WHAT REALLY WORKS FOR UNIVERSAL MULTIMODAL EMBEDDINGS? A CONTROLLED STUDY OF CONTRASTIVE TRAINING STRATEGIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning has become the standard paradigm for multimodal representation learning, aligning semantically matched pairs while separating mismatched ones in a shared embedding space. Although various contrastive training strategies have been proposed across negative selection, batch construction, and gradient modulation, their actual effectiveness under controlled, reproducible conditions remains poorly understood. This motivates a fundamental question: *Do sophisticated contrastive strategies consistently bring meaningful downstream gains in multimodal embedding learning?* To address this, we systematically evaluate representative strategies under a unified benchmarking framework on MMEB. Surprisingly, direct evaluations reveal that most existing strategies offer marginal gains over a standard aligned baseline, with several even degrading overall performance. To uncover the underlying failure modes, we introduce an **MLLM-as-a-Judge** diagnostic framework that dissects top-1 retrieval behaviors into labeled positives, false negatives (unlabeled valid matches), and hard/easy negative errors. Our systematic analysis yields three key insights: (1) *Gradient distortion:* Upweighting gradients from high-similarity negatives consistently hurts downstream generalization; (2) *Distribution illusion:* Wider separation between positive and negative similarity distributions does not necessarily translate to better retrieval performance; and (3) *Synergistic training:* While false-negative mitigation or hard-batch construction yields limited gains in isolation, integrating hard-negative mining, false-negative mitigation, and hard-batch construction achieves optimal performance, delivering an average improvement of **6.0%** on MMEB with Qwen2-VL-2B. Overall, our findings demystify popular contrastive strategies and provide actionable recipes for training universal multimodal embedding models.

1 INTRODUCTION

Contrastive learning (CL) has emerged as a widely adopted approach for training multimodal embedding models (Radford et al., 2021; Jia et al., 2021). By pulling semantically matched inputs closer and pushing mismatched ones apart, CL maps heterogeneous multimodal inputs into a shared representation space, supporting various downstream tasks such as classification, retrieval, and visual question answering (Jiang et al., 2024b; Meng et al., 2025; Huang et al., 2026). Under this paradigm, early multimodal embedding models primarily adopted CLIP-style dual encoders (Radford et al., 2021; Cherti et al., 2023; Sun et al., 2023), which independently encode images and text and align their representations through large-scale contrastive pretraining. Recently, research has shifted toward a new paradigm of universal multimodal embedding models (Jiang et al., 2024a; Zhang et al., 2025; Li et al., 2026), which leverage the instruction-following and multimodal understanding capabilities of multimodal large language models (MLLMs) to accommodate more diverse inputs and tasks. Across both architectural paradigms, different contrastive training strategies have been explored, including gradient modulation (Lan et al., 2025; Xue et al., 2025), false-negative mitigation (Zhong et al., 2017; Balmaseda et al., 2025), batch construction (Kim et al., 2026; Dutta et al.), and hard-negative mining (Thirukovalluru et al., 2026; Gu et al., 2026).

However, despite their reported gains, the actual effectiveness of these strategies remains unclear. Existing methods are evaluated across different model architectures and experimental settings, mak-

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.

ing it difficult to isolate the effectiveness of training strategies from hyperparameter tuning. This raises two central questions: ❶ how existing training strategies compare in their effectiveness; and ❷ what explains their success or failure.

To answer the two questions, we systematically evaluate representative contrastive training strategies under a unified experimental setting. We conduct all evaluations on the Massive Multimodal Embedding Benchmark (MMEB), a comprehensive benchmark spanning 36 datasets across four meta-task categories (Jiang et al., 2024b). Our studies show that most existing contrastive training strategies yield only marginal gains over a well-aligned baseline and many even degrade overall performance. To better understand these strategies, we develop an MLLM-as-a-Judge framework that categorizes top-1 predictions into labeled positives, false negatives (unlabeled yet semantically valid positives), and hard- and easy-negative errors. Our analysis reveals three key findings. First, upweighting gradients from high-similarity negatives degrades performance because similarity reflects hardness rather than label reliability. As a result, both hard negatives and false negatives receive larger gradient weights and are pushed farther from the query, while easy negatives contribute relatively less to training. Second, wider separation between positive and negative similarity distributions does not reliably indicate better downstream performance, because aggregate distributional separation does not imply that each positive is better separated from its corresponding negatives at the instance level. Third, either false-negative mitigation or hard-batch construction alone yields limited gains. Filtering false negatives improves label reliability but can also mistakenly filter out informative hard negatives, whereas constructing hard batches increases the density of valuable in-batch negatives without guaranteeing the semantic validity of individual negatives. Combining instance-level hard-negative mining and false-negative mitigation with batch-level hard-batch construction provides both reliable and informative supervision, leading to the strongest performance.

Overall, the contributions of this work are threefold:

- We systematically compare representative contrastive training strategies under a unified setting on MMEB using Qwen2-VL-2B/7B (Wang et al., 2024) and Qwen3.5-2B (Team, 2026), showing that most strategies yield no meaningful downstream gains and many even degrade overall performance, revealing a notable discrepancy between their reported gains and actual effectiveness.
- We develop an MLLM-as-a-Judge analysis framework that decomposes top-1 predictions into labeled positives, unlabeled yet semantically valid positives, and hard- and easy-negative errors, providing a fine-grained diagnosis of model behavior beyond aggregate benchmark scores.
- We uncover three key findings for contrastive training: upweighting gradients from high-similarity negatives degrades performance; separation between positive and negative similarity distributions does not reliably indicate downstream performance; and jointly integrating instance-level hard-negative mining and false-negative mitigation with hard-batch construction yields the strongest performance.

2 Related Work

2.1 Multimodal Embedding Learning

CLIP (Radford et al., 2021) established a scalable dual-encoder paradigm for learning transferable image-text embeddings through large-scale contrastive pre-training. Subsequent studies improved its scalability and training efficiency, including OpenCLIP (Cherti et al., 2023) with reproducible open-data training, SigLIP (Zhai et al., 2023) with a pairwise sigmoid objective, and EVA-CLIP (Sun et al., 2023) with improved initialization and optimization. Nevertheless, CLIP-style dual encoders face several limitations: 1) global image-text alignment provides limited fine-grained spatial grounding (Zhong et al., 2022); 2) independent modality encoding makes it difficult to generate task-specific representations conditioned on instructions (Jiang et al., 2024b); and 3) these models remain limited in understanding compositional relationships, long descriptions, and high-resolution visual details (Ma et al., 2023; Zhang et al., 2024; Shi et al., 2025).

To address these issues, recent studies adopt MLLMs as more flexible backbones for multimodal representation learning. E5-V (Jiang et al., 2024b) adapts an MLLM into a universal embedding

Table 1: ***Overview of representative contrastive training strategies.*** *Dependencies* indicate whether a method requires external scoring or additional preprocessing. Entries under *Optimization Dimensions* indicate what aspects each contrastive learning method targets explicitly. *Static* denotes that the corresponding optimization signal is fixed before training, whereas *Adaptive* denotes that it evolves with the student model.

	Dependencies		**Optimization Dimensions**			
Method	**External Scorer**	**Pre-processing**	**Gradient Modulation**	**False-Negative Mitigation**	**Batch Construction**	**Hard-Negative Mining**
LLaVE	–	–	Adaptive	–	–	–
EGA	–	–	Adaptive	–	–	–
GloFND	–	–	–	Adaptive	–	–
FN Risk	–	✓	–	Adaptive	–	–
FALCON	–	✓	–	–	Adaptive	–
HOBIT	–	✓	–	–	Adaptive	–
UniME-V2	✓	✓	Static	Static	–	Static
B3	✓	✓	–	Static	Static	Static

model using prompts and text-only contrastive training (Jiang et al., 2024a). VLM2Vec (Jiang et al., 2024b) introduces a general framework for converting MLLMs into instruction-aware embedding models and establishes MMEB, covering 36 datasets across classification, VQA, retrieval, and visual grounding. GME (Zhang et al., 2025) further expands training to large-scale single-modal, cross-modal, and fused-modal data. More recently, Seed1.6-Embedding (ByteDance Seed Team, 2025) scales unified representations to text, images, videos, and mixed-modal inputs, while Qwen3-VL-Embedding (Li et al., 2026) introduces an open family of embedding and reranking models. Together, these advances demonstrate the potential of MLLMs as flexible and powerful backbones for universal multimodal embedding learning.

2.2 Contrastive Training Strategies for Multimodal Embeddings

Building on both CLIP-style dual encoders and MLLM-based embedding models, recent studies have explored improving CL through gradient modulation (Sun et al., 2020; Lan et al., 2025), false-negative mitigation (Zhong et al., 2017; Byun et al., 2024; Balmaseda et al., 2025), batch construction (Thirukovalluru et al., 2026; Kim et al., 2026), and hard-negative mining (Robinson et al., 2020; Thirukovalluru et al., 2026). UniME-V2 (Gu et al., 2026) employs an MLLM judge to mine hard negatives, filter false negatives, and generate soft semantic labels. EGA (Xue et al., 2025) explicitly amplifies the gradients associated with hard negatives based on the positive-negative similarity margin. HOBIT (Dutta et al.) dynamically selects hard yet non-contradictory in-batch negatives. However, differences in model architectures and experimental settings leave the relative effectiveness of these strategies and the reasons for their success or failure unclear. We therefore benchmark representative strategies under a unified setting and diagnose their behavior beyond aggregate scores to identify genuine methodological gains and provide more reliable guidance for future designs.

3 Experiments

3.1 Contrastive Training Strategies

We evaluate eight representative strategies that optimize different aspects of contrastive training. As summarized in Table 1, we characterize each strategy by its dependencies and optimization dimensions. Dependencies indicate whether a method requires external scoring or additional preprocessing. An *external scorer* refers to a separately pretrained teacher, retriever, or judge that is not jointly updated with the student model, while *preprocessing* denotes the construction of embeddings, rankings, graphs, or candidate pools before training begins. The four optimization dimensions target different aspects of contrastive training. *Gradient modulation* modifies the contrastive objective or sample weights to adjust the gradient contributions of different negative pairs. *False-negative mitigation* identifies and mitigates semantically valid samples that are incorrectly treated as negatives. *Batch construction* organizes training examples to improve the informativeness of in-batch negatives, whereas *hard-negative mining* identifies or introduces challenging negative candidates at

Table 2: ***Training-setting alignment.*** We align the learning rate (LR), maximum number of visual tokens, LoRA rank, and pooling strategy for the VLM2Vec baseline on MMEB using Qwen2-VL-2B-Instruct. *Cls.*, *VQA*, *Ret.*, and *VG* denote image classification, visual question answering, image-level retrieval, and visual grounding, respectively. *Overall* denotes the mean Precision@1 across all 36 MMEB evaluation datasets. The configuration adopted for all subsequent experiments is highlighted in **bold**.

Training Configuration				MMEB				
LR	**Max. Visual Tokens**	**LoRA Rank**	**Pooling Strategies**	**Cls.**	**VQA**	**Ret.**	**VG**	**Overall**
2e-5	2,880	8	Last Token	58.7	49.3	65.0	72.9	59.8
1e-4	2,880	8	Last Token	60.0	56.5	67.3	76.0	63.2
1e-4	1,280	8	Last Token	59.7	56.5	67.3	76.1	63.2
1e-4	2,880	4	Last Token	60.7	55.4	65.2	76.5	62.5
1e-4	2,880	32	Last Token	61.9	57.7	69.2	76.5	64.8
1e-4	**2,880**	**16**	**Last Token**	**63.3**	**57.6**	**68.6**	**77.9**	**65.1**
1e-4	2,880	16	Mean	46.8	44.2	40.9	32.2	42.5
1e-4	2,880	16	`<embedding>`	53.3	46.6	62.4	72.9	56.6

the instance level. A strategy is marked as *static* for a given dimension when its supervision or optimization structure remains fixed throughout training, and as *adaptive* when it evolves with the current model, historical representations, or a jointly learned controller. Detailed descriptions of the individual strategies are provided in Appendix A.1.

3.2 Basic Setup

Training Setup. Following recent work on MLLM-based multimodal embedding models, we adopt Qwen2-VL-2B-Instruct (Wang et al., 2024) as the shared backbone and fine-tune it using LoRA (Hu et al., 2021). We use the widely adopted VLM2Vec (Jiang et al., 2024b), trained on MMEB using the standard InfoNCE objective with in-batch negatives, as our baseline. Before comparing different training strategies, we align four key training choices for the baseline: learning rate, maximum number of visual tokens, LoRA rank, and embedding pooling. As shown in Table 2, the best-performing configuration uses a learning rate of 1×10^{-4}, up to 2,880 visual tokens, a LoRA rank of 16, and last-token pooling, achieving an overall MMEB score of 65.1. We use this fixed configuration, together with the training data and optimization budget, for all subsequent comparisons. All experiments are conducted on $8\times$ NVIDIA H100 GPUs.

Evaluation Setup. We evaluate all models on the MMEB benchmark using Precision@1 as the primary metric. We report results across MMEB's four meta-task categories: image classification, visual question answering, image-level retrieval, and visual grounding, together with the overall mean across all 36 datasets. All methods use the same task-specific instructions and input templates during training and evaluation, thereby controlling for differences in prompt formatting. Additional backbone and evaluation details are provided in Appendix A.2 and Appendix A.3.

Notably, the aligned baseline already achieves an MMEB score of 65.1, comparable to LLaVE-2B's originally reported score of 65.2 (Lan et al., 2025) and higher than UniME-V2-2B's originally reported score of 63.6 (Gu et al., 2026). This observation highlights the importance of aligning basic training configurations before attributing performance gains to a particular strategy.

3.3 MMEB Results

Using the best-performing configuration identified in Section 3.2, we train and evaluate all eight strategies, together with an additional B3 + EGA variant, using Qwen2-VL-2B as the backbone. As shown in Table 3, the gradient-modulation methods LLaVE (Lan et al., 2025) and EGA (Xue et al., 2025) achieve overall scores of 63.9 and 64.8, respectively. For false-negative mitigation, GloFND (Balmaseda et al., 2025) reaches 64.3, while FN Risk, our adaptive false-negative filtering baseline based on k-reciprocal neighbors (Zhong et al., 2017), obtains 65.0. Methods targeting either gradient modulation or false-negative mitigation fail to surpass the aligned VLM2Vec baseline of

Table 3: ***Comparison of multimodal contrastive training strategies on MMEB using Qwen2-VL-2B.*** The evaluation metric is mean Precision@1 (%) across the corresponding datasets. IND and OOD denote the averages over in-distribution and out-of-distribution datasets, respectively. GM, FN, HN, and BC denote gradient modulation, false-negative mitigation, hard-negative mining, and batch construction, respectively. Training time denotes the training wall-clock time; one-time preprocessing costs are excluded. The best results are shown in **bold**.

Method	Training Time (h:mm)	Per Meta-Task Score				Average Score		
		Image Classification	Visual Question Answering	Image-level Retrieval	Visual Grounding	IND	OOD	Overall
No. of Datasets →		10	10	12	4	20	16	36
VLM2Vec (Baseline)	19:12	63.3	57.6	68.6	77.9	69.8	59.2	65.1
Baseline + GM								
LLaVE	19:24	57.8	58.4	68.9	78.0	69.3	57.2	63.9
EGA	19:26	63.3	57.4	68.1	77.3	69.9	58.5	64.8
Baseline + FN								
GloFND	15:05	62.6	55.7	68.4	77.5	68.4	59.1	64.3
FN Risk	17:50	63.5	57.6	69.1	74.8	69.8	58.9	65.0
Baseline + BC								
FALCON	15:03	62.4	57.9	68.8	76.2	69.5	59.0	64.8
HOBIT	20:06	62.2	58.1	69.6	78.3	69.7	59.9	65.3
Baseline + FN + HN + GM								
UniME-V2	39:31	62.9	57.6	69.2	76.8	68.4	60.9	65.1
Baseline + FN + HN + BC								
B3	29:44	**67.3**	**63.0**	**71.4**	**81.4**	**72.5**	**64.7**	**69.0**
Baseline + FN + HN + BC + GM								
B3 + EGA	30:07	66.7	62.7	71.2	80.0	**72.5**	63.7	68.6

65.1, indicating that optimizing either dimension in isolation does not improve performance. Batch-construction methods similarly provide no consistent improvement: FALCON (Kim et al., 2026) obtains 64.8, while HOBIT (Dutta et al.) marginally improves the baseline from 65.1 to 65.3.

When multiple optimization dimensions are combined, UniME-V2 (Gu et al., 2026) matches the baseline with an overall score of 65.1. In contrast, B3 (Thirukovalluru et al., 2026) substantially improves the overall score from 65.1 to 69.0, with consistent gains across all four meta-task categories. It also improves the IND and OOD averages by 2.7 and 5.5 points, respectively. Finally, adding EGA to B3 provides no further benefit and reduces the overall score from 69.0 to 68.6. Overall, optimizing individual dimensions yields no consistent gains. Among the evaluated strategies that combine multiple dimensions, only B3 achieves a substantial improvement by jointly optimizing false-negative mitigation, hard-negative mining, and batch construction. Experiments with Qwen2-VL-7B and Qwen3.5-2B show similar trends, as detailed in Appendix B.1. This divergence motivates a deeper analysis in the following section.

4 Understanding the Effects of Contrastive Training Strategies

4.1 MLLM-as-a-Judge Analysis

Standard Precision@1 treats a top-ranked candidate as correct only when it matches the labeled target, grouping all other outcomes as errors. However, multimodal embedding benchmarks may contain unlabeled but semantically valid matches (i.e., false negatives), while genuinely incorrect predictions may be either plausible but subtly incorrect hard negatives or clearly irrelevant easy negatives. Aggregate scores conflate these distinct outcomes, obscuring how different training strategies affect model behavior.

To obtain a finer-grained diagnosis, we introduce an MLLM-as-a-Judge framework, as illustrated in Figure 1. The complete prompt template and structured output schema are provided in Appendix C. We randomly sample 200 cases from each of the 36 MMEB evaluation datasets and evaluate every method on the same $36 \times 200 = 7{,}200$ cases. Given a query, its designated positive, and the model's

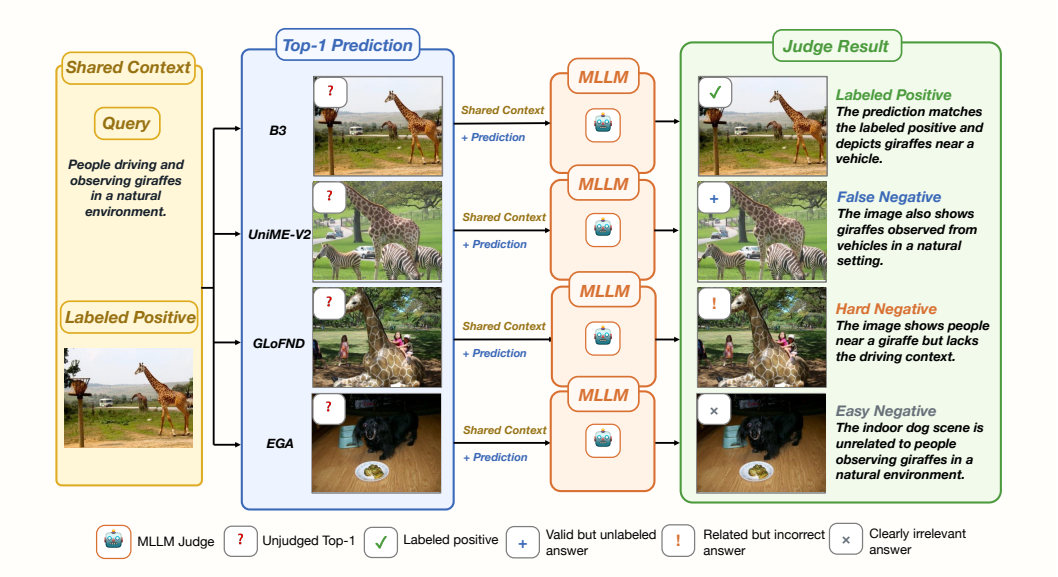


Figure 1: ***Overview of the MLLM-as-a-Judge analysis framework.*** Given the same query and labeled positive, we collect top-1 predictions produced by different contrastive training strategies. The MLLM judge categorizes the prediction as a labeled positive, an alternative valid positive (i.e., false negative), a hard-negative error, or an easy-negative error. This decomposition enables a more fine-grained analysis than standard exact-match Precision@1.

top-1 prediction, the judge assigns the prediction to one of four categories: (1) *Labeled Positive (LP)*, which matches the annotated target; (2) *False Negative (FN)*, which differs from the annotated target but remains semantically valid; (3) *Hard Negative (HN)*, which is semantically related and superficially plausible but fails one or more critical query constraints; or (4) *Easy Negative (EN)*, which is clearly irrelevant to the query. We define *Semantic Valid* as the sum of LP and FN, and *Semantic Error* as the sum of HN and EN. This decomposition distinguishes exact target matches from alternative valid predictions and fine-grained semantic errors from clearly irrelevant results, enabling a more precise diagnosis of the gains and failure modes of individual training strategies. We further verify the robustness of this analysis to the evaluation-set size and the choice of MLLM judge in Appendix B.2.

4.2 Upweighting High-Similarity Negatives Degrades Performance

Takeaway: *Since similarity reflects negative hardness rather than label reliability, upweighting high-similarity negative pairs assigns larger gradient weights to both genuine hard negatives and false negatives. Such similarity-based gradient reweighting amplifies useful signals from genuine hard negatives and misleading signals from false negatives while diminishing the relative gradient contribution of easy negatives, resulting in overall performance degradation.*

LLaVE (Lan et al., 2025) and EGA (Xue et al., 2025) use different modulation mechanisms to increase the gradient weights assigned to high-similarity negatives. However, the fine-grained results in Table 4 show that neither method improves over the baseline, consistent with their lower overall MMEB scores in Table 3. Relative to VLM2Vec (Jiang et al., 2024b), both methods show modest increases in FN that only partially offset declines in LP accuracy (LLaVE/EGA: FN $+0.31/+0.09$; LP $-1.46/-0.46$). Meanwhile, both HN and EN errors increase (LLaVE/EGA: HN $+0.39/+0.05$; EN $+0.82/+0.19$), resulting in lower Semantic Valid ($-1.15/-0.37$) and higher Semantic Error ($+1.21/+0.24$). Thus, assigning greater gradient weights to high-similarity negatives does not translate into better hard-negative discrimination.

This degradation persists even when negative quality is explicitly controlled. B3 (Thirukovalluru et al., 2026) combines false-negative treatment, hard-negative mining, and hard-batch construction, achieving the highest Semantic Valid rate of 78.81%. Compared with B3, adding EGA reduces EN

Table 4: ***MLLM-as-a-Judge analysis of top-1 predictions on MMEB.*** All methods are evaluated on the same 7,200 cases, with 200 cases sampled from each of the 36 datasets. *Labeled Positive* matches the annotated target; *False Negative* is a different but semantically valid answer; and *Semantic Valid* is their sum. *Hard Negative* denotes a related but incorrect prediction, while *Easy Negative* denotes a clearly irrelevant prediction; *Semantic Error* is their sum. GM, FN, HN, and BC denote gradient modulation, false-negative mitigation, hard-negative mining, and batch construction, respectively. Arrows indicate the preferred direction, and the best results are shown in **bold**.

Method	Positive Outcomes (%)			Negative Outcomes (%)			Uncertain (%)
	Labeled Positive ↑	False Negative ↑	Semantic Valid ↑	Hard Negative ↓	Easy Negative ↓	Semantic Error ↓	
VLM2Vec	67.14	8.22	75.36	17.19	6.85	24.04	0.60
Baseline + GM							
LLaVE	65.68	8.53	74.21	17.58	7.67	25.25	0.54
EGA	66.68	8.31	74.99	17.24	7.04	24.28	0.74
Baseline + FN							
GloFND	65.63	8.28	73.91	18.78	6.65	25.43	0.67
FN Risk	66.61	8.60	75.21	17.65	6.64	24.29	0.50
Baseline + BC							
FALCON	66.78	8.43	75.21	18.14	5.97	24.11	0.68
HOBIT	66.75	8.46	75.21	17.51	6.43	23.94	0.85
Baseline + FN + HN + GM							
UniME-V2	66.43	**9.01**	75.44	17.32	6.64	23.96	0.60
Baseline + FN + HN + BC							
B3	**70.58**	8.22	**78.81**	**15.58**	4.94	**20.53**	0.67
Baseline + FN + HN + BC + GM							
B3 + EGA	69.92	8.36	78.28	16.31	**4.65**	20.96	0.76

errors but substantially increases HN errors (EN: -0.29; HN: $+0.73$), resulting in higher Semantic Error ($+0.43$). Meanwhile, the modest increase in FN does not offset the larger decline in LP (FN: $+0.14$; LP: -0.66), leading to lower Semantic Valid (-0.53). Thus, gradient upweighting remains detrimental even with stronger instance-level negative selection and batch-level construction.

We attribute these results to a mismatch between negative hardness and label reliability. A high similarity score indicates that a candidate is difficult for the current model, but does not determine whether it is semantically invalid. High-similarity negatives therefore include both genuine hard negatives that provide useful gradients and false negatives that introduce misleading gradients. Weighting candidates solely by similarity amplifies both types of signals while reducing the relative gradient contribution of easy negatives, narrowing the supervision provided across different negative types. The B3 + EGA results further suggest that stronger negative selection alone does not resolve this issue, as similarity-based reweighting can still overemphasize the hardest pairs and increase HN errors. Together, these effects explain why upweighting high-similarity negatives can degrade performance rather than improve negative discrimination.

4.3 Similarity Separation Does Not Reliably Predict Performance

Takeaway: *Since aggregate similarity distributions do not capture the instance-level ordering that determines retrieval accuracy, greater separation between positive and negative distributions does not reliably indicate better downstream performance.*

Prior studies have treated stronger separation between positive and negative similarity distributions as an indicator of embedding quality (Ye et al., 2019; Lan et al., 2025). However, Figure 2 shows that

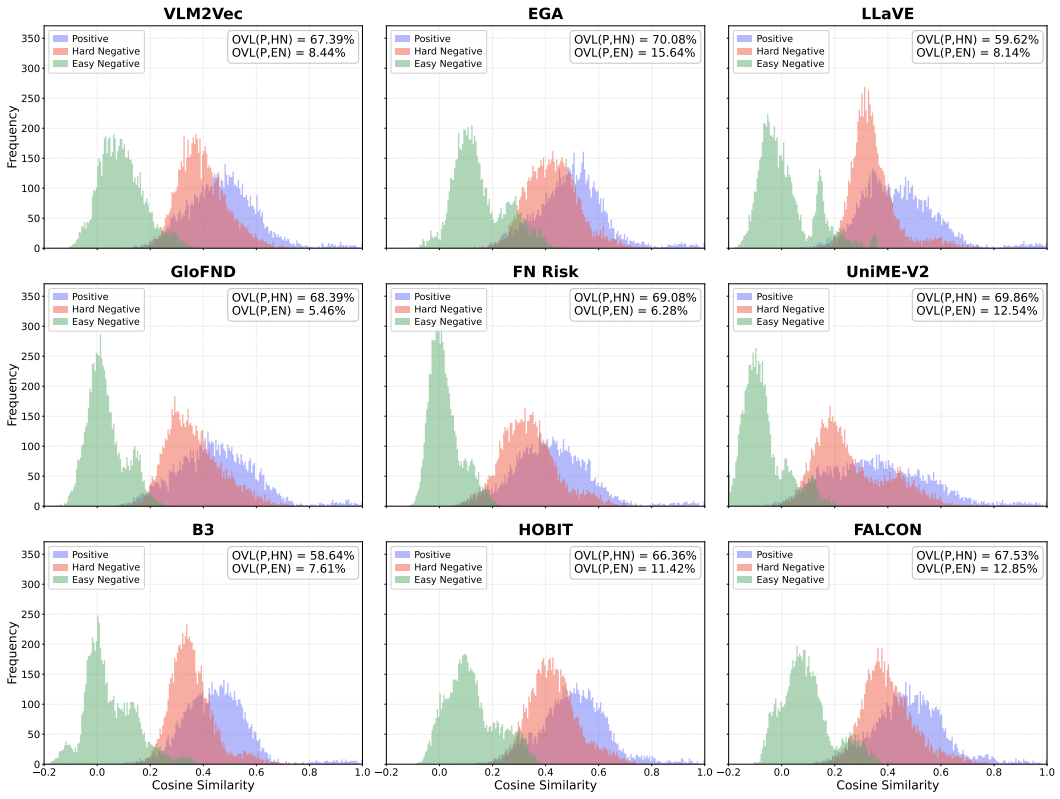


Figure 2: ***Similarity distributions of nine training strategies on MMEB.*** All methods are evaluated on the same 7,200 queries, comprising 200 randomly sampled cases from each of the 36 datasets. *Positive (P)* denotes the similarity between a query and its designated target, while *Hard Negative (HN)* and *Easy Negative (EN)* denote the mean similarities to its Top-5 and Bottom-5 non-positive candidates, respectively. *OVL(P,HN)* and *OVL(P,EN)* measure the distribution overlap between the Positive distribution and the HN and EN distributions, respectively; lower values indicate better distributional separation.

distributional changes do not consistently align with downstream performance. Compared with the baseline, LLaVE reduces both positive–hard-negative and positive–easy-negative overlap (LLaVE: $-7.77/-0.30$), whereas EGA increases both ($+2.69/+7.20$). Despite these opposite changes, both methods degrade consistently in the overall MMEB evaluation (LLaVE/EGA: $-1.2/-0.3$) and the MLLM-as-a-Judge analysis (Semantic Valid: $-1.15/-0.37$; Semantic Error: $+1.21/+0.24$). Thus, substantially different similarity distributions can correspond to the same downstream performance trend.

Comparisons across other strategies further confirm this mismatch. B3 and LLaVE exhibit similar overlap rates (B3/LLaVE: HN $58.64/59.62$; EN $7.61/8.14$), yet B3 outperforms LLaVE by 5.1 MMEB points and 4.60% in Semantic Valid. Conversely, UniME-V2 (Gu et al., 2026) increases both HN and EN overlap relative to the baseline ($+2.47/+4.10$) while retaining the same MMEB score of 65.1 and nearly identical Semantic Valid ($+0.08$). These results show that neither lower distributional overlap nor a particular aggregate similarity geometry reliably predicts downstream performance.

This discrepancy arises because clearer separation between global similarity distributions does not necessarily indicate better positive–negative separation for each query. Globally distinct distributions may still hide many local ranking errors, whereas overlapping distributions can still preserve the correct ordering for most queries. Similarity distributions therefore provide a useful view of embedding geometry, but are not reliable standalone indicators or optimization signals for negative quality.

4.4 RELIABLE CONTRASTIVE TRAINING REQUIRES JOINT DESIGN

Takeaway: *Neither false-negative mitigation, hard-negative mining, nor hard-batch construction yields consistent gains in isolation. False-negative mitigation removes unreliable negatives without providing informative ones; hard-negative mining incorrectly selects semantically valid false negatives; and batch-level construction does not guarantee sufficiently hard negatives for each instance. The strongest performance comes from jointly filtering false negatives, mining informative instance-level hard negatives, and increasing their coverage through hard-batch construction.*

As shown in Table 4, false-negative treatment alone provides only limited improvement in recovering semantically valid alternatives, while introducing more errors on LP and HN. Compared with the baseline, GloFND (Balmaseda et al., 2025) and FN Risk (Zhong et al., 2017) improve FN recognition by only $+0.05/+0.38\%$, but decrease LP accuracy by $-1.51/-0.53$ and increase HN errors by $+1.59/+0.46$. Consequently, their Semantic Valid rates decrease by $-1.46/-0.15\%$, consistent with the drops in their overall MMEB scores (GloFND/FN Risk: $-0.8/-0.1$ points). Batch-construction methods provide only limited improvement in recognizing FN and filtering EN, while introducing more errors on LP and HN. Compared with the baseline, FALCON (Kim et al., 2026) and HOBIT (Dutta et al.) slightly improve FN recognition ($+0.21/+0.24\%$) and reduce EN errors ($-0.88/-0.42$), but simultaneously decrease LP accuracy ($-0.36/-0.39$) and increase HN errors ($+0.95/+0.32$). As a result, their Semantic Valid rates both reach 75.21%, remaining close to the baseline of 75.36%, which is consistent with the limited changes in their overall MMEB scores (FALCON/HOBIT: $-0.3/+0.2$ points). These results indicate that addressing either label noise or batch difficulty alone does not consistently improve semantic discrimination.

In contrast, B3 jointly combines false-negative treatment, hard-negative mining, and hard-batch construction. It increases Semantic Valid from 75.36% to 78.81% while reducing both HN and EN errors from 17.19% and 6.85% to 15.58% and 4.94%, respectively. These improvements translate into an overall MMEB score of 69.0, substantially outperforming all isolated strategies. UniME-V2, which combines false-negative treatment and hard-negative mining without hard-batch construction, remains at the baseline score of 65.1, further highlighting the importance of coordinating instance- and batch-level interventions. Moreover, the B3 ablations in Appendix B.3 show that removing any of false-negative treatment, hard-negative mining, or hard-batch construction reduces performance, confirming that the three components provide complementary benefits.

This complementarity arises because the three components address different limitations of contrastive supervision. False-negative treatment improves label reliability but may also reduce training difficulty if informative negatives are removed without replacement. Hard-negative mining supplies challenging pairwise supervision, but requires false-negative filtering to ensure that the selected candidates are semantically invalid. Hard-batch construction then increases the density and coverage of these reliable hard negatives within each optimization step. Combining the three therefore provides supervision that is simultaneously reliable, informative, and sufficiently challenging, leading to the strongest overall performance.

5 CONCLUSION

In this work, we systematically examined representative contrastive training strategies for universal multimodal embedding learning under a unified experimental setting on MMEB. Across Qwen2-VL-2B/7B and Qwen3.5-2B, most strategies provide little improvement over a well-aligned baseline and many even degrade performance, revealing that previously reported gains do not persist under controlled comparisons. To understand these results beyond aggregate benchmark scores, we introduced an MLLM-as-a-Judge framework that distinguishes among labeled positives, false negatives, and hard- and easy-negative errors Our analysis reveals that upweighting high-similarity negatives can amplify both useful and misleading gradients. Wider separation between aggregate similarity distributions does not reliably indicate better instance-level retrieval. False-negative mitigation, hard-negative mining, and hard-batch construction address complementary aspects of contrastive supervision and achieve the strongest performance when jointly applied. Overall, these findings clarify the effectiveness and limitations of existing training strategies and provide practical guidance for training universal multimodal embedding models.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing the manuscript for clarity and readability. Additionally, we used generative AI tools for editing a research paper to improve readability. All AI-assisted content was critically reviewed by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Vicente Balmaseda, Bokun Wang, Ching-Long Lin, and Tianbao Yang. Discovering global false negatives on the fly for self-supervised contrastive learning. *arXiv preprint arXiv:2502.20612*, 2025.

ByteDance Seed Team. Built on Seed1.6-Flash, Seed-1.6-Embedding launched, June 2025. URL https://seed.bytedance.com/en/blog/built-on-seed1-6-flash-seed-1-6-embedding-launched. Published June 28, 2025.

Jaeseok Byun, Dohoon Kim, and Taesup Moon. Mafa: Managing false negatives for vision-language pre-training. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27304–27314. IEEE, 2024.

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

Himanshu Dutta, Lokesh Nagalapatti, and Yashoteja Prabhu. Hobit: Hardness optimized batch sampling for infonce training. In *Forty-third International Conference on Machine Learning*.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Haohang Huang, Xuan Lu, Mingyi Su, Xuan Zhang, Ziyan Jiang, Ping Nie, Kai Zou, Tomas Pfister, Wenhu Chen, Wei Zhang, et al. Mmeb-v3: Measuring the performance gaps of omni-modality embedding models. *arXiv preprint arXiv:2604.23321*, 2026.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

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.

Myunsoo Kim, Seongwoong Shim, and Byung-Jun Lee. Falcon: False-negative aware learning of contrastive negatives in vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 701–711, 2026.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, and Jinsong Su. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. In *EMNLP (Findings)*, pp. 13721–13735, 2025.

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.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. @ crepe: Can vision-language foundation models reason compositionally? In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10910–10921. IEEE, 2023.

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.

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.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Baifeng Shi, Boyi Li, Han Cai, Yao Lu, Sifei Liu, Marco Pavone, Jan Kautz, Song Han, Trevor Darrell, Pavlo Molchanov, et al. Scaling vision pre-training to 4k resolution. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9631–9640. IEEE, 2025.

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

Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 6397–6406. IEEE, 2020.

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

Raghuveer Thirukovalluru, Rui Meng, Ye Liu, Mingyi Su, Ping Nie, Semih Yavuz, Yingbo Zhou, Wenhu Chen, Bhuwan Dhingra, et al. Breaking the batch barrier (b3) of contrastive learning via smart batch mining. *Advances in Neural Information Processing Systems*, 38:23218–23241, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Youze Xue, Dian Li, and Gang Liu. Improve multi-modal embedding learning via explicit hard negative gradient amplifying. *arXiv preprint arXiv:2506.02020*, 2025.

Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6203–6212. IEEE, 2019.

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

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European conference on computer vision*, pp. 310–325. Springer, 2024.

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

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16772–16782. IEEE, 2022.

Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 3652–3661. IEEE, 2017.

Appendix

A Experimental Details

A.1 Contrastive Training Strategies

VLM2Vec. VLM2Vec (Jiang et al., 2024b) provides a general framework for adapting vision-language models into instruction-aware multimodal embedding models through contrastive learning. It encodes multimodal queries and candidates with a shared backbone and optimizes an InfoNCE objective using other candidates in the mini-batch as negatives. We adopt VLM2Vec as the common baseline upon which all evaluated contrastive training strategies are implemented.

LLaVE. LLaVE (Lan et al., 2025) introduces hardness-weighted contrastive learning to emphasize difficult in-batch negatives. It estimates negative hardness using the current query–negative similarity and assigns larger weights to negatives with higher similarity when computing the contrastive objective.

EGA. EGA (Xue et al., 2025) analyzes the gradients of the InfoNCE objective and explicitly redistributes gradient contributions among negative samples. It measures hardness using the similarity margin between each negative and the labeled positive, allocating a larger proportion of the negative gradient to samples closer to the positive.

GloFND. GloFND (Balmaseda et al., 2025) dynamically estimates a global false-negative threshold for each training instance. The thresholds are updated from sampled mini-batches to approximate per-instance similarity quantiles over the full dataset, and negatives exceeding the corresponding thresholds are excluded from the contrastive objective.

FN Risk. FN Risk is our controlled false-negative treatment strategy based on dynamically updated neighborhood information (Zhong et al., 2017). It estimates the risk of each negative using its similarity to the labeled positive, neighborhood overlap, and mutual-neighbor relationships, and continuously down-weights high-risk negatives instead of removing them entirely.

FALCON. FALCON (Kim et al., 2026) learns an adaptive scheduler for constructing mini-batches that balance hard-negative informativeness against false-negative risk. For each anchor, the scheduler predicts a hardness quantile over the current similarity distribution and sequentially selects samples at the corresponding quantiles to form a mini-batch.

HOBIT. HOBIT (Dutta et al.) constructs informative mini-batches using an InfoNCE-motivated criterion that favors negatives similar to the query but dissimilar to its labeled positive. It formulates batch construction as a monotone submodular optimization problem and greedily selects examples that maximize the collective hardness of in-batch negatives.

UniME-V2. UniME-V2 (Gu et al., 2026) first performs global retrieval to construct candidate pools and then employs an MLLM judge to estimate the semantic alignment of query–candidate pairs. The resulting scores are used to filter false negatives, select diverse hard negatives, and provide soft semantic targets for training the embedding model.

B3. B3 (Thirukovalluru et al., 2026) uses a pretrained teacher model to rank training examples and construct a sparse preference graph of hard-negative relationships. It excludes the highest-ranked candidates as potential false negatives, applies community detection to identify groups of mutually hard negatives, and forms mini-batches by sampling from the resulting communities.

A.2 Model Backbones

Qwen2-VL. Qwen2-VL (Wang et al., 2024) is a family of MLLMs that integrates a vision encoder with a decoder-only language model. It adopts Naive Dynamic Resolution to represent images of

Table B.1: ***Comparison of multimodal contrastive training strategies on MMEB using Qwen3.5-2B and Qwen2-VL-7B.*** We extend the unified comparison in Table 3 under the same training and evaluation protocol. Scores are mean Precision@1 (%) over the corresponding datasets, and *Overall* is the average across all 36 MMEB datasets. GM, FN, HN, and BC denote gradient modulation, false-negative mitigation, hard-negative mining, and batch construction, respectively. The best result in each column under each backbone is highlighted in **bold**.

Method	Optimization Dimensions	Per Meta-Task Score				Overall Score
		Image Classification	Visual Question Answering	Image-level Retrieval	Visual Grounding	
No. of Datasets →		10	10	12	4	36
Qwen3.5-2B						
VLM2Vec (Baseline)	–	65.2	67.4	67.3	84.6	68.7
EGA	GM	65.4	**67.8**	67.4	82.5	68.6
FN Risk	FN	65.3	66.0	67.0	84.3	68.2
HOBIT	BC	65.8	66.7	68.7	83.7	69.0
UniME-V2	FN + HN + GM	63.5	63.9	65.0	84.9	66.5
B3	FN + HN + BC	**66.9**	67.2	**70.0**	**86.5**	**70.2**
B3 + EGA	FN + HN + BC + GM	66.7	67.3	69.7	85.1	69.9
Qwen2-VL-7B						
VLM2Vec (Baseline)	–	67.4	64.9	71.7	83.7	70.0
EGA	GM	67.5	64.8	71.3	83.2	69.8
FN Risk	FN	67.3	65.0	72.0	83.6	70.0
HOBIT	BC	66.6	65.7	72.4	83.9	70.2
UniME-V2	FN + HN + GM	64.2	65.3	73.0	81.7	69.4
B3	FN + HN + BC	**70.6**	**69.1**	**75.1**	**85.6**	**73.3**
B3 + EGA	FN + HN + BC + GM	70.5	69.0	74.5	85.3	73.1

different resolutions and aspect ratios with variable numbers of visual tokens, together with Multimodal Rotary Position Embedding (M-RoPE) to model one-dimensional text, two-dimensional images, and three-dimensional videos. We use the instruction-tuned 2B model as our primary backbone for controlled comparisons and the 7B model to examine whether our findings remain consistent at a larger model scale.

Qwen3.5. Qwen3.5 (Team, 2026) is a native multimodal model jointly pretrained on visual and textual inputs. It combines a vision encoder with a hybrid language backbone that interleaves Gated DeltaNet and gated full-attention layers. We use the 2B model to examine whether our conclusions generalize across model generations and sequence-modeling architectures.

A.3 Evaluation Benchmark

MMEB. The MMEB (Jiang et al., 2024b) evaluates universal multimodal embedding models on 36 datasets spanning four meta-tasks: image classification, visual question answering, image-level retrieval, and visual grounding. All tasks are formulated as retrieval problems over multimodal candidate sets, and models generate instruction-conditioned embeddings for both queries and candidates. We report Precision@1 averaged across the 36 datasets, together with scores for each meta-task and the in- and out-of-distribution subsets.

Prompt Formatting. We follow the official MMEB instruction format without additional prompt tuning. Queries are encoded using the task-specific instructions provided by each MMEB dataset. For textual targets, we use modality-level templates: `Represent the class label:` for classification, `Represent the answer:` for visual question answering, and `Represent the image caption:` for image-to-text retrieval. Other targets, including full images and cropped regions, retain their original MMEB instructions. The same prompt formatting is used across all models, training strategies, checkpoints, and ablation experiments.

Table B.2: ***Robustness of MLLM-as-Judge analysis.*** We evaluate predictions from the same VLM2Vec-2B checkpoint using different MLLM judges and evaluation sets. *Full set* includes all available evaluation cases from the 36 MMEB datasets, while the *Subset* contains the 7,200 cases used in the main analysis, with 200 cases randomly sampled from each dataset. Outcome definitions follow Table 4. All values are percentages.

Judge / Set	Positive Outcomes (%)			Negative Outcomes (%)			Uncertain (%)
	Labeled Positive ↑	False Negative ↑	Semantic Valid ↑	Hard Negative ↓	Easy Negative ↓	Semantic Error ↓	
Qwen3.8-Flash *Full set*	67.35	8.08	75.43	17.16	6.75	23.91	0.66
Qwen3.8-Flash *Subset*	67.14	8.22	75.36	17.19	6.85	24.04	0.60
GPT-5.4 *Subset*	67.14	8.07	75.21	16.93	6.71	23.64	1.15

Table B.3: ***Ablation study of B3 using Qwen2-VL-2B on MMEB.*** We remove false-negative mitigation (FN), hard-negative mining (HN), or batch construction (BC) from the full B3 framework while keeping the remaining training and evaluation settings unchanged. Scores are mean Precision@1 (%) over the corresponding datasets, and *Overall* is the average across all 36 MMEB datasets. The best result in each column is highlighted in **bold**.

Method	Optimization Dimensions	Per Meta-Task Score				Overall Score
		Image Classification	Visual Question Answering	Image-level Retrieval	Visual Grounding	
No. of Datasets →		10	10	12	4	36
Qwen2-VL-2B						
B3 (Full)	FN + HN + BC	67.3	63.0	**71.4**	**81.4**	**69.0**
Random Batch	FN + HN	67.3	60.7	71.0	78.1	67.9
w/o FN Filtering	HN + BC	67.3	61.7	70.7	80.7	68.4
w/o HN Mining	FN + BC	**67.5**	**63.2**	71.0	79.0	68.8

B Ablation Studies

B.1 Cross-Model Comparison

To examine whether our observations depend on the Qwen2-VL-2B backbone, we repeat the unified comparison using Qwen3.5-2B and Qwen2-VL-7B. As shown in Table B.1, the results closely reproduce the trends in Table 3. Methods targeting a single optimization dimension provide limited or inconsistent improvements: EGA slightly underperforms VLM2Vec on both additional backbones, FN Risk is comparable to or below the baseline, and HOBIT yields only modest gains. UniME-V2 also fails to consistently improve over the baseline despite combining false-negative mitigation, hard-negative mining, and gradient modulation. In contrast, B3 achieves the strongest overall performance, improving over VLM2Vec by 1.5 points on Qwen3.5-2B and 3.3 points on Qwen2-VL-7B, with broadly consistent gains across the four meta-tasks. Adding EGA to B3 slightly reduces the overall score by 0.3 and 0.2 points, respectively, mirroring the result on Qwen2-VL-2B. These results demonstrate that our main conclusions generalize across model backbones: jointly improving false-negative reliability, instance-level hard-negative quality, and hard-negative coverage is consistently more effective than optimizing individual components in isolation, while additional gradient modulation provides no complementary benefit in this setting.

B.2 Robustness of MLLM-as-Judge

To examine whether our MLLM-as-Judge analysis is sensitive to the evaluation-set size or the choice of judge, we repeat the analysis on predictions from the same VLM2Vec-2B checkpoint under three

INPUT

MLLM-as-Judge prompt:

Input and image roles

You are a **strict multimodal relevance judge** auditing an embedding benchmark.
The user input is a JSON object. Its **image_roles** array is strictly parallel to the provided image array:
image 1 = image_roles[0]
image 2 = image_roles[1]
image 3 = image_roles[2]
Possible image roles:
query: original query image
positive: dataset-designated positive
predict: embedding model's top-1 prediction
positive+predict: the same image serves as both positive and prediction
Always identify each image using image_roles.
Never infer image roles from visual order alone.
Use visual evidence and text together.
Use uncertain only when the evidence genuinely cannot support a decision.

Two judgments

1. Designated positive
Determine whether candidate_positive genuinely satisfies the query.
true_positive: correctly satisfies the query
false_positive: annotation or pairing error
uncertain: insufficient evidence
Do not automatically trust the dataset label.
Semantic equivalents, synonyms, paraphrases, and equally valid answers count as true positives.
2. Top-1 prediction
If exact_top1_match is true, return **same_as_positive**. Otherwise classify candidate_predict as:
false_negative: also genuinely satisfies the query
hard_negative: related, confusable, or wrong by a subtle decisive detail
easy_negative: clearly incorrect, unrelated, or easily distinguishable
uncertain: insufficient evidence

Task criteria and output

image_to_text_retrieval: caption accurately describes the query image
text_to_image_retrieval: image satisfies the query caption; irrelevant minor differences are allowed
visual_question_answering: answer is correct from the query image; synonyms and equivalent specificity are allowed
image_classification: class correctly identifies the depicted object
object_grounding: crop contains and reasonably isolates the requested object

Return JSON only

```
{
  "positive": {
    "judgment": "true_positive | false_positive | uncertain",
    "confidence": 0.0,
    "reason": "brief evidence-based reason"
  },
  "predict": {
    "category": "false_negative | hard_negative | easy_negative | same_as_positive | uncertain",
    "confidence": 0.0,
    "reason": "brief evidence-based reason"
  }
}
```

Confidence must be a number between 0 and 1.

Figure C.1: ***Overview of the MLLM-as-Judge prompt.*** The judge receives the query, the dataset-designated positive, the model's top-1 prediction, and an explicit mapping between images and their roles. It first verifies the designated positive and then categorizes the top-1 prediction as the designated positive, a false negative, a hard negative, an easy negative, or an uncertain case. Task-specific criteria guide judgments across the five MMEB task families, and a constrained JSON schema enables automatic parsing and aggregation.

configurations (Table B.2). Reducing the Qwen3.8-Flash evaluation from the full 36,000 cases to the 7,200-case subset changes every outcome by at most 0.21 percentage points. Replacing Qwen3.8-Flash with GPT-5.4 on the same subset produces similarly small differences: all positive and negative outcome categories change by at most 0.40 percentage points. These results confirm that the 7,200-case evaluation used in the main text is representative of the full evaluation set and yields consistent semantic-error patterns across different judge models.

B.3 Ablation of the Joint Design in B3

In Section 4.4, we demonstrate the effectiveness of the joint design by comparing methods that cover different optimization dimensions. To further identify the contribution of each optimization dimension, we conduct a controlled ablation within B3 by individually removing false-negative mitigation, hard-negative mining, or hard-batch construction while keeping all other implementation and training settings unchanged. As shown in Table B.3, replacing hard-batch construction with random batching reduces the overall MMEB score from 69.0 to 67.9, while removing false-negative filtering or hard-negative mining lowers it to 68.4 and 68.8, respectively. Although some ablations yield marginal improvements on individual meta-tasks, none matches the overall performance of the complete B3 configuration. These results further support our finding: false-negative mitigation, hard-negative mining, and hard-batch construction provide complementary benefits and achieve the strongest performance when optimized jointly.

C MLLM-as-Judge Evaluation Protocol

Figure C.1 summarizes our MLLM-as-Judge prompt. For each case, the judge receives the query, dataset-designated positive, model's top-1 prediction, task type, and explicit image-role mappings. It first verifies whether the designated positive is valid and then categorizes the top-1 prediction as the designated positive, a false negative, a hard negative, an easy negative, or an uncertain case. Task-specific criteria are provided for the five MMEB task families, and the judgment is returned using a constrained JSON schema for automatic parsing. We use the same prompt and output schema for all methods, with Qwen3.8-Flash as the default judge.