SCALE: TASK-CONDITIONED CALIBRATION OF HETEROGENEOUS MLLM SCORES FOR UNIVERSAL MULTIMODAL EMBEDDING LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing approaches to universal multimodal embedding learning adopt an MLLM-as-teacher paradigm but often overlook how the meanings of teacher scores and score gaps vary across tasks and instructions, leading to misaligned negative-sample and positive-pair supervision. To address this issue, we propose **SCALE**, which uses task-conditioned score calibration to coordinate negative routing and positive-pair ranking for reliable cross-task supervision. Specifically, **Task-Conditioned Score Calibration** establishes task-specific score references using task quantiles and scales, yielding *positive reliability* and a *negative boundary* to guide the two learning objectives. Guided by the *negative boundary*, **Collaboration-Based Negative-Sample Routing** assigns negatives to logit enhancement or suppression based on calibrated teacher score gaps, thereby matching their training influence to task-relative relevance. Complementing this candidate-level supervision, **Ranking-Based Positive-Pair Alignment** uses *positive reliability* to weight the alignment of student similarities with the teacher's ordering of positive pairs within each task, preserving differences in match quality across queries. Experiments on MMEB demonstrate the effectiveness of SCALE, with overall Precision@1 gains of **2.3** and **2.8** percentage points for 2B and 8B students, respectively.

1 INTRODUCTION

Universal multimodal embedding (UME) supports classification, visual question answering, retrieval, and grounding by mapping instruction-conditioned queries and candidates into a shared metric space (Jiang et al., 2025; Li et al., 2026; Gu et al., 2026). Multimodal large language models (MLLMs) enrich this training with continuous relevance scores that capture candidate differences and match quality beyond binary labels (Gu et al., 2026; Li et al., 2026). However, these scores depend on the task, modality, and even instruction (Chen et al., 2024; Xu et al., 2025; Lee et al., 2026). The same score or score gap may therefore carry different meanings across tasks, making uniform supervision difficult to justify. This raises a central question: *How can teacher scores be interpreted in task context to guide shared representation learning?*

Our empirical analysis reveals task-conditioned bias in both teacher score distributions and positive–negative score gaps (Figure 2). This bias affects supervision in two complementary ways. For negatives, the teacher's positive–negative score gap indicates how strongly it favors the positive, helping distinguish valid hard negatives from potential false negatives beyond student similarity alone. Yet gap distributions vary across tasks, so a single threshold can misinterpret teacher preference and assign inappropriate negative competition. For positives, teacher scores capture differences in match quality that binary labels do not express, but task-conditioned score scales make raw magnitudes an inconsistent basis for weighting pairs. However, existing teacher-based UME methods (Gu et al., 2026; Li et al., 2026) use these scores directly for supervision without explicitly accounting for task-conditioned bias. This leaves task-conditioned score variations entangled with relevance, potentially compromising the supervision transferred to the embedding model.

To address this challenge, we propose **SCALE** (Figure 1), which calibrates teacher scores against task-specific references before using them for representation learning. ***Task-Conditioned Score Calibration*** uses a lower quantile of each task's score gaps to set the negative routing boundary,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

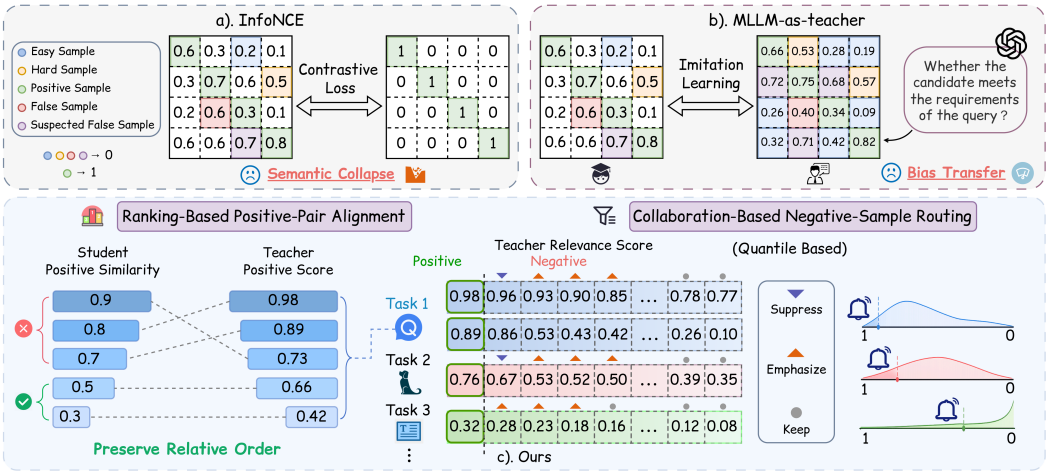


Figure 1: **Two uses of teacher scores.** Distribution distillation matches candidate probabilities within each query. SCALE uses task-specific score references to learn candidate differences through negative routing and match quality through positive-pair ranking within each task.

capturing weak teacher separation that the task average may obscure. A positive-score quantile provides a reference for positive-pair weighting. Together with task-specific scales, these references produce task-aware signals for two complementary objectives. Within each query, ***Collaboration-Based Negative-Sample Routing*** routes negatives by calibrated teacher gaps, with student similarity determining larger adjustments for harder negatives. Across queries within each task, ***Ranking-Based Positive-Pair Alignment*** aligns student similarities with teacher ordering, weighted by score gaps and calibrated positive-pair weights. Together, these objectives translate teacher judgments into supervision for candidate differences and match quality while preserving single-vector retrieval.

On MMEB, SCALE improves overall Precision@1 over distillation with the same backbones and teacher scores by **2.3** and **2.8** percentage points for 2B and 8B students, respectively, with gains across all four task groups. Ablations examine the contributions of negative routing and positive-pair ranking. The t-SNE visualization further suggests more compact within-task embeddings and clearer separation between retrieval tasks, consistent with stronger within-task coherence while retaining task-specific structure.

Our contributions are:

- We identify task-conditioned bias in MLLM teacher scores and score gaps, highlighting the need for task-specific interpretation when learning from candidate differences and positive-pair match quality.
- We develop SCALE with task-conditioned score calibration, collaboration-based negative-sample routing, and ranking-based positive-pair alignment. Our analysis shows how teacher score gaps and student similarity jointly adjust each negative's competition with the positive.
- On MMEB, SCALE improves overall Precision@1 by 2.3 and 2.8 percentage points for 2B and 8B students over distillation with the same backbones and teacher scores. Ablations examine the contributions of negative routing and positive-pair ranking.

2 Related Work

2.1 Universal Multimodal Embeddings

Large-scale vision-language pre-training established shared image–text representations through contrastive or generative supervision (Radford et al., 2021; Jia et al., 2021; Li et al., 2022). UME extends this setting to instruction-conditioned queries, mixed modalities, and heterogeneous retrieval tasks. VLM2Vec and MM-Embed develop general frameworks for this setting (Jiang et al., 2025; Lin

et al., 2025). Recent approaches address complementary aspects: QQMM amplifies gradients from explicit hard negatives (Xue et al., 2025), PUMA combines layer pruning with modality-adaptive contrastive learning (Lyu et al., 2025), and MetaEmbed supports flexible late interaction at inference (Xiao et al., 2026). Qwen3-VL-Embedding and UniME-V2 use MLLM supervision to improve embedding learning (Li et al., 2026; Gu et al., 2026). SCALE uses task-specific score references to learn from candidate differences through negative routing and match quality through positive-pair ranking.

2.2 MLLMs-as-Judges and Teachers

LLMs and MLLMs provide scalable evaluation and preference annotations. MT-Bench studies LLM-based model comparison, while MLLM-as-a-Judge and VL-RewardBench evaluate multimodal scoring and preference judgments (Zheng et al., 2023; Chen et al., 2024; Li et al., 2025). Prometheus-Vision and LLaVA-Critic develop open critic models, and RLAIF uses AI feedback for alignment (Lee et al., 2024b; Xiong et al., 2025; Lee et al., 2024a). Judge outputs can be sensitive to response order, external context, and visual manipulation (Wang et al., 2024; Xu et al., 2025; Hwang et al., 2025). This motivates care when transferring numerical scores across heterogeneous tasks. SCALE interprets teacher scores in task context: task-specific score references guide negative routing and positive-pair weighting, while within-task score gaps weight positive-pair ranking.

2.3 Robust Contrastive and Relation Learning

The effectiveness of contrastive learning depends on the negative set (Oord et al., 2018). Vision-language methods address false negatives through connection mining or label smoothing (Byun et al., 2024; Bulat et al., 2024), while TripletCLIP synthesizes hard negatives for fine-grained discrimination (Patel et al., 2024). Pairwise ranking objectives supervise relative similarities rather than exact targets (Burges et al., 2005). SCALE combines negative routing and positive-pair ranking using teacher scores in task context. Teacher score gaps set the routing direction, student difficulty controls its magnitude, and within-task teacher scores supervise match quality through positive-pair ranking.

3 Problem Setup and Motivation

UME training. Let $\mathcal{B} = \{(q_i, c_i^+, t_i)\}_{i=1}^B$ contain instruction-conditioned queries, annotated positives, and task identifiers. Queries and candidates can contain text, images, or their combinations. An embedding model f_θ maps each input to a unit vector in a shared space. Write s_i^+ for the cosine similarity of the paired query and positive, and s_{ij}^- for the similarity of q_i to another pair's candidate c_j^+. With temperature $\tau > 0$, the per-query InfoNCE loss is

$$\mathcal{L}_{\mathrm{NCE}} = -\log \frac{\exp(s_i^+/\tau)}{\exp(s_i^+/\tau) + \sum_{j \neq i} \exp(s_{ij}^-/\tau)}. \tag{1}$$

Binary labels do not distinguish valid hard negatives from relevant but unannotated candidates.

Fine-grained supervision from MLLM. An MLLM judge converts its "yes" and "no" logits for a query-candidate pair into a relevance score:

$$\widehat{r}_{ik} = \frac{\exp(\ell_{ik}^{\mathrm{yes}})}{\exp(\ell_{ik}^{\mathrm{yes}}) + \exp(\ell_{ik}^{\mathrm{no}})}. \tag{2}$$

A pretrained embedding model mines negatives, which the MLLM then scores (Gu et al., 2026; Li et al., 2026). Write p_i for the annotated positive's score and h_{ik} for the k-th mined negative's score. These scores are teacher judgments, not ground-truth relevance probabilities. The baseline, labeled KL-Distillation in our experiments, matches teacher and student candidate distributions using twice the Jensen–Shannon divergence:

$$\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}(\mathbf{p}_i^T \| \mathbf{m}_i) + D_{\mathrm{KL}}(\mathbf{p}_i^S \| \mathbf{m}_i), \tag{3}$$

Here $\mathbf{p}_i^T$ and $\mathbf{p}_i^S$ are normalized teacher and student distributions over the same candidate support, and $\mathbf{m}_i = (\mathbf{p}_i^T + \mathbf{p}_i^S)/2$. This objective transfers relative candidate probabilities within a query; it does not explicitly compare match quality across positive pairs from different queries.

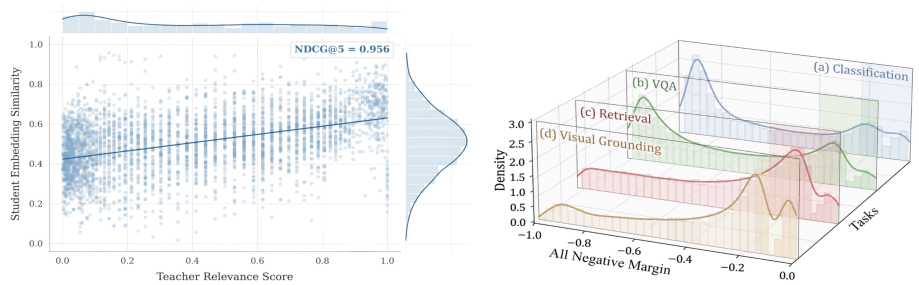


Figure 2: **Teacher–student heterogeneity.** (a) Different score distributions despite strong ranking agreement; (b) task-conditioned score gaps motivate task-specific score references.

Heterogeneous score spaces. Teacher scores and student similarities have different distributions despite NDCG@5 agreement above 0.9 (Figure 2(a)). Positive–negative score gaps also vary by task (Figure 2(b)). To express how task context can affect scores, consider the working model

$$\widehat{r}_{ik} = g_{t_i}(y_{ik}) + \epsilon_{ik}, \tag{4}$$

where y_{ik} is latent relevance, g_t is a strictly increasing task-conditioned mapping, and ϵ_{ik} is residual error. Without noise, g_t preserves within-task orders but changes score values and gaps. This model motivates our design; the distribution analysis does not verify it.

Why task context and score gaps matter. Within-task scores share g_t, whereas cross-task scores need not. If two pairs admit comparable latent relevance values and their residuals satisfy $|\epsilon_i|, |\epsilon_j| \leq e_t$, then

$$g_t(y_i) - g_t(y_j) \geq \widehat{r}_i - \widehat{r}_j - 2e_t.$$

A score gap above $2e_t$ implies $y_i > y_j$, motivating greater weight on clearer within-task orders. SCALE uses soft gap-based weights without estimating e_t; their usefulness depends on teacher quality and within-task comparability.

4 SCALE

4.1 Overview

SCALE learns from candidate differences through negative routing and match quality through positive-pair ranking (Figure 3). Task-specific score references guide both forms of supervision. Calibration first defines the negative routing boundary and positive-pair weights. Negative routing then combines teacher score gaps with student difficulty, while positive-pair ranking transfers within-task differences in match quality. Section 5.1 specifies auxiliary training terms.

4.2 Task-Conditioned Score Calibration

Negative routing boundary and scale. We compare each mined negative with its paired positive in log-odds space to expand score differences near zero and one. Let $a(u) = \log[u/(1-u)]$, with scores clipped away from zero and one (Section 5.1). The resulting score gap and its task-specific score reference are

$$d_{ik} = a(p_i) - a(h_{ik}), \qquad G_t = Q_{1-\rho_{\text{neg}}}(\{d_{ik} : t_i = t\}). \tag{5}$$

Here Q_α is the empirical quantile over teacher-scored training pairs in task t. The lower quantile G_t summarizes cases with weak teacher separation within the task. A mean score gap can obscure this region when other candidates are clearly separated. Using the same quantile level across tasks locates each reference at the same relative position in its own distribution. With $\rho_{\text{neg}} = 0.95$, this is the fifth percentile. Let $\bar{G}$ be the median of these references. Our boundary follows a conservative rule: tasks with weaker teacher separation require a larger score gap before strengthening negative competition.

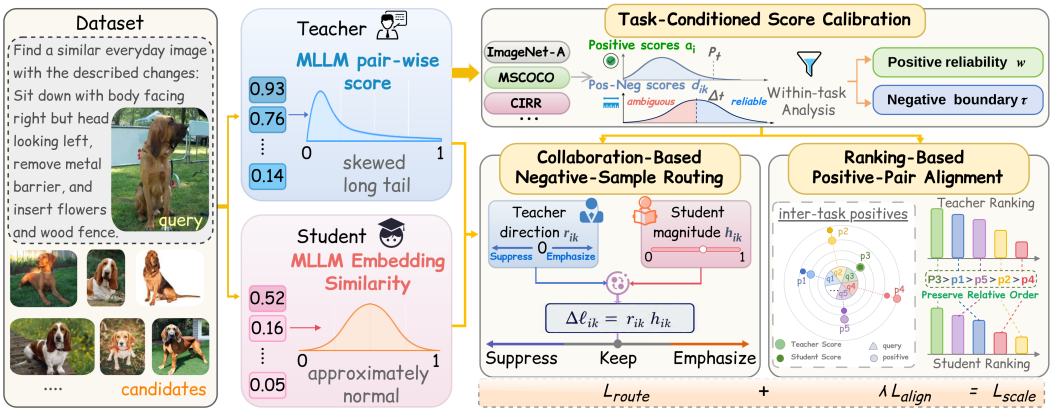


Figure 3: **SCALE framework.** Task-conditioned score references guide negative-sample routing within queries and positive-pair alignment across queries from the same task.

We implement this rule by adjusting a base boundary δ_0 according to the task reference, and estimate a robust task scale:

$$\delta_t = \delta_0 - (G_t - \bar{G}), \qquad S_t = c_{\mathrm{MAD}}\,\mathrm{MAD}(\{d_{ik} : t_i = t\}). \tag{6}$$

The offset $\bar{G} - G_t$ raises the boundary when teacher separation is weaker than the median task reference and lowers it when separation is stronger. A task at the median retains δ_0. This ties the boundary to observed task statistics through a shared adjustment rule. The robust scale S_t accounts for the spread of score gaps; MAD denotes median absolute deviation with fixed factor c_{MAD}. These statistics guide routing decisions without identifying false negatives or calibrating relevance probabilities.

Positive-pair weights. The positive also serves as the reference for negative supervision. We reduce the weight of training pairs with weak teacher support for the positive. Let $P_t = Q_{\rho_{\mathrm{pos}}}(\{p_i : t_i = t\})$ be the task-specific score reference for positives and σ_t^+ the standard deviation of positive scores in task t. We assign

$$w_i = \max\left(w_{\min},\ \sigma\left(\frac{p_i - P_{t_i}}{\max(\sigma_{t_i}^+, \varepsilon_s)}\right)\right). \tag{7}$$

Here σ is the sigmoid. Higher scores relative to the task reference receive larger weights; the floor $w_{\min}$ retains supervision from low-scoring positives. In contrastive learning, w_i controls the contribution of the entire query and its candidates. In positive-pair ranking, it controls the contribution of relations involving that positive pair. It therefore weights teacher supervision without removing the annotated positive. Numerical tolerances and scale details are given in Section 5.1.

4.3 Collaboration-Based Negative-Sample Routing

Teacher routing weight. We use one smooth function to control both the direction and strength of the teacher-guided adjustment. It should suppress negatives below the routing boundary, strengthen competition moderately above it, and reduce the adjustment when the teacher already separates the candidates clearly. With the normalized score gap z_{ik}, we define

$$z_{ik} = \frac{d_{ik} - \delta_{t_i}}{S_{t_i}}, \qquad r_{ik} = \tanh(z_{ik})\exp[-\mathrm{ReLU}(z_{ik})]. \tag{8}$$

The $\tanh$ term changes sign at the boundary and bounds the adjustment. The exponential term acts only above the boundary, so positive adjustments first increase and then decay as the teacher score gap grows. Below the boundary, the exponential term equals one, preserving suppression when the teacher provides weak or reversed support for the positive. Their product gives a bounded routing weight, $|r_{ik}| \le 1$, that smoothly connects suppression, emphasis, and diminishing adjustment.

Student routing magnitude. Student difficulty changes during training. Let s^-_{ik} be the cosine similarity to an explicit negative and $\ell^-_{ik} = s^-_{ik}/\tau$ its logit. The set $\mathcal{H}_i$ contains the selected explicit negatives in the gathered batch. We measure each negative against the current pool:

$$\bar{\ell}^-_i = \log\left(\frac{1}{|\mathcal{H}_i|}\sum_{h\in\mathcal{H}_i} e^{\ell^-_{ih}}\right), \qquad v_{ik} = \sigma(\ell^-_{ik} - \bar{\ell}^-_i). \tag{9}$$

The log-mean-exp reference emphasizes hard negatives, and v_{ik} increases with a negative's logit relative to this reference. A negative that is already easy for the student receives a smaller adjustment, even if the teacher score gap supports stronger competition. For each query's own teacher-scored negatives, we apply

$$\widetilde{\ell}^-_{ik} = \ell^-_{ik} + \frac{\eta}{\tau}\,\mathrm{sg}(r_{ik}v_{ik}), \tag{10}$$

Here η controls the adjustment and sg stops gradients through its weight. All gathered positives and selected negatives remain in the candidate pool; only the query's own scored-negative logits change. With the paired positive as target, cross-entropy over the routed logits $\widehat{\boldsymbol{\ell}}_i$ gives

$$\mathcal{L}_{\mathrm{route}} = \frac{\sum_i w_i\ \mathrm{CE}_i(\widehat{\boldsymbol{\ell}}_i)}{\sum_i w_i}. \tag{11}$$

Effect on negative competition. Let P_i and $\widetilde{P}_i$ be softmax probabilities before and after routing. Since the positive logit is unchanged, each negative's probability relative to the positive changes by

$$\frac{\widetilde{P}_i(k)/\widetilde{P}_i(+)}{P_i(k)/P_i(+)} = \exp\Big(\frac{\eta}{\tau}r_{ik}v_{ik}\Big). \tag{12}$$

The teacher term r_{ik} determines whether this ratio increases or decreases; student difficulty controls the magnitude through $v_{ik} > 0$. For a hard negative, a score gap below the boundary gives a larger downward logit adjustment; a moderately larger gap gives an upward adjustment. Thus, student difficulty amplifies the teacher-guided correction in either direction. Equation (12) isolates this effect even when other negatives are adjusted. Routing preserves negative labels while adjusting competition with the positive.

4.4 Ranking-Based Positive-Pair Alignment

Learning match quality. Negative routing learns from teacher score gaps between candidates. Positive-pair ranking uses the same scores to learn differences in match quality across queries. Two positive pairs may both outrank their own negatives while their student similarities disagree with the teacher's match-quality ordering. Positive-pair ranking supplies this missing comparison. We compare positive pairs only within the same task, assuming their teacher scores are sufficiently comparable.

Confidence-weighted relations. For two positive pairs, we use the raw teacher score gap $\Delta^T_{ij} = p_i - p_j$. Let m be a reference gap and τ_r the smoothing temperature. We define

$$C_{ij} = \mathbf{1}[t_i = t_j, i \neq j]\,\sigma\left(\frac{\Delta^T_{ij} - m}{\tau_r}\right), \qquad w_{ij} = \sqrt{w_i w_j}. \tag{13}$$

C_{ij} favors larger teacher gaps and excludes cross-task pairs; w_{ij} combines the positive-pair weights. Both directions retain soft weights near a tie, avoiding an abrupt choice of ordering from a small teacher score difference. The teacher scores determine which ordering receives more weight; they do not specify an exact target similarity. Let u_i denote positive-pair student similarity in the centered ranking view (Section 5.1). To favor pair i over pair j, we use

$$\ell^{\mathrm{rank}}_{ij} = \mathrm{softplus}\left(\frac{u_j - u_i}{\tau_r}\right). \tag{14}$$

This penalty grows when u_j exceeds u_i. We average it with $q_{ij} = C_{ij}w_{ij}$. A release factor A_{ij} reduces pressure as the student satisfies the relation, and D decays positive-pair ranking during training:

$$\mathcal{L}_{\mathrm{align}} = D\,\frac{\sum_{i,j} q_{ij}A_{ij}\,\ell^{\mathrm{rank}}_{ij}}{\sum_{i,j} q_{ij}}. \tag{15}$$

Table 1: **Results on MMEB.** IND/OOD denote in-/out-of-distribution averages. Parenthesized gains are relative to our KL-Distillation reproduction with the same backbone and teacher scores. ReMatch† uses additional mmE5 hard negatives. Published comparisons are reproduced from UniME-V2.

Model	Backbone	Per Meta-Task Score				Average Score		
		Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
No. of evaluation datasets		10	10	12	4	20	16	36
Without MMEB fine-tuning								
CLIP (Radford et al., 2021)	ViT-L	42.8	9.1	53.0	51.8	37.1	38.7	39.2
SigLIP (Zhai et al., 2023)	So/14	40.3	8.4	31.6	59.5	32.3	38.0	35.0
BLIP-2 (Li et al., 2023)	ViT-L	27.0	4.2	33.9	47.0	25.3	25.1	28.0
CLIP (Radford et al., 2021)	ViT-BigG/14	52.3	14.0	50.5	60.3	38.9	45.8	44.3
E5-V (Jiang et al., 2024)	LLaVA-1.6	39.7	10.8	39.4	60.2	34.2	33.4	37.5
FreeRet (Zhu et al., 2025)	Qwen2.5-VL-7B	59.7	52.8	49.2	54.7	–	–	53.7
With MMEB fine-tuning								
CLIP (Radford et al., 2021)	ViT-L	55.2	19.7	53.2	62.2	47.6	42.8	47.6
VLM2Vec (Jiang et al., 2025)	Qwen2-VL-2B	59.0	49.4	65.4	73.4	66.0	52.6	60.1
VLM2Vec (Jiang et al., 2025)	Qwen2-VL-7B	62.6	57.8	69.9	81.7	72.2	57.8	65.8
QQMM (Xue et al., 2025)	LLaVA-OV-7B	66.8	66.8	70.5	90.4	74.7	65.6	70.7
UniME-V2 (Gu et al., 2026)	Qwen2-VL-2B	62.1	56.3	68.0	72.7	67.4	58.9	63.6
UniME-V2 (Gu et al., 2026)	Qwen2-VL-7B	64.0	60.1	73.1	82.8	72.0	63.0	68.0
ReMatch† (Liu et al., 2026)	Qwen2.5-VL-3B	62.4	69.6	70.0	92.0	75.1	64.0	70.2
ReMatch† (Liu et al., 2026)	Qwen2.5-VL-7B	65.8	73.6	74.1	92.5	78.1	68.2	73.7
SCALE								
KL-Distillation	Qwen3-VL-2B	64.7	64.5	69.6	84.5	71.4	64.8	68.4
SCALE (Ours)	Qwen3-VL-2B	**66.1**(+1.4)	**67.8**(+3.3)	**71.6**(+2.0)	**86.9**(+2.4)	**75.3**(+3.9)	**65.0**(+0.2)	**70.7**(+2.3)
KL-Distillation	Qwen3-VL-8B	67.9	68.9	72.8	88.5	75.4	67.9	72.1
SCALE (Ours)	Qwen3-VL-8B	**69.6**(+1.7)	**73.4**(+4.5)	**75.3**(+2.5)	**91.1**(+2.6)	**79.6**(+4.2)	**69.1**(+1.2)	**74.9**(+2.8)

The denominator normalizes by the total relation weight. Release and decay appear only in the numerator, so their attenuation is preserved. They reduce continued ranking pressure after the student satisfies a relation and later in training. We set the loss to zero when no eligible relation exists. Section 5.1 defines release and decay.

4.5 Core Supervision

The core SCALE objective combines negative routing with positive-pair ranking, weighted by λ:

$$\mathcal{L}_{\text{scale}} = \mathcal{L}_{\text{route}} + \lambda \mathcal{L}_{\text{align}}. \tag{16}$$

Training also uses a forward-KL anchor: $\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{scale}} + \beta(s)\mathcal{L}_{\text{anchor}}$, where s is the training step. Section 5.1 specifies this auxiliary term.

5 Experiments

5.1 Experimental Setup

Data and evaluation. We evaluate on MMEB (Jiang et al., 2025), using its training split for embedding fine-tuning. The benchmark covers classification, visual question answering, retrieval, and visual grounding, with 20 in-distribution (IND) and 16 out-of-distribution (OOD) evaluation datasets. We report Precision@1 for the four meta-tasks and the IND, OOD, and Overall aggregates. No additional embedding-training datasets are used. All reported improvements are in percentage points.

Implementation Details. Following VLM2Vec (Jiang et al., 2025) and UniME-V2 (Gu et al., 2026), we train models in the main comparison for 2,000 steps on 8 NVIDIA A800 GPUs with a learning rate of 1×10^{-4}. We use LoRA with $r = 16$ and $\alpha = 64$. We set $\rho_{\text{neg}} = 0.95$ and $\rho_{\text{pos}} = 0.25$ in our experiments. We use the same MLLM teacher scores as UniME-V2. Our main experiments use Qwen3-VL-2B and Qwen3-VL-8B as the embedding backbones.

Training refinements. MAD is the median absolute deviation from the task median, with fixed scale factor $c_{\text{MAD}} \approx 1.4826$. Teacher scores are clipped to $[\varepsilon_p, 1 - \varepsilon_p]$ before the log-odds transform,

Table 2: **Component ablation (%).** Base.: distillation; T.: teacher routing; S.: student routing; R.: positive-pair ranking.

	Component settings				
	Base.	w/o T.	w/o S.	w/o R.	SCALE
Cls.	**64.2**	64.0	63.8	63.3	64.0
VQA	62.2	64.6	64.5	64.2	**65.6**
Ret.	68.0	67.5	68.0	68.6	**68.8**
Grnd.	82.3	**84.8**	**84.8**	84.0	83.8
IND	69.5	**71.5**	70.8	71.2	71.4
OOD	63.7	62.9	63.9	63.1	**64.2**
Overall	66.9	67.7	67.7	67.6	**68.2**

Table 3: **Boundary settings (%).** Parentheses give tightness. Supp.: fraction of mined negatives receiving suppression.

δ_0	0.6 (0.1)	0.8 (0.1)	1.0 (0.2)	1.2 (0.1)
Supp. (%)	∼1	∼2	∼4	∼8
Cls.	63.2	63.9	**64.0**	63.5
VQA	64.3	64.9	**65.6**	64.6
Ret.	67.3	68.0	**68.8**	67.6
Grnd.	83.9	84.1	83.8	**85.2**
IND	70.5	70.8	**71.4**	70.5
OOD	63.0	64.1	**64.2**	63.9
Overall	67.2	67.8	**68.2**	67.6

with $\varepsilon_p = \varepsilon_s = 10^{-4}$. Degenerate MAD scales fall back to the median absolute log-odds score gap, with a positive numerical floor.

Positive-pair ranking uses the detached release factor $A_{ij} = \sigma((p_i - p_j + \gamma - \mathrm{sg}(u_i - u_j))/\tau_r)$. The factor D decays linearly to zero at training fraction κ. Both factors appear only in the numerator of the ranking loss, preserving their attenuation. The reference configuration uses release margin $\gamma = 0.03$ and $\kappa = 0.70$.

Positive-pair ranking uses separately centered and renormalized query and positive embeddings, with moving-average momentum 0.99; contrastive logits remain uncentered. The forward-KL anchor uses teacher candidate targets and unrouted student logits, with its weight ramped linearly to 0.1 over 500 steps. This auxiliary term is distinct from the JS-based distillation baseline in Eq. (3). These refinements belong to the training configuration; the main comparison evaluates the resulting system rather than isolating each refinement.

5.2 Main Results

Comparison with distillation. Table 1 compares SCALE with the distillation baseline using the same backbones and teacher scores. SCALE improves Overall from 68.4 to **70.7** with Qwen3-VL-2B and from 72.1 to **74.9** with Qwen3-VL-8B. Both models improve across all seven reported aggregates. The largest meta-task gains occur in VQA: 3.3 points for 2B and 4.5 points for 8B. These results support the combined use of negative routing, positive-pair ranking, and auxiliary training terms across both student sizes.

Generalization and broader comparisons. IND gains are 3.9 and 4.2 points for the 2B and 8B students, compared with OOD gains of 0.2 and 1.2 points. The benefit is therefore concentrated in the training-task families, with more modest transfer to unseen datasets. Among the published methods in Table 1, SCALE-2B matches QQMM's 70.7 Overall, and SCALE-8B reaches 74.9 compared with ReMatch's 73.7. These results contextualize performance across different backbones and training pipelines; the same-backbone distillation comparison is the more direct evidence for SCALE. Inference uses single-vector similarity search without teacher calls or pairwise reranking.

5.3 Component and Qualitative Analyses

Ablation on Individual Components. In Table 2, removing teacher routing, student routing, or positive-pair ranking lowers Overall from 68.2 to 67.7, 67.7, and 67.6 within the same training budget, respectively. The full configuration exceeds distillation by 1.3 points, supporting complementary aggregate contributions. It leads VQA, retrieval, and OOD, although some ablated variants perform better on grounding or IND.

Ablation on the Base Boundary. In Table 3, the intermediate setting reaches 68.2 Overall and leads six of seven aggregates; grounding favors the largest boundary. Its approximately 4% suppression rate measures routing decisions, not true false-negative prevalence. Since the parenthesized tightness also varies, these are joint configurations rather than an isolated test of δ_0.

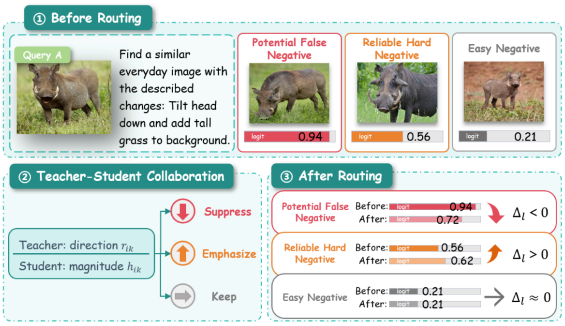


Figure 4: **Negative-sample routing.** The teacher sets routing direction, while student difficulty sets its magnitude (h_{ik} in the illustration, v_{ik} in the text). The three cases show suppression, emphasis, and minimal adjustment of negative competition for selected negatives.

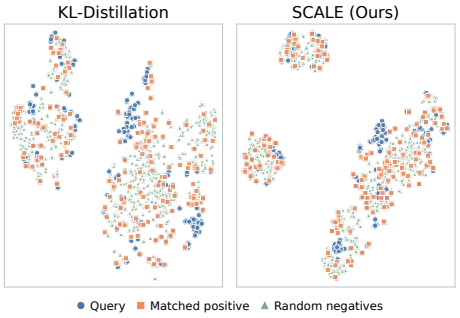


Figure 5: **Positive-pair similarities.** Independent t-SNE projections of 180 matched groups from the same 1,000 records; each contains a query, its positive, and three negatives. Both panels use identical records.

Sensitivity to the teacher. SCALE scores 68.2, 67.8, and 68.0 with the three teachers in Table 4, a range of 0.4 points. This suggests limited sensitivity to the tested teacher substitutions. The reproduced UniME-V2 results provide context but use a different student backbone, so the two columns do not constitute a controlled comparison of robustness.

Table 4: **Teacher sensitivity.** KL-Distill is from UniME-V2 (Qwen2-VL-2B); student configurations differ across columns.

Teacher Model	MMEB Overall	
	KL-Distill	SCALE
Qwen2.5-VL-7B	63.6	**68.2**
InternVL3-8B	58.5(-5.1)	**67.8**(-0.4)
InternVL3-14B	63.2(-0.4)	**68.0**(-0.2)

Routing regimes. Figure 4 illustrates suppression of an ambiguous negative, emphasis of a teacher-supported hard negative, and little adjustment to a negative with a large teacher score gap. These are training-time score corrections, not measured changes in encoder similarities. All selected negatives remain in the contrastive denominator. Equation (12) explains the mechanism: a negative's odds against the positive decrease or increase according to the teacher's direction, with the student controlling the strength of that change.

Positive-pair similarities. Figure 5 compares identical records under KL-Distillation and SCALE. The SCALE representations form a more compact embedding distribution, while samples from different retrieval tasks remain organized into more clearly separated regions. This structure suggests that SCALE promotes within-task coherence without collapsing the task-specific distinctions needed for universal retrieval. In the original 2,048-dimensional space, mean matched-pair cosine increases from 0.529 to 0.579, and the positive outranks all three exported negatives for 97.3% versus 98.8% of records. Since each t-SNE panel is fitted independently, the visual comparison is qualitative; cross-panel distances are not quantitatively comparable.

6 Conclusion

We present SCALE to address task-conditioned bias overlooked when MLLM teacher scores are used directly for universal multimodal embedding learning. Task-conditioned calibration coordinates negative-sample routing, which combines teacher gaps with student difficulty, and positive-pair alignment, which transfers match-quality ordering within each task. On MMEB, SCALE improves overall Precision@1 by 2.3 and 2.8 percentage points over distillation with the same backbones and teacher scores for 2B and 8B students, respectively. Gains across all four task groups support interpreting teacher judgments in task context, while single-vector retrieval is preserved without a teacher at inference.

Reproducibility Statement

To support reproducibility, we describe the score calibration procedure and learning objectives in the main text, with numerical safeguards and auxiliary training terms detailed in Section 5.1. Appendix C documents the MMEB dataset composition and input construction, and Appendix D provides the backbone configurations, teacher-score source, training hyperparameters, hardware, and evaluation protocol. The main comparisons use the same backbones and teacher scores for SCALE and the distillation baseline. Each reported configuration was trained once, and repeated-seed variability is not assessed.

AI use statement

Generative AI tools were used to assist with writing and language polishing, as well as literature retrieval and discovery, including finding related work. All technical content, experimental results, claims, and citations were reviewed by the authors, who take responsibility for the final manuscript.

References

Adrian Bulat, Yassine Ouali, and Georgios Tzimiropoulos. Fff: Fixing flawed foundations in contrastive pre-training results in very strong vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14172–14182, 2024.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

Jaeseok Byun, Dohoon Kim, and Taesup Moon. Mafa: Managing false negatives for vision-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27314–27324, 2024.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

Tiancheng Gu, Kaicheng Yang, Kaichen Zhang, Xiang An, Ziyong Feng, Yueyi Zhang, Weidong Cai, Jiankang Deng, and Lidong Bing. Unime-v2: Mllm-as-a-judge for universal multimodal embedding learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21378–21386, 2026.

Yerin Hwang, Dongryeol Lee, Kyungmin Min, Taegwan Kang, Yongil Kim, and Kyomin Jung. Fooling the lvlm judges: Visual biases in lvlm-based evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23197–23216, 2025.

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

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

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TE0KOzWYAF.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. In *Forty-first International Conference on Machine Learning*, 2024a.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11286–11315, 2024b.

Sua Lee, Sanghee Park, and Jinbae Im. Mm-judgebias: A benchmark for evaluating compositional biases in mllm-as-a-judge. In *The 64th Annual Meeting of the Association for Computational Linguistics, ACL 2026*. Association for Computational Linguistics, 2026.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

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

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. Vl-rewardbench: A challenging benchmark for vision-language generative reward models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24657–24668. IEEE, 2025.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pp. 44215–44234, 2025.

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

Yibo Lyu, Rui Shao, Gongwei Chen, Yijie Zhu, Weili Guan, and Liqiang Nie. Puma: Layer-pruned language model for efficient unified multimodal retrieval with modality-adaptive learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7653–7662, 2025.

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

Maitreya Patel, Naga Sai Abhiram kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. TripletCLIP: Improving compositional reasoning of CLIP via synthetic vision-language negatives. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ZfRGRK5Kxl.

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

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, 2024.

Zilin Xiao, Qi Ma, Mengting Gu, Chun cheng Jason Chen, Xintao Chen, Vicente Ordonez, and Vijai Mohan. Metaembed: Scaling multimodal retrieval at test-time with flexible late interaction. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=yKDqg9HwZX.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Lllava-critic: Learning to evaluate multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13618–13628. IEEE Computer Society, 2025.

Austin Xu, Srijan Bansal, Yifei Ming, Semih Yavuz, and Shafiq Joty. Does context matter? contextualjudgebench for evaluating llm-based judges in contextual settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9541–9564, 2025.

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

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=uccHPGDlao.

Yuhan Zhu, Xiangyu Zeng, Chenting Wang, Xinhao Li, Chunxu Liu, Yicheng Xu, Ziang Yan, Yi Wang, and Limin Wang. Freeret: Mllms as training-free retrievers. *arXiv preprint arXiv:2509.24621*, 2025.

A Additional Analysis of Teacher Supervision

A.1 Pairwise Judgments and Independent Embeddings

The MLLM teacher judges a query and a candidate jointly. Its yes/no output logits give the bounded score $\widehat{r}_{ik}$ defined in the main text. A representative judgment prompt supplies both inputs, asks whether the candidate meets the query's requirements, and restricts the answer to yes or no. In contrast, the student independently embeds the query and candidate before computing their cosine similarity. The two outputs refer to the same pair but have different numerical meanings: the teacher score is a probability over two answer tokens, whereas the student similarity is a geometric relation between separately encoded vectors.

This distinction also applies when the judge and student share a model family. Joint processing can recognize a relation that an independently encoded, fixed-size representation does not yet preserve. The teacher can therefore provide useful supervision without its score being a suitable numerical target for the student. SCALE uses the teacher's comparisons to change contrastive competition and to supervise positive-pair alignment, while retrieval still uses one vector per input. The working model $\widehat{r}_{ik} = g_{t_i}(y_{ik}) + \epsilon_{ik}$ in the main text expresses a possible source of score heterogeneity; it is a motivation, not an identified measurement model.

A.2 Why Full Distribution Matching Can Transfer Score Variation

For one query, let $\mathbf{q}$ be the normalized teacher distribution over its candidates and $\mathbf{p} = \mathrm{softmax}(\mathbf{a})$ the student distribution, where $\mathbf{a}$ contains student candidate logits. The experimental KL-Distillation baseline minimizes the twice-Jensen–Shannon objective in Eq. (3), namely

$$\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}(\mathbf{q}\|\mathbf{m}) + D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{m}). \tag{17}$$

Here, $\mathbf{m}$ is the arithmetic mean of the teacher distribution $\mathbf{q}$ and the student distribution $\mathbf{p}$. With strictly positive softmax probabilities, differentiation with respect to a student logit gives

$$\frac{\partial \mathcal{L}_{\mathrm{KL}}}{\partial a_k} = p_k \left[\log \frac{p_k}{m_k} - \sum_j p_j \log \frac{p_j}{m_j} \right]. \tag{18}$$

Thus the objective responds to differences between the complete teacher and student distributions, not only to their orderings. For comparison, a forward-KL objective $D_{\mathrm{KL}}(\mathbf{q}\|\mathbf{p})$ has gradient $p_k - q_k$ with respect to a_k. This forward-KL expression illustrates the same issue but is not the experimental baseline of Eq. (17).

To see the possible effect of heterogeneous scores, suppose an idealized student distribution $\mathbf{p}^\star$ reflects semantic relevance, while the teacher distribution is formed from scores $g_t(y_k) + \epsilon_k$. Even at $\mathbf{p} = \mathbf{p}^\star$, Eq. (18) generally remains nonzero when $\mathbf{q} \neq \mathbf{p}^\star$. A task-conditioned mapping or residual error can therefore influence the fitted student probabilities. This is a mechanism analysis, not proof that a particular teacher's errors cause the measured performance gap. A strictly increasing g_t preserves noiseless within-task order, but it need not preserve score gaps or cross-task comparability. SCALE uses task-conditioned references and within-task relations to reduce dependence on those absolute values.

A.3 Relation to Reward-Based Learning

An MLLM can also score generated responses as rewards. Online reward optimization changes the probability of sampled actions or responses and requires a policy and a sampling process; offline preference objectives still operate on generated-output likelihoods. SCALE instead trains an embedding model over observed query–candidate sets. Its teacher-dependent operation is a bounded shift of selected negative logits, followed by weighted contrastive learning and positive-pair alignment. It does not sample actions from the student or optimize an expected reward. The same distinction suggests possible use of task-conditioned judge relations in other ranking settings, although this paper evaluates only MMEB.

B Properties of Negative-Sample Routing

B.1 Task-Conditioned Boundary and Bounded Correction

For each task, the negative-sample routing reference G_t is the lower quantile of teacher positive–negative log-odds gaps. The boundary $\delta_t = \delta_0 - (G_t - \bar{G})$ rises when that task has weaker lower-tail teacher separation than the median task. The MAD-based scale S_t makes the normalized gap $z_{ik} = (d_{ik} - \delta_{t_i})/S_{t_i}$ less dependent on the spread of gaps in one task. These statistics are computed from available teacher-scored training pairs; the boundary is a routing rule, not an estimate of false-negative prevalence.

The teacher coefficient $r_{ik} = \tanh(z_{ik}) \exp[-\operatorname{ReLU}(z_{ik})]$ is negative below the boundary, positive above it, and tends back toward zero for large positive gaps. In particular, $|r_{ik}| \leq 1$. Because $0 < v_{ik} < 1$, the magnitude of the stopped-gradient logit shift in Eq. (10) is bounded by η/τ. All explicit negatives remain in the contrastive denominator, including those receiving a negative shift. A low teacher gap therefore changes the strength of competition rather than relabeling or removing the candidate.

B.2 Gradient and Odds Views

Write $\widetilde{P}_i(k)$ for a negative's softmax probability after routing, and treat the coefficient $r_{ik}v_{ik}$ as detached, as in Eq. (10). For the query's own selected negative logit, the direct contribution to the weighted cross-entropy gradient is

$$\frac{\partial \mathcal{L}_{\text{route}}}{\partial \ell_{ik}^{-}} = \frac{w_i}{\sum_j w_j} \widetilde{P}_i(k). \tag{19}$$

The corresponding derivative with respect to its cosine similarity includes the additional factor $1/\tau$. This direct contribution does not include any other role the same gathered candidate may play in another query's loss. Without routing, the analogous factor is $P_i(k)$. Increasing one negative logit while holding other logits fixed increases its own softmax probability; decreasing it has the reverse effect. This connects routing to hard-negative gradient amplification (Xue et al., 2025), with the teacher deciding the sign of the correction and the student controlling its size.

When several negatives are routed at once, their shared softmax denominator changes, so the absolute probability of an individual negative cannot be inferred from its shift alone. The comparison with the positive is exact because its logit is unchanged:

$$\frac{\widetilde{P}_i(k)/\widetilde{P}_i(+)}{P_i(k)/P_i(+)} = \exp\Big(\frac{\eta}{\tau} r_{ik} v_{ik}\Big). \tag{20}$$

A reliable hard negative above the boundary has increased odds against the positive, whereas an ambiguous negative below it has decreased odds. A very easy negative receives little change. This statement concerns training logits and gradients, not a measured change in encoder cosine similarity after training.

C MMEB Data and Input Construction

MMEB evaluates instruction-conditioned ranking across four meta-tasks (Jiang et al., 2025). Table 5 lists its 20 in-distribution (IND) and 16 out-of-distribution (OOD) datasets. The modality entries summarize each group; visual-grounding candidates are cropped image regions.

Table 5: **MMEB dataset composition.** I and T denote image and text; I+T denotes an image–text input. The query/candidate modality entries summarize forms present in each group.

Meta-task	Split	Datasets	Query	Candidate
Classification	IND	ImageNet-1K, N24News, HatefulMemes, VOC2007, SUN397	I / I+T	T / I
	OOD	Places365, ImageNet-A, ImageNet-R, ObjectNet, Country211	I	T
VQA	IND	OK-VQA, A-OKVQA, DocVQA, InfographicsVQA, ChartQA, Visual7W	I+T	T
	OOD	ScienceQA, VizWiz, GQA, TextVQA	I+T	T
Retrieval	IND	VisDial, CIRR, VisualNews-t2i, VisualNews-i2t, MSCOCO-t2i, MSCOCO-i2t, NIGHTS, WebQA	T / I / I+T	T / I / I+T
	OOD	OVEN, FashionIQ, EDIS, Wiki-SS-NQ	T / I+T	I / I+T
Grounding	IND	MSCOCO	I+T	I
	OOD	Visual7W-Pointing, RefCOCO, RefCOCO-Matching	I+T	I / I+T

Instruction format. An MMEB query includes a task instruction and available image and text content:

```
[IMAGE_TOKEN] Instruct: {task definition}
Query: {query}
```

The image token appears only for visual queries. The task definition describes the required relation; the query field carries the original question, caption, or referring expression when available. Candidates contain their original image, text, or both and are encoded independently. The backbone's standard input template serializes these fields, and embeddings are normalized before similarity computation. MMEB instructions are used without generation or rewriting.

D Training and Evaluation Details

The main comparisons train Qwen3-VL-2B and Qwen3-VL-8B for 2,000 steps on eight NVIDIA A800 GPUs, with learning rate 10^{-4} and LoRA rank/scale 16/64. Teacher scores are produced offline from the UniME-V2 annotation pool (Gu et al., 2026); no teacher call is needed during student training or inference. We set $\rho_{\text{neg}} = 0.95$ and $\rho_{\text{pos}} = 0.25$. Section 5.1 specifies clipping, the MAD fallback, centering, release and decay, and the forward-KL anchor. The KL-Distillation baseline instead uses Eq. (17).

Task-conditioned quantiles and scales can be prepared from offline teacher annotations. Student difficulty v_{ik} and routed logits are recomputed from current similarities. Routing changes only a query's own scored explicit negatives; all gathered candidates remain in the denominator. Positive-pair alignment compares same-task pairs and is zero without an eligible relation. Inference uses single-vector similarity search without a teacher or reranker.

Reported configurations were trained once; repeated-seed standard deviations are unavailable. Fixed MMEB candidate sets make checkpoint evaluation deterministic, but do not measure training variation. The same-backbone comparison in Table 1 is the direct controlled comparison. Component, boundary, and teacher-sensitivity tables have the qualifications stated in the main text and are not a single matched factorial experiment.

E CASE STUDY: CORRECTING CANDIDATE RANKINGS

Figure 6 compares KL-Distillation and SCALE on four selected queries: one composed image retrieval example and three visual grounding examples. Each row shows the query, its annotated positive, and three negatives. Blue and red values denote student cosine similarities under KL-Distillation and SCALE, respectively; arrows indicate the change between models. We assess the positive's margin over the highest-scoring displayed negative. These margins describe the displayed candidate subset rather than the full evaluation pool.

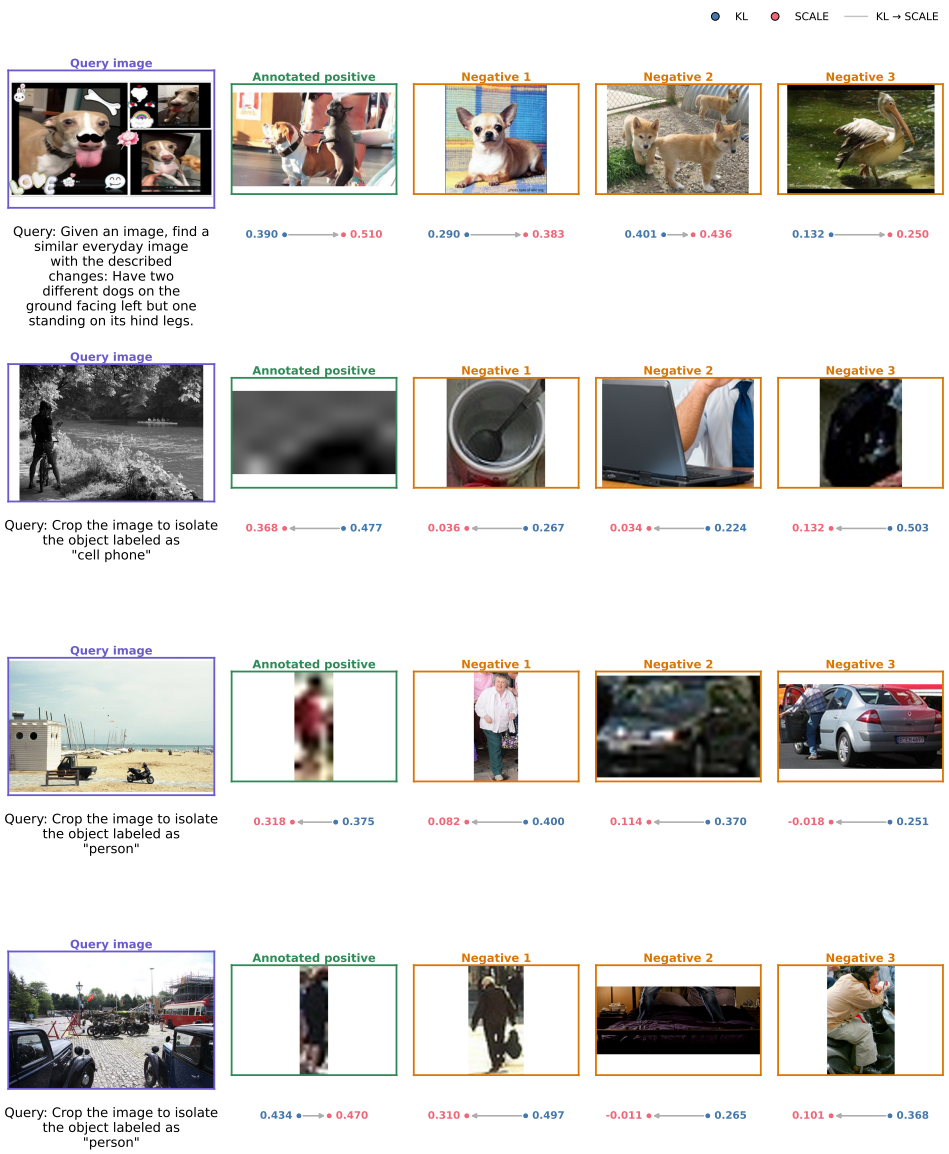


Figure 6: **Qualitative ranking corrections.** From top to bottom: composed retrieval with a dog-related modification, cell-phone grounding, and two person-grounding queries. Blue denotes KL-Distillation and red denotes SCALE, regardless of the values' left-to-right positions. SCALE ranks the annotated positive above all three displayed negatives in every row. The examples illustrate both stronger positive matching and reduced similarity to competing negatives.

Case 1: Compositional requirements beyond object identity. The first query requests two different dogs facing left, with one standing on its hind legs. The annotated positive depicts two

dogs, including an upright dog, whereas Negative 2 contains several dogs without the requested pose. KL-Distillation slightly favors this distractor over the positive (0.401 versus 0.390), suggesting that shared object content can outweigh the more specific modification. Under SCALE, the positive increases to 0.510, while Negative 2 reaches only 0.436. The positive's margin therefore changes from -0.011 to $+0.074$. All displayed candidates increase in similarity, so the correction comes from a larger gain for the positive rather than a uniform reduction of negative scores. This example illustrates improved discrimination of a compositional match from a semantically related alternative.

Case 2: Grounding a visually ambiguous small object. The second query asks for the cell phone in an outdoor scene. The annotated crop is highly blurred, making its identity difficult to verify from the crop alone. KL-Distillation assigns 0.477 to this positive but ranks Negative 3 higher at 0.503. SCALE lowers the positive score to 0.368 while reducing Negative 3 to 0.132; the other two negative scores fall to 0.036 and 0.034. The margin improves from -0.026 to $+0.236$. Thus, a lower positive cosine similarity can accompany a better ranking when misleading candidates decrease more strongly. The result is consistent with learning query-specific compatibility rather than relying only on an isolated crop's apparent object identity.

Case 3: Distinguishing the target person from a generic person image. The third query asks for a person in a beach scene. Its annotated positive is a small, low-resolution person crop, while Negative 1 clearly depicts a person in a different setting. KL-Distillation ranks this negative at 0.400, above the positive at 0.375; Negative 2 is also close at 0.370. SCALE assigns 0.318 to the positive and reduces these two negatives to 0.082 and 0.114, respectively. The strongest displayed negative switches from Negative 1 to Negative 2, and the positive's margin improves from -0.025 to $+0.204$. The example highlights the distinction between recognizing the requested category and matching the particular instance specified by the query image.

Case 4: Separating competing person instances. In the street-scene example, KL-Distillation favors Negative 1, another person crop, over the annotated positive (0.497 versus 0.434). SCALE both raises the positive to 0.470 and lowers Negative 1 to 0.310, changing the margin from -0.063 to $+0.160$. Negative 2 decreases from 0.265 to -0.011, and Negative 3 from 0.368 to 0.101. Unlike Cases 2 and 3, this correction combines stronger positive similarity with weaker negative similarity. Together, the grounding examples suggest that the learned representations better distinguish the annotated target from candidates that share broad category cues.

Relation to SCALE's learning objectives. Across the four cases, the relevant improvement is the positive's ordering relative to competing candidates, rather than an increase in every positive score. This behavior is consistent with SCALE's joint use of negative competition and positive-pair match-quality supervision. However, the figure reports final student similarities, not teacher scores, calibrated routing weights, or training trajectories. It therefore does not isolate which component causes an individual correction or establish that a particular negative was suppressed during training. These selected examples complement the aggregate evaluation and ablations by illustrating how successful ranking corrections can arise through different changes in the learned similarity space.