LLaVE: Large Language and Vision Embedding Models with Hardness-Weighted Contrastive Learning

Zhibin Lan[1*], **Liqiang Niu**[2], **Fandong Meng**[2], **Jie Zhou**[2], **Jinsong Su**[1,3†]
[1]School of Informatics, Xiamen University, China,
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, China,
[3]Shanghai Artificial Intelligence Laboratory, China
lanzhibin@stu.xmu.edu.cn, jssu@xmu.edu.cn
{poetniu, fandongmeng, withtomzhou}@tencent.com

LLaVE LLaVE-0.5B LLaVE-2B LLaVE-7B

Abstract

Universal multimodal embedding models play a critical role in tasks such as interleaved image-text retrieval, multimodal RAG, and multimodal clustering. However, our empirical results indicate that existing LMM-based embedding models trained with the standard InfoNCE loss exhibit a high degree of overlap in similarity distribution between positive and negative pairs, making it challenging to distinguish hard negative pairs effectively. To deal with this issue, we propose a simple yet effective framework that dynamically improves the embedding model's representation learning for negative pairs based on their discriminative difficulty. Within this framework, we train a series of models, named LLaVE, and evaluate them on the MMEB benchmark, which covers 4 meta-tasks and 36 datasets. Experimental results show that LLaVE establishes stronger baselines that achieve state-of-the-art (SOTA) performance while demonstrating strong scalability and efficiency. Specifically, LLaVE-2B surpasses the previous SOTA 7B models, while LLaVE-7B achieves a further performance improvement of 6.2 points. Although LLaVE is trained on image-text data, it can generalize to text-video retrieval tasks in a zero-shot manner and achieve strong performance, demonstrating its remarkable potential for transfer to other embedding tasks.

1 Introduction

Multimodal embedding models aim to encode inputs from any modality into vector representations, which then facilitate various multimodal tasks, such as image-text retrieval (Wu et al., 2021a; Zhang et al., 2024a), automatic evaluation (Hessel et al., 2021), and retrieval-augmented generation (RAG) (Zhao et al., 2023). Although advanced pretrained vision-language models such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and SigLIP (Zhai et al., 2023) can provide unified representations for text and images, they face difficulties when dealing with more complex tasks. Particularly, they adopt a dual-encoder architecture that encodes images and text separately, leading to poor performance in tasks such as interleaved image-text retrieval (Wu et al., 2021a).

In recent years, the rapid development and exceptional performance of large multimodal models (LMMs) have prompted researchers to focus increasingly on LMM-based multimodal embedding models. Compared to traditional pretrained vision-language models, LMMs not only demonstrate superior multimodal semantic understanding capabilities but also naturally support interleaved text-image inputs (OpenAI, 2023; Liu et al., 2023, 2024a; Lan et al., 2025; Hu et al., 2025). This advantage makes them more flexible and efficient in handling multimodal embedding tasks. As a representative work, Jiang et al. (2024b) construct the Massive Multimodal Embedding Benchmark (MMEB), which encompasses 4 meta-tasks and 36 datasets, and train multimodal embedding models based on LMMs. Experiment results demonstrate that by providing suitable task instructions and employing contrastive learning for training, LMM can significantly outperform existing multimodal embedding models and generalize effectively across diverse tasks.

However, as shown in Figure 1(a), our preliminary study finds that when training an LMM as a multimodal embedding model using the standard InfoNCE loss (van den Oord et al., 2018), the query-target similarity distribution of positive and negative pairs exhibits significant overlap. This indicates that the model struggles to learn discriminative multimodal representations for positive and hard negative pairs.

Building on the above observation and insights from preference learning (Rafailov et al., 2023;

[*] Work was done when Zhibin Lan was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.
[†] Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 13721–13735
November 4-9, 2025

Song et al., 2024), we propose a simple yet effective framework to encourage the model to focus more on hard negative pairs, forcing it to learn more discriminative multimodal representations. Under our framework, we consider the embedding model as a policy model and introduce a reward model to assign an adaptive weight to each negative pair, where harder pairs are assigned with larger weights. This ensures that harder negative pairs play a more significant role in model training. In addition, by decoupling the reward model from the policy model, our framework can not only use different models for hardness estimation but also leverage manually annotated hardness to enhance the representation learning for specific samples. Inspired by SigLIP (Zhai et al., 2023), we introduce a cross-device negative sample gathering strategy, which significantly alleviates the issue of limited negative samples in LMMs caused by excessive memory usage. As shown in Figure 1, we observe that our framework increases the query-target similarity gap between positive and negative pairs, indicating its effectiveness in helping the model learn more discriminative multimodal representations.

To evaluate the effectiveness of our framework, we train a series of multimodal embedding models, referred to as LLaVE (Large Language and Vision Embedding Models), within the proposed framework. These models are based on advanced open-source LMMs of varying scales, including LLaVA-OV-0.5B (Li et al., 2024), Aquila-VL-2B (Gu et al., 2024), and LLaVA-OV-7B (Li et al., 2024). Experimental results on MMEB demonstrate that LLaVE-0.5B achieves comparable results to that of the previous VLM2Vec (phi-3.5-V-4B). When scaled up to LLaVE-2B, the model requires only about 17 hours of training on a single machine equipped with 8 A100 GPUs (40GB) to surpass the state-of-the-art (SOTA) model MMRet-7B, which is pretrained on 27 million image-text pairs. Furthermore, when expanded to LLaVE-7B, its performance is even more impressive, surpassing the previous SOTA model by 6.2 points. Meanwhile, when being scaled to different sizes, LLaVE still outperforms the models trained with InfoNCE loss based on the same LMM, demonstrating the effectiveness of our framework. These results fully validate that our framework is both easily scalable and resource-efficient. In addition, despite being trained exclusively on image-text data, LLaVE generalizes effectively to text-video retrieval tasks, showcasing its strong potential for transferring to other embedding tasks.

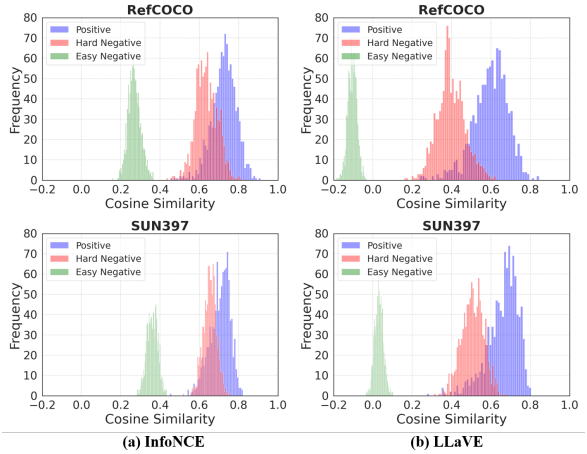

Figure 1: Similarity distributions of learned embeddings on the SUN397 (Xiao et al., 2010) and RefCOCO (Kazemzadeh et al., 2014) dataset. We present the query-target cosine similarity histograms of positive, hard negative, and easy negative pairs for the model trained with the standard InfoNCE loss and LLaVE.

2 Preliminary Study

In this section, we first briefly formulate the multimodal embedding task and review the standard InfoNCE loss. Then, we analyze the similarity distributions of positive and negative pairs in LMM-based embedding models trained with the standard InfoNCE loss.

2.1 Contrastive Learning for LMM-based Multimodal Embedding Models

Following VLM2Vec (Jiang et al., 2024b), we address the challenge of universal retrieval with LMMs. Specifically, given a query-target pair, it can be represented as (q, t^+), where both q and t^+ could be an image, text, or interleaved image-text input. Note that q will be equipped with corresponding task instructions for different tasks. The objective of this task is to ensure that the similarity between q and t^+ is greater than the similarities between q and other negative candidates $\{t^-\}$.

The aim of contrastive learning is to learn discriminative multimodal representations by pulling closer the representations of queries and targets in positive pairs while pushing apart the representations of queries and targets in negative pairs (Hadsell et al., 2006). Given an LMM, we input the query and the target separately into the model and obtain their representations by extracting the vector representations of the last token in the final

Type	Classification		VQA		Retrieval		Visual Grounding	
	SUN397	ImageNet-R	DocVQA	TextVQA	CIRR	Wiki-SS-NQ	MSCOCO	RefCOCO-Matching
InfoNCE								
Positive	0.71	0.68	0.70	0.69	0.78	0.57	0.82	0.73
Hard Negative	0.65(-0.06)	0.59(-0.09)	0.62(-0.08)	0.62(-0.07)	0.76(-0.02)	0.55(-0.02)	0.76(-0.06)	0.64(-0.09)
Easy Negative	0.36(-0.35)	0.38(-0.30)	0.30(-0.40)	0.32(-0.37)	0.41(-0.37)	0.29(-0.28)	0.39(-0.43)	0.27(-0.46)
Precision@1 ↑	66.2	86.4	81.2	76.3	39.7	56.6	76.1	83.8
LLaVE								
Positive	0.66	0.58	0.61	0.59	0.63	0.45	0.62	0.93
Hard Negative	0.51(-0.15)	0.39(-0.19)	0.42(-0.19)	0.45(-0.14)	0.56(-0.07)	0.38(-0.07)	0.52(-0.10)	0.64(-0.29)
Easy Negative	0.03(-0.63)	0.07(-0.51)	-0.04(-0.65)	0.01(-0.58)	0.01(-0.62)	0.03(-0.42)	-0.02(-0.64)	0.05(-0.88)
Precision@1 ↑	**75.5**	**89.1**	**88.5**	**78.8**	**50.0**	**64.4**	**80.0**	**85.5**

Table 1: The average cosine similarity between queries and targets is reported for positive, hard negative, and easy negative pairs across eight datasets. The numbers in parentheses represent the similarity difference between negative and positive pairs, where lower values indicate a smaller overlap between the two similarity distributions. It can be observed that our method effectively increases the similarity gap between negative and positive pairs, resulting in higher precision.

layer. Formally, for a mini-batch of training data $\{(q_1, t_1), ..., (q_N, t_N)\}$, the standard InfoNCE loss is defined as

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \underbrace{-log\frac{e^{s_{i,i}/\tau}}{e^{s_{i,i}/\tau} + \sum_{j\neq i}^{N} e^{s_{i,j}/\tau}}}_{\mathcal{L}_i}, \quad (1)$$

where $s_{i,j} = \text{cosine}(\text{LMM}(q_i), \text{LMM}(t_i))$, $\text{LMM}(\cdot)$ denotes the use of LMM for obtaining the representation, and τ represents the temperature hyper-parameter. In this work, τ is always set to 0.02 following the setting of VLM2Vec (Jiang et al., 2024b).

2.2 Analysis

We use Aquila-VL-2B (Gu et al., 2024) as the base model, which builds upon the LLaVA-OneVision architecture, and perform contrastive learning on the MMEB dataset following the VLM2Vec setup. Then, we define the five pairs with the highest query-target similarities (excluding the positive pairs) as hard negative pairs, and the five pairs with the lowest similarities as easy ones. Subsequently, we use the cosine function to calculate the average query-target similarity for the two groups, respectively. As shown in the left part of Figure 1, we visualize the similarity distributions of the trained model on SUN397 and RefCOCO. It can be observed that the distributions of positive and hard negative pairs exhibit significant overlap, while easy negative pairs also demonstrate relatively high similarities.

To further explore, we randomly select one in-distribution dataset and one out-of-distribution dataset from the four meta-tasks in MMEB to evaluate the model's average query-to-target similarity on positive, hard negative, and easy negative pairs, respectively. As shown in Table 1, the similarity difference between positive and negative pairs in models trained with InfoNCE loss is relatively small (no more than 0.09), especially on CIRR and Wiki-SS-NQ, where the difference is only 0.02. Besides, the model exhibits the lowest precision on these two datasets. **Empirically, we observe that the smaller the similarity difference between positive and negative pairs, the lower the final precision tends to be**. This observation validates the necessity of enhancing learning on hard negative pairs during training, motivating us to explore a simple and effective approach to strengthen the model's learning of negative pairs with varying difficulty levels.

3 Our Framework

In this section, we first illustrate the inherent consistency between preference learning and contrastive learning. We then propose a simple yet effective framework that incorporates hardness-weighted contrastive learning (See Section 3.1) and cross-device negative sample gathering (See Section 3.2).

3.1 Hardness-Weighted Contrastive Learning

Preference learning and contrastive learning share a fundamental goal: modeling relationships between pairs based on relative preference or similarity of the target within the pairs. Generally, preference learning involves a reward model and a policy model. The reward model scores the outputs of the policy model, while the policy model

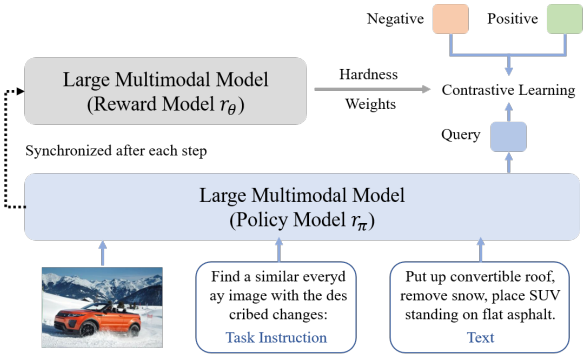

Figure 2: Overview of hardness-weighted contrastive learning. Please note that the policy and reward models are identical in our work. The dashed line indicates directly copying the parameters of the policy model to the reward model.

updates its parameters using the feedback from the reward model to produce higher-reward outputs.

As a typical representative of preference learning, the Bradley-Terry (BT) model (Bradley and Terry, 1952) captures pairwise relationships through probabilistic comparisons. To directly optimize the embedding model (i.e. the policy model), we follow Song et al. (2024) to consider the embedding model as both the reward model and policy model. Formally, given a query q_1 and two targets t_1 and t_2, the BT model defines the training objective of preferring t_1 over t_2 as

$$\mathcal{L}_1 = -\log \frac{e^{r_\pi(q_1,t_1)}}{e^{r_\pi(q_1,t_1)} + e^{r_\pi(q_1,t_2)}}, \quad (2)$$

where, $r_\pi(\cdot)$ denotes the function of reward/policy model. Naturally, we can extend the Bradley-Terry (Bradley and Terry, 1952) model to a one-to-N contrast setting (Song et al., 2024), which is essentially consistent with the InfoNCE loss. As a result, Equation 2 is derived as

$$\mathcal{L}_i = -\log \frac{e^{r_\pi(q_i,t_i)}}{e^{r_\pi(q_i,t_i)} + \sum_{j\neq i}^{N} e^{r_\pi(q_i,t_j)}}, \quad (3)$$

where $r_\pi(q_i,t_j) = s_{i,j}/\tau$, $s_{i,j}$ and τ have been defined in Section 2.1. Based on observations from the preliminary study, we propose hardness-weighted contrastive learning that assigns weight according to the learning difficulty of the negative pair. Higher weights indicate greater difficulty and incur heavier penalties, encouraging the model to learn more from challenging negative pairs. To this end, the training objective is revised as

$$\mathcal{L}_i = -\log \frac{e^{r_\pi(q_i,t_i)}}{e^{r_\pi(q_i,t_i)} + \sum_{j\neq i}^{N} w_{ij} \cdot e^{r_\pi(q_i,t_j)}}, \quad (4)$$

where w_{ij} represents the weight of learning difficulty. As shown in Figure 2, to estimate the learning difficulty of pairs, we introduce a reward model r_θ and set $w_{ij} = e^{r_\theta(q_i,t_j)}$. Accordingly, the policy model adjusts its learning of different negative pairs based on the feedback from the reward model. In this work, to achieve higher training efficiency and simpler implementation, we update the reward model r_θ to keep aligned with the policy model r_π after each step. The reward model does not perform backpropagation, i.e., $r_\theta(q_i,t_j) = \alpha \cdot \mathrm{sg}(s_{ij})$, where α is a hyperparameter and $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. When the reward model is set to be the same as the policy model, another advantage is that assigning a higher reward to a negative sample indicates that the policy model finds it more difficult to distinguish, thereby enhancing its learning on currently hard samples. Note that r_θ can also adopt model structures other than the policy model. Finally, $\mathcal{L}_i$ is defined as

$$\mathcal{L}_i = -\log \frac{e^{r_\pi(q_i,t_i)}}{e^{r_\pi(q_i,t_i)} + \sum_{j\neq i}^{N} e^{(r_\pi(q_i,t_j)+r_\theta(q_i,t_j))}}. \quad (5)$$

We further analyze the gradients with respect to the $r_\pi(q_i,t_j)$ $(j \neq i)$, which are formulated as

$$\frac{\partial L_i}{\partial r_\pi(q_i,t_j)} = e^{r_\theta(q_i,t_j)} \cdot \frac{e^{r_\pi(q_i,t_j)}}{Z_i}, \quad (6)$$

$$Z_i = e^{r_\pi(q_i,t_i)} + \sum_{j\neq i}^{N} e^{(r_\pi(q_i,t_j)+r_\theta(q_i,t_j))}. \quad (7)$$

From Equation 6, we can observe that the gradients of the negative pairs are proportional to the product $r_\theta(q_i,t_j)$. which implies that the greater the learning difficulty of a negative pair, the more significant its role in the gradient update.

3.2 Cross-Device Negative Sample Gathering

The number of negative pairs in contrastive learning has an important effect on model training. However, LMM-based embedding models face the challenge of high memory consumption, making it difficult to use a large batch size directly. To alleviate this issue, inspired by OpenCLIP (Cherti et al., 2023) and SigLIP (Zhai et al., 2023), we adopt a cross-device negative sample gathering strategy, which increases the number of negative pairs by a factor of the device number K. As illustrated in Figure 3, we expand the number of negative pairs on each device by gathering samples from

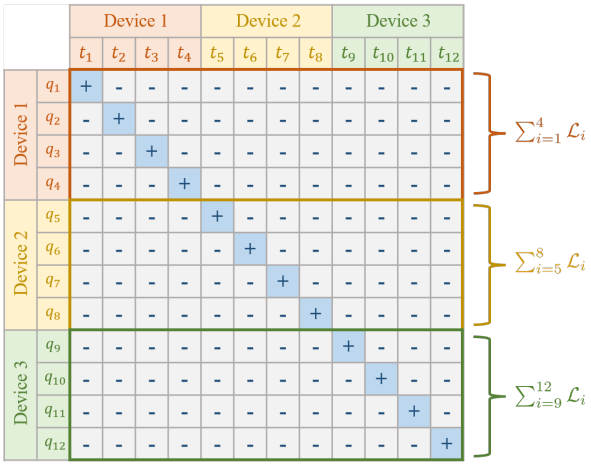

Figure 3: An example of cross-device negative sample gathering (N=4 and K=3). The plus signs represent positive pairs, and the minus signs represent negative pairs. Each device calculates the similarity between its own queries and the targets on all other devices, which is then used for loss computation.

other devices. Consequently, $\mathcal{L}_i$ is reformulated as follows:

$$\mathcal{L}_i = -\log \frac{e^{r_\pi(q_i,t_i)}}{e^{r_\pi(q_i,t_i)} + \sum_{j\neq i}^{N\cdot K} e^{(r_\pi(q_i,t_j)+r_\theta(q_i,t_j))}}. \tag{8}$$

With this strategy, we can effectively increase the number of negative pairs without significantly increasing memory consumption.

4 Experiments

4.1 Setup

Datasets and Metrics. In this study, we follow VLM2Vec (Jiang et al., 2024b) to train our model on 20 in-distribution datasets from MMEB. These datasets encompass four meta-tasks: classification, VQA, multimodal retrieval, and visual grounding, with a total of 662K training pairs. The model is then evaluated on both 20 in-distribution and 16 out-of-distribution test sets from MMEB. We report Precision@1 on each dataset, which measures the proportion of top-ranked candidates that are positive samples.

Implementation Details. Our trained model, LLaVE, includes three scales: 0.5B, 2B, and 7B, based on LLaVA-OV-0.5B (Li et al., 2024), Aquila-VL-2B (Gu et al., 2024), and LLaVA-OV-7B (Li et al., 2024), respectively. To facilitate community use, the training code for LLaVE is built on the widely-used Transformers (Wolf et al., 2020) and DeepSpeed (Rasley et al., 2020) packages. We use a batch size of 256 by gradient accumulation, set the weighting hyperparameter α to 9 [1], and impose a total length limit of 4096. Furthermore, we employ the Higher Anyres technique (Li et al., 2024) to support high-resolution images, setting the maximum image resolution to 672 $\times$ 672. The learning rate is set to 1e-5 for LLaVE-0.5B and LLaVE-2B, and 5e-6 for LLaVE-7B. For efficient training, we freeze the vision encoder and train the model for one epoch using the DeepSpeed ZeRO-3 strategy. Regarding training costs, LLaVE-0.5B and LLaVE-2B are trained on 8 NVIDIA A100 GPUs (40GB) for 12 and 17 hours, respectively, while LLaVE-7B is trained on 16 Ascend 910B GPUs (64GB) for 33 hours. More details can be found in Appendix A.1.

Baselines. Following VLM2Vec, we compare our model with CLIP (Radford et al., 2021), OpenCLIP (Cherti et al., 2023), BLIP2 (Li et al., 2023a), SigLIP (Zhai et al., 2023), UniIR (Wei et al., 2024), E5-V (Jiang et al., 2024a), and Magiclens (Zhang et al., 2024b). Additionally, we include two powerful models: VLM2Vec (Jiang et al., 2024b) and MMRet-MLLM (Zhou et al., 2024). Among them, MMRet-MLLM enhances downstream task performance through pretraining on a self-built retrieval dataset consisting of 26 million pairs. To ensure a fairer comparison, we also compare the VLM2Vec trained using the same base LMM, including VLM2Vec (LLaVA-OV-0.5B), VLM2Vec (Aquila-VL-2B), and VLM2Vec (LLaVA-OV-7B).

4.2 Main Results

Table 2 presents the performance comparison of our proposed LLaVE series (LLaVE-0.5B, LLaVE-2B, LLaVE-7B) against existing baseline models. Among the baseline models, our trained VLM2Vec (LLaVA-OV-7B) achieves the highest overall average score of 65.8, surpassing the current state-of-the-art model, MMRet, which achieves the second-best score of 64.1. This indicates that a more powerful foundational LMM can lead to better performance in the embedding models. Notably, MMRet excels in retrieval tasks with a score of 69.9, which is attributed to its pretraining on its self-constructed 26M image-text retrieval dataset. VLM2Vec (LLaVA-NeXT-7B-HR) exhibits superior performance in grounding tasks, likely due to its higher input image resolution, which achieves a 22.1-point improvement over VLM2Vec (LLaVA-

[1] We empirically analyze the impact of α on model performance in Appendix A.4.

Model	Per Meta-Task Score				Average Score		
	Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# Datasets	10	10	12	4	20	16	36
Baselines							
CLIP (Radford et al., 2021)	42.8	9.1	53.0	51.8	37.1	38.7	37.8
BLIP2 (Li et al., 2023a)	27.0	4.2	33.9	47.0	25.3	25.1	25.2
SigLIP (Zhai et al., 2023)	40.3	8.4	31.6	59.5	32.3	38.0	34.8
OpenCLIP (Cherti et al., 2023)	47.8	10.9	52.3	53.3	39.3	40.2	39.7
UniIR (BLIP$_{FF}$) (Wei et al., 2024)	42.1	15.0	60.1	62.2	44.7	40.4	42.8
UniIR (CLIP$_{SF}$) (Wei et al., 2024)	44.3	16.2	61.8	65.3	47.1	41.7	44.7
Magiclens (Zhang et al., 2024b)	38.8	8.3	35.4	26.0	31.0	23.7	27.8
LMM-based Baselines							
E5-V (Jiang et al., 2024a)	21.8	4.9	11.5	19.0	14.9	11.5	13.3
VLM2Vec (Phi-3.5-V-4B) (Jiang et al., 2024b)	54.8	54.9	62.3	79.5	66.5	52.0	60.1
VLM2Vec (LLaVA-NeXT-7B-LR) (Jiang et al., 2024b)	54.7	50.3	56.2	64.0	61.0	47.5	55.0
VLM2Vec (LLaVA-NeXT-7B-HR) (Jiang et al., 2024b)	61.2	49.9	67.4	86.1	67.5	57.1	62.9
MMRet (LLaVA-NeXT-7B) (Zhou et al., 2024)	56.0	57.4	69.9	83.6	68.0	59.1	64.1
Our trained LMM-based Baselines							
VLM2Vec (LLaVA-OV-0.5B)	54.6	44.7	56.8	76.5	59.8	49.1	55.0
VLM2Vec (Aquila-VL-2B)	61.1	57.3	62.1	85.5	67.2	58.1	63.1
VLM2Vec (LLaVA-OV-7B)	63.5	61.1	64.5	87.3	69.7	61.0	65.8
Ours							
LLaVE-0.5B	57.4	50.3	59.8	82.9	64.7	52.0	59.1
LLaVE-2B	62.1	60.2	65.2	84.9	69.4	59.8	65.2
LLaVE-7B	**65.7**	**65.4**	**70.9**	**91.9**	**75.0**	**64.4**	**70.3**
△ - Best baseline	+2.2	+4.3	+1.0	+4.6	+5.3	+3.4	+4.5

Table 2: Results on the MMEB benchmark. IND represents the in-distribution dataset, and OOD represents the out-of-distribution dataset. In UniIR, the FF and SF subscripts under CLIP or BLIP represent feature-level fusion and score-level fusion, respectively. LLaVA-NeXT-7B-LR indicates the use of low-resolution (336×336) image inputs, while LLaVA-NeXT-7B-HR refers to the use of high-resolution (1344×1344) image inputs. The reported scores are the average Precision@1 over the corresponding datasets. The best results are marked in bold, and the second-best results are underlined. Part of the baseline results are sourced from (Jiang et al., 2024b) and (Zhou et al., 2024). Detailed results and qualitative evaluations can be found in Appendix A.6 and Section 4.5.

NeXT-7B-LR).

Although previous models have achieved strong performance, our LLaVE series demonstrates consistent improvements over the best baseline across all metrics. LLaVA-NeXT-7B achieves a state-of-the-art overall score of 70.3, outperforming the previous SOTA model, MMRet, by 6.2 points and surpassing VLM2Vec (LLaVA-OV-7B) by 4.5 points. In grounding, LLaVE-7B attains an impressive score of 91.9, a +4.6 point improvement over VLM2Vec. LLaVE-7B also leads in VQA (65.4) and classification (65.7), with improvements of +4.3 and +2.2 points, respectively. In addition, we observe that the performance of LLaVE models scales consistently with model size, indicating that our framework has excellent scalability. It is worth mentioning that the performance of LLaVE-0.5B is already comparable to VLM2Vec (Phi-3.5-V-4B), while LLaVE-2B achieves an overall score of 65.2, surpassing the previously pretrained MMRet-7B that utilizes an additional 27M image-text pairs.

4.3 Ablation Study

Table 3 provides an ablation study analyzing the impact of various design choices on the performance of LLaVE-2B across IND, OOD, and overall metrics. The baseline model (ID 1) uses the standard InfoNCE with up to 50K samples per training dataset for balanced fine-tuning.

Freezing the image encoder helps generalize to out-of-distribution datasets. Freezing the image encoder ($1 \rightarrow 2$) can significantly improve performance on unseen datasets at the cost of a slight decrease in in-distribution performance. This is likely because the original vision encoder possesses stronger generalization capabilities, and instruction fine-tuning on smaller datasets may affect this ability. However, freezing the projector ($1 \rightarrow 3$) harms the model's performance because transforming the LMM into an embedding model requires re-adaptation.

ID	Model	IND	OOD	Overall
0	Previous SOTA (MMRet)	68.0	59.1	64.1
1	Aquila-VL-2B + InfoNCE	60.6	56.4	58.7
2	1 + Freeze image encoder	60.5 (-0.1)	58.3 (+2.1)	59.5 (+0.8)
3	2 + Freeze projector	60.3 (-0.2)	57.0 (-1.3)	58.8 (-0.7)
4	2 + Less training data (Each training dataset samples up to 20K data)	60.4 (-0.1)	56.4 (-1.9)	58.6 (-0.9)
5	2 + More training data (Each training dataset samples up to 100K data)	61.2 (+0.7)	57.4 (-0.9)	59.5 (+0.0)
6	2 + Cross-device negative sample gathering	68.6 (+8.1)	58.4 (+0.1)	64.0 (+4.5)
7	6 + Focal-InfoNCE loss (Hou and Li, 2023)	67.9 (-0.7)	59.5 (+1.1)	64.2 (+0.2)
8	6 + Hardness-weighted contrastive learning (LLaVE-2B)	69.4 (+0.8)	59.8 (+1.4)	65.1 (+1.1)

Table 3: Ablation results on MMEB. The model with ID 1 is configured to use the standard InfoNCE for full fine-tuning, with each training dataset sampling up to 50K examples, so as to ensure balance across different datasets. The numbers in parentheses indicate the impact of the changes on performance compared to the model corresponding to the previously selected ID.

When the data is sufficient, a balanced distribution of various data is more important than simply having more data. Table 3 (2 $\rightarrow$ 4) demonstrates that with limited data, the model's generalization ability is constrained. When the data sampling limit is increased to 100K (2 $\rightarrow$ 5), certain training datasets expand further, while others remain unchanged due to limited availability. The results show that although this approach improves in-distribution performance, it negatively impacts generalization, highlighting the importance of maintaining a balanced distribution across different meta-tasks.

The number of negatives is crucial for training LMM-based embedding models. By analyzing the performance of the model (ID 6), we observe that the introduction of the cross-device negative sample gathering strategy leads to substantial gains in IND (+8.1) with negligible impact of OOD (+0.1), resulting in a notable overall improvement of +4.5, which highlights the importance of diverse negative samples.

Hardness-weighted contrastive learning can further enhance the performance of powerful models on both in-distribution and out-of-distribution datasets. Although the model (ID 6) already achieves near-SOTA performance, hardness-weighted contrastive learning further enhances its effectiveness, particularly with a 1.4-point improvement on out-of-distribution datasets, demonstrating its complementarity. Besides, we also compare the Focal-InfoNCE loss (ID 7), which also weights positive and negative pairs. Although it slightly improves performance on the out-of-distribution dataset, it reduces performance on the in-distribution dataset. As shown in Figure 1 and Table 1, our framework significantly increases the similarity gap between positive and negative pairs, thereby improving the model's discriminative capability.

Model	MSR-VTT			MSVD		
	R@1	R@5	R@10	R@1	R@5	R@10
Zero-shot (finetuned with text-video data)						
InternVideo	40.0	65.3	74.1	43.4	69.9	79.1
ViCLIP	42.4	-	-	49.1	-	-
UMT-L	42.6	64.4	73.1	49.9	77.7	85.3
InternVideo2-6B	55.9	78.3	85.1	59.3	84.4	89.6
Zero-shot (finetuned only with text-image data)						
VLM2Vec	43.5	69.3	78.9	49.5	77.5	85.7
LamRA	44.7	68.6	78.6	52.4	79.8	87.0
LLaVE-7B	**46.8**	**71.1**	**80.0**	**52.9**	**80.1**	**87.0**

Table 4: Results of zero-shot text-to-video retrieval. The gray font indicates that the model is trained using contrastive learning on tens of millions of text-video data.

4.4 Zero-shot Video Retrieval

We also evaluate LLaVE on the widely used text-video retrieval datasets: MSR-VTT (Xu et al., 2016) and MSVD (Chen and Dolan, 2011), to explore its generalization capability. There are two types of comparative models considered. The first type includes models trained on tens of millions of video-text data, such as InternVideo (Wang et al., 2022), ViCLIP (Wang et al., 2024a), UMT-L (Li et al., 2023b), and InternVideo2-6B (Wang et al., 2024b). The second type is completely zero-shot, trained only on text-image data using contrastive learning and directly evaluated on text-video data. Particularly, we focus on the two strongest models, VLM2Vec (LLaVA-OV-7B) and LamRA (Liu

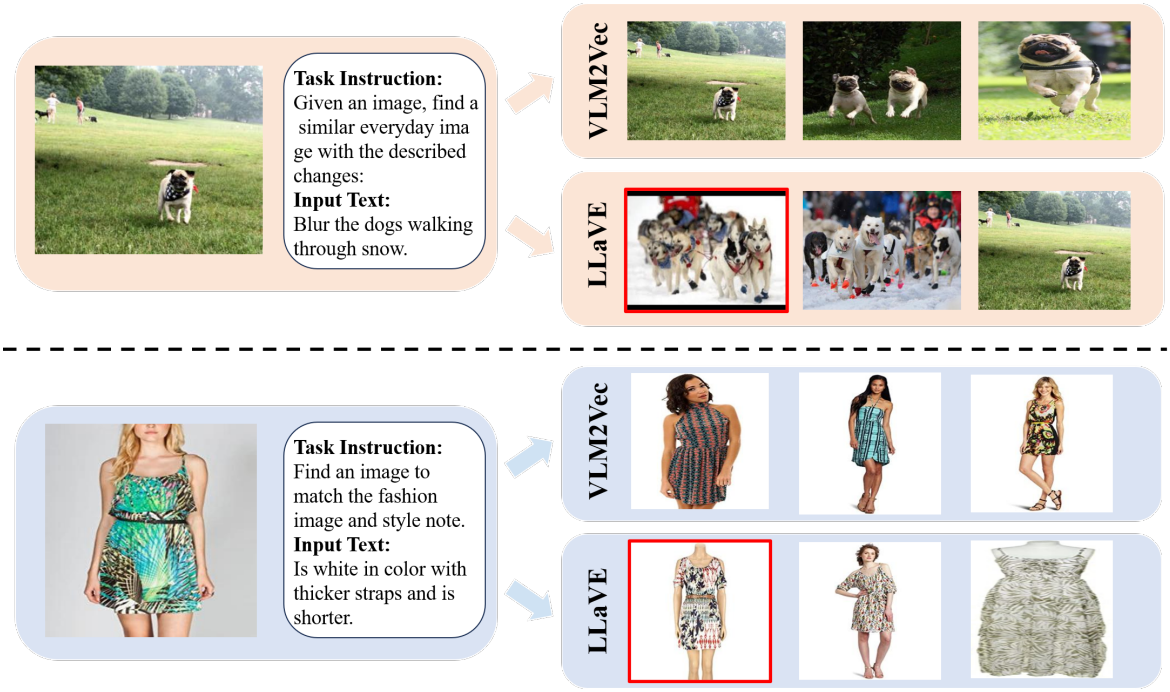

Figure 4: Qualitative evaluation comparing LLaVE and VLM2Vec. Retrievals consistent with the ground truth are highlighted with red borders. From left to right, the images represent the top-1 to top-3 retrieval results.

et al., 2024b), which is based on Qwen2-VL-7B and consists of two 7B models: LamRA-Ret and LamRA-Rank. LamRA-Ret retrieves the top-K candidates, while LamRA-Rank further re-ranks these retrieved candidates.

To enable video embedding, we set the maximum number of sampled frames to 32, expand the total input length of the model to 8192, and reduce the visual features of the video by 4 times through bilinear interpolation. It is observed that, compared to LamRA, LLaVE-7B requires only a single model and shows consistent improvements across all metrics. Notably, on MSR-VTT, the R@1, R@5, and R@10 scores increase by 2.1, 2.5, and 1.4, respectively. Moreover, although LLaVE-7B does not utilize text-video data for contrastive training, its performance still surpasses most video retrieval models except for InternVideo2-6B, which are trained on tens of millions of video-text pairs. These results demonstrate that LLaVE-7B has strong potential for transferring to other embedding tasks.

4.5 Qualitative Evaluation

In Figure 4, we present the qualitative evaluation results of LLaVE and the strongest baseline model, VLM2Vec (LLaVA-OV-7B). As shown in the upper part of the figure, LLaVE successfully identifies and retrieves the modified target images that meet the specified requirement ("*dogs walking through snow*") in the Top-2 retrieval, while VLM2Vec only retrieves images similar to the original image. Similarly, in the lower part of the figure, LLaVE fulfills the requirement of "*white in color and is shorter*" in the Top-3 retrieval. These examples demonstrate that our framework effectively facilitates the model to capture complex intents in challenging samples and enhances the discriminability of hard samples.

5 Related Work

Multimodal Embeddings. As a significant research direction, multimodal embeddings aim to integrate the information from multiple modalities (e.g., vision and language) into a shared representation space, which enables seamless understanding across modalities. Early research primarily focuses on leveraging dual-stream architectures to separately encode texts and images. For instance, CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), BLIP (Li et al., 2022), and SigLIP (Zhai et al., 2023) all adopt dual-encoder frameworks, learning universal representations from large-scale weakly supervised image-text pairs through contrastive learning. To learn the more universal multimodal representations, UniIR (Wei et al., 2024) proposes two fusion mechanisms to combine the visual and textual representations generated by the

dual-encoder model. Although these models have achieved impressive results, they still face challenges in handling tasks such as interleaved image-text retrieval (Wu et al., 2021a) and instruction-following multimodal retrieval.

LMM-based Multimodal Embeddings. To address the aforementioned issue, E5-V (Jiang et al., 2024a) and VLM2Vec (Jiang et al., 2024b) transform LMM into multimodal embedding models through contrastive learning, fully leveraging LMM's powerful multimodal understanding capability and its inherent advantage in handling interleaved text-image input. Recently, a few concurrent studies (Zhang et al., 2025; Gu et al., 2025; Thirukovalluru et al., 2025) further explored the application of LMMs in multimodal embeddings. For example, LamRA (Liu et al., 2024b) adopts the retrieval model to select the top-K candidates, which are then scored by the reranking models. Finally, the scores from the retrieval and reranking models are combined using a weighted sum to produce the final score for retrieval. MMRet (Zhou et al., 2024) creates a large-scale multimodal instruction retrieval dataset called MegaPairs. By pretraining on this dataset, MMRet achieves SOTA results on MMEB.

Contrastive Learning. Contrastive learning enables models to learn effective representations by distinguishing between positive and negative samples, and it has been widely applied across various domains (Kipf et al., 2020; Wu et al., 2021b; Lan et al., 2023; Yin et al., 2023; Zhang et al., 2024c). Negative samples play a crucial role in contrastive learning, Chen et al. (2020) show that incorporating more negative samples can enhance the performance of contrastive learning. Awasthi et al. (2022) further explore the impact of the number of negative samples from both theoretical and empirical perspectives. Moreover, Cai et al. (2020) demonstrate that hard negative samples are both necessary and sufficient to learn more discriminative representation, and Robinson et al. (2021) propose a hard negative sampling strategy where the user can control the hardness. The study most similar to ours is Focal-InfoNCE (Hou and Li, 2023), which weighs both positive and negative pairs based on their query-target similarities. Specifically, it uses a fixed threshold to determine the hardness of negative pairs, increasing the weight if the similarity exceeds the threshold and decreasing it otherwise. Unlike this work, our hardness-weighted contrastive learning introduces a reward model to dynamically estimate the hardness of negative pairs and applies weighting only to the negative pairs based on the estimated hardness. Notably, the reward model can be decoupled from the policy model.

6 Conclusion

In this paper, we conduct a preliminary study to find that LMM-based embedding models trained with the standard InfoNCE loss face significant challenges in handling hard negative pairs. To address this issue, we propose a simple yet effective framework that includes the hardness-weighted contrastive learning and the cross-device negative sample gathering strategy to enhance the model's learning of negative pairs with varying difficulty levels. This framework significantly improves the model's capacity to distinguish between positive and negative pairs. Experimental results and in-depth analyses validate the effectiveness of the proposed framework. In the future, we plan to collect and construct a universal multimodal embedding benchmark for video-text retrieval, aiming to investigate the more universal multimodal embedding models. We will open-source all models and code, hoping to inspire further research in this field.

Limitations

LLaVE is trained only on embedding datasets that contain arbitrary combinations of text and image modalities. Although it can generalize to embedding tasks that include the video modality in a zero-shot manner, there is still significant room for improvement. Constructing a multimodal embedding benchmark that incorporates the video modality will be a crucial direction for training more generalizable embedding models.

Acknowledgments

The project was supported by National Key R&D Program of China (No. 2022ZD0160501), Natural Science Foundation of Fujian Province of China (No. 2024J011001), and the Public Technology Service Platform Project of Xiamen (No.3502Z20231043). We also thank the reviewers for their insightful comments.

References

Pranjal Awasthi, Nishanth Dikkala, and Pritish Kamath. 2022. Do more negative samples necessarily hurt in contrastive learning? In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 1101–1116. PMLR.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tiffany Tianhui Cai, Jonathan Frankle, David J. Schwab, and Ari S. Morcos. 2020. Are all negatives created equal in contrastive instance discrimination? *CoRR*, abs/2010.06682.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 190–200. The Association for Computer Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2818–2829. IEEE.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, Zhenchong Hu, Bo-Wen Zhang, Jijie Li, Dong Liang, Yingli Zhao, Yulong Ao, Yaoqi Liu, Fangxiang Feng, and Guang Liu. 2024. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *CoRR*, abs/2410.18558.

Tiancheng Gu, Kaicheng Yang, Ziyong Feng, Xingjun Wang, Yanzhao Zhang, Dingkun Long, Yingda Chen, Weidong Cai, and Jiankang Deng. 2025. Breaking the modality barrier: Universal embedding learning with multimodal llms. *CoRR*, abs/2504.17432.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 1735–1742. IEEE Computer Society.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7514–7528. Association for Computational Linguistics.

Pengyue Hou and Xingyu Li. 2023. Improving contrastive learning of sentence embeddings with focal infonce. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 4757–4762. Association for Computational Linguistics.

Qingguo Hu, Ante Wang, Jia Song, Delai Qiu, Qingsong Liu, and Jinsong Su. 2025. Boosting visual knowledge-intensive training for lvlms through causality-driven visual object completion. In *IJCAI 2025*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024a. E5-V: universal embeddings with multimodal large language models. *CoRR*, abs/2407.12580.

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

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798. ACL.

Thomas N. Kipf, Elise van der Pol, and Max Welling. 2020. Contrastive learning of structured world models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhibin Lan, Wei Li, Jinsong Su, Xinyan Xiao, Jiachen Liu, Wenhao Wu, and Yajuan Lyu. 2023. Factgen: Faithful text generation by factuality-aware pre-training and contrastive ranking fine-tuning. *J. Artif. Intell. Res.*, 76:1281–1303.

Zhibin Lan, Liqiang Niu, Fandong Meng, Wenbo Li, Jie Zhou, and Jinsong Su. 2025. AVG-LLaVA: An efficient large multimodal model with adaptive visual granularity. In *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria. Association for Computational Linguistics.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *CoRR*, abs/2408.03326.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. 2023b. Unmasked teacher: Towards training-efficient video foundation models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19891–19903. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yikun Liu, Pingan Chen, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. 2024b. Lamra: Large multimodal model as your advanced retrieval assistant. *Preprint*, arXiv:2412.01720.

OpenAI. 2023. Gpt-4v(ision) system card.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3505–3506. ACM.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive learning with hard negative samples. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18990–18998. AAAI Press.

Raghuveer Thirukovalluru, Rui Meng, Ye Liu, Karthikeyan K, Mingyi Su, Ping Nie, Semih Yavuz, Yingbo Zhou, Wenhu Chen, and Bhuwan Dhingra. 2025. Breaking the batch barrier (B3) of contrastive learning via smart batch mining. *CoRR*, abs/2505.11293.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. 2024a. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. 2024b. Internvideo2: Scaling foundation models for multimodal video understanding. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXV*, volume 15143 of *Lecture Notes in Computer Science*, pages 396–416. Springer.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. 2022. Internvideo: General video foundation models via generative and discriminative learning. *CoRR*, abs/2212.03191.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXVII*, volume 15145 of *Lecture Notes in Computer Science*, pages 387–404. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogério Feris. 2021a. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11307–11317. Computer Vision Foundation / IEEE.

Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021b. Self-supervised graph learning for recommendation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 726–735. ACM.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3485–3492. IEEE Computer Society.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society.

Yongjing Yin, Jiali Zeng, Jinsong Su, Chulun Zhou, Fandong Meng, Jie Zhou, Degen Huang, and Jiebo Luo. 2023. Multi-modal graph contrastive encoding for neural machine translation. *Artif. Intell.*, 323:103986.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. 2024a. Magiclens: Self-supervised image retrieval with open-ended instructions. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. 2024b. Magiclens: Self-supervised image retrieval with open-ended instructions. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Liang Zhang, Zhen Yang, Biao Fu, Ziyao Lu, Liangying Shao, Shiyu Liu, Fandong Meng, Jie Zhou, Xiaoli Wang, and Jinsong Su. 2024c. Multi-level cross-modal alignment for speech relation extraction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 11975–11986. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9274–9285. Computer Vision Foundation / IEEE.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Do Xuan Long, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. 2023. Retrieving multimodal information for augmented generation: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 4736–4756. Association for Computational Linguistics.

Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. 2024. Megapairs: Massive data

synthesis for universal multimodal retrieval. *Preprint*, arXiv:2412.14475.

Hyperparameter	LLaVE-0.5B	LLaVE-2B	LLaVE-7B
#Data	662K	662K	662K
Batch size	256	256	256
lr	1e-5	1e-5	5e-6
lr schedule		cosine decay	
lr warmup ratio		0.03	
Weight decay		0	
Epoch		1	
Optimizer		AdamW	
DeepSpeed stage		3	
Precision	Bf16 and TF32	Bf16 and TF32	Bf16
GPU	8 × A100	8 × A100	16 × 910B
Training cost (#Hours)	12	17	33

Table 5: Training details of LLaVE.

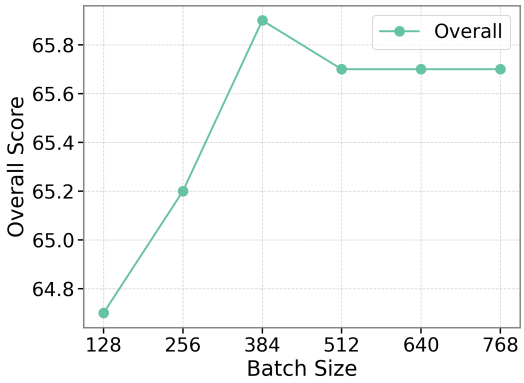

Figure 5: The impact of batch size on overall model performance when training for one epoch.

A Appendix

A.1 Training Details

We present the training details of LLaVE in Table 5. To save memory and accelerate training, we adopt Gradient Checkpointing and Flash Attention (Dao et al., 2022) techniques. Due to resource constraints, LLaVE-7B is trained on 910B GPUs, and the training time will significantly decrease when conducted on A100 GPUs. In addition, using more GPUs to adopt a larger batch size or higher resolution will improve the model's performance.

A.2 The Impact of Different Batch Sizes

As shown in Figure 5, we also explore the impact of different batch sizes on the overall performance of the model when training for one epoch. It can be observed that as the batch size increases, the overall performance of the model improves. However, since the total number of trained epochs does not increase, larger batch sizes do not achieve further performance improvements. This is similar to the trend observed by Zhai et al. (2023), as large batch sizes need a sufficiently long schedule to ramp up. Besides, we find that the similarity distribution does not differ significantly across different batch

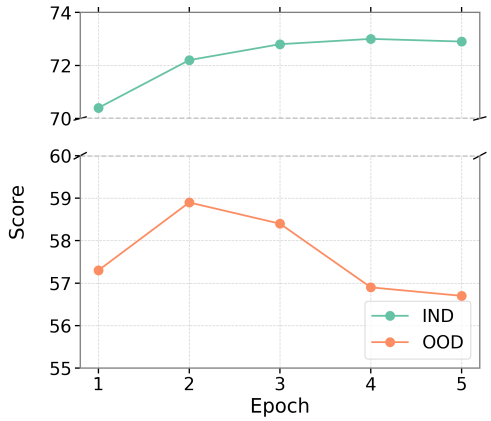

Figure 6: Impact of training epochs on IND and OOD performance.

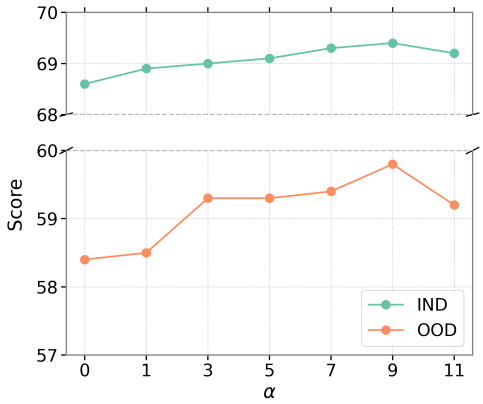

Figure 7: Influence of the α on model performance, measured on IND and OOD datasets, respectively.

sizes, which means that the similarity distribution remains relatively stable using different batch sizes.

A.3 The Impact of Training Epochs

To further investigate the impact of training duration, we fine-tune LLaVE-2B for up to five epochs. The results in Figure 6 show that increasing the number of epochs initially enhances in-domain performance and overall accuracy. However, after three epochs, overfitting appears, which degrades out-of-domain performance and limits further overall improvements. For the sake of training efficiency and generalization, we therefore train for one epoch throughout this paper.

A.4 Hyperparameter Analysis

We experimentally explore the influence of the hyperparameter α on model performance. Figure 7 shows that increasing α positively influences both IND and OOD performance, stopping at a certain point. This indicates that appropriately weighting hard negative samples helps the model learn effectively. Moreover, it can be observed that the model's performance is robust to the α parameter, consistently outperforming the results obtained without hardness-weighted contrastive learning (i.e., when α=0).

A.5 Generalization of LLaVE Across LMMs

To further validate the generality of our proposed methods, we extend LLaVE to other widely used LMMs and evaluated it on MMEB. Specifically, we apply our methods to Qwen2-VL models of different scales (2B and 7B) and employ the same training settings as VLM2Vec, including identical batch sizes and training steps. To further validate the generalizability of our proposed methods, we extend LLaVE to other widely used LMMs and evaluate its performance on MMEB. Specifically, we apply our methods to Qwen2-VL models of varying scales (2B and 7B) and employ the same training settings as VLM2Vec. As shown in Table 5, when using the same LMM as the backbone, LLaVE still significantly outperforms VLM2Vec. In particular, LLaVE-2B achieves an improvement of 4.7 over VLM2Vec-2B, and LLaVE-7B achieves an improvement of 3.4 over VLM2Vec-7B. These results fully demonstrate the strong generalization capability of our method.

A.6 Detailed results on MMEB

We present the detailed results of each model on various datasets of MMEB in Table 7.

Model	Per Meta-Task Score				Average Score		
	Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# Datasets	10	10	12	4	20	16	36
VLM2Vec (Qwen2-VL-2B)	59.0	49.4	65.4	73.4	66.0	52.6	60.1
VLM2Vec (Qwen2-VL-7B)	62.6	57.8	69.9	81.7	72.2	57.8	65.8
LLaVE-2B (Qwen2-VL-2B)	64.3	58.5	66.5	76.9	71.1	57.1	64.8
LLaVE-7B (Qwen2-VL-7B)	**65.4**	**65.2**	**70.6**	**84.8**	**74.8**	**62.3**	**69.2**

Table 6: Comparison of LLaVE and VLM2Vec based on different LMMs.

	CLIP	OpenCLIP	SigLIP	BLIP2	MagicLens	E5-V	UniIR	VLM2Vec	MMRet	LLaVE
Classification (10 tasks)										
ImageNet-1K	55.8	63.5	45.4	10.3	48.0	9.6	58.3	67.8	58.8	77.1
N24News	34.7	38.6	13.9	36.0	33.7	23.4	42.5	76.3	71.3	82.1
HatefulMemes	51.1	51.7	47.2	49.6	49.0	49.7	56.4	65.8	53.7	74.3
VOC2007	50.7	52.4	64.3	52.1	51.6	49.9	66.2	88.9	85.0	90.3
SUN397	43.4	68.8	39.6	34.5	57.0	33.1	63.2	74.4	70.0	79.1
Place365	28.5	37.8	20.0	21.5	31.5	8.6	36.5	43.0	43.0	45.1
ImageNet-A	25.5	14.2	42.6	3.2	8.0	2.0	9.8	51.4	36.1	51.6
ImageNet-R	75.6	83.0	75.0	39.7	70.9	30.8	66.2	86.3	71.6	90.9
ObjectNet	43.4	51.4	40.3	20.6	31.6	7.5	32.2	59.5	55.8	46.2
Country-211	19.2	16.8	14.2	2.5	6.2	3.1	11.3	21.4	14.7	20.1
All Classification	42.8	47.8	40.3	27.0	38.8	21.8	44.3	63.5	56.0	65.7
VQA (10 tasks)										
OK-VQA	7.5	11.5	2.4	8.7	12.7	8.9	25.4	67.3	73.3	71.1
A-OKVQA	3.8	3.3	1.5	3.2	2.9	5.9	8.8	63.6	56.7	70.8
DocVQA	4.0	5.3	4.2	2.6	3.0	1.7	6.2	86.6	78.5	90.3
InfographicsVQA	4.6	4.6	2.7	2.0	5.9	2.3	4.6	51.9	39.3	53.5
ChartQA	1.4	1.5	3.0	0.5	0.9	2.4	1.6	54.9	41.7	62.2
Visual7W	4.0	2.6	1.2	1.3	2.5	5.8	14.5	48.7	49.5	55.8
ScienceQA	9.4	10.2	7.9	6.8	5.2	3.6	12.8	46.6	45.2	54.4
VizWiz	8.2	6.6	2.3	4.0	1.7	2.6	24.3	48.3	51.7	48.5
GQA	41.3	52.5	57.5	9.7	43.5	7.8	48.8	66.8	59.0	68.4
TextVQA	7.0	10.9	1.0	3.3	4.6	8.2	15.1	76.0	79.0	79.4
All VQA	9.1	10.9	8.4	4.2	8.3	4.9	16.2	61.1	57.4	65.4
Retrieval (12 tasks)										
VisDial	30.7	25.4	21.5	18.0	24.8	9.2	42.2	73.8	83.0	83.0
CIRR	12.6	15.4	15.1	9.8	39.1	6.1	51.3	50.3	61.4	54.5
VisualNews_t2i	78.9	74.0	51.0	48.1	50.7	13.5	74.3	69.7	74.2	76.6
VisualNews_i2t	79.6	78.0	52.4	13.5	21.1	8.1	76.8	72.3	78.1	81.2
MSCOCO_t2i	59.5	63.6	58.3	53.7	54.1	20.7	68.5	74.5	78.6	78.9
MSCOCO_i2t	57.7	62.1	55.0	20.3	40.0	14.0	72.1	71.7	72.4	74.7
NIGHTS	60.4	66.1	62.9	56.5	58.1	4.2	66.2	66.5	68.3	67.0
WebQA	67.5	62.1	58.1	55.4	43.0	17.7	89.6	87.5	90.2	90.4
FashionIQ	11.4	13.8	20.1	9.3	11.2	2.8	40.2	20.6	54.9	23.3
Wiki-SS-NQ	55.0	44.6	55.1	28.7	18.7	8.6	12.2	53.7	24.9	63.9
OVEN	41.1	45.0	56.0	39.5	1.6	5.9	69.4	67.0	87.5	68.0
EDIS	81.0	77.5	23.6	54.4	62.6	26.8	79.2	66.6	65.6	89.1
All Retrieval	53.0	52.3	31.6	33.9	35.4	11.5	61.8	64.5	69.9	70.9
Visual Grounding (4 tasks)										
MSCOCO	33.8	34.5	46.4	28.9	22.1	10.8	46.6	80.6	76.8	87.0
RefCOCO	56.9	54.2	70.8	47.4	22.8	11.9	67.8	92.3	89.8	95.4
RefCOCO-matching	61.3	68.3	50.8	59.5	35.6	38.9	62.9	85.3	90.6	92.8
Visual7W-pointing	55.1	56.3	70.1	52.0	23.4	14.3	71.3	91.0	77.0	92.5
All Visual Grounding	51.8	53.3	59.5	47.0	26.0	19.0	65.3	87.3	83.6	91.9
Final Score (36 tasks)										
All	37.8	39.7	34.8	25.2	27.8	13.3	44.7	65.8	64.1	70.3
All IND	37.1	39.3	32.3	25.3	31.0	14.9	47.1	61.0	59.1	64.4
All OOD	38.7	40.2	38.0	25.1	23.7	11.5	41.7	69.7	68.0	75.0

Table 7: The detailed results of the baselines and LLaVE on MMEB. The out-of-distribution datasets are highlighted with a yellow background in the table. We only include the best version of each series of models in the table, such as LLaVE-7B and VLM2Vec (LLaVA-OV-7B).