LEARNING MULTIMODAL EMBEDDINGS WITH EVIDENCE-ALIGNED READOUT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models can expose task-relevant evidence through generation, but producing useful evidence does not by itself determine how it enters a retrieval embedding. We study whether the semantic organization of that evidence can also specify where representations are read. To address this question, we introduce **EviAlign**, which couples *Semantic Evidence Generation* with *Boundary Readout* in a shared multimodal large language model. It organizes evidence into five semantic units, reads the contextualized state at each unit boundary, and aggregates these states into a single normalized embedding. Generation and contrastive retrieval objectives jointly train this shared structure. With the same trailing readout, semantic evidence and free-form CoT yield nearly identical retrieval performance, suggesting that evidence organization alone does not explain the full gain. A controlled 2×3 study compares consistent and permuted evidence organization across three readout strategies, using training targets with matched evidence spans. With five readout states and the same mean pooling, the advantage of consistent semantic organization grows from 0.65 points at length-based training positions to 2.39 at evidence boundaries, yielding a 1.74-point co-design interaction. Across 12 MMEB retrieval tasks, EviAlign achieves 76.9 average Recall@1 with 500K training pairs while retaining single-vector indexing and scoring.

1 INTRODUCTION

Multimodal embedding models map text, images, and interleaved inputs into a shared space for retrieval (Jiang et al., 2025; Zhang et al., 2025). Across tasks, a relevant match may depend on different aspects of the input, including entities, attributes, relations, and task-specific details. Multimodal large language models (MLLMs) can expose these cues through generation. We refer to such task-relevant cues as *retrieval evidence* and study how they can be incorporated into a single retrieval embedding. For example, a composed-image query can pair a photograph of one bottle with an instruction to retrieve three. The relevant evidence must express the requested count and arrangement while preserving identifying visual cues from the photograph. The central challenge is that generating useful evidence does not determine how it contributes to the embedding.

Recent work introduces explicit reasoning, generated context, retrieval-oriented rewrites, or latent computation before embedding (Lan et al., 2026; Cui et al., 2026; Wu et al., 2026a;b), while flexible readouts use learned aggregation or compact multi-vector representations (Cui et al., 2026; Xiao et al., 2026; Faysse et al., 2025). Some of these systems jointly optimize generation and embedding. We study a more specific design question: *Can the semantic units produced during generation also specify where retrieval representations are read?*

Figure 1(a) illustrates three ways to construct an embedding from generated evidence: reading a trailing state, pooling distributed states whose locations are independent of evidence boundaries, and pooling states at evidence-unit boundaries. In a controlled study whose Semantic and Mixed training targets contain the same evidence spans, semantic organization yields its largest advantage when readouts follow these boundaries (Figure 1(b)). This motivates *evidence-aligned readout*: using the semantic organization of generated evidence to determine where contextualized states are read for embedding construction.

In this paper, we introduce **EviAlign**, which couples *Semantic Evidence Generation* with *Boundary Readout* in a shared MLLM. It organizes retrieval evidence into five semantic units—Entity, Attribute,

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.

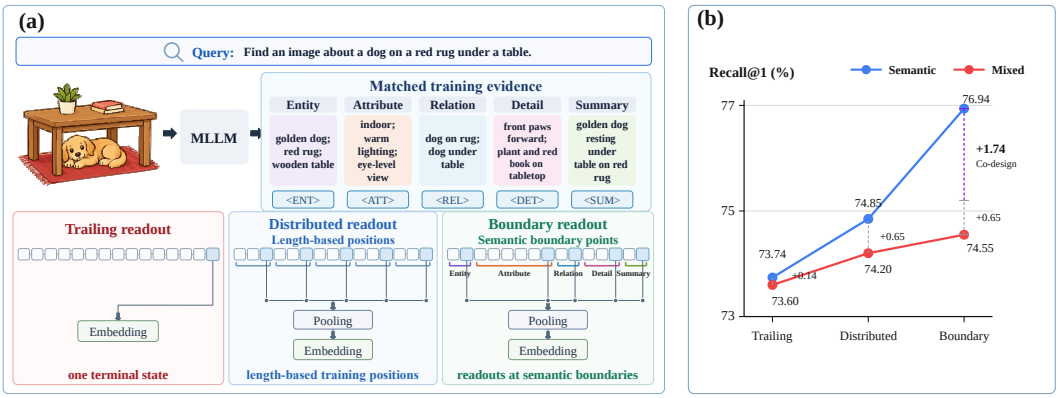


Figure 1: **Evidence–readout co-design in EviAlign. (a)** Training targets share five evidence spans across readout conditions. Distributed uses length-based training positions; at inference, readouts follow the emitted tokens. **(b)** Mixed permutes labeled evidence spans while fixing boundary-token identities and order; Semantic preserves role-to-boundary correspondence. Boundary readout enlarges the semantic–mixed gap from 0.65 to 2.39 points, yielding a 1.74-point co-design interaction.

Relation, Detail, and Summary—and aggregates the contextualized states at their boundaries into a single normalized embedding. Each boundary state is read after its evidence unit has been completed, with access to the input and preceding evidence. Generation supervision and contrastive retrieval jointly optimize the model.

A controlled 2×3 study yields a **1.74-point co-design interaction**; readout analyses further characterize the shared and complementary information retained by the boundary states. On 12 MMEB retrieval tasks, EviAlign reaches **76.9** average Recall@1 with **500K training pairs**. These results support using evidence structure as an interface for representation construction while retaining conventional single-vector indexing and scoring.

Overall, our contributions are as follows:

- We formulate *evidence-aligned readout*: the semantic organization of generated evidence explicitly determines the roles and locations of representation readouts.
- We introduce **EviAlign**, which implements this principle through Semantic Evidence Generation and Boundary Readout while retaining a single-vector indexing and scoring interface.
- Through matched controls, we show that the benefit of semantic evidence depends on how its states are read: neither semantic generation with a trailing readout nor additional readout tokens alone reproduces the full gain.

2 Related Work

Multimodal Embedding Learning. Contrastive vision–language models such as CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and SigLIP (Zhai et al., 2023) learn shared representation spaces for cross-modal matching. Building on MLLMs, unified multimodal embedders support task-conditioned retrieval across text, images, and interleaved multimodal inputs (Jiang et al., 2025; Lin et al., 2025; Zhang et al., 2025). UniME-V2 and Qwen3-VL-Embedding further scale direct multimodal embedding through stronger supervision and multi-stage, multi-source training; Qwen3-VL-Embedding reports 80.2 on MMEB-V2 Image RET (Gu et al., 2026; Li et al., 2026). These systems construct a retrieval representation directly from contextualized model states, without using generated semantic units to specify its readout locations.

Generation-Enhanced Embedding Learning. Generation-enhanced embedders expose retrieval-relevant information before constructing the final representation. UME-R1 and Think-Then-Embed generate reasoning traces or task-conditioned context, while Reasoning Guided Embeddings and RIME produce retrieval-oriented descriptions or rewrites (Lan et al., 2026; Cui et al., 2026; Liu et al., 2025a; Wu et al., 2026a). PLUME and LaME instead perform embedding-oriented computation in

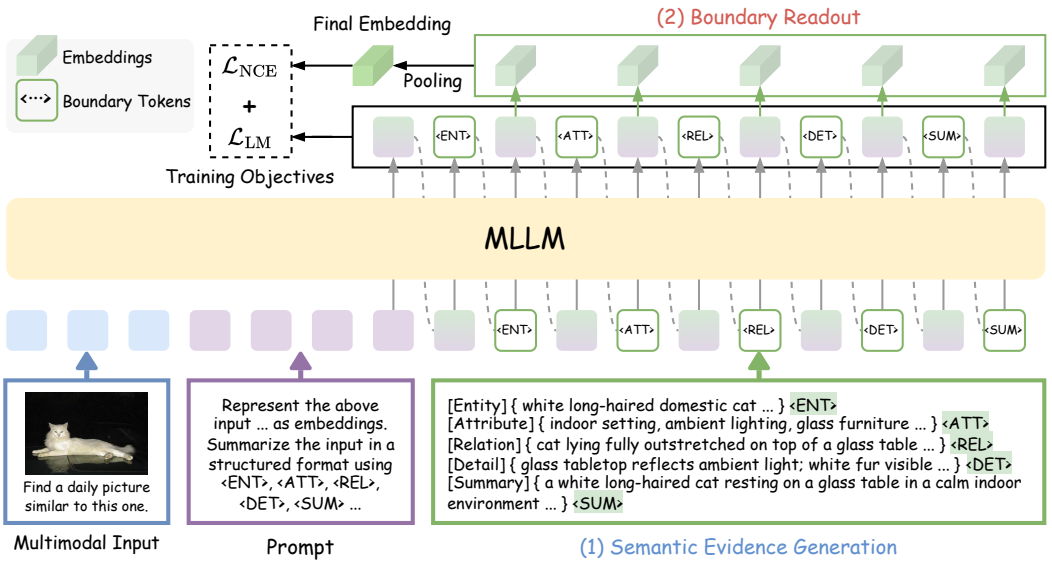


Figure 2: **Overview of EviAlign's evidence–readout co-design.** Semantic Evidence Generation defines evidence units and their boundaries; Boundary Readout extracts states at those boundaries and pools them into one normalized embedding.

latent states rather than readable textual evidence (He et al., 2026; Wu et al., 2026b). Several of these methods jointly optimize generation and embedding. EviAlign focuses on a different structural question: whether generated semantic units can explicitly determine the roles and locations of the states used to construct the retrieval embedding.

Representation Readout for Retrieval. Readout locations may be trailing, learned implicitly, or associated with predefined spans. Most single-vector MLLM embedders use a trailing hidden state (Jiang et al., 2025); other designs apply learned attention, latent-context pooling, or query-based aggregation to contextualized features (Cui et al., 2026). Late chunking constructs representations from contextualized token states over predefined text spans (Günther et al., 2024). Boundary Readout uses boundaries defined by generated, task-relevant semantic evidence. Multi-vector systems such as ColBERT, ColPali, and MetaEmbed retain several representations for late interaction (Khattab & Zaharia, 2020; Faysse et al., 2025; Xiao et al., 2026); EviAlign aggregates its internal readouts into one vector for standard indexing and similarity scoring.

3 EviAlign

We introduce **EviAlign**, a generation-assisted multimodal embedding framework that co-designs evidence generation and representation readout through a shared boundary interface. As shown in Figure 2, Semantic Evidence Generation produces task-relevant units with explicit boundaries, and Boundary Readout extracts and aggregates the contextualized states at those boundaries into a single retrieval embedding. A shared multimodal large language model (MLLM) performs both operations.

3.1 Problem Formulation and Design Principle

Let $\mathcal{D} = \{(q_i, c_i^+)\}_{i=1}^{|\mathcal{D}|}$ denote a collection of query–candidate training pairs, where c_i^+ is relevant to query q_i. A query or candidate may contain text, an image, an interleaved image–text input, and its associated task instruction. We learn a shared encoder

$$\mathbf{h}_\theta(x) \in \mathbb{R}^d, \qquad \|\mathbf{h}_\theta(x)\|_2 = 1, \tag{1}$$

and rank candidates using

$$s_\theta(q, c) = \mathbf{h}_\theta(q)^\top \mathbf{h}_\theta(c). \tag{2}$$

Queries and candidates are independently encoded by the same model. In particular, candidate representations are not conditioned on the query against which they are subsequently compared, allowing them to be precomputed and indexed.

A generation-assisted embedding model can be expressed as

$$\mathbf{z} = G_\theta(\mathtt{P}(x)), \qquad \mathbf{h}_\theta(x) = R_\theta(\mathtt{P}(x), \mathbf{z}), \tag{3}$$

where $\mathtt{P}(x)$ combines the multimodal input with its task instruction, G_θ denotes autoregressive generation, and R_θ denotes representation readout. A single MLLM performs both operations. Equations 5–6 below instantiate R_θ for EviAlign.

Equation 3 exposes two coupled design choices: how retrieval evidence is expressed during generation and how that evidence contributes to the resulting embedding. These choices interact: richer generated content need not preserve its organization in a trailing readout, and additional readout tokens need not align with evidence units.

We therefore formulate *evidence-aligned readout* as a co-design principle with three requirements. First, generation should select task-relevant evidence that supports the intended match or distinguishes it from misleadingly similar candidates. Second, the organization of this evidence should explicitly determine the semantic roles and locations of representation readouts. Third, the resulting readouts should be jointly integrated into an embedding optimized for retrieval.

EviAlign realizes these requirements through two components that share the same evidence structure. Semantic Evidence Generation defines the evidence units and their boundaries, while Boundary Readout uses these boundaries as representation readout locations. Their aggregation yields one normalized vector per input.

3.2 Semantic Evidence Generation

Semantic Evidence Generation selects task-relevant information and organizes it into units that directly guide representation readout. The generated content focuses on evidence that affects relevance under the given instruction. For composed retrieval, for example, the prompt contains both the reference input and the requested modification, and the generated evidence describes the intended target after applying that modification rather than merely captioning the reference image.

Given an input x, the MLLM generates a token sequence partitioned into evidence spans and their following boundary tokens:

$$\mathbf{z} = (\mathbf{z}_1, s_1, \mathbf{z}_2, s_2, \ldots, \mathbf{z}_K, s_K), \tag{4}$$

where $\mathbf{z}$ denotes the complete generated token sequence, $\mathbf{z}_k$ is the contiguous span of the k-th task-relevant evidence unit, and s_k is a boundary token placed immediately after that span. Each evidence-unit type specifies a semantic role; the following token s_k marks the end of the unit, and its final-layer state serves as the corresponding boundary readout.

In our implementation, $K = 5$: Entity ends with <ENT>, Attribute with <ATT>, Relation with <REL>, Detail with <DET>, and Summary with <SUM>. Entity records objects or concepts together with identifying properties; Attribute captures scene-level characteristics; Relation describes actions, interactions, and spatial relations; Detail retains locally discriminative cues such as visible text and fine-grained patterns; and Summary provides a retrieval-focused global description.

The boundary tokens are new vocabulary items initialized from the [EOS] embedding. Their fixed role-to-boundary correspondence lets the same structure organize evidence and provide readout states. Each query or candidate is encoded with its own task instruction, following the independent encoding interface in Section 3.1. Section 4.4 evaluates joint evidence-schema and readout configurations; Appendix D provides the complete generation prompt.

3.3 Boundary Readout

Boundary Readout makes this evidence structure part of embedding construction by forming a contextualized representation at each boundary defined in Equation 4.

Let p_k denote the position of s_k in the generated sequence, with the input-prefix offset implicit. We extract

$$\mathbf{e}_k = H_\theta(\mathtt{P}(x), \mathbf{z}_{\le p_k})\,[p_k], \qquad k = 1, \ldots, K, \tag{5}$$

where $H_\theta(\cdot)[p_k]$ denotes the final-layer hidden state at position p_k. Under causal attention, s_k is the first designated readout position after the complete unit $\mathbf{z}_k$, so $\mathbf{e}_k$ summarizes the input-conditioned evidence prefix available at that position rather than only the immediately preceding unit. The semantic roles specify where representations are read, while retrieval training determines how these contextualized states jointly contribute to the final embedding.

We aggregate the evidence-aligned readouts and normalize the result:

$$\bar{\mathbf{h}}_\theta(x) = \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_k, \qquad \mathbf{h}_\theta(x) = \frac{\bar{\mathbf{h}}_\theta(x)}{\|\bar{\mathbf{h}}_\theta(x)\|_2}. \tag{6}$$

Mean pooling assigns equal coefficients to the five contextualized boundary states, and L2 normalization produces a single cosine-scored vector for indexing and retrieval.

Here *alignment* denotes the within-input correspondence between evidence units and representation readout locations. Retrieval operates on the aggregated global embeddings in Equation 2.

3.4 Training and Inference

EviAlign jointly learns evidence generation and retrieval representation construction. During supervised training, each input x is associated with a target semantic-evidence sequence $\mathbf{y} = (y_1, \ldots, y_T)$, which contains both the evidence content and its boundary tokens. Under teacher forcing, the representation associated with the k-th evidence boundary is

$$\mathbf{e}_k^{\text{train}} = H_\theta\big(\mathtt{P}(x), \mathbf{y}_{\le p_k(\mathbf{y})}\big)\,[p_k(\mathbf{y})]. \tag{7}$$

Here $p_k(\mathbf{y})$ is the boundary-token position in the teacher-forced target, rather than in a generated sequence. These readouts are aggregated by Equation 6.

Given a batch $\mathcal{B} = \{(q_i, c_i^+)\}_{i=1}^{|\mathcal{B}|}$, we optimize a symmetric in-batch contrastive objective. Let $s_{ij} = \mathbf{h}_\theta(q_i)^\top \mathbf{h}_\theta(c_j^+)/\tau$. The loss averages query-to-candidate and candidate-to-query retrieval:

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left[\log\frac{\exp(s_{ii})}{\sum_j \exp(s_{ij})} + \log\frac{\exp(s_{ii})}{\sum_j \exp(s_{ji})}\right], \tag{8}$$

where τ is a temperature parameter. This objective trains the aggregated representation to preserve evidence that distinguishes the relevant candidate from in-batch negatives.

We sum the query- and candidate-side language-modeling losses, each averaged over non-masked supervised target tokens:

$$\mathcal{L}_{\text{LM}} = \mathcal{L}_{\text{LM}}^{q} + \mathcal{L}_{\text{LM}}^{c}. \tag{9}$$

The training objective is

$$\mathcal{L} = \mathcal{L}_{\text{NCE}} + \mathcal{L}_{\text{LM}}. \tag{10}$$

The language-modeling objective teaches the model to express task-relevant evidence under the prescribed organization, whereas the contrastive objective trains the boundary readouts to form a discriminative retrieval embedding. Together they optimize the two sides of the evidence structure shared by generation and readout.

At inference time, no target evidence sequence is provided. The MLLM first autoregressively generates $\hat{\mathbf{z}} = G_\theta(\mathtt{P}(x))$, after which the model reads the hidden states at the generated evidence boundaries and aggregates them into $\mathbf{h}_\theta(x)$. If a required boundary token is absent, its readout falls back to the final-layer hidden state of the last generated token. Queries and candidates follow the same generate-and-read procedure but are encoded independently. Candidate representations are computed offline and stored in the retrieval index, while query-side generation remains part of online encoding. Retrieval then uses Equation 2. Algorithm 1 in Appendix A summarizes the complete procedure.

Table 1: Comparison on the 12 MMEB retrieval tasks (Recall@1, %). In-Domain and Out-of-Domain average eight and four tasks, respectively. Split averages are computed from the corresponding per-task results.†UniME-V2 and LaME report the 12-task retrieval mean without the 8/4 split.

Method	Backbone	Data	In-Domain	Out-of-Domain	Overall
Direct embedding					
GME (Zhang et al., 2025)	Qwen2-VL-7B	∼8M	70.9	**71.8**	71.2
LamRA-Ret (Liu et al., 2025b)	Qwen2-VL-7B	∼1.4M	70.0	69.9	70.0
VLM2Vec (Jiang et al., 2025)	Qwen2-VL-7B	∼662K	75.2	57.9	69.4
VLM2Vec-V2 (Meng et al., 2026)	Qwen2-VL-2B	∼1.7M	74.8	58.7	69.5
UniME-V2 (Gu et al., 2026)	Qwen2-VL-7B	∼662K	–	–	73.1†
Explicit generation or reasoning					
UME-R1 (Lan et al., 2026)	Qwen2-VL-7B	∼1.5M	75.4	64.8	71.9
RIME (Wu et al., 2026a)	Qwen2-VL-7B	∼1.5M	77.0	65.6	73.2
Think-Then-Embed$_t$ (Cui et al., 2026)	Qwen2-VL-7B	∼662K	–	–	75.9
Latent reasoning					
PLUME (He et al., 2026)	Qwen2-VL-2B	∼1.5M	71.5	59.8	67.6
LaME (Wu et al., 2026b)	Qwen2-VL-7B	∼1.55M	–	–	73.1†
EviAlign	Qwen2-VL-7B	**500K**	80.7	65.1	75.5
EviAlign	Qwen3-VL-8B	**500K**	**81.6**	67.7	**76.9**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and Metrics. We focus training and evaluation on MMEB's 12 retrieval tasks, which directly match our study of representation construction for candidate ranking (Jiang et al., 2025). Following the standard protocol, each dataset contains 1000 test queries, and each query is evaluated against a candidate pool of 1000 items (one positive and 999 negatives). The benchmark defines eight in-domain and four out-of-domain retrieval datasets, listed in Appendix B.1. We report Recall@1 and its average across all 12 datasets. With one positive per query, Recall@1 equals Precision@1 under this protocol.

Implementation Details. Our main model and all controlled analyses reported in the main text use 500K query–candidate training pairs from the MMEB-V1 retrieval training split. Semantic-evidence targets for EviAlign are generated by GLM-4.1V-9B-Thinking (GLM-V Team et al., 2025); Appendix B.3 describes the construction and filtering procedure. The default backbone is "`Qwen3-VL-8B-Instruct`" (Bai et al., 2025). We freeze the vision encoder and fine-tune the LLM and projector; all variants within each comparison share the same optimization schedule and single-vector evaluation protocol. We train each variant for one epoch and report its final checkpoint. During training, Boundary Readout operates on teacher-forced semantic-evidence target sequences. During inference, the model generates the evidence sequence with greedy decoding and constructs readouts at the designated positions. The readouts are averaged and L2-normalized to obtain the final embedding. Detailed hyperparameter settings are listed in Appendix B.4.

Baselines. We compare EviAlign with three families of MLLM-based multimodal embedders. Direct embedding baselines include GME (Zhang et al., 2025), VLM2Vec (Jiang et al., 2025), VLM2Vec-V2 (Meng et al., 2026), UniME-V2 (Gu et al., 2026), and LamRA (Liu et al., 2025b). Explicit generation or reasoning baselines include UME-R1 (Lan et al., 2026), RIME (Wu et al., 2026a), and Think-Then-Embed (Cui et al., 2026); latent-reasoning baselines include PLUME (He et al., 2026) and LaME (Wu et al., 2026b).

Appendix B.2 summarizes each baseline's reported training scale and representation extraction.

4.2 MAIN RESULTS

Table 1 reports final retrieval performance and the backbone used by each method. With 500K query–candidate training pairs, EviAlign reaches 75.5 average Recall@1 with Qwen2-VL-7B and 76.9 with Qwen3-VL-8B. Under the Qwen2-VL-7B backbone, EviAlign exceeds UME-R1 and

RIME. Think-Then-Embed$_t$ reports 75.9 using a Qwen2-VL-7B embedder together with a separate Qwen2.5-VL-72B reasoner. The Qwen3-VL-8B model reaches 81.6 in-domain and 67.7 out-of-domain, while GME records the strongest out-of-domain average at 71.8. The controlled studies below use Qwen3-VL-8B and hold the data budget, optimization, and evaluation protocol fixed across variants. Figure 3a shows scaling within EviAlign.

4.3 Controlled Study of Evidence–Readout Co-design

Factorial Evidence–Readout Comparison. We vary evidence organization and representation readout in a 2×3 study under the same 500K budget. For each training example, the Semantic and Mixed targets contain the same five labeled evidence spans. Semantic maintains a fixed correspondence between evidence roles and boundary positions; Mixed permutes the labeled spans for each example while keeping the boundary-token identities and order fixed. The readout factor compares (i) *Trailing*, where only the final-layer state at <SUM> serves as the embedding for contrastive training and retrieval; (ii) *Distributed*, which inserts five distinct readout tokens at positions $\bar{p}_k = \lfloor kT/5 \rfloor$ for $k = 1, \ldots, 5$, where T is the evidence-target length before insertion; and (iii) *Boundary*, which places the same five token types at the five evidence-unit ends. Both five-readout conditions mean-pool the token states. Trailing provides the single-state reference; Distributed matches Boundary in readout count and pooling while placing those states independently of evidence boundaries. All six configurations are trained independently under the same data budget and optimization schedule. At inference, each model generates its own sequence and constructs the embedding from the resulting readout states. Appendix B.6 gives the complete construction.

Let $\mathrm{R@1}(o, r)$ denote average Recall@1 for evidence organization $o \in \{M, S\}$ and readout $r \in \{T, D, B\}$, corresponding to mixed or semantic evidence and trailing, distributed, or boundary readout. We define the organization advantage under readout r as $A_r = \mathrm{R@1}(S, r) - \mathrm{R@1}(M, r)$. Using the distributed condition as the matched five-readout control, the co-design interaction is

$$\Delta_{\mathrm{co}} = A_B - A_D = \big[\mathrm{R@1}(S, B) - \mathrm{R@1}(M, B)\big] - \big[\mathrm{R@1}(S, D) - \mathrm{R@1}(M, D)\big]. \tag{11}$$

Table 2: Evidence organization $\times$ readout on MMEB (average Recall@1, %).

Evidence organization	**Readout**			*Boundary − Distributed*
	Trailing	**Distributed**	**Boundary**	
Mixed	73.60	74.20	74.55	+0.35
Semantic	73.74	74.85	**76.94**	+2.09
Semantic − Mixed	+0.14	+0.65	**+2.39**	**+1.74**

Table 2 reports the six results visualized in Figure 1(b). Its bottom row compares Semantic with Mixed under each readout, while its rightmost column compares Boundary with Distributed under each evidence organization. Under Distributed, the Semantic–Mixed advantage is 0.65 points; with Boundary, it grows to 2.39. Viewed from the other direction, moving the readouts from distributed positions to boundaries adds 0.35 points for Mixed but 2.09 for Semantic. Both comparisons give the same co-design interaction, $\Delta_{\mathrm{co}} = 2.39 - 0.65 = 2.09 - 0.35 = 1.74$ points. Across the six end-to-end configurations, stable role-to-boundary correspondence produces its largest Semantic–Mixed advantage when the model constructs the embedding from the emitted boundary states.

In EviAlign, the boundary states form the interface between generated evidence and the retrieval embedding. Each state is read after its evidence unit has been completed, with the preceding units available through causal context. Contrastive learning trains the pooled boundary states as the retrieval representation, and inference applies the same readout rule to the model's generated evidence. Distributed preserves the token inventory and aggregation without aligning readouts to evidence-unit ends; Mixed preserves the labeled evidence spans without consistent role-to-boundary assignments. The evidence schema thus specifies both what the model generates and which states become the vector used for ranking.

4.4 Analysis of Representation Construction

Table 3: Controlled decomposition of generation and representation readout on MMEB (average Recall@1, %).

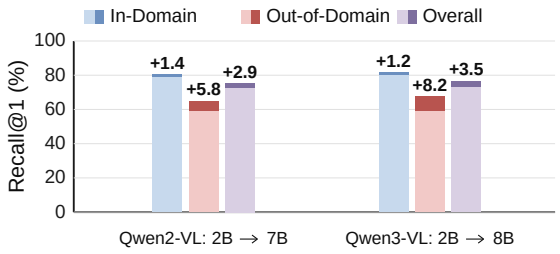
Evidence Generation	Representation Readout	Teacher	Overall
None	Single trailing	–	68.30
None	Multiple readouts	–	70.30
Free-form CoT	Single trailing	GLM	73.75
Semantic evidence	Single trailing	GLM	73.74
Semantic evidence	Multiple trailing readouts (MLTR)	GLM	73.07
Free-form CoT + semantic evidence	Boundary	GLM	76.59
Semantic evidence	Boundary	Qwen	75.60
Semantic evidence	Boundary	GLM	**76.94**

All variants use the same Qwen3-VL-8B student, 500K training pairs, and optimization settings. GLM and Qwen denote GLM-4.1V-9B-Thinking and Qwen3-VL-8B-Thinking. MLTR appends five dedicated readout tokens to the generated sequence end.

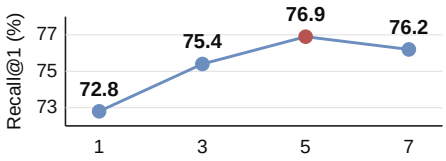
Generation, Readout Capacity, and Evidence Source. Table 3 decomposes the contributions of generation, readout capacity, and evidence source under the same 500K training budget. Adding free-form CoT before a trailing readout raises average Recall@1 from 68.30 to 73.75. With the same trailing readout, semantic evidence and free-form CoT are nearly tied (73.74 vs. 73.75).

Readout multiplicity is likewise insufficient by itself: five readouts without generation reach 70.30, while five trailing readouts after semantic evidence reach 73.07. Using the same semantic-evidence supervision, the boundary-readout model reaches 76.94, outperforming the separately trained trailing-readout model by 3.20 points. Prepending free-form CoT to the semantic evidence gives 76.59 with the same boundary readout, while changing the evidence teacher from GLM-4.1V-9B-Thinking (GLM-V Team et al., 2025) to Qwen3-VL-8B-Thinking (Bai et al., 2025) gives 75.60. Together, these results show that structured evidence realizes its retrieval advantage when its semantic units also determine where representations are read.

Representation-Construction Cost and Format. At batch size 4, end-to-end batch times are 94.6 ms for direct encoding, 1,502.5 ms for EviAlign, and 9,825.3 ms for CoT-plus-semantic Boundary Readout. The latter two average 114.3 and 702.1 generated tokens and reach 76.94 and 76.59 Recall@1, respectively: the longer CoT prefix increases time without improving retrieval. Of 12,000 query generations, 99.52% contain all five boundary tokens exactly once and in order. Appendices C.1 and C.8 report per-task format validity and full efficiency results.



(a) Backbone scaling.



(b) Evidence-unit count.

Figure 3: Analysis of EviAlign's representation construction under the 500K training budget. (a) Performance across Qwen-VL model generations and scales; darker regions show the absolute gains from the larger model within each generation. (b) Effect of the joint evidence-unit and boundary-readout configuration on MMEB Recall@1.

Backbone Analysis. Figure 3a shows that scaling from 2B to 7B/8B improves the overall result from 72.6 to 75.5 for Qwen2-VL and from 73.4 to 76.9 for Qwen3-VL. Out-of-domain gains are 5.8 and 8.2 points, respectively, compared with in-domain gains of 1.4 and 1.2 points. EviAlign improves with backbone scale in both model families, with the largest gains on the out-of-domain tasks. Appendix Table 12 gives the per-task 7B/8B results.

Effect of Evidence-Unit Configuration. We train $K \in \{1, 3, 5, 7\}$ evidence-schema and readout configurations under identical settings. They respectively use Entity; Entity/Attribute/Relation; the default five units; and two additional units, Context (scene context) and Emotion (emotional tone). Figure 3b rises from 72.8 ($K = 1$) to 75.4 ($K = 3$) and 76.9 ($K = 5$), then falls to 76.2 ($K = 7$). The five-unit schema performs best among these joint configurations; adding further units does not necessarily improve retrieval. Appendix D gives the prompts.

Readout Complementarity. Without retraining, we evaluate individual boundary readouts and leave-one-out aggregates from the full five-unit model. These are contextualized states rather than isolated semantic features: each has access to the input and the evidence generated up to its boundary. The one-slot and leave-one-out evaluations therefore offer complementary views of a boundary's retrieval information and its contribution to the pooled vector. ENT alone reaches 76.12 average Recall@1, while aggregating all five readouts reaches 76.94; removing any one reduces the average by 0.23–0.36 points. Different tasks favor different individual readouts: Detail is the strongest single readout on CIRR (78.6), Attribute on VisualNews image-to-text (84.8), and Relation on OVEN (70.9). On OVEN, the full aggregate reaches 72.2. Full aggregation exceeds every individual readout on seven tasks and yields the highest 12-task average. Fixed aggregation thus combines cues across tasks without selecting a task-specific readout. These inference-time interventions differ from the K-unit comparisons in Figure 3b, whose models are trained separately. Appendices C.2 and C.2.1 report per-task and geometry results; Appendices C.4–C.8 and C.10 provide the remaining analyses and examples.

Qualitative Example. Figure 4 pairs a reference image of one bottle with a text request for three. Entity records the requested count, Relation specifies the side-by-side arrangement, and Detail carries over visual cues from the reference bottle. Summary combines the preserved appearance with the new count and arrangement. The resulting evidence describes the intended retrieval target rather than the observed image alone, and its boundary states are pooled into the embedding used for ranking. Appendix C.10 provides examples from two additional retrieval settings.



Query image

"Show three bottles of soft drink."

Query text



Retrieved image

Entity: three vintage soft-drink glass bottles with crown caps
Attribute: warm indoor, retro product-photography setting
Relation: three bottles standing upright side by side
Detail: clear glass, dark liquid, printed labels, metal caps
Summary: three vintage soft-drink bottles arranged upright side by side

Figure 4: A CIRR example. The query pairs a reference image with a text modification; the generated evidence describes the requested three-bottle target while retaining visual attributes of the reference.

5 Conclusion

We introduced **EviAlign**, which uses the same evidence boundaries to organize generation and construct a multimodal retrieval embedding. Its five contextualized boundary readouts are aggregated into one vector for standard indexing and scoring, so generation changes representation construction without changing the retrieval index. A controlled 2×3 study shows that semantic evidence yields its largest advantage when readouts follow the corresponding boundaries, producing a 1.74-point co-design interaction. Complementary controls show that semantic evidence and free-form CoT are nearly tied under a trailing readout, while additional readout states alone do not reproduce the boundary result. With 500K training pairs, EviAlign achieves 76.9 average Recall@1 across 12 MMEB retrieval tasks. Together, these results support using semantic evidence structure not only to organize generation, but also to determine where retrieval representations are read.

Reproducibility Statement

Section 4.1 specifies the MMEB retrieval evaluation protocol. Appendix A provides the training and inference algorithm; Appendix B documents data construction, filtering criteria, hyperparameters, and the controlled experimental configurations. Appendix D presents the evidence-generation, annotation, evidence-configuration ablation, and judge prompt templates. Per-task results and additional analyses are reported in Appendix C.2–C.10.

AI Use Statement

We used GLM-4.1V-9B-Thinking and Qwen3-VL-8B-Thinking to generate structured evidence annotations for the training experiments, and GPT-4o to assess training-target quality. Generative AI tools also assisted with language polishing. The authors reviewed AI-assisted content and take responsibility for the final paper.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token? *arXiv preprint arXiv:2504.02732*, 2025.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16495–16504, 2022.

Xuanming Cui, Jianpeng Cheng, Hong-You Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Kumar Mishra, Taipeng Tian, Qi Guo, Ser-Nam Lim, Aashu Singh, and Xiangjun Fan. Think then embed: Generative context improves multimodal embedding. In *The Fourteenth International Conference on Learning Representations*, 2026.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 326–335, 2017.

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

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50742–50768, 2023.

GLM-V Team, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Haochen Li, Jiale Zhu, Jiali Chen, Jiaxing Xu, Jiazheng Xu, Jing Chen, Jinghao Lin, Jinhao Chen, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Ruiliang Lyu, Shangqin Tu, Sheng Yang, Shengbiao Meng, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wei Jia, Wenkai Li, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyu

Zhang, Xinyue Fan, Xuancheng Huang, Yadong Xue, Yanfeng Wang, Yanling Wang, Yanzi Wang, Yifan An, Yifan Du, Yiheng Huang, Yilin Niu, Yiming Shi, Yu Wang, Yuan Wang, Yuanchang Yue, Yuchen Li, Yusen Liu, Yutao Zhang, Yuting Wang, Yuxuan Zhang, Zhao Xue, Zhengxiao Du, Zhenyu Hou, Zihan Wang, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. GLM-4.5V and GLM-4.1V-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Tiancheng Gu, Kaicheng Yang, Kaichen Zhang, Xiang An, Ziyong Feng, Yueyi Zhang, Weidong Cai, Jiankang Deng, and Lidong Bing. Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21378–21386, 2026.

Michael Günther, Isabelle Mohr, Daniel James Williams, Bo Wang, and Han Xiao. Late chunking: Contextual chunk embeddings using long-context embedding models. *arXiv preprint arXiv:2409.04701*, 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.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12065–12075, 2023.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020. doi: 10.1145/3397271.3401075.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022.

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

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Chunxu Liu, Jiyuan Yang, Ruopeng Gao, Yuhan Zhu, Feng Zhu, Rui Zhao, and Limin Wang. Reasoning guided embeddings: Leveraging mllm reasoning for improved multimodal retrieval. *arXiv preprint arXiv:2511.16150*, 2025a.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6761–6771, 2021a.

Siqi Liu, Weixi Feng, Tsu-Jui Fu, Wenhu Chen, and William Yang Wang. Edis: Entity-driven image search over multimodal web content. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4877–4894, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4015–4025, June 2025b.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2125–2134, 2021b.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6492–6505, 2024.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Raghuveer Thirukovalluru, Xuan Zhang, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *Transactions on Machine Learning Research*, 2026.

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

Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9840–9855, 2023.

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

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *Computer Vision – ECCV 2024*, volume 15145 of *Lecture Notes in Computer Science*, pp. 387–404. Springer, 2024. doi: 10.1007/978-3-031-73021-4_23.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11307–11317, 2021.

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

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

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2024.

Zilin Xiao, Qi Ma, Mengting Gu, Chun cheng Jason Chen, Xintao Chen, Vicente Ordóñez, and Vijai Mohan. Metaembed: Scaling multimodal retrieval at test-time with flexible late interaction. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

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

A EviAlign Algorithm

Algorithm 1 summarizes the training and inference procedure of EviAlign.

Algorithm 1: EviAlign Training and Inference

Initialize MLLM $\mathcal{M}$.

Function `GetEmbedding(`$x, \mathbf{y}$ *(optional)*`)`**:**

  Use target evidence $\mathbf{y}$ under teacher forcing during training; otherwise generate $\mathbf{z}$; `// Eq. 4`

  Read boundary representations $\{\mathbf{e}_k\}_{k=1}^{K}$; `// Eq. 5`

  Mean-pool and normalize the readouts to obtain $\mathbf{h}$; `// Eq. 6`

  return $\mathbf{h}, \{\mathbf{e}_k\}_{k=1}^{K}$

if *is training* **then**

  Input : Batch of query–candidate pairs and target evidence sequences.

  foreach *pair* $(q_n, c_n) \in \mathcal{B}$ **do**

    $\mathbf{h}_{q_n}, \{\mathbf{e}_k^{(q)}\}_{k=1}^{K} \leftarrow$ `GetEmbedding(`$q_n, \mathbf{y}_{q_n}$`)`;

    $\mathbf{h}_{c_n}, \{\mathbf{e}_k^{(c)}\}_{k=1}^{K} \leftarrow$ `GetEmbedding(`$c_n, \mathbf{y}_{c_n}$`)`;

  end

  Compute contrastive loss $\mathcal{L}_{\text{NCE}}$ over $\{\mathbf{h}_{q_n}, \mathbf{h}_{c_n}\}_{n=1}^{|\mathcal{B}|}$; `// Eq. 8`

  Compute language modeling loss $\mathcal{L}_{\text{LM}}$ over semantic-evidence sequences; `// Eq. 9`

  Compute total loss $\mathcal{L} \leftarrow \mathcal{L}_{\text{NCE}} + \mathcal{L}_{\text{LM}}$;

  Update MLLM parameters.

end

if *is inference* **then**

  Input : Query q, candidate corpus $\mathcal{C}$, and precomputed candidate embeddings $\{\mathbf{h}_c : c \in \mathcal{C}\}$.

  $\mathbf{h}_q, _ \leftarrow$ `GetEmbedding(`$q, \varnothing$`)`;

  Retrieve top-k candidates $\mathcal{C}^\star$ using $\mathbf{h}_q^\top \mathbf{h}_c$; `// Eq. 2`

  return $\mathcal{C}^\star$

end

B Experimental Details

B.1 Benchmark Details

We provide a comprehensive overview of the 12 benchmarks that constitute the MMEB retrieval evaluation suite. These datasets collectively span a broad spectrum of multimodal retrieval challenges—from fundamental image-text alignment and composed image retrieval to knowledge-intensive entity grounding, fine-grained visual attribute matching, and visual document understanding. A summary of each benchmark's modality configurations and domain categorization is presented in Table 4; detailed per-dataset descriptions follow.

B.1.1 In-Domain Datasets

VisDial (Visual Dialogue) (Das et al., 2017). Originating from a conversational AI challenge, the dataset features dialogues between a "Questioner" (who sees only a caption) and an "Answerer" (who sees the image). In the MMEB benchmark, this dataset is adapted into a discriminative retrieval task. Instead of generating responses, the model is evaluated on its ability to retrieve the correct target image given the full dialogue history, effectively testing its ability to resolve visual references within multi-turn conversational contexts.

CIRR (Composed Image Retrieval on Real-life images) (Liu et al., 2021b). The dataset is a standard benchmark for Composed Image Retrieval (CIR) containing open-domain, real-life photography. Each query consists of a reference image and a relative modification text (e.g., "change the dog to a cat"). The task requires the model to retrieve a target image that visually reflects the textual instructions applied to the reference. This setup challenges the model to understand fine-grained visual semantics and compositional logic beyond simple keyword matching.

VisualNews (Liu et al., 2021a). The dataset is a large-scale entity-aware dataset consisting of news photographs paired with their original captions. The MMEB suite evaluates this dataset under two distinct bidirectional setups: **VisualNews t2i** (retrieving the relevant image given a caption) and **VisualNews i2t** (retrieving the correct caption given an image). These tasks assess the model's capability to ground complex, event-driven textual concepts—often rich in named entities and context—into the visual domain.

MSCOCO (Microsoft Common Objects in Context) (Lin et al., 2014). As a cornerstone benchmark in vision-language research, the dataset contains high-quality images of everyday scenes annotated with five human-written captions. We evaluate on the retrieval splits defined by the MMEB benchmark, covering both **MSCOCO t2i** and **MSCOCO i2t**. Performance on this dataset serves as a baseline for the model's fundamental ability to perform semantic alignment in general visual scenarios with complex spatial layouts.

NIGHTS (Fu et al., 2023). The dataset focuses on visual similarity across diverse environmental conditions. The original dataset contains triplets with human judgments on similarity. Following the protocol established in M-BEIR (Wei et al., 2024) and adopted by MMEB, this is formulated as a pairwise retrieval task. The reference image serves as the query, and the specific perturbed version identified by human annotators as the "match" serves as the ground-truth target, evaluating robustness to domain shifts.

WebQA (Chang et al., 2022). The dataset is originally a multihop, multimodal question-answering dataset mimicking web search behavior. For the retrieval evaluation, MMEB isolates the evidence retrieval stage, where the goal is to select the correct candidate that contains the answer to a natural language query. This formulation tests the model's capacity for knowledge-intensive retrieval and multihop reasoning.

B.1.2 Out-of-Domain Datasets

FashionIQ (Wu et al., 2021). As a domain-specific counterpart to CIRR, the dataset focuses on the fashion industry (Dresses, Shirts, Tops). It uses crowd-sourced relative captions describing specific attribute differences between a reference product and a target product (e.g., "is darker blue and has a V-neck"). This benchmark specifically tests the model's ability to comprehend fine-grained attribute manipulation and visual feedback within a specialized vertical domain.

Wiki-SS-NQ (Wikipedia ScreenShot Natural Questions) (Ma et al., 2024). Derived from the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019), the dataset introduces a "visual document understanding" component. Instead of retrieving plain text, the task involves retrieving screenshots of Wikipedia pages (Wiki-SS) that answer a user's question. This unique format requires the model to process not just textual content but also structural and layout information embedded in the page screenshot.

OVEN (Open-domain Visual Entity Recognition) (Hu et al., 2023). The dataset evaluates entity grounding in an open-world setting. Each instance pairs a visual query with a recognition-based question. The retrieval task defined in MMEB involves linking this query to a specific knowledge entry. The candidate pool consists of Wikipedia images accompanied by their textual descriptions (titles and summaries). Models must effectively bridge visual features with encyclopedic textual knowledge to identify the correct entity within the MMEB candidate pool.

EDIS (Entity-Driven Image Search) (Liu et al., 2023). The dataset addresses cross-modal retrieval in the dynamic news domain. Unlike generic caption retrieval, EDIS queries are typically rich in specific named entities and event descriptions. The evaluation setup matches these queries against a pool of news images paired with their headlines. This requires the model to go beyond surface-level matching and perform deep semantic reasoning to link textual entities with their visual representations.

Table 4: Summary of the MMEB retrieval evaluation suite. "Query" and "Target" denote input modalities: T = text, I = image, I+T = image-text pair.

Split	Dataset	Query	Target	Retrieval Task
In Domain	VisDial	T	I	Multi-turn dialogue resolution
	CIRR	I+T	I	Composed image retrieval
	VisualNews (t2i)	T	I	Entity-grounded news matching
	VisualNews (i2t)	I	T	Visual-to-caption grounding
	MSCOCO (t2i)	T	I	General visual-language alignment
	MSCOCO (i2t)	I	T	General visual-language alignment
	NIGHTS	I	I	Fine-grained visual similarity
	WebQA	T	I+T	Knowledge-intensive multi-hop QA
Out of Domain	OVEN	I+T	I+T	Open-domain visual entity recognition
	FashionIQ	I+T	I	Fine-grained attribute manipulation
	EDIS	T	I+T	Named-entity news image search
	Wiki-SS-NQ	T	I	Visual document understanding

B.2 Baseline Details

We provide detailed specifications for the baselines compared in our experiments, focusing on differences in training data scale, representation construction, and retrieval protocol that contextualize comparisons with EviAlign.

GME (Zhang et al., 2025). GME is a data-intensive approach built on the Qwen2-VL (Wang et al., 2024) backbone. To achieve broad coverage across arbitrary modality combinations, it relies on a composite dataset of approximately 8 million samples, integrating diverse open-source benchmarks and $\sim$1.1M self-synthesized fused-modal data. Embeddings are extracted from the hidden state of the last (EOS) token using standard instruction-based prompting.

LamRA-Ret (Liu et al., 2025b). Table 1 uses the released LamRA-Ret checkpoint based on Qwen2-VL-7B. Its in-domain, out-of-domain, and overall retrieval means are computed from the corresponding per-task results in the public MMEB leaderboard. LamRA empowers MLLMs with retrieval capabilities through a multi-stage adaptation process using lightweight LoRA modules. Its training pipeline involves a total of approximately 1.4 million samples, combining a language-only pre-training stage on the NLI dataset (275k) with a multimodal instruction tuning stage on M-BEIR (Wei et al., 2024) (1.1M). For embedding extraction, the model utilizes specific summarization prompts ("Summarize above image in one word"), and the representation is obtained from the hidden state immediately preceding a specific placeholder token.

VLM2Vec (Jiang et al., 2025) & VLM2Vec-V2 (Meng et al., 2026). These models represent the standard contrastive fine-tuning paradigm for MLLMs. VLM2Vec (V1) uses a balanced 662k-sample subset of MMEB. V2 significantly expands the corpus to 1.7 million samples by incorporating video instruction data, visual documents, and the original image data, with an interleaved sub-batching strategy to handle modality heterogeneity. Both extract embeddings from the last token's hidden state.

UniME-V2 (Gu et al., 2026). UniME-V2 trains on 662K pairs from the 20 in-domain MMEB datasets and uses MLLM judgments for hard-negative mining and soft semantic supervision. Its Qwen2-VL-7B encoder reports 73.1 average Precision@1 over MMEB's 12 retrieval tasks.

UME-R1 (Lan et al., 2026). UME-R1 introduces a two-stage reasoning-driven paradigm. The Cold-Start SFT stage uses 1.5 million samples augmented with free-form CoT reasoning traces and summaries from a thinking model. A subsequent RL stage (11k balanced samples across modalities) further optimizes the reasoning process via GRPO. Embeddings are extracted from a dedicated generative token appended after the reasoning sequence, bridging generation and retrieval.

RIME (Wu et al., 2026a). RIME replaces free-form chain-of-thought with a retrieval-oriented rewrite and jointly optimizes generation and embedding through Cross-Mode Alignment and Refine-RL.

Think-Then-Embed (Cui et al., 2026). Think-Then-Embed first generates embedding-oriented context with a reasoner and then constructs the retrieval representation with an embedder conditioned on both the input and generated context. The reported TTE_t variant uses a separate teacher reasoner.

PLUME (He et al., 2026). PLUME replaces explicit textual reasoning with a short autoregressive rollout of continuous latent states and transfers explicit reasoning behavior into latent computation through a progressive curriculum.

LaME (Wu et al., 2026b). LaME performs embedding-oriented reasoning with a fixed number of learnable reason tokens in a single forward pass, using them as a fixed-capacity latent bottleneck.

B.3 EviAlign-Evidence Construction

To teach the MLLM to directly generate semantic retrieval evidence, we construct **EviAlign-Evidence**, a supervised training corpus of ⟨multimodal input, semantic-evidence target⟩ pairs from the MMEB-V1 retrieval training split. Specifically, we employ "GLM-4.1V-9B-Thinking" (GLM-V Team et al., 2025) to convert each training input from the MMEB-V1 (Jiang et al., 2025) retrieval subset into the semantic-evidence format consumed by EviAlign. The four out-of-domain evaluation datasets are excluded from EviAlign-Evidence construction. Given an input, a specialized prompt organizes task-relevant evidence into explicit semantic units and appends the corresponding boundary tokens (see Appendix D for the complete template). To ensure annotation quality and training stability, we apply automatic filtering to remove samples exhibiting excessive repetition, invalid formats, or evidence sequences exceeding the 8192-token hard safety limit. After filtering, the retained semantic evidence is combined with the corresponding boundary tokens and aligned back to the original MMEB retrieval examples using their source identifiers. This process retains approximately 500K query–candidate training pairs. Their evidence targets supervise Semantic Evidence Generation and provide the teacher-forced sequences used by Boundary Readout during training.

B.4 Hyperparameter Configuration

To facilitate reproducibility, we detail the hyperparameter settings and hardware configurations used for EviAlign in Table 5.

B.5 Generation, Readout-Capacity, and Evidence-Teacher Controls

Table 3 reports the controlled generation and readout ablations discussed in Section 4.4. The no-generation multi-readout control uses five dedicated readout tokens in place of a generated evidence sequence. All five tokens participate in training, and their final-layer states are mean-pooled and L2-normalized. MLTR places the five dedicated readout tokens together after the generated semantic evidence. The free-form-CoT trailing control generates an unstructured reasoning sequence and uses a single trailing readout. The CoT-plus-semantic control generates free-form reasoning followed by structured semantic evidence and retains the five boundary readouts. Both CoT controls use GLM-generated targets, as specified in Table 3.

B.6 Factorial Evidence–Readout Construction

The 2×3 study in Section 4.3 matches evidence content in the training targets while varying cross-example role consistency and readout location. In the *semantic* condition, Entity, Attribute, Relation, Detail, and Summary evidence maintain fixed assignments to <ENT>, <ATT>, <REL>, <DET>, and <SUM>. In the *mixed* condition, one sample-specific permutation moves the same five evidence spans together with their textual role labels across these fixed boundary-token positions. The permutation is sampled once during target construction and retained across training epochs. At inference, each separately trained model autoregressively generates its own five-unit sequence, and representations are collected from the emitted token positions according to the assigned readout condition. The

Table 5: Hyperparameter settings and hardware configurations for EviAlign.

Hyperparameter	**Value**
Model Architecture	
Base Backbone	Qwen3-VL-8B-Instruct
Vision Encoder	Frozen
LLM & Projector	Full Fine-Tuning
Boundary Token Initialization	`[EOS]` embedding
Min / Max Pixels	768 / $1,572,864$
Max Sequence Length	12,288 tokens
Precision	BF16
Optimization	
Optimizer	AdamW
Learning Rate	5×10^{-5}
Contrastive Temperature τ	0.02
LM Loss Reduction	Per-side mean over supervised target tokens
NCE / Query LM / Candidate LM Weights	1.0 / 1.0 / 1.0
LR Schedule	Cosine Decay
Warmup Ratio	0.03
Weight Decay	0
Gradient Clipping	1.0
DeepSpeed Strategy	ZeRO Stage 3
Training Setup	
Batch Size per GPU	4 pairs
Gradient Accumulation Steps	2
Effective Global Batch Size	256 pairs
Training Epoch	1
Random Seed	42
Hardware	32 $\times$ NVIDIA H20 (80GB)

training targets therefore preserve the labeled evidence spans, unit count, boundary-token identities and order, and length distribution while varying the cross-example role-to-boundary correspondence.

For representation readout, the *trailing* condition retains all five boundary tokens but uses only the final-layer state at `<SUM>` for contrastive training and retrieval. The *distributed* condition places five dedicated readout tokens at content-independent positions

$$\bar{p}_k = \left\lfloor \frac{kT}{5} \right\rfloor, \qquad k = 1, \ldots, 5, \tag{12}$$

where T is the complete evidence-sequence length before readout-token insertion. During supervised-target construction, T is known and the five readout tokens are inserted at these positions before training; the first four are evenly spaced through the sequence and the fifth follows the complete sequence. The model learns to generate the readout tokens jointly with the evidence sequence; at inference, their hidden states are collected when the model autoregressively emits the corresponding tokens in the same generation pass. Their target placement is determined only by sequence length and is not selected or adjusted using evidence boundaries. The *boundary* condition instead places the same five distinct token types—`<ENT>`, `<ATT>`, `<REL>`, `<DET>`, and `<SUM>`—at the five evidence-unit ends. The two five-readout conditions share token identities, order, count, initialization, training treatment, and mean aggregation followed by L2 normalization; they differ in the rule used to place the readout tokens in the training targets. All six configurations retain the same five evidence units and are trained separately using the same Qwen3-VL-8B-Instruct backbone, 500K training pairs, optimization schedule, and greedy inference procedure. The comparison tests whether stable semantic evidence organization is most useful when representation readout follows the corresponding organization.

Table 6: Schematic of the 2×3 evidence order and readout for a CIRR query requesting cookies on a counter. The five evidence spans are abbreviated, and Mixed shows one example permutation.

Organization	Readout	Evidence order and readout
Semantic	Trailing	E<ENT> A<ATT> R<REL> D<DET> S<SUM>; <SUM> state only
Semantic	Distributed	$\mathcal{U}(E \mid A \mid R \mid D \mid S)$; five inserted-token states
Semantic	Boundary	E<ENT> A<ATT> R<REL> D<DET> S<SUM>
Mixed	Trailing	D<ENT> S<ATT> E<REL> A<DET> R<SUM>; <SUM> state only
Mixed	Distributed	$\mathcal{U}(D \mid S \mid E \mid A \mid R)$; five inserted-token states
Mixed	Boundary	D<ENT> S<ATT> E<REL> A<DET> R<SUM>

E: cookies, pastries, and display cases; A: warm indoor bakery; R: cookies on counter and pastries below; D: colorful icing and tiered glass displays; S: bakery counter with cookies. Each abbreviation includes its textual role label (e.g., E includes `[Entity]`), which moves with the span in Mixed. $\mathcal{U}(X)$ inserts <ENT>, <ATT>, <REL>, <DET>, and <SUM> after the first $\lfloor T/5 \rfloor, \ldots, \lfloor 5T/5 \rfloor$ tokens of X, respectively.

C Additional Analyses

C.1 Boundary-Format Validity

We evaluate 1,000 query generations from each of the 12 MMEB retrieval tasks using the final Qwen3-VL-8B model trained on 500K pairs. As shown in Table 7, 11,942 of 12,000 generations (99.52%) contain all five boundary tokens exactly once and in the prescribed order. The remaining 58 generations are incomplete; none contains duplicated or reordered boundary tokens.

Table 7: Query-side boundary-format validity of the final 500K EviAlign model. A valid output contains each of the five boundary tokens exactly once and in the prescribed order.

Dataset	Queries	Valid outputs	Valid format (%)
In-domain			
VisDial	1,000	998	99.8
CIRR	1,000	994	99.4
VisualNews-t2i	1,000	991	99.1
VisualNews-i2t	1,000	997	99.7
MSCOCO-t2i	1,000	993	99.3
MSCOCO-i2t	1,000	999	99.9
NIGHTS	1,000	995	99.5
WebQA	1,000	992	99.2
Out-of-domain			
FashionIQ	1,000	998	99.8
Wiki-SS-NQ	1,000	994	99.4
OVEN	1,000	998	99.8
EDIS	1,000	993	99.3
Overall	**12,000**	**11,942**	**99.52**

C.2 Representation Analysis of Evidence-Aligned Readouts

Starting from the final 500K EviAlign model, we evaluate each evidence-aligned readout alone. We also remove one readout at a time and form the embedding by averaging the remaining four followed by L2 normalization, without retraining. Table 8 summarizes the 12-task averages: the individual states carry retrieval information, but none matches the full five-state aggregate. Removing any one readout lowers the overall average by 0.23–0.36 points. Tables 9 and 10 give the corresponding

per-task results, showing how the five contextualized states contribute to the final single-vector representation.

Table 8: Readout contribution on MMEB (average Recall@1, %).

Setting	ENT	ATT	REL	DET	SUM
One-slot	76.12	75.60	74.94	73.52	74.37
Leave-one-out	76.58	76.63	76.71	76.66	76.68
Full (all five readouts): 76.94					

Table 9: Per-task single-readout results on MMEB (Recall@1, %; 500K model). Each readout is evaluated without retraining; Full averages all five.

Task	ENT	ATT	REL	DET	SUM	Full
VisDial	85.5	**85.7**	84.3	84.3	81.4	85.5
CIRR	74.7	75.2	75.8	**78.6**	75.7	77.2
VisualNews-t2i	76.0	72.7	77.3	77.9	76.1	**78.9**
VisualNews-i2t	82.7	**84.8**	78.1	77.6	79.1	83.3
MSCOCO-t2i	81.0	79.9	80.1	79.3	79.2	**82.1**
MSCOCO-i2t	79.2	78.7	75.2	74.2	76.3	**79.8**
NIGHTS	**75.4**	73.7	72.7	72.6	74.2	74.6
WebQA	90.5	90.0	89.9	88.5	90.5	**91.0**
FashionIQ	31.7	31.7	31.9	30.0	29.5	**32.8**
Wiki-SS-NQ	**75.3**	73.6	72.3	67.2	71.6	73.6
OVEN	69.3	70.1	70.9	61.8	68.6	**72.2**
EDIS	92.1	91.1	90.8	90.2	90.2	**92.3**
Average	76.1	75.6	74.9	73.5	74.4	**76.9**

Table 10: Per-task leave-one-out results on MMEB (Recall@1, %; 500K model). Each column omits one readout without retraining; Full averages all five.

Task	w/o ENT	w/o ATT	w/o REL	w/o DET	w/o SUM	Full
VisDial	85.1	85.1	85.4	85.4	85.4	**85.5**
CIRR	77.1	77.1	76.9	76.2	76.9	**77.2**
VisualNews-t2i	78.8	78.6	78.5	78.1	78.8	**78.9**
VisualNews-i2t	82.7	82.2	83.2	83.1	82.9	**83.3**
MSCOCO-t2i	81.9	82.0	82.0	82.0	81.9	**82.1**
MSCOCO-i2t	79.2	79.5	79.7	79.5	79.7	**79.8**
NIGHTS	74.1	74.5	74.5	74.5	74.1	**74.6**
WebQA	90.9	90.9	90.9	90.9	90.6	**91.0**
FashionIQ	32.5	32.7	32.6	32.6	32.4	**32.8**
Wiki-SS-NQ	73.0	73.0	73.3	73.4	73.5	**73.6**
OVEN	71.8	71.8	71.3	72.0	71.9	**72.2**
EDIS	91.9	92.1	92.2	92.2	92.1	**92.3**
Average	76.58	76.63	76.71	76.66	76.68	**76.94**

C.2.1 Readout-Level Representation Geometry

We examine the five readouts before their aggregation into a single retrieval embedding, using EviAlign and two matched five-readout controls from the factorial study. For this diagnostic, each readout is L2-normalized as $\tilde{e}_k(x) = e_k(x)/\|e_k(x)\|_2$. Positive-pair same-readout alignment is the mean cosine similarity between $\tilde{e}_k(q)$ and $\tilde{e}_k(c^+)$, averaged over the five readouts and matched query–candidate pairs. We quantify within-input dispersion with a Gaussian-potential statistic adapted from Wang & Isola (2020), applied to distinct readout pairs $i < j$ within the same input:

$$\mathcal{D}_{\text{readout}} = \log \mathbb{E}_{x,\, i<j}\left[\exp\left(-2\|\tilde{e}_i(x) - \tilde{e}_j(x)\|_2^2\right)\right]. \tag{13}$$

As shown in Table 11, positive-pair alignment increases from 0.53 and 0.56 in the two controls to 0.63 with semantic evidence and boundary readout, following the same ordering as retrieval performance.

Table 11: Readout geometry under matched five-readout controls on MMEB. Dispersion is diagnostic (more negative means greater separation); Recall@1 is the 12-task average.

Setting	Positive-pair alignment	Within-input dispersion	Recall@1 (%)
Mixed + boundary	0.53	−3.18	74.55
Semantic + distributed	0.56	−3.34	74.85
Semantic + boundary (EviAlign)	**0.63**	−3.27	**76.94**

All three settings retain comparable within-input dispersion. Together with the leave-one-out results in Table 10, these measurements characterize the evidence-aligned readouts that contribute to EviAlign's aggregated representation.

C.3 Per-Task Backbone Comparison

Table 12 reports the complete per-task results of the Qwen2-VL-7B and Qwen3-VL-8B variants summarized in the main comparison, together with their in-domain, out-of-domain, and overall averages.

Table 12: Per-task EviAlign results with Qwen2-VL-7B and Qwen3-VL-8B under the 500K training budget.

In-domain		
Dataset	**Qwen3-8B**	**Qwen2-7B**
VisDial	85.5	84.7
CIRR	77.2	76.3
VisualNews-t2i	78.9	78.0
VisualNews-i2t	83.3	82.5
MSCOCO-t2i	82.1	81.2
MSCOCO-i2t	79.8	78.9
NIGHTS	74.6	73.7
WebQA	91.0	90.3
Average	**81.6**	**80.7**

Out-of-domain		
Dataset	**Qwen3-8B**	**Qwen2-7B**
FashionIQ	32.8	31.6
Wiki-SS-NQ	73.6	70.4
OVEN	72.2	69.0
EDIS	92.3	89.4
Average	**67.7**	**65.1**
Overall (12)	**76.9**	**75.5**

C.4 Structured vs. Free-Form Training Targets

To characterize the generation supervision used by the controlled models, we compare corresponding structured-evidence and free-form CoT targets from the training data. We sample 1,000 MMEB training inputs in a stratified manner and use GPT-4o ("gpt-4o-2024-11-20") as the judge with temperature 0.0. Each target is scored independently from 1 to 5 for retrieval relevance, faithfulness, and specificity. For each input and dimension, the target with the higher score is preferred, and equal scores are counted as ties. We report score-derived win rates among non-tied comparisons together with mean scores over all inputs. The prompt template is provided in Appendix D.

As shown in Figure 5, EviAlign's structured training targets receive higher scores than the free-form CoT training targets across all three dimensions. Models trained with the two target formats nevertheless remain nearly tied under the same trailing readout (73.74 vs. 73.75). Higher target-quality scores therefore do not, by themselves, translate into a better trailing-readout retrieval representation; the boundary readout lets the semantic organization contribute to embedding construction.

C.5 Embedding Visualization

We further visualize query and target embeddings with t-SNE (van der Maaten & Hinton, 2008) on four representative MMEB subsets: CIRR and MSCOCO image-to-text for in-domain evaluation, and EDIS and OVEN for out-of-domain evaluation. For each subset, 800 query embeddings and

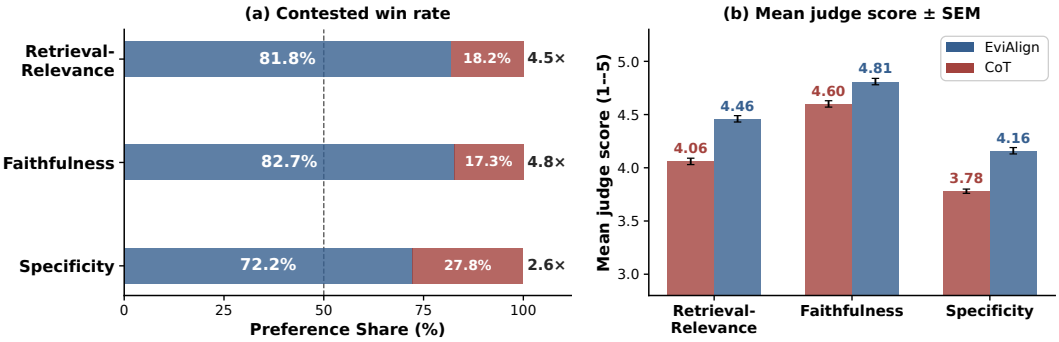


Figure 5: LLM-as-a-Judge comparison of EviAlign structured training targets and free-form CoT training targets. Win rates compare independently assigned 1–5 scores and exclude ties; mean scores use all sampled inputs. Multipliers denote the ratio of EviAlign wins to CoT wins among non-tied comparisons.

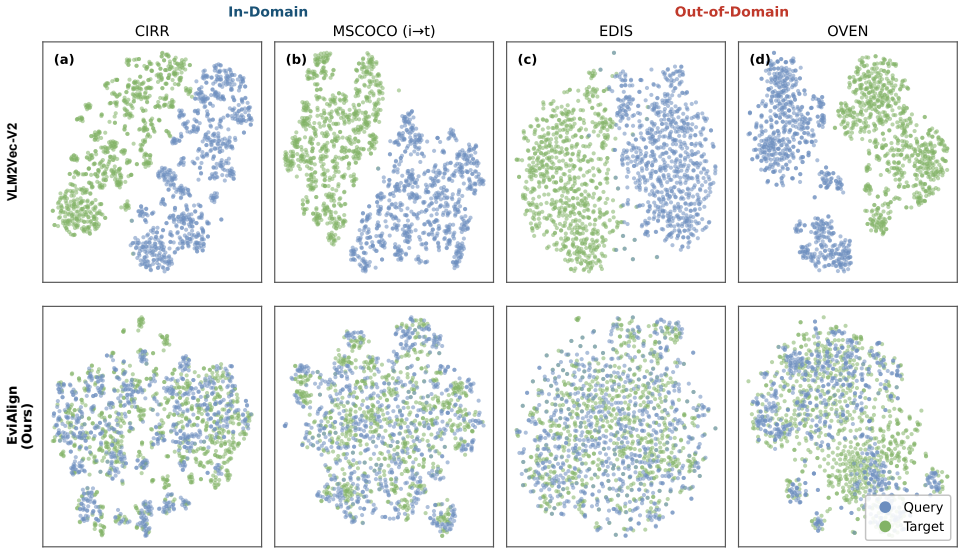


Figure 6: t-SNE visualization of query and target embeddings on four representative MMEB subsets.

800 target embeddings are jointly projected into a two-dimensional space. As shown in Figure 6, VLM2Vec-V2 produces visibly separated query and target clusters, whereas EviAlign yields more interleaved query and target distributions across both in-domain and out-of-domain tasks.

C.6 Attention Visualization

We visualize attention from the last evidence-boundary token (<SUM>) when free-form CoT precedes the semantic evidence, and compare it with the trailing readout of free-form CoT. Figure 7 shows attention concentrated near semantic evidence boundaries, with layer-wise patterns over the surrounding sequence positions. These qualitative patterns echo prior observations that specific token positions can serve as attention or information-flow anchors (Xiao et al., 2024; Barbero et al., 2025; Wang et al., 2023).

C.7 Data Scaling Analysis

We further study how EviAlign scales with training data. Starting from the same training pool, we construct subsets ranging from 1K to 500K samples and train EviAlign with the same backbone,

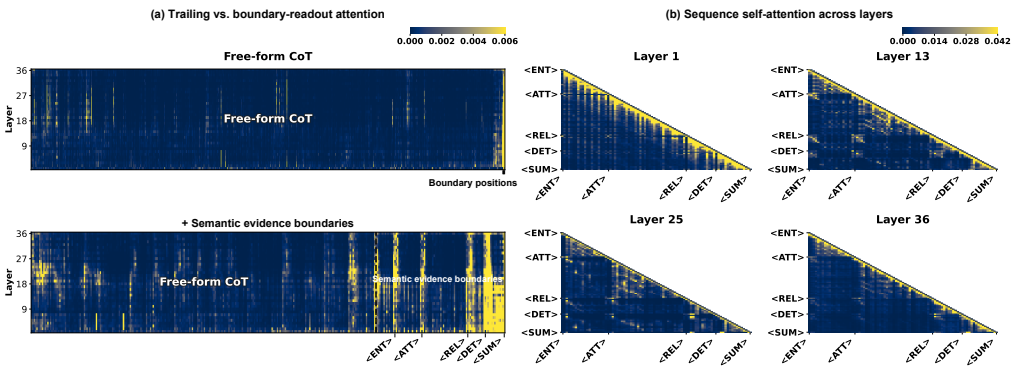


Figure 7: Attention patterns in the reasoning-prefix diagnostic. (a) Attention from the last evidence-boundary token (<SUM>) to preceding generated positions, compared with the trailing token of free-form CoT. (b) Layer-wise self-attention over sequence positions, with the five evidence-boundary positions marked. Panels (a) and (b) use separate color scales, shown above each panel.

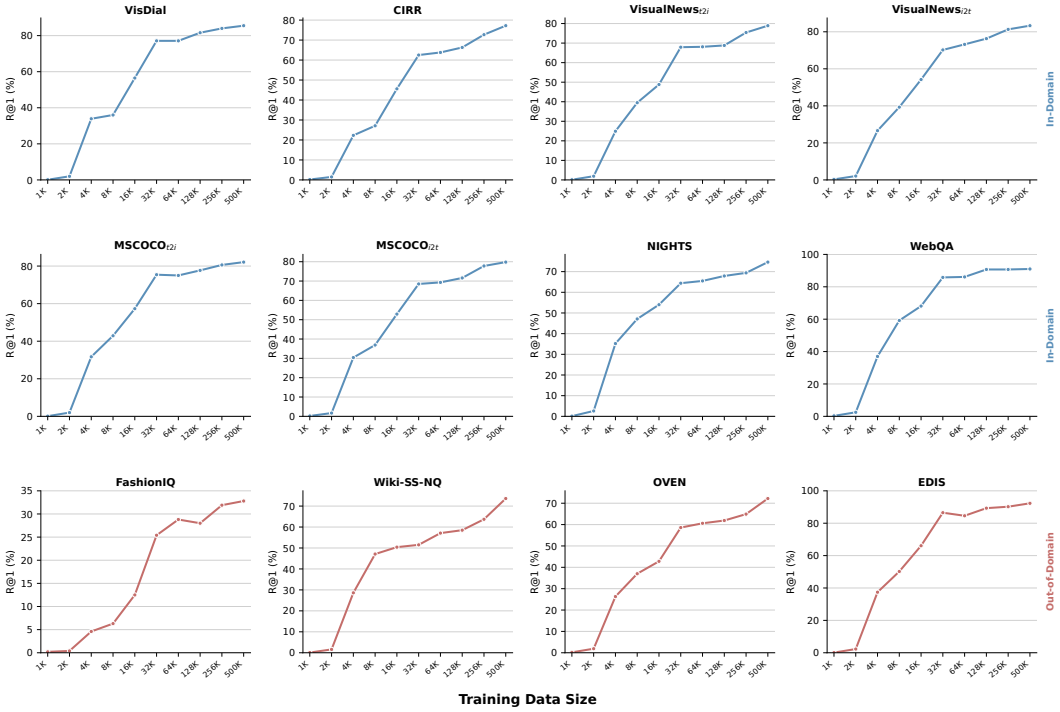


Figure 8: Data scaling curves of EviAlign on MMEB from 1K to 500K training pairs. We report Recall@1 across in-domain and out-of-domain retrieval benchmarks.

evidence configuration, and optimization settings. All models are evaluated on the same 12 MMEB retrieval tasks, including eight in-domain and four out-of-domain tasks.

As shown in Figure 8, performance remains low at 1K–2K training pairs. Inspection of generated outputs reveals frequent boundary-token omissions, causing the affected readouts to fall back to the final hidden state rather than the intended evidence boundaries. From 4K onward, the model progressively learns to follow the prescribed boundary-token format more reliably, allowing boundary-aligned readout to operate as intended. Alongside this transition, retrieval performance rises sharply and continues to improve with additional training data across both in-domain and out-of-domain tasks. Scaling from 256K to 500K further improves all 12 tasks, with particularly clear gains on CIRR, NIGHTS, Wiki-SS-NQ, and OVEN.

Table 13: Query-side representation-construction efficiency and retrieval quality. Recall@1 is averaged over the 12 MMEB retrieval tasks. Batch time measures end-to-end representation construction for four queries, including multimodal encoding, autoregressive generation when applicable, and embedding aggregation. Memory reports the peak GPU allocation under the same run.

Setting	Recall@1	Batch time	Tokens	Memory
VLM2Vec-V2 (2B)	69.5	67.8 ms	–	17.9 GB
GME-7B	71.2	89.1 ms	–	15.5 GB
Qwen3-VL-8B direct (ours)	68.30	94.6 ms	–	17.4 GB
EviAlign	**76.94**	1,502.5 ms	114.3	19.49 GB
Free-form CoT + semantic evidence	76.59	9,825.3 ms	702.1	44.46 GB

C.8 Complexity and Scalability Analysis

We analyze the two deployment stages at which EviAlign differs from conventional embedding systems: representation construction and index storage. The single-vector interface preserves index storage and similarity scoring; query-side evidence generation remains part of online encoding.

C.8.1 Representation-Construction Efficiency

EviAlign learns to generate semantic evidence under teacher forcing and reads the K evidence-boundary states to construct its retrieval representation. At inference, it generates the evidence and aggregates these states into one embedding. Table 13 profiles this representation-construction path and a control that generates free-form CoT before the same semantic evidence while retaining the boundary readout. Under BF16 inference on NVIDIA H20 GPUs with a per-device batch size of 4, EviAlign reaches 76.94 average Recall@1 with 114.3 generated tokens on average, an end-to-end representation-construction time of 1,502.5 ms per batch, and 19.49 GB peak memory per GPU. Adding free-form CoT before the evidence yields 76.59 while increasing these costs to 702.1 tokens, 9,825.3 ms, and 44.46 GB. Both variants produce a single vector scored with cosine similarity.

C.8.2 Index Storage

EviAlign aggregates its five internal readouts into one dense vector before indexing. For a corpus of $|\mathcal{C}|$ items and embedding dimension d, it therefore stores $|\mathcal{C}|d$ scalars, without a multiplicative dependence on the number of evidence units. Table 14 calculates raw FP16 embedding storage for 1M items from the vector count and dimensionality of each configuration; compression, index metadata, and system-level overhead are excluded. EviAlign uses $d = 4096$ with Qwen3-VL-8B and $d = 2048$ with Qwen3-VL-2B. ColPali (Faysse et al., 2025) and ColBERT are included as reference multi-vector configurations. EviAlign remains compatible with standard dense-retrieval systems such as FAISS (Johnson et al., 2021).

Table 14: Calculated raw FP16 embedding storage for 1M retrieval items under the listed vector counts and dimensions; index overhead and compression are excluded.

Method	Vecs/Doc	Dim	Storage/1M	Retrieval
ColPali	1,030	128	263.7 GB	MaxSim
ColBERT (text)	$\sim$30	128	7.7 GB	MaxSim
EviAlign (8B)	1	4,096	8.2 GB	Cosine
EviAlign (2B)	1	2,048	4.1 GB	Cosine

C.9 Evidence Schema Details

EviAlign organizes generated evidence into explicit units with corresponding readout boundaries. Our implementation uses five units—Entity, Attribute, Relation, Detail, and Summary—as a compact empirical configuration. Entity covers objects or concepts together with identifying properties;

Case 2 — VisualNews: Image-to-Caption Retrieval



Query image

A Maryland State Police cruiser sits at a blocked southbound entrance on the Baltimore–Washington Parkway that accesses Fort Meade.

Retrieved caption

Entity. brown rectangular NSA directional sign, white "Restricted Entrance" placard, Maryland State Police cruiser with emergency light bar, orange traffic cones

Attribute. outdoor, overcast daytime, rural road setting, law-enforcement / security checkpoint atmosphere

Relation. NSA sign and Restricted Entrance sign co-mounted on same roadside pole; police cruiser blocking road at entrance; traffic cones positioned along road edge

Detail. text: "NSA" with arrow on brown sign; text: "Restricted Entrance" on white sign; text: "Road Closed" on smaller placard; red/blue emergency light bar on police cruiser

Summary. Police cruiser blocking NSA restricted entrance on Baltimore–Washington Parkway near Fort Meade

Figure 9: VisualNews example: an image query, structured retrieval evidence, and the retrieved caption.

Attribute captures scene-level characteristics; Relation describes actions, interactions, and spatial relations; Detail retains locally discriminative cues such as visible text, signs, textures, and patterns; and Summary provides a retrieval-focused global description. These roles follow the data-construction prompt used in our experiments.

These units provide stable semantic roles and explicit boundaries shared by generation and representation readout. Section 4.4 evaluates joint evidence-schema and readout configurations and supports the selected implementation among the tested settings.

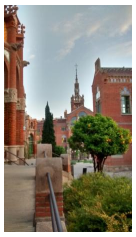
C.10 Case Studies

Figure 4 presents the CIRR composed-image example in the main text. Here we provide two additional cases from image-to-text retrieval on VisualNews and out-of-domain knowledge-grounded retrieval on OVEN. Each case presents the query and target first, followed by the generated evidence at readable single-column scale.

Case 2: Cross-Modal Image-to-Text Retrieval (VisualNews). Image-to-text retrieval connects visual evidence with a natural-language caption. Entity records the roadside signs, police cruiser, and traffic cones; Detail captures text such as "NSA," "Restricted Entrance," and "Road Closed"; and Relation describes the blocked entrance. Summary integrates these cues into a description consistent with the retrieved caption.

Case 3: Knowledge-Grounded Retrieval (OVEN). This case uses an out-of-domain query format that asks the model to identify a place from an image. The retrieved candidate is a knowledge entry with both image and text; the figure shows its image. Entity and Attribute describe the architectural components and Catalan Modernista style, Relation preserves the multi-pavilion layout, and Detail records tile patterns, mosaic spires, and sculptural reliefs. Summary integrates these local and global cues into a landmark-level description of the retrieved entry.

Case 3 — OVEN: Knowledge-Grounded Retrieval



Query image

"Where is this place?"

Query text



Retrieved entry image

Entity. ornate red-brick pavilion buildings with Catalan Modernista sculptural facades, stone railings with brick bases, tall tree with yellow blossoms, paved garden walkway

Attribute. outdoor, daytime, clear sky, historic Catalan Modernista architectural style, well-maintained heritage garden setting

Relation. multi-pavilion complex enclosing a central garden; main tower with decorated dome visible at far end; stone railings lining the central walkway; paved path leading toward main building

Detail. polychrome tile and brick ornamental motifs on pavilion facades; conical spire with mosaic finish on central dome; decorative sculptural reliefs above archways; geometric stone-and-brick railing bases

Summary. Hospital de Sant Pau–style Catalan Modernista brick complex with ornate multi-pavilion layout and central heritage garden

Figure 10: OVEN example: the query combines an image and a place-identification question. The retrieved candidate is an image–text knowledge entry; its image is shown.

D Prompt Templates

We document four categories of prompt templates used in experiments: (1) the inference prompt of EviAlign, applied identically to every query and candidate, (2) the data-construction prompt of EviAlign-Evidence, used to organize semantic-evidence targets, (3) the prompts used in evidence-configuration ablations, and (4) the judgment prompt used in the evidence-quality analysis.

D.1 Inference Prompt of EviAlign

The evidence-generation template is shared across all 12 MMEB datasets. {{`Multimodal Input`}} contains the input text and/or image together with its retrieval task instruction. The ASSISTANT OUTPUT block shows the structure of the model-generated evidence. The five boundary tokens `<ENT>`, `<ATT>`, `<REL>`, `<DET>`, and `<SUM>` are registered as new vocabulary items initialized from the `[EOS]` embedding; their final-layer states provide the five boundary readouts (Section 3.3).

Inference Prompt of EviAlign

USER

{{`Multimodal Input`}}

Represent the above input texts, images, or any combination of the two as embeddings. Directly summarize the input in a structured format, using the tokens `<ENT>`, `<ATT>`, `<REL>`, `<DET>`, `<SUM>` to respectively represent the Entity, Attribute, Relation, Detail, and Summary.

ASSISTANT OUTPUT

[Entity] {...} <ENT>
[Attribute] {...} <ATT>
[Relation] {...} <REL>
[Detail] {...} <DET>
[Summary] {...} <SUM>

D.2 Training Data Construction Prompt of EviAlign-Evidence

We use GLM-4.1V-9B-Thinking (GLM-V Team et al., 2025) to convert each training input from the MMEB-V1 (Jiang et al., 2025) retrieval subset into five-unit semantic-evidence targets. The resulting annotations are filtered to remove excessive repetition, invalid formats, or outputs exceeding the 8192-token hard safety limit and are associated with approximately 500K query–candidate training pairs used for supervised fine-tuning.

Training Data Construction Prompt of EviAlign-Evidence

USER

{{Multimodal Input }}

The above contains a retrieval input (text, image, or both). Use them for reference, but verify all details against the original input—the input may contain errors. Identify the retrieval task instruction (e.g., "find a matching image", "find a similar image with the described changes") and focus extraction on information most relevant to that task.

Extract key information and output **EXACTLY** in this format (5 lines only):

[Entity] *entities with bound identifying attributes; every entity needs at least one attribute (e.g., "orange tabby cat", not "cat")* <ENT>

[Attribute] *scene-wide attributes only: lighting, camera angle, weather, time of day, image style, mood* <ATT>

[Relation] *actions and spatial relationships (e.g., "cat sitting on counter", "bus parked at hillside")* <REL>

[Detail] *discriminative visual cues: visible text as text:''...'', logos as logo:''...''; also patterns, materials, signs, etc.* <DET>

[Summary] *concise summary of the retrieval target* <SUM>

Rules: (i) Output only the 5 lines—no commentary or analysis.
(ii) [Entity] binds attributes: "red dress" in Entity, not "dress" + "red" in Attribute.
(iii) [Attribute] is scene-wide only, never entity-specific.
(iv) [Detail] must be grounded in the actual input; output [] only if truly absent.
(v) [Summary] must be specific—avoid generic descriptions.

ASSISTANT

[Entity] {...} <ENT>
[Attribute] {...} <ATT>
[Relation] {...} <REL>
[Detail] {...} <DET>
[Summary] {...} <SUM>

D.3 Evidence-Configuration Ablation Prompts

For the $K=7$ ablation, we retain the five default unit names and insert Context (<CTX>, scene context) and Emotion (<EMO>, emotional tone) between Detail and Summary. Mood is assigned to Emotion rather than Attribute in this template. The $K=1$ and $K=3$ inference prompts remove the

unused units from the standard EviAlign template. Data construction and inference otherwise follow the same format.

Inference Prompt in Ablation Studies (K=7)

USER

{{Multimodal Input}}

Represent the above input texts, images, or any combination of the two as embeddings. Directly summarize the input in a structured format, using the tokens <ENT>, <ATT>, <REL>, <DET>, <CTX>, <EMO>, <SUM> to respectively represent the Entity, Attribute, Relation, Detail, Context, Emotion, and Summary.

ASSISTANT

[Entity] {...} <ENT>
[Attribute] {...} <ATT>
[Relation] {...} <REL>
[Detail] {...} <DET>
[Context] {...} <CTX>
[Emotion] {...} <EMO>
[Summary] {...} <SUM>

Training Data Construction Prompt in Ablation Studies (K=7)

USER

{{Multimodal Input }}

The above contains a retrieval input (text, image, or both). Use them for reference, but verify all details against the original input—the input may contain errors. Identify the retrieval task instruction (e.g., "find a matching image", "find a similar image with the described changes") and focus extraction on information most relevant to that task.

Extract key information and output **EXACTLY** in this format (7 lines only):

[Entity] *entities with bound identifying attributes; every entity needs at least one attribute (e.g., "orange tabby cat", not "cat")* <ENT>

[Attribute] *scene-wide attributes only: lighting, camera angle, weather, time of day, image style* <ATT>

[Relation] *actions and spatial relationships (e.g., "cat sitting on counter", "bus parked at hillside")* <REL>

[Detail] *discriminative visual cues: visible text as text:''...'', logos as logo:''...''; also patterns, materials, signs, etc.* <DET>

[Context] *situational or event background: where/when/what activity frames the scene (e.g., "outdoor festival", "sports match at stadium")—not entity-level traits* <CTX>

[Emotion] *mood, sentiment, or subjective feeling conveyed by the scene (e.g., "cheerful", "tense", "somber")—must not duplicate neutral scene descriptions in Attribute* <EMO>

[Summary] *concise summary of the retrieval target* <SUM>

Rules: (i) Output only the 7 lines—no commentary or analysis.

(ii) [Entity] binds attributes: "red dress" in Entity, not "dress" + "red" in Attribute.

(iii) [Attribute] is scene-wide only, never entity-specific.

(iv) [Context] covers narrative situation, not physical attributes; output [] if the scene has no discernible event context.

(v) [Emotion] covers evaluative or affective tone only; output [] if the scene is affectively neutral.

(vi) [Summary] must be specific—avoid generic descriptions.

ASSISTANT

[Entity] {...} <ENT>
[Attribute] {...} <ATT>
[Relation] {...} <REL>
[Detail] {...} <DET>
[Context] {...} <CTX>
[Emotion] {...} <EMO>
[Summary] {...} <SUM>

D.4 LLM-AS-JUDGE PROMPT

The judge independently evaluates each structured-evidence or free-form CoT training target for faithfulness, specificity, and retrieval relevance on a 1–5 scale. Per-input preferences are obtained by comparing the two scores for each dimension; equal scores are ties and are omitted from contested win rates.

LLM-as-Judge Prompt

USER

You are evaluating the quality of a training target for a multimodal retrieval model. The target provides intermediate textual supervision before the retrieval embedding is constructed.

Rate the following training target on three dimensions (1–5 scale):

1. Faithfulness (1–5): Does the target faithfully satisfy the retrieval requirements expressed by the input, including its text, images, and task instruction? Higher = more faithful to the input's retrieval requirements.

2. Specificity (1–5): Does the target contain specific, discriminative information that would help distinguish the correct match from similar candidates? Higher = more specific and discriminative.

3. Retrieval-Relevance (1–5): Is the target directly useful for matching this input to its correct target in a retrieval task? Higher = more retrieval-aligned; lower = contains verbose or irrelevant content.

{multimodal input}
Training target to evaluate:

{trace}

Output **ONLY** a JSON object:
{"faithfulness": <int>, "specificity": <int>, "retrieval_relevance": <int>}