REVISITING VISION–LANGUAGE EMBEDDING DISTILLATION THROUGH NEGATIVE COMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning is a core objective for vision–language embedding models, where in-batch examples define the negative relations that shape the learned space. In heterogeneous training, globally mixed batches introduce negatives from datasets whose candidate pools are evaluated separately. We study this negative-source mismatch in vision–language embedding distillation. Across FastVLM-0.5B and LLaVA-OneVision-0.5B, classification and visual question answering, five objectives, and three training seeds, task-consistent batches improve overall Precision@1 by 1.54–3.72 points relative to the corresponding mixed-task objective. For contrastive-only training, the gains are 2.03–3.04 points. By comparison, under task-consistent batching, auxiliary distillation changes the overall score by only -0.60 to $+0.57$ points relative to contrastive-only training. Dataset-level effects are heterogeneous, with large gains on many datasets and regressions on a few. Because the samplers handle per-dataset remainders differently, the comparison does not isolate the causal effect of negative-source composition. Nevertheless, the aggregate association is consistent across settings, identifying negative-source composition as a first-order experimental axis and motivating matched-exposure and gradient-level tests of the proposed mechanism.

1 INTRODUCTION

Universal vision–language embedding models encode images, text, and interleaved inputs into a shared space, and are trained with in-batch contrastive learning on large unions of heterogeneous datasets (Jiang et al., 2025; Zhang et al., 2025; Lin et al., 2025; Meng et al., 2026). Deploying them under tight compute budgets has motivated distillation into compact students, usually by adding an auxiliary loss that transfers the teacher's embeddings, pairwise relations, or hierarchical structure (Park et al., 2019; Tian et al., 2020; Yang et al., 2024; Le et al., 2026). These objectives are compared on top of a shared contrastive recipe, which is implicitly treated as a neutral backbone.

But that shared recipe is not neutral: under in-batch InfoNCE, the batch itself defines the negative set. Each query is contrasted against every other candidate in the global batch, whereas multi-dataset benchmarks such as MMEB evaluate each dataset against its own candidate pool: an ImageNet-1K query is never ranked against SUN397 scene labels or N24News categories. A globally shuffled sampler therefore replaces some same-dataset comparisons with candidates the query will never face at test time, creating a *negative-source mismatch*. Practice is divided on how to handle this mismatch. Several text and multimodal embedders draw each batch from a single source (Su et al., 2023; Li et al., 2023; Merrick et al., 2024; Zhou et al., 2026), whereas others deliberately mix tasks, or homogenize only sub-batches, to avoid false negatives and unstable updates (Lee et al., 2025; Hu et al., 2025; Meng et al., 2026). Since existing recipe-level ablations conflate sampling with other design choices, they leave unclear both the magnitude of the sampler effect relative to auxiliary distillation and how sampling changes the contrastive update.

To answer these questions, we cross five objectives—contrastive-only, MSE, RKD, HieRD, and Merge KD—with mixed-task and task-consistent sampling while distilling the B3-Qwen2-2B teacher (Thirukovalluru et al., 2025) into FastVLM-0.5B (Vasu et al., 2025) and LLaVA-OneVision-0.5B (Li et al., 2025) on the CLS and VQA families of MMEB (Jiang et al., 2025). The mixed sampler shuffles the dataset union, whereas the task-consistent sampler builds each global batch from one dataset. We use the latter as an intervention on negative composition, not as a new method.

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.

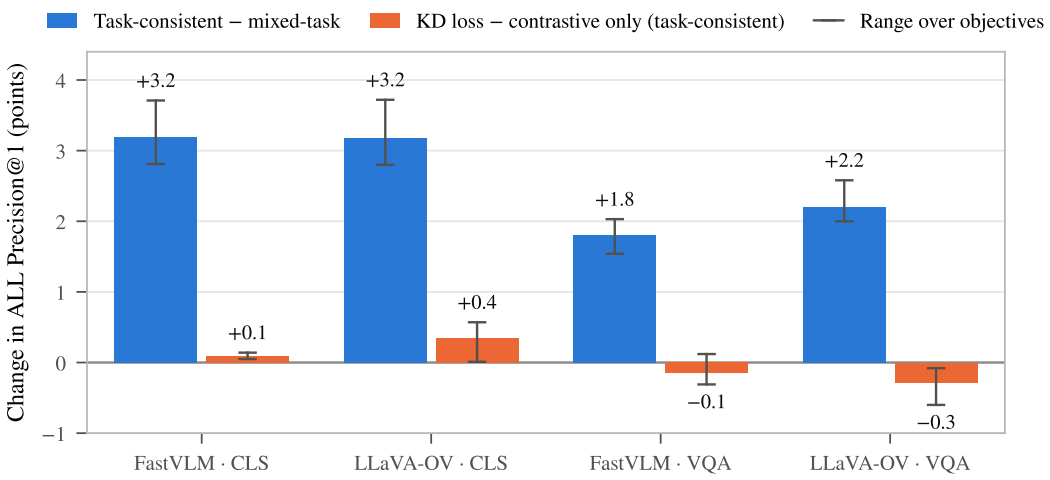


Figure 1: **The sampler moves performance more than the distillation objective.** Change in ALL Precision@1 for FastVLM-0.5B and LLaVA-OneVision-0.5B students on MMEB CLS and VQA. Blue: task-consistent minus mixed-task batching, with the objective fixed. Orange: effect of adding auxiliary losses (B3-Qwen2-2B teacher) under task-consistent batching. Bars show the mean over objectives and whiskers their range (Tables 1 and 2).

Our contributions are threefold.

1. **Negative composition outweighs auxiliary distillation.** Across two students, two task families, five objectives, and three seeds, task-consistent batching improves overall Precision@1 by 1.54–3.72 points, while auxiliary distillation changes it by only -0.60 to $+0.57$ with the sampler fixed (Figure 1).
2. **A theoretical analysis of sampler-dependent supervision.** At the population level, we characterize how the negative distribution shifts the optimal InfoNCE scoring rule. At the finite-batch level, we exactly decompose the mixed–homogeneous gradient gap into cross-task repulsion, softmax dilution, and lost same-task coverage.
3. **Empirical evidence points to lost same-task coverage.** We keep each query and its positive fixed, swap only the other candidates in its batch between mixed-task and same-task ones, and measure how much a single update reduces a fixed within-task reference loss. Mixed batches make less within-task progress after initialization, and our decomposition attributes most of this deficit to cross-task candidates displacing same-task comparisons.

2 Related Work

Universal embeddings and their distillation. Multimodal LLMs are turned into universal encoders by contrastive training over multi-task mixtures (Jiang et al., 2025; Zhang et al., 2025; Lin et al., 2025; Li et al., 2026), evaluated on MMEB and its successors (Jiang et al., 2025; Meng et al., 2026). Compressing them draws on relational (Park et al., 2019), contrastive (Tian et al., 2020), and topological (Kim et al., 2024) distillation, specialized to CLIP-style encoders (Yang et al., 2024; Vasu et al., 2024) and to compact VLM embedders (Le et al., 2026; Cho et al., 2026). Controlled comparisons repeatedly find that differences among such losses are small or fragile, and that implicit training choices can dominate the outcome (Tian et al., 2020; Yang et al., 2024; Stanton et al., 2021; Beyer et al., 2022). We identify one such choice, the candidate set of the student's own contrastive term, and analyze when an auxiliary loss added on top of it can still contribute a new update direction.

Batch composition in multi-dataset embedding training. Whether a batch should be drawn from a single dataset remains an unresolved recipe choice. Many embedders use single-source batches so that negatives cannot be separated by task identity (Su et al., 2023; Li et al., 2023;

Merrick et al., 2024; Zhou et al., 2026); others homogenize only sub-batches (Meng et al., 2026; Ovis-Embedding Team, 2026) or deliberately mix tasks to avoid "zigzag" updates and false in-batch negatives (Lee et al., 2025; Hu et al., 2025; Calvo et al., 2025). This evidence comes from recipe-level ablations entangled with other design choices and rests on competing intuitions about shortcuts, hardness, and false negatives. We instead hold the objective fixed, change only the sampler, and ask which part of the contrastive update it changes.

Negatives, density ratios, and gradient interference. Negative selection affects contrastive learning through false negatives (Chuang et al., 2020), hardness (Robinson et al., 2021; Thirukovalluru et al., 2025), and the coverage–collision trade-off (Ash et al., 2022). Ranking-based noise-contrastive estimation recovers $\log p(y \mid x) - \log r(y)$ up to a query-dependent constant (Gutmann & Hyvärinen, 2010; Oord et al., 2018; Ma & Collins, 2018), so the negative law shapes the learned score; logit corrections (Bengio & Senécal, 2003; Yi et al., 2019) can reweight observed candidates but cannot add missing same-task comparisons. Multi-task work instead explains joint-training deficits through conflicting task gradients (Yu et al., 2020; Wang et al., 2021), though strong baselines often match such corrections (Kurin et al., 2022). With in-batch InfoNCE, mixing changes both the gradients being averaged and each query's comparison set. We therefore compare mixed and within-dataset negative laws, decompose their finite-batch gap into cross-task repulsion, softmax dilution, and lost same-task coverage, and use a hard-negative control to separate source compatibility from hardness.

3 How Negative Composition Shapes the Contrastive Update

Problem formulation. Let d index one training dataset within either CLS or VQA, and let $\mathcal{B} = \{(x_i, y_i, d_i)\}_{i=1}^{B}$ be a *global* batch after distributed gathering. The student encodes query x_i and its positive candidate y_i as normalised embeddings q_i and c_i. With $z_{ij} = q_i^\top c_j/\tau$, the row InfoNCE loss is

$$\ell_i = -z_{ii} + \log \sum_{j=1}^{B} \exp z_{ij}, \qquad \mathcal{L}_{\mathrm{CL}} = \frac{1}{B} \sum_i \ell_i. \tag{1}$$

Every candidate other than the positive acts as a negative. Because this term is the student's primary objective and inference ranks by the same similarity, the choice of negatives shapes the embedding space itself rather than acting as an auxiliary regularizer (Wang & Isola, 2020). The *mixed* arm shuffles the concatenated datasets. The *task-consistent* (homogeneous) arm forms each global batch inside one dataset before sharding it over workers (Su et al., 2023), so that after all-gather every query is contrasted only against candidates from its own dataset.

3.1 Same-task coverage under the two samplers

For a query from dataset d, let K_d be the number of other examples from d in its mixed batch. Under uniform shuffling from a large concatenated mixture, with source share π_d,

$$K_d \approx \mathrm{Binomial}(B-1, \pi_d), \qquad \mathbb{E}[K_d] \approx (B-1)\pi_d. \tag{2}$$

Task-consistent batching instead gives $K_d = B - 1$. These are counts of eligible same-task comparisons, not necessarily useful ones, since repeated labels and small candidate pools can still create false or uninformative negatives (Ash et al., 2022; Chuang et al., 2020). With $p_{ij} = \exp(z_{ij})/\sum_k \exp(z_{ik})$ and $h_{ij} = \nabla_\theta z_{ij}$,

$$\nabla_\theta \ell_i = \sum_{j \neq i} p_{ij}(h_{ij} - h_{ii}). \tag{3}$$

The sampler therefore changes both which relations are available and the softmax weights on the relations already in the batch, so cross-task candidates can alter the update without being individually antagonistic. Task consistency is also a separate axis from hardness. We therefore compare mixed-random, homogeneous-random, and homogeneous-hard batches, with hard negatives mined by B3 (Thirukovalluru et al., 2025). The first contrast isolates task compatibility, and the second asks whether harder within-task negatives still help once compatibility is fixed.

3.2 POPULATION VIEW: THE SAMPLER SHIFTS THE INFONCE TARGET

Ranking-based contrastive estimation identifies a density ratio (Gutmann & Hyvärinen, 2010; Oord et al., 2018; Ma & Collins, 2018). We use this fact to state what a change of negative law can and cannot explain.

Proposition 1 (sampler-dependent InfoNCE target). Fix dataset d and suppose that, conditional on almost every query x, $p_d(y \mid x)$ and each negative distribution being compared have strictly positive densities on the same candidate support Ω. Draw $K \geq 1$ negatives i.i.d. from an anchor-independent distribution r, and consider population ranking InfoNCE with an unrestricted, finite score function. Then *every* population-optimal score, up to equality almost everywhere, has the form

$$s_r^*(x, y) = \log p_d(y \mid x) - \log r(y) + C_r(x), \qquad y \in \Omega, \tag{4}$$

where C_r is independent of the candidate. In particular, for a within-dataset distribution r_d and the mixed distribution $r_{\text{mix}} = \sum_k \pi_k r_k$, any two optima satisfy

$$s_{\text{mix}}^*(x, y) - s_d^*(x, y) = \log \frac{r_d(y)}{r_{\text{mix}}(y)} + C'(x), \qquad y \in \Omega. \tag{5}$$

The two optimum sets have no common element on Ω if and only if $\log(r_d/r_{\text{mix}})$ is not constant almost everywhere on Ω.

Proof sketch. The Bayes posterior over which of the $K+1$ candidates is the positive is proportional to $p_d(y \mid x)/r(y)$, and cross-entropy is minimized exactly when the model reproduces it. The full argument is in Appendix A.1.

A sampler can therefore shift the relative scores of candidates, but the shift need not help. Proposition 1 also has clear limits. It idealizes negatives as i.i.d. and the scorer as unrestricted, whereas in-batch negatives are shared across rows and embedding scores are bounded and low-rank. If $r_{\text{mix}} \propto r_d$ on the candidates of dataset d, as with disjoint task vocabularies, Eq. (5) is constant within the task and predicts no ranking shift. Our CLS evaluation also ranks against a fixed label bank rather than i.i.d. draws. For these reasons, Proposition 1 alone does not imply that homogeneous training is better. Appendix A.1 states when the population target matches such an evaluation, including a conditional top-1 stability bound. The finite-batch analysis below explains the gap without requiring the population optimum to change.

3.3 FINITE-BATCH VIEW: A PAIRED STEP COMPARISON

To see how mixed negatives change an actual update, we hold the anchors fixed and rebatch only their companions. For a focal task, we construct

$$\mathcal{B}_{\text{mix}} = A \cup S \cup X, \qquad \mathcal{B}_{\text{rest}} = A \cup S, \qquad \mathcal{B}_{\text{hom}} = A \cup S \cup F, \tag{6}$$

where S contains same-task companions shared by both arms, X contains cross-task companions present only in the mixed batch, and F contains the same-task replacements that fill those slots in the homogeneous batch. Both arms thus have the same anchor–positive pairs and the same batch size. For each focal anchor, we compute g_{mix} and g_{hom} with respect to the *same query embedding*. We also take $g_{\text{ref}} = \nabla_q L_{\text{ref}}$, where L_{ref} is a within-dataset contrastive probe over a separate fixed pool of 256 task-matched candidates. The primary statistic, estimated by paired resampling over independent anchor sets, is

$$D_{\text{full}} = \langle g_{\text{ref}}, g_{\text{mix}} - g_{\text{hom}} \rangle \tag{7}$$

Proposition 2 (reference-loss step comparison). For a *realized*, paired mixed/homogeneous draw at a fixed query embedding q, suppose $\nabla_q L_{\text{ref}}$ is L-Lipschitz for $L > 0$ in a neighborhood containing both candidate steps. Then

$$\begin{aligned} &L_{\text{ref}}(q - \eta g_{\text{mix}}) - L_{\text{ref}}(q - \eta g_{\text{hom}}) \\ &\qquad = -\eta D_{\text{full}} + R_\eta, \quad |R_\eta| \leq \frac{L\eta^2}{2}(\|g_{\text{mix}}\|^2 + \|g_{\text{hom}}\|^2). \end{aligned} \tag{8}$$

If $D_{\text{full}} < 0$, the mixed step leaves strictly larger reference loss whenever

$$0 < \eta < \frac{2|D_{\text{full}}|}{L(\|g_{\text{mix}}\|^2 + \|g_{\text{hom}}\|^2)}. \tag{9}$$

The guarantee concerns steps in embedding space, not full-model parameter updates, and it becomes vacuous as $D_{\text{full}} \to 0$.

Proof sketch. Apply the smoothness bound to both steps at the same q and subtract; see Appendix A.2.

3.4 Decomposing the mixed–homogeneous gap

To attribute D_{full} to mechanisms, we follow the allocation used by the probe: cross-task candidates contribute *repulsion only*, and the positive attraction is assigned entirely to the same-task part. Appendix A.3 compares this with the per-negative allocation of Eq. (3). For focal anchor i, work in query-embedding space and write $h_{ij} = \nabla_{q_i} z_{ij}$ and $u_i = h_{ii}$. Let $R = A \cup S$ be the restricted candidate set, including the positive index i. With the mixed softmax weights,

$$g^{\text{rep}}_{X,i} = \sum_{j \in X} p^{\text{mix}}_{ij} h_{ij}, \qquad g^{R,\text{code}}_{\text{mix},i} = \sum_{j \in R} p^{\text{mix}}_{ij} h_{ij} - u_i, \qquad g_{\text{rest},i} = \sum_{j \in R} p^{\text{rest}}_{ij} h_{ij} - u_i. \tag{10}$$

Then $g_{\text{mix},i} = g^{R,\text{code}}_{\text{mix},i} + g^{\text{rep}}_{X,i}$, and the signed, path-dependent components are

$$\delta_{X,i} = g^{\text{rep}}_{X,i}, \qquad \text{(cross-task repulsion under mixed weights)}, \tag{11}$$

$$\delta_{D,i} = g^{R,\text{code}}_{\text{mix},i} - g_{\text{rest},i}, \qquad \text{(restricted candidate-weight dilution)}, \tag{12}$$

$$\delta_{F,i} = g_{\text{rest},i} - g_{\text{hom},i}, \qquad \text{(additional same-task coverage)}. \tag{13}$$

The restricted and cross-task partition sums, their odds, and the cross-task softmax mass are

$$Z_{R,i} = \sum_{j \in R} e^{z_{ij}}, \qquad Z_{X,i} = \sum_{j \in X} e^{z_{ij}}, \qquad \epsilon_i = \frac{Z_{X,i}}{Z_{R,i}}, \qquad m_{X,i} = \frac{Z_{X,i}}{Z_{R,i} + Z_{X,i}} = \frac{\epsilon_i}{1 + \epsilon_i}. \tag{14}$$

For $j \in R$, $p^{\text{mix}}_{ij} = (1 - m_{X,i}) p^{\text{rest}}_{ij}$. Define the restricted softmax-weighted candidate gradient

$$\mu_{R,i} = \sum_{j \in R} p^{\text{rest}}_{ij} h_{ij} = g_{\text{rest},i} + u_i. \tag{15}$$

The probe's allocation then gives the *exact* identity

$$\begin{aligned} g^{R,\text{code}}_{\text{mix},i} &= (1 - m_{X,i}) \mu_{R,i} - u_i, \\ \boxed{\delta_{D,i}} &= \boxed{-m_{X,i} \mu_{R,i}} = -m_{X,i}(g_{\text{rest},i} + u_i), \\ D_{D,i} &= -m_{X,i} \langle g_{\text{ref},i}, \mu_{R,i} \rangle. \end{aligned} \tag{16}$$

Dilution therefore shrinks the candidate gradient $\mu_{R,i}$, not the full restricted gradient $g_{\text{rest},i} = \mu_{R,i} - u_i$. Its sign against the reference is set by $b_i = \langle g_{\text{ref},i}, \mu_{R,i} \rangle$, and its mean splits as

$$-\mathbb{E}[D_{D,i}] = \mathbb{E}[m_X b] = \mathbb{E}[m_X]\,\mathbb{E}[b] + \text{Cov}(m_X, b). \tag{17}$$

A positive mean D_D therefore indicates negative mass-weighted alignment between the candidate gradient and the reference. It does not indicate that g_{rest} is anti-aligned with g_{ref}. The components telescope (Figure 2):

$$\begin{aligned} g_{\text{mix},i} - g_{\text{hom},i} &= \delta_{X,i} + \delta_{D,i} + \delta_{F,i}, \\ D_{\text{full},i} &= D_{X,i} + D_{D,i} + D_{F,i}, \quad D_{k,i} = \langle g_{\text{ref},i}, \delta_{k,i} \rangle. \end{aligned} \tag{18}$$

Only the split between D_X and D_D depends on how positive attraction is allocated; D_{full} and the coverage term D_F do not. Because δ_F compares batches of different sizes, the component shares are descriptive rather than a unique causal attribution.

3.5 When does auxiliary distillation add a new direction?

Relational, contrastive, and hierarchical KD transfer pairwise or higher-order teacher structure (Park et al., 2019; Tian et al., 2020; Le et al., 2026). Our Merge KD comparator aligns the teacher and student minimax-connectivity matrices of the indexed query–candidate graph. We ask whether such losses still add a new update direction once the contrastive term is well constructed. Let

$$g_C = \nabla_\theta L_{\text{con}}, \qquad d = \lambda \nabla_\theta L_{\text{KD}}, \qquad g_{\text{tot}} = g_C + d, \tag{19}$$

and decompose $d = d_\parallel + d_\perp$ relative to g_C.

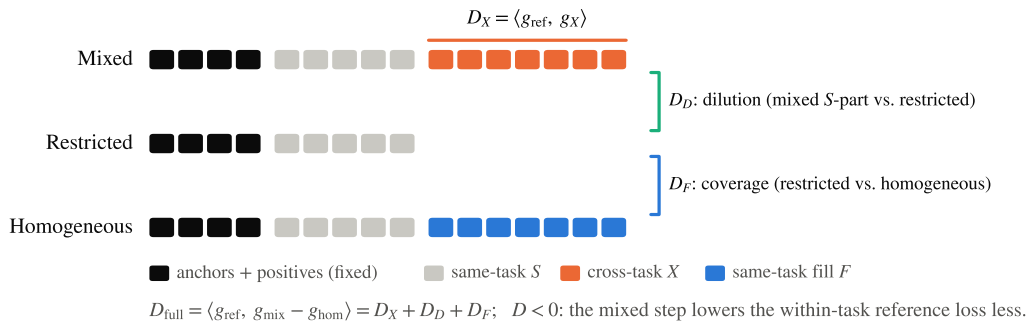


Figure 2: **Rebatching probe (schematic, not measured data).** The anchor–positive pairs and final batch size are fixed. The cross-task companions X of the mixed batch are removed to form the restricted batch and are replaced by same-task companions F in the homogeneous batch. The resulting directional gap decomposes exactly as $D_{\text{full}} = D_X + D_D + D_F$. A negative D_{full} predicts less first-order descent on the within-task reference loss under the mixed update.

Lemma 1 (local directional effect of auxiliary KD). Let $g_C \neq 0$, $u = g_C/\|g_C\|$, and write $d = au + d_\perp$, where $a = \langle d, u \rangle$ and $\langle d_\perp, u \rangle = 0$. If $\|g_C\| + a > 0$, the angle $\phi \in [0, \pi/2)$ between g_C and g_{tot} satisfies

$$\tan \phi = \frac{\|d_\perp\|}{\|g_C\| + a}. \tag{20}$$

A small orthogonal component therefore rotates the update only slightly, provided the forward component $\|g_C\| + a$ stays bounded away from zero. This is a geometric identity (Appendix A.4); it does not claim that auxiliary gradients are small.

To separate directional novelty from the KD coefficient, we evaluate both gradients at the same checkpoint and on the same batch. Writing $k = \nabla_\theta L_{\text{KD}}$, we report the signed cosine $\rho_{\text{KD}} = \cos(g_C, k)$ together with

$$S_{\text{KD}} = \sin \angle(g_C, k) = \frac{\|P_{g_C^\perp} k\|}{\|k\|}, \qquad M_{\text{KD}} = \frac{\lambda \|k\|}{\|g_C\|}, \qquad R_{\text{KD}} = M_{\text{KD}} S_{\text{KD}}. \tag{21}$$

S_{KD} does not depend on λ, whereas M_{KD} and R_{KD} measure the actual weighted auxiliary update. The signed cosine separates reinforcing gradients from oppositely parallel ones, which have the same zero sine.

Corollary 1 (illustrative soft-target KL case). Consider one-hot contrastive cross-entropy and teacher-to-student KL on the *same candidate logits and temperature*. Let π_T denote the teacher soft targets, e_i the one-hot positive target, and $J = \partial z/\partial \theta$ the shared student-logit Jacobian. Then

$$\|d_\perp\| \leq \lambda \|J\|_{\text{op}} \|e_i - \pi_T\|_2 \leq 2\lambda \|J\|_{\text{op}} (1 - \pi_{T,i}). \tag{22}$$

The bound holds because the two logit gradients, $\pi_S - e_i$ and $\pi_S - \pi_T$, differ by $e_i - \pi_T$ (Appendix A.5). A confident teacher therefore makes this KD gradient nearly parallel to the contrastive one. The corollary is illustrative only: it does not cover MSE, RKD, HieRD, or topology-based KD, whose gradient alignment must be measured directly.

4 Experiments

4.1 Setup

Training data and evaluation. We train CLS and VQA as separate models on unions of five and six in-distribution datasets, respectively, and evaluate each query only against the candidate pool of its dataset. The primary metric is Precision@1 (%). Each table entry is the mean±standard deviation over seeds 42, 43, and 44. IND, OOD, and ALL are unweighted dataset averages; ALL contains ten datasets for each task family.

Table 1: CLS Precision@1 (%), mean±std over seeds 42–44. KD rows use B3-Qwen2-2B; Contrastive only uses no teacher. IND, OOD, and ALL are unweighted means over datasets.

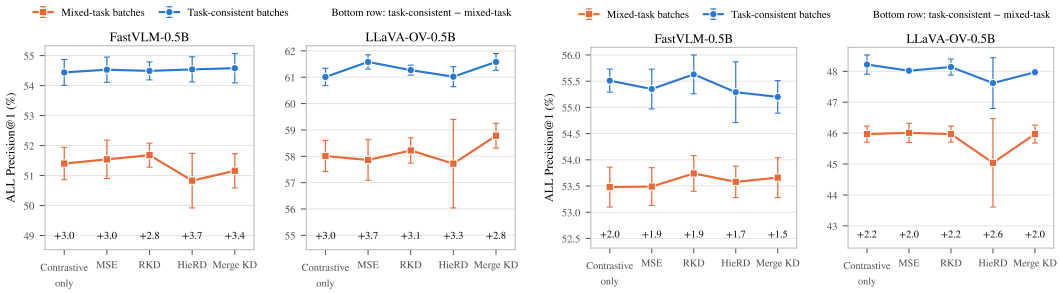
	FastVLM-0.5B			LLaVA-OV-0.5B		
Objective	IND	OOD	ALL	IND	OOD	ALL
Mixed-task batches						
Contrastive only	70.96 ±0.14	31.84 ±1.10	51.40 ±0.54	74.05 ±0.18	41.96 ±1.03	58.01 ±0.59
MSE	71.23 ±0.23	31.85 ±1.29	51.54 ±0.64	73.90 ±0.40	41.81 ±1.21	57.86 ±0.77
RKD	71.17 ±0.29	32.19 ±0.96	51.68 ±0.40	74.35 ±0.29	42.10 ±1.00	58.22 ±0.48
HieRD	70.72 ±0.02	30.94 ±1.84	50.83 ±0.91	73.23 ±1.48	42.21 ±1.98	57.72 ±1.68
Merge KD	71.05 ±0.47	31.26 ±1.49	51.16 ±0.57	74.23 ±0.16	43.32 ±0.83	58.78 ±0.47
Task-consistent batches						
Contrastive only	72.45 ±0.26	36.43 ±0.83	54.44 ±0.43	76.52 ±0.36	45.51 ±0.39	61.01 ±0.33
MSE	72.55 ±0.21	36.51 ±0.84	54.53 ±0.42	76.50 ±0.13	**46.66** ±0.64	**61.58** ±0.27
RKD	72.61 ±0.30	36.37 ±0.69	54.49 ±0.30	76.29 ±0.19	46.25 ±0.51	61.27 ±0.19
HieRD	**72.81** ±0.25	36.26 ±0.75	54.54 ±0.42	75.77 ±0.65	46.27 ±0.75	61.02 ±0.38
Merge KD	72.26 ±0.31	**36.89** ±0.79	**54.58** ±0.49	**76.59** ±0.19	46.56 ±0.45	**61.58** ±0.32
Task-consistent hard batches (B3)						
Contrastive only	71.96 ±0.21	35.46 ±0.97	53.71 ±0.49	76.72 ±0.28	45.14 ±0.78	60.93 ±0.48
MSE	72.30 ±0.22	35.40 ±1.09	53.85 ±0.54	76.02 ±0.30	44.90 ±0.97	60.46 ±0.58
RKD	71.88 ±0.30	35.34 ±0.84	53.61 ±0.35	75.64 ±0.25	43.82 ±0.79	59.73 ±0.37
HieRD	72.04 ±0.18	35.40 ±1.41	53.72 ±0.71	76.42 ±1.14	44.00 ±1.50	60.21 ±1.22
Merge KD	71.46 ±0.40	35.74 ±1.19	53.60 ±0.53	76.52 ±0.18	44.00 ±0.67	60.26 ±0.40

Table 2: VQA Precision@1 (%), mean±std over seeds 42–44. KD rows use B3-Qwen2-2B; Contrastive only uses no teacher. IND, OOD, and ALL are unweighted means over datasets.

	FastVLM-0.5B			LLaVA-OV-0.5B		
Objective	IND	OOD	ALL	IND	OOD	ALL
Mixed-task batches						
Contrastive only	53.79 ±0.35	53.01 ±0.54	53.48 ±0.38	44.49 ±0.23	48.19 ±0.34	45.97 ±0.26
MSE	53.86 ±0.29	52.94 ±0.47	53.49 ±0.36	44.57 ±0.48	48.17 ±0.09	46.01 ±0.31
RKD	54.07 ±0.43	53.24 ±0.23	53.74 ±0.34	44.45 ±0.22	48.25 ±0.74	45.97 ±0.26
HieRD	53.89 ±0.15	53.12 ±0.51	53.58 ±0.30	43.69 ±1.19	47.06 ±1.80	45.04 ±1.43
Merge KD	53.96 ±0.48	53.21 ±0.26	53.66 ±0.38	44.55 ±0.22	48.11 ±0.39	45.97 ±0.29
Task-consistent batches						
Contrastive only	55.94 ±0.26	54.85 ±0.42	55.51 ±0.22	46.89 ±0.22	**50.20** ±0.49	**48.22** ±0.31
MSE	55.82 ±0.35	54.64 ±0.77	55.35 ±0.38	46.87 ±0.03	49.74 ±0.22	48.02 ±0.08
RKD	**56.11** ±0.20	**54.92** ±0.62	**55.63** ±0.37	**47.02** ±0.30	49.82 ±0.20	48.14 ±0.26
HieRD	55.63 ±0.57	54.77 ±0.79	55.29 ±0.58	46.14 ±0.84	49.85 ±0.88	47.62 ±0.82
Merge KD	55.71 ±0.46	54.44 ±0.08	55.20 ±0.31	46.86 ±0.23	49.64 ±0.13	47.97 ±0.09
Task-consistent hard batches (B3)						
Contrastive only	26.35 ±0.31	36.25 ±0.48	30.31 ±0.31	47.35 ±0.23	50.45 ±0.42	48.59 ±0.29
MSE	25.70 ±0.32	35.68 ±0.64	29.69 ±0.37	47.30 ±0.34	50.55 ±0.17	48.60 ±0.23
RKD	26.50 ±0.34	36.55 ±0.47	30.52 ±0.36	46.98 ±0.26	50.58 ±0.54	48.42 ±0.26
HieRD	22.22 ±0.42	25.88 ±0.66	23.68 ±0.46	47.33 ±1.03	50.75 ±1.42	48.70 ±1.17
Merge KD	24.30 ±0.47	34.30 ±0.19	28.30 ±0.35	47.10 ±0.23	50.95 ±0.29	48.64 ±0.21

Implementation details. All experiments run on four NVIDIA H200 GPUs. We train FastVLM-0.5B and LLaVA-OneVision-0.5B with a global batch size of 128 (32 per GPU). The mixed-task sampler shuffles the concatenated training sets, whereas the task-consistent sampler constructs each global batch from one dataset. A third condition uses B3 within-task hard-negative batches with cluster size 8 at the same global batch size.

Baselines. We compare contrastive-only training with four auxiliary distillation objectives, each added to the same contrastive loss: embedding matching with an MSE loss (Romero et al., 2015; Kim et al., 2023), relational knowledge distillation (RKD; Park et al., 2019), hierarchical relational



(a) CLS (b) VQA

Figure 3: **The sampler moves the whole curve; the objective barely moves along it.** ALL Precision@1 of each objective under mixed-task and task-consistent batches for (a) CLS and (b) VQA, averaged over seeds 42–44 with ± 1 std whiskers. The bottom row of each panel gives the task-consistent minus mixed-task gap.

distillation (HieRD; Le et al., 2026), and Merge KD, a correspondence-aware structural objective related to topological distillation (Kim et al., 2024). All KD runs use the B3-Qwen2-2B teacher (Thirukovalluru et al., 2025), whereas contrastive-only runs use no teacher. Crossing every objective with the sampling conditions allows objective-wise comparisons under a shared contrastive task loss.

4.2 Main Results

CLS: the sampler, not the KD loss, drives the gain. Table 1 shows that task-consistent batching improves ALL Precision@1 for every objective by 2.80–3.72 points across both students, while under task-consistent batching, auxiliary distillation adds at most 0.57 points over contrastive-only training.

VQA: task consistency is robust; hardness is not. Table 2 shows the same sampler advantage for every objective: 1.54–2.03 ALL points for FastVLM and 2.00–2.58 for LLaVA-OneVision, while under task-consistent batching, auxiliary distillation adds at most 0.12 points and often reduces performance.

Hard negatives expose a student-dependent failure mode. Relative to random task-consistent batching, the available B3 hard-batch runs reduce CLS ALL by 0.68–0.98 points for FastVLM and 0.08–1.54 for LLaVA-OneVision. On VQA, they improve LLaVA-OneVision by 0.28–1.08 points but degrade FastVLM by 25.11–31.61 across objectives. Task consistency therefore transfers across students, whereas additional hardness does not.

The loss moves a student along a curve; the sampler moves the curve. Figure 3 shows that changing the distillation objective shifts ALL Precision@1 by at most 1.06 points within a sampler, whereas task-consistent batching lifts the entire curve by 1.54–3.72 points. Across all four student–task panels, the weakest task-consistent run still beats the strongest mixed run by 1.46–2.76 points. Thus, no auxiliary loss recovers under mixed batching what plain task-consistent batching provides, motivating the probe in Section 4.3.

4.3 Ablation Studies

Mixed batches slow within-task descent. Figure 4 compares coupled mixed and task-consistent rebatchings while holding each anchor–positive pair fixed. From step 2893 onward, D_{full} is negative for all five datasets; at the final checkpoint, four of five unadjusted intervals exclude zero, and equal-length mixed updates leave a higher reference loss on all five datasets.

The gap points to lost same-task coverage. Figure 5 decomposes the mixed–task-consistent difference along the selected repulsion-only path. Same-task coverage accounts for most of the signed deficit on N24News, SUN397, and VOC2007, while cross-task repulsion becomes small after ini-

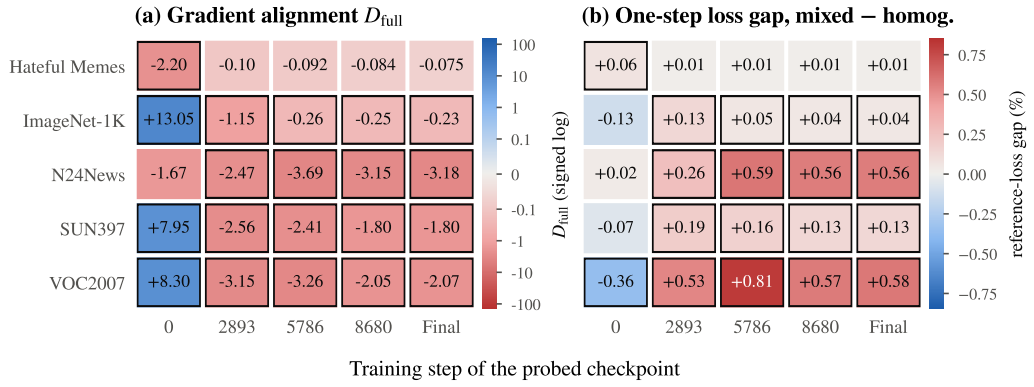


Figure 4: **Mixed batches make less within-task progress after initialization.** (a) Negative D_{full} predicts less first-order descent under the mixed update. (b) Positive mixed-minus-homogeneous gaps mean that an equal-length mixed step leaves a larger reference loss.

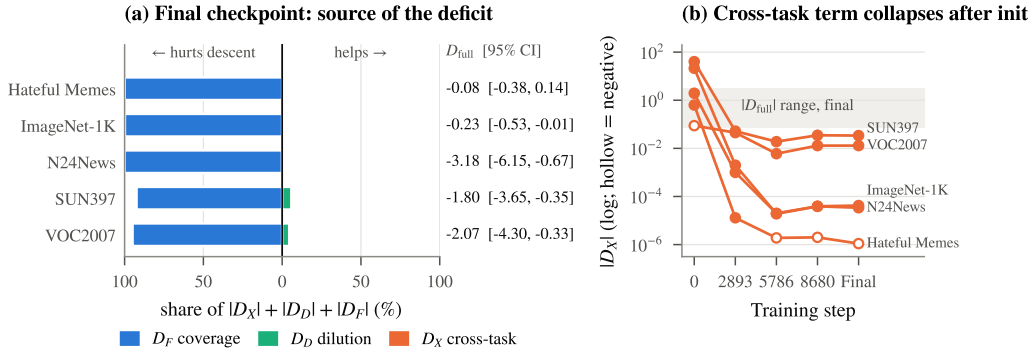


Figure 5: **Exploratory decomposition of the seed-42 probe.** (a) Signed shares of same-task coverage (D_F), dilution (D_D), and cross-task repulsion (D_X). (b) The $|D_X|$ trajectory. These path-dependent shares are descriptive rather than independent causal effects.

tialization. Because this path-dependent probe covers one student, task family, and seed, we treat it as descriptive rather than causal evidence.

5 Conclusion

Across two compact students, two task families, five objectives, and three seeds, task-consistent batches improve aggregate Precision@1 by 1.54–3.72 points, while auxiliary distillation changes the task-consistent score by only -0.60 to $+0.57$. Negative-source composition is therefore a first-order experimental axis that must be controlled when comparing distillation objectives. The exploratory decomposition suggests that lost same-task coverage is a major contributor along the selected decomposition path. However, the repulsion-only allocation does not rule out antagonistic contributions under alternative gradient allocations.

AI Use Statement

Generative AI tools were used only to identify and summarize related literature and to improve the clarity and readability of the manuscript. They were not used to formulate research hypotheses, develop theoretical results or write proofs, design the methodology or experiments, write or debug code, process datasets, run analyses, or interpret experimental results. All AI-assisted material was

reviewed by the authors, who take full responsibility for the theoretical claims, experimental results, code, and written content of this work.

Reproducibility Statement

Section 4 specifies the students, teachers, task families, training objectives, samplers, seeds, and the per-dataset evaluation protocol. Section 3 states the assumptions of each analytical result, and Appendix A gives the complete proofs. Appendix B reports the full per-dataset results, teacher-sensitivity controls, and their implementation details. Upon acceptance, we will release the training and evaluation code, configuration files, and instructions needed to reproduce all reported experiments.

Ethics Statement

This work studies knowledge distillation for vision–language embedding models using existing public benchmark datasets and does not collect new human-subject data. As with other distillation approaches, the student may inherit biases present in the teacher representations and training data. We therefore encourage further evaluation of fairness and robustness before deployment in sensitive applications.

References

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra. Investigating the role of negatives in contrastive representation learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 7187–7209. PMLR, 2022. URL https://proceedings.mlr.press/v151/ash22a.html.

Yoshua Bengio and Jean-Sébastien Senécal. Quick training of probabilistic neural nets by importance sampling. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume R4 of *Proceedings of Machine Learning Research*, pp. 17–24, 2003. URL https://proceedings.mlr.press/r4/bengio03a.html.

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10925–10934, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Beyer_Knowledge_Distillation_A_Good_Teacher_Is_Patient_and_Consistent_CVPR_2022_paper.html.

Miguel Romero Calvo, Shuoyang Ding, Corey D. Barrett, Georgiana Dinu, and George Karypis. Beyond instruction-conditioning, MoTE: Mixture of task experts for multi-task embedding models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22731–22746, 2025. doi: 10.18653/v1/2025.findings-acl.1168.

HanYeong Cho, Changwoo Kim, Taeuk Chu, and Jimin Park. Eddy-VL 1.9B: Structural pruning and layered distillation for edge-deployable multimodal embedding. *arXiv preprint arXiv:2607.16316*, 2026. doi: 10.48550/arXiv.2607.16316.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33: 8765–8775, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/63c3ddcc7b23daa1e42dc41f9a44a873-Abstract.html.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 297–304. PMLR, 2010. URL https://proceedings.mlr.press/v9/gutmann10a.html.

Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, Haofen Wang, Jun Yu, and Min Zhang. KaLM-Embedding: Superior training data brings a stronger embedding model. *arXiv preprint arXiv:2501.01028*, 2025. doi: 10.48550/arXiv.2501.01028.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TE0KOzWYAF.

Jungeun Kim, Junwon You, Dongjin Lee, Ha Young Kim, and Jae-Hun Jung. Do topological characteristics help in knowledge distillation? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24674–24693. PMLR, 2024. URL https://proceedings.mlr.press/v235/kim24aj.html.

Seungyeon Kim, Ankit Singh Rawat, Manzil Zaheer, Sadeep Jayasumana, Veeranjaneyulu Sadhanala, Wittawat Jitkrittum, Aditya Krishna Menon, Rob Fergus, and Sanjiv Kumar. EmbedDistill: A geometric knowledge distillation for information retrieval. *arXiv preprint arXiv:2301.12005*, 2023. doi: 10.48550/arXiv.2301.12005.

Vitaly Kurin, Alessandro De Palma, Ilya Kostrikov, Shimon Whiteson, and M. Pawan Kumar. In defense of the unitary scalarization for deep multi-task learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/4f301ae934f396086bfefd1139039dbd-Abstract-Conference.html.

Vinh Le, Nguyen Hong Dang, Tu Vu, Linh Ngo Van, Duc Anh Nguyen, and Trung Le. Hierd: Hierarchical relational distillation for vision-language embedding models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://openreview.net/forum?id=ZyqAd4xpT5.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=lgsyLSsDRe.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=zKv8qULV6n.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 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. doi: 10.48550/arXiv.2601.04720.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. doi: 10.48550/arXiv.2308.03281.

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*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6d5d6afa9957cfc9142ba60e78a467e9-Abstract-Conference.html.

Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3698–3707. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1405.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. VLM2Vec-V2: Advancing multimodal embedding for videos, images, and visual documents. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=TpU38jbKIJ. arXiv:2507.04590.

Luke Merrick, Danmei Xu, Gaurav Nuti, and Daniel Campos. Arctic-embed: Scalable, efficient, and accurate text embedding models. *arXiv preprint arXiv:2405.05374*, 2024. doi: 10.48550/arXiv.2405.05374.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. doi: 10.48550/arXiv.1807.03748.

Ovis-Embedding Team. Ovis-Embedding: Pushing the frontiers of universal omni-modal embeddings. *arXiv preprint arXiv:2609.25165*, 2026. doi: 10.48550/arXiv.2609.25165.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971.

IEEE, 2019. doi: 10.1109/CVPR.2019.00409.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=CR1XOQ0UTh-.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6550.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/376c6b9ff3bedbbea56751a84fffc10c-Abstract.html.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1102–1121, 2023. doi: 10.18653/v1/2023.findings-acl.71.

Raghuveer Thirukovalluru, Rui Meng, Ye Liu, Karthikeyan K, Mingyi Su, Ping Nie, Semih Yavuz, Yingbo Zhou, Wenhu Chen, and Bhuwan Dhingra. Breaking the batch barrier (b3) of contrastive learning via smart batch mining. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://neurips.cc/virtual/2025/poster/118493.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkgpBJrtvS.

Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. MobileCLIP: Fast image-text models through multi-modal reinforced training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15963–15974, 2024. doi: 10.1109/cvpr52733.2024.01511.

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, and Hadi Pouransari. FastVLM: Efficient vision encoding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19769–19780, 2025. doi: 10.1109/cvpr52734.2025.01841.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, volume 119, pp. 9929–9939. PMLR, 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=F1vEjWK-lH_.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. CLIP-KD: An empirical study of CLIP model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15952–15962, 2024. doi: 10.1109/cvpr52733.2024.01510.

Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, 2019. doi: 10.1145/3298689.3346996.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9274–9285, 2025. doi: 10.1109/cvpr52734.2025.00866.

Junjie Zhou, Ke Mei, Lei Li, Tianyi Wang, Fengyun Rao, and Jing Lyu. WeMM-Embedding: WeChat multi-modal embedding technical report. *arXiv preprint arXiv:2608.24053*, 2026. doi: 10.48550/arXiv.2608.24053.

A THEORETICAL PROOFS

These statements concern the *student's* contrastive learning objective, not the teacher's training history. They do not imply universal superiority of homogeneous sampling or universal redundancy of auxiliary distillation losses.

A.1 PROOF OF PROPOSITION 1: POPULATION RANKING INFONCE

Assumptions. Fix dataset d and a query x. Assume an idealized candidate space Ω with a common dominating measure such that the positive conditional $p_d(y \mid x)$ and each negative density under comparison are *strictly positive almost everywhere on* Ω and zero outside it. Consider fixed $K \geq 1$; draw K negatives independently from $r(y)$ and independently of x, draw one positive from $p_d(y \mid x)$, and uniformly randomize the positive's position I in the $K+1$ candidate tuple $Y = (y_0, \ldots, y_K)$. Allow any finite measurable scoring function $s(x, y)$. These conditions exclude zero-probability Bayes classes that require infinite optimal logits; they ensure identification on Ω, not outside it.

Bayes posterior and necessity. The tuple-generation density conditional on the positive index is

$$\Pr(I = i, Y = y \mid x) = \frac{1}{K+1} p_d(y_i \mid x) \prod_{j \neq i} r(y_j). \tag{23}$$

Using $r(y_i) > 0$ and canceling the common $\prod_j r(y_j)$ gives

$$Q(i \mid x, Y) := \Pr(I = i \mid x, Y) = \frac{p_d(y_i \mid x)/r(y_i)}{\sum_{m=0}^{K} p_d(y_m \mid x)/r(y_m)}. \tag{24}$$

Let $\widehat{Q}_s$ be the model softmax posterior formed from logits $s(x, y_i)$. The conditional cross-entropy obeys the exact identity

$$H(Q, \widehat{Q}_s) = H(Q) + \mathrm{KL}(Q \| \widehat{Q}_s). \tag{25}$$

The KL term vanishes if and only if $\widehat{Q}_s = Q$ for almost every candidate tuple. Since an unrestricted finite scorer $s(x, y) = \log p_d(y \mid x) - \log r(y)$ realizes the Bayes posterior, the population minimum is attained and *every* population minimizer must realize the Bayes posterior almost everywhere.

Identification up to a query-only constant. Equality of two softmax distributions implies that their logits differ by a constant *on each observed tuple*. To promote this to a single candidate-independent constant on Ω, consider an arbitrary pair of candidate values (y, y') drawn from $r \otimes r$. Strict positivity of $p_d(\cdot \mid x)$ relative to r ensures the corresponding pair can occur as the positive and a negative even when $K = 1$. Thus almost every pair is linked by a candidate tuple of positive sampling density. Bayes-posterior equality on such tuples forces

$$s(x, y) - s(x, y') = \log \frac{p_d(y \mid x)}{r(y)} - \log \frac{p_d(y' \mid x)}{r(y')} \quad \text{for } (r \otimes r)\text{-a.e. } (y, y'). \tag{26}$$

Equivalently, the measurable function $s(x, y) - \log p_d(y \mid x) + \log r(y)$ has equal values for $(r \otimes r)$-almost every pair and is therefore almost surely constant in y. Denoting that constant by $C_r(x)$ establishes

$$s_r^*(x, y) = \log p_d(y \mid x) - \log r(y) + C_r(x) \qquad \text{for } r\text{-almost every } y \in \Omega. \tag{27}$$

Comparing optimum sets. Apply the characterization to task-specific r_d and mixture r_{mix}. On their common support, every pair of minimizers satisfies

$$s_{\mathrm{mix}}^*(x, y) - s_d^*(x, y) = \log \frac{r_d(y)}{r_{\mathrm{mix}}(y)} + C'(x). \tag{28}$$

Their optimum sets intersect if and only if the log density ratio is constant almost everywhere in y on this shared support. Under the common support and normalized-density assumptions, a constant ratio across all of Ω implies the distributions are identical. If practical datasets have different supports, these statements only identify relative scores where the corresponding density ratios are well defined; no global optimum-set assertion is justified without additional assumptions.

Duplicate candidates and modeling limitations. Repeated candidate *values* do not invalidate the posterior: identical candidates receive identical model logits and identical Bayes odds, so the posterior splits its probability symmetrically among tied positions. Such ties can impose an irreducible cross-entropy floor, but do not alter the population optimum characterization under the positive-density assumptions. By contrast, correlated in-batch negatives, finite temperature, normalized bounded similarities, and a low-dimensional bilinear student score violate the unrestricted idealization. A changed population target need not imply different constrained embedding optima or improved downstream accuracy. Notably, if r_{mix} is proportional to r_d over the relevant within-dataset candidates, their relative target scores there agree; finite-batch same-task coverage, analyzed next, can still differ strongly.

Conditional link to evaluation ranking. If evaluation *also* draws one positive from $p_d(y \mid x)$ and K_{eval} independent distractors from $r_{\mathrm{eval}}(y)$, the same Bayes-posterior argument gives $s^*_{\mathrm{eval}} = \log p_d(y \mid x) - \log r_{\mathrm{eval}}(y) + C(x)$. Only with the additional assumptions of matched train/evaluation positive conditionals and $r_{\mathrm{eval}} = r_d$ can the homogeneous score be identified as evaluation-optimal in this *random-candidate* model. A fixed classification label bank admits a *separate*, conditional analysis; it is not an i.i.d. distractor sample. Suppose the bank $\mathcal{Y}_d$ contains each eligible label exactly once, a correct prediction is its highest-scored label, the test positive conditional on the bank agrees with $p_d(y \mid x)$, and all candidate densities under consideration are positive there. The Bayes top-1 decision on that fixed bank is $\arg\max_{y \in \mathcal{Y}_d} p_d(y \mid x)$. For any two labels y, y' in the bank, the idealized InfoNCE score margin is

$$s^*_r(x, y) - s^*_r(x, y') = \log \frac{p_d(y \mid x)}{p_d(y' \mid x)} - \log \frac{r(y)}{r(y')}. \tag{29}$$

Thus if r_d is uniform on the label bank (an *empirical assumption* requiring a check of training candidate frequencies), the homogeneous optimum has the Bayes-optimal *ranking* for this idealized fixed-bank task. If r_{mix} is also uniform on the bank, its ranking is identical. Conversely, nonuniform mixed candidate mass can change some score margins, and one can construct positive conditionals for which such nonuniformity reverses a pairwise ranking; it does *not* follow that a reversal occurs for the actual p_d, or that top-1 accuracy must decrease. A useful candidate-level diagnostic is the dispersion of $\log r_{\mathrm{mix}}(y)$ over $y \in \mathcal{Y}_d$, compared with the conditional log-odds margins in Eq. (29). If other datasets have disjoint candidate vocabularies, then $r_{\mathrm{mix}}(y) = \pi_d r_d(y)$ on the focal bank, and the population margin shift vanishes there even though mixed batches still lose within-task comparison slots.

A conservative sufficient condition for fixed-bank top-1 stability. Under the preceding additional assumption that r_d is uniform on the fixed bank, suppose the Bayes-optimal label $y^\star(x) = \arg\max_{y \in \mathcal{Y}_d} p_d(y \mid x)$ is unique almost surely. Define its smallest positive conditional log-odds margin and the mixed marginal's log-density oscillation by

$$M_d(x) = \min_{y \in \mathcal{Y}_d \setminus \{y^\star\}} \log \frac{p_d(y^\star \mid x)}{p_d(y \mid x)}, \qquad \Delta_d = \sup_{y, y' \in \mathcal{Y}_d} \left| \log r_{\mathrm{mix}}(y) - \log r_{\mathrm{mix}}(y') \right|. \tag{30}$$

Both are finite under the positive finite-bank assumptions. A tighter query-conditioned *signed* distortion compares the Bayes-best label only with its competitors:

$$\Delta_d(x) = \max_{y \in \mathcal{Y}_d \setminus \{y^\star(x)\}} \log \frac{r_{\mathrm{mix}}(y^\star(x))}{r_{\mathrm{mix}}(y)} = \log r_{\mathrm{mix}}(y^\star(x)) - \min_{y \neq y^\star(x)} \log r_{\mathrm{mix}}(y) \leq \Delta_d. \tag{31}$$

Crucially, the minimum excludes $y^\star$. Including it would force $\Delta_d(x) \geq 0$ and erase the useful case where the Bayes-best label has *lower* mixed candidate mass than every competitor. By Eq. (29), for every competing y,

$$\begin{aligned} s^*_{\mathrm{mix}}(x, y^\star) - s^*_{\mathrm{mix}}(x, y) &= \log \frac{p_d(y^\star \mid x)}{p_d(y \mid x)} - \log \frac{r_{\mathrm{mix}}(y^\star)}{r_{\mathrm{mix}}(y)} \\ &\geq M_d(x) - \Delta_d(x). \end{aligned} \tag{32}$$

Therefore $M_d(x) > \Delta_d(x)$ suffices to preserve the unique Bayes-best top-1 decision. In particular, $M_d(x) > 0$, so $\Delta_d(x) < 0$ ensures stability (and $\Delta_d(x) = 0$ does too). The probability bound becomes

$$\Pr_x[\text{mixed-population top-1 differs from Bayes top-1}] \leq \Pr_x[M_d(x) \leq \Delta_d(x)] \leq \Pr_x[M_d(x) \leq \Delta_d]. \tag{33}$$

The query-specific bound is still conservative: the most distorted competitor and the smallest Bayes margin need not be the same label. It becomes uninformative when margins are too small or ties arise. In particular, if other datasets' candidate vocabularies are disjoint and r_d is uniform on this bank, $r_{\text{mix}} = \pi_d r_d$ there, $\Delta_d = 0$, and no *population* top-1 flip is induced by mixing in this idealized setting; finite-batch coverage effects can still occur. Outside uniform r_d, disjoint vocabularies still preserve the *relative* homogeneous-versus-mixed scores on the focal bank, but neither is necessarily Bayes optimal. Crucially, the true $p_d(y \mid x)$ is not supplied by the training candidate counts or by the learned model. An empirical substitute for $M_d(x)$ estimated from calibrated held-out predictions is a *proxy*, not a verified numerical value of the population bound. The argument also cannot be transferred unchanged to OOD evaluation with shifted positive conditionals.

These conclusions additionally depend on an unrestricted realizable scorer and matched train/test positive conditionals; neither is established for our actual in-batch-trained models and OOD tasks. They are testable predictions under stated conditions, not explanations of the measured accuracy gains without a candidate-frequency audit. □

A.2 Proof of Proposition 2: the realized reference-loss effect

This proposition uses a *query embedding* q as its optimization variable; the actual model parameters and the associated Jacobian are held fixed in the embedding intervention. Fix a realized, paired set of anchor and companion draws and put $g_{\text{ref}} = \nabla_q L_{\text{ref}}(q)$ and $D_{\text{full}} = \langle g_{\text{ref}}, g_{\text{mix}} - g_{\text{hom}} \rangle$. If $\nabla_q L_{\text{ref}}$ is L-Lipschitz on a convex neighborhood containing both steps, the two-sided smoothness bound yields, for each $g \in \{g_{\text{mix}}, g_{\text{hom}}\}$,

$$|L_{\text{ref}}(q - \eta g) - L_{\text{ref}}(q) + \eta \langle g_{\text{ref}}, g \rangle| \le \frac{L\eta^2}{2} \|g\|^2. \tag{34}$$

Subtracting the two expansions gives

$$L_{\text{ref}}(q - \eta g_{\text{mix}}) - L_{\text{ref}}(q - \eta g_{\text{hom}}) = -\eta D_{\text{full}} + R_\eta, \tag{35}$$

with

$$|R_\eta| \le \frac{L\eta^2}{2} (\|g_{\text{mix}}\|^2 + \|g_{\text{hom}}\|^2). \tag{36}$$

For $D_{\text{full}} < 0$, the leading term dominates for step sizes satisfying Eq. (9). The result is deterministic conditional on the realized draw; D_{full} is itself random over anchor and companion sampling. Practical uncertainty estimation should preserve the mixed/homogeneous pairing and resample independent anchor sets rather than treat correlated negative draws as independent experiments. The same algebra applies in parameter space if *all three* gradients are consistently taken with respect to θ, but that is a distinct measured statistic. For equal-length embedding interventions the directions are normalized and their loss difference is measured separately. □

A.3 Exact path-dependent decomposition using repulsion-only cross-task terms

The full row-wise InfoNCE gradient has the equivalent expressions

$$g_{\text{mix},i} = \sum_j p_{ij}^{\text{mix}} h_{ij} - u_i = \sum_{j \ne i} p_{ij}^{\text{mix}} (h_{ij} - u_i), \qquad h_{ij} = \nabla_{q_i} z_{ij}, \quad u_i = h_{ii}. \tag{37}$$

Although the second expression distributes positive attraction across negative pairs, the *archived probe implementation* uses the first: it assigns the whole $-u_i$ to the restricted/same-task component and uses only the repulsive candidate gradients for the cross-task component. Keep the same anchor, positive, logits, and restricted set $R = A \cup S$ (including i) throughout. Define

$$\begin{aligned} g_{X,i}^{\text{rep}} &= \sum_{j \in X} p_{ij}^{\text{mix}} h_{ij}, \\ g_{\text{mix},i}^{R,\text{code}} &= \sum_{j \in R} p_{ij}^{\text{mix}} h_{ij} - u_i, \\ g_{\text{rest},i} &= \sum_{j \in R} p_{ij}^{\text{rest}} h_{ij} - u_i. \end{aligned} \tag{38}$$

With the homogeneous batch gradient $g_{\mathrm{hom},i}$,

$$\delta_{X,i} = g^{\mathrm{rep}}_{X,i}, \qquad \delta_{D,i} = g^{R,\mathrm{code}}_{\mathrm{mix},i} - g_{\mathrm{rest},i}, \qquad \delta_{F,i} = g_{\mathrm{rest},i} - g_{\mathrm{hom},i}. \tag{39}$$

Consequently,

$$\delta_{X,i} + \delta_{D,i} + \delta_{F,i} = g^{\mathrm{rep}}_{X,i} + g^{R,\mathrm{code}}_{\mathrm{mix},i} - g_{\mathrm{hom},i} \tag{40}$$

$$= g_{\mathrm{mix},i} - g_{\mathrm{hom},i}. \tag{41}$$

Taking the dot product with the same reference gradient establishes $D_{\mathrm{full},i} = D_{X,i} + D_{D,i} + D_{F,i}$, and summing/averaging over anchors preserves this identity under common weights. The restricted intermediate batch has fewer candidates, making this a path-dependent coverage decomposition rather than a unique causal attribution.

Exact dilution identity for the implemented convention. Define

$$Z_{R,i} = \sum_{j\in R} e^{z_{ij}}, \quad Z_{X,i} = \sum_{j\in X} e^{z_{ij}}, \quad \epsilon_i = Z_{X,i}/Z_{R,i}, \quad m_{X,i} = Z_{X,i}/(Z_{R,i}+Z_{X,i}) = \epsilon_i/(1+\epsilon_i). \tag{42}$$

Every $j \in R$ has $p^{\mathrm{mix}}_{ij} = (1 - m_{X,i})p^{\mathrm{rest}}_{ij}$. Set

$$\mu_{R,i} = \sum_{j\in R} p^{\mathrm{rest}}_{ij} h_{ij} = g_{\mathrm{rest},i} + u_i. \tag{43}$$

The implemented restricted component therefore satisfies

$$g^{R,\mathrm{code}}_{\mathrm{mix},i} = (1 - m_{X,i})\mu_{R,i} - u_i, \tag{44}$$

and subtracting $g_{\mathrm{rest},i} = \mu_{R,i} - u_i$ yields

$$\boxed{\delta_{D,i} = -m_{X,i}\mu_{R,i}} = -m_{X,i}(g_{\mathrm{rest},i} + u_i), \qquad \|\delta_{D,i}\| = m_{X,i}\|\mu_{R,i}\|, \tag{45}$$

$$D_{D,i} = -m_{X,i}\langle g_{\mathrm{ref},i}, \mu_{R,i}\rangle. \tag{46}$$

Thus $\delta_{D,i}$ is exactly anti-parallel to $\mu_{R,i}$ for a nondegenerate anchor with $m_{X,i} > 0$, *not* necessarily to the restricted loss gradient $g_{\mathrm{rest},i} = \mu_{R,i} - u_i$. Likewise, this is shrinkage of the candidate-weighted term, not a no-rotation claim about the full restricted loss gradient.

A numerically checkable per-anchor consequence (if the original probe logs are available) is

$$\|\delta_{D,i} + m_{X,i}\mu_{R,i}\| \approx 0, \qquad \cos(\delta_{D,i}, \mu_{R,i}) \approx -1 \tag{47}$$

for nonzero component norms. The latter collinearity does *not* transfer to averages over anchors when $m_{X,i}$ varies. The code's repulsion-only allocation is known; passing this numerical test in the actual archived probe remains unverified until its per-anchor values are available.

With $b_i = \langle g_{\mathrm{ref},i}, \mu_{R,i}\rangle$ and a single common averaging population and weight rule for all terms, the mean signed dilution projection obeys

$$-\mathbb{E}[D_{D,i}] = \mathbb{E}[m_X b] = \mathbb{E}[m_X]\,\mathbb{E}[b] + \mathrm{Cov}(m_X, b). \tag{48}$$

A positive mean D_D requires negative *mass-weighted* reference/candidate-gradient alignment; it need not indicate negative unweighted mean alignment of either μ_R or g_{rest}. The three summaries $\mathbb{E}[m_X]$, $\mathbb{E}[b]$, and $\mathrm{Cov}(m_X, b)$ distinguish negative average candidate alignment from preferential weighting of negatively aligned anchors. These quantities are not recoverable from component-share plots alone.

Dependence on positive-attraction allocation. For comparison only, if one instead allocated $-u_i$ proportionally across negatives via the per-negative-difference expression, then

$$\delta^{\mathrm{pair}}_{X,i} = \delta^{\mathrm{rep}}_{X,i} - m_{X,i}u_i, \qquad \delta^{\mathrm{pair}}_{D,i} = \delta^{\mathrm{rep}}_{D,i} + m_{X,i}u_i = -m_{X,i}g_{\mathrm{rest},i}. \tag{49}$$

The two conventions share $\delta_X + \delta_D$ and δ_F exactly. This transformation is a mathematical comparison, *not* a claim that archived experimental curves were recomputed under the second convention. In the *implemented* convention, the triangle inequality provides

$$\|\delta_{X,i}\| \le m_{X,i} \max_{j\in X} \|h_{ij}\|, \qquad \|\delta_{D,i}\| = m_{X,i}\|\mu_{R,i}\|. \tag{50}$$

Small cross-task mass implies both components are small when the candidate derivatives are bounded; small measured D_X alone does not establish small mass because weighted gradients may cancel. □

A.4 Proof of Lemma 1: auxiliary-gradient geometry

Assume $g_C \neq 0$ and define $u = g_C/\|g_C\|$. Write $d = au + d_\perp$ with $a = \langle d, u\rangle$ and $\langle d_\perp, u\rangle = 0$. Then

$$g_{\text{tot}} = (\|g_C\| + a)u + d_\perp. \tag{51}$$

If $\|g_C\| + a > 0$, the angle to g_C is acute and

$$\tan\phi = \frac{\|d_\perp\|}{\|g_C\| + a}. \tag{52}$$

If $\|g_C\| + a < 0$, the total step points into the opposite half-space even when $d_\perp = 0$; if it equals zero, angle computation requires the remaining orthogonal component to be nonzero. The lemma is local and does not constrain accumulation of orthogonal components over multiple training steps. □

A.5 Proof of Corollary 1: illustrative soft-target bound

This illustrative bound is *not* used to infer gradient novelty for MSE, RKD, HieRD, or the topology objective in the actual experiments. Only in this subsection assume the auxiliary objective is teacher-to-student KL on exactly the same student logits z, candidate set, and temperature as the contrastive cross-entropy, and assume equal per-row normalization in both loss terms. Denote the student candidate distribution by π_S, the teacher distribution by π_T, the one-hot true-positive distribution by e_i, and the shared Jacobian by $J = \partial z/\partial\theta$. For one anchor,

$$\nabla_z L_{\text{con}} = \pi_S - e_i, \qquad \nabla_z L_{\text{KD}} = \pi_S - \pi_T. \tag{53}$$

Hence $g_C = J^\top(\pi_S - e_i)$ and $d = \lambda J^\top(\pi_S - \pi_T)$. Let $u = g_C/\|g_C\|$. Orthogonal projection minimizes distance to the one-dimensional span of u, so

$$\|d_\perp\| = \min_{b\in\mathbb{R}} \|d - bu\| \le \|d - \lambda g_C\| \tag{54}$$

$$= \lambda\|J^\top(e_i - \pi_T)\| \tag{55}$$

$$\le \lambda\|J\|_{\text{op}}\|e_i - \pi_T\|_2 \tag{56}$$

$$\le 2\lambda\|J\|_{\text{op}}(1 - \pi_{T,i}). \tag{57}$$

The last step uses $\|v\|_2 \le \|v\|_1$ and $\|e_i - \pi_T\|_1 = 2(1 - \pi_{T,i})$. For multiple anchors with summed/averaged gradients, an analogous bound follows using a block Jacobian and a corresponding norm of the stacked target differences; it need not equal the single-anchor expression. This proof has no implication for generic structural KD, which need not share the same student logits or have teacher-soft-target gradients. □

B Additional Evaluation

Task consistency wins broadly, but not uniformly. Table 3 shows a positive task-consistent–minus–mixed difference on 18, 19, and 18 of the 20 datasets in the three model–teacher columns. The few regressions are small and localized, the largest being -1.31 points for FastVLM on Hateful Memes, whereas the largest effect in the table is a $+11.29$-point improvement for LLaVA-OneVision on ObjectNet. These objective-averaged values are descriptive summaries rather than paired uncertainty estimates.

Table 3: Per-dataset sampler effect: Task-consistent − Mixed-task Precision@1 points, averaged over the available objectives. B3-Qwen2 entries average objective-level three-seed means; †Qwen3 uses one seed and four objectives. These are descriptive summaries, not uncertainty estimates.

		B3-Qwen2-2B		Qwen3-VL-Emb-2B
Split	Dataset	FastVLM	LLaVA-OV	FastVLM†
CLS IND	IN-1K	+0.83	+2.25	+3.00
	N24News	+0.19	+0.11	−0.68
	Hateful Memes	−1.31	+2.68	+4.38
	VOC2007	+2.73	+1.84	+2.95
	SUN397	+5.10	+5.04	+4.75
CLS OOD	Places365	+6.39	+4.27	+3.15
	IN-A	+1.65	+1.21	+0.55
	IN-R	+6.43	+3.54	+7.60
	ObjectNet	+8.87	+11.29	+1.72
	Country211	+1.06	−0.48	−0.57
VQA IND	OK-VQA	+3.09	+3.09	+0.42
	A-OKVQA	+0.44	+3.02	+0.98
	DocVQA	+1.82	+3.70	+2.60
	InfoVQA	+2.04	+1.98	+0.67
	ChartQA	+2.09	+1.63	+0.40
	Visual7W	+2.07	+1.01	+0.20
VQA OOD	ScienceQA	+1.91	+1.09	+1.75
	VizWiz	+1.65	+2.18	+0.23
	GQA	−0.08	+1.52	+0.75
	TextVQA	+3.01	+2.79	+0.20

B.1 PER-DATASET BENCHMARK RESULTS

FastVLM classification gains are largest out of distribution. Table 4 shows that random task-consistent batching improves the OOD average for every objective by 4.18–5.63 points, compared with 1.21–2.09 points in distribution. Hard batches remain below random task-consistent batches on every OOD objective, so compatibility is more reliable than hardness for this student.

Table 4: Per-dataset classification Precision@1 (%), B3-Qwen2-2B → FastVLM-0.5B, mean±std over seeds 42–44. Avg is the unweighted mean over the panel's datasets.

(a) In-distribution

Objective	IN-1K	N24News	Hateful Memes	VOC2007	SUN397	**Avg**
Mixed-task batches						
Contrastive only	65.17 ±0.71	74.27 ±0.29	64.63 ±1.07	80.20 ±0.80	70.53 ±0.49	70.96 ±0.14
MSE	65.10 ±0.53	74.53 ±0.15	**65.47** ±1.25	80.07 ±0.29	70.97 ±0.70	71.23 ±0.23
RKD	64.90 ±0.17	74.73 ±0.06	65.20 ±1.11	80.20 ±0.36	70.80 ±0.26	71.17 ±0.29
HieRD	64.30 ±0.62	74.10 ±0.61	63.87 ±0.93	80.43 ±0.49	70.90 ±1.22	70.72 ±0.02
Merge KD	65.23 ±0.83	74.47 ±0.35	64.97 ±1.56	79.93 ±0.51	70.67 ±0.25	71.05 ±0.47
Task-consistent batches						
Contrastive only	65.63 ±0.70	74.50 ±0.60	62.50 ±1.47	**83.37** ±0.31	**76.23** ±0.68	72.45 ±0.26
MSE	65.67 ±0.46	74.60 ±0.62	63.63 ±0.93	82.93 ±0.67	75.90 ±0.10	72.55 ±0.21
RKD	65.60 ±0.10	74.40 ±0.56	64.07 ±1.44	82.97 ±0.42	76.00 ±0.52	72.61 ±0.30
HieRD	**66.10** ±0.70	**74.90** ±0.10	64.40 ±1.13	82.93 ±0.76	75.73 ±0.59	**72.81** ±0.25
Merge KD	65.87 ±0.76	74.67 ±0.29	62.97 ±0.93	82.30 ±0.44	75.50 ±0.46	72.26 ±0.31
Task-consistent hard batches (B3)						
Contrastive only	64.4 ±0.71	74.6 ±0.47	63.4 ±1.29	81.7 ±0.61	75.7 ±0.59	71.96 ±0.21
MSE	64.7 ±0.50	74.7 ±0.45	64.6 ±1.10	81.2 ±0.52	76.3 ±0.50	72.30 ±0.22
RKD	64.3 ±0.14	74.7 ±0.40	63.1 ±1.29	81.9 ±0.39	75.4 ±0.41	71.88 ±0.30
HieRD	65.1 ±0.66	74.5 ±0.44	63.0 ±1.03	82.4 ±0.64	75.2 ±0.96	72.04 ±0.18
Merge KD	63.6 ±0.80	73.8 ±0.32	62.1 ±1.28	82.0 ±0.48	75.8 ±0.37	71.46 ±0.40

(b) Out-of-distribution

Objective	Places365	IN-A	IN-R	ObjectNet	Country211	**Avg**
Mixed-task batches						
Contrastive only	31.53 ±1.55	28.33 ±1.00	57.47 ±2.55	35.13 ±2.28	6.73 ±0.40	31.84 ±1.10
MSE	31.77 ±1.63	28.10 ±0.80	57.33 ±3.50	35.33 ±1.84	6.70 ±0.70	31.85 ±1.29
RKD	32.13 ±1.31	27.90 ±0.40	58.03 ±3.45	35.70 ±1.28	7.17 ±0.86	32.19 ±0.96
HieRD	32.20 ±1.14	28.40 ±1.44	53.30 ±6.43	33.67 ±3.35	7.13 ±0.51	30.94 ±1.84
Merge KD	32.50 ±1.85	27.63 ±1.29	54.93 ±2.74	33.90 ±2.86	7.33 ±1.04	31.26 ±1.49
Task-consistent batches						
Contrastive only	38.70 ±0.26	29.47 ±1.00	62.90 ±2.23	43.33 ±1.46	7.77 ±0.47	36.43 ±0.83
MSE	38.30 ±0.17	29.83 ±0.42	**63.27** ±2.50	43.10 ±1.48	8.03 ±0.12	36.51 ±0.84
RKD	37.87 ±0.67	29.60 ±1.25	62.43 ±1.62	43.70 ±1.59	**8.27** ±0.40	36.37 ±0.69
HieRD	38.30 ±0.53	29.50 ±0.35	61.93 ±2.58	43.30 ±2.00	**8.27** ±0.32	36.26 ±0.75
Merge KD	**38.93** ±0.72	**30.20** ±0.75	62.67 ±2.57	**44.67** ±0.95	8.00 ±0.10	**36.89** ±0.79
Task-consistent hard batches (B3)						
Contrastive only	37.3 ±1.11	30.3 ±1.00	57.9 ±2.40	44.3 ±1.91	7.5 ±0.44	35.46 ±0.97
MSE	37.3 ±1.16	29.1 ±0.64	59.1 ±3.04	44.0 ±1.67	7.5 ±0.50	35.40 ±1.09
RKD	36.8 ±1.04	29.7 ±0.93	59.3 ±2.70	43.4 ±1.44	7.5 ±0.67	35.34 ±0.84
HieRD	38.1 ±0.89	29.2 ±1.05	57.3 ±4.90	44.6 ±2.76	7.8 ±0.43	35.40 ±1.41
Merge KD	37.4 ±1.40	30.0 ±1.06	59.1 ±2.66	45.2 ±2.13	7.0 ±0.74	35.74 ±1.19

FastVLM VQA exposes a sharp hard-negative failure mode. In Table 5, random task-consistent batches improve both IND and OOD averages for every objective. B3 hard batches instead reduce IND averages to 22.22–26.50 and OOD averages to 25.88–36.55, far below the corresponding random task-consistent ranges of 55.63–56.11 and 54.44–54.92.

Table 5: Per-dataset visual question answering Precision@1 (%), B3-Qwen2-2B → FastVLM-0.5B, mean±std over seeds 42–44. Avg is the unweighted mean over the panel's datasets.

(a) In-distribution

Objective	OK-VQA	A-OKVQA	DocVQA	InfoVQA	ChartQA	Visual7W	**Avg**
Mixed-task batches							
Contrastive only	55.63 ±0.64	53.47 ±0.32	77.13 ±0.25	34.97 ±0.45	52.43 ±0.21	49.13 ±1.25	53.79 ±0.35
MSE	55.30 ±0.17	53.13 ±0.51	77.17 ±0.12	35.57 ±0.50	52.73 ±0.32	49.23 ±1.19	53.86 ±0.29
RKD	55.50 ±0.35	53.40 ±0.26	77.37 ±0.32	35.60 ±1.13	52.87 ±0.59	49.67 ±1.11	54.07 ±0.43
HieRD	55.50 ±0.44	52.93 ±0.49	77.00 ±0.20	35.73 ±0.60	52.73 ±0.51	49.43 ±0.60	53.89 ±0.15
Merge KD	55.67 ±0.51	53.83 ±0.21	77.43 ±0.49	35.00 ±1.25	52.63 ±0.57	49.17 ±1.20	53.96 ±0.48
Task-consistent batches							
Contrastive only	58.43 ±0.65	**54.13** ±0.90	78.83 ±0.25	37.73 ±0.40	54.83 ±0.25	51.70 ±0.30	55.94 ±0.26
MSE	58.77 ±0.67	53.93 ±0.68	78.93 ±0.38	37.30 ±0.50	54.73 ±0.61	51.23 ±0.35	55.82 ±0.35
RKD	**59.20** ±0.56	53.73 ±0.49	79.13 ±0.32	**37.93** ±0.51	**54.90** ±0.70	**51.77** ±0.31	**56.11** ±0.20
HieRD	58.13 ±0.45	53.33 ±0.25	78.93 ±0.67	37.13 ±1.19	54.77 ±1.00	51.47 ±0.68	55.63 ±0.57
Merge KD	58.53 ±0.93	53.83 ±0.76	**79.40** ±0.70	37.00 ±0.79	54.63 ±0.49	50.83 ±1.07	55.71 ±0.46
Task-consistent hard batches (B3)							
Contrastive only	35.0 ±0.65	30.1 ±0.68	34.6 ±0.25	15.2 ±0.43	17.8 ±0.23	25.4 ±0.91	26.35 ±0.31
MSE	33.7 ±0.49	28.7 ±0.60	34.4 ±0.28	14.8 ±0.50	17.3 ±0.49	25.3 ±0.88	25.70 ±0.32
RKD	36.2 ±0.47	30.3 ±0.39	34.9 ±0.32	15.3 ±0.88	16.8 ±0.65	25.5 ±0.81	26.50 ±0.34
HieRD	30.9 ±0.45	26.8 ±0.39	31.7 ±0.49	12.0 ±0.94	14.4 ±0.79	17.5 ±0.64	22.22 ±0.42
Merge KD	34.9 ±0.75	29.4 ±0.56	27.9 ±0.60	13.4 ±1.05	15.3 ±0.53	24.9 ±1.14	24.30 ±0.47

(b) Out-of-distribution

Objective	ScienceQA	VizWiz	GQA	TextVQA	**Avg**
Mixed-task batches					
Contrastive only	38.17 ±1.22	41.37 ±0.31	63.33 ±0.68	69.17 ±0.55	53.01 ±0.54
MSE	38.00 ±0.82	40.93 ±0.23	63.53 ±0.55	69.30 ±0.56	52.94 ±0.47
RKD	39.13 ±0.80	40.97 ±0.21	63.70 ±1.25	69.17 ±0.46	53.24 ±0.23
HieRD	38.90 ±0.61	41.37 ±0.12	63.33 ±1.27	68.87 ±0.68	53.12 ±0.51
Merge KD	38.33 ±0.57	41.53 ±0.15	63.67 ±0.90	69.30 ±0.10	53.21 ±0.26
Task-consistent batches					
Contrastive only	40.57 ±1.22	42.77 ±0.65	63.53 ±0.95	**72.53** ±0.49	54.85 ±0.42
MSE	40.13 ±2.32	42.60 ±0.17	**63.80** ±0.95	72.03 ±1.10	54.64 ±0.77
RKD	40.60 ±2.29	43.03 ±0.15	63.63 ±0.68	72.40 ±0.52	**54.92** ±0.62
HieRD	**41.00** ±2.26	**43.30** ±0.87	63.07 ±0.42	71.73 ±0.55	54.77 ±0.79
Merge KD	39.77 ±1.36	42.70 ±0.66	63.13 ±1.88	72.17 ±0.74	54.44 ±0.08
Task-consistent hard batches (B3)					
Contrastive only	24.7 ±1.22	26.5 ±0.51	62.1 ±0.83	31.7 ±0.52	36.25 ±0.48
MSE	24.1 ±1.74	26.6 ±0.20	61.2 ±0.78	30.8 ±0.87	35.68 ±0.64
RKD	24.2 ±1.72	26.5 ±0.18	64.0 ±1.01	31.5 ±0.49	36.55 ±0.47
HieRD	15.0 ±1.66	13.6 ±0.62	46.3 ±0.95	28.6 ±0.62	25.88 ±0.66
Merge KD	21.3 ±1.04	25.2 ±0.48	61.3 ±1.47	29.4 ±0.53	34.30 ±0.19

LLaVA-OneVision classification repeats the OOD sampler advantage. Table 6 shows that random task-consistent batching raises the OOD average by 3.24–4.85 points for every objective. Hard batches remain competitive in distribution but trail random task-consistent batches on all five OOD objective averages.

Table 6: Per-dataset classification Precision@1 (%), B3-Qwen2-2B → LLaVA-OneVision-0.5B, mean±std over seeds 42–44. Avg is the unweighted mean over the panel's datasets.

(a) In-distribution

Objective	IN-1K	N24News	Hateful Memes	VOC2007	SUN397	**Avg**
Mixed-task batches						
Contrastive only	70.43 ±0.76	76.20 ±0.30	63.03 ±1.22	86.83 ±0.67	73.77 ±0.65	74.05 ±0.18
MSE	69.80 ±0.75	76.10 ±0.46	62.73 ±0.50	87.13 ±0.70	73.73 ±0.67	73.90 ±0.40
RKD	70.57 ±0.32	**76.57** ±0.38	63.83 ±1.58	86.97 ±0.87	73.80 ±0.35	74.35 ±0.29
HieRD	68.80 ±3.98	75.57 ±0.61	62.63 ±1.53	86.50 ±1.06	72.63 ±2.23	73.23 ±1.48
Merge KD	70.30 ±0.61	76.07 ±0.67	62.93 ±1.02	87.33 ±0.38	74.53 ±0.67	74.23 ±0.16
Task-consistent batches						
Contrastive only	**72.87** ±0.51	76.13 ±0.58	65.83 ±1.10	88.87 ±0.29	78.90 ±0.36	76.52 ±0.36
MSE	72.37 ±1.36	76.13 ±0.68	**66.37** ±1.27	88.97 ±0.72	78.67 ±0.32	76.50 ±0.13
RKD	72.77 ±0.96	76.53 ±0.40	64.83 ±0.59	88.67 ±0.49	78.67 ±0.59	76.29 ±0.19
HieRD	70.87 ±2.72	76.43 ±0.51	65.17 ±1.99	88.17 ±0.38	78.20 ±0.78	75.77 ±0.65
Merge KD	72.27 ±0.85	75.83 ±0.21	**66.37** ±1.36	**89.30** ±0.00	**79.20** ±1.01	**76.59** ±0.19
Task-consistent hard batches (B3)						
Contrastive only	74.4 ±0.65	75.0 ±0.46	66.8 ±1.16	88.0 ±0.52	79.4 ±0.53	76.72 ±0.28
MSE	74.2 ±1.10	75.1 ±0.58	63.7 ±0.97	88.6 ±0.71	78.5 ±0.53	76.02 ±0.30
RKD	74.1 ±0.72	75.9 ±0.39	61.8 ±1.19	87.9 ±0.71	78.5 ±0.49	75.64 ±0.25
HieRD	73.7 ±3.41	76.1 ±0.56	65.4 ±1.77	88.1 ±0.80	78.8 ±1.67	76.42 ±1.14
Merge KD	73.7 ±0.74	75.3 ±0.50	65.4 ±1.20	89.2 ±0.27	79.0 ±0.86	76.52 ±0.18

(b) Out-of-distribution

Objective	Places365	IN-A	IN-R	ObjectNet	Country211	**Avg**
Mixed-task batches						
Contrastive only	37.67 ±1.99	39.47 ±2.08	77.73 ±2.35	44.60 ±2.55	10.33 ±0.55	41.96 ±1.03
MSE	38.13 ±1.59	38.53 ±1.62	79.23 ±1.65	42.90 ±5.60	10.27 ±0.45	41.81 ±1.21
RKD	37.90 ±1.73	39.47 ±1.76	79.43 ±1.66	43.20 ±4.86	10.50 ±0.53	42.10 ±1.00
HieRD	37.77 ±0.78	38.50 ±1.45	81.57 ±2.62	41.60 ±8.87	11.60 ±0.36	42.21 ±1.98
Merge KD	37.80 ±1.87	40.03 ±2.42	79.40 ±1.32	47.53 ±1.80	**11.83** ±0.61	43.32 ±0.83
Task-consistent batches						
Contrastive only	42.00 ±0.98	39.77 ±2.15	83.50 ±0.20	52.10 ±2.69	10.17 ±0.51	45.51 ±0.39
MSE	42.23 ±1.02	40.20 ±1.01	**83.57** ±1.02	**56.90** ±2.04	10.40 ±0.96	**46.66** ±0.64
RKD	42.20 ±0.62	39.83 ±1.20	82.90 ±0.87	55.90 ±1.20	10.40 ±1.20	46.25 ±0.51
HieRD	41.53 ±0.87	**41.40** ±2.35	81.97 ±1.01	56.07 ±0.98	10.37 ±1.12	46.27 ±0.75
Merge KD	**42.67** ±0.98	40.87 ±1.33	83.13 ±0.92	55.33 ±0.93	10.80 ±0.80	46.56 ±0.45
Task-consistent hard batches (B3)						
Contrastive only	41.6 ±1.57	38.2 ±2.12	82.5 ±1.67	52.3 ±2.62	11.1 ±0.53	45.14 ±0.78
MSE	41.4 ±1.34	39.0 ±1.35	81.5 ±1.37	51.6 ±4.21	11.0 ±0.75	44.90 ±0.97
RKD	41.3 ±1.30	37.4 ±1.51	81.5 ±1.33	48.0 ±3.54	10.9 ±0.93	43.82 ±0.79
HieRD	42.1 ±0.83	37.1 ±1.95	81.0 ±1.99	48.1 ±6.31	11.7 ±0.83	44.00 ±1.50
Merge KD	42.2 ±1.49	37.7 ±1.95	82.3 ±1.14	44.9 ±1.43	12.9 ±0.71	44.00 ±0.67

Hard batches become competitive for LLaVA-OneVision VQA. Table 7 confirms that both task-consistent variants outperform mixed batching across the aggregate panels. Unlike FastVLM VQA, hard batches raise the OOD average to 50.45–50.95, compared with 49.64–50.20 under random task-consistent batching, showing that the value of hardness is student dependent.

Table 7: Per-dataset visual question answering Precision@1 (%), B3-Qwen2-2B → LLaVA-OneVision-0.5B, mean±std over seeds 42–44. Avg is the unweighted mean over the panel's datasets.

(a) In-distribution

Objective	OK-VQA	A-OKVQA	DocVQA	InfoVQA	ChartQA	Visual7W	**Avg**
Mixed-task batches							
Contrastive only	53.60 ±0.79	48.50 ±0.30	53.83 ±0.31	26.17 ±0.81	37.40 ±1.49	47.43 ±0.76	44.49 ±0.23
MSE	53.13 ±0.99	48.40 ±0.53	54.10 ±1.05	25.83 ±0.83	38.50 ±0.61	47.47 ±0.75	44.57 ±0.48
RKD	53.30 ±0.35	48.10 ±0.44	53.87 ±0.72	25.67 ±0.67	38.03 ±1.00	47.73 ±0.74	44.45 ±0.22
HieRD	52.70 ±1.00	47.47 ±1.96	52.43 ±2.61	25.77 ±0.49	36.83 ±1.66	46.97 ±0.75	43.69 ±1.19
Merge KD	53.53 ±0.31	48.00 ±0.89	53.97 ±0.21	26.33 ±0.38	37.83 ±0.38	47.63 ±0.84	44.55 ±0.22
Task-consistent batches							
Contrastive only	56.23 ±0.46	51.37 ±0.25	57.37 ±0.32	28.00 ±0.17	**39.60** ±0.52	**48.80** ±1.23	46.89 ±0.22
MSE	**56.73** ±0.38	51.30 ±0.62	57.63 ±0.59	27.67 ±0.42	39.30 ±0.80	48.60 ±0.69	46.87 ±0.03
RKD	**56.73** ±0.59	**51.47** ±0.50	**57.80** ±1.01	**28.27** ±0.47	39.53 ±0.12	48.30 ±1.01	**47.02** ±0.30
HieRD	55.77 ±0.76	50.23 ±1.76	56.33 ±1.42	27.60 ±0.46	39.10 ±0.17	47.80 ±0.80	46.14 ±0.84
Merge KD	56.23 ±0.32	51.20 ±0.20	57.57 ±0.58	28.13 ±0.21	39.23 ±0.45	**48.80** ±0.79	46.86 ±0.23
Task-consistent hard batches (B3)							
Contrastive only	56.5 ±0.65	51.3 ±0.28	58.0 ±0.32	29.0 ±0.59	41.2 ±1.12	48.1 ±1.02	47.35 ±0.23
MSE	56.5 ±0.75	51.7 ±0.58	58.8 ±0.85	29.1 ±0.66	40.1 ±0.71	47.6 ±0.72	47.30 ±0.34
RKD	56.3 ±0.49	51.1 ±0.47	57.8 ±0.88	28.7 ±0.58	39.9 ±0.71	48.1 ±0.89	46.98 ±0.26
HieRD	56.9 ±0.89	50.8 ±1.86	57.7 ±2.10	29.3 ±0.48	41.2 ±1.18	48.1 ±0.78	47.33 ±1.03
Merge KD	56.3 ±0.32	51.4 ±0.65	57.6 ±0.44	29.6 ±0.31	40.9 ±0.42	46.8 ±0.82	47.10 ±0.23

(b) Out-of-distribution

Objective	ScienceQA	VizWiz	GQA	TextVQA	**Avg**
Mixed-task batches					
Contrastive only	31.30 ±0.10	41.87 ±1.21	59.50 ±1.90	60.10 ±0.80	48.19 ±0.34
MSE	31.33 ±0.47	41.47 ±0.87	59.93 ±0.51	59.93 ±0.42	48.17 ±0.09
RKD	31.93 ±0.74	41.67 ±1.63	59.33 ±2.08	60.07 ±0.31	48.25 ±0.74
HieRD	32.67 ±1.11	41.10 ±2.99	55.40 ±2.35	59.07 ±1.12	47.06 ±1.80
Merge KD	30.90 ±0.26	41.87 ±1.55	59.67 ±2.84	60.00 ±0.87	48.11 ±0.39
Task-consistent batches					
Contrastive only	**33.27** ±0.83	43.70 ±0.36	60.63 ±1.76	**63.20** ±0.60	**50.20** ±0.49
MSE	32.47 ±0.76	43.53 ±0.55	60.27 ±1.70	62.70 ±0.72	49.74 ±0.22
RKD	32.80 ±0.56	43.67 ±0.50	60.07 ±1.07	62.77 ±0.40	49.82 ±0.20
HieRD	32.50 ±0.72	**44.33** ±1.51	**60.70** ±0.82	61.87 ±1.38	49.85 ±0.88
Merge KD	32.53 ±0.78	43.67 ±0.60	59.77 ±0.99	62.60 ±0.40	49.64 ±0.13
Task-consistent hard batches (B3)					
Contrastive only	35.0 ±0.59	43.9 ±0.89	59.9 ±1.83	63.0 ±0.71	50.45 ±0.42
MSE	34.1 ±0.63	44.3 ±0.73	60.6 ±1.26	63.2 ±0.59	50.55 ±0.17
RKD	35.1 ±0.66	43.7 ±1.21	60.2 ±1.65	63.3 ±0.36	50.58 ±0.54
HieRD	34.4 ±0.94	44.7 ±2.37	60.6 ±1.76	63.3 ±1.26	50.75 ±1.42
Merge KD	33.6 ±0.58	44.5 ±1.18	62.0 ±2.13	63.7 ±0.68	50.95 ±0.29

B.2 TEACHER SENSITIVITY

Implementation details. The teacher-sensitivity check keeps FastVLM-0.5B, the training data, evaluation pools, global batch size of 128, and the mixed-task and task-consistent samplers fixed, changing the teacher from B3-Qwen2-2B to Qwen3-VL-Embedding-2B. Tables 8 and 9 therefore report single-seed scores without standard deviations for the four available objectives.

The CLS sampler advantage survives the teacher change. Table 8 shows higher task-consistent IND and OOD averages for every available objective, yielding a $+2.29$ to $+3.09$ point improvement on the combined ALL score. The effect therefore persists with Qwen3-VL-Embedding-2B, although the single-seed design makes this corroborating rather than replicated evidence.

Table 8: Per-dataset classification Precision@1 (%), Qwen3-VL-Embedding-2B → FastVLM-0.5B, one seed (42). Avg is the unweighted mean over the panel's datasets. Bold marks the best value in each column. HieRD was not run with this teacher.

(a) In-distribution

Objective	IN-1K	N24News	Hateful Memes	VOC2007	SUN397	**Avg**
Mixed-task batches						
Contrastive only	52.90	57.50	56.50	75.60	67.20	61.94
MSE	53.00	57.20	56.60	75.30	66.60	61.74
RKD	52.50	**57.70**	55.20	75.10	68.50	61.80
Merge KD	53.10	57.50	55.00	74.60	67.90	61.62
Task-consistent batches						
Contrastive only	55.80	56.50	60.20	**78.20**	**72.70**	64.68
MSE	**56.10**	57.10	60.40	78.10	72.00	**64.74**
RKD	55.90	56.70	59.50	77.90	72.40	64.48
Merge KD	55.70	56.90	**60.70**	**78.20**	72.10	64.72

(b) Out-of-distribution

Objective	Places365	IN-A	IN-R	ObjectNet	Country211	**Avg**
Mixed-task batches						
Contrastive only	31.50	19.80	48.40	38.70	4.50	28.58
MSE	32.30	20.50	48.90	37.60	**5.00**	28.86
RKD	31.30	19.70	49.30	39.70	4.50	28.90
Merge KD	32.20	20.70	43.40	37.80	**5.00**	27.82
Task-consistent batches						
Contrastive only	34.80	21.00	**56.10**	39.40	4.00	31.06
MSE	35.00	**21.10**	55.80	40.40	4.50	**31.36**
RKD	35.00	20.80	54.10	39.90	4.20	30.80
Merge KD	**35.10**	20.00	54.40	**41.00**	4.00	30.90

The VQA advantage also survives, but narrows. Table 9 shows a positive task-consistent ALL difference for every objective, ranging from $+0.56$ to $+1.06$ points. The consistent direction across CLS and VQA argues against the main sampler result being specific to B3-Qwen2-2B, while the smaller VQA gains caution against assuming a teacher-invariant effect size.

Table 9: Per-dataset visual question answering Precision@1 (%), Qwen3-VL-Embedding-2B $\rightarrow$ FastVLM-0.5B, one seed (42). Avg is the unweighted mean over the panel's datasets. Bold marks the best value in each column. HieRD was not run with this teacher.

(a) In-distribution

Objective	OK-VQA	A-OKVQA	DocVQA	InfoVQA	ChartQA	Visual7W	**Avg**
Mixed-task batches							
Contrastive only	13.10	9.50	6.40	3.90	2.50	9.60	7.50
MSE	13.70	9.50	6.50	4.20	2.50	10.10	7.75
RKD	13.00	10.20	6.40	3.80	2.30	9.30	7.50
Merge KD	**14.30**	10.60	6.10	4.00	2.80	10.10	7.98
Task-consistent batches							
Contrastive only	14.00	**11.10**	**9.40**	4.60	2.80	**10.20**	**8.68**
MSE	13.70	10.80	9.30	4.60	3.00	10.00	8.57
RKD	14.00	10.80	8.60	**5.00**	**3.10**	9.80	8.55
Merge KD	14.10	11.00	8.50	4.40	2.80	9.90	8.45

(b) Out-of-distribution

Objective	ScienceQA	VizWiz	GQA	TextVQA	**Avg**
Mixed-task batches					
Contrastive only	9.90	**12.60**	10.80	8.20	10.38
MSE	9.20	11.90	10.70	8.30	10.03
RKD	7.60	11.80	10.70	8.30	9.60
Merge KD	9.40	12.30	10.90	8.10	10.18
Task-consistent batches					
Contrastive only	10.10	12.40	**11.90**	8.40	10.70
MSE	**11.40**	12.40	11.30	8.30	10.85
RKD	10.60	12.40	11.40	8.30	10.67
Merge KD	11.00	12.30	11.50	**8.70**	**10.88**