UniME-V2: MLLM-as-a-Judge for Universal Multimodal Embedding Learning

Tiancheng Gu[1,2*], **Kaicheng Yang**[3*], **Kaichen Zhang**[1,4], **Xiang An**[3], **Ziyong Feng**[3], **Yueyi Zhang**[1‡], **Weidong Cai**[2], **Jiankang Deng**[5‡], **Lidong Bing**[1]

[1] MiroMind AI
[2] The University of Sydney
[3] M.R.L. Team
[4] LMMs-Lab Team
[5] Imperial College London

Abstract

Universal multimodal embedding models are essential in various tasks. Existing approaches typically use in-batch mining to identify hard negatives by measuring the similarity of query-candidate pairs. However, these methods often struggle to capture subtle semantic differences among candidates and lack diversity in negative samples. Moreover, the embeddings exhibit limited discriminative ability in distinguishing false and hard negatives. In this paper, we leverage the advanced understanding capabilities of MLLMs to enhance representation learning, and present a novel **Uni**versal **M**ultimodal **E**mbedding (**UniME-V2**) model. Our approach first constructs a potential hard negative set through global retrieval. We then introduce the MLLM-as-a-Judge mechanism, which utilizes MLLMs to assess the semantic alignment of query-candidate pairs and generate soft semantic matching scores. These scores serve as a foundation for hard negative mining, mitigating the impact of false negatives and enabling the identification of diverse, high-quality hard negatives. Furthermore, the semantic matching scores are used as soft labels to mitigate the rigid one-to-one mapping constraint. By aligning the similarity matrix with the soft semantic matching score matrix, the model learns semantic distinctions among candidates, significantly enhancing its discriminative capacity. To further improve performance, we propose **UniME-V2-Reranker**, a reranking model trained on our mined hard negatives through a joint pairwise and listwise optimization approach. We conduct comprehensive experiments on the MMEB benchmark and multiple retrieval tasks, demonstrating that our method achieves state-of-the-art performance across all tasks.

Code — https://garygutc.github.io/UniME-v2
Extended version — https://arxiv.org/pdf/2510.13515

Introduction

Multimodal embedding models aim to encode heterogeneous multimodal data into a unified dense representation space, enabling a wide range of downstream applications such as visual question answering (Dong et al. 2025; Hamza et al. 2025) and multimodal retrieval. With the increasing adoption of these models, multimodal representation learning has attracted significant attention from researchers. Among these models, CLIP (Radford et al. 2021) stands out as a pioneering approach, achieving remarkable performance in text-image retrieval by leveraging cross-modal contrastive learning on large-scale web-collected image-text pairs. However, its effectiveness is hindered by three major limitations: (1) CLIP enforces a strict text token limit of 77, which restricts its ability to process detailed or lengthy descriptions (Zhang et al. 2024a; Cao, Wei, and Ma 2025; Huang et al. 2024); (2) Its dual-encoder design processes images and text independently, which reduces its effectiveness in handling complex tasks, such as instruction-following multimodal retrieval (Jiang et al. 2025; Liu et al. 2024; Gu et al. 2025; Jiang et al. 2025); and (3) CLIP exhibits limited proficiency in advanced language understanding, struggles with handling compositionality, and often demonstrates bag-of-words behavior (Yuksekgonul et al.

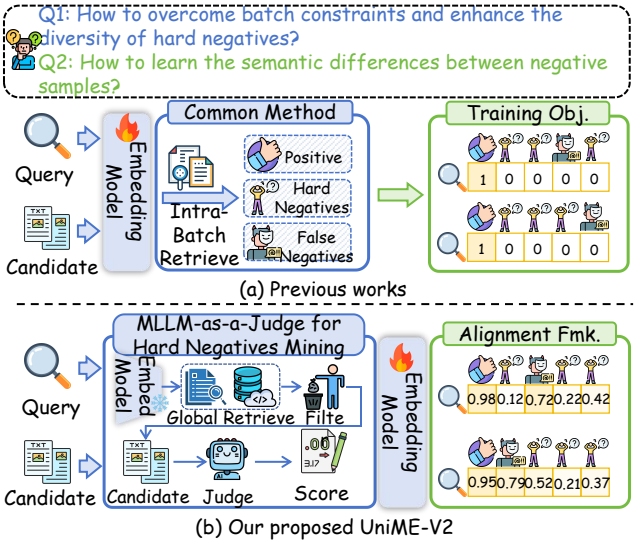

Figure 1: Comparison between previous works and UniME-V2. UniME-V2 exploits the understanding capabilities of MLLMs for hard negatives mining and generates a soft semantic matching score to supervise the model in learning the semantic difference among candidates.

[*] Equal contribution.
[‡] Corresponding Author

2022; Tschannen et al. 2023; Hu et al. 2025).

Recent advances in Large Language Models (LLMs) have achieved state-of-the-art performance on the MTEB benchmark (Muennighoff et al. 2023). Motivated by these developments (Lee et al. 2024; BehnamGhader et al. 2024), researchers are currently exploring how to utilize Multimodal Large Language Models (MLLMs) to learn universal multimodal representation. E5-V (Jiang et al. 2024) adopts a unimodal contrastive learning approach, training the language component of MLLMs on sentence pairs to address disparities between cross-modal representations. VLM2Vec (Jiang et al. 2025) introduces the Massive Multimodal Embedding Benchmark (MMEB), comprising 36 datasets across four meta-tasks, and develops a contrastive learning framework that adapts state-of-the-art vision-language models into embedding models by training on the MMEB dataset. QQMM (Xue, Li, and Liu 2025a) provides an in-depth analysis of the gradients derived from the InfoNCE loss and proposes amplifying gradients associated with hard negative samples to encourage the model to learn more discriminative embeddings. UniME (Gu et al. 2025) presents a two-stage framework that leverages a powerful LLM-based teacher model to improve the embedding capabilities of the language component in MLLMs. Furthermore, it incorporates a hard negative sampling strategy that selects multiple challenging negatives per instance within each batch. Despite these advances, existing methods fail to fully exploit the semantic differences between candidates and are limited by the lack of diversity in negative samples. Additionally, directly utilizing embeddings proves insufficient to effectively distinguish between false negatives and hard negatives.

In this paper, we propose a novel **Uni**versal **M**ultimodal **E**mbedding (**UniME-V2**) model, which leverages the robust understanding capabilities of MLLMs to enhance representation learning. As shown in Fig. 1, we first construct a potential hard negative set through global retrieval. Then, we introduce MLLM-as-a-Judge to assess the semantic alignment of query-candidate pairs, producing semantic matching scores. This score serves as the foundation for hard negative mining, effectively reducing interference from false negatives and enabling the identification of high-quality, diverse hard negatives. Additionally, we use the scores as soft labels to mitigate strict one-to-one mapping constraints. Aligning the similarity matrix with the semantic score matrix enables the model to capture semantic distinctions among candidates, significantly improving its discriminative ability. To further enhance performance, we introduce **UniME-V2-Reranker**, a reranking model trained on our mined hard negatives through a joint pairwise and listwise optimization approach. Extensive experiments on the MMEB benchmark and various retrieval tasks, including short/long caption retrieval and compositional retrieval, demonstrate that our method achieves state-of-the-art performance across all tasks. The main contributions of this paper are summarized as follows:

- We introduce **a MLLM-as-a-Judge pipeline for hard negative mining** that uses the advanced understanding capabilities of MLLM to assess the semantic alignment of each query-candidate pair within a globally retrieved potential hard negative set.
- We present **UniME-V2**, a novel universal multimodal embedding model trained with **an MLLM judgment based distribution alignment framework**. By leveraging semantic matching scores as soft labels, the model effectively captures semantic differences between candidates, significantly enhancing its discriminative capability.
- We propose **UniME-V2-Reranker**, a reranking model trained on high-quality, diverse hard negatives through a joint pairwise and listwise optimization approach.
- We conduct **extensive experiments** on the MMEB benchmark and various retrieval tasks, including short and long caption retrieval as well as compositional retrieval. The results demonstrate that our method achieves state-of-the-art performance across all tasks.

Related Work

Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) extend traditional LLMs to process and integrate information across multiple modalities (Xie et al. 2024; Bai et al. 2025). As a foundational contribution, LLaVA (Liu et al. 2023) leverages a subset of the CC3M (Changpinyo et al. 2021) dataset to achieve more balanced conceptual coverage. In this approach, the visual encoder and language model remain frozen, while only the projection layer is trained to align visual features with language tokens. Subsequently, numerous MLLM variants (Peng et al. 2024; Lin et al. 2024; Zhang et al. 2024b; Zhu et al. 2023) have achieved remarkable results in multimodal understanding and reasoning tasks. For instance, CogVLM (Wang et al. 2024b) incorporates a trainable visual expert module into the attention and feed-forward layers of the language model, achieving significant improvements on 17 standard cross-modal benchmarks. Similarly, Qwen2-VL (Wang et al. 2024a) introduces the Naive Dynamic Resolution mechanism and integrates M-RoPE to enhance positional information fusion, yielding competitive performance across diverse benchmarks. LLaVA-OneVision (Li et al. 2024) pushes the boundaries of open MLLMs by excelling in single-image, multi-image, and video tasks, showcasing robust video understanding through effective task transfer from image-based training. Although these advances have significantly improved the understanding capabilities of MLLMs, further research is needed to explore how MLLMs can effectively learn unified multimodal representations.

Multimodal Representation Learning

CLIP (Radford et al. 2021) demonstrates strong image-text retrieval performance through large-scale cross-modal contrastive learning but faces three key limitations: (1) A 77-token text truncation restricts fine-grained semantic alignment (Zhang et al. 2024a; Cao, Wei, and Ma 2025; Huang et al. 2024); (2) Its dual-encoder architecture limits effective cross-modal fusion, particularly for instruction-sensitive tasks (Jiang et al. 2025; Liu et al. 2024; Gu et al. 2025); and (3) Simplistic language modeling results in bag-of-words representations (Yuksekgonul et al. 2022; Tschannen et al. 2023; Hu et al. 2025). To address these issues, recent studies have incorporated MLLMs for enhanced multimodal representation

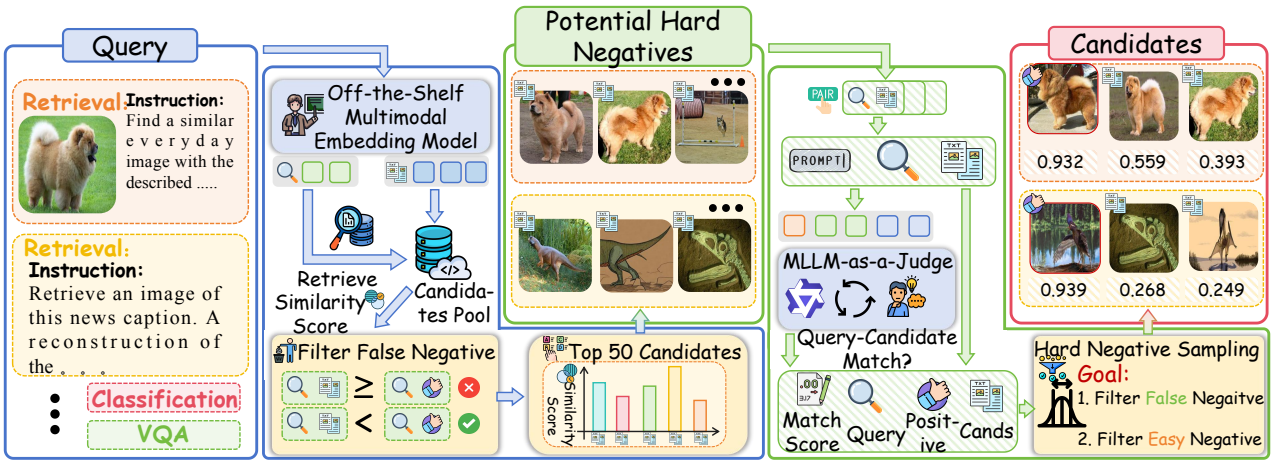

Figure 2: The MLLM-as-a-Judge pipeline for Hard Negatives Mining. We first utilize an existing multimodal embedding model for global retrieval to construct a potential hard negative set. We then leverage the powerful understanding capabilities of MLLM to score query-candidate pairs based on their semantic alignment, enabling precise identification of hard negatives.

learning. E5-V (Jiang et al. 2024) employs unimodal contrastive learning, training the language component of MLLMs on sentence pairs to reduce cross-modal representation gaps. VLM2Vec (Jiang et al. 2025) introduces the Massive Multimodal Embedding Benchmark (MMEB) and adapts state-of-the-art vision-language models into embedding models using a contrastive framework trained on MMEB. QQMM (Xue, Li, and Liu 2025a) analyzes InfoNCE loss gradients and proposes enhancing gradients associated with hard negatives to improve embedding discrimination. UniME (Gu et al. 2025) adopts a two-stage framework with an LLM-based teacher model to refine language embeddings and employs a hard negative sampling strategy, selecting multiple challenging negatives per batch. Despite these advancements, existing methods still under-utilize the semantic differences among candidates and struggle to effectively identify and leverage hard negatives during retrieval.

Methodology

Task Definition

Unlike CLIP, which employs separate encoders to generate embeddings for each modality, we investigate leveraging the unified architecture of MLLM to extract embeddings across multiple modalities and improve retrieval performance through reranking. Specifically, given a query q and a set of candidates $\Omega_c = \{c_1, c_2, \ldots, c_n\}$, which may include images, text, and interleaved image-text data, the universal embedding model Φ_{emb} encodes the query and candidates, retrieving the top-k most relevant candidates $\Omega_k = \Phi_{\text{emb}}(q, \Omega_c)$. To further enhance retrieval performance, a reranker model Φ_{rank} refines this subset through a reranking process, producing the final ranked output $\hat{\Omega}_k = \Phi_{\text{rank}}(q, \Omega_k)$.

MLLM-as-a-Judge for Hard Negatives Mining

Previous works (Jiang et al. 2025; Liu et al. 2024) primarily rely on in-batch hard negative mining, where query-candidate embedding similarities are computed to sample negatives. However, this method often suffers from limited negative sample diversity and insufficient embedding discriminative power to effectively distinguish false and hard negatives. To overcome these challenges, as shown in Fig. 2, we first construct a potential hard negative set using global retrieval. After that, inspired by previous work (Zheng et al. 2023; Chen et al. 2024a), we leverage the robust understanding capabilities of MLLMs to assess the semantic alignment of each query-candidate pair and generate a soft semantic matching score. This score guides hard negative mining, enabling the identification of diverse and high-quality hard negatives while reducing the impact of false negatives.

Potential Hard Negative Set. To extract higher-quality hard negatives from global samples, we first use VLM2Vec to generate embeddings for both queries and candidates. We then retrieve the top 50 most relevant candidates for each query. To address false negatives and improve diversity, we apply a similarity threshold (δ) based on the query-candidate similarity scores and select the top 50 candidates as the potential hard negative set (Ω_p):

$$\Omega_p = \text{Rank}_{50}\left(\{x_1, \ldots, x_n\}\right), \text{where } x_i < \delta, \qquad (1)$$

where x_i is the similarity score of the query q and candidates $\hat{\Omega}_c$ calculated by the VLM2Vec model.

Semantic Matching Score. After constructing the potential hard negative set (Ω_p), we employ an MLLM as a judge to compute a semantic matching score for each query-candidate pair in Ω_p, guided by the following instruction:

> *I will provide you with a query and a candidate. Please evaluate whether the candidate meets the requirements of the query. If it does, respond with 'Yes'; if it doesn't, respond with 'No'. Query:<Query>, Candidates:<Candidate>.*

After that, we compute the semantic matching score $\text{S} = \{s_1, s_2, \ldots, s_m\}$ based on the logits of the `Yes` (e_y) and

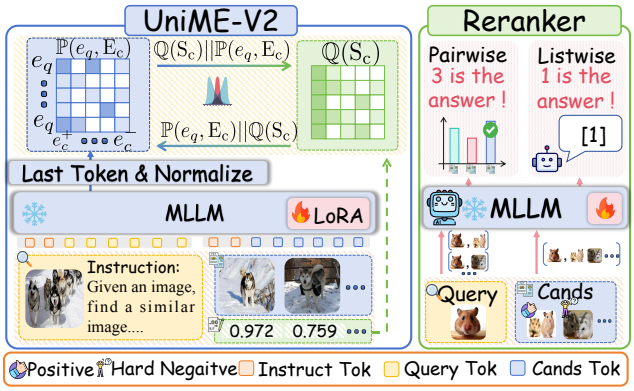

Figure 3: The architecture of the MLLM Judgment Based Training Framework. UniME-V2 uses soft semantic matching scores as supervised signals to enhance semantic distinction learning between candidates. UniME-V2-Reranker employs joint pairwise and listwise optimization to enhance reranking performance. Best viewed by zooming in.

NO (e_n) tokens, where $s_i = \frac{e_y^i}{e_y^i + e_n^i}$, where $\mathrm{S} \in \mathbb{R}^{n_q \times 50}$ and n_q denotes the number of queries. Leveraging the advanced understanding capabilities of MLLMs, the semantic matching score S effectively captures the degree of semantic alignment between queries and candidates.

Hard Negative Sampling. To enhance the quality of hard negatives, we refine candidates using the semantic matching score (S). False negatives are excluded if their score exceeds a threshold $\alpha = \sigma_{q,c_t} - \beta$, where c_t denotes the positive sample and β is a hyperparameter controlling the threshold margin. To maintain diversity, we apply a cyclical sampling strategy with five-step intervals. If the refined set contains fewer than ten candidates, we duplicate selections to ensure a minimum of ten. In the rare case where no candidates meet the criteria ($< 1\%$), we randomly select 10 candidates from the initial pool of fifty and assign each a semantic matching score of 1.0. Finally, for each query q, we obtain the hard negative set $\Omega_h = \{c_1, ..., c_k\}$ along with the corresponding semantic matching scores $\mathrm{S_h} = \{s_{q,c_1}, ..., s_{q,c_k}\}$.

MLLM Judgment Based Training Framework

UniME-V2. Previous works (Jiang et al. 2025; Gu et al. 2025) are limited by a rigid one-to-one mapping, which restricts the ability to learn distinctions among diverse negative samples. To address this, as shown in Fig.3, we propose an MLLM judgment based distribution alignment framework, leveraging soft semantic matching scores as supervised signals to improve representation performance. Specifically, given a query q and its candidate set $\Omega_c = \{c_t, c_1, ..., c_k\}$, we input them into the MLLM and extract the last token as embeddings for the query e_q and candidates $\mathrm{E_c} = \{e_c^+, e_{c_1}^-, .., e_{c_k}^-\}$, where e_c^+ is the embedding of target candidate and k is the hard negative number for each query. We then compute the relation score matrix between the query embedding e_q and candidate embeddings $\mathrm{E_c}$ as follows:

$$\mathbb{P}(e_q, \mathrm{E_c}) = \frac{exp(cos(e_q, e_c^+)/\tau)}{exp(cos(e_q, e_c^+)/\tau) + \sum_{i=1}^{k} exp(cos(e_q, e_{c_i^-})/\tau)}. \tag{2}$$

Based on the semantic matching scores $\mathrm{S_c} = \{s_{q,c_t}, s_{q,c_1}, ..., s_{q,c_k}\}$, we compute the semantic matching score matrix $\mathbb{Q}(e_q, \Omega_c)$ derived from the MLLM judgment as follows:

$$\mathbb{Q}(\mathrm{S_c}) = \frac{exp(s_{q,c_t}/\tau)}{exp(s_{q,c_t}/\tau) + \sum_{i=1}^{k} exp(s_{q,c_i}/\tau)}. \tag{3}$$

To enhance learning robustness and ensure matrix symmetry, we employ JS-Divergence, a symmetric alternative to KL-Divergence (Nielsen 2020). The final loss function $\mathcal{L}$ is defined as:

$$\mathcal{L} = \frac{1}{2}(\frac{1}{N}\sum_{i=1}^{N} \mathrm{KL}(\mathbb{P}(e_i, \mathrm{E_c})||\mathbb{Q}(\mathrm{S_c})) + \frac{1}{N}\sum_{i=1}^{N} \mathrm{KL}(\mathbb{Q}(\mathrm{S_c})||\mathbb{P}(e_i, \mathrm{E_c}))). \tag{4}$$

UniME-V2-Reranker. Following previous works (Liu et al. 2024; Lin et al. 2025), we train a reranking model to enhance retrieval precision following initial embedding-based retrieval. Specifically, we train UniME-V2-Reranker using joint pairwise and listwise approaches to enhance its reranking capability (refer to Fig. 3). In pairwise training, we construct two pairs for each query q by combining with the positive candidate c_t and the hardest negatives c_h. We then instruct UniME-V2-Reranker to output YES for the positive and NO for the negative. The pairwise loss $\mathcal{L}_{pair}$ is computed using the cross-entropy loss function as:

$$\mathcal{L}_{pair} = \mathcal{L}_{ce}(\mathrm{YES}, \eta(q, c_t)) + \mathcal{L}_{ce}(\mathrm{NO}, \eta(q, c_h)), \tag{5}$$

where η denotes the autoregressive output process of UniME-V2-Reranker. For listwise training, based on the semantic matching score, we choose top-x candidates ($\{c_1, ...c_x\}$) from the hard negative candidates, insert the target candidate c_t at a random position and get its index I_{c_t}. The UniME-V2-Reranker is then prompted to output the position of the ground truth, formulated as:

$$\mathcal{L}_{list} = \mathcal{L}_{ce}(I_{c_t}, \eta(q, c_t, \{c_1, ...c_x\})). \tag{6}$$

The final loss function is defined as $\mathcal{L} = \mathcal{L}_{pair} + \mathcal{L}_{list}$. Detailed descriptions of the prompts used for pairwise and listwise training are provided in the supplementary material.

Inference Pipeline

After obtaining UniME-V2 and UniME-V2-Reranker, we integrate them during inference to improve retrieval performance. We initially use UniME-V2 embed query and candidate into features and utilize cosine similarity scores to retrieve the top-10 most relevant candidates. Subsequently, UniME-V2-Reranker reranks these candidates based on the following instruction:

Models	#Parameters	Per Meta-Task Score				Average Score		
		Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# of Datasets →		10	10	12	4	20	16	36
Zero-shot on MMEB								
CLIP (ViT-L)(Jiang et al. 2025)	0.4B	42.8	9.1	53.0	51.8	37.1	38.7	39.2
OpenCLIP (ViT-L)(Radford et al. 2021)	0.4B	41.5	6.9	44.6	53.5	32.8	36.0	36.6
Magiclens (ViT-L)(Zhang et al. 2024c)	0.4B	38.8	8.3	35.4	26.0	31.0	23.7	27.1
SigLIP (So/14)(Zhai et al. 2023)	0.9B	40.3	8.4	31.6	59.5	32.3	38.0	35.0
BLIP2 (ViT-L)(Li et al. 2023)	1.2B	27.0	4.2	33.9	47.0	25.3	25.1	28.0
CLIP (ViT-BigG/14)(Cherti et al. 2022)	2.5B	52.3	14.0	50.5	60.3	38.9	45.8	44.3
EVA-CLIP(Sun et al. 2023)	8B	56.0	10.4	49.2	58.9	38.1	45.6	43.7
E5-V (Phi3.5-V)(Jiang et al. 2024)	4.2B	39.1	9.6	38.0	57.6	33.1	31.9	36.1
E5-V (LLaVA-1.6)(Jiang et al. 2024)	7B	39.7	10.8	39.4	60.2	34.2	33.4	37.5
Fine-tuning on MMEB								
CLIP (ViT-L)(Jiang et al. 2025)	0.4B	55.2	19.7	53.2	62.2	47.6	42.8	47.6
VLM2Vec (Qwen2-VL)(Jiang et al. 2025)	2B	59.0	49.4	65.4	73.4	66.0	52.6	60.1
VLM2Vec (Qwen2-VL)(Jiang et al. 2025)	7B	62.6	57.8	69.9	81.7	72.2	57.8	65.8
LLaVE (LLaVA-OV)(Lan et al. 2025)	7B	65.7	65.4	70.9	**91.9**	75.0	64.4	70.3
QQMM (LLaVA-OV)(Xue, Li, and Liu 2025b)	7B	66.8	66.8	70.5	90.4	74.7	65.6	70.7
UniME (Qwen2-VL)(Gu et al. 2025)	2B	59.0	53.4	64.9	69.6	65.5	54.6	60.6
UniME (Qwen2-VL)(Gu et al. 2025)	7B	64.7	59.0	71.6	82.7	72.2	61.4	67.4
UniME (LLaVA-OV)(Gu et al. 2025)	7B	**66.8**	66.6	70.5	90.9	74.6	65.8	70.7
UniME-V2(Qwen2-VL)	2B	62.1(+3.1)	56.3(+2.9)	68.0(+3.1)	72.7(+3.1)	67.4(+1.9)	58.9(+4.3)	63.6(+3.0)
UniME-V2(Qwen2-VL)	7B	64.0(-0.7)	60.1(+1.1)	73.1(+1.5)	82.8(+0.1)	72.0(-0.2)	63.0(+1.6)	68.0(+0.6)
UniME-V2(LLaVA-OV)	7B	65.3(-1.5)	**67.6**(+1.0)	**72.9**(+2.4)	90.2(-0.7)	**74.8**(+0.2)	**66.7**(+0.9)	**71.2**(+0.5)

Table 1: Results on the MMEB benchmark. IND: in-distribution, OOD: out-of-distribution. Scores are average Precision@1. Detailed results in the Appendix.

I will provide you with a query followed by multiple candidates in the format: (1) candidate1 (2) candidate2, etc. Each candidate is independent of the others. Evaluate each candidate against the query, and respond with the number corresponding to the candidate that best meets the requirements of the query. Query:<Query>, Candidates:<Candidate list>.

Experiments and Results

Implementation

We extract query and candidate embeddings using VLM2Vec (Qwen2-VL-7B) to construct a potential hard negative set. We use the Qwen2.5VL-7B to generate the soft semantic matching score. We train UniME-V2 using two different multimodal large language models: Qwen2-VL (Wang et al. 2024a) and LLaVA-OneVision (Li et al. 2024). To optimize GPU memory, we implement LoRA (rank=16) with DeepSpeed ZeRO stage-2 (Aminabadi et al. 2022). The training of UniME-V2 is conducted on 8×NVIDIA A800 (80GB) GPUs to accommodate the substantial computational demands. We use 336×336 resolution image inputs, set the accumulated batch size to 1024, with learning rates of 1e-4 (Qwen2-VL) and 2e-5 (LLaVA-OneVision). We set the temperature of the Symmetric KL loss $\tau = 0.02$ and sample $k = 8$ hard negatives, and train each model for 2,000 steps.

Datasets and Evaluation

Training Data. Following VLM2Vec (Jiang et al. 2025) and UniME (Gu et al. 2025), we employ 20 in-distribution datasets from the MMEB benchmark, which cover four core multimodal tasks: classification, visual question answering, multimodal retrieval, and visual grounding. This comprehensive training corpus, incorporating both unimodal and multimodal input data, totals 662k carefully curated training pairs, ensuring robust model adaptation across diverse multimodal tasks.

Evaluation. In this study, we evaluate UniME-V2 across both in-distribution (20 test sets) and out-of-distribution (16 test sets) benchmarks from MMEB (Jiang et al. 2025) to assess its multimodal embedding capabilities across diverse retrieval tasks. Following standard evaluation protocols (Liu et al. 2024; Jiang et al. 2025), we report Precision, measuring the proportion of correct matches among the top-ranked candidates for each dataset. To further examine the unimodal embedding performance of UniME-V2, we conduct experiments on multiple cross-modal retrieval tasks, including short-caption image-text retrieval on Flickr30K (Plummer et al. 2015) and COCO2014 (Lin et al. 2014), long-caption image-text retrieval on ShareGPT4V (Chen et al. 2024b) and Urban1K (Zhang et al. 2024a), and compositional retrieval on SugarCrepe (Hsieh et al. 2023). Consistent with the MMEB benchmark, we use Precision as the primary evaluation metric across all datasets.

Main Results

Multi-Modal Retrieval. In Tab. 1, we present the performance of the proposed UniME-V2 compared to existing baseline models. Under identical training data and configurations, UniME-V2 consistently achieves notable performance

Models	#Parameters	Short Caption				Long Caption				Compositional		
		Flickr30K		COCO		ShareGPT4V		Urban1K		SugarCrepe		
		$q^i \to c^t$	$q^t \to c^i$	$q^i \to c^t$	$q^t \to c^i$	$q^i \to c^t$	$q^t \to c^i$	$q^i \to c^t$	$q^t \to c^i$	Replace	Swap	Add
OpenCLIP (ViT-L)(Radford et al. 2021)	0.4B	67.3	87.2	37.0	58.1	81.8	84.0	47.0	47.0	79.5	62.7	74.9
CLIP (ViT-BigG/14)(Cherti et al. 2022)	2.5B	79.5	92.9	51.3	67.3	90.1	93.6	77.8	80.7	86.5	68.9	88.4
EVA-CLIP(Sun et al. 2023)	8B	80.3	**94.5**	52.0	70.1	93.1	91.2	80.4	77.8	85.9	70.3	86.7
E5-V (Phi3.5-V)(Jiang et al. 2024)	4.2B	72.2	79.6	44.7	53.4	86.0	88.5	83.8	83.6	88.2	66.6	75.3
E5-V (LLaVA-1.6)(Jiang et al. 2024)	7B	77.3	85.7	49.1	57.6	85.1	82.1	88.9	83.2	86.3	68.7	66.9
VLM2Vec (Qwen2-VL)(Jiang et al. 2025)	2B	69.3	89.6	40.0	62.5	78.1	88.2	78.7	83.9	67.2	46.5	66.4
VLM2Vec (Qwen2-VL)(Jiang et al. 2025)	7B	80.0	94.2	49.2	68.5	78.5	90.4	94.0	94.2	70.0	51.7	72.2
UniME (Qwen2-VL)(Gu et al. 2025)	2B	74.9	90.6	44.0	63.5	83.6	88.6	83.3	83.2	65.6	45.2	65.7
UniME (Qwen2-VL)(Gu et al. 2025)	7B	80.8	92.7	50.9	69.8	86.5	93.8	95.3	94.0	68.8	53.0	69.8
UniME (LLaVA-OV)(Gu et al. 2025)	7B	83.3	94.4	54.8	74.0	93.9	89.3	94.3	95.5	80.5	65.5	82.2
UniME-V2 (Qwen2-VL)	2B	79.8(+4.9)	89.9(-0.7)	53.7(+9.7)	65.1(+1.6)	91.6(+8.0)	94.(+5.6)	95.6(+12.3)	92.2(+9.0)	70.9(+5.3)	51.2(+6.0)	70.2(+4.5)
UniME-V2 (Qwen2-VL)	7B	84.6(+3.8)	93.5(+0.8)	57.3(+6.4)	70.3(+0.5)	94.3(+0.8)	**95.2**(+1.4)	**97.2**(+1.9)	96.3(+2.3)	77.8(+9.0)	62.2(+9.2)	79.0(+9.2)
UniME-V2 (LLaVA-OV)	7B	**85.5**(+2.2)	93.7(-0.7)	**60.9**(+6.1)	**74.1**(+0.1)	**95.1**(+1.2)	94.1(+4.8)	96.3(+2.0)	**96.7**(+1.2)	**88.6**(+8.1)	**73.7**(+8.2)	**90.5**(+8.3)

Table 2: Zero-shot text-image retrieval results on short caption (Flickr30K, MS-COCO), long caption (ShareGPT4V, Urban1K) and compositional (SugarCrepe) datasets. Scores are Recall@1.

improvements across various foundation models. Specifically, UniME-V2 outperforms VLM2Vec by 3.5% and 2.2% on the Qwen2-VL-2B and 7B models, respectively. When built on LLaVA-OneVision as the foundation, UniME-V2 achieves a 0.5%-0.9% improvement over previous state-of-the-art models, including QQMM, LLaVE, and UniME. Furthermore, UniME-V2 attains a score of 66.7 on out-of-distribution datasets, significantly exceeding all prior approaches, highlighting its robustness and superior transferability.

Short & Long Caption Cross-Modal Retrieval. We evaluate UniME-V2 on zero-shot cross-modal retrieval tasks. For short-caption datasets, including Flickr30K and MS-COCO, UniME-V2 demonstrates a 2.2%-9.7% performance improvement in image-to-text retrieval compared to UniME. In text-to-image retrieval, its performance is comparable to UniME, primarily due to two factors: (1) the limited proportion of text-to-image data in the MMEB training set and (2) the insufficient semantic information in short captions. For long-caption cross-modal retrieval tasks, UniME-V2 achieves significant improvements on ShareGPT4V and Urban1K, benefitting from its enhanced discriminative capability and the richer semantic content provided by detailed captions. Notably, compared to EVA-CLIP-8B, UniME-V2 demonstrates more robust retrieval performance. This is primarily due to its universal multimodal embedding can significantly reduce the modality gap (as shown in Fig. 4).

Compositional Cross-Modal Retrieval. We evaluate the capacity of the UniME-V2 model to discriminate hard negative samples using the compositional benchmark SugarCrepe. As shown in Tab. 2, UniME-V2 consistently delivers superior performance across all evaluated metrics. Compared with UniME, our model achieves 5.3%, 6.0%, 4.5% performance improvement using Qwen2-VL-2B. After scaling the model from 2B to 7B, our model also achieves 9.0%, 9.2%, and 9.2% performance improvement. Additionally, UniME-V2 exhibits improvements of 2.7%, 3.4%, and 3.8% compared to EVA-CLIP-8B, underscoring its robust capability to discriminate against hard negative samples.

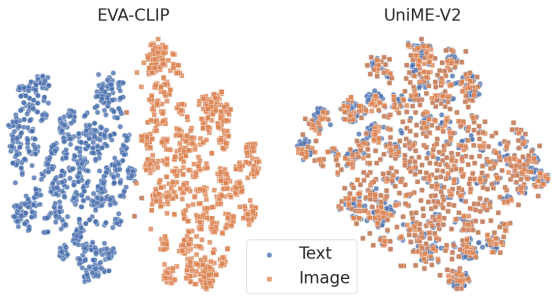

Figure 4: Comparison of representation distributions between EVA-CLIP-8B (left) and UniME-V2 (LLaVA-OV) (right).

Embedding Model	Reranker	#Data	MMEB	R_{Short}	R_{Long}	R_{Compos}
UniME(2B)	—	—	60.6	68.3	84.7	58.8
UniME-V2(2B)	—	—	63.6	72.1	93.4	64.1
UniME-V2(2B)	LamRA(7B)	1.1M	67.3	76.4	96.4	87.4
UniME-V2(2B)	UniME-V2(7B)	0.6M	**67.6**	**76.4**	**96.9**	**94.8**
UniME(7B)	—	—	67.4	73.6	92.4	63.9
UniME-V2(7B)	—	—	68.0	76.4	95.8	73.0
UniME-V2(7B)	LamRA(7B)	1.1M	69.1	78.3	97.2	87.4
UniME-V2(7B)	UniME-V2(7B)	0.6M	**69.6**	**78.7**	**97.5**	**94.8**

Table 3: Comparison of reranking performance between LamRA and UniME-V2-Reranker using UniME-V2 (Qwen2-VL-7B) and UniME-V2 (Qwen2-VL-2B).

Reranking Comparison. In Tab. 3, we compare the performance between LamRA and UniME-V2-Reranker using listwise reranking on the top-5 retrieval results. To ensure fairness, we use the same training parameters and base model (Qwen2.5-VL-7B) as LamRA. When UniME-V2 (Qwen2-VL-2B) is used for retrieval, both LamRA and UniME-V2-Reranker improve performance across four downstream tasks, with UniME-V2-Reranker consistently achieving superior results while utilizing only half the data. Similarly, with UniME-V2 (Qwen2-VL-7B) for retrieval, UniME-V2-Reranker outperforms LamRA, achieving performance gains of 0.5%, 0.4%, 0.3%, and 7.4% across the four tasks. Notably, UniME-V2-Reranker demonstrates a significant advantage over LamRA in compositional understanding retrieval tasks,

Hard Negatives	Soft Score	MMEB	R_{Short}	R_{Long}	R_{Compos}
✘	✘	60.1	63.4	82.2	60.0
✔	✘	61.6	68.9	89.8	63.7
✔	✔	**63.6**	**72.1**	**93.4**	**64.1**

Table 4: Ablation study on our proposed MLLM-as-a-Judge hard negatives mining method and MLLM judgment based training framework.

Judge Model	MMEB	R_{Short}	R_{Long}	R_{Compos}
Qwen2.5VL-7B	**63.6**	**72.1**	**93.4**	**64.1**
InternVL3-8B	58.5	70.2	91.3	64.1

Table 5: Ablation study on different MLLM-based judges.

#Negatives	MMEB	R_{Short}	R_{Long}	R_{Compos}
4	61.3	69.2	91.0	62.4
6	61.8	70.8	91.7	61.2
8	**63.6**	**72.1**	**93.4**	**64.1**
10	63.0	72.0	93.4	63.4

Table 6: Ablation study on the number of hard negatives.

attributed to its use of MLLM's understanding capabilities to extract diverse and high-quality hard samples, which effectively enhance the model's discriminative power.

Analysis

Ablation on Different Components. We evaluate the effectiveness of UniME-V2 through ablation studies on the proposed MLLM-as-a-Judge hard negatives mining method and the MLLM judgment based training framework, utilizing Qwen2-VL-2B. As shown in Tab. 4, our proposed hard negatives mining method achieves performance improvements of 1.5%, 5.5%, 7.6%, and 3.7% over direct contrastive learning (e.g., VLM2Vec) on the MMEB, short-retrieval, long-retrieval, and composed-retrieval tasks, respectively. Building on this, the introduction of the MLLM judgment based training framework further enhances the model's discriminative ability by capturing finer semantic distinctions among candidate samples, leading to additional performance gains of 2.0%, 3.2%, 3.6%, and 0.4% for the corresponding tasks.

Ablation on Different MLLM-based Judges. The comprehension ability of the MLLM acting as a judge directly impacts the accuracy of the generated semantic matching scores, thereby influencing the final model performance. Therefore, based on Qwen2-VL-2B, we compare two influential MLLMs in the current open-source community: Qwen2.5-VL-7B and InterVL3-8B. As shown in Tab. 5, under the same inference settings, the quality of semantic matching scores produced by Qwen2.5-VL is significantly superior to that of InterVL3-8B, particularly on the MMEB (63.6 v.s. 58.5). The primary reason can be attributed to differences in the distribution of instruction data used during their SFT phase.

Ablation on the Number of Hard Negatives. Tab. 6 presents the impact of varying the number of hard negatives based on Qwen2-VL-2B. When the number of hard negative samples increases from 4 to 8, UniME-V2 demonstrates consistent improvements across all evaluation metrics: +2.3% on MMEB, +2.9% on short retrieval, +2.4% on long retrieval, and +1.7% on composed retrieval. These gains can be attributed to the model's enhanced ability to discriminate between candidates during training. However, further increasing to 10 introduces easier negatives, diminishing discriminative learning, and slightly reducing performance.

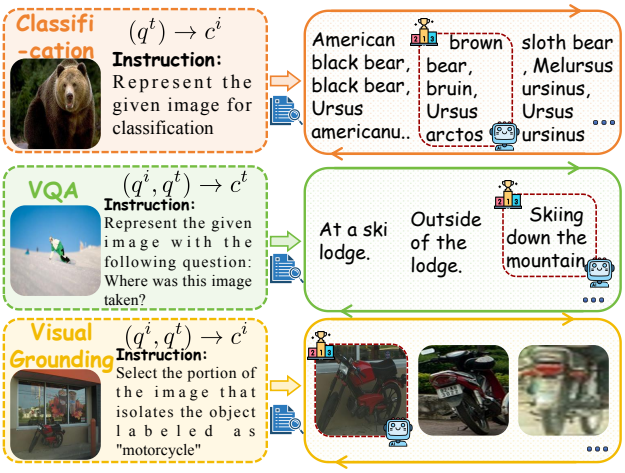

Figure 5: Qualitative examples. We present the retrieval and reranking results of our method across different tasks.

Qualitative Results. Fig. 5 illustrates the qualitative results of our method across various tasks. Retrieval results from UniME-V2 are shown, with the top-1 candidate refined by UniME-V2-Reranker highlighted in red dashed boxes. UniME-V2 effectively retrieves query-relevant candidates, such as "black bear" and "brown bear" in the first example, while UniME-V2-Reranker further refines the ranking of retrieved results, prioritizing "brown bear" over "black bear".

Conclusion

In this paper, we explore how to leverage the advanced understanding capabilities of MLLMs to enhance representation learning and propose a novel Universal Multimodal Embedding model (UniME-V2). Specifically, we first construct a potential hard negative set using global retrieval. We then introduce MLLM-as-a-Judge, which utilizes the robust semantic understanding of MLLMs to assess the alignment of query-candidate pairs and generate soft semantic matching scores. These scores guide hard negative mining by reducing false negative interference and identifying high-quality, diverse hard negatives. Additionally, the scores serve as soft labels, relaxing the rigid one-to-one mapping constraint. By aligning the similarity matrix with the soft semantic matching score matrix, the model learns finer-grained semantic distinctions among candidates, thereby enhancing its discriminative power. To further improve performance, we propose UniME-V2-Reranker, which incorporates joint pairwise and listwise reranking optimization based on the mined hard negatives. We conduct extensive experiments on the MMEB benchmark and various retrieval tasks, including short and long caption retrieval as well as compositional retrieval. Our method achieves state-of-the-art performance across all tasks.

References

Aminabadi, R. Y.; Rajbhandari, S.; Awan, A. A.; Li, C.; Li, D.; Zheng, E.; Ruwase, O.; Smith, S.; Zhang, M.; Rasley, J.; et al. 2022. Deepspeed-inference: enabling efficient inference of transformer models at unprecedented scale. In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–15. IEEE.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

BehnamGhader, P.; Adlakha, V.; Mosbach, M.; Bahdanau, D.; Chapados, N.; and Reddy, S. 2024. Llm2vec: Large language models are secretly powerful text encoders. *COLM*.

Cao, A.; Wei, X.; and Ma, Z. 2025. FLAME: Frozen Large Language Models Enable Data-Efficient Language-Image Pre-training. In *CVPR*.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 3558–3568.

Chen, D.; Chen, R.; Zhang, S.; Wang, Y.; Liu, Y.; Zhou, H.; Zhang, Q.; Wan, Y.; Zhou, P.; and Sun, L. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *ICML*.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2024b. Sharegpt4v: Improving large multi-modal models with better captions. In *ECCV*.

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2022. Reproducible scaling laws for contrastive language-image learning. *arXiv:2212.07143*.

Dong, G.; Song, X.; Zhu, Y.; Qiao, R.; Dou, Z.; and Wen, J.-R. 2025. Toward general instruction-following alignment for retrieval-augmented generation. In *AAAI*.

Gu, T.; Yang, K.; Feng, Z.; Wang, X.; Zhang, Y.; Long, D.; Chen, Y.; Cai, W.; and Deng, J. 2025. Breaking the Modality Barrier: Universal Embedding Learning with Multimodal LLMs. In *ACM MM*.

Hamza, A.; Ahn, Y. H.; Lee, S.; Kim, S. T.; et al. 2025. Llava needs more knowledge: Retrieval augmented natural language generation with knowledge graph for explaining thoracic pathologies. In *AAAI*.

Hsieh, C.-Y.; Zhang, J.; Ma, Z.; Kembhavi, A.; and Krishna, R. 2023. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *NeurIPS*.

Hu, X.; Yang, K.; Wang, J.; Xu, H.; Feng, Z.; and Wang, Y. 2025. Decoupled Global-Local Alignment for Improving Compositional Understanding. In *ACM MM*.

Huang, W.; Wu, A.; Yang, Y.; Luo, X.; Yang, Y.; Hu, L.; Dai, Q.; Dai, X.; Chen, D.; Luo, C.; et al. 2024. Llm2clip: Powerful language model unlock richer visual representation. *arXiv:2411.04997*.

Jiang, T.; Song, M.; Zhang, Z.; Huang, H.; Deng, W.; Sun, F.; Zhang, Q.; Wang, D.; and Zhuang, F. 2024. E5-v: Universal embeddings with multimodal large language models. *arXiv:2407.12580*.

Jiang, Z.; Meng, R.; Yang, X.; Yavuz, S.; Zhou, Y.; and Chen, W. 2025. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *ICLR*.

Lan, Z.; Niu, L.; Meng, F.; Zhou, J.; and Su, J. 2025. Llave: Large language and vision embedding models with hardness-weighted contrastive learning. *arXiv preprint arXiv:2503.04812*.

Lee, C.; Roy, R.; Xu, M.; Raiman, J.; Shoeybi, M.; Catanzaro, B.; and Ping, W. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *ICLR*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. Vila: On pre-training for visual language models. In *CVPR*.

Lin, S.-C.; Lee, C.; Shoeybi, M.; Lin, J.; Catanzaro, B.; and Ping, W. 2025. Mm-embed: Universal multimodal retrieval with multimodal llms. In *ICLR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *NeurIPS*.

Liu, Y.; Chen, P.; Cai, J.; Jiang, X.; Hu, Y.; Yao, J.; Wang, Y.; and Xie, W. 2024. LamRA: Large Multimodal Model as Your Advanced Retrieval Assistant. *CVPR*.

Muennighoff, N.; Tazi, N.; Magne, L.; and Reimers, N. 2023. MTEB: Massive Text Embedding Benchmark. In *EACL*.

Nielsen, F. 2020. On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid. *Entropy*.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2024. Kosmos-2: Grounding multimodal large language models to the world. In *ICLR*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Sun, Q.; Wang, J.; Yu, Q.; Cui, Y.; Zhang, F.; Zhang, X.; and Wang, X. 2023. EVA-CLIP-18B: Scaling CLIP to 18 Billion Parameters. *arXiv:2402.04252*.

Tschannen, M.; Kumar, M.; Steiner, A.; Zhai, X.; Houlsby, N.; and Beyer, L. 2023. Image captioners are scalable vision learners too. *NeurIPS*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; Xu, J.; Xu, B.; Li, J.; Dong, Y.; Ding, M.; and Tang, J. 2024b. CogVLM: Visual Expert for Pretrained Language Models. In *Neurips*.

Xie, Y.; Yang, K.; Yang, N.; Deng, W.; Dai, X.; Gu, T.; Wang, Y.; An, X.; Zhao, Y.; Feng, Z.; et al. 2024. Croc: Pretraining large multimodal models with cross-modal comprehension. *arXiv preprint arXiv:2410.14332*.

Xue, Y.; Li, D.; and Liu, G. 2025a. Improve Multi-Modal Embedding Learning via Explicit Hard Negative Gradient Amplifying. *arXiv preprint arXiv:2506.02020*.

Xue, Y.; Li, D.; and Liu, G. 2025b. Improve Multi-Modal Embedding Learning via Explicit Hard Negative Gradient Amplifying. *arXiv preprint arXiv:2506.02020*.

Yuksekgonul, M.; Bianchi, F.; Kalluri, P.; Jurafsky, D.; and Zou, J. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv:2210.01936*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *ICCV*.

Zhang, B.; Zhang, P.; Dong, X.; Zang, Y.; and Wang, J. 2024a. Long-clip: Unlocking the long-text capability of clip. In *ECCV*.

Zhang, H.; Li, H.; Li, F.; Ren, T.; Zou, X.; Liu, S.; Huang, S.; Gao, J.; Leizhang; Li, C.; et al. 2024b. Llava-grounding: Grounded visual chat with large multimodal models. In *ECCV*.

Zhang, K.; Luan, Y.; Hu, H.; Lee, K.; Qiao, S.; Chen, W.; Su, Y.; and Chang, M.-W. 2024c. Magiclens: Self-supervised image retrieval with open-ended instructions. In *ICML*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeuriPS*, 36: 46595–46623.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*.