9 $\uparrow$4.8	43.6 $\downarrow$4.1	5.8 $\uparrow$4.7	63.2 $\downarrow$3.3	6.9 $\uparrow$5.8
$L = 64$	7	68.6 $\downarrow$3.0	4.4 $\uparrow$3.2	72.2 $\downarrow$0.6	2.3 $\uparrow$1.2	45.0 $\downarrow$2.7	2.8 $\uparrow$1.7	64.3 $\downarrow$2.2	3.2 $\uparrow$2.1
$L = 128$	7	70.5 $\downarrow$1.1	2.3 $\uparrow$1.1	72.8 $\uparrow$0.0	2.2 $\uparrow$1.1	46.3 $\downarrow$1.4	2.5 $\uparrow$1.4	65.6 $\downarrow$0.9	2.3 $\uparrow$1.2
RAI	7	71.3 $\downarrow$0.3	4.5 $\uparrow$3.3	72.5 $\downarrow$0.3	3.3 $\uparrow$2.2	47.3 $\downarrow$0.4	3.4 $\uparrow$2.3	66.1 $\downarrow$0.4	3.7 $\uparrow$2.6

C.2 DETAILED RESULTS ON MMEB-V2

Tables 10 and 11 report detailed results on all 78 MMEB-V2 datasets. We compare four ReWAM configurations with UME-R1 (Lan et al., 2026), Embed-RL (Jiang et al., 2026), RIME (Wu et al., 2026b), TTE$_s$ (Cui et al., 2026), and PLUME (He et al., 2026), using reported baseline scores. Embed-RL uses Qwen3-VL embedders; the other baselines use Qwen2-VL. For ReWAM, Q2 and Q3 denote Qwen2-VL and Qwen3-VL, respectively.

Table 10: Detailed MMEB-V2 results: overall and task-group averages, followed by the 36 image datasets. Video and visual-document results continue in Table 11. Best and second-best scores are bolded and underlined, respectively.

Dataset / task group	UME-R1		Embed-RL		RIME		TTE_s	PLUME	ReWAM			
	2B	7B	2B	4B	2B	7B	2B	2B	Q2-2B	Q3-2B	Q3-4B	Q2-7B
Avg - All (78 tasks)	60.1	64.5	66.8	68.1	64.1	68.6	63.1	61.6	66.5	67.4	68.7	**69.0**
Avg - Image (36 tasks, Hit@1)	66.6	71.3	69.2	70.1	69.1	73.4	70.1	66.3	71.6	71.9	73.4	**74.2**
Avg - Video (18 tasks, Hit@1)	42.2	47.5	52.1	**53.0**	43.7	49.4	41.3	44.1	47.7	49.7	49.8	49.5
Avg - VisDoc (24 tasks, NDCG@5)	63.9	67.1	74.1	74.7	71.4	75.6	68.8	67.5	72.8	74.2	**75.8**	75.7
I-CLS (10)	64.8	67.1	62.8	63.7	67.9	**70.3**	67.9	66.5	68.8	68.9	**70.3**	**70.3**
I-QA (10)	62.8	69.2	67.9	70.5	64.4	71.7	66.6	59.2	70.3	70.7	71.8	**72.5**
I-RET (12)	67.6	71.9	68.6	71.3	69.8	73.2	70.2	67.6	71.2	70.5	72.3	**74.2**
I-VG (4)	77.2	84.9	90.4	**91.4**	82.1	86.3	84.1	79.7	82.7	86.0	88.4	87.8
V-CLS (5)	44.3	48.6	57.0	**57.6**	48.0	52.6	47.3	45.0	53.0	53.6	52.2	53.6
V-QA (5)	51.2	60.7	55.9	58.4	52.1	62.0	49.1	52.3	61.3	**63.3**	63.2	62.6
V-RET (5)	32.9	38.2	**45.1**	**45.1**	33.6	38.4	33.2	33.5	36.2	38.0	39.5	38.6
V-MR (3)	39.7	39.3	49.4	**49.5**	39.2	41.6	32.1	46.7	35.4	39.9	40.4	38.6
VD-ViDoRe-V1 (10)	72.4	75.7	79.9	80.2	76.4	**80.9**	77.5	72.1	77.4	78.6	80.1	**80.9**
VD-ViDoRe-V2 (4)	46.2	50.5	52.0	53.4	51.4	55.6	53.2	49.8	51.8	55.4	**58.2**	56.1
VD-VisRAG (6)	79.2	83.7	84.6	84.9	81.7	**85.8**	83.2	78.1	84.2	84.2	85.7	85.7
VD-OOD (4)	37.2	37.6	65.7	67.1	63.9	66.9	41.1	57.4	65.2	66.7	**68.1**	67.3
ImageNet-1K	75.3	80.4	78.0	79.5	81.2	80.9	**83.3**	74.1	82.0	78.5	80.1	81.4
N24News	81.1	82.3	44.9	48.3	80.0	**82.7**	78.6	81.1	81.0	79.7	81.3	**82.7**
HatefulMemes	75.2	**79.0**	65.0	66.2	68.4	76.2	64.0	75.5	71.4	71.6	72.5	75.9
VOC2007	80.0	90.8	78.7	79.5	90.4	91.0	86.3	86.1	91.6	**93.3**	**93.3**	92.2
SUN397	79.4	80.3	75.4	79.2	80.1	80.6	77.5	76.9	78.9	81.2	**81.3**	79.3
Place365	42.6	46.8	43.9	43.1	45.3	45.5	45.7	42.4	45.7	46.8	**47.7**	**47.7**
ImageNet-A	50.4	53.9	59.2	58.1	52.1	57.4	50.9	50.8	51.8	57.3	**61.2**	56.7
ImageNet-R	88.7	90.1	88.5	88.2	89.9	89.9	89.7	87.5	90.8	90.5	**92.4**	90.9
ObjectNet	52.0	42.3	74.8	**75.4**	66.1	72.7	74.1	61.5	72.5	70.6	73.5	71.1
Country211	23.4	25.0	20.0	19.4	25.5	26.4	**28.5**	25.0	22.0	19.8	19.6	24.8
OK-VQA	62.4	71.7	61.4	67.3	65.8	74.1	68.4	60.5	68.4	69.6	72.3	**74.5**
A-OKVQA	51.1	58.7	54.7	59.3	56.4	61.8	57.1	49.9	60.0	60.5	62.3	**62.7**
DocVQA	92.2	93.8	92.4	94.3	93.4	94.4	94.2	89.9	94.8	94.6	94.9	**95.6**
InfographicsVQA	67.7	**79.2**	76.7	77.5	63.6	79.1	65.6	59.6	74.6	74.4	76.1	77.3
ChartQA	64.9	75.1	80.7	80.9	61.7	77.4	57.5	49.8	81.6	83.9	**85.1**	84.0
Visual7W	54.1	55.2	52.7	55.3	54.8	54.9	54.1	47.6	54.6	**58.5**	56.6	58.3
ScienceQA	42.7	53.7	57.3	61.6	45.7	59.0	50.7	42.9	58.8	58.3	**62.6**	61.9
VizWiz	46.8	51.6	54.5	**56.2**	47.7	55.3	55.1	46.5	53.9	51.4	53.4	52.5
GQA	67.3	69.3	64.9	68.5	73.1	73.6	**77.0**	69.1	71.4	70.3	69.3	72.2
TextVQA	78.6	83.5	83.8	84.3	81.3	**87.0**	86.2	78.9	84.4	85.9	85.8	86.4
VisDial	76.6	80.7	81.5	**84.9**	80.7	82.6	81.2	72.6	82.8	82.5	83.5	83.2
CIRR	53.7	55.3	47.6	61.2	57.0	60.0	59.4	54.6	55.6	60.8	**61.7**	60.1
VisualNews_t2i	71.7	76.8	71.9	73.7	72.4	**79.8**	72.8	71.3	75.4	71.1	74.8	79.2
VisualNews_i2t	74.2	82.0	73.6	73.9	78.2	**83.5**	76.5	72.7	78.2	73.5	77.0	82.7
MSCOCO_t2i	75.1	78.3	79.4	78.9	76.1	77.8	75.2	74.1	76.8	78.7	**79.5**	78.8
MSCOCO_i2t	68.9	71.4	75.3	76.3	70.0	72.6	71.1	69.8	73.0	75.7	**76.5**	76.3
NIGHTS	67.2	68.1	66.3	66.4	67.9	68.6	**70.8**	68.0	68.2	66.9	69.1	68.9
WebQA	90.0	90.9	89.3	90.5	90.7	90.7	90.4	89.1	90.8	90.3	**92.2**	91.5
FashionIQ	17.1	23.4	24.0	**31.9**	19.8	26.0	26.3	20.3	22.7	25.5	**31.9**	30.8
Wiki-SS-NQ	62.0	72.5	68.9	69.6	68.9	**76.3**	64.2	68.6	72.4	70.1	71.9	76.2
OVEN	66.9	**71.4**	61.4	60.7	67.5	68.6	67.6	68.4	68.7	62.9	63.1	69.7
EDIS	88.0	92.0	84.5	87.4	88.9	91.6	87.0	81.8	89.5	88.4	86.1	**92.7**
MSCOCO	69.5	72.7	92.9	**93.6**	69.0	72.0	67.7	66.9	73.3	78.1	80.5	76.1
RefCOCO	83.3	91.4	94.9	**95.9**	88.9	91.7	91.4	86.5	89.2	93.0	94.8	93.8
RefCOCO-Matching	84.4	91.1	85.8	88.0	90.1	93.7	**95.0**	88.4	89.6	88.1	89.6	93.5
Visual7W-Pointing	71.5	84.2	88.0	87.9	80.5	87.9	82.5	74.9	78.7	84.6	**88.8**	87.8

Table 11: Detailed MMEB-V2 results on the 18 video datasets (Hit@1) and 24 visual-document datasets (NDCG@5). Model abbreviations follow Table 10. Best and second-best scores are bolded and underlined, respectively.

Dataset / task group	UME-R1		Embed-RL		RIME		TTE_s	PLUME	ReWAM			
	2B	7B	2B	4B	2B	7B	2B	2B	Q2-2B	Q3-2B	Q3-4B	Q2-7B
K700	35.8	42.8	55.8	**56.8**	47.7	55.0	49.6	42.2	52.1	50.7	53.7	55.3
SmthSmthV2	44.1	50.4	56.7	**59.5**	48.5	55.1	50.4	44.8	53.5	53.0	55.2	55.1
HMDB51	54.4	58.3	56.7	**60.1**	56.7	58.9	52.5	51.2	58.5	55.9	54.3	55.9
UCF101	67.2	70.0	**79.3**	78.5	66.4	68.5	58.3	66.5	72.0	76.3	75.3	74.7
Breakfast	20.1	21.5	**36.7**	33.0	20.6	25.4	25.4	20.1	28.9	32.3	22.3	26.8
MVBench	49.9	58.2	50.8	55.9	49.1	**59.9**	48.5	47.4	56.5	58.8	59.0	59.0
Video-MME	41.7	47.3	47.1	**50.5**	41.6	49.6	45.8	40.0	48.9	50.4	50.4	49.5
NExTQA	59.9	69.6	53.9	58.2	58.9	69.7	53.8	57.3	70.3	**71.1**	69.2	71.0
EgoSchema	45.4	52.4	53.0	52.8	40.4	55.6	36.4	47.8	55.2	59.6	**62.1**	58.2
ActivityNetQA	57.8	76.0	74.8	74.4	70.5	75.1	60.8	69.2	75.4	**76.5**	75.5	75.4
DiDeMo	32.4	40.0	45.3	**46.8**	33.8	38.4	33.5	32.7	35.4	37.5	38.6	39.0
MSR-VTT	34.3	38.9	45.7	**46.2**	36.4	41.5	34.8	36.2	40.4	42.2	42.6	42.6
MSVD	55.4	60.8	**67.2**	65.8	56.6	59.4	56.5	56.1	57.8	60.6	63.8	59.1
VATEX	29.9	32.6	**43.6**	43.4	28.1	32.7	25.6	28.2	32.3	34.0	35.4	35.8
YouCook2	12.7	18.5	**23.5**	23.3	13.3	19.9	15.8	14.5	15.1	15.9	17.1	16.3
QVHighlight	57.5	54.9	70.7	**73.6**	55.1	56.9	38.9	57.1	43.8	54.7	57.7	50.3
Charades-STA	20.4	21.9	**26.4**	25.0	19.4	21.6	19.5	19.4	18.1	19.3	20.0	19.0
MomentSeeker	41.2	41.1	50.9	49.9	43.2	46.2	37.7	**63.5**	44.2	45.6	43.4	46.6
ViDoRe_arxivqa	73.9	73.6	86.1	**88.7**	82.1	84.1	80.7	72.6	85.1	86.2	87.6	86.8
ViDoRe_docvqa	37.9	41.1	45.7	47.5	45.4	46.9	44.5	36.2	44.7	**48.7**	48.6	47.9
ViDoRe_infovqa	76.2	80.8	86.8	86.9	81.2	85.4	84.8	79.0	83.4	85.0	86.9	**88.0**
ViDoRe_tabfquad	86.1	90.2	94.5	94.7	85.5	**95.3**	88.4	88.8	93.2	92.9	94.0	93.5
ViDoRe_tatdqa	40.6	46.7	54.6	54.8	48.9	52.6	50.4	36.6	50.1	50.8	52.4	**55.5**
ViDoRe_shiftproject	66.8	65.0	70.7	69.0	69.9	73.7	65.2	64.8	69.5	63.9	68.0	**74.2**
ViDoRe_artificial_intelligence	85.9	89.5	94.0	91.6	89.8	**96.1**	91.9	83.8	88.5	92.2	93.1	92.0
ViDoRe_energy	83.3	85.7	86.7	88.1	86.2	88.5	88.7	82.6	84.7	86.4	**88.8**	87.4
ViDoRe_government_reports	82.6	89.8	89.0	90.7	86.4	**91.7**	86.9	83.2	84.6	89.2	89.7	89.4
ViDoRe_healthcare_industry	90.8	94.3	91.1	90.4	88.3	**94.9**	92.8	91.1	90.2	90.9	91.9	94.0
ViDoRe_esg_reports_human_labeled_v2	50.2	50.4	56.9	59.8	56.5	60.2	59.0	52.1	54.5	59.0	59.5	**60.5**
ViDoRe_biomedical_lectures_v2_multilingual	46.2	50.7	51.0	50.1	47.6	51.5	52.0	48.2	52.5	55.3	**58.3**	52.6
ViDoRe_economics_reports_v2_multilingual	45.7	57.8	53.0	53.9	47.0	59.2	49.8	49.6	49.3	55.1	59.5	**60.2**
ViDoRe_esg_reports_v2_multilingual	42.7	43.2	46.9	49.7	54.3	51.6	52.1	49.0	51.0	52.0	**55.3**	51.0
VisRAG_ArxivQA	74.3	80.5	84.9	**86.9**	79.4	84.0	78.5	71.6	83.0	83.0	85.3	84.9
VisRAG_ChartQA	86.0	85.0	88.3	**88.5**	83.3	85.4	84.4	80.8	88.2	86.1	86.6	85.0
VisRAG_MP-DocVQA	75.6	83.4	79.1	79.3	82.3	**87.4**	79.2	74.9	81.8	82.1	84.4	86.6
VisRAG_SlideVQA	87.1	91.5	92.3	92.6	91.0	**94.1**	92.3	88.9	91.4	93.2	94.0	94.0
VisRAG_InfoVQA	84.4	89.2	90.0	89.6	82.4	**91.8**	87.2	85.7	87.7	88.5	89.5	89.5
VisRAG_PlotQA	68.0	72.7	73.0	72.4	71.6	72.1	**77.5**	66.2	73.3	72.4	74.2	74.4
ViDoSeek-page	21.2	21.3	82.0	84.4	80.7	**85.6**	22.6	80.6	81.8	81.5	82.2	81.1
ViDoSeek-doc	75.9	75.3	82.6	82.4	79.7	80.9	82.0	76.9	80.5	82.3	**83.1**	82.7
MMLongBench-page	11.9	12.3	47.7	51.0	48.3	52.4	12.9	39.9	49.5	52.5	**55.2**	53.7
MMLongBench-doc	39.7	41.3	50.3	50.7	46.9	48.8	47.0	32.0	49.1	50.4	**51.9**	51.8

C.3 Qualitative Case Studies

Reasoning quality and adaptive inference. Figures 4–10 illustrate the reasoning quality of the RASD-trained reasoner. Its full CoTs focus on salient entities, actions, document statistics, and spatial relations, distilling multimodal inputs into informative descriptions relevant to retrieval. RAI further shortens these CoTs by omitting later restatements and secondary elaborations while retaining the core matching cues. Query–positive cosine similarities remain close to the full-CoT values, suggesting that the shortened reasoning largely preserves semantic correspondence.

01 Image-to-Text Retrieval (Text Rewriting) — MSCOCO i2t



Query image

QUERY
Find an image caption describing the given everyday image.

POSITIVE
A close up of dozens of of oranges stacked.

Full CoT	Q-P cosine: 0.721539	RAI	Q-P cosine: 0.721461
262 tokens (Q + P)		126 tokens (Q + P)	Δ: −0.000078

Query CoT — Image + text description

140 tokens

Step 1: image description
The image shows a close-up of **numerous oranges piled together**, displaying their bright orange color and textured rinds. Some oranges have minor blemishes or spots. The composition is dense, with oranges overlapping, suggesting abundance, possibly at a market or harvest setting.
Step 2: text analysis
The text requests a caption for this everyday image. The caption should be concise and descriptive, highlighting the subject (oranges) and context (abundance, freshness). A suitable caption might be: "A vibrant pile of fresh oranges, ready for sale at the market." This aligns with the visual content and fulfills the text's request for an image caption.

61 tokens

Step 1: image description
The image shows a close-up of **numerous oranges piled together**, displaying their bright orange color and textured rinds. Some oranges have minor blemishes or spots. The composition is dense, with oranges overlapping, suggesting abundance, possibly at a market or harvest setting.
Step

Positive CoT — Text

122 tokens

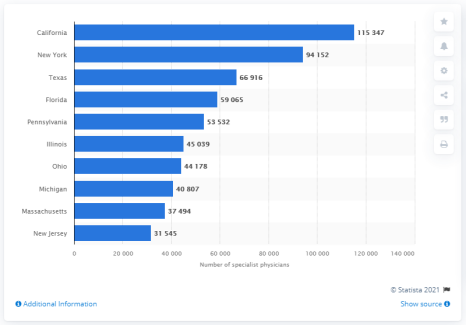
Step 1: Text Type Identification
Conventional complete Sentence or Paragraph
Step 2: Analysis Results
Transformation 1: **Dozens of oranges are stacked in a close-up view.** (Passive voice, focus on arrangement.)
Transformation 2: A close-up shot captures oranges stacked in dozens. (Restructured focus on visual capture.)
Transformation 3: Stacked dozens of oranges are shown in close-up. (Reordered subject-verb-object for emphasis.)
Step 3: Summary
The text describes a visual composition of stacked oranges, with variations preserving meaning through syntactic restructuring.

65 tokens

Step 1: Text Type Identification
Conventional complete Sentence or Paragraph
Step 2: Analysis Results
Transformation 1: **Dozens of oranges are stacked in a close-up view.** (Passive voice, focus on arrangement.)
Transformation 2: A close-up shot captures oranges stacked in dozens. (Restruct

Figure 4: An image-to-text retrieval case on MSCOCO. The RAI panels show the generated CoTs up to their actual stopping positions.

02 Text-to-Document Retrieval (Chart QA) — VisRAG ChartQA



Positive document

QUERY
Find a document image that matches the given query: What state has the 3rd most physicians in 2021?

POSITIVE
Understand the content of the provided document image.

Full CoT
Q–P cosine: 0.591429
250 tokens (Q + P)

RAI
Q–P cosine: 0.588116
183 tokens (Q + P)
Δ: −0.003312

Query CoT

Text

131 tokens

Step 1: Text Type Identification
Conventional complete Sentence or Paragraph
Step 2: Analysis Results
Transformation 1: Identify a document image matching the query about the state with the **third-highest physician count in 2021**.
Transformation 2: Locate the image of a document that answers which state ranked third in physician numbers in 2021.
Transformation 3: Search for a document image specifying the state with the third-most physicians in 2021.
Step 3: Summary
The query seeks a visual document matching a specific factual question about physician distribution by state in 2021.

78 tokens

Step 1: Text Type Identification
Conventional complete Sentence or Paragraph
Step 2: Analysis Results
Transformation 1: Identify a document image matching the query about the state with the **third-highest physician count in 2021**.
Transformation 2: Locate the image of a document that answers which state ranked third in physician numbers in 2021.
Transformation

Positive CoT

Image only

119 tokens

The image is a horizontal bar chart from Statista (2021) showing the number of specialist physicians in ten U.S. states. **California leads with 115,347, followed by New York (94,152) and Texas (66,916).** The chart ranks states from highest to lowest, with New Jersey at the bottom with 31,545. Bars are blue, labeled with state names and exact numbers, against a grid background. A source link and "Additional Information" are visible at the bottom.

105 tokens

The image is a horizontal bar chart from Statista (2021) showing the number of specialist physicians in ten U.S. states. **California leads with 115,347, followed by New York (94,152) and Texas (66,916).** The chart ranks states from highest to lowest, with New Jersey at the bottom with 31,545. Bars are blue, labeled with state names and exact numbers, against a grid background.

Figure 5: A text-to-document retrieval case on VisRAG-ChartQA. The RAI panels show the generated CoTs up to their actual stopping positions.

03 Image-to-Text Retrieval (Image Description) — MSCOCO i2t



Query image

QUERY
Find an image caption describing the given everyday image.

POSITIVE
Bunches of carrots and broccoli on display stand.

Full CoT — Q–P cosine: 0.617568
251 tokens (Q + P)

RAI — Q–P cosine: 0.616503
98 tokens (Q + P) — Δ: −0.001064

Query CoT

Image + text description

127 tokens

Step 1: image description
A vibrant farmers market stall displays fresh produce: **bright orange carrots bundled in the foreground**, dark purple beets to the right, and **green broccoli and leafy greens in the background**. A woven basket and a handwritten sign are partially visible, suggesting a rustic, local setting.
Step 2: text analysis
The text requests a caption for this everyday scene. The image perfectly fits the context of a farmers market, so a caption like "Freshly harvested carrots and beets at a local farmers market" would accurately and concisely describe the visual, aligning with the text's intent.

63 tokens

Step 1: image description
A vibrant farmers market stall displays fresh produce: **bright orange carrots bundled in the foreground**, dark purple beets to the right, and **green broccoli and leafy greens in the background**. A woven basket and a handwritten sign are partially visible, suggesting a rustic, local setting.
Step

Positive CoT

Text

124 tokens

Step 1: Text Type Identification
Conventional complete Sentence or Paragraph
Step 2: Analysis Results
Transformation 1: **Carrots and broccoli** are displayed in bunches on a stand. (Passive voice, same meaning)
Transformation 2: A display stand holds bunches of carrots and broccoli. (Subject inversion, retains meaning)
Transformation 3: Bunches of carrots and broccoli are arranged for display on a stand. (Verb variation, same core idea)
Step 3: Summary
The sentence describes a visual arrangement of vegetables, with three equivalent rephrasings preserving original meaning.

35 tokens

Step 1: Text Type Identification
Conventional complete Sentence or Paragraph
Step 2: Analysis Results
Transformation 1: **Carrots and broccoli** are displayed in bunch

Figure 6: An image-to-text retrieval case on MSCOCO. The RAI panels show the generated CoTs up to their actual stopping positions.

04 (Image + Text)-to-Text Retrieval (Visual QA) GQA



Query image

QUERY
Represent the given image with the following question: Is the freezer to the right or to the left of the appliance near the trashcan?

POSITIVE
The freezer is to the right of the stove.

Full CoT Q–P cosine: 0.484031
227 tokens (Q + P)

RAI Q–P cosine: 0.510906
150 tokens (Q + P) Δ: +0.026874

Query CoT

Image + text QA

117 tokens

Step 1: Image Information Assessment
Low-information image.
Step 2: Image Content Analysis
The image shows a kitchen with a white stove (appliance) near a white trashcan on the left side. To the right of the stove is a white refrigerator with a freezer compartment on top. The spatial relationship is clear: **the freezer is to the right of the stove.**
Step 3: Question Answering
The freezer is to the right of the appliance near the trashcan.
Confidence: High (the spatial arrangement is clearly visible and unambiguous in the image).

84 tokens

Step 1: Image Information Assessment
Low-information image.
Step 2: Image Content Analysis
The image shows a kitchen with a white stove (appliance) near a white trashcan on the left side. To the right of the stove is a white refrigerator with a freezer compartment on top. The spatial relationship is clear: **the freezer is to the right of the stove.**
Step 3: Question Answering

Positive CoT

Text

110 tokens

Step 1: Text Type Identification
Conventional complete Sentence
Step 2: Analysis Results
Transformation 1: **The stove lies to the left of the freezer.** (Reversed spatial relation)
Transformation 2: Positioned to the right of the stove is the freezer. (Inversion with prepositional phrase)
Transformation 3: The freezer, located right beside the stove, is on its right side. (Added descriptive detail)
Step 3: Summary
Original sentence describes spatial arrangement; transformations preserve meaning via syntactic restructuring.

66 tokens

Step 1: Text Type Identification
Conventional complete Sentence
Step 2: Analysis Results
Transformation 1: **The stove lies to the left of the freezer.** (Reversed spatial relation)
Transformation 2: Positioned to the right of the stove is the freezer. (Inversion with prepositional phrase)
Transformation

Figure 7: A visual question answering case on GQA. The RAI panels show the generated CoTs up to their actual stopping positions.

05 Text-to-Video Retrieval (Video Description) MSR-VTT

Positive video: all 8 input frames, in order



QUERY
Find a video that contains the following visual content: a gospel band playing

POSITIVE
Understand the content of the provided video.

Full CoT Q–P cosine: 0.392119
234 tokens (Q + P)

RAI Q–P cosine: 0.398745
184 tokens (Q + P) Δ: +0.006626

Query CoT Text

138 tokens

Step 1: Text Type Identification
Conventional complete Sentence or Paragraph
Step 2: Analysis Results
Transformation 1: Locate a video featuring a gospel band performing live. (Rephrased for clarity and active voice.)
Transformation 2: Search for a video showing a gospel band in action. (Simplified expression, retains core meaning.)
Transformation 3: Retrieve a video with visual footage of **a gospel band playing music**. (Emphasizes "visual content" and "playing music.")
Step 3: Summary
The query seeks a video with specific visual content: a gospel band performing. Three rephrasings preserve meaning while varying structure or emphasis.

104 tokens

Step 1: Text Type Identification
Conventional complete Sentence or Paragraph
Step 2: Analysis Results
Transformation 1: Locate a video featuring a gospel band performing live. (Rephrased for clarity and active voice.)
Transformation 2: Search for a video showing a gospel band in action. (Simplified expression, retains core meaning.)
Transformation 3: Retrieve a video with visual footage of **a gospel band playing music**. (Emphasizes "visual content" and "playing music.")
Step

Positive CoT Video only

96 tokens

The video opens with **a drummer energetically playing a drum set**, surrounded by other musicians in a dimly lit venue. The scene transitions to **a choir of women** dressed in white and rust-colored robes, singing passionately into microphones. They are positioned in front of **a brick wall with columns**, suggesting a church or performance hall setting. Their synchronized movements and expressive faces convey deep emotion and musical engagement. The lighting highlights their faces and robes, enhancing the spiritual or celebratory atmosphere.

80 tokens

The video opens with **a drummer energetically playing a drum set**, surrounded by other musicians in a dimly lit venue. The scene transitions to **a choir of women** dressed in white and rust-colored robes, singing passionately into microphones. They are positioned in front of **a brick wall with columns**, suggesting a church or performance hall setting. Their synchronized movements and expressive faces convey deep emotion and musical engagement

Figure 8: A text-to-video retrieval case on MSR-VTT. The RAI panels show the generated CoTs up to their actual stopping positions.

06 Text-to-Video Retrieval (Video + Text Description) MSR-VTT

Positive video: all 8 input frames, in order



QUERY
Find a video that contains the following visual content: a woman pours water into her pot of meat then tomato sauce and stirs it all around while talking

POSITIVE
Understand the content of the provided video.

Full CoT	Q–P cosine: 0.525334
280 tokens (Q + P)	

RAI	Q–P cosine: 0.516076
155 tokens (Q + P)	Δ: −0.009258

Query CoT — Text

121 tokens

Step 1: Text Type Identification
Conventional complete Sentence or Paragraph
Step 2: Analysis Results
Transformation 1: Locate a video showing a woman **adding water, tomato sauce to meat, stirring, and speaking**.
Transformation 2: A woman in the video combines water and tomato sauce with meat, stirs, and talks.
Transformation 3: In a video, a woman adds water to a pot of meat, then tomato sauce, and stirs while talking.
Step 3: Summary
The text describes a video scene with sequential actions and speech; transformations preserve core meaning with varied phrasing.

70 tokens

Step 1: Text Type Identification
Conventional complete Sentence or Paragraph
Step 2: Analysis Results
Transformation 1: Locate a video showing a woman **adding water, tomato sauce to meat, stirring, and speaking**.
Transformation 2: A woman in the video combines water and tomato sauce with meat, stirs, and talks.
Transformation

Positive CoT — Video + text description

159 tokens

Step 1: video description
The video shows a close-up of a pot on a stove containing chopped meat and onions. A hand pours a red sauce, likely ketchup or tomato-based, into the pot. **The sauce mixes with the meat and onions**, and the hand **uses a spoon to stir the contents**, ensuring even distribution. The scene is set in a kitchen, with the focus on the cooking process.
Step 2: text analysis
The text "Understand the content of the provided video" serves as an instruction to analyze the video's content, which is fulfilled by the detailed description above. The video clearly depicts a cooking process involving meat, onions, and a red sauce, aligning with the directive to comprehend the visual content.

85 tokens

Step 1: video description
The video shows a close-up of a pot on a stove containing chopped meat and onions. A hand pours a red sauce, likely ketchup or tomato-based, into the pot. **The sauce mixes with the meat and onions**, and the hand **uses a spoon to stir the contents**, ensuring even distribution. The scene is set in a kitchen, with the focus on the cooking process.
Step

Figure 9: A text-to-video retrieval case on MSR-VTT. The RAI panels show the generated CoTs up to their actual stopping positions.

07 (Video + Text)-to-Text Retrieval (Video QA) — ActivityNetQA

Query video: all 8 input frames, in order



QUERY
Given a video and a question, select the most accurate answer from the provided candidates. Return only the exact text of your chosen answer. Question: do you use a basket in the course of playing volleyball? (A) yes; (B) no.

POSITIVE
no

Full CoT — Q–P cosine: 0.514007
214 tokens (Q + P)

RAI — Q–P cosine: 0.501112
168 tokens (Q + P) — Δ: −0.012895

Query CoT — Video + text QA

117 tokens

Step 1: Video Information Assessment
Low-information video. The video shows a simple beach volleyball game with repetitive actions and minimal details.
Step 2: Video Content Analysis
The video depicts a beach volleyball match at sunset. Players are seen hitting the ball over a net, diving for digs, and moving around the sandy court. The background includes palm trees and distant buildings. **No basket is visible or used in any part of the video.**
Step 3: Question Answering
(B) no.
Confidence: High (the video clearly shows volleyball play without any basket involved).

96 tokens

Step 1: Video Information Assessment
Low-information video. The video shows a simple beach volleyball game with repetitive actions and minimal details.
Step 2: Video Content Analysis
The video depicts a beach volleyball match at sunset. Players are seen hitting the ball over a net, diving for digs, and moving around the sandy court. The background includes palm trees and distant buildings. **No basket is visible or used in any part of the video.**
Step 3: Question Answering

Positive CoT — Text

97 tokens

Step 1: Text Type Identification
Independent Noun Phrase
Step 2: Analysis Results
Ambiguity: Could be a command, negation, or absence marker.
Semantics: **May imply "not" in context or serve as a standalone response.**
Part-of-Speech: Functions as an interjection or adverbial negation.
Step 3: Summary
"no" is a concise, context-dependent negation or response with multiple interpretive layers.

72 tokens

Step 1: Text Type Identification
Independent Noun Phrase
Step 2: Analysis Results
Ambiguity: Could be a command, negation, or absence marker.
Semantics: **May imply "not" in context or serve as a standalone response.**
Part-of-Speech: Functions as an interjection or adverbial negation.
Step

Figure 10: A video question answering case on ActivityNetQA. The RAI panels show the generated CoTs up to their actual stopping positions.

CEPI-conditioned teacher probabilities. Figures 11–14 visualize the teacher's token probabilities across four cases. We score the saved CoTs offline under fixed prefixes using the same reference reasoner conditioned on CEPI. Superscripts report the CEPI-induced change in token log-probability, Δ; $\bar{\Delta}$ denotes the mean over tokens in an annotated span. Positive and negative values indicate increased and decreased model support, respectively. In the VisDial examples, high-probability regions cover several retrieval-relevant visual details, including clothing and scene attributes, while some low-probability regions correspond to generic evaluations and task restatements. Their image-description steps also have higher mean teacher probabilities than their task-analysis steps. The VisualNews and NIGHTS cases illustrate factual conflicts: a custodian helmet is mislabeled as a "peaked cap," and seating on a flat floor is described as a "tiered layout." CEPI explicitly identifies these discrepancies, and the flagged tokens "peaked" and "layout" receive low teacher probabilities, consistent with the conflict explanations. CEPI increases support for several grounded descriptions while substantially reducing support for the conflicting headwear and seating-layout claims.

Text-to-Image Retrieval (Visual Dialogue) — VisDial

Tennis player



Positive image

Contrastive Evidence Privileged Information

Evidence The primary image shows a male figure wearing a white shirt, dark shorts, and a yellowish cap, playing tennis on an outdoor clay court.

Boundary The subject is a male wearing dark shorts and a cap, distinguishing him from the female subjects or those wearing white shorts described in the comparison contents.

Support The reasoning correctly identifies the subject as a male tennis player in mid-air executing a forehand stroke on a clay court. It accurately notes the attire: white shirt, dark shorts, cap, and athletic shoes. The background elements (chain-link fence, trees, foliage, daylight) are consistent with the image. The interpretation of the text task as a request to visually depict the scene is reasonable.

Conflict None identified.

Teacher probability 0 1

<think>

Step 1: image description

A tennis player in mid-air, executing a **forehand stroke**$^{\bar{\Delta}=+0.415}$ on a clay court. He wears a white shirt, **dark shorts**,$^{\bar{\Delta}=+0.024}$ a cap, and athletic shoes. The background shows a chain-link fence, trees, and foliage under daylight.

Step 2: text analysis

The text "Represent the given image" confirms the task is to visually depict or describe the scene. The image aligns perfectly with this instruction, showing a dynamic, authentic tennis moment. The text's simplicity underscores the image's role as the core visual element to be conveyed.

</think>

Probability: deeper pink denotes higher likelihood. The gray prefill is not scored.

Figure 11: CEPI-conditioned teacher probabilities on VisDial (tennis).

Text-to-Image Retrieval (Visual Dialogue) VisDial

Skier



Positive image

Contrastive Evidence Privileged Information

Evidence man wearing blue snowsuit, hat, and skis on a snowy slope with no trees or clouds visible

Boundary single skier in blue snowsuit wearing a hat (not a helmet) on a clear day with no trees or clouds

Support The skier wears a blue snowsuit and a hat (beanie), not a helmet. The scene is a snowy slope with no trees or clouds visible. Sunlight casts a sharp shadow. A red pole is visible on the left. The skier is carving a turn, kicking up snow spray.

Conflict None identified.

Teacher probability 0 1

<think>

Step 1: image description

A skier in a **blue** **snowsuit**$^{\bar{\Delta}=+0.438}$ and **beanie**$^{\bar{\Delta}=+0.152}$ descends a snowy slope, poles in hand, carving a turn that kicks up snow spray. Sunlight casts a sharp shadow, emphasizing motion. Background shows a wide, snow-covered expanse with faint tracks and a red pole on the left.

Step 2: text analysis

The text "Represent the given image" directs the analysis to visually interpret and describe the scene accurately. The image aligns perfectly with this directive, presenting a clear, dynamic depiction of skiing in a snowy landscape. The text serves as a prompt to focus on visual representation, which the image fulfills by capturing action, environment, and detail.

</think>

Probability: deeper pink denotes higher likelihood. The gray prefill is not scored.

Figure 12: CEPI-conditioned teacher probabilities on VisDial (skiing).

Text-to-Image Retrieval (News Retrieval) — VisualNews

Police headwear



Positive image

Contrastive Evidence Privileged Information

Evidence The image displays two police officers in high-visibility jackets, with the back of one officer clearly showing a 'METROPOLITAN POLICE' patch.

Boundary The image depicts the Metropolitan Police force specifically, whereas the comparison contents refer to general police issues, specific crimes, or police tape without identifying the specific force.

Support The reasoning correctly identifies two Metropolitan Police officers outdoors on grass. It accurately notes the officer facing left wears a checkered hat and holds a radio, and the officer seen from behind wears a high-visibility vest labeled 'METROPOLITAN POLICE.' The interpretation that their uniforms and posture suggest active duty or training is reasonable and explicitly framed.

Conflict The reasoning incorrectly states the officer seen from behind wears a 'peaked cap.' The image clearly shows this officer wearing a traditional custodian helmet (bobby hat), not a peaked cap. This is a factual error regarding the headwear.

Teacher probability 0 1

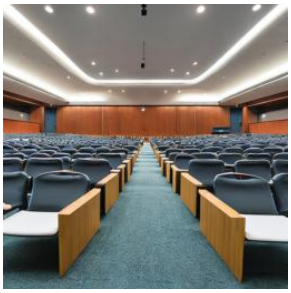
<think>
Step 1: image description
Two **Metropolitan Police** $^{\Delta=+0.002}$ officers are shown outdoors on grass. One officer, facing left, wears a checkered hat and holds a radio. The other, seen from behind, wears a **peaked** $^{\Delta=-11.455}$ cap and a high-visibility vest labeled "METROPOLITAN POLICE." Their uniforms and posture suggest active duty or training.
Step 2: text analysis
The text "Represent the given image" is a directive to depict or describe the visual. The image accurately fulfills this by showing two identifiable Metropolitan Police officers in uniform, providing a clear visual representation as requested.
</think>

Probability: deeper pink denotes higher likelihood. The gray prefill is not scored. Underlines mark conflicts.

Figure 13: A headwear conflict case on VisualNews.

Image-to-Image Retrieval (Visual Similarity) NIGHTS

Seating layout



Positive image

Contrastive Evidence Privileged Information

Evidence Rows of blue chairs with wooden armrests, teal carpet, and a flat ceiling with recessed lighting.

Boundary A modern, flat-ceilinged lecture hall with blue seating, distinct from tiered auditoriums, classrooms with desks, or ornate theaters.

Support The reasoning correctly identifies the room as a modern auditorium with dark blue chairs on wooden armrests, teal carpet, and a flat ceiling with recessed lighting and linear LED strips. It accurately notes the wood paneling on walls and stage backdrop. The interpretation of a formal, institutional atmosphere is consistent with the visual evidence.

Conflict The claim that the chairs are arranged in a 'tiered layout' is incorrect; the primary image shows a flat floor with rows on the same level, unlike the tiered auditoriums in the comparison images.

Teacher probability 0 1

<think>

Step 1: image description

An expansive, modern auditorium with rows of dark blue, individual **chairs**$^{\Delta=+1.279}$ on wooden armrests, arranged in a symmetrical, **tiered**$^{\Delta=-6.189}$ **layout**$^{\Delta=-4.322}$ facing a stage area. The floor is covered in teal carpet, and the ceiling features recessed lighting and linear LED strips. The walls and stage backdrop are clad in wood paneling, conveying a formal, institutional atmosphere.

Step 2: text analysis

The text "Represent the given image" is a directive to depict or describe the visual content. The image fulfills this by presenting a clear, detailed representation of an auditorium, suitable for visual or textual representation as requested. The text does not add contextual meaning but confirms the task's intent.

</think>

Probability: deeper pink denotes higher likelihood. The gray prefill is not scored. Underlines mark conflicts.

Figure 14: A seating layout conflict case on NIGHTS.

D CEPI PROMPT TEMPLATES

The condensed prompts below describe how the analyzer in Figure 2(b) extracts Evidence and Boundary, then verifies each sampled CoT to identify Support and Conflict.

CEPI Prompt Templates

Step 1. Extract Evidence and Boundary

You analyze and compare multimodal content.
Inputs, in order: the primary content (the input), matching content (its positive), and comparison contents (its retrieved hard negatives). Use their complete multimodal context; no generated reasoning is provided in this call.

Return a JSON object containing:
(1) Evidence — evidence: State facts and cues grounded in the primary content.

(2) Boundary — boundary: State why the matching content is consistent with the primary content and which semantic distinction separates it from the comparisons.

Grounding and abstention: Never borrow another item's exclusive details as evidence about the primary content. Set **usable=false** if the primary and matching contents conflict, a comparison is equivalent to the matching content, or the context is insufficient or ambiguous for a grounded distinction. Then set boundary="No reliable boundary." and explain abstain_reason; otherwise set usable=true and abstain_reason="None."

Step 2. Verify Reasoning Claims for Support and Conflict

You check one generated reasoning text against complete multimodal content.
Inputs: the same ordered original contents, the shared Evidence and Boundary from Step 1, and **one sampled reasoning text** about the primary content.

First verify the shared guidance against the complete contents. Set **shared_guidance_consistent=false** and describe shared_guidance_issue if it contradicts the contents or asserts an unsupported distinction; otherwise return true and "None."

Return these claim-analysis fields in the same JSON object:
(3) Support — supported_claims: Identify statements grounded in the primary content, reasonable explicitly framed interpretations consistent with it and the analytical purpose, or reasoning that correctly uses the semantic boundary.

(4) Conflict — incorrect_or_borrowed_claims: Identify contradicted facts, materially specific invented details, information incorrectly borrowed from another content item, materially implausible interpretations, or reasoning inconsistent with the boundary.

Neutral omissions — missing_but_relevant_information: Record omitted information only when grounded in the primary content and relevant to the task. **Omissions and brevity are not errors.** Do not reject reasonable, clearly qualified interpretations merely because they are not directly observable.

Recheck claims against the original contents; shared guidance is only a reference. Use "None identified." when a claim-analysis field has no applicable content.

Rules for Both Calls

Output valid JSON only, without Markdown fences, within **512 generated tokens per call**. Do not mention data labels, evaluation signals, or how contents were selected. Do not provide final answers or generate rewritten reasoning.

CEPI Assembled for the Self-Teacher

After both validity checks pass, concatenate **Evidence + Boundary + Support + Conflict**. Exclude validation fields and neutral omission notes from this context.

E MORE DISCUSSIONS

As shown in Tables 1 and 2, ReWAM achieves strong overall performance on MMEB-V2 and MRMR. For a fixed corpus and encoding configuration, each target's CoT and embedding can be computed once offline, with the embedding stored in a retrieval index. Subsequent requests then require CoT generation and embedding extraction only for the incoming query, followed by a search over the cached target vectors. RAI accelerates both initial corpus encoding and online query processing. However, query-side CoT generation still adds latency, and CEPI construction requires additional analyzer calls during training. Future work could distill the analyzer's evidence extraction and claim verification capabilities into the reasoner to generate its own training-time privileged guidance, reducing external analyzer calls. Another future direction is to learn a unified policy that balances retrieval gains against decoding costs to decide whether to generate CoT, when to stop, and how many tokens to draft at each step.