TSEmbed: Unlocking Task Scaling in Universal Multimodal Embeddings

Yebo Wu[1†], Feng Liu[2†], Ziwei Xie[2], Changwang Zhang[2], Jun Wang[2⋆], and Li Li[1⋆]

[1] State Key Laboratory of IOTSC, University of Macau, China
[2] OPPO Research Institute, China
{yc37926,llili}@um.edu.mo, {liufeng4hit,junwang.lu}@gmail.com
xieziwei@oppo.com, changwangzhang@foxmail.com

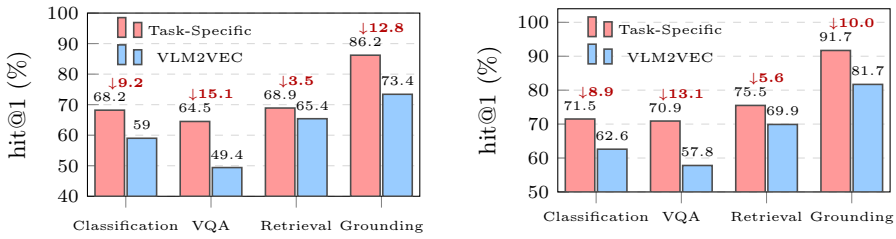
Abstract. Despite the exceptional reasoning capabilities of Multimodal Large Language Models (MLLMs), their adaptation into universal embedding models is significantly impeded by task conflict. To address this, we propose `TSEmbed`, a universal multimodal embedding framework that synergizes Mixture-of-Experts (MoE) with Low-Rank Adaptation (LoRA) to explicitly disentangle conflicting task objectives. Moreover, we introduce Expert-Aware Negative Sampling (EANS), a novel strategy that leverages expert routing distributions as an intrinsic proxy for semantic similarity. By dynamically prioritizing informative hard negatives that share expert activation patterns with the query, EANS effectively sharpens the model's discriminative power and refines embedding boundaries. To ensure training stability, we further devise a two-stage learning paradigm that solidifies expert specialization before optimizing representations via EANS. `TSEmbed` achieves state-of-the-art performance on both the Massive Multimodal Embedding Benchmark (MMEB) and real-world industrial production datasets, laying a foundation for task-level scaling in universal multimodal embeddings.

Keywords: Representation Learning · Multimodal Embedding · MLLMs

1 Introduction

Multimodal embedding models [15,50] project heterogeneous inputs into a shared semantic manifold, underpinning a wide range of downstream applications such as image-text retrieval [1,30,41], retrieval-augmented generation (RAG) [47,55], and multimodal recommendation [22, 46]. While foundational vision-language models such as CLIP [28] and SigLIP [33] have established robust cross-modal alignment, their dual-encoder architectures inherently process visual and textual streams in isolation. This late-fusion paradigm creates a structural barrier to deep cross-modal interaction, severely constraining their performance on complex tasks [36, 42, 43] that require fine-grained, inter-modal reasoning.

[†] Equal contribution.
[⋆] Corresponding author.



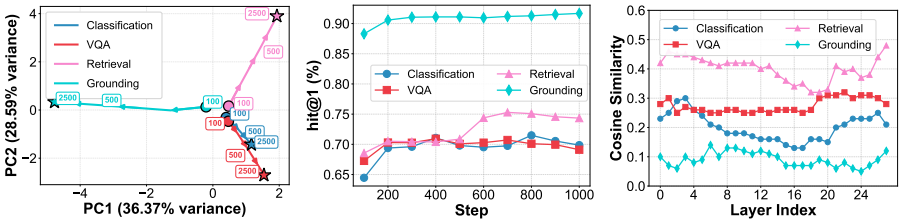
(a) Qwen2-VL-2B. (b) Qwen2-VL-7B.

Fig. 1: Impact of task conflict on model performance. Red annotations (↓) indicate the performance drop when switching from task-specific models to the unified VLM2VEC.

The rapid evolution of Multimodal Large Language Models (MLLMs), exemplified by GPT-4V [45] and Qwen-VL [34], has catalyzed a promising paradigm shift. Pioneering works such as VLM2VEC [16] repurpose these generative architectures as universal embedding models, harnessing their expansive world knowledge and compositional reasoning for multimodal representation learning. To further enhance their representational capacity, subsequent research has coalesced around three directions: (1) Data Synthesis [21, 52] to overcome the scarcity of high-quality supervised data; (2) Hard Negative Mining [19, 44] to refine contrastive boundaries; and (3) Reasoning Elicitation [6, 14] to extract deeper semantic logic via Chain-of-Thought or reinforcement learning. However, these strategies overlook a fundamental bottleneck: **task conflict**. Forcing diverse semantic objectives into a monolithic parameter space inevitably induces severe gradient interference, significantly degrading model performance.

To empirically quantify the impact of task conflict, we compare the performance of individual task-specific training against joint training (VLM2VEC) on the MMEB. As illustrated in Figure 1, the jointly trained model consistently underperforms its task-specific counterparts across all four meta-task categories. Notably, on VQA, the VLM2VEC suffers severe performance drops of 15.1% and 13.1% for the 2B and 7B models, respectively. This performance gap confirms that a shared parameter space struggles to satisfy competing optimization objectives, fundamentally constraining the representational capability of MLLMs.

In this paper, we propose `TSEmbed`, a universal multimodal embedding framework that synergizes MoE with LoRA to resolve task conflicts through conditional computation. Unlike existing methods that force compromises among divergent objectives, `TSEmbed` partitions the optimization landscape into semantically decoupled subspaces. By dynamically routing queries to specialized experts, it transforms destructive conflict into collaborative specialization. Furthermore, we introduce Expert-Aware Negative Sampling (EANS), which leverages the MoE routing distribution as an intrinsic proxy to identify hard negatives, thereby enhancing the model's discriminative power. To harness the full potential of this synergy, we devise a progressive two-stage learning paradigm that



(a) Spatial Conflict. (b) Temporal Mismatch. (c) Ecological Bias.

Fig. 2: A Multidimensional Anatomy of Task Conflict in Monolithic Adapters. (a) Divergent optimization trajectories of isolated task-specific adapters. (b) Heterogeneous convergence dynamics of individual tasks during training. (c) Layer-wise cosine similarity between the jointly trained adapter and its isolated counterparts.

stabilizes expert specialization before deploying EANS to precisely refine the embedding boundaries. Our main contributions are summarized as follows:

- We systematically analyze task conflict in universal multimodal embeddings across Spatial, Temporal, and Ecological dimensions, exposing the limitations of monolithic adapters in satisfying divergent semantic objectives.
- We propose `TSEmbed`, a novel architecture that synergizes MoE with LoRA to mitigate task conflicts through conditional computation. This design explicitly decouples the optimization landscape, transforming destructive gradient interference into collaborative specialization, thereby establishing a foundation for task-level scaling in universal multimodal embeddings.
- We introduce Expert-Aware Negative Sampling (EANS), an efficient strategy that leverages intrinsic MoE routing distributions as a proxy for semantic similarity to dynamically up-weight hard negatives. To ensure the reliability of these routing signals, we further devise a progressive two-stage learning paradigm that stabilizes expert specialization prior to EANS refinement, effectively sharpening embedding boundaries.
- We conduct extensive evaluations on both the MMEB and real-world production datasets. The results demonstrate that `TSEmbed` consistently achieves state-of-the-art performance across both public benchmarks and practical applications, yielding a notable **21.87%** gain in advertising scenarios.

2 Anatomy of Task Conflict in Multimodal Embeddings

To mechanistically understand task conflicts in universal multimodal embeddings, we empirically analyze the MMEB using Qwen2-VL-7B [34], deconstructing the conflict into three dimensions: Spatial, Temporal, and Ecological.

2.1 Spatial Dimension: Divergent Gradient Trajectories

To uncover the geometric nature of task conflict, we first examine the optimization landscape of isolated tasks. Specifically, we independently train four distinct

LoRA adapters, each dedicated to a specific meta-task category. To map the divergence of their optimization paths, we extract model checkpoints at 100-step intervals and project their high-dimensional parameter states into a 2D subspace via PCA [9], which captures a substantial 64.96% of the total variance.

As illustrated in Figure 2(a), although all task-specific adapters originate from a shared initialization, their training trajectories rapidly splinter into opposing quadrants. Visual Grounding evolves along the negative PC1 axis, nearly orthogonal to Retrieval's sharp divergence along the positive PC2 axis. Concurrently, Classification and VQA traverse toward the bottom-right quadrant, in stark geometric opposition to Retrieval. These highly divergent trajectories confirm that the optimal parameter configurations for distinct tasks reside in entirely disparate regions of the solution space.

Takeaway I: *The heterogeneous optimization trajectories reveal that the optimal solutions for distinct tasks reside in disjoint parameter spaces.*

2.2 Temporal Dimension: Heterogeneous Convergence Dynamics

Beyond spatial conflicts, we investigate the temporal optimization dynamics of isolated tasks. As illustrated in Figure 2(b), the four meta-tasks exhibit a pronounced temporal misalignment, bifurcating into two divergent patterns. Early-converging tasks, such as Visual Grounding and VQA, exhibit rapid initial gains but experience severe degradation as training progresses. Visual Grounding, for instance, swiftly captures coarse spatial localization, reaching 88.30% by step 100 and plateauing at 90.58% by step 200. Conversely, late-converging tasks like Retrieval and Classification demand sustained optimization horizons.

This temporal heterogeneity imposes a strict synchronization bottleneck on monolithic architectures. A single, shared learning schedule cannot simultaneously accommodate these divergent patterns: halting training prematurely to preserve VQA performance leaves Retrieval severely underfitted, while extending the optimization horizon causes early-converging tasks to overfit and degrade.

Takeaway II: *Distinct tasks converge at markedly different speeds, making a single shared adapter incapable of synchronized optimization.*

2.3 Ecological Dimension: Data Imbalance and Task Dominance

Finally, we examine the ecological dimension of task conflict stemming from inherent data imbalances across the training corpus. While a universal embedding model should ideally cultivate balanced representations for all semantic objectives, a monolithic adapter is inevitably hijacked by data-rich tasks. Figure 2(c) illustrates the layer-wise cosine similarity between the jointly trained adapter and its isolated, task-specific counterparts, exhibiting a profound ecological bias that pervades every layer. The joint representations maintain high alignment with the data-abundant Retrieval task (peaking near 0.48), yet leave the data-scarce Visual Grounding task severely underrepresented (hovering near 0.10).

This severe representational gap demonstrates that dominant tasks unavoidably co-opt the shared parameter space throughout the entire feature extraction

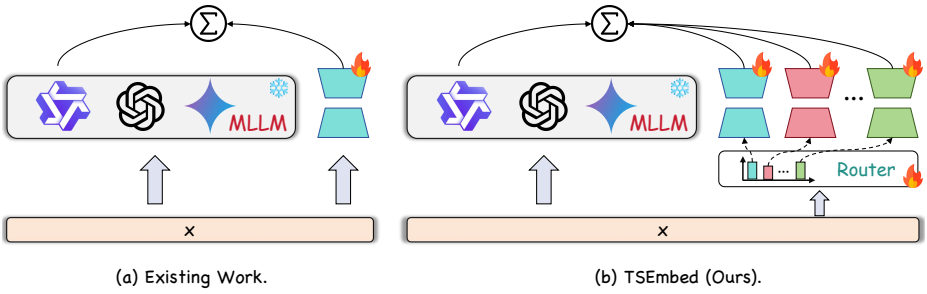


Fig. 3: Overview of existing work and our proposed TSEmbed.

hierarchy. Monolithic adapters inherently fail to equitably balance heterogeneous objectives. Consequently, the model's limited capacity is monopolized by data-abundant tasks, marginalizing minority objectives and thereby undermining the core premise of truly scalable and universal multimodal representation learning. ***Takeaway III:*** *Joint training facilitates data-abundant tasks to hijack the optimization process, suppressing the representation learning of data-scarce ones.*

3 TSEmbed: Task Scaling Multimodal Embeddings

3.1 Preliminary

MLLM-based Embedding. Given an input $\mathbf{x}$, the representation is obtained by extracting the hidden state of the last token at the final layer:

$$\mathbf{h} = f_\theta(\mathbf{x})_{[\texttt{EOS}]} \in \mathbb{R}^d, \tag{1}$$

where f_θ denotes the MLLM backbone and [EOS] indexes the final token. For a query $\mathbf{q}$ and a positive target $\mathbf{k}^+$, the model is fine-tuned via the InfoNCE [27] contrastive objective over a batch of M negatives $\{\mathbf{k}_i^-\}_{i=1}^M$:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(s^+/\tau)}{\exp(s^+/\tau) + \sum_{i=1}^M \exp(s_i^-/\tau)}, \tag{2}$$

where $s^+ = \mathbf{h}_q \cdot \mathbf{h}_{k^+}$ and $s_i^- = \mathbf{h}_q \cdot \mathbf{h}_{k_i^-}$ denote the query-positive and query-negative similarities, respectively, and τ is a temperature hyperparameter.

Low-Rank Adaptation (LoRA). To efficiently adapt the backbone, LoRA [12, 38] freezes the pre-trained weights $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$ and introduces a trainable low-rank residual parameterized by $\mathbf{A} \in \mathbb{R}^{r\times k}$ and $\mathbf{B} \in \mathbb{R}^{d\times r}$ ($r \ll \min(d, k)$). The adapted forward pass is defined as: $\mathbf{h}' = \mathbf{W}_0\mathbf{x} + \mathbf{BAx}$. While parameter-efficient, standard LoRA applies a *uniform* transformation $\mathbf{BA}$ to all inputs regardless of their semantic type [39], fundamentally inducing severe task conflict.

3.2 Conflict Decoupling: MoE-LoRA

To resolve task conflicts caused by shared parameters, we introduce the MoE-LoRA for universal multimodal embeddings (Figure 3). By utilizing conditional computation to route queries to specialized experts [37], it decouples the semantic space and prevents mutual interference. Specifically, for each layer, the adapted forward pass is reformulated as:

$$\mathbf{h}' = \mathbf{W}_0\mathbf{x} + \sum_{i=1}^{N} g_i(\mathbf{x}) \cdot \mathbf{B}_i\mathbf{A}_i\mathbf{x}, \tag{3}$$

where N is the number of experts, and $g_i(\mathbf{x})$ denotes the routing weight for the i-th expert. Formally, the gating network employs a linear projection parameterized by a weight matrix $\mathbf{W}_g \in \mathbb{R}^{N\times d}$ to dynamically compute input-dependent weights for each expert. These weights are subsequently normalized via softmax:

$$\mathbf{g}(\mathbf{x}) = \text{Softmax}\left(\frac{\mathbf{W}_g\mathbf{x}}{\tau'}\right), \quad g_i(\mathbf{x}) = [\mathbf{g}(\mathbf{x})]_i, \tag{4}$$

where τ' controls the sharpness of the routing distribution, and $[\cdot]_i$ indexes the i-th component of the resulting weight vector.

3.3 Boundary Refinement: Expert-Aware Negative Sampling

Existing hard negative mining methods [44] typically rely on computationally prohibitive metrics or auxiliary models. In contrast, we observe that the MoE routing mechanism naturally encodes task-level semantic topology, providing a highly effective proxy for identifying hard negatives.

Routing Distribution as Semantic Proxy. For each sample processed by MoE-LoRA, we extract its routing distributions across all L layers. Given that the MoE-LoRA adaptation is applied to G distinct projection matrices (e.g., `Query`, `Key`, and `Value`) within each transformer block, we concatenate the routing probabilities from all G matrices at layer ℓ to form a joint layer-wise distribution $\mathbf{g}^{(\ell)} \in [0,1]^{G\times N}$. We then aggregate these layer-wise distributions across all L layers to construct a comprehensive global routing signature:

$$\mathbf{r} = [\mathbf{g}^{(1)}, \mathbf{g}^{(2)}, \ldots, \mathbf{g}^{(L)}] \in [0,1]^{L\times G\times N}. \tag{5}$$

The flattened vector $\text{vec}(\mathbf{r})$ comprehensively encapsulates the sample's activation trajectory throughout the network depth, explicitly mapping its task-specific expert utilization. Given a query $\mathbf{q}$ and a negative sample $\mathbf{k}_i^-$ with their routing signatures $\mathbf{r}_q$ and $\mathbf{r}_i$, we quantify their semantic divergence as:

$$d(\mathbf{r}_q, \mathbf{r}_i) = \frac{\|\text{vec}(\mathbf{r}_q) - \text{vec}(\mathbf{r}_i)\|_1}{L \cdot G \cdot N}. \tag{6}$$

Normalization by the total dimension $L{\cdot}G{\cdot}N$ ensures the distance metric is scale-invariant and strictly bounded. A smaller distance indicates highly overlapping

expert activation patterns, strongly suggesting that the samples share task-level semantics and thereby serve as exceptionally informative hard negatives.

Exponential Decay Weighting. To translate the semantic distance into an effective learning signal, we design an exponential decay weighting function, which emphasizes hard negatives while smoothly suppressing trivial ones:

$$w_i = w_{\min} + (w_{\max} - w_{\min}) \cdot \exp\left(-\frac{d(\mathbf{r}_q, \mathbf{r}_i)}{\sigma}\right), \tag{7}$$

where $w_{\min}$ and $w_{\max}$ establish the lower and upper bounds of the penalty, respectively, and σ serves as a scale parameter governing the decay sensitivity. This exponential formulation exhibits superior sensitivity to minute distance variations within the high-similarity regime. Specifically, as the semantic distance approaches zero ($d \to 0$), the weight converges to $w_{\max}$, imposing a stringent penalty on hard negatives. Conversely, for semantically distant samples ($d \gg 0$), the weight rapidly decays towards $w_{\min}$, safely marginalizing trivial negatives.

Weight Normalization for Gradient Consistency. To ensure that the dynamic weighting mechanism does not inadvertently distort the global gradient scale relative to the standard InfoNCE loss, we normalize the raw weights:

$$\tilde{w}_i = w_i \cdot \frac{M}{\sum_{j=1}^{M} w_j}. \tag{8}$$

This ensures the aggregate negative contribution in the loss denominator remains strictly invariant. As a result, EANS purely redistributes the gradient budget.

EANS Loss. The final Expert-Aware Negative Sampling loss is:

$$\mathcal{L}_{\text{EANS}} = -\log \frac{\exp(s^+/\tau)}{\exp(s^+/\tau) + \sum_{i=1}^{M} \tilde{w}_i \cdot \exp(s_i^-/\tau)}. \tag{9}$$

EANS effectively compels hard negatives with high $\tilde{w}_i$ to exert a greater influence, thereby sharpening the model's representational capability.

3.4 Two-Stage Learning Paradigm

The efficacy of EANS hinges on a critical prerequisite: routing distributions must serve as reliable semantic proxies. Since randomly initialized routers are inherently stochastic, applying EANS prematurely would inject spurious gradients and destabilize training. Therefore, we design a progressive two-stage paradigm.

Stage 1: Expert Warm-up ($t < T_{\text{warmup}}$). Optimized exclusively via standard InfoNCE, this phase allows MoE-LoRA to naturally disentangle the heterogeneous semantic space. Guided purely by data distributions, experts autonomously carve out distinct functional niches to resolve task conflicts.

Stage 2: EANS Refinement ($t \geq T_{\text{warmup}}$). Upon the stabilization of the routing topology, we engage the dynamically weighted EANS loss. Leveraging the reliable routing trajectories as semantic cues, EANS imposes targeted penalties on hard negatives, effectively sharpening the embedding boundaries.

Overall Objective. Formally, we optimize the piecewise combined loss:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{\text{InfoNCE}} & \text{if } t < T_{\text{warmup}}, \\ \mathcal{L}_{\text{EANS}} & \text{if } t \geq T_{\text{warmup}}. \end{cases} \tag{10}$$

4 Experiments

4.1 Experimental Setup

Datasets and Models. We evaluate `TSEmbed` on the MMEB [16] alongside proprietary real-world industrial datasets to rigorously assess both in-domain alignment and zero-shot generalization capabilities. Qwen2-VL [34] is employed as the backbone, with evaluations scaled across its 2B and 7B variants.
Implementation Details. We implement `TSEmbed` in PyTorch, executing all experiments on 8 NVIDIA A800 GPUs. During visual preprocessing, images are resized with a minimum resolution of 401,408 pixels. We extract the sequence embedding from the final hidden state corresponding to the `<EOS>` token. To facilitate large-batch contrastive learning within hardware memory constraints, we integrate Gradient Caching [8], configuring a sub-batch chunk size of 2.
Hyperparameter Settings. We optimize the model using AdamW with a learning rate of 5×10^{-5} and a linear decay scheduler [40]. The training process spans 2,200 steps with a global batch size of 1,024 (128 per GPU). Moreover, our MoE layers feature $N = 4$ experts, with each LoRA configured with a rank of 16 and a scaling factor of 64. The exponential decay weighting function is parameterized by $w_{\min} = 0.1$, $w_{\max} = 10.0$, and a decay sensitivity $\sigma = 0.002$. T_{warmup} is set to 600 and 1,200 steps for the 2B and 7B models, respectively.

4.2 Baselines

We compare `TSEmbed` against a wide range of multimodal embedding models:

- **Encoder-Only Models:** We include representative dual-encoder architectures such as CLIP [28], OpenCLIP [5], SigLIP [33], and BLIP-2 [20], as well as retrieval-specialized models like UniIR [35] and Magiclens [52].
- **MLLM-Based Embedding Models:** We benchmark against state-of-the-art generative embedding models across various scales, including VLM2VEC [16], UniME-V2 [10], LLaVE [19], B3 [32], and QQMM-embed [44].
- **External Data-Augmented Models:** We also include the performance of the proprietary Seed-1.6-embedding model as a commercial benchmark, alongside a suite of models augmented with massive external corpora beyond MMEB. These include Ops-MM-embedding-v1, GME [53], UNITE [18], CAFe [48], ColPali-v1.3 [7], LamRA [23], RzenEmbed [13], and Qwen3-VL-Embedding [21]. It is important to note that these models serve solely as high-resource references rather than direct competitors. To ensure a fair assessment of algorithmic contributions, our primary evaluation focuses on models trained exclusively on the MMEB dataset, thereby isolating the architectural efficacy of our proposed framework.

Table 1: Performance comparison of various multimodal embedding models on the MMEB [16]. **Bold** and underlined values denote the best and second-best results within each parameter scale, respectively. † indicates a link to the model's official homepage.

Model	Backbone	Model Size	Per Meta-Task Score				Average Score		
			Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
# of datasets →			10	10	12	4	20	16	36
Encoder-Only Models									
CLIP [28]	-	0.428B	42.8	9.1	53.0	51.8	37.1	38.7	37.8
BLIP-2 [20]	-	3.74B	27.0	4.2	33.9	47.0	25.3	25.1	25.2
SigLIP [33]	-	0.203B	40.3	8.4	31.6	59.5	32.3	38.0	34.8
OpenCLIP [5]	-	0.428B	47.8	10.9	52.3	53.3	39.3	40.2	39.7
UniIR (BLIP_FF) [35]	-	0.247B	42.1	15.0	60.1	62.2	44.7	40.4	42.8
UniIR (CLIP_SF) [35]	-	0.428B	44.3	16.2	61.8	65.3	47.1	41.7	44.7
Magiclens [52]	-	0.428B	38.8	8.3	35.4	26.0	31.0	23.7	27.8
External Data-Augmented Models									
Seed-1.6-embedding†	Seed1.6-flash	unknown	76.1	74.0	77.9	91.3	-	-	77.8
Ops-MM-embedding-v1†	Qwen2-VL	2.21B	68.1	65.1	69.2	80.9	-	-	69.0
Ops-MM-embedding-v1†	Qwen2-VL	8.29B	69.7	69.6	73.1	87.2	-	-	72.7
GME [53]	Qwen2-VL	2.21B	54.4	29.9	66.9	55.5	-	-	51.9
GME [53]	Qwen2-VL	8.29B	57.7	34.7	71.2	59.3	-	-	56.0
UNITE [18]	Qwen2-VL	2.21B	63.2	55.9	65.4	75.6	65.8	60.1	63.3
UNITE [18]	Qwen2-VL	8.29B	68.3	65.1	71.6	84.8	73.6	66.3	70.3
CAFe [48]	LLaVA-OV	0.894B	59.1	49.1	61.0	83.0	64.3	53.7	59.6
CAFe [48]	LLaVA-OV	8.03B	65.2	65.6	70.0	91.2	75.8	62.4	69.8
ColPali-v1.3 [7]	PaliGemma	2.92B	40.3	11.5	48.1	40.3	-	-	34.9
LamRA [23]	Qwen2-VL	8.29B	59.2	26.5	70.0	62.7	-	-	54.1
LamRA [23]	Qwen2.5-VL	8.29B	51.7	34.1	66.9	56.7	-	-	52.4
RzenEmbed [13]	Qwen2-VL	2.21B	68.5	66.3	74.5	90.3	76.1	67.4	72.3
RzenEmbed [13]	Qwen2-VL	8.29B	70.6	71.7	78.5	92.1	78.5	72.7	75.9
Qwen3-VL-Embedding [21]	Qwen3-VL	2.13B	70.3	74.3	74.8	88.5	75.3	74.5	75.0
Qwen3-VL-Embedding [21]	Qwen3-VL	8.14B	74.2	81.1	80.2	92.3	81.9	78.0	80.1
∼ 2B Models									
VLM2VEC [16]	Phi-3.5-V	4.15B	54.8	54.9	62.3	79.5	66.5	52.0	60.1
VLM2VEC [16]	Qwen2-VL	2.21B	59.0	49.4	65.4	73.4	66.0	52.6	59.3
VLM2VEC-V2 [24]	Qwen2-VL	2.21B	62.9	56.3	69.5	77.3	-	-	64.9
UniME-V2 [10]	Qwen2-VL	2.21B	62.1	56.3	68.0	72.7	67.4	58.9	63.6
LLaVE [19]	Aquila-VL	1.95B	62.1	60.2	65.2	84.9	69.4	59.8	65.2
B3 [32]	Qwen2-VL	2.21B	67.0	61.2	70.9	79.9	72.1	63.1	68.1
`TSEmbed` **(ours)**	Qwen2-VL	2.26B	**68.8**	**64.3**	**72.1**	**85.7**	**74.5**	**65.6**	**70.5**
∼ 7B Models									
VLM2VEC [16]	LLaVA-1.6	7.57B	61.2	49.9	67.4	86.1	67.5	57.1	62.9
VLM2VEC [16]	Qwen2-VL	8.29B	62.6	57.8	69.9	81.7	65.2	56.3	65.8
UniME-V2 [10]	LLaVA-OV	8.03B	65.3	67.6	72.9	90.2	74.8	66.7	71.2
UniME-V2 [10]	Qwen2-VL	8.29B	64.0	60.1	73.1	82.8	72.0	63.0	68.0
LLaVE [19]	LLaVA-OV	8.03B	65.7	65.4	70.9	**91.9**	75.0	64.4	70.3
B3 [32]	Qwen2-VL	8.29B	70.0	66.5	74.1	84.6	75.9	67.1	72.0
QQMM-embed [44]	LLaVA-OV	8.297B	66.8	66.8	70.5	90.4	74.7	65.6	70.7
`TSEmbed` **(ours)**	Qwen2-VL	8.40B	**71.1**	**70.3**	**75.9**	91.3	**78.8**	**69.6**	**74.7**

4.3 Main Results

Table 1 summarizes the comparative results on the MMEB, where `TSEmbed` demonstrates consistent superiority over baselines across all evaluation settings. **Overall Performance.** Among MLLM-based models trained exclusively on the MMEB dataset, `TSEmbed` achieves new state-of-the-art performance with **74.7%** at the 7B scale and **70.5%** at the 2B scale. At the 7B scale, `TSEmbed` surpasses the previous best method, B3 (72.0%), by a margin of **2.7%**, while

consistently outperforming other strong baselines such as UniME-V2 (+3.5%), QQMM-embed (+4.0%), and LLaVE (+4.4%). This superiority extends to the 2B scale, where `TSEmbed` improves over B3 (+2.4%) and LLaVE (+5.3%). Most notably, compared to the VLM2VEC (65.8% at 7B), `TSEmbed` yields a substantial improvement of **8.9%**. Crucially, `TSEmbed` achieves superior performance compared to several baselines augmented with external data: it outperforms Ops-MM-embedding-v1 by **2.0%** at the 7B scale (74.7% vs. 72.7%) and by **1.5%** at the 2B scale (70.5% vs. 69.0%), and substantially exceeds data-augmented models such as UNITE (74.7% vs. 70.3%) and CAFe (74.7% vs. 69.8%) at the 7B scale. These results demonstrate that our principled architectural design not only mitigates task conflict, but also unlocks exceptional data efficiency, effectively bridging the performance gap with models trained on external corpora.

Approaching Task-Specific Oracle Performance. Notably, `TSEmbed` achieves strong per-task performance: at the 7B scale, it attains 71.1% on Classification, 70.3% on VQA, 75.9% on Retrieval, and 91.3% on Visual Grounding, closely approaching or even surpassing the oracle performance set by dedicated task-specific models (71.5%, 70.9%, 75.5%, and 91.7%, respectively; Figure 1). These results collectively demonstrate that `TSEmbed` effectively preserves task-specific specialization within a unified framework, validating the task-scaling capability of our MoE-LoRA for universal multimodal embeddings.

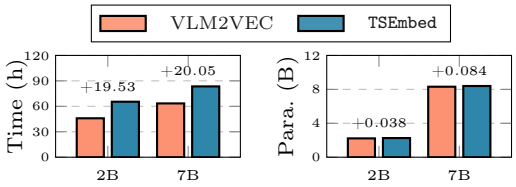
Robustness Across Data Distributions. Beyond merely achieving strong in-domain alignment, `TSEmbed` demonstrates an exceptional zero-shot generalization capability. At the 7B scale, the model attains 78.8% on in-distribution (IND) and 69.6% on out-of-distribution (OOD) tasks, outperforming the formidable B3 baseline by 2.9% and 2.5%, respectively. This robust performance seamlessly extends to the 2B scale, where `TSEmbed` achieves 74.5% IND and 65.6% OOD, surpassing B3 by 2.4% and 2.5%. These consistent performance gains across varying model scales and data regimes confirm that our MoE-driven decomposition encourages experts to cultivate intrinsically transferable semantic patterns, rather than overfitting to the training distribution.

4.4 Generalization and Efficiency Analysis

Generalization Analysis. To evaluate its real-world efficacy, we validate `TSEmbed` on proprietary production datasets encompassing diverse commercial domains, including advertising, theme, lockscreen, and gaming. Advertising evaluates fine-grained content alignment, theme assesses cross-modal intent-to-content matching, and lockscreen and gaming measure sequential preference modeling for next-item prediction. `TSEmbed` consistently exhibits robust zero-shot generalization without requiring any domain-specific fine-tuning, surpassing VLM2VEC by a significant margin (Table 2). Specifically, it achieves a substantial 21.87% gain in advertising, alongside steady improvements in theme (+2.05%), lockscreen (+5.23%), and gaming (+3.99%). These results substantiate that `TSEmbed` effectively learns transferable multimodal representations rather than merely memorizing patterns from academic datasets.

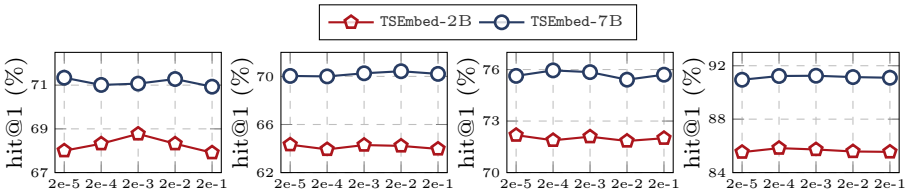
Table 2: Performance evaluation on proprietary production datasets sourced from a large-scale technology enterprise.

Domain	Metric	VLM2VEC	TSEmbed
Advertising	Recall	11.33%	**33.20%** (↑ 21.87)
Theme	NDCG@5	84.17%	**86.22%** (↑ 2.05)
Lockscreen	NDCG@1	56.31%	**61.54%** (↑ 5.23)
Gaming	NDCG@1	36.07%	**40.06%** (↑ 3.99)



(a) Training Efficiency. (b) Model Size.

Fig. 4: Efficiency analysis of TSEmbed.



(a) Classification. (b) VQA. (c) Retrieval. (d) Grounding.

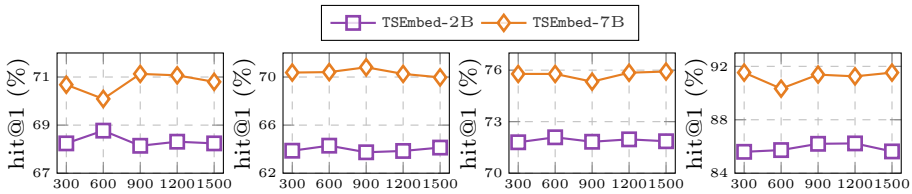
Fig. 5: Sensitivity analysis of the EANS decay parameter σ. We evaluate the impact of σ on four meta-task categories using both 2B and 7B models.

Training Efficiency. Figure 4(a) demonstrates that TSEmbed introduces minimal training overhead, incurring only an additional 19.53 and 20.05 hours of training time for the 2B and 7B models, respectively. This slight increase is primarily attributed to the computations required for expert routing and EANS. Given its computational efficiency and stable scaling behavior, TSEmbed is conducive to large-scale industrial deployment.

Parameter Efficiency. Furthermore, TSEmbed demonstrates exceptional structural scalability. As shown in Figure 4(b), it requires a mere 0.038B (+1.7%) and 0.084B (+1.0%) additional parameters for the 2B and 7B models, respectively. Given the significant performance gain of up to 11.8% over VLM2VEC at the 7B scale, TSEmbed delivers an optimal capacity-to-performance trade-off, effectively amplifying representational power with negligible parameter expansion.

4.5 Sensitivity Analysis

EANS Decay Sensitivity (σ). The decay parameter σ (Equation 7) governs hard-negative weighting sharpness. Remarkably, TSEmbed demonstrates exceptional stability across a vast range of $\sigma \in [2\times10^{-5}, 2\times10^{-1}]$ (Figure 5). Across all meta-tasks and model scales, performance fluctuations are tightly bounded within a $\sim$1% margin. Although extreme values induce slight degradations, stemming from weight homogenization at large σ and gradient over-concentration at small σ, the overall performance remains remarkably resilient. This striking consistency validates the inherent robustness of our framework.



(a) Classification. (b) VQA. (c) Retrieval. (d) Grounding.

Fig. 6: Sensitivity analysis of the warm-up duration T_{warmup}. We evaluate the impact of T_{warmup} on four meta-task categories using both 2B and 7B models.

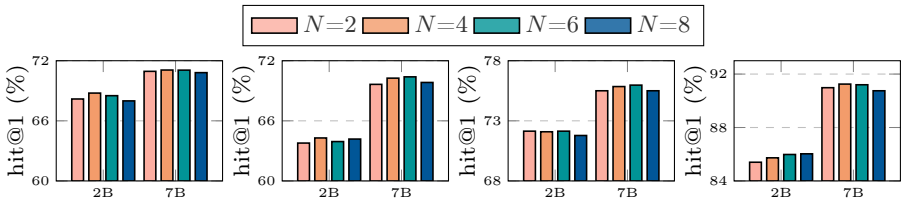
Table 3: Ablation study of `TSEmbed`. ✓/✗ indicate enabled/disabled modules.

Configuration	Components			Per Meta-Task Score				Overall
	MoE	EANS	Two-Stage	Cls.	VQA	Ret.	Grd.	
Qwen2-VL-2B [34]								
Baseline (VLM2VEC)	✗	✗	✗	59.00	49.40	65.40	73.40	59.30
+ MoE	✓	✗	✗	67.87	64.17	72.03	85.65	70.20
+ MoE + EANS	✓	✓	✗	68.31	64.26	71.59	85.08	70.14
`TSEmbed` (Full)	✓	✓	✓	**68.77**	**64.29**	**72.09**	**85.73**	**70.52**
Qwen2-VL-7B [34]								
Baseline (VLM2VEC)	✗	✗	✗	62.60	57.80	69.90	81.70	65.80
+ MoE	✓	✗	✗	70.25	70.15	75.22	91.23	74.21
+ MoE + EANS	✓	✓	✗	69.99	70.16	75.33	91.20	74.18
`TSEmbed` (Full)	✓	✓	✓	**71.07**	**70.26**	**75.85**	**91.25**	**74.68**

Warm-Up Period (T_{warmup}). The parameter T_{warmup} dictates when EANS activates, balancing router stabilization against refinement capacity. Figure 6 reveals a capacity-dependent stabilization dynamic. The 2B model, with its smaller parameter space, rapidly establishes stable routing patterns, optimally converging at just 600 steps (peaking at 68.77% for Classification, 64.29% for VQA, 72.09% for Retrieval). In contrast, the complex feature space of the 7B model demands a longer warm-up to solidify expert specialization, achieving optimal cross-task balance at 900–1200 steps, with Retrieval exhibiting a slight continued improvement up to 1,500 steps. Crucially, despite these distinct optimal points, both models exhibit relatively stable performance across adjacent step ranges, demonstrating the robustness of our two-stage learning paradigm.

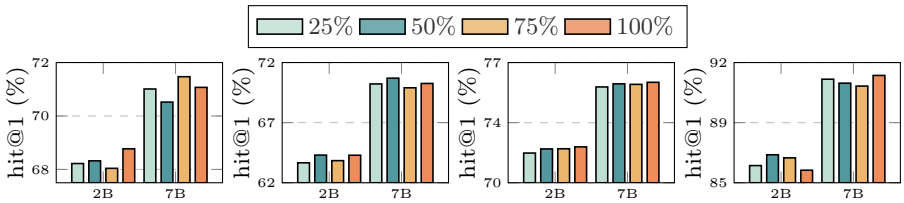
4.6 Ablation Study

Table 3 presents a detailed ablation analysis of `TSEmbed`. This breakdown analysis reveals three critical insights: **(1) MoE-LoRA effectively mitigates the problem of task conflict:** Transitioning from dense adaptation (VLM2VEC)



(a) Classification. (b) VQA. (c) Retrieval. (d) Grounding.

Fig. 7: Impact of expert number N on model performance. We evaluate $N \in \{2, 4, 6, 8\}$ across four meta-task categories on both 2B and 7B models.



(a) Classification. (b) VQA. (c) Retrieval. (d) Grounding.

Fig. 8: Impact of layer coverage on EANS routing aggregation. We evaluate the effect of aggregating router distributions from 25%, 50%, 75%, and 100% of layers.

to MoE-LoRA yields massive initial gains (+10.9% at 2B, +8.41% at 7B), proving that parameter-level expert specialization is essential to disentangle multimodal semantics. **(2) EANS strictly necessitates the two-stage warm-up:** Crucially, applying EANS without the two-stage paradigm marginally *degrades* performance (e.g., dropping from 70.20% to 70.14% at 2B). This validates our hypothesis that premature, unstabilized routing distributions inject unreliable noise rather than helpful hard-negative signals. **(3) Component synergy unlocks peak performance:** Only when the two-stage paradigm first establishes a stable routing topology does EANS function as intended. This indispensable synergy yields consistent improvements across all meta-tasks, pushing the full `TSEmbed` framework to a peak performance of 70.52% (2B) and 74.68% (7B).

4.7 More Analysis

Impact of Expert Number. As illustrated in Figure 7, we sweep the number of experts $N \in \{2, 4, 6, 8\}$ to evaluate scaling behaviors. The results reveal a clear performance "sweet spot" at $N = 4$, which achieves peak or near-peak performance. We attribute this to an organic alignment between the network capacity and the underlying taxonomy of the MMEB benchmark, which inherently comprises four distinct meta-task clusters. A smaller N (e.g., $N = 2$) forces disparate tasks into shared parameters, failing to fully disentangle conflicts. Conversely, scaling to $N = 8$ generally degrades performance, suggesting that an excessive

Table 4: Inference latency (min) of the Qwen2-VL model evaluated on an A100 GPU.

Model	Advertising (5,630)		Gaming (6,532)	
	Qwen2-VL-2B	Qwen2-VL-7B	Qwen2-VL-2B	Qwen2-VL-7B
VLM2VEC	35.72	42.02	49.58	59.43
`TSEmbed`	44.67	47.08	70.02	74.11

Table 5: Zero-shot performance on VisDoc (MMEB-v2).

Model	VDR-arxivQA	VDR-docvqa	VDR-infovqa	VDR-tabfquad
VLM2VEC	18.12 (↓ 38.28)	14.01 (↓ 10.16)	39.53 (↓ 24.24)	36.03 (↓ 32.54)
`TSEmbed`	**56.40**	**24.17**	**63.77**	**68.57**

number of experts over-fragments the semantic space, introducing severe routing ambiguity that outweighs any marginal capacity gains. While $N = 6$ offers slight domain-specific advantages (e.g., 75.97% on 7B retrieval), $N = 4$ yields the most robust cross-task generalization. Based on these findings, we recommend configuring N to roughly match the number of macro-task clusters in the target domain to prevent semantic over-partitioning.

Impact of Layer Coverage for EANS. Figure 8 demonstrates the effect of varying the EANS routing aggregation coverage from 25% to 100% of the backbone layers. The benefit of full coverage is explicitly evident in retrieval, which exhibits a near-monotonic performance trajectory, culminating at 72.09% (2B) and 75.85% (7B). This confirms that holistic, cross-layer routing effectively amalgamates the low-level and high-level semantic signals necessary for discerning hard negatives. While intermediate coverage levels (e.g., 50%) occasionally yield slight localized peaks in VQA and grounding, 100% coverage consistently maintains robust, top-tier performance across all meta-tasks and model scales.

Inference Latency. Table 4 demonstrates that `TSEmbed` introduces negligible inference latency over VLM2VEC in practical settings. This minimal cost is attributed to our streamlined design, which requires only lightweight gating projections and expert transformations during inference.

Performance on VisDoc. We further evaluate `TSEmbed` on VisDoc from MMEB-v2 [24] under a zero-shot setting. As shown in Table 5, `TSEmbed` consistently outperforms VLM2VEC across all four VisDoc retrieval tasks, demonstrating its effectiveness on visually rich document understanding scenarios.

5 Related Work

Universal Multimodal Embedding. Early dual-encoder models, such as CLIP [28] and SigLIP [33], excel at coarse-grained alignment but struggle with complex reasoning due to their shallow cross-modal interactions. Recently, the

proliferation of MLLMs has catalyzed a paradigm shift toward unified generative embeddings. For instance, Qwen3-VL-Embedding [21] leverages its foundational model to synthesize task-specific multimodal annotations, augmented by positive refinement and hard-negative mining. QQMM-embed [44] amplifies the gradients of hard negatives to enforce sharper discriminative boundaries. However, they suffer from severe task conflict, which `TSEmbed` is designed to address.

Task Conflict in Multi-Task Learning. Task conflict, where gradients from divergent objectives destructively interfere, is a fundamental bottleneck in multi-task learning [2, 54]. Traditional mitigations typically involve explicit gradient manipulation (e.g., PCGrad [49], GradNorm [4], Nash-MTL [26]) or static architecture routing (e.g., Cross-stitch [25], AdaShare [31]). However, these methods become computationally prohibitive and structurally intractable for massive, open-ended multimodal embeddings. Our MoE-based architecture provides an elegant alternative, achieving automatic task decomposition and gradient isolation without the exorbitant costs of explicit gradient engineering.

Hard Negative Sampling. Negative sample selection critically dictates the efficacy of contrastive representation learning [3, 11]. Standard approaches to hard negative mining often rely on maintaining large memory banks (e.g., MoCHi [17]), density-aware resampling [29], or complex curriculum schedules [51]. While these methods effectively improve cluster separation, they inevitably introduce significant computational overhead by requiring explicit feature-level similarity computations or external memory mechanisms. In contrast, our EANS strategically repurposes the MoE routing mechanism as a free intrinsic semantic proxy.

6 Conclusion

In this paper, we introduce `TSEmbed` to address the challenge of task conflict in universal multimodal embeddings. Our framework uniquely combines MoE with LoRA to achieve task-specific parameter specialization, effectively disentangling conflicting semantic objectives. Furthermore, we propose Expert-Aware Negative Sampling (EANS), a novel strategy that leverages routing distributions as an intrinsic proxy to dynamically emphasize hard negatives. To ensure training stability, we devise a two-stage learning paradigm that solidifies expert specialization before deploying EANS, guaranteeing robust optimization. Extensive evaluations on both the MMEB and real-world production datasets demonstrate the effectiveness of `TSEmbed` while preserving training efficiency.

Acknowledgements

This work is supported in part by the Science and Technology Development Fund of Macau (0107/2024/RIA2 and 0061/2025/RIB2), Joint Science and Technology Research Project with Hong Kong and Macau in Key Areas of Nansha District's Science and Technology Plan (EF2024-00180-IOTSC), and the Multi-Year Research Grant of University of Macau (MYRG-GRG2023-00211-IOTSC-UMDF and MYRG-GRG2024-00180-IOTSC).

References

1. Cao, M., Li, S., Li, J., Nie, L., Zhang, M.: Image-text retrieval: A survey on recent research and development. arXiv preprint arXiv:2203.14713 (2022)
2. Chen, S., Zhang, Y., Yang, Q.: Multi-task learning in natural language processing: An overview. ACM Computing Surveys **56**(12), 1–32 (2024)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
4. Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A.: Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: International conference on machine learning. pp. 794–803. PMLR (2018)
5. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2818–2829 (2023)
6. Cui, X., Cheng, J., Chen, H.y., Shukla, S.N., Awasthi, A., Pan, X., Ahuja, C., Mishra, S.K., Yang, Y., Xiao, J., et al.: Think then embed: Generative context improves multimodal embedding. arXiv preprint arXiv:2510.05014 (2025)
7. Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., Colombo, P.: Colpali: Efficient document retrieval with vision language models. In: International Conference on Learning Representations. vol. 2025, pp. 61424–61449 (2025)
8. Gao, L., Zhang, Y., Han, J., Callan, J.: Scaling deep contrastive learning batch size under memory limited setup. In: Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021). pp. 316–321 (2021)
9. Greenacre, M., Groenen, P.J., Hastie, T., d'Enza, A.I., Markos, A., Tuzhilina, E.: Principal component analysis. Nature Reviews Methods Primers **2**(1), 100 (2022)
10. Gu, T., Yang, K., Zhang, K., An, X., Feng, Z., Zhang, Y., Cai, W., Deng, J., Bing, L.: Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 21378–21386 (2026)
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
13. Jian, W., Zhang, Y., Liang, D., Xie, C., He, Y., Leng, D., Yin, Y.: Rzenembed: Towards comprehensive multimodal retrieval. arXiv preprint arXiv:2510.27350 (2025)
14. Jiang, H., Wang, Y., Zhu, Y., Lu, X., Qin, W., Wang, M., Wan, P., Tang, Y.: Embed-rl: Reinforcement learning for reasoning-driven multimodal embeddings. arXiv preprint arXiv:2602.13823 (2026)
15. Jiang, T., Song, M., Zhang, Z., Huang, H., Deng, W., Sun, F., Zhang, Q., Wang, D., Zhuang, F.: E5-v: Universal embeddings with multimodal large language models. arXiv preprint arXiv:2407.12580 (2024)
16. Jiang, Z., Meng, R., Yang, X., Yavuz, S., Zhou, Y., Chen, W.: Vlm2vec: Training vision-language models for massive multimodal embedding tasks. arXiv preprint arXiv:2410.05160 (2024)
17. Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D.: Hard negative mixing for contrastive learning. Advances in neural information processing systems **33**, 21798–21809 (2020)

18. Kong, F., Zhang, J., Liu, Y., Zhang, H., Feng, S., Yang, X., Wang, D., Tian, Y., Zhang, F., Zhou, G., et al.: Modality curation: Building universal embeddings for advanced multimodal information retrieval. arXiv preprint arXiv:2505.19650 (2025)
19. Lan, Z., Niu, L., Meng, F., Zhou, J., Su, J.: Llave: Large language and vision embedding models with hardness-weighted contrastive learning. arXiv preprint arXiv:2503.04812 (2025)
20. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
21. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720 (2026)
22. Liu, Q., Zhu, J., Yang, Y., Dai, Q., Du, Z., Wu, X.M., Zhao, Z., Zhang, R., Dong, Z.: Multimodal pretraining, adaptation, and generation for recommendation: A survey. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 6566–6576 (2024)
23. Liu, Y., Zhang, Y., Cai, J., Jiang, X., Hu, Y., Yao, J., Wang, Y., Xie, W.: Lamra: Large multimodal model as your advanced retrieval assistant. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4015–4025 (2025)
24. Meng, R., Jiang, Z., Liu, Y., Su, M., Yang, X., Fu, Y., Qin, C., Chen, Z., Xu, R., Xiong, C., et al.: Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. arXiv preprint arXiv:2507.04590 (2025)
25. Misra, I., Shrivastava, A., Gupta, A., Hebert, M.: Cross-stitch networks for multi-task learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3994–4003 (2016)
26. Navon, A., Shamsian, A., Achituve, I., Maron, H., Kawaguchi, K., Chechik, G., Fetaya, E.: Multi-task learning as a bargaining game. arXiv preprint arXiv:2202.01017 (2022)
27. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
29. Robinson, J., Chuang, C.Y., Sra, S., Jegelka, S.: Contrastive learning with hard negative samples. arXiv preprint arXiv:2010.04592 (2020)
30. Song, X., Lin, H., Wen, H., Hou, B., Xu, M., Nie, L.: A comprehensive survey on composed image retrieval. ACM Transactions on Information Systems **44**(1), 1–54 (2025)
31. Sun, X., Panda, R., Feris, R., Saenko, K.: Adashare: Learning what to share for efficient deep multi-task learning. Advances in Neural Information Processing Systems **33**, 8728–8740 (2020)
32. Thirukovalluru, R., Meng, R., Liu, Y., Su, M., Nie, P., Yavuz, S., Zhou, Y., Chen, W., Dhingra, B., et al.: Breaking the batch barrier (b3) of contrastive learning via smart batch mining. Advances in Neural Information Processing Systems **38**, 23218–23241 (2026)
33. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2:

Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
34. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
35. Wei, C., Chen, Y., Chen, H., Hu, H., Zhang, G., Fu, J., Ritter, A., Chen, W.: Uniir: Training and benchmarking universal multimodal information retrievers. In: European Conference on Computer Vision. pp. 387–404. Springer (2024)
36. Wu, Y., Jin, H., Guo, Z., Li, L.: Spotlight and shadow: Attention-guided dual-anchor introspective decoding for mllm hallucination mitigation. arXiv preprint arXiv:2604.10071 (2026)
37. Wu, Y., Li, J., Guo, Z., Li, L.: Elastic mixture of rank-wise experts for knowledge reuse in federated fine-tuning. arXiv preprint arXiv:2512.00902 (2025)
38. Wu, Y., Li, J., Guo, Z., Li, L.: Developmental federated tuning: A cognitive-inspired paradigm for efficient llm adaptation. In: The Fourteenth International Conference on Learning Representations (2026)
39. Wu, Y., Li, J., Tian, C., Guo, Z., Li, L.: Memory-efficient federated fine-tuning of large language models via layer pruning. arXiv preprint arXiv:2508.17209 (2025)
40. Wu, Y., Li, J., Tian, C., Tam, K., Guo, Z., Li, L.: Beyond end-to-end: Dynamic chain optimization for private llm adaptation on the edge. arXiv preprint arXiv:2604.06819 (2026)
41. Xu, N., Li, C., Du, T., Li, M., Luo, W., Liang, J., Li, Y., Zhang, X., Han, M., Yin, J., et al.: Copyrightmeter: Revisiting copyright protection in text-to-image models. arXiv preprint arXiv:2411.13144 (2024)
42. Xu, N., Zhang, J., Li, C., An, H., Zhou, C., Wang, J., Xu, B., Li, Y., Du, T., Ji, S.: Bridging the copyright gap: Do large vision-language models recognize and respect copyrighted content? In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 35949–35957 (2026)
43. Xu, N., Zhang, J., Li, C., Chen, Z., Zhou, C., Li, Q., Du, T., Ji, S.: Videoeraser: Concept erasure in text-to-video diffusion models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 5965–5994 (2025)
44. Xue, Y., Li, D., Liu, G.: Improve multi-modal embedding learning via explicit hard negative gradient amplifying. arXiv preprint arXiv:2506.02020 (2025)
45. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421 (2023)
46. Ye, Y., Zheng, Z., Shen, Y., Wang, T., Zhang, H., Zhu, P., Yu, R., Zhang, K., Xiong, H.: Harnessing multimodal large language models for multimodal sequential recommendation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 13069–13077 (2025)
47. Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., Liu, Z.: Evaluation of retrieval-augmented generation: A survey. In: CCF Conference on Big Data. pp. 102–120. Springer (2024)
48. Yu, H., Zhao, Z., Yan, S., Korycki, L., Wang, J., He, B., Liu, J., Zhang, L., Fan, X., Yu, H.: Cafe: Unifying representation and generation with contrastive-autoregressive finetuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6286–6297 (2025)
49. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. Advances in neural information processing systems **33**, 5824–5836 (2020)

50. Zhang, C., Zhang, H., Wu, S., Wu, D., Xu, T., Zhao, X., Gao, Y., Hu, Y., Chen, E.: Notellm-2: Multimodal large representation models for recommendation. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1. pp. 2815–2826 (2025)
51. Zhang, C., Zhang, K., Pham, T.X., Niu, A., Qiao, Z., Yoo, C.D., Kweon, I.S.: Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14441–14450 (2022)
52. Zhang, K., Luan, Y., Hu, H., Lee, K., Qiao, S., Chen, W., Su, Y., Chang, M.W.: Magiclens: Self-supervised image retrieval with open-ended instructions. arXiv preprint arXiv:2403.19651 (2024)
53. Zhang, X., Zhang, Y., Xie, W., Li, M., Dai, Z., Long, D., Xie, P., Zhang, M., Li, W., Zhang, M.: Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9274–9285 (2025)
54. Zhang, Y., Yang, Q.: A survey on multi-task learning. IEEE transactions on knowledge and data engineering **34**(12), 5586–5609 (2021)
55. Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., Cui, B.: Retrieval-augmented generation for ai-generated content: A survey. Data Science and Engineering pp. 1–29 (2026)