SELF-BOOTSTRAPPED MULTIMODAL EMBEDDINGS FOR ROBUST TEXT-TO-MULTIMODAL RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-multimodal retrieval aims to retrieve documents containing both images and text in response to textual queries. Existing multimodal embedding models typically rely on large-scale heterogeneous training corpora, external teachers, or auxiliary data-generation models, while task-specific text-to-multimodal training data remain scarce. We present Self-Bootstrapped Multimodal Embeddings, a framework that exploits the backbone MLLM's own generation, reasoning, and retrieval capabilities to improve multimodal representations without external teachers or auxiliary annotation models during retrieval training. First, the backbone generates detailed descriptions for document images, self-annotating existing text-to-image corpora and converting them into text-enriched multimodal retrieval data. Second, reasoning instruction conditioning during contrastive training elicits and internalizes the model's native reasoning-oriented computation into single-vector embeddings, requiring neither explicit reasoning traces nor additional reasoning tokens at inference. Finally, the trained retriever uses its own retrieval capability to identify highly ranked non-relevant documents and adopts them as self-mined hard negatives for targeted representation refinement. The resulting framework improves the average performance of both Qwen3-VL-Instruct and VLM2Vec-V2 on text-to-multimodal and text-to-video retrieval. With only 300K training pairs, our VLM2Vec-V2-based model achieves an average Recall@1 of 81.3% across six text-to-multimodal benchmarks—the best among models fine-tuned with no more than 300K task-specific pairs—and reaches 94.1% on EDIS. We further introduce EDIS-Noisy and WebQA-Noisy to simulate realistic retrieval scenarios. Without noise-specific training, our model achieves average Recall@1 scores of 91.4% and 80.4%, respectively, outperforming both Qwen3-VL-Embedding and WeMM-Embedding. These results demonstrate that self-bootstrapping the backbone's existing capabilities provides an effective approach to data-efficient and robust multimodal retrieval.

1 INTRODUCTION

Text-to-multimodal retrieval—searching for documents that contain both images and text in response to a text query—is central to modern search engines, knowledge bases, and question-answering systems. Recent advances in multimodal large language models (MLLMs) have enabled a new generation of embedding models that encode text queries and multimodal documents into a unified latent space (Jiang et al., 2024b; Meng et al., 2025; Li et al., 2026). However, state-of-the-art methods typically rely on large-scale pretraining followed by extensive fine-tuning on multimodal datasets (Meng et al., 2025; Zhang et al., 2024b; Lan et al., 2025b; Jiang et al., 2026; Wu et al., 2026b). For instance, Embed-RL (Jiang et al., 2026) trains on a large-scale corpus spanning millions of image-text pairs, while models like Qwen3-VL-Embedding and WeMM-Embedding benefit from massive multimodal pretraining and finetuning (Li et al., 2026; WeMM Team, 2025).

This reliance on scale is particularly problematic for text-to-multimodal retrieval, where task-specific training data remain scarce. Existing datasets whose documents contain semantically related images and text are limited in size; for example, WebQA (Chang et al., 2022) contains only tens of thousands of training examples. Existing methods compensate by incorporating auxiliary image-query retrieval tasks over large-scale image-text corpora such as CC3M (Sharma et al., 2018; Li et al., 2026), incurring substantial computational cost without directly optimizing the target task. We take a

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.

different perspective: a pretrained MLLM already possesses the generation, reasoning, and retrieval capabilities needed to bootstrap its own task-specific embedding training. Rather than relying on large-scale heterogeneous supervision or external teachers, we investigate how these inherent capabilities can be elicited and coordinated to improve multimodal representations with limited task-specific supervision. Our framework addresses this problem from these three complementary perspectives.

Self-generated descriptions for document images. We exploit the base MLLM's generation capability to bridge the mismatch between abundant text-to-image data (Sharma et al., 2018) and the target text-to-multimodal setting. For each document image, the MLLM generates a detailed description rather than a simple caption, providing richer and more effective document-side textual supervision. In this way, the model self-annotates existing text-to-image pairs and converts them into text-enriched multimodal training documents without relying on a stronger teacher or auxiliary annotation model. These self-generated descriptions work in concert with reasoning instructions to enrich document-side supervision and elicit reasoning-oriented representations, yielding an initial text-to-multimodal retrieval model.

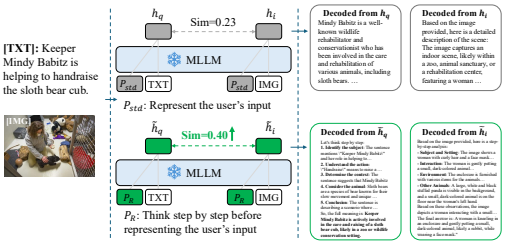


Figure 1: Reasoning instructions can effectively modulate the embeddings produced by pretrained MLLMs (e.g., Qwen3-VL-Instruct). The top and bottom panels use the standard and reasoning instructions, respectively; even before retrieval fine-tuning, the reasoning instruction yields finer-grained hidden representations and better query-positive alignment.

Native reasoning training for embedding learning. Standard embedding models typically use generic instructions to produce single-vector representations. Existing reasoning-enhanced methods often require explicit reasoning-token generation or distillation from a stronger teacher (Cui et al., 2025; Wu et al., 2026a; Liu et al., 2025a). We instead elicit the MLLM's native reasoning through a reasoning instruction. As shown in Fig. 1, this instruction yields finer-grained hidden representations and better query–positive alignment even before retrieval fine-tuning. Our subsequent experiments further show that training with reasoning instructions enlarges the similarity margin between positive documents and hard negatives, indicating that native reasoning helps capture fine-grained distinctions rather than coarse shared semantics. We therefore use reasoning instructions during contrastive training to internalize this capability into single-vector embeddings, without external reasoning supervision or additional reasoning tokens at inference.

Self-retrieval representation refinement. The resulting model is then used to refine its own representation space. While random in-batch negatives encourage overall query–document alignment, they provide limited supervision for distinguishing among confusable multimodal documents. We therefore treat highly ranked but non-relevant documents retrieved by the model from the enriched corpus as hard negatives. Further contrastive training on these self-mined examples directly targets the model's own retrieval errors, improving fine-grained document discrimination without additional annotations or external mining models.

We evaluate our framework on six standard text-to-multimodal retrieval benchmarks from MMEB (Jiang et al., 2024b)—VisualNews, MSCOCO, VisDial, Wiki-SS-NQ, EDIS, and WebQA—as well as five text-to-video retrieval benchmarks (e.g., DiDeMo, MSRVTT, and MSVD). Our framework generalizes across both Qwen3-VL-Instruct and VLM2Vec-V2, consistently improving their average performance on text-to-multimodal and text-to-video retrieval. Built on VLM2Vec-V2 and trained with only small-scale data, our model achieves an average Recall@1 of 81.3% across the six text-to-multimodal benchmarks, outperforming several methods trained on substantially larger datasets. On EDIS, it achieves 94.1% Recall@1, surpassing the large-scale WeMM-Embedding (93.3%) and Qwen3-VL-Embedding (92.7%).

To assess robustness under noisy conditions commonly encountered in real-world retrieval, we construct EDIS-Noisy and WebQA-Noisy by introducing three types of query-level perturbations and four types of document-level perturbations. Our model establishes new state-of-the-art results on these challenging benchmarks. In particular, on EDIS-Noisy, it outperforms the large-scale Qwen3-VL-Embedding and WeMM-Embedding models by 5.6 and 6.5 Recall@1 points, respectively.

These results demonstrate that our framework produces multimodal representations that are both accurate and robust. Our contributions are three-fold:

1. We propose Self-Bootstrapped Multimodal Embeddings, a framework that exploits the backbone MLLM's own generation, reasoning, and retrieval capabilities. The MLLM self-annotates document images, internalizes its native reasoning into embeddings, and refines its representations using self-mined retrieval errors.
2. We construct EDIS-Noisy and WebQA-Noisy based on existing multimodal retrieval benchmarks to simulate noisy conditions in real-world retrieval. The two benchmarks cover three query-level and four document-level perturbations, evaluating robustness to incomplete, corrupted, and cross-modally inconsistent inputs.
3. Extensive experiments demonstrate the generalizability of our framework across Qwen3-VL-Instruct and VLM2Vec-V2, six text-to-multimodal and five text-to-video benchmarks. Without noise-specific training, our model also outperforms the large-scale Qwen3-VL-Embedding and WeMM-Embedding models on both EDIS-Noisy and WebQA-Noisy.

2 Related Work

2.1 Universal Multimodal Embedding Models

Universal multimodal embedding models map heterogeneous inputs into a shared representation space for any-to-any retrieval. Early approaches mainly adopt dual-encoder architectures and contrastive learning (Oord et al., 2018), including CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), and BLIP-2 (Li et al., 2023). UniIR (Wei et al., 2024) unifies diverse retrieval tasks within a single framework, while MagicLens (Zhang et al., 2024a) trains self-supervised image-retrieval models using large-scale web-mined data. However, conventional dual encoders are often tied to fixed input formats, limiting their flexibility across heterogeneous multimodal retrieval scenarios.

Recent work therefore adapts multimodal large language models (MLLMs) into universal embedding models (Jiang et al., 2024a;b; Lin et al., 2025). These studies improve multimodal embeddings through training strategies (Jiang et al., 2024b; Liu et al., 2025b), data synthesis (Zhou et al., 2025b), architectural design (Chen et al., 2026), and knowledge distillation (Lan et al., 2025a). VLM2Vec (Jiang et al., 2024b) introduces MMEB and evaluates MLLM-based embeddings across diverse multimodal tasks, while GME (Zhang et al., 2025b) explores scalable multimodal data construction. ColPali (Faysse et al., 2025) adopts multi-vector representations for visual document retrieval, MoCa (Chen et al., 2026) introduces bidirectional attention for scalable embedding learning, LLaVE (Lan et al., 2025a) proposes hardness-weighted contrastive learning, and UniME (Gu et al., 2025) distills a bidirectional teacher into a causal-attention student.

2.2 Reasoning Embedding Models

Inspired by advances in reasoning models (Shao et al., 2024) and multimodal reasoning (Xu et al., 2025), recent studies incorporate reasoning into embedding learning through either explicit (Cui et al., 2025) or implicit (He et al., 2026) mechanisms. Explicit methods generate or transform textual reasoning before embedding extraction. Think-Then-Embed (Cui et al., 2025) relies on an external MLLM to produce reasoning tokens, whereas RGE (Liu et al., 2025a) uses the embedding model itself. UME-R1 (Lan et al., 2025b) and MOON3.0 (Wu et al., 2026a) combine CoT supervised fine-tuning with GRPO-based reinforcement learning (Shao et al., 2024), while Embed-RL (Jiang et al., 2026) jointly optimizes reasoning and embedding capabilities. TRACE (Hao et al., 2026) adaptively controls the amount of reasoning, and RIME (Wu et al., 2026b) rewrites queries before embedding extraction. Implicit methods avoid generating explicit reasoning text at inference. PLUME (He et al., 2026) performs iterative latent-state transitions with progressive explicit-to-latent training, while TTE-Flash (Cheng et al., 2026) trains latent tokens using CoT generation and contrastive objectives. LaME (Wu et al., 2026c) instead confines reasoning to bottleneck tokens, enabling single-pass embedding extraction. In contrast, our method requires neither explicit reasoning-token generation nor dedicated multi-stage latent-token training, but directly exploits the intrinsic reasoning capability of the underlying MLLM to improve its representations.

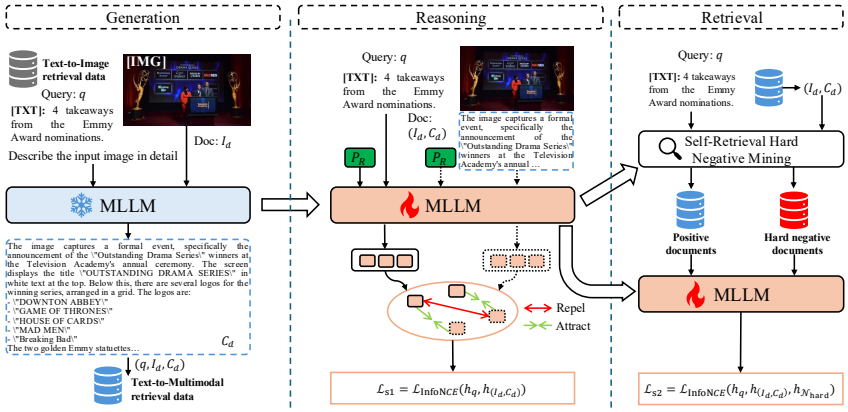


Figure 2: Overview of the self-bootstrapped training framework. Left: the pretrained MLLM generates detailed descriptions of document images to enrich text-to-image pairs. Middle: reasoning instructions elicit its reasoning capability during contrastive embedding training. Right: retrieval-based hard-negative mining refines the learned representations.

2.3 Multimodal Retrieval Benchmarks

Following text embedding benchmarks such as MTEB (Muennighoff et al., 2023), recent work has expanded multimodal evaluation across diverse domains, modalities, and tasks. MBEIR (Wei et al., 2024) evaluates heterogeneous query formats and task instructions across eight tasks and sixteen datasets, while MMEB (Jiang et al., 2024b) covers thirty-six datasets spanning retrieval, visual question answering, and visual grounding. MIEB (Xiao et al., 2025) further includes multimodal clustering and visual semantic textual similarity, and MMEB-V2 (Meng et al., 2025) extends evaluation to video. More recent benchmarks emphasize reasoning-intensive retrieval: MR^2-Bench (Zhou et al., 2025a) focuses on vision-centric reasoning, MRMR (Zhang et al., 2025a) targets expert-level retrieval across academic and professional domains, and ARK (Lin et al., 2026) examines the roles of multimodal knowledge and reasoning.

For text-to-multimodal retrieval, EDIS (Liu et al., 2023) and WebQA (Chang et al., 2022) are widely used benchmarks. EDIS contains entity-rich textual queries and multimodal candidates, whereas WebQA requires retrieving multimodal evidence for natural-language questions. However, these benchmarks primarily evaluate retrieval under relatively clean conditions and provide limited evidence of robustness to noise commonly encountered in real-world applications. To address this gap, we construct EDIS-Noisy and WebQA-Noisy by introducing three types of query-level perturbations and four types of document-level perturbations. Accordingly, we evaluate our method on both standard text-to-multimodal retrieval benchmarks and the proposed noisy benchmarks, assessing retrieval effectiveness under clean conditions and robustness to query- and document-level noise.

3 Method

We present a two-stage framework that bootstraps a pretrained MLLM's own generation, reasoning, and retrieval capabilities for data-efficient text-to-multimodal embedding learning as shown in Fig. 2. Given an MLLM (e.g., Qwen3-VL-Instruct), the first stage trains the embedding model with reasoning instructions and self-generated descriptions for document images. The resulting model uses its own retrieval capability to mine hard negatives and refine its representation space.

3.1 Preliminaries

We consider the text-to-multimodal retrieval task where a text query q is matched against a document d that may contain an image, text, or both. An embedding model $f_{\boldsymbol{\theta}}(\cdot)$ parameterized by $\boldsymbol{\theta}$ encodes both queries and documents into a shared h-dimensional embedding space. The relevance score is computed as cosine similarity: $\text{sim}(q, d) = \langle f_{\boldsymbol{\theta}}(q), f_{\boldsymbol{\theta}}(d) \rangle / (\|f_{\boldsymbol{\theta}}(q)\| \, \|f_{\boldsymbol{\theta}}(d)\|)$, and retrieval is performed via nearest-neighbor search.

We adopt the standard contrastive learning objective with in-batch negatives. Given a training batch $\mathcal{B} = \{(\mathbf{q}_i, \mathbf{d}_i)\}_{i=1}^{N}$ of query-document pairs, the InfoNCE loss (Oord et al., 2018) is:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(\mathbf{q}_i, \mathbf{d}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{q}_i, \mathbf{d}_j)/\tau)}, \tag{1}$$

where τ is a learned or fixed temperature parameter.

3.2 Reasoning Instruction Embedding Training with Self-Generated Descriptions

3.2.1 Self-Generated Descriptions

A major challenge in text-to-multimodal retrieval is the scarcity of training data where documents contain interleaved images and text. We observe that text-to-image (T2I) retrieval corpora (e.g., CC3M) are abundant and provide high-quality query–image alignment. Although training on image-only documents can already yield effective representations, it does not fully exploit the semantic information available on the document side and remains mismatched with the multimodal documents encountered at test time.

As a result, there remains substantial room for improvement on text-to-multimodal retrieval. We bridge this distribution gap by enriching documents in T2I datasets. Specifically, for each document image I_d in a T2I corpus, we use a pretrained MLLM $\mathcal{M}$ (e.g., Qwen3-VL-Instruct) to generate a detailed description $C_d = \mathcal{M}(I_d; \mathbf{p}_{\text{S}})$, where $\mathbf{p}_{\text{S}}$ denotes "Describe the input image in detail."

The document is then represented as a text-enriched multimodal document $\tilde{d} = (I_d, C_d)$, where C_d serves as the document's text component. For each retriever, descriptions are generated by its underlying pretrained generative MLLM offline before training. The descriptions are independent per image and do not refer to the query. This preserves the retrieval task's challenge: the model must still learn to match the query to the joint image and text representation.

3.2.2 Reasoning Instruction Training

Standard embedding training typically uses a fixed instruction to encode both queries and documents. During training, we replace the standard instruction with a reasoning instruction: $\mathbf{p}_{\text{R}}$ = "Think step by step before representing the input."

The key intuition is that a reasoning instruction activates the MLLM's native reasoning capabilities during the forward pass, encouraging the model to produce richer and more structured hidden representations, as illustrated in Fig. 1. Unlike explicit reasoning-based approaches, our method does not require the autoregressive generation of intermediate reasoning tokens. Instead, the induced reasoning process is implicitly encoded in the hidden states and subsequently compressed into a single-vector embedding.

Given a text query q and a text-enriched multimodal document $\tilde{d} = (I_d, C_d)$, where I_d denotes the document image and C_d denotes its generated description, their training-time representations are computed as:

$$\mathbf{h}_q = f_{\boldsymbol{\theta}}(q; \mathbf{p}_{\text{R}}), \qquad \mathbf{h}_{(I_d, c_d)} = f_{\boldsymbol{\theta}}(\tilde{d}; \mathbf{p}_{\text{R}}). \tag{2}$$

In this stage, we combine reasoning instruction training with self-generated descriptions. This allows the model to capture both visual content and explicit textual semantics. The training objective is defined as follows:

$$\mathcal{L}_{\text{s1}} = \mathcal{L}_{\text{InfoNCE}}(\mathbf{h}_q, \mathbf{h}_{(I_d, c_d)}), \tag{3}$$

with all other documents in the same mini-batch treated as in-batch negatives.

Our method internalizes reasoning-oriented computation into the embedding function during training. Therefore, at inference time, both queries and documents can be encoded with the standard prompt, $\mathbf{p}_{\text{std}}$ = "Represent the user's input.", without relying on reasoning instructions or generating additional reasoning tokens.

Table 1: Main results on text-to-multimodal retrieval benchmarks. The model defaults to 2B parameters, or the parameters specified in the name. † denotes our reproduction.

Model	Train Data	VisualNews	MSCOCO	VisDial	Wiki-SS-NQ	EDIS	WebQA	Avg
Closed-Source or Large-scale training								
WeMM-Emb (WeMM Team, 2025)	–	76.9	77.6	85.1	83.9	93.3	89.3	84.7
Qwen3-VL-Emb (Li et al., 2026)	–	73.0	77.8	81.3	83.6	92.7	90.1	83.0
ColPali-v1.3 (Faysse et al., 2025)	–	50.5	58.6	41.1	**80.5**	65.2	58.6	59.1
VLM2Vec-V1 (Jiang et al., 2024b)	662K	73.4	73.1	75.1	54.6	81.4	85.8	73.9
GME (Zhang et al., 2024b)	8M	74.8	68.3	47.7	73.8	91.7	88.8	74.1
UME-R1 (Lan et al., 2025b)	1M	71.1	75.1	76.6	62.0	88.0	90.0	77.1
ReMatch-3B (Liu et al., 2026)	662K	75.1	76.5	**83.6**	66.3	82.9	89.9	79.0
EMbed-RL (Jiang et al., 2026)	2M	71.9	**79.4**	81.5	68.9	84.5	89.3	79.2
RGE (Liu et al., 2025a)	662K	**77.5**	76.7	79.8	62.9	90.1	88.4	79.2
VLM2Vec-V2 (Meng et al., 2025)	–	74.7	75.9	82.7	67.7	84.2	90.6	79.3
RIME (Wu et al., 2026b)	1.5M	72.4	76.1	80.7	68.9	88.9	90.7	79.6
Ops-MM-embedding-v1 (OpenSearch-AI Team, 2025)	–	76.2	74.1	79.4	68.8	90.6	**91.1**	80.0
Qwen3-VL-Ins† (Bai et al., 2025) (Baseline)	100K	64.0	42.2	22.2	49.3	77.1	61.3	52.7
+ Ours	100K	69.2	55.6	40.0	68.3	90.6	80.9	67.4
VLM2Vec-V2†	100K	76.0	44.8	36.9	62.7	90.7	73.1	64.0
+ Ours	100K	77.3	48.2	37.3	63.5	92.6	68.2	64.5
Qwen3-VL-Ins† (Baseline)	300K	68.1	62.2	80.1	67.6	86.6	83.7	74.7
+ Ours	300K	70.3	73.8	81.0	67.9	90.4	80.2	77.3
ReMatch† (Liu et al., 2026)	300K	73.5	75.2	76.8	66.5	92.0	87.6	78.6
VLM2Vec-V2†	300K	**77.5**	76.7	81.5	69.4	86.0	89.1	80.0
+ Ours	300K	77.2	76.5	79.2	70.7	**94.1**	90.2	**81.3**

3.3 Self-Retrieval Representation Refinement

While Stage 1 produces strong embedding models, standard contrastive learning with random in-batch negatives does not specifically prepare the model for difficult distractors. In realistic retrieval scenarios, many documents are superficially similar to the correct answer but differ in key details, such as contradictory facts or partially matching phrases.

To address this issue, this stage introduces hard negative mining. Unlike prior approaches that mine negatives using unimodal documents, we perform hard negative mining over multimodal documents, jointly accounting for the fine-grained semantics conveyed by both the image and the accompanying text. This enables us to identify challenging multimodal examples that expose the remaining weaknesses of the Stage-1 model and provide more informative supervision for further optimization. Specifically, we use the Stage-1 model to encode all documents in the training corpus and, for each query $\mathbf{q}_i$, retrieve the top-K documents with the highest cosine similarity that are not the ground-truth document $\mathbf{d}_i$:

$$\mathcal{N}_{\text{hard}}(\mathbf{q}_i) = \underset{\mathbf{d}\in\mathcal{D}\setminus\{\mathbf{d}_i\},\text{top-}K}{\arg\max} \ \text{sim}(\mathbf{q}_i, \mathbf{d}). \tag{4}$$

We construct new training batches pairing each query with its ground-truth document and k hard negatives, then optimize a modified InfoNCE loss that explicitly contrasts against these negatives:

$$\mathcal{L}_{\text{s2}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{q}_i, \mathbf{d}_i)/\tau)}{\exp(\text{sim}(\mathbf{q}_i, \mathbf{d}_i)/\tau) + \sum_{\mathbf{d}\in\mathcal{N}_{\text{hard}}(\mathbf{q}_i)} \exp(\text{sim}(\mathbf{q}_i, \mathbf{d})/\tau)}. \tag{5}$$

Specifically, self-retrieval enables the model to mine its own highly ranked but non-relevant documents as hard negatives, directly exposing its retrieval errors and providing targeted supervision for fine-grained representation refinement.

3.4 Noisy Benchmark Construction

Existing benchmarks largely assume clean queries and aligned image–text documents, whereas practical retrieval often involves incomplete, corrupted, or mismatched inputs. We use Qwen3.5-27B (Qwen Team, 2026) to construct EDIS-Noisy and WebQA-Noisy while retaining the original candidate pools and relevance labels. Manual inspection confirms perturbation validity, target preservation, and no target-information leakage; details are provided in the supplementary material.

Query-side perturbations. For each query, we generate three variants conditioned on the original query and its positive document context: (i) *short queries*, which retain only one to five salient entities

or phrases; (ii) *noisy queries*, which introduce realistic typos, grammatical errors, colloquial expressions, or word-order changes; and (iii) *synonym queries* using semantically equivalent expressions. Positive document context is used only to preserve the original retrieval target.

Document-side perturbations. Given the original image–text pair, we construct four text variants: (i) *partial tags*, containing only two to six keywords or short phrases; (ii) *noisy text*, where approximately half of the original text is corrupted by OCR artifacts, misspellings, garbled fragments, or irrelevant metadata; (iii) *conflicting text*, which remains plausible but contradicts one to three visible attributes; and (iv) *shuffled text*, obtained by randomly pairing an image with text from another document. These perturbations test robustness to incomplete information, surface noise, lexical variation, and cross-modal inconsistency. Tab. 2 reports the average query/document text length, noisy-to-original text similarity, and noisy-text-to-image similarity for each split.

Table 2: Statistics of EDIS-Noisy and WebQA-Noisy.

EDIS/WebQA	Word Num.	Text Sim. with GT	Image–Text Sim.
Original Query	19.84/16.14	–	–
Short Query	5.67/6.31	0.82/0.86	–
Noisy Query	19.49/15.76	0.96/0.95	–
Synonym Query	19.43/16.12	0.82/0.92	–
Original Doc Text	13.13/13.71	1.00/1.00	0.36/0.54
Partial Doc Text	4.49/4.31	0.84/0.88	0.38/0.53
Noisy Doc Text	19.42/19.81	0.92/0.91	0.39/0.59
Conflicting Doc Text	13.00/13.09	0.59/0.66	0.35/0.45
Shuffled Doc Text	13.13/13.71	–	–

Table 3: Generalization results in Text-to-Video retrieval.

Method	DiDeMo	MSRVTT	MSVD	YouCook2	VATEX	Avg
Closed-Source or Large-scale training						
WeMM-Emb (WeMM Team, 2025)	64.1	54.2	71.8	32.6	49.9	54.5
Qwen3-VL-Emb (Li et al., 2026)	51.5	49.7	71.0	24.9	42.9	48.0
VLM2Vec-V1 (Jiang et al., 2024b)	19.1	25.6	37.5	4.37	15.7	20.5
ColPali-v1.3 (Faysse et al., 2025)	22.5	17.7	45.3	5.2	16.7	21.5
GME (Zhang et al., 2024b)	21.5	27.0	47.3	7.9	23.1	25.4
VLM2Vec-V2 (Meng et al., 2025) (w/ video)	29.9	27.8	47.3	10.6	26.1	28.3
VLM2Vec-V2† (w/o video)	26.9	24.9	38.8	3.7	23.9	23.6
+ Ours	28.5	28.8	43.5	6.2	25.6	26.5
Qwen3-VL-Ins† (Bai et al., 2025)	33.6	**42.1**	56.5	12.2	30.0	34.9
+ Ours	**34.8**	40.2	**60.0**	**12.9**	**31.7**	**35.9**

4 Experiment

4.1 Experimental Setting

Our training data are drawn from publicly available text-to-image retrieval corpora. We fine-tune two baselines, Qwen3-VL-Instruct and VLM2Vec-V2, under two data settings: 100K training examples from VisualNews-T2I (Liu et al., 2021), and 300K training examples from VisualNews-T2I (Liu et al., 2021), MSCOCO-T2I (Lin et al., 2014), VisDial (Das et al., 2017), and WebQA (Chang et al., 2022). We report Recall@1 on each benchmark. Under the 100K setting, we also summarize out-of-distribution (OOD) generalization using the average Recall@1 across the OOD benchmarks. Detailed training configurations and benchmark examples are provided in the Appendix A.

Table 4: Evaluation on the EDIS-Noisy benchmark. SQ: Short Query, NQ: Noisy Query, SynQ: Synonyms Query, PD: Partial Doc Text, ND: Noisy Doc Text, CD: Conflict Doc Text, SD: Shuffled Doc Text.

Method	Origin	SQ	NQ	SynQ	PD	ND	CD	SD	Avg (Noisy)
Qwen3-VL-Emb	92.7	80.5	92.5	90.4	88.8	89.7	84.6	74.1	85.8
WeMM-Emb	93.3	81.6	93.1	92.3	93.6	92.0	81.8	59.7	84.9
Qwen3-VL-Ins† (Baseline)	77.1	61.5	76.4	74.9	78.8	77.1	59.8	48.2	68.1
+ Ours	90.6	76.7	89.8	89.0	90.2	89.7	85.8	74.7	85.1
VLM2Vec-V2†	86.0	77.0	85.1	84.9	87.3	86.0	76.5	47.3	77.7
+ Ours	**94.1**	**86.0**	**93.8**	**93.4**	**94.4**	**93.6**	**91.5**	**87.0**	**91.4**

Table 5: Evaluation on the WebQA-Noisy benchmark.

Method	Origin	SQ	NQ	SynQ	PD	ND	CD	SD	Avg (Noisy)
Qwen3-VL-Emb	90.1	**89.0**	**90.0**	**89.6**	**88.4**	**89.0**	67.8	25.9	77.1
WeMM-Emb	89.3	85.5	88.2	89.0	87.6	87.9	59.2	27.5	75.0
Qwen3-VL-Ins† (Baseline)	61.3	56.3	61.1	59.6	60.3	57.7	39.5	9.5	49.1
+ Ours	80.9	78.0	81.6	80.1	78.3	77.7	63.2	39.0	71.1
VLM2Vec-V2†	89.1	87.9	89.0	89.1	87.0	87.0	60.1	10.1	72.9
+ Ours	**90.2**	88.0	89.7	89.0	87.5	88.0	**74.0**	**46.6**	**80.4**

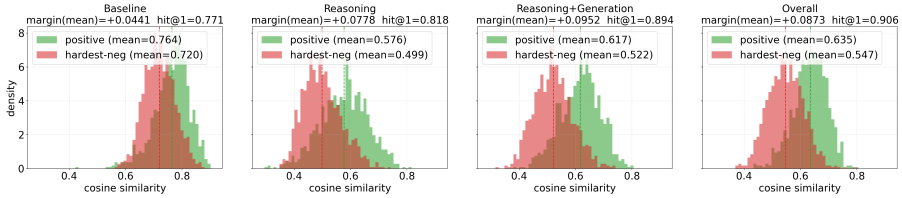


Figure 3: Similarity distributions of learned embeddings on EDIS. Our framework increases the query-target similarity gap between positive and hard negative pairs, indicating its effectiveness in helping the model learn more discriminative representations.

4.2 Main Results

Tab. 1 summarizes the text-to-multimodal retrieval results. With 100K training pairs, our framework improves Qwen3-VL-Ins from 52.7 to 67.4 average Recall@1 and yields consistent gains on VLM2Vec-V2, demonstrating its effectiveness across different MLLM-based embedding backbones. Our model achieves an average Recall@1 of 81.3 across six benchmarks—the best among models fine-tuned with comparable scale data, and higher than several methods trained on much larger corpora such as RIME (79.6) and EMbed-RL (79.2). On EDIS it reaches 94.1, surpassing the large-scale WeMM-Emb (93.3) and Qwen3-VL-Emb (92.7).

Tab. 3 reports the text-to-video retrieval results. The two backbones exhibit different strengths, with Qwen3-VL-Ins outperforming VLM2Vec-V2 overall, suggesting that video retrieval depends substantially on the pretrained multimodal knowledge of the underlying MLLM. Nevertheless, our framework improves both backbones, with average gains of 1.0 and 2.9 points, respectively.

4.3 Robustness on Noisy Benchmarks

As shown in Tab. 2, WebQA has higher image–text similarity than EDIS (0.54 vs. 0.36). Tabs. 4 and 5 show that model rankings remain broadly consistent under noise. On EDIS-Noisy, shuffled documents and short queries cause the largest drops, highlighting the impact of irrelevant document text and reduced query information. On WebQA-Noisy, shuffled and conflict perturbations are particularly challenging because they disrupt the stronger image–text correspondence. Our method achieves the best average performance on both benchmarks, reaching 91.4 on EDIS-Noisy and 80.4 on WebQA-Noisy.

4.4 Ablation Analysis

We conduct controlled ablations to examine the key components and design choices of our framework. Unless otherwise specified, all experiments use Qwen3-VL-Instruct-2B trained on the VisualNews-t2i training set and report Recall@1.

Effect of each component. Tab. 6 ablates reasoning instruction training (Rea), self-generated description (Gen), and self-retrieval representation refinement (Ret). Rea provides the largest individual gain, improving the OOD average from 50.4 to 62.6, while Gen and Ret alone reach 53.7 and 51.3, respectively. The components are complementary: $Rea + Gen$ achieves 65.2, and adding Ret further improves the result to 67.1. As illustrated in Fig. 3, Rea increases the separation between positive and hard-negative documents, Gen further enlarges this margin. Overall, Rea is the primary source of improvement, while Gen and Ret provide complementary gains in document discrimination and alignment.

Table 6: Ablation study of key components.

Rea	*Gen*	*Ret*	VisualNews (IID)	OOD Avg.
			64.0	50.4
✓			68.8	62.6
	✓		65.4	53.7
		✓	63.9	51.3
✓	✓		69.3	65.2
✓	✓	✓	69.2	67.1

Table 7: Ablation study of different prompts.

Train	Eval	VisualNews (IID)	OOD Avg.
$\mathbf{p}_{\text{std}}$	$\mathbf{p}_{\text{std}}$	64.0	50.4
$\mathbf{p}_{\text{std}}$	$\mathbf{p}_{\text{R}}$	63.4	50.4
$\mathbf{p}_{\text{R}}$	$\mathbf{p}_{\text{std}}$	68.8	62.6
$\mathbf{p}_{\text{R}}$	$\mathbf{p}_{\text{R}}$	69.9	62.5
$\mathbf{p}_{\text{desc}}$	$\mathbf{p}_{\text{std}}$	63.2	48.4
$\mathbf{p}_{\text{ent}}$	$\mathbf{p}_{\text{std}}$	66.1	50.1
$\mathbf{p}_{\text{reason}}$	$\mathbf{p}_{\text{std}}$	70.1	62.0
$\mathbf{p}_{\text{think}}$	$\mathbf{p}_{\text{std}}$	69.1	62.7

Table 8: Ablation study of different document texts in reasoning-instruction training.

Doc Text	VisualNews (IID)	OOD Avg.
Baseline	64.0	50.4
None	68.8	62.6
Placeholder text	65.0	51.7
Short caption	64.1	54.3
Detailed description	69.3	65.2
+ 50% noisy	68.1	62.3
+ 40% noisy	67.6	63.1
+ 30% noisy	68.1	63.5
+ 20% noisy	68.6	64.9
+ 10% noisy	67.5	65.0

Table 9: Ablation study on noisy benchmarks.

Rea	*Gen*	*Ret*	EDIS	EDIS-Noisy	WebQA	WebQA-Noisy
			86.0	77.7	89.1	72.9
✓			85.2	76.0	89.1	72.4
✓	✓		93.8	89.5	90.0	77.8
✓	✓	✓	94.1	91.4	90.2	80.4

Different instructions during training and inference. Tab. 7 separates the effects of reasoning instruction during training and inference. Applying a reasoning instruction only at inference does not improve a model trained with the standard prompt, leaving the OOD average unchanged at 50.4. In contrast, training raises the OOD average to 62.6. After training, the standard and reasoning instructions perform comparably at inference (62.6 vs. 62.5), suggesting that reasoning-oriented computation has been internalized by the model. We further compare several prompt variants: $\mathbf{p}_{\text{desc}}$ = "Describe the image.", $\mathbf{p}_{\text{ent}}$ = "Focus on key entities.", $\mathbf{p}_{\text{reason}}$ = "Reason step by step before representing the input.", and $\mathbf{p}_{\text{think}}$ = "Think step by step." The non-reasoning variants, $\mathbf{p}_{\text{desc}}$ and $\mathbf{p}_{\text{ent}}$, perform similarly to the standard prompt, suggesting that prompt length or additional wording is not the main cause of improvement. By contrast, reasoning instructions $\mathbf{p}_{\text{reason}}$ and $\mathbf{p}_{\text{think}}$ yield comparable gains, indicating that the benefit generalizes across different reasoning-oriented formulations.

Document description quality. Tab. 8 compares different document descriptions under reasoning instruction training. Training without document text already achieves an OOD average of 62.6. A fixed placeholder ("This is an image") and short captions reduce the average to 51.7 and 54.3, respectively. In contrast, detailed descriptions achieve the best OOD average of 65.2, indicating that document enrichment is most effective when the added text contains detailed and retrieval-relevant visual semantics. To simulate realistic annotation noise, $x\%$ noisy denotes randomly shuffling that proportion of texts while keeping the remainder correctly matched. Compared with using no document text, description enrichment remains beneficial when the noise ratio is below 40%, and performance generally improves as the noise level decreases. This suggests that reasoning instruction training can tolerate moderate noise, while higher-quality annotations yield stronger gains.

Robustness on noisy benchmarks. Tab. 9 evaluates each component based on VLM2Vec-V2 with 300K data. Rea alone does not improve robustness, whereas adding description enrichment increases EDIS-Noisy from 76.0 to 89.5 and WebQA-Noisy from 72.4 to 77.8. Ret further improves them to 91.4 and 80.4, with smaller gains on the original benchmarks. Overall, description enrichment provides the main robustness gains, while Ret offers complementary refinement, enabling the framework to achieve the best performance.

5 Conclusion

We propose Self-Bootstrapped Multimodal Embeddings, which exploits the MLLM's own generation, reasoning, and retrieval capabilities without external teachers or auxiliary annotation models. The backbone generates detailed document descriptions, internalizes native reasoning through reasoning instruction training, and mines hard negatives for representation refinement. Our framework steadily improves Qwen3-VL-Instruct and VLM2Vec-V2 across text-to-multimodal and text-to-video retrieval. Without noise-specific training, it also outperforms Qwen3-VL-Embedding and WeMM-Embedding on EDIS-Noisy and WebQA-Noisy, demonstrating the effectiveness and robustness of self-bootstrapping across diverse retrieval scenarios. These results highlight the potential of harnessing MLLMs' native capabilities to learn effective and robust multimodal embeddings.

AI USE STATEMENT

Generative AI tools were used solely for grammar correction and language polishing and were not used for any other tasks. All AI-assisted revisions were reviewed and approved by two authors to ensure accuracy, appropriate wording, and consistency with the intended meaning. The authors take full responsibility for the final content of this work.

REFERENCES

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pp. 5803–5812, 2017.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 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.

David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 190–200, 2011.

Haonan Chen, Hong Liu, Yuping Luo, Liang Wang, Nan Yang, Furu Wei, and Zhicheng Dou. Moca: Modality-aware continual pre-training makes better bidirectional multimodal embeddings. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 810–823, 2026.

Jianpeng Cheng, Xian Wu, Jiangfan Zhang, Wentao Bao, Chaitanya Ahuja, Shlok Kumar Mishra, Hanchao Yu, Yang Gao, Fan Xia, Qi Guo, et al. Tte-flash: Accelerating reasoning-based multimodal representations via think-then-embed tokens. *arXiv preprint arXiv:2605.16638*, 2026.

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

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF 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 *International Conference on Learning Representations*, volume 2025, pp. 61424–61449, 2025.

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

Xiangzhao Hao, Shijie Wang, Tianyu Yang, Tianyue Wang, Haiyun Guo, and Jinqiao Wang. Trace: Task-adaptive reasoning and representation learning for universal multimodal retrieval. *arXiv preprint arXiv:2603.02929*, 2026.

Chenwei He, Xiangzhao Hao, Tianyu Yang, Yuxiang Ma, Yuheng Jia, Lingxiang Wu, Chaoyang Zhao, Haiyun Guo, and Jinqiao Wang. Plume: Latent reasoning based universal multimodal embedding. *arXiv preprint arXiv:2604.02073*, 2026.

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

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

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

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

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

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

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. 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 *International Conference on Learning Representations*, volume 2025, pp. 44215–44234, 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.

Yijie Lin, Guofeng Ding, Haochen Zhou, Haobin Li, Mouxing Yang, and Xi Peng. Ark: A dual-axis multimodal retrieval benchmark along reasoning and knowledge. *arXiv preprint arXiv:2602.09839*, 2026.

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, 2021.

Qianying Liu, Xiao Liang, Zhiqiang Zhang, Yibo Chen, Xu Tang, Zhongfei Qing, Fengfan Zhou, Yao Hu, and Paul Henderson. Rematch: Boosting representation through matching for multimodal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16823–16833, 2026.

Siqi Liu, Weixi Feng, Tsu-Jui Fu, Wenhu Chen, and William 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 Computer Vision and Pattern Recognition Conference*, pp. 4015–4025, 2025b.

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, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

OpenSearch-AI Team. Ops-mm-embedding-v1-2b: A multimodal embedding model. `https://huggingface.co/OpenSearch-AI/Ops-MM-embedding-v1-2B`, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

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

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

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4581–4591, 2019.

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 *European Conference on Computer Vision*, pp. 387–404. Springer, 2024.

WeMM Team. WeMM-Embedding. MMEB leaderboard entry, 2025.

Junxian Wu, Chenghan Fu, Zhanheng Nie, Daoze Zhang, Bowen Wan, Wanxian Guan, Chuan Yu, Jian Xu, and Bo Zheng. Moon3. 0: Reasoning-aware multimodal representation learning for e-commerce product understanding. *arXiv preprint arXiv:2604.00513*, 2026a.

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

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

Chenghao Xiao, Isaac Chung, Imene Kerboua, Jamie Stirling, Xin Zhang, Márton Kardos, Roman Solomatin, Noura Al Moubayed, Kenneth Enevoldsen, and Niklas Muennighoff. Mieb: Massive image embedding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22187–22198, 2025.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

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

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. *arXiv preprint arXiv:2403.19651*, 2024a.

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

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*, 2024b.

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 Computer Vision and Pattern Recognition Conference*, pp. 9274–9285, 2025b.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. URL https://arxiv.org/abs/2408.05517.

Junjie Zhou, Ze Liu, Lei Xiong, Jin-Ge Yao, Yueze Wang, Shitao Xiao, Fenfen Lin, Miguel Hu Chen, Zhicheng Dou, Siqi Bao, et al. MR^2-Bench: Going beyond matching to reasoning in multimodal retrieval. *arXiv preprint arXiv:2509.26378*, 2025a.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. Megapairs: Massive data synthesis for universal multimodal retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19076–19095, 2025b.

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

A Extended Ablation of Reasoning Instructions

Beyond the ablation study of different instructions presented in the main paper, we evaluate additional reasoning instructions while keeping the training recipe and the standard evaluation instruction unchanged. The results are reported in Table A1. Compared with $\mathbf{p}_{\text{R}}$, the generic $\mathbf{p}_{\text{process}}$ instruction improves the VisualNews (IID) score by 1.2 points and the OOD average by 0.5 points. More importantly, the two relation-aware instructions, $\mathbf{p}_{\text{cond}}$ and $\mathbf{p}_{\text{mm}}$, raise the OOD average by 2.3 and 3.0 points, respectively, while maintaining comparable IID performance. The concise multimodal-specific instruction achieves the best OOD average of 65.6, outperforming the longer conditional instruction by 0.7 points. These results suggest that instructions whose reasoning steps are explicit and aligned with the multimodal structure of the input can elicit more transferable representations than generic step-by-step cues; adding more prompt text alone is not necessarily beneficial.

Table A1: Extended ablation of reasoning training instructions. All models use $\mathbf{p}_{\text{std}}$ at evaluation.

Train	Eval	VisualNews(IID)	OOD Avg.
$\mathbf{p}_{\text{R}}$	$\mathbf{p}_{\text{std}}$	68.8	62.6
$\mathbf{p}_{\text{think}}$	$\mathbf{p}_{\text{std}}$	69.1	62.7
$\mathbf{p}_{\text{process}}$	$\mathbf{p}_{\text{std}}$	**70.0**	63.1
$\mathbf{p}_{\text{cond}}$	$\mathbf{p}_{\text{std}}$	68.8	64.9
$\mathbf{p}_{\text{mm}}$	$\mathbf{p}_{\text{std}}$	69.0	**65.6**

Here, $\mathbf{p}_{\text{R}}$ is "Think step by step before representing the input."; $\mathbf{p}_{\text{think}}$ is "Think step by step."; $\mathbf{p}_{\text{process}}$ is "Process the input step by step before representing it."; $\mathbf{p}_{\text{cond}}$ is "If the user input is text, represent the input text. If the user input consists of both images and text, first analyze the relationship between them and then represent them."; and $\mathbf{p}_{\text{mm}}$ is "If the user input consists of both images and text, first analyze the relationship between them and then represent them." The OOD average is computed over the five benchmarks unseen during training: MSCOCO, VisDial, Wiki-SS-NQ, EDIS, and WebQA.

Our goal is not to identify an optimal prompt or exhaustively search the instruction space. Rather, these results provide further evidence that an instruction can serve as a lightweight mechanism for eliciting the latent reasoning capability of an embedding model during representation learning. They also point to instruction optimization—and potentially automatic instruction evolution tailored to embedding objectives—as promising directions for future work.

B Public Benchmarks and Qualitative Evaluation

Table A2: Query and candidate statistics of the six text-to-multimodal retrieval datasets. Candidate occurrences count candidates across all query-specific retrieval pools, while unique candidates are deduplicated by image path within each dataset.

Dataset	Retrieval Task	Queries	Unique Queries	Candidates / Query	Candidate Occurrences	Unique Candidates	Unique Ground Truths
VisualNews_t2i	Text → Image	1,000	1,000	1,000	1,000,000	5,000	1,000
MSCOCO_t2i	Text → Image	1,000	1,000	1,000	1,000,000	1,002	201
VisDial	Text → Image	1,000	1,000	1,000	1,000,000	2,064	1,000
Wiki-SS-NQ	Text → Image	1,000	1,000	1,000	1,000,000	10,694	985
EDIS	Text → Text+Image	1,000	1,000	1,000	1,000,000	3,241	1,000
WebQA	Text → Text+Image	1,000	1,000	1,000	1,000,000	2,360	974
Total	–	6,000	6,000	–	6,000,000	24,361	5,160

Each query is evaluated against a local pool of 1,000 distinct candidates, consisting of one ground-truth candidate and 999 negatives. The totals for unique candidates and unique ground truths are sums of per-dataset counts rather than cross-dataset deduplicated counts.

Text-to-multimodal retrieval. We evaluate on six public benchmarks from MMEB (Jiang et al., 2024b): VisualNews-T2I (Liu et al., 2021), MSCOCO-T2I (Lin et al., 2014), VisDial (Das et al., 2017), Wiki-SS-NQ (Ma et al., 2024), EDIS (Liu et al., 2023), and WebQA (Chang et al., 2022). All six benchmarks use textual queries but differ in the modality of their candidate documents. The first four are text-to-image retrieval benchmarks, where every candidate contains only an image. In contrast, EDIS and WebQA evaluate text-to-multimodal retrieval, where each candidate comprises both text

and an image. EDIS focuses on entity-rich news retrieval, for which successful matching often requires jointly grounding named entities, events, and locations in the accompanying visual evidence. Its evaluation set contains 1,000 queries and 3,241 unique multimodal candidates. WebQA instead formulates natural-language questions as queries and requires retrieving the multimodal evidence needed to answer them; its evaluation set contains 1,000 queries and 2,360 unique multimodal candidates. For all datasets, we retain the official query-specific candidate pools and relevance annotations. Table A2 provides detailed query and candidate statistics. Figures A1 and A2 provide qualitative examples for the EDIS and WebQA benchmarks. As shown in Figs. A1 and A2, the same pattern emerges on both EDIS and WebQA: the baseline tends to confuse the positive document with hard-negative documents. Both the Reasoning and Generation components increase the separation between the positive and hard negatives. The complete method achieves the highest similarity for the positive document while maintaining the largest positive–hard-negative margin. Figure A9 provides examples of other T2I retrieval benchmarks.

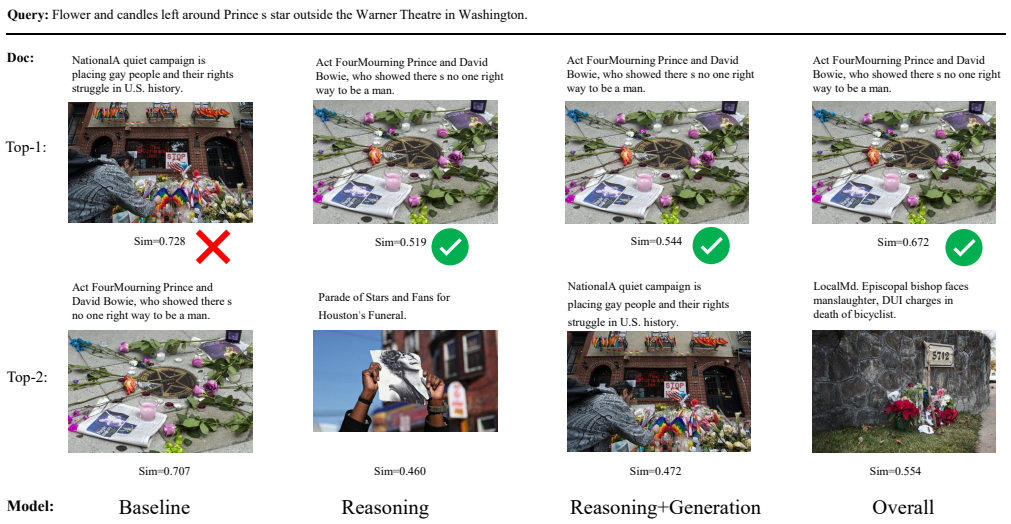


Figure A1: Qualitative retrieval example on EDIS. We compare the baseline with individual components of our proposed method. The example shows that our full model effectively enlarges the similarity margin between the positive and hard-negative documents.

Table A3: Query and candidate statistics for the text-to-video retrieval datasets. All datasets use global retrieval, where each query is evaluated against the entire candidate video corpus.

Dataset	Retrieval Task	Queries	Unique Queries	Candidates / Query	Candidate Occurrences	Unique Candidates	Unique Ground Truths
MSR-VTT	Text → Video	1,000	995	1,000	1,000,000	1,000	1,000
MSVD	Text → Video	670	660	670	448,900	670	670
DiDeMo	Text → Video	1,004	1,004	1,004	1,008,016	1,004	1,004
YouCook2	Text → Video	3,179	3,105	3,179	10,106,041	3,179	3,179
VATEX	Text → Video	4,478	4,478	4,478	20,052,484	4,478	4,478
Total	–	10,331	10,242	–	32,615,441	10,331	10,331

Unique queries are counted by exact matching of the query text used by the evaluation data loader. Candidate occurrences equal the number of queries multiplied by the size of the global candidate corpus. The totals for unique candidates and unique ground truths are sums of the per-dataset counts, without cross-dataset deduplication.

Text-to-video retrieval. We additionally evaluate zero-shot transfer on MSR-VTT (Xu et al., 2016), MSVD (Chen & Dolan, 2011), DiDeMo (Anne Hendricks et al., 2017), YouCook2 (Zhou et al., 2018), and VATEX (Wang et al., 2019). These datasets cover open-domain activities, temporally localized descriptions, instructional procedures, and multilingual captions. We use no video data when fine-tuning the models in the main comparison. Each video is represented by eight temporally sampled frames. Following the global-retrieval protocol, every text query is ranked against the complete candidate-video corpus of the corresponding dataset. Table A3 reports the resulting query

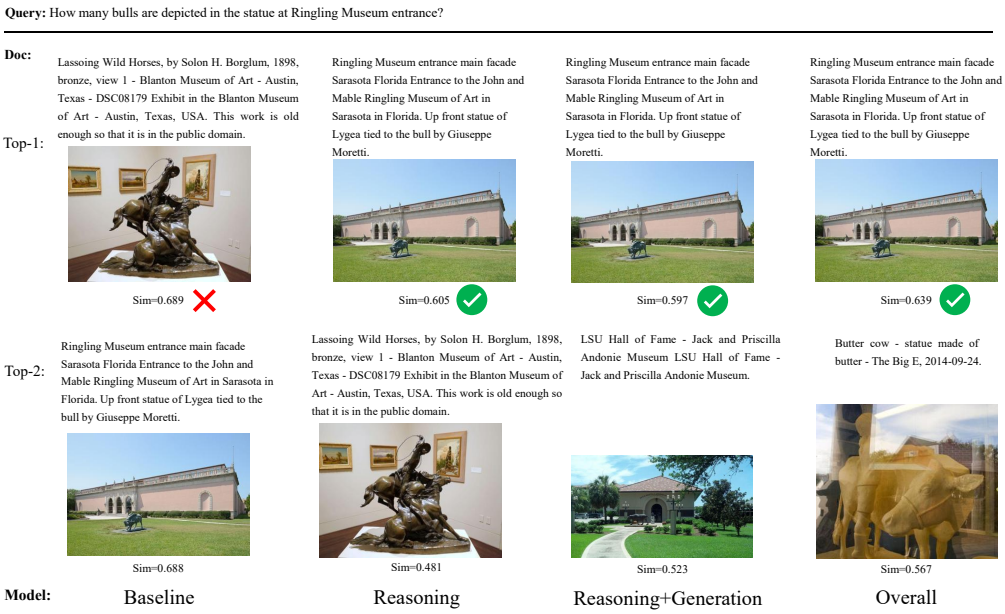


Figure A2: Qualitative retrieval example on WebQA. We compare the baseline with individual components of our proposed method. The example shows that our full model effectively enlarges the similarity margin between the positive and hard-negative documents.

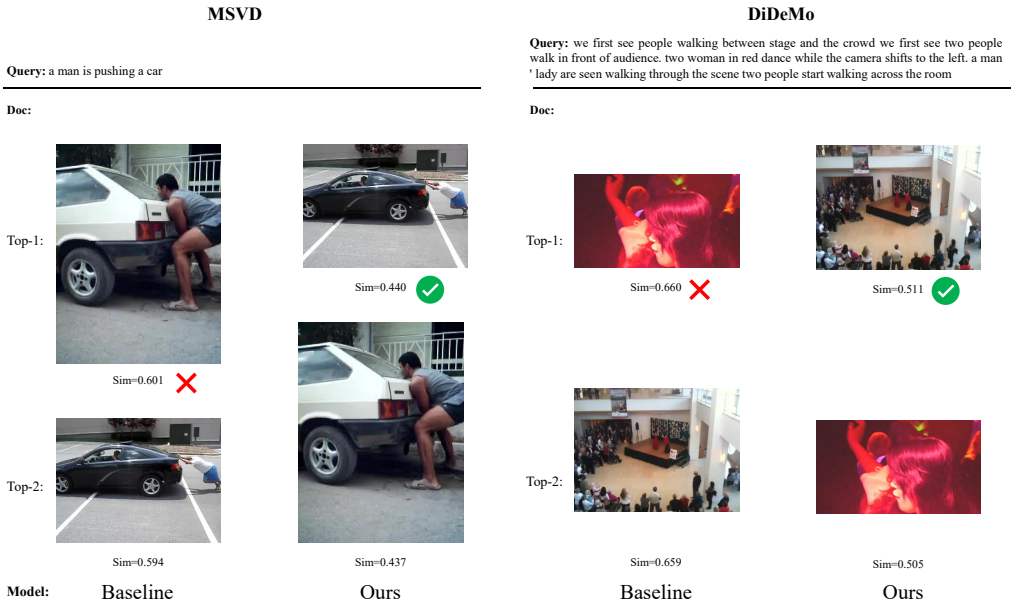


Figure A3: Qualitative text-to-video retrieval examples comparing the baseline with our proposed model. The examples show that our model accurately retrieves the correct videos for both simple and complex queries.

and candidate statistics, and Fig. A3 presents qualitative retrieval examples. As illustrated in the figure, our method successfully retrieves the correct videos for both concise and complex queries, indicating that it can handle varying levels of query specificity and semantic complexity.

Qualitative analysis on Noisy Benchmark. Figures A4 and A5 compare Qwen3-VL-Embedding-2B with our method on the Conflict Doc and Partial Doc splits of EDIS-Noisy, respectively. In both cases, the document text is either semantically conflicting or incomplete, making text-dominant matching unreliable. Nevertheless, our method assigns a high similarity score to the correct image and retrieves the positive document at rank 1. These examples demonstrate that our model can rely effectively on visual evidence when document text is unreliable, yielding robust multimodal retrieval under document-side noise.

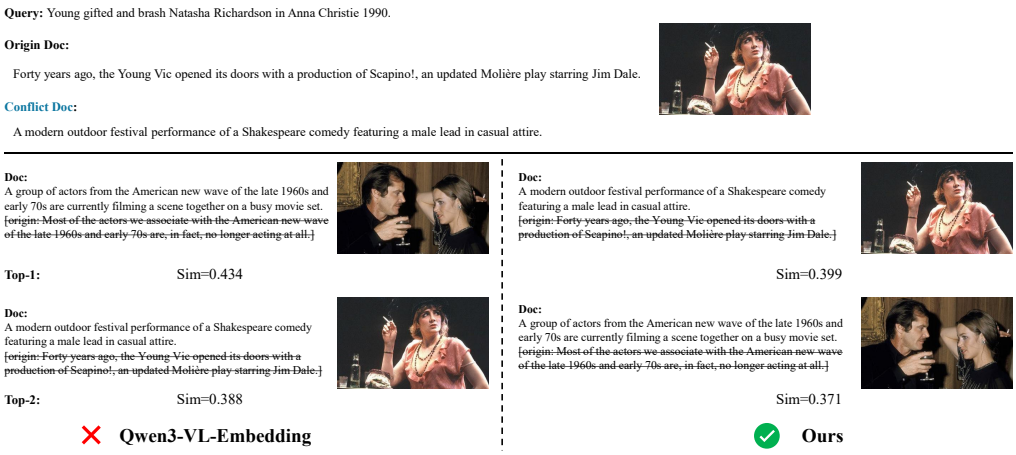


Figure A4: Qualitative comparison between Qwen3-VL-Embedding-2B and our model on the Conflict Doc split of EDIS-Noisy. Despite the conflicting document text, our method assigns a high similarity score to the correct image and ranks the positive document first.

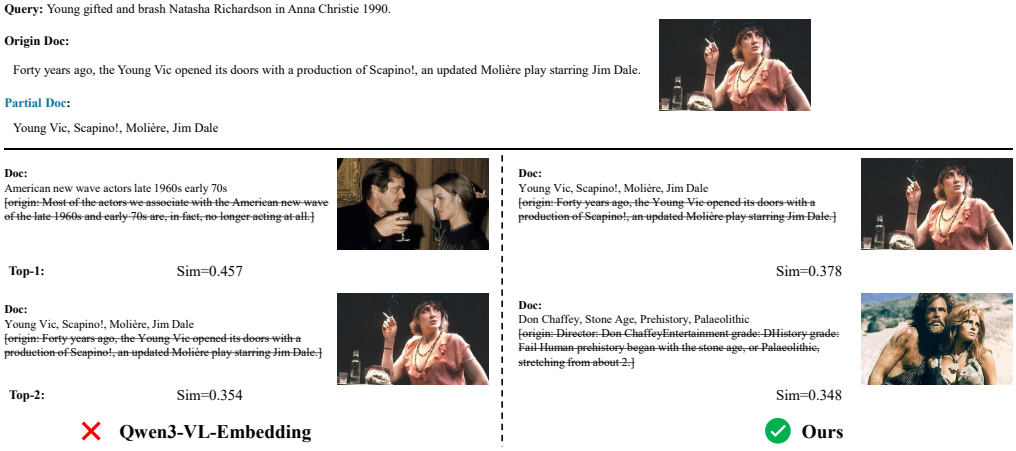


Figure A5: Qualitative comparison between Qwen3-VL-Embedding-2B and our model on the Partial Doc split of EDIS-Noisy. Despite the incomplete document text, our method effectively uses the visual evidence to assign a high similarity score to the positive document and retrieve it at rank 1.

C Construction of EDIS-Noisy and WebQA-Noisy

We construct both noisy benchmarks from the original EDIS and WebQA test sets. The images, candidate pools, relevance annotations, and train/test boundaries remain fixed; only the designated query or document text is perturbed. Performance differences from the clean benchmark can therefore

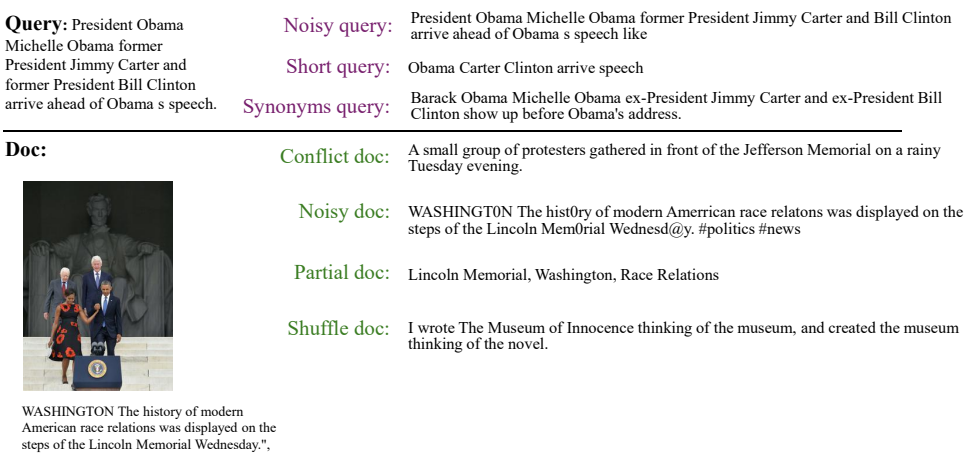


Figure A6: Example visualization from EDIS-Noisy.

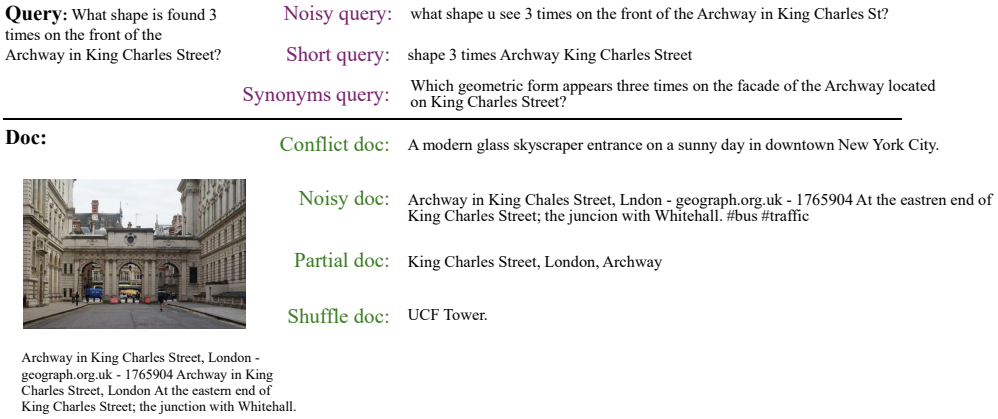


Figure A7: Example visualization from WebQA-Noisy.

be attributed to input corruption rather than a changed retrieval pool. We generate query-side and document-side variants independently and attach the generated text back to every occurrence of the corresponding unique query or document.

Document-side generation. We deploy Qwen3.5-27B (Qwen Team, 2026) as a multimodal inference service and use the document-side generation prompt provided in Listing 1. Before generation, we deduplicate the original candidate collection by the pair (`tgt_text`, `tgt_img_path`). Each unique pair is sent to the service exactly once: the original target text is inserted as `original_gt_text`, `noise_ratio` is set to 50, and the target image is provided in the same multimodal request. One response jointly produces three variants in strict JSON format: `partial_tag`, `noisy_text`, and `conflicting_text`. The generated variants are then mapped back to all candidate records sharing that pair. Generating all three variants in one request keeps the source image and text fixed across noise types while avoiding redundant calls for duplicated candidates.

The document prompt defines *partial tags* as two to six keywords or short phrases that preserve some main entities or attributes but omit important details. *Noisy text* corrupts approximately 50% of the original text using OCR-like errors, misspellings, garbled fragments, irrelevant entities, or noisy metadata, while preserving at least one semantic clue and avoiding intentional contradiction.

Conflicting text remains a plausible caption for a similar scene but contradicts one to three visible attributes, such as color, count, action, location, material, weather, or object identity. The response schema additionally records an English rationale, noise level, image–text semantic relation, and quality flags for each variant. The prompt requires strict JSON, preservation of the original language when possible, and no explanatory text outside the JSON object.

Document-side shuffling. `shuffle_text` is not generated by the MLLM. We collect the target texts of the full document pool and apply one global random permutation across all documents (using seed 42), then assign the permuted texts back to the original images. Thus, the image, document identifier, candidate pool, and relevance annotation stay fixed, while the accompanying text comes from the globally shuffled document collection. This construction directly breaks image–text alignment rather than introducing local lexical corruption.

Query-side generation. For query-side noise, we use the deployed multimodal LLM service with the query-side generation prompt provided in Listing 2. We deduplicate the test data by `qry_text`; each unique clean query is submitted once, and its three generated variants are reused for every record containing that query. The user-prompt template inserts the dataset name, query identifier, original query, positive-document text, and an optional positive visual description when available; this context guides intent preservation, while only the query text is replaced in the resulting benchmark. The prompt requests a single strict-JSON response containing `short_query`, `noisy_query`, and `synonyms_query`. *Short query* retains only one to five important keywords or phrases, removes modifiers or constraints, and preserves the broad search intent. *Noisy query* preserves the target but adds one to three realistic input errors, including typos, colloquial wording, redundant words, minor grammar errors, or word-order changes. *Synonyms query* paraphrases the query with synonyms, aliases, hypernyms, hyponyms, or equivalent expressions without changing the target entity, critical attributes, or answer intent. The returned metadata includes an English rationale, noise level, semantic-preservation level, and type-specific quality flags. As on the document side, the prompt requires the generated query to remain in the original language when possible and forbids text outside the JSON response.

These six generated variants isolate information loss, surface-form corruption, and lexical variation on the query side, and incomplete evidence, surface noise, and semantic contradiction on the document side. Global text shuffling separately evaluates robustness to severe cross-modal misalignment.

Human quality control. We manually inspect every model-generated query and document variant to ensure that it satisfies the intended perturbation definition, remains linguistically coherent, and preserves or alters the required semantics. When an output is invalid or unreasonable, we first request regeneration from the model; if regeneration still fails to meet the criteria, a human annotator corrects the output manually. Only variants that pass this inspection are retained in the final benchmark. The resulting dataset statistics are reported in Table A4.

Table A4: Numbers of queries and candidate documents in the clean and noisy sets.

	Queries				Documents				
Dataset	Orig.	Short	Noisy	Syn.	Orig.	Partial	Noisy	Conflict	Shuffle
EDIS	1,000	1,000	1,000	1,000	3,241	3,227	3,227	3,227	3,241
WebQA	1,000	993	993	993	2,360	2,332	2,332	2,332	2,360

C.1 Full Generation Prompts

For reproducibility, Listings 1 and 2 reproduce the complete prompts used by the two generation pipelines, including the system instructions, user templates, output schemas, and quality-control fields.

Illustrative case. Figure A6 and Figure A7 reserve the layout for representative EDIS-Noisy and WebQA-Noisy examples, while Table A5 gives a schematic summary of the perturbation operators.

Query variants preserve the retrieval target, whereas conflict and shuffle deliberately make document modalities inconsistent. The image and relevance label never change.

Table A5: Schematic noisy-benchmark case. Bracketed spans denote semantic slots retained or modified by each operator.

Level	Variant	Transformation
Query	Original	`[entity] [performs event] at [location]`
	Short	`[entity]; [location]; [event keyword]`
	Noisy	Retain all semantic slots but insert typos, disfluency, or reordered words.
	Synonym	Replace `[performs event]` with an equivalent paraphrase.
Document	Original	Image and text describe the same `[entity]`, `[attribute]`, and `[event]`.
	Partial	Retain only `[entity]; [attribute]; [event keyword]`.
	Noisy	Preserve event semantics while corrupting roughly half of the surface text.
	Conflict	Change a visible slot, e.g., `[attribute A]` to `[attribute B]`.
	Shuffled	Keep the image but replace its text with text from another candidate.

D SELF-ANNOTATION DETAILS

The same MLLM backbone (e.g., Qwen3-VL-Instruct-2B) used to initialize retrieval training generates the document annotations; we require neither an external captioner nor human-written auxiliary labels. For each training image, we run one offline generation pass using `Describe this image in detail.` The generated caption C_d becomes the textual component of the enriched document (I_d, C_d). Generation depends only on the image and never on its paired query, preventing query leakage and allowing a caption to be reused when an image has multiple associated queries. Figure A8 shows a representative annotation result.

Query: Turkish President Tayyip Erdogan looks on after arriving at Esenboga Airport in Ankara Turkey June 8 2015 Turkey faced the prospect of weeks of political turmoil after the ruling AK Party lost its parliamentary majority REUTERSUmit Bektas.

Doc:



Self-generated annotation:

This is a close-up photograph of a man in a formal black suit, white shirt, and a purple patterned tie. He has a mustache and is looking down, appearing to be reading or examining something. A small, dark pin is visible on his left lapel. The man is positioned in the foreground, with another man in a dark suit and a grey tie standing slightly behind and to his right. The background is blurred, suggesting a shallow depth of field, and the lighting appears to be natural daylight. The image is framed by the top of a vehicle's roof, which is visible at the bottom of the frame.

Figure A8: Example of document-side self-annotation. The MLLM generates a detailed textual description from the document image without access to the paired retrieval query.

The 100K setting contains 89,859 valid VisualNews-T2I pairs. The nominal 300K setting contains approximately 289K pairs: 90K VisualNews-T2I, 100K MSCOCO-T2I, 17K WebQA, and 82K VisDial examples.

We retain model-generated captions without query-conditioned rewriting. To assess sensitivity to annotation quality, we construct controlled variants by randomly permuting captions for an $x\%$ subset while preserving the remaining image–caption alignments. The resulting annotation-noise ablation shows that document enrichment remains useful under moderate noise, whereas clean detailed captions provide the strongest performance.

E Training Configuration

We train two backbone baselines using their respective codebases, both of which are fully open-source frameworks. The Qwen3-VL-Instruct-2B baseline is implemented with `ms-swift` (Zhao et al., 2024), whereas the VLM2Vec-V2 baseline is trained with the official VLM2Vec-V2 framework (Meng et al., 2025). Both follow the same two-stage recipe: Stage 1 performs contrastive training on the caption-enriched data, and Stage 2 continues training with self-mined hard negatives. We summarize only the key hyperparameters shown in Table A6 and Table A7.

Table A6: Key hyperparameters for the Qwen3-VL-Instruct-2B baseline implemented with `ms-swift`. Parameters not listed are unchanged between the two stages.

Hyperparameter	Stage 1	Stage 2
Training framework	`ms-swift`	`ms-swift`
Backbone	Qwen3-VL-Instruct-2B	Qwen3-VL-Instruct-2B
Fine-tuning method	LoRA (all linear layers)	LoRA (all linear layers)
LoRA rank / alpha	8 / 32	8 / 32
Learning rate	2×10^{-5}	1×10^{-5}
InfoNCE temperature	0.02	0.02
Number of GPUs	8	16
Per-GPU batch size	24	8
Gradient accumulation	1	1
Training epochs	20	20
Warmup ratio	0.05	0.05
Maximum sequence length	1,024	1,024
Maximum visual tokens	768	768
Numerical precision	bfloat16	bfloat16
Attention / distributed training	SDPA / ZeRO-2	SDPA / ZeRO-2
Hard negatives per query	–	8

For the `ms-swift` baseline, Stage 1 uses standard in-batch InfoNCE. Stage 2 initializes from the Stage-1 model and introduces eight mined hard negatives per query. Relative to Stage 1, we reduce the learning rate and per-GPU batch size and scale training from 8 to 16 GPUs; the remaining key settings are unchanged.

For VLM2Vec-V2, Stage 1 trains the LoRA adapter with gradient caching and in-batch negatives. Stage 2 merges and initializes from the Stage-1 adapter, switches to the hard-negative dataset, and trains with eight hard negatives per query using a smaller learning rate and passage-cache chunk size.

F Evaluation Configuration

Evaluation implementation. Unless otherwise specified, we evaluate all models with the official Qwen3-VL-Embedding evaluation repository (`https://github.com/QwenLM/Qwen3-VL-Embedding`). The dataset-specific preprocessing, query and target instructions, test splits, candidate-pool construction, relevance annotations, frame-sampling rules, and retrieval metrics follow the VLM2Vec-V2 benchmark configuration (Meng et al., 2025). We apply the same evaluation pipeline and benchmark configuration to every compared checkpoint, changing only the model weights and model-specific input processor. This controlled setup avoids differences caused by independently reimplemented data loaders or metrics.

Embedding and ranking. At inference, query and document inputs use `Represent the user's input.`; no CoT text is generated. We extract one embedding from the final valid

Table A7: Key hyperparameters for the VLM2Vec-V2 baseline trained with the VLM2Vec-V2 framework.

Hyperparameter	Stage 1	Stage 2
Backbone	Qwen2-VL-2B-Instruct	Qwen2-VL-2B-Instruct
Fine-tuning method	LoRA	LoRA initialized from Stage 1
LoRA rank	16	16
Learning rate	5×10^{-5}	5×10^{-6}
Learning-rate scheduler	Linear	Linear
InfoNCE temperature	0.02	0.02
Number of GPUs	8	8
Per-GPU batch size	256	128
Maximum training steps	5,000	2,000
Warmup steps	100	5
Gradient cache	Enabled	Enabled
Query / passage cache chunk	8 / 8	8 / 4
Interleave batch size	64	64
Pooling / normalization	EOS / L2	EOS / L2
Numerical precision	bfloat16	bfloat16
Hard negatives per query	–	8

hidden state and L2-normalize it. Matrix multiplication between query and candidate embeddings therefore gives cosine similarity, without cross-encoder reranking.

Text-to-multimodal retrieval. Following VLM2Vec-V2, we use the official test split, candidate pool, task instructions, and relevance labels for each benchmark. EDIS, WebQA, and their noisy variants share exactly the same candidates and relevance labels; only the designated query or document text is replaced.

Text-to-video retrieval. We likewise retain the VLM2Vec-V2 configuration for MSR-VTT, MSVD, DiDeMo, YouCook2, and VATEX. Each video is represented by 8 uniformly sampled frames selected from at most 64 pre-extracted frames, and retrieval is performed globally over the complete candidate pool of each dataset.

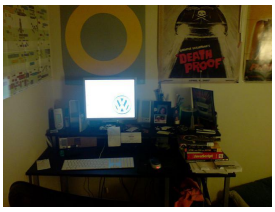
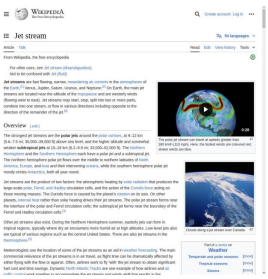
Benchmark	Query	Doc
VisualNews_t2i	A Maryland State Police cruiser sits at a blocked southbound entrance on the BaltimoreWashington Parkway that accesses Fort Meade.	
MSCOCO_t2i	A man with a red helmet on a small moped on a dirt road.	
VisDial	Q:is the photo in color\nA:yes\nQ:is it a professional photo\nA:no\nQ:is it well lit\nA:no\nQ:is it daytime\nA:i don't see windows\nQ:does this look like an adults bedroom\nA:maybe\nQ:is the room clean\nA:no\nQ:can you tell what kind of computer it is\nA:no not really\nQ:is it a flat screen\nA:yes\nQ:what's the desk made out of\nA:cheap plastic or wood\nQ:is there a computer chair\nA:yes	
Wiki-SS-NQ	who needs to know about the jet stream	

Figure A9: Example visualization from VisualNews-T2I, MSCOCO-T2I, VisDial, and Wiki-SS-NQ benchmarks.

Listing 1: Complete document-side noise-generation prompt used with Qwen3.5-27B.

```
## System Prompt

You are a data construction assistant for a noisy multimodal retrieval benchmark named Noisy-
    MMR.

We are constructing noisy document-side variants for text-to-multimodal retrieval datasets
    such as EDIS and WebQA. Each document contains an image and an original ground-truth text/
    caption. Your job is to generate controlled noisy text variants for the document.

Generate exactly the following three document-side variants:
1. partial_tag
2. noisy_text
3. conflicting_text

Definitions:

1. partial_tag:
- A short, incomplete but partially useful tag-like text.
- Extract only 2 to 6 keywords or short phrases from the original document text.
- Preserve some main entities or attributes, but intentionally drop important details.
- Do not introduce false information.
- Simulate weak business metadata, sparse tags, hashtags, or title fragments.

2. noisy_text:
- A corrupted version of the original document text.
- It should still be partially recognizable but contain realistic OCR errors, typos, garbled
    fragments, random irrelevant entities, or noisy metadata.
- Corrupt about 50% of the original text unless otherwise specified.
- Preserve at least one main semantic clue.
- Do not intentionally contradict the image; this is noisy text, not conflicting text.

3. conflicting_text:
- A plausible caption/title that intentionally conflicts with visible image attributes.
- It should still look like a natural caption for a similar type of scene, but contradict 1 to
     3 key visual attributes, such as color, number, action, location, material, weather, or
    object identity.
- Do not make it completely unrelated.
- Do not mention that it is wrong or conflicting.

Safety and quality rules:
- Do not output pornography, hate speech, violence, political content, personal private
    information, or other unsafe content.
- Do not answer the retrieval query.
- Do not explain outside the JSON.
- Output strict JSON only. No markdown, no comments.
- Keep generated text in the same language as the original text when possible.
- The rationale field must be in English.

## User Prompt Template

Construct document-side noisy variants for the following multimodal document.

Input:
{
  "original_gt_text": "{ORIGINAL_GT_TEXT}",
  "noise_ratio": 50
}

The document image is provided together with this message. Please generate exactly three
    variants:
1. partial_tag
2. noisy_text
3. conflicting_text

Return strict JSON in the following format:
{
  "variants": [
    {
      "noise_type": "partial_tag",
      "generated_text": "...",
      "rationale": "Briefly explain why this is a partial tag.",
      "noise_level": "medium",
      "semantic_relation_to_image": "partial",
      "quality_flags": {
        "preserve_main_entity": true,
        "contains_random_or_corrupted_tokens": false,
        "intentionally_conflicts_with_image": false,
        "safe_to_use": true
      }
```

```
    },
    {
      "noise_type": "noisy_text",
      "generated_text": "...",
      "rationale": "Briefly explain what OCR, typo, garbled-token, or random-entity noise was
     added.",
      "noise_level": "medium",
      "semantic_relation_to_image": "partial",
      "quality_flags": {
        "preserve_main_entity": true,
        "contains_random_or_corrupted_tokens": true,
        "intentionally_conflicts_with_image": false,
        "safe_to_use": true
      }
    },
    {
      "noise_type": "conflicting_text",
      "generated_text": "...",
      "rationale": "Briefly explain which visible attributes conflict with the image.",
      "noise_level": "high",
      "semantic_relation_to_image": "conflicting",
      "quality_flags": {
        "preserve_main_entity": false,
        "contains_random_or_corrupted_tokens": false,
        "intentionally_conflicts_with_image": true,
        "safe_to_use": true
      }
    }
  ]
}
```

Listing 2: Complete query-side noise-generation prompt used with Qwen3.5-27B.

```
## System Prompt

You are a data construction assistant for a noisy multimodal retrieval benchmark named Noisy-
     MMR.

We are constructing noisy query-side variants for text-to-multimodal retrieval datasets such
     as EDIS and WebQA. Each input query is a clean original retrieval query. Your job is to
     generate controlled noisy query variants while preserving the intended retrieval target.

Generate exactly the following three query-side variants:
1. short_query
2. noisy_query
3. synonyms_query

Definitions:

1. short_query:
- A much shorter and incomplete version of the original query.
- Keep only 1 to 5 important keywords or short phrases.
- Drop some modifiers, context, or constraints so the query becomes underspecified.
- Preserve the broad search intent.
- Do not introduce new entities.

2. noisy_query:
- A realistic noisy user query.
- Preserve the original search intent.
- Add 1 to 3 realistic user input noises, such as typos, colloquial words, redundant words,
     minor grammar errors, or word order changes.
- Do not change the target entity, key attribute, or intended answer.
- Do not make the query impossible to understand.

3. synonyms_query:
- A paraphrased version of the original query using synonyms, aliases, hypernyms/hyponyms, or
     natural alternative expressions.
- Preserve the exact retrieval target and answer intent.
- Do not remove critical attributes unless they are replaced by equivalent expressions.
- Do not introduce a different entity or contradictory attribute.

Safety and quality rules:
- Do not answer the query.
- Do not output pornography, hate speech, violence, political content, personal private
     information, or unsafe content.
- Output strict JSON only. No markdown, no comments.
- Keep generated query in the same language as the original query when possible.
- The rationale field must be in English.
```

```
## User Prompt Template

Construct query-side noisy variants for the following retrieval query.

Input:
{
  "dataset": "{DATASET_NAME}",
  "query_id": "{QUERY_ID}",
  "original_query": "{ORIGINAL_QUERY}",
  "positive_doc_text": "{POSITIVE_DOC_TEXT}",
  "optional_positive_visual_description": "{OPTIONAL_POSITIVE_VISUAL_DESCRIPTION}"
}

Please generate exactly three variants:
1. short_query
2. noisy_query
3. synonyms_query

Return strict JSON in the following format:
{
  "query_id": "{QUERY_ID}",
  "variants": [
    {
      "noise_type": "short_query",
      "generated_query": "...",
      "rationale": "Briefly explain what information was removed and why the broad retrieval
     intent is still preserved.",
      "noise_level": "medium",
      "semantic_preservation": "medium",
      "quality_flags": {
        "preserve_search_intent": true,
        "is_shorter_than_original": true,
        "contains_typo_or_colloquial_noise": false,
        "uses_synonym_or_paraphrase": false,
        "safe_to_use": true
      }
    },
    {
      "noise_type": "noisy_query",
      "generated_query": "...",
      "rationale": "Briefly explain what typo, colloquial, redundant, or grammatical noise was
      added.",
      "noise_level": "medium",
      "semantic_preservation": "high",
      "quality_flags": {
        "preserve_search_intent": true,
        "is_shorter_than_original": false,
        "contains_typo_or_colloquial_noise": true,
        "uses_synonym_or_paraphrase": false,
        "safe_to_use": true
      }
    },
    {
      "noise_type": "synonyms_query",
      "generated_query": "...",
      "rationale": "Briefly explain which words were replaced by synonyms or paraphrases and
     confirm that the retrieval intent is unchanged.",
      "noise_level": "low",
      "semantic_preservation": "high",
      "quality_flags": {
        "preserve_search_intent": true,
        "is_shorter_than_original": false,
        "contains_typo_or_colloquial_noise": false,
        "uses_synonym_or_paraphrase": true,
        "safe_to_use": true
      }
    }
  ]
}
```