TE-VLM: TRANSFER ENTROPY-INSPIRED DISTILLATION FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transfer entropy (TE) is a principled measure of directed information flow, but it is intractable to estimate in high-dimensional multimodal representation spaces. We use optimization-step TE, the information a frozen teacher provides about a student's next-step representations beyond the student's current state, as a conceptual lens for vision–language model (VLM) distillation and derive from it two tractable proxy objectives, TE1 and TE2. Both reward the student for reproducing the teacher's local embedding geometry by matching within-batch finite-difference directions up to orientation through squared cosine similarity, per modality (TE1) or jointly across image and text (TE2), with no density estimation, no extra teacher passes, and negligible compute. In CLIP-style distillation with RN50, ViT-B/16, and RN50×16 teachers and RN34/RN18 students, the proxies consistently improve image-to-text retrieval over contrastive, KL, MSE, interactive-contrastive, mutual-information, and relational Gram-matching baselines while remaining competitive on text-to-image retrieval. With the base objective fixed, adding TE lifts MSCOCO image-to-text R@1 from 6.1% to 10.3%, against 8.7% for the strongest competing regularizer. The gains persist across temperatures, batch sizes, datasets (MSCOCO, Flickr8k, Flickr30k), and five random seeds; transfer better under distribution shift (MSCOCO→Flickr8k); and extend to classification, where the distilled student surpasses its teacher in label-free Food-101 evaluation and generalizes better to ImageNet-1k zero-shot classification. Representation-level analyses show that TE-distilled students preserve the teacher's neighborhood structure and joint image–text geometry more faithfully than mutual-information distillation.

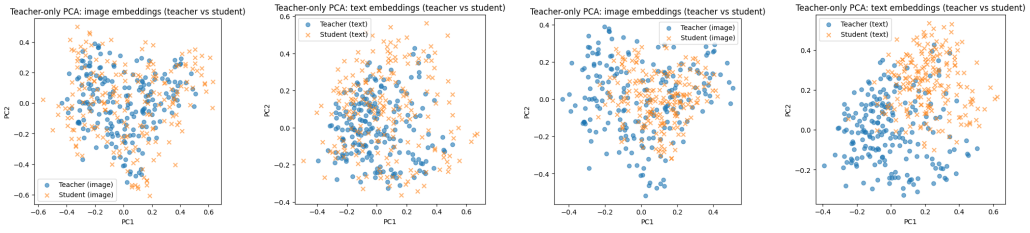
1 INTRODUCTION

Vision–language models (VLMs) learn joint image–text representations that support captioning, visual question answering, and cross-modal retrieval (Radford et al., 2021; Jia et al., 2021), but their size limits deployment in resource-constrained settings. Knowledge distillation (Hinton et al., 2015) transfers knowledge from a large teacher to a compact student while aiming to preserve performance.

Existing VLM distillation methods rely on contrastive objectives (Li et al., 2022; Yang et al., 2024) or divergence-based losses such as KL divergence (Li et al., 2024a), which match the teacher's embeddings, similarity scores, or batch-level distributions at a given optimization step. They constrain *where* the student's representations should be, but not *how* they should change relative to the teacher's. For CLIP-style retrieval this matters: a low-capacity student can match individual embeddings or similarity scores while still distorting the local geometry on which retrieval depends.

We approach this gap through *transfer entropy* (TE) (Schreiber, 2000). At the level of optimization dynamics, teacher-to-student TE measures how much the teacher explains the student's next-step representations beyond the student's current state, a natural description of teacher-guided update geometry. Estimating it in high-dimensional embedding spaces is intractable, so, following the scalable cosine-similarity approximation of Liang & Shi (2026), we use TE only as motivation for two geometric proxies: TE1 matches the one-dimensional subspaces spanned by teacher and student within-batch finite differences separately per modality, and TE2 matches the corresponding concatenated image–text subspaces jointly. Both require no Jacobian construction or additional teacher passes and add negligible compute.

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.



(a) TE: image (b) TE: text (c) MI: image (d) MI: text

Figure 1: Per-modality teacher–student embedding geometry on MSCOCO val after 10 epochs of distillation (CLIP RN50→RN34, $\tau = 0.07$; $\alpha = 1$, $\beta = 50$, $\delta = 1$). For each modality, PCA is fit on the teacher validation embeddings and teacher and student are projected onto this shared 2D basis. (a)–(b) TE-based student ($\gamma = 7.5$, $\eta = 0$); (c)–(d) MI-based student ($\eta = 5$, $\gamma = 0$). In both modalities the TE student tracks the teacher's geometry more closely than the MI student; quantitative agreement metrics are in Section 5.2.

With the base objective fixed, adding TE lifts image-to-text R@1 on MSCOCO from 6.1% to 10.3%, compared with 8.7% for a mutual-information objective and at most 7.0% for a relational Gram-matching baseline that aligns static pairwise similarities. Matching *finite-difference geometry up to orientation* thus transfers substantially more than aligning *static similarities*. The gains hold across teachers, students, datasets, temperatures, batch sizes, and seeds; transfer from MSCOCO to Flickr8k; and extend to classification, where the distilled student surpasses its teacher in label-free Food-101 evaluation and, without ever seeing ImageNet images, generalizes better than the mutual-information student on ImageNet-1k zero-shot classification. Figure 1 previews the mechanism: the TE-distilled student reproduces the teacher's per-modality embedding geometry far more faithfully than a mutual-information student.

Our contributions are as follows.

- A TE-inspired perspective on VLM distillation that frames the teacher's role as guiding the student's update geometry, while clearly separating the ideal optimization-step TE from the tractable proxy that is optimized.
- Two lightweight objectives, TE1 and TE2, that enforce per-modality and joint orientation-invariant alignment of teacher–student within-batch finite-difference geometry.
- Controlled comparisons showing that these proxies outperform contrastive, KL, MSE, interactive-contrastive, mutual-information, and Gram-matching regularizers on image-to-text retrieval while remaining competitive on text-to-image retrieval, with recipe-level, cross-dataset, seed-level, and representation-level analyses on MSCOCO, Flickr8k, Flickr30k, Food-101, and ImageNet-1k.

2 Related Work

Knowledge distillation. Knowledge distillation transfers knowledge from a large teacher to a smaller student (Hinton et al., 2015) by matching outputs, intermediate features (Romero et al., 2015), attention maps (Zagoruyko & Komodakis, 2017; Touvron et al., 2021), contrastive agreement (Tian et al., 2020), or relational structure (Park et al., 2019; Passalis & Tefas, 2018). Information-theoretic variants maximize teacher–student mutual information (Ahn et al., 2019) or preserve cross-modal dependencies (Chen et al., 2023); other extensions include data-free (Huang & Wang, 2017), frequency-aware (Pham et al., 2024; Zhang et al., 2024), self- (Chen et al., 2024), synthetic-data (Liu et al., 2024), dual-teacher (Li et al., 2024b), and projection-based (Miles et al., 2024) distillation.

Vision–language model distillation. Building on joint image–text representations (Li et al., 2020; Chen et al., 2020), prior work compresses multimodal encoders (Fang et al., 2021), transfers vision–language knowledge downstream (Dai et al., 2022; Wang et al., 2022), distills CLIP via affinity mimicking and weight inheritance (Wu et al., 2023), and targets reasoning, open-vocabulary recog-

nition, robustness, and transfer (Hu et al., 2024; Gu et al., 2021; Li et al., 2023; Addepalli et al., 2024). Recent CLIP distillation adds multiple teachers (Cao et al., 2025), earlier cross-modal alignment (Feng et al., 2025), interactive contrastive learning with MSE matching (Yang et al., 2024), and KL-based alignment (Li et al., 2024a).

Positioning. All of the above constrain the student at a fixed optimization step by matching positions, similarity distributions, or intra-model similarity structure. Our proxies instead constrain the *one-dimensional subspaces spanned by representation changes*: student and teacher within-batch finite differences are compared directly in the shared embedding space using squared cosine similarity, giving an orientation-invariant Jacobian-action signal without forming Jacobians, with TE2 coupling both modalities in one jointly normalized score. TE motivates the objective but is never estimated (Section 4.1), and the comparison with Gram matching (Section 5.1) shows that matching finite-difference secant geometry transfers far more than aligning static similarities.

3 Introduction to Transfer Entropy

Transfer entropy (TE) is an information-theoretic measure introduced by Schreiber (Schreiber, 2000) to quantify the directed transfer of information between two stochastic processes. It is well suited to detecting asymmetric interactions because it measures how much the past of one process, X, informs the future of another, Y, beyond what is already explained by the past of Y itself.

For two discrete-time stochastic processes $X(t)$ and $Y(t)$, the transfer entropy from X to Y is defined as (Schreiber, 2000)

$$T_{X\to Y} = \sum_{y_{t+1},\, y_t,\, x_t} p(y_{t+1}, y_t, x_t) \log \frac{p(y_{t+1} \mid y_t, x_t)}{p(y_{t+1} \mid y_t)}, \tag{1}$$

where $p(\cdot)$ denotes the probability distribution of the respective random variables.

Transfer entropy is closely related to conditional mutual information: it can be rewritten as the mutual information between Y_{t+1} and X_t conditioned on Y_t (Shahsavari Baboukani et al., 2020),

$$T_{X\to Y} = I(Y_{t+1}; X_t \mid Y_t). \tag{2}$$

The proof is given in Appendix A. This formulation makes explicit that TE measures the additional information that X_t provides about the future state Y_{t+1}, over and above Y's own history. Appendix B reviews prior work on mutual information and TE estimation.

Computing TE exactly for Transformer hidden states is generally intractable, since it requires estimating high-dimensional conditional densities. Following Liang & Shi (2026), we use a finite-difference surrogate that captures directed dependence through the alignment of perturbation-induced changes across layers:

$$\widehat{\mathrm{TE}}^{(\ell)}(a \to b) \approx \frac{1}{2(B-1)} \sum_{m=1}^{B-1} \cos^2\left(\Delta z_{b,m}^{(\ell+1)}, \Delta z_{a,m}^{(\ell)}\right), \tag{3}$$

where B is the mini-batch size and $\Delta z_{*,m}$ is the finite difference between adjacent batch elements, e.g.,

$$\Delta z_{b,m}^{(\ell+1)} = \tilde{h}_{b,m+1}^{(\ell+1)} - \tilde{h}_{b,m}^{(\ell+1)}. \tag{4}$$

This surrogate ranks directed information flow between components efficiently, without explicitly estimating the underlying high-dimensional conditional distributions.

4 Method

4.1 Why Transfer Entropy for VLM Distillation

In VLM distillation, the goal is not only to match the teacher's representations but also to encourage the student's updates to move in teacher-consistent directions across both modalities. TE offers a conceptual framing for this idea. For a mini-batch $\mathcal{B}$, let $(U^{(T)}, V^{(T)})$ denote the frozen teacher's

text and image representations, and let $(U_t^{(S)}, V_t^{(S)})$ and $(U_{t+1}^{(S)}, V_{t+1}^{(S)})$ denote the student's representations on the same mini-batch at optimization steps t and $t+1$. We define the teacher→student TE at step t as

$$T_{\mathrm{T}\to\mathrm{S}}(t;\mathcal{B}) = I\Big(\big(U_{t+1}^{(S)}, V_{t+1}^{(S)}\big)\,;\,\big(U^{(T)}, V^{(T)}\big)\;\Big|\;\big(U_t^{(S)}, V_t^{(S)}\big)\Big), \tag{5}$$

the additional predictive information the teacher provides about the student's next-step representations beyond the student's current state. A high value indicates that the teacher is meaningfully shaping the student's refinement, as expected early in distillation. Here t is the optimization step, not wall-clock time, so (5) assumes no temporal correlation among samples in a shuffled mini-batch; similar information-theoretic views of learning dynamics appear in prior work (Goldfeld et al., 2019; Achille & Soatto, 2018).

At CLIP-scale dimensions, (5) is intractable to estimate directly. We therefore use geometric proxies (Section 4.2) that align teacher and student finite-difference vectors up to orientation via squared cosine similarity. In practice, this enforces *local secant-line alignment* rather than causal information transfer. The proxies add only inexpensive difference and cosine computations to standard KD.

4.2 TE-Inspired Proxies via Squared Cosine Similarity

Direct TE estimation for modern vision–language models is challenging because their representations are high-dimensional (e.g., 1024-dimensional for CLIP ResNet-50 (Radford et al., 2021)). Since TE is a conditional mutual information (Eq. 2), estimation requires modeling high-dimensional conditional dependencies, which suffers from the curse of dimensionality and often requires expensive density estimation or restrictive assumptions (Köppen, 2000; Shahsavari Baboukani et al., 2020; Gowri et al., 2025).

Following Liang & Shi (2026), we instead use finite differences and normalized directional similarity as scalable proxies, specialized here to teacher–student representation alignment.

From local variations to finite differences. For a small displacement δx, a first-order expansion gives

$$f_S(x+\delta x) - f_S(x) \approx J_S(x)\delta x, \qquad f_T(x+\delta x) - f_T(x) \approx J_T(x)\delta x, \tag{6}$$

Comparing teacher and student finite differences aligns their local response subspaces without forming Jacobians. For nearby pairs, this has an orientation-invariant Jacobian-action interpretation; for within-batch pairs, it acts as a stochastic *secant-line* regularizer. Appendix G.6 shows that adjacent and nearest-neighbor pairing perform similarly. Since batch index is not time, these are TE-inspired geometric rewards rather than temporal TE estimators. We define TE1 and TE2 in Sections 4.2.1–4.2.2.

4.2.1 TE1: Per-Modality Squared Cosine Similarity of Differences

Let $\mathbf{v}^{(S)}$ and $\mathbf{u}^{(S)}$ denote the student's image and text embeddings, and $\mathbf{v}^{(T)}$ and $\mathbf{u}^{(T)}$ the corresponding teacher embeddings. Since the teacher is frozen, any variation in its embeddings arises from changing inputs rather than training dynamics. After an internal random permutation of the mini-batch, we compute differences between adjacent embeddings for each modality:

$$\Delta\mathbf{v}_i^{(S)} = \mathbf{v}_{i+1}^{(S)} - \mathbf{v}_i^{(S)}, \quad \Delta\mathbf{v}_i^{(T)} = \mathbf{v}_{i+1}^{(T)} - \mathbf{v}_i^{(T)}, \tag{7}$$

$$\Delta\mathbf{u}_i^{(S)} = \mathbf{u}_{i+1}^{(S)} - \mathbf{u}_i^{(S)}, \quad \Delta\mathbf{u}_i^{(T)} = \mathbf{u}_{i+1}^{(T)} - \mathbf{u}_i^{(T)}. \tag{8}$$

We first compute the cosine similarity between student and teacher finite-difference vectors (Xia et al., 2015),

$$\cos\theta_i^{(v)} = \frac{\langle \Delta\mathbf{v}_i^{(S)}, \Delta\mathbf{v}_i^{(T)} \rangle}{\|\Delta\mathbf{v}_i^{(S)}\|\|\Delta\mathbf{v}_i^{(T)}\| + \epsilon}, \qquad \cos\theta_i^{(u)} = \frac{\langle \Delta\mathbf{u}_i^{(S)}, \Delta\mathbf{u}_i^{(T)} \rangle}{\|\Delta\mathbf{u}_i^{(S)}\|\|\Delta\mathbf{u}_i^{(T)}\| + \epsilon}, \tag{9}$$

where ϵ is a small constant that prevents division by zero. Averaging the squared cosines over all adjacent pairs yields per-modality scores,

$$\mathrm{TE}_{\mathrm{img}} = \frac{1}{B-1}\sum_{i=1}^{B-1}\cos^2\theta_i^{(v)}, \qquad \mathrm{TE}_{\mathrm{txt}} = \frac{1}{B-1}\sum_{i=1}^{B-1}\cos^2\theta_i^{(u)}, \tag{10}$$

and TE1 is their mean:

$$\mathrm{TE1} = \frac{1}{2}\left(\mathrm{TE}_{\mathrm{img}} + \mathrm{TE}_{\mathrm{txt}}\right). \tag{11}$$

Because the score uses $\cos^2$, it is invariant to the sign of either finite-difference vector: parallel and anti-parallel vectors receive the same reward. TE1 therefore matches the one-dimensional subspaces (secant lines) spanned by teacher and student finite differences, rather than their oriented directions. The two terms carry modality-specific signal: rewarding only $\mathrm{TE}_{\mathrm{img}}$ mainly improves I2T retrieval (R@1 6.08→7.46%) and rewarding only the text-side term mainly improves T2I (5.92→7.03%), while the full TE1 improves both (Table 2).

4.2.2 TE2: Squared Cosine Similarity of Concatenated Differences

TE2 instead combines the two modalities before computing similarity. Using the same differences as in (7)–(8), we concatenate them along the feature dimension,

$$\Delta\mathbf{c}_i^{(S)} = \left[\Delta\mathbf{v}_i^{(S)} \,\|\, \Delta\mathbf{u}_i^{(S)}\right], \tag{12}$$

$$\Delta\mathbf{c}_i^{(T)} = \left[\Delta\mathbf{v}_i^{(T)} \,\|\, \Delta\mathbf{u}_i^{(T)}\right], \tag{13}$$

and compute

$$\cos\theta_i^{(\mathrm{cat})} = \frac{\langle \Delta\mathbf{c}_i^{(S)}, \Delta\mathbf{c}_i^{(T)} \rangle}{\|\Delta\mathbf{c}_i^{(S)}\|\,\|\Delta\mathbf{c}_i^{(T)}\| + \epsilon}, \qquad \mathrm{TE2} = \frac{1}{B-1}\sum_{i=1}^{B-1} \cos^2\theta_i^{(\mathrm{cat})}. \tag{14}$$

TE2 thus provides a single score reflecting how well the student's combined multimodal secant-line structure matches the teacher's, up to orientation.

Why concatenation differs from averaging. Because cosine similarity is nonlinear, TE2 is not algebraically equivalent to TE1. Geometrically, TE2 uses the squared cosine of the angle between the concatenated finite-difference vectors in the direct-sum space $\mathbb{R}^{d_v} \oplus \mathbb{R}^{d_u}$, thereby matching the one-dimensional subspaces they span up to orientation, whereas TE1 treats the modalities independently. Writing $a = \Delta\mathbf{v}_i^{(S)}$, $b = \Delta\mathbf{u}_i^{(S)}$, $c = \Delta\mathbf{v}_i^{(T)}$, $d = \Delta\mathbf{u}_i^{(T)}$,

$$\cos\theta_i^{(\mathrm{cat})} = \frac{\langle a, c\rangle + \langle b, d\rangle}{\sqrt{\|a\|^2 + \|b\|^2}\,\sqrt{\|c\|^2 + \|d\|^2} + \epsilon}. \tag{15}$$

The shared normalization couples the modalities: a modality with large update magnitude (e.g., image) dominates the joint direction unless the other (e.g., text) is also aligned. This coupling encourages the student to match the teacher's *combined* multimodal update geometry rather than achieving high alignment in only one modality. Empirically the joint score is at least as effective as the per-modality one: TE2 alone gives the best MSCOCO I2T result (10.32% R@1) and TE1, TE2, and their sum agree to within 0.2 points (Table 2), so the variants are interchangeable in practice and TE2 is preferable when a single scalar is wanted.

In all experiments, TE proxies are computed on the *global* CLIP embeddings: the pooled visual embedding (e.g., the CLS token for ViT) for images, and the end-of-text (EOT) token embedding from the final Transformer layer for text; we do not operate on token-level features. Appendix C analyzes the computational cost of direct TE estimation versus the proposed proxies.

4.3 Loss Functions for VLM Distillation

The total objective combines distillation signals at different granularities: a contrastive loss (CL) for CLIP-style global alignment within a batch, KL divergence on the teacher's soft similarity distribution, feature regression (MSE) that anchors the student to the teacher in representation space, and interactive contrastive learning (ICL), which compares student image embeddings with teacher text embeddings and vice versa. For controlled comparison we also include two geometric baselines: a CRD-style mutual-information (MI) loss (Tian et al., 2020), an InfoNCE objective that scores matched teacher–student pairs above mismatched ones, and an MSE loss on finite differences (MSE-Δ). All of these match embeddings or similarity structure at a fixed optimization step; the TE proxy instead rewards teacher-consistent update geometry. Treating the proxy score as a reward,

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{contrastive}} + \alpha\,\mathcal{L}_{\mathrm{KL}} + \beta\,\mathcal{L}_{\mathrm{MSE}} + \delta\,\mathcal{L}_{\mathrm{ICL}} + \eta\,\mathcal{L}_{\mathrm{MI}} + \lambda\,\mathcal{L}_{\mathrm{MSE\text{-}\Delta}} - \gamma\,\mathrm{TE}, \tag{16}$$

Table 1: Retrieval (Recall@K) of an RN34 student distilled from CLIP RN50 on **MSCOCO** under different loss combinations. All Loss Function: CL + KL + MSE + ICL - TE1 - TE2. Weights: $\alpha = 1.0$, $\beta = 50$, $\delta = 1.0$, $\gamma = 1.0$.

Model and Loss Function	I2T Retrieval (R)			T2I Retrieval (R)		
	R@1	R@5	R@10	R@1	R@5	R@10
Teacher Model (RN50)	25.29%	45.68%	55.37%	12.02%	26.44%	34.51%
Student Models (RN34)						
CL Only (Oord et al., 2018)	4.94%	14.60%	22.51%	3.96%	12.67%	19.45%
CL + MSE (Yang et al., 2024)	5.13%	15.41%	23.17%	4.00%	12.79%	19.53%
CL + KL (Li et al., 2024a)	5.42%	16.20%	24.55%	5.06%	15.35%	22.92%
CL + ICL (Yang et al., 2024)	5.75%	16.86%	24.83%	5.07%	15.14%	22.44%
CL - TE1	6.91%	19.48%	28.12%	5.68%	16.33%	23.93%
CL - TE2	7.04%	19.22%	27.97%	5.46%	15.90%	23.49%
CL - TE1 - TE2	**8.24%**	**22.43%**	**31.73%**	6.53%	18.13%	26.02%
CL + KL - TE1	7.81%	21.31%	30.65%	6.42%	18.18%	26.33%
CL + KL - TE2	7.77%	21.10%	30.22%	6.21%	17.95%	26.03%
CL + KL + MSE - TE1	7.62%	21.05%	30.34%	6.53%	18.48%	26.66%
CL + KL + ICL - TE1	7.59%	20.87%	29.95%	6.78%	19.11%	27.33%
CL + KL + MSE + ICL - TE1	7.51%	20.62%	29.81%	6.76%	19.02%	27.32%
All Loss Function	8.11%	22.05%	31.57%	**7.18%**	**19.75%**	**28.14%**

where α, β, δ, η, λ, and γ weight the respective terms. All terms except TE are defined in Appendix D, and a full training step is given in Appendix E.

5 Experiments

Setup. We evaluate the proposed TE-inspired proxies under four CLIP-style teacher $\rightarrow$ student configurations: RN50 $\rightarrow$ RN34, ViT-B/16 $\rightarrow$ RN34, RN50$\times$16 $\rightarrow$ RN34, and RN50 $\rightarrow$ RN18 (background on the CLIP teachers is given in Appendix F). The student pairs a ResNet image encoder with a lightweight two-layer Transformer text encoder, for a total of roughly 55–60M parameters (RN34) or 45–50M (RN18). Before vector-space losses are evaluated, teacher and student representations are placed in a common comparison dimension by their architecture-specific output/projection heads; for the RN50$\rightarrow$RN34 configuration this dimension is 1024. Retrieval is evaluated on MSCOCO 2014 (Lin et al., 2014), Flickr8k (Hodosh et al., 2013; Marco et al., 2023), and Flickr30k (Young et al., 2014); transfer to classification is evaluated on Food-101 (Bossard et al., 2014) and, out of domain, on ImageNet-1k (Deng et al., 2009; Russakovsky et al., 2015). We report Recall@K ($K \in \{1, 5, 10\}$, in %) for image-to-text (I2T) and text-to-image (T2I) retrieval (Manning et al., 2008) and, where indicated, mean reciprocal rank (MRR; defined in Appendix G.12). Unless otherwise stated, the student is trained for 10 epochs with Adam ($\rho = 10^{-4}$), $\tau = 0.07$, $|B| = 64$, and $\alpha = 1$, $\beta = 50$, $\delta = 1$, $\lambda = 0$ in Eq. (16); TE-based settings use $\eta = 0, \gamma > 0$ and MI-based baselines use $\eta > 0, \gamma = 0$. Loss weights are first scaled so that no term dominates by magnitude alone and then tuned by a small grid on MSCOCO validation Recall@K; the heuristic, the full sweep, architectural details, and compute budgets are given in Appendix G.1. We stress that TE1/TE2 are tractable proxies motivated by optimization-step TE (Section 4.1), not estimators of it.

5.1 Teacher: RN50, Student: RN34, Dataset: MSCOCO

Which losses does the proxy complement? Table 1 compares loss combinations at a fixed, conservative TE weight ($\gamma = 1$). Adding TE1 or TE2 to the base contrastive loss yields the largest single-step gain among all tested additions (I2T R@1 from 4.94% to 6.91%/7.04%), exceeding CL+KL, CL+MSE, and CL+ICL. `CL - TE1 - TE2` gives the best I2T result, while the full objective is best for T2I, indicating that the proxies act as strong regularizers on their own and also complement conventional distillation losses. Every best-performing configuration includes a TE component.

Controlled comparison with geometry-matching baselines. Table 2 fixes the base objective CL+KL+MSE+ICL ($\alpha = 1$, $\beta = 50$, $\delta = 1$) and varies only the added regularizer, isolating the

Table 2: Controlled comparison of regularizers added to a fixed base objective (CL+KL+MSE+ICL; $\alpha = 1$, $\beta = 50$, $\delta = 1$, $\tau = 0.07$, $|B| = 64$) for RN50→RN34 on MSCOCO (Recall@K, %). Only the added term changes across rows. Gram matching aligns within-batch teacher–student similarity matrices (Appendix G.4); MI is the CRD-style objective. Best per column in bold.

	Image-to-Text			Text-to-Image		
Added term	R@1	R@5	R@10	R@1	R@5	R@10
None (base)	6.08	18.14	26.93	5.92	17.06	24.89
MSE-Δ ($\lambda = 50$)	6.67	18.48	27.30	6.01	17.27	25.24
Gram ($\mu = 50$)	7.01	19.35	28.03	5.92	17.13	24.94
Gram ($\mu = 100$)	6.91	19.50	28.16	6.05	17.37	25.30
MI ($\eta = 5$)	8.65	23.29	33.05	**7.65**	**20.63**	**29.05**
TE_1 ($\gamma = 7.5$)	10.10	25.97	35.76	7.33	20.22	28.59
TE_2 ($\gamma = 7.5$)	**10.32**	**26.44**	**36.43**	7.33	20.03	28.46
TE_1+TE_2 ($\gamma = 7.5$)	10.27	26.36	36.30	6.98	18.95	26.95
image-side TE_1 only	7.46	20.92	30.36	5.69	16.67	24.38
text-side TE_1 only	7.49	20.27	28.96	7.03	19.40	27.60

proxy from every other term. TE1, TE2, and their sum give the largest I2T gains by a wide margin; the MSE-Δ loss on finite differences, a relational Gram-matching baseline that aligns within-batch teacher–student cosine-similarity matrices (Tung & Mori, 2019; Park et al., 2019), and the CRD-style MI objective (Tian et al., 2020) all trail under identical conditions, although MI remains the strongest option for T2I at this temperature. The TE–Gram gap is the informative one: both are relational geometric regularizers, but Gram matching preserves *static* pairwise similarities whereas TE1/TE2 match *finite-difference secant lines up to orientation*, and the latter is markedly more effective. Restricting the proxy to a single modality (last two rows) shows that the image-side term mainly drives I2T and the text-side term T2I, consistent with the proxies transferring orientation-invariant finite-difference geometry per modality. Details of the Gram baseline and this ablation are in Appendices G.4 and G.5.

Sensitivity to the TE weight. In the full sweep (Appendix Table 5), γ has the clearest effect of any weight: I2T R@1 rises monotonically from 6.08% ($\gamma = 0$) to 10.27% ($\gamma = 7.5$) and dips to 9.77% at $\gamma = 10$, indicating mild over-regularization when the reward is too strong, while T2I peaks earlier ($\gamma = 2.5$, 7.57%). Varying α, β, and δ produces comparatively small fluctuations, and stacking auxiliary terms is not additive. We therefore use $\gamma = 7.5$ (TE) and $\eta = 5$ (MI) as reference settings in the remaining MSCOCO experiments.

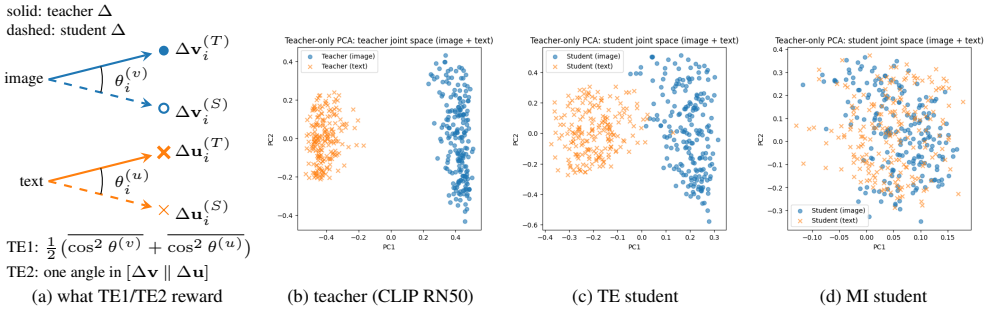
Why are gains larger for I2T than for T2I? The asymmetry is already present in the teacher (25.29% vs. 12.02% R@1 in Table 1) and is partly inherited by a student trained to preserve the teacher's cross-modal geometry. Moreover, I2T ranks captions for an image whereas T2I ranks images for a caption, so the two directions weight image-side, text-side, and joint geometry differently, while TE1/TE2 are shared rather than direction-specific regularizers; the modality-specific rows of Table 2 support this reading.

5.2 Robustness to the Training Recipe and Representation-level Diagnostics

Robustness to the training recipe. To check that the gains are not an artifact of one training recipe, Table 3 varies the contrastive temperature $\tau \in \{0.03, 0.07, 0.15\}$ and the batch size $|B| \in \{64, 128\}$, comparing the reference TE ($\gamma = 7.5$) and MI ($\eta = 5$) settings. TE improves I2T over MI at every temperature and batch size, with matching gains at R@5 and R@10; on T2I the picture is mixed, with TE slightly ahead at $\tau = 0.03$ and MI ahead at the two higher temperatures. Increasing $|B|$ also improves the TE model itself, so the regularizer benefits from a stronger negative pool rather than relying on a weak one. The strongest TE result occurs at $\tau = 0.03$: the training dynamics in Appendix G.2 show that this lower temperature yields substantially larger KL reduction for both objectives, while TE gives a markedly stronger MSE reduction than MI at both temperatures, so the weak KL trend at $\tau = 0.07$ is a recipe effect rather than a property of the objective.

Table 3: Recipe ablation for RN50→RN34 on MSCOCO (Recall@K, %) under $\alpha = 1$, $\beta = 50$, $\delta = 1$. TE denotes $(\gamma = 7.5, \eta = 0)$ and MI denotes $(\eta = 5, \gamma = 0)$. Top: temperature sweep at $|B| = 64$; bottom: larger batch at $\tau = 0.07$. Best within each setting and metric in bold.

τ	$\lvert B\rvert$	Method	Image-to-Text			Text-to-Image		
			R@1	R@5	R@10	R@1	R@5	R@10
0.03	64	TE	**10.85**	**26.47**	**36.29**	**6.84**	**18.76**	**26.94**
0.03	64	MI	9.14	23.67	33.14	6.49	18.23	26.44
0.07	64	TE	**10.27**	**26.36**	**36.30**	6.98	18.95	26.95
0.07	64	MI	8.65	23.29	33.05	**7.65**	**20.63**	**29.05**
0.15	64	TE	**9.68**	**24.92**	**34.64**	6.86	18.14	25.80
0.15	64	MI	8.73	23.10	32.42	**6.96**	**18.54**	**26.23**
0.07	128	TE	**10.69**	**26.94**	**37.18**	7.07	19.25	27.52
0.07	128	MI	9.64	25.60	35.50	**7.62**	**20.46**	**29.01**



(a) what TE1/TE2 reward (b) teacher (CLIP RN50) (c) TE student (d) MI student

Figure 2: The proxy of Section 4.2 and its effect on embedding geometry. (a) For a within-batch pair, TE1 and TE2 reward the student and teacher finite-difference vectors for spanning the same one-dimensional subspace, irrespective of orientation. Secant-line geometry rather than pointwise positions is compared, so the reward targets the teacher's local geometric structure (Section 4.1). (b)–(d) Teacher-only PCA of pooled image (circles) and text (crosses) embeddings on MSCOCO val after 10 epochs at $\tau = 0.07$ (RN50→RN34; $\alpha = 1$, $\beta = 50$, $\delta = 1$; PCA fit per panel, so only within-panel structure is comparable). The TE student ($\gamma = 7.5$, with both rewards active, TE = TE1 + TE2) retains the teacher's bimodal image–text organization, the property the reward targets. The MI student ($\eta = 5$) shows heavier cross-modal mixing: image and text embeddings interleave where the teacher keeps them apart, so the student has rearranged the teacher's neighborhoods rather than reproduced them. This departure is reflected numerically in lower neighborhood overlap and cosine agreement with the teacher (Appendix Table 6).

Representation-level diagnostics. The retrieval gains come with a more teacher-faithful embedding geometry. On MSCOCO validation (RN50→RN34), TE preserves the teacher's local neighborhoods (kNN overlap@10 on image embeddings) better than MI at both temperatures and yields far higher per-sample cosine agreement in each modality (Appendix Tables 6 and 7). Figures 1 and 2(b)–(d) show the same picture per modality and in the joint space: the TE student retains the teacher's bimodal image–text organization, whereas the MI student shows heavier cross-modal mixing ($\tau = 0.03$ counterparts in Appendix G.3). The proxy thus reshapes the representation, not only scalar losses, and Appendix G.6 shows the results are insensitive to the finite-difference pairing rule.

5.3 Generalization Across Teachers, Datasets, and Seeds

Appendix Table 22 reports five-seed means and standard deviations for TE- and MI-regularized students across four teacher–dataset combinations, including a ViT-B/16 teacher and the larger Flickr30k benchmark. TE achieves higher I2T Recall@K than MI in every combination, with gains that are many times the seed standard deviation on the larger benchmarks. On T2I, TE is competitive with or better than MI on most metrics, with MI occasionally matching or slightly exceeding it on a single Recall@K value. The same ordering holds for MRR (Appendix Table 23).

Table 4: Cross-dataset zero-shot retrieval (%) for RN34 distilled from RN50 on MSCOCO and evaluated on Flickr8k ($\alpha = 1$, $\beta = 50$, $\delta = 1$, $\lambda = 0$). TE rows use $\eta = 0$; MI rows use $\gamma = 0$. Best TE and best MI within each direction are boldfaced separately.

τ	Method	Image-to-Text				Text-to-Image			
		R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
0.07	TE ($\gamma = 2.5$)	15.36	34.33	44.77	25.03	11.40	27.59	36.84	19.90
0.07	TE ($\gamma = 5$)	16.35	35.85	46.41	26.23	**12.17**	**28.44**	**37.98**	**20.74**
0.07	TE ($\gamma = 7.5$)	16.66	36.02	46.22	26.43	11.79	27.41	36.95	20.15
0.03	TE ($\gamma = 5$)	17.19	36.15	46.67	26.85	11.22	27.14	36.47	19.59
0.03	TE ($\gamma = 7.5$)	**17.74**	**38.26**	**48.76**	**27.92**	11.80	27.50	36.88	20.10
0.07	MI ($\eta = 2.5$)	13.85	31.95	42.62	23.26	11.08	26.45	35.62	19.23
0.07	MI ($\eta = 5$)	**15.51**	33.58	44.14	**24.82**	**11.89**	27.70	36.93	20.23
0.07	MI ($\eta = 7.5$)	14.58	**33.68**	**45.20**	24.48	11.79	**27.79**	**37.31**	**20.24**
0.03	MI ($\eta = 5$)	13.51	30.08	41.19	22.41	9.70	23.96	32.70	17.38
0.03	MI ($\eta = 7.5$)	13.22	30.08	40.09	22.11	9.68	23.81	32.62	17.33

Flickr8k shows larger variance ($\approx$1 point), as expected from its small size, but TE still leads on I2T. Full loss-combination and weight-sweep tables for ViT-B/16$\rightarrow$RN34 on MSCOCO, Flickr8k, and Flickr30k (Appendix G.7), RN50$\rightarrow$RN34 on Flickr8k (Appendix G.8), the smaller RN18 student (Appendix G.9), and the stronger RN50$\times$16 teacher (Appendix G.10) show the same pattern: TE1/TE2 give the largest single gain among the tested loss terms. The proxies are also essentially free: the TE terms account for less than 10^{-6} of total training FLOPs and change 10-epoch wall-clock time by about one minute (Appendix Table 21).

5.4 Cross-Dataset Transfer and Classification

MSCOCO $\rightarrow$ Flickr8k retrieval. To test whether the proxies preserve transferable structure rather than fitting the source distribution, we distill RN34 from RN50 on MSCOCO and evaluate zero-shot retrieval on Flickr8k without fine-tuning (Table 4). TE outperforms the best MI setting.

Classification. Appendix G.11 reports label-free classification on Food-101 and zero-shot classification on ImageNet-1k for RN34 distilled from RN50 without ground-truth labels.

6 Conclusions and Future Work

We introduced a TE-inspired regularization framework for VLM distillation. Since optimization-step transfer entropy is intractable in high-dimensional multimodal embedding spaces, we use it only to motivate two lightweight geometric proxies, TE1 and TE2, that match teacher and student within-batch finite-difference geometry up to orientation, per modality and jointly, at negligible cost and alongside standard CLIP-style losses. With a fixed base objective, the proxies outperform contrastive, KL, MSE, interactive-contrastive, mutual-information, and Gram-matching regularizers on image-to-text retrieval while remaining competitive on text-to-image retrieval; the gains hold across teachers, students, datasets, temperatures, batch sizes, and seeds, transfer from MSCOCO to Flickr8k, and extend to label-free classification on Food-101 and zero-shot classification on ImageNet-1k. TE-distilled students also preserve the teacher's neighborhood structure and joint image–text geometry more faithfully than mutual-information distillation. The central lesson is the contrast with Gram matching: matching orientation-invariant finite-difference geometry transfers far more than aligning static pairwise similarities, and the TE view is a useful guide to such objectives even when TE itself cannot be estimated.

Our study is limited to fixed-teacher distillation and first-order surrogates on pooled embeddings. Natural extensions include co-distillation with jointly evolving teacher and student, surrogates capturing nonlinear or higher-order dependencies, and evaluation on larger VLMs, stronger students, and fusion-based or generative tasks such as visual question answering and instruction tuning.

AI USE STATEMENT

In preparing this paper, we used ChatGPT and Claude to support both manuscript preparation and experimental development. Specifically, it was used to refine the writing, including grammar, spelling, word choice, and overall style. It was also used during experimentation to assist with identifying and debugging issues in Python code.

ETHICS STATEMENT

This work studies training objectives for distilling compact vision–language models and does not involve human subjects, user studies, or the collection of new data. All experiments use publicly available benchmarks (MSCOCO 2014, Flickr8k, Flickr30k, Food-101, and ImageNet-1k) and publicly released CLIP checkpoints, used under their respective licenses and only for research purposes. These datasets are known to contain demographic, geographic, and cultural biases, and the teacher models inherit biases from their web-scale pretraining data. Because distillation transfers the teacher's representation geometry to the student, and our objective explicitly encourages the student to follow the teacher's local geometry, it may propagate such biases rather than correct them; we do not claim that the distilled students are fairer, more robust, or better calibrated than their teachers. The primary intended benefit is more efficient multimodal models, which can lower the compute and energy cost of deployment and broaden access, but the same efficiency could also lower the barrier to misuse in surveillance or content analysis. We discuss these risks and possible safeguards, including fairness audits and transparent reporting, in Appendix I. We do not foresee harms specific to the proposed objective beyond those shared by knowledge distillation and CLIP-style models in general.

REPRODUCIBILITY STATEMENT

We have taken several steps to make our results reproducible. The complete code used for all experiments, in the form of Google Colab notebooks covering data preparation, distillation training with every loss combination reported in the paper, and retrieval and classification evaluation, is provided in the supplementary material; the notebooks run on the Colab GPU tiers listed in Appendix G.1 without additional infrastructure. The full training objective is given in Eq. (16), the TE-inspired proxies are defined in Sections 4.2.1 and 4.2.2, and Algorithm 1 in Appendix E specifies one complete training step, including how within-batch finite differences are formed. Student architectures, dataset splits, optimizer settings, default hyperparameters, the heuristic used to set loss weights, and the complete hyperparameter sweeps are reported in Appendix G.1, with per-configuration training times and GPU types listed alongside each experiment. All datasets (MSCOCO 2014, Flickr8k, Flickr30k, Food-101, and ImageNet-1k) and teacher checkpoints (CLIP RN50, ViT-B/16, and RN50$\times$16) are publicly available. To quantify variability, the main comparisons are repeated over five random seeds and reported as mean and standard deviation in Table 22 and Appendix Table 23.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018.

Sravanti Addepalli, Ashish Ramayee Asokan, Lakshay Sharma, and R Venkatesh Babu. Leveraging vision-language models for improving domain generalization in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23922–23932, 2024.

Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Jiajun Cao, Yuan Zhang, Tao Huang, Ming Lu, Qizhe Zhang, Ruichuan An, Ningning Ma, and Shanghang Zhang. Move-kd: Knowledge distillation for vlms with mixture of visual encoders. *CVPR*, 2025.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16296–16305, 2021.

Mengxi Chen, Linyu Xing, Yu Wang, and Ya Zhang. Enhanced multimodal representation learning with cross-modal kd. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11766–11775, 2023.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. In *European Conference on Computer Vision*, pp. 104–120. Springer, 2020.

Paweł Czyż, Frederic Grabowski, Julia Vogt, Niko Beerenwinkel, and Alexander Marx. Beyond normal: On the evaluation of mutual information estimators. *Advances in Neural Information Processing Systems*, 36:16957–16990, 2023.

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1428–1438, 2021.

Qianhan Feng, Wenshuo Li, Tong Lin, and Xinghao Chen. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language model. *CVPR*, 2025.

Chaochen Gao, Xing Wu, Peng Wang, Jue Wang, Liangjun Zang, Zhongyuan Wang, and Songlin Hu. Distilcse: Effective knowledge distillation for contrastive sentence embeddings. In *arXiv preprint arXiv:2112.05638*, 2021.

Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *Artificial intelligence and statistics*, pp. 277–286. PMLR, 2015.

Sahil Garg, Umang Gupta, Yu Chen, Syamantak Datta Gupta, Yeshaya Adler, Anderson Schneider, and Yuriy Nevmyvaka. Estimating transfer entropy under long ranged dependencies. In *Uncertainty in Artificial Intelligence*, pp. 685–695. PMLR, 2022.

Ziv Goldfeld and Kristjan Greenewald. Sliced mutual information: A scalable measure of statistical dependence. *Advances in Neural Information Processing Systems*, 34:17567–17578, 2021.

Ziv Goldfeld, Ewout van den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. *ICML*, 2019.

Gokul Gowri, Xiaokang Lun, Allon Klein, and Peng Yin. Approximating mutual information of high-dimensional variables using learned representations. *Advances in Neural Information Processing Systems*, 37:132843–132875, 2025.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

Jinyang Guo, Jiaheng Liu, Zining Wang, Yuqing Ma, Ruihao Gong, Ke Xu, and Xianglong Liu. Adaptive contrastive knowledge distillation for bert compression. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8941–8953, 2023.

Taku Hasegawa, Kyosuke Nishida, Koki Maeda, and Kuniko Saito. Duet: Image-text contrastive transfer learning with dual-adapter tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13607–13624, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9590–9601, 2024.

Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. In *Advances in Neural Information Processing Systems*, 2017.

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.

Mario Köppen. The curse of dimensionality. In *5th online world conference on soft computing in industrial applications (WSC5)*, volume 1, pp. 4–8, 2000.

Joon Lee, Shamim Nemati, Ikaro Silva, Bradley A Edwards, James P Butler, and Atul Malhotra. Transfer entropy estimation and directional coupling change detection in biomedical time series. *Biomedical engineering online*, 11:1–17, 2012.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16420–16429, 2022.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pp. 121–137. Springer, 2020.

Xuanlin Li, Yunhao Fang, Minghua Liu, Zhan Ling, Zhuowen Tu, and Hao Su. Distilling large vision-language model with out-of-distribution generalizability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2492–2503, 2023.

Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1234–1243, 2024a.

Zheng Li, Xiang Li, Lingfeng Yang, Renjie Song, Jian Yang, and Zhigeng Pan. Dual teachers for self-knowledge distillation. *Pattern Recognition*, 151:110422, 2024b.

Jessica E. Liang and Jianbo Shi. Transfer entropy as a measure of information flow in llms and vlms. OpenReview, 2026. URL `https://openreview.net/forum?id=1iXWOZYgHF`.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *European conference on computer vision (ECCV)*, pp. 740–755, 2014.

He Liu, Yikai Wang, Huaping Liu, Fuchun Sun, and Anbang Yao. Small scale data-free knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Omer Luxembourg, Dor Tsur, and Haim Permuter. Treet: Transfer entropy estimation via transformer. *arXiv preprint arXiv:2402.06919*, 2024.

Jian Ma. Estimating transfer entropy via copula entropy. *arXiv preprint arXiv:1910.04375*, 2019.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Rubén Gaspar Marco, Sofia Strukova, Félix Gómez Mármol, and José A Ruipérez-Valiente. Annotated flickr dataset for identification of professional photographers. *Data in Brief*, 50:109511, 2023.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884. PMLR, 2020.

Sachin Mehta, Marjan Ghazvininejad, Srinivasan Iyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Delight: Deep and light-weight transformer. *arXiv preprint arXiv:2008.00623*, 2020.

Roy Miles, Ismail Elezi, and Jiankang Deng. Vkd: Improving knowledge distillation using orthogonal projections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15720–15730, 2024.

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

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. *arXiv preprint arXiv:1803.10837*, 2018.

Zhiyuan Peng, Xuanji He, Ke Ding, Tan Lee, and Guanglu Wan. Label-free knowledge distillation with contrastive loss for light-weight speaker recognition. In *arXiv preprint arXiv:2212.03090*, 2022.

Cuong Pham, Van-Anh Nguyen, Trung Le, Dinh Phung, Gustavo Carneiro, and Thanh-Toan Do. Frequency attention for knowledge distillation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2277–2286, 2024.

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.

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.

Jakob Runge, Jobst Heitzig, Vladimir Petoukhov, and Jürgen Kurths. Escaping the curse of dimensionality in estimating multivariate transfer entropy. *Physical review letters*, 108(25):258701, 2012.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

Payam Shahsavari Baboukani, Carina Graversen, Emina Alickovic, and Jan Østergaard. Estimating conditional transfer entropy in time series using mutual information and nonlinear prediction. *Entropy*, 22(10):1124, 2020.

Yizheng Sun, Hao Li, ChengHua Lin, and Riza Batista-Navarro. Lanvikd: Cross-modal language-vision knowledge distillation for egocentric action recognition. *In Proceedings of the 1st International Workshop on Egocentric Perception, Interaction and Computing*, 2024.

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

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1365–1374, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Zhecan Wang, Noel Codella, Yen-Chun Chen, Luowei Zhou, Jianwei Yang, Xiyang Dai, Bin Xiao, Haoxuan You, Shih-Fu Chang, and Lu Yuan. Clip-td: Clip targeted distillation for vision-language tasks. *arXiv preprint arXiv:2201.05729*, 2022.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21970–21980, 2023.

Peipei Xia, Li Zhang, and Fanzhang Li. Learning similarity with cosine similarity ensemble. *Information sciences*, 307:39–52, 2015.

Xin Xiao, Bohong Wu, Jiacong Wang, Chunyuan Li, Xun Zhou, and Haoyuan Guo. Seeing the image: Prioritizing visual correlation by contrastive alignment. *arXiv preprint arXiv:2405.17871*, 2024.

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*, 2024.

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19163–19173, 2022.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4133–4141, 2017.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

Jian Zhang. Low-dimensional approximation searching strategy for transfer entropy from non-uniform embedding. *PloS one*, 13(3):e0194382, 2018.

Yuan Zhang, Tao Huang, Jiaming Liu, Tao Jiang, Kuan Cheng, and Shanghang Zhang. Freekd: Knowledge distillation via semantic frequency prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15931–15940, 2024.

Jie Zhu, Jean-Jacques Bellanger, Huazhong Shu, and Régine Le Bouquin Jeannès. Contribution to transfer entropy estimation via the k-nearest-neighbors approach. *Entropy*, 17(6):4173–4201, 2015.

Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Aijun Yang, Mingzhe Rong, and Xiaohua Wang. Complementary relation contrastive distillation. In *arXiv preprint arXiv:2103.16367*, 2021.

A Relations between Transfer Entropy, Entropy, and Mutual Information

For two discrete-time stochastic processes $X(t)$ and $Y(t)$, the transfer entropy from X to Y is formally defined as (Schreiber, 2000):

$$T_{X\to Y} = \sum_t p(y_{t+1}, y_t, x_t) \log \frac{p(y_{t+1} \mid y_t, x_t)}{p(y_{t+1} \mid y_t)}, \tag{17}$$

where $p(\cdot)$ represents probability distributions of the respective random variables.

Transfer entropy can also be expressed in terms of conditional entropy and mutual information (Shahsavari Baboukani et al., 2020). Specifically, the transfer entropy from X to Y, denoted $T_{X\to Y}$, measures the reduction in uncertainty about the future state Y_{t+1} given the joint past of X and Y, compared to the uncertainty given the past of Y alone. Based on (17), this difference can be expressed as:

$$\begin{aligned}
T_{X\to Y} &= \sum_t p(y_{t+1}, y_t, x_t) \log \frac{1}{p(y_{t+1} \mid y_t)} + \\
&\quad \sum_t p(y_{t+1}, y_t, x_t) \log p(y_{t+1} \mid y_t, x_t) \\
&= \sum_t p(y_{t+1}, y_t) \log \frac{1}{p(y_{t+1} \mid y_t)} + \\
&\quad \sum_t p(y_{t+1}, y_t, x_t) \log p(y_{t+1} \mid y_t, x_t) \\
&= H(Y_{t+1} \mid Y_t) - H(Y_{t+1} \mid Y_t, X_t) \qquad (18) \\
&= I(Y_{t+1}; X_t \mid Y_t) \qquad (19)
\end{aligned}$$

where $H(Y_{t+1} \mid Y_t)$ is the conditional entropy of Y_{t+1} given its own history Y_t, $H(Y_{t+1} \mid Y_t, X_t)$ is the conditional entropy of Y_{t+1} given both the history of Y and the history of X, and $I(Y_{t+1}; X_t \mid Y_t)$ is the mutual information between Y_{t+1} and the history of X, conditioned on the history of Y. In this formulation, the transfer entropy quantifies the amount by which the uncertainty about the future of Y is reduced by incorporating information from X.

B Prior Work on Mutual Information and Transfer Entropy Estimation

Mutual Information (MI) techniques have been employed to capture shared information between variables (Hjelm et al., 2018)(Oord et al., 2018)(Tian et al., 2020). MINE (Belghazi et al., 2018) offers a differentiable estimator for mutual information, and information-theoretic regularization has been applied in generative models for disentanglement and improved control (Chen et al., 2016). In (Gao et al., 2015), a mutual information estimator was proposed based on modified k-nearest neighbor (KNN) that is robust to local non-uniformity with limited data. A diverse set of distributions with known MI values were introduced to evaluate the performance of different MI estimators beyond traditional normal distributions (Czyż et al., 2023). McAllester and Stratos (McAllester & Stratos, 2020) highlighted the inherent difficulties in estimating mutual information from finite data, demonstrating that any distribution-free high-confidence lower bound on MI cannot exceed $O(\ln N)$, thereby underscoring the fundamental challenges in accurate mutual information estimation without strong assumptions about the data distribution. Goldfeld and Greenewald (Goldfeld & Greenewald, 2021) introduced Sliced Mutual Information, a scalable measure that projects high-dimensional distributions onto one-dimensional subspaces, effectively capturing complex dependencies while reducing computational complexity. Approximating mutual information of high-dimensional variables using learned representations was studied in (Gowri et al., 2025).

Transfer entropy is a conditional mutual information from two stochastic processes, so it's more challenging in TE estimation. In (Zhang, 2018), Low-dimensional approximation in the searching procedure was applied to transfer entropy from non-uniform embedding. In (Zhu et al., 2015), KNN

was used for TE estimation. However, KNN-based approach doesn't work well if the data are noisy and long ranged. To overcome this weakness, a perturbation model based on locality sensitive hash function was proposed for TE estimation (Garg et al., 2022). Three estimators were used for TE estimation (Lee et al., 2012), namely fixed-binning with ranking, kernel density estimation, and the Darbellay-Vajda (D-V) adaptive partitioning algorithm extended to three dimensions. In (Ma, 2019), copula entropy was applied to TE estimation. To overcome the curse of dimensionality in TE estimation, TE was decomposed into a sum of finite-dimensional contributions in (Runge et al., 2012). Recently, transformer was used for TE estimation (Luxembourg et al., 2024). In this paper, we propose TE approximation approaches which can tremendously reduce the computation cost and overcome the curse of dimensionality.

C Computational Cost Analysis: Exact TE versus TE-Inspired Proxies

Computational Complexity:

Direct estimation of transfer entropy in high-dimensional multimodal representation spaces is often computationally expensive. In general, estimating conditional mutual information or transfer entropy requires modeling dependencies among multiple variables, which may involve density estimation, kernel methods, or nearest-neighbor statistics whose cost scales poorly as the number of samples and feature dimensionality increase. This makes direct TE-style computation impractical for large-scale VLM distillation, where both the batch size and embedding dimension can be substantial.

In contrast, the TE-inspired proxy measures used in this paper are much simpler computationally. TE Proxy Method 1 in Section 4.2.1 uses squared cosine similarity between within-batch teacher and student finite differences. Concretely, after computing teacher and student embeddings for a batch, the method forms adjacent-pair differences and compares the one-dimensional subspaces they span, making the score invariant to orientation. This avoids constructing any joint probability model or density estimate. For a batch of size B and embedding dimension d, the additional computation scales approximately as $O(Bd)$, since it consists mainly of vector differences and cosine-similarity calculations on already-computed embeddings.

TE Proxy Method 2 in Section 4.2.2 similarly uses squared cosine similarity after concatenating image and text finite differences into a joint multimodal representation. This also avoids explicit TE estimation and remains linear in the effective feature dimension of the concatenated vectors. In practice, both methods add only lightweight operations on top of the standard forward pass of teacher and student encoders, making their cost negligible relative to the encoder computation itself.

Memory Usage: Exact TE-style estimation can require storing large collections of joint samples, pairwise distances, or density-estimation structures, and these costs can grow rapidly with dimensionality. This makes direct computation infeasible for high-dimensional embeddings, since it may demand substantial storage for probability tables, kernel statistics, or large neighborhood structures. The TE-inspired proxies mitigate this issue by avoiding explicit density estimation altogether. TE Proxy Method 1 stores only the teacher and student embeddings for the current batch, together with the corresponding within-batch difference vectors, keeping memory usage approximately at $O(Bd)$. TE Proxy Method 2 operates on the same batch features and their concatenated directional differences, leading to a comparable memory requirement. These proxy measures therefore enable scalable computation in large-scale deep learning settings without overwhelming memory constraints.

Scalability in High Dimensions: Direct TE estimation suffers from the curse of dimensionality. As dimensionality increases, conditional density estimation becomes increasingly unreliable because high-dimensional data points become sparse, making the estimated information-theoretic quantities unstable and sample-inefficient. In contrast, squared-cosine proxy measures are far more scalable, since they remain well-defined and easy to compute even in high dimensions. TE Proxy Method 1 relies only on orientation-invariant comparisons of teacher and student finite-difference subspaces within a batch, while TE Proxy Method 2 extends this idea to the joint multimodal space through concatenation. Because neither method requires explicit probabilistic modeling in the full feature space, both remain practical even for very large embeddings.

In summary, using TE-inspired cosine-similarity proxies enables analysis and regularization of high-dimensional, large-scale VLM distillation settings that would be computationally prohibitive for direct TE estimation. These proxy measures significantly improve computational feasibility and remain stable in settings where exact information-theoretic estimation would be difficult or unreliable. The trade-off, however, is that they provide only a geometric surrogate rather than an exact measure of information flow, and may therefore overlook more complex nonlinear dependencies in the data. Since the primary goal in CLIP-style distillation is to handle rich multimodal embeddings efficiently while obtaining a practical teacher-consistent alignment signal, these TE-inspired proxy measures provide a useful and scalable alternative to direct TE computation.

D Loss Functions in VLM Distillation

D.1 Logit Representation in VLM Distillation

In our framework, logits represent the similarity scores between image and text embeddings, which are fundamental to contrastive learning. Given a batch of image-text pairs, let $\mathbf{v}^{(S)}, \mathbf{u}^{(S)}$ denote the image and text embeddings from the student model, and $\mathbf{v}^{(T)}, \mathbf{u}^{(T)}$ denote the corresponding embeddings from the teacher model. The logit computation follows these steps.

First, we normalize the embeddings to unit norm:

$$\hat{\mathbf{v}}^{(S)} = \frac{\mathbf{v}^{(S)}}{\|\mathbf{v}^{(S)}\|_2}, \quad \hat{\mathbf{u}}^{(S)} = \frac{\mathbf{u}^{(S)}}{\|\mathbf{u}^{(S)}\|_2}, \tag{20}$$

$$\hat{\mathbf{v}}^{(T)} = \frac{\mathbf{v}^{(T)}}{\|\mathbf{v}^{(T)}\|_2}, \quad \hat{\mathbf{u}}^{(T)} = \frac{\mathbf{u}^{(T)}}{\|\mathbf{u}^{(T)}\|_2}. \tag{21}$$

The similarity logits for the student and teacher models are then computed as the dot product between the corresponding image and text embeddings, scaled by a temperature parameter τ:

$$\mathbf{z}^{(S)} = \frac{\hat{\mathbf{v}}^{(S)} \cdot (\hat{\mathbf{u}}^{(S)})^\top}{\tau}, \quad \mathbf{z}^{(T)} = \frac{\hat{\mathbf{v}}^{(T)} \cdot (\hat{\mathbf{u}}^{(T)})^\top}{\tau}. \tag{22}$$

Here, $\mathbf{z}^{(S)}$ and $\mathbf{z}^{(T)}$ are $|B| \times |B|$ matrices, where each entry $z_{ij}^{(S)}$ represents the similarity between the i-th image embedding and the j-th text embedding in the batch for the student model, and similarly for the teacher model. The temperature parameter τ controls the sharpness of the similarity distribution, with lower values making the distribution more peaky.

These logits are subsequently used in the contrastive loss and KL divergence computation to align the student's feature representations with those of the teacher, ensuring effective knowledge transfer during distillation. Several studies have explored the computation and utilization of these logits in image-text contrastive frameworks (Radford et al., 2021)(Jia et al., 2021)(Yang et al., 2022)(Hasegawa et al., 2023)(Xiao et al., 2024).

D.2 Contrastive Loss for VLM Distillation

We employ a contrastive loss based on the InfoNCE loss formulation to align the student model's image and text representations effectively. Given a batch of $|B|$ image-text pairs, we define the contrastive loss using the computed logits. The contrastive loss for image-to-text alignment is defined as (Oord et al., 2018):

$$\mathcal{L}_{I \to T} = -\frac{1}{|B|} \sum_{k=1}^{|B|} \log \frac{\exp(z_{kk}^{(S)})}{\sum_{j=1}^{|B|} \exp(z_{kj}^{(S)})} \tag{23}$$

$z_{kj}^{(S)}$ represents the similarity between the k-th image embedding and the j-th text embedding in the batch for the student model, and $z_{kk}^{(S)}$ represents the similarity logit between the k-th image and its corresponding text in the batch for the student model.

Similarly, the contrastive loss for text-to-image alignment is given by:

$$\mathcal{L}_{T\to I} = -\frac{1}{|B|}\sum_{k=1}^{|B|}\log\frac{\exp(z_{kk}^{(S)})}{\sum_{j=1}^{|B|}\exp(z_{jk}^{(S)})} \tag{24}$$

$z_{jk}^{(S)}$ represents the similarity between the j-th image embedding and the k-th text embedding in the batch for the student model, and $z_{kk}^{(S)}$ is the same as that in (23).

The total contrastive loss, which balances both image-to-text and text-to-image objectives, is computed as:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2}(\mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}). \tag{25}$$

This loss function encourages the student model to align its multi-modal representations by bringing matching pairs closer in the embedding space while pushing apart non-matching pairs. Contrastive loss has been extensively applied to knowledge distillation (Chen et al., 2021)(Gao et al., 2021)(Peng et al., 2022)(Zhu et al., 2021)(Guo et al., 2023).

To enhance the effectiveness of distillation, we extend this contrastive loss with additional terms such as KL divergence and transfer entropy-based regularization. These terms further refine the student model's learning dynamics by ensuring information flow from the teacher's embeddings to the student's representations while preserving structural consistency across modalities.

D.3 KL Divergence for VLM Distillation

To ensure that the student model effectively mimics the probability distributions of the teacher model, we include a Kullback-Leibler (KL) divergence loss term. KL divergence measures how much the student's predicted distribution deviates from the teacher's distribution, enforcing a closer alignment between their logits. KL divergence has been applied to VLM distillation (Li et al., 2024a)(Sun et al., 2024).

For a given batch of image-text pairs, let $\mathbf{z}^{(S)}$ and $\mathbf{z}^{(T)}$ represent the similarity logits of the student and teacher models, respectively. The soft probability distributions are obtained via the softmax function:

$$P_i^{(S)} = \frac{\exp(z_i^{(S)}/\tau)}{\sum_{j=1}^{|B|}\exp(z_j^{(S)}/\tau)}, \tag{26}$$

$$P_i^{(T)} = \frac{\exp(z_i^{(T)}/\tau)}{\sum_{j=1}^{|B|}\exp(z_j^{(T)}/\tau)}, \tag{27}$$

where τ is the temperature parameter that controls the sharpness of the distributions.

The KL divergence loss is computed as:

$$\mathcal{L}_{\text{KL}} = \frac{1}{2}\left(D_{\text{KL}}\left(P_{\text{image}}^{(T)} \parallel P_{\text{image}}^{(S)}\right) + D_{\text{KL}}\left(P_{\text{text}}^{(T)} \parallel P_{\text{text}}^{(S)}\right)\right), \tag{28}$$

where the teacher-to-student KL divergence is defined as:

$$D_{\text{KL}}(P^{(T)} \parallel P^{(S)}) = \sum_{i=1}^{|B|} P_i^{(T)} \log\frac{P_i^{(T)}}{P_i^{(S)}}. \tag{29}$$

This loss encourages the student model to produce probability distributions that closely resemble those of the teacher, effectively preserving the knowledge distilled from the teacher while allowing the student to generalize efficiently.

D.4 MSE Loss Function for VLM Distillation

To further align the feature representations of the teacher and student models, we include MSE loss that minimizes the discrepancy between their intermediate embeddings (Yang et al., 2024).

The MSE loss is computed as the sum of the squared differences between the student and teacher embeddings for both modalities:

$$\mathcal{L}_{\text{MSE}} = \mathcal{L}_{\text{MSE}}^{\text{image}} + \mathcal{L}_{\text{MSE}}^{\text{text}}, \tag{30}$$

where

$$\mathcal{L}_{\text{MSE}}^{\text{image}} = \frac{1}{|B|} \sum_{i=1}^{|B|} \left\| \hat{\mathbf{v}}_i^{(S)} - \hat{\mathbf{v}}_i^{(T)} \right\|^2, \tag{31}$$

$$\mathcal{L}_{\text{MSE}}^{\text{text}} = \frac{1}{|B|} \sum_{i=1}^{|B|} \left\| \hat{\mathbf{u}}_i^{(S)} - \hat{\mathbf{u}}_i^{(T)} \right\|^2. \tag{32}$$

Here, $|B|$ represents the batch size, and $\|\cdot\|^2$ denotes the squared Euclidean norm. This loss ensures that the student model's learned embeddings remain close to the teacher's representations in the feature space, facilitating effective knowledge transfer. MSE has been applied to VLM loss function in (Yang et al., 2024), and was called feature distillation.

D.5 Interactive Contrastive Learning

Interactive Contrastive Learning (ICL) was proposed in (Yang et al., 2024) to aligns the student model's feature representations with those of the teacher by treating the student embeddings as anchors and contrasting them with the teacher embeddings.

Given a batch of image-text pairs, let $\mathbf{v}_k^{(S)}$ be the image embedding from the student model, and $\{\mathbf{u}_b^{(T)}\}_{b=1}^{|B|}$ denote the contrastive text embeddings from the teacher model. The image-to-text ICL loss is formulated as:

$$\mathcal{L}_{\text{ICL}}^{I \to T} = -\log \frac{\exp(\mathbf{v}_k^{(S)} \cdot \mathbf{u}_k^{(T)}/\tau)}{\sum_{b=1}^{|B|} \exp(\mathbf{v}_k^{(S)} \cdot \mathbf{u}_b^{(T)}/\tau)}, \tag{33}$$

where τ is the temperature parameter.

Similarly, for a student text embedding $\mathbf{u}_k^{(S)}$ and contrastive image embeddings from the teacher model $\{\mathbf{v}_b^{(T)}\}_{b=1}^{|B|}$, the text-to-image ICL loss is:

$$\mathcal{L}_{\text{ICL}}^{T \to I} = -\log \frac{\exp(\mathbf{u}_k^{(S)} \cdot \mathbf{v}_k^{(T)}/\tau)}{\sum_{b=1}^{|B|} \exp(\mathbf{u}_k^{(S)} \cdot \mathbf{v}_b^{(T)}/\tau)}. \tag{34}$$

The final ICL loss is a combination of the two:

$$\mathcal{L}_{\text{ICL}} = \frac{1}{2} \left(\mathcal{L}_{\text{ICL}}^{I \to T} + \mathcal{L}_{\text{ICL}}^{T \to I} \right). \tag{35}$$

By integrating ICL, the student model effectively learns from the teacher's structured feature space, leading to improved representation learning and knowledge transfer.

D.6 Mutual-Information (MI) Objective

In addition to the losses above, we also evaluate mutual-information objectives as an alternative teacher-student supervision signal. This baseline helps distinguish improvements due to directed, conditional information-flow regularization from improvements that may arise from standard contrastive representation matching.

MI baseline loss (CRD-style). We implement an MI-driven alignment loss following the contrastive mutual-information maximization principle used in Contrastive Representation Distillation (CRD) (Tian et al., 2020). Using the same notation as above, given a batch of $|B|$ image-text pairs, let $\mathbf{v}_k^{(S)}$ and $\mathbf{u}_k^{(S)}$ denote the student image and text embeddings, and $\mathbf{v}_k^{(T)}$ and $\mathbf{u}_k^{(T)}$ the corresponding teacher embeddings. We ℓ_2-normalize the embeddings, yielding $\hat{\mathbf{v}}_k^{(\cdot)}$ and $\hat{\mathbf{u}}_k^{(\cdot)}$, and use temperature τ.

Unlike the CLIP-style contrastive loss in Eq. (23) and ICL (Yang et al., 2024), which both involve *cross-modal* matching (image↔text) either within the student or across teacher-student, this MI baseline performs *modality-matched* teacher-student alignment (image↔image and text↔text). Concretely, the image-side MI loss treats $(\hat{\mathbf{v}}_k^{(T)}, \hat{\mathbf{v}}_k^{(S)})$ as the positive pair and $(\hat{\mathbf{v}}_k^{(T)}, \hat{\mathbf{v}}_b^{(S)})$ for $b \neq k$ as negatives:

$$\mathcal{L}_{\mathrm{MI}}^{V} = -\frac{1}{|B|}\sum_{k=1}^{|B|} \log \frac{\exp\left(\hat{\mathbf{v}}_k^{(T)} \cdot \hat{\mathbf{v}}_k^{(S)}/\tau\right)}{\sum_{b=1}^{|B|} \exp\left(\hat{\mathbf{v}}_k^{(T)} \cdot \hat{\mathbf{v}}_b^{(S)}/\tau\right)}. \tag{36}$$

Similarly, the text-side MI loss treats $(\hat{\mathbf{u}}_k^{(T)}, \hat{\mathbf{u}}_k^{(S)})$ as the positive pair:

$$\mathcal{L}_{\mathrm{MI}}^{U} = -\frac{1}{|B|}\sum_{k=1}^{|B|} \log \frac{\exp\left(\hat{\mathbf{u}}_k^{(T)} \cdot \hat{\mathbf{u}}_k^{(S)}/\tau\right)}{\sum_{b=1}^{|B|} \exp\left(\hat{\mathbf{u}}_k^{(T)} \cdot \hat{\mathbf{u}}_b^{(S)}/\tau\right)}. \tag{37}$$

The final MI objective is the symmetric combination

$$\mathcal{L}_{\mathrm{MI}} = \tfrac{1}{2}\left(\mathcal{L}_{\mathrm{MI}}^{V} + \mathcal{L}_{\mathrm{MI}}^{U}\right). \tag{38}$$

Minimizing $\mathcal{L}_{\mathrm{MI}}$ encourages matched teacher-student representations to score higher than mismatched pairs within the batch, which corresponds to maximizing an InfoNCE lower bound on mutual information (Tian et al., 2020).

D.7 MSE-Difference Objective

MSE-difference baseline loss. To mirror the finite-difference construction used by TE1/TE2, we use the same adjacent pairs from the randomly permuted mini-batch:

$$\Delta\mathbf{v}_i^{(T)} = \mathbf{v}_{i+1}^{(T)} - \mathbf{v}_i^{(T)}, \qquad \Delta\mathbf{v}_i^{(S)} = \mathbf{v}_{i+1}^{(S)} - \mathbf{v}_i^{(S)}, \tag{39}$$

with analogous definitions for $\Delta\mathbf{u}_i^{(T)}$ and $\Delta\mathbf{u}_i^{(S)}$. We then define

$$\mathcal{L}_{\text{MSE-}\Delta} = \frac{1}{2}\Big(\mathbb{E}_i\Big[\|\Delta\mathbf{v}_i^{(T)} - \Delta\mathbf{v}_i^{(S)}\|_2^2\Big] + \mathbb{E}_i\Big[\|\Delta\mathbf{u}_i^{(T)} - \Delta\mathbf{u}_i^{(S)}\|_2^2\Big]\Big). \tag{40}$$

This baseline therefore uses the same sampled secants as TE1/TE2 but penalizes Euclidean mismatch in both direction and magnitude, rather than the orientation-invariant angular geometry captured by squared cosine similarity.

Geometric interpretation of the squared-cosine proxy. The proposed proxy uses squared cosine similarity rather than ordinary cosine similarity. Consequently, it matches finite-difference directions up to orientation: parallel and anti-parallel vectors receive the same score. For nonzero vectors $\mathbf{a}$ and $\mathbf{b}$, let $\hat{\mathbf{a}} = \mathbf{a}/\|\mathbf{a}\|_2$ and $\hat{\mathbf{b}} = \mathbf{b}/\|\mathbf{b}\|_2$. Then

$$\left\|\hat{\mathbf{a}}\hat{\mathbf{a}}^\top - \hat{\mathbf{b}}\hat{\mathbf{b}}^\top\right\|_F^2 = 2\left(1 - (\hat{\mathbf{a}}^\top\hat{\mathbf{b}})^2\right) = 2\left(1 - \cos^2(\mathbf{a},\mathbf{b})\right). \tag{41}$$

Thus, maximizing the squared-cosine proxy is equivalent, up to an affine transformation, to minimizing the distance between the rank-one projectors onto the one-dimensional subspaces spanned by the teacher and student finite-difference vectors. Unlike raw MSE-Δ, this criterion is invariant to vector magnitude and orientation and focuses on matching local secant-line geometry.

E Description of a Full Training Step with TE Surrogates

One training step with TE surrogates (within-batch finite differences). Algorithm 1 summarizes one full distillation iteration and makes explicit how the TE-inspired surrogate terms (TE1/TE2) are computed in practice. At each iteration, we sample a mini-batch $\mathcal{B} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{B}$ and perform one student parameter update using Adam. The teacher f_T is frozen, so its embeddings provide a stable reference geometry for the current batch, while the student f_S is updated by backpropagation through the full objective.

Algorithm 1: One distillation step with TE1/TE2 (within-batch finite differences)

Input: Student f_S, frozen teacher f_T, batch $\mathcal{B} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{B}$, Adam optimizer with learning rate ρ
Output: Updated student parameters θ_S
Sample mini-batch $\mathcal{B}$;
Sample a random permutation π over $\{1, \dots, B\}$;
Form permuted batch $\mathcal{B}_\pi = \{(\mathbf{x}_{\pi(i)}, y_{\pi(i)})\}_{i=1}^{B}$;
Teacher forward (once, no grad): $(\mathbf{v}^{(T)}, \mathbf{u}^{(T)}) \leftarrow f_T(\mathcal{B}_\pi)$;
Detach teacher features;
Student forward: $(\mathbf{v}^{(S)}, \mathbf{u}^{(S)}) \leftarrow f_S(\mathcal{B}_\pi)$;
Compute base distillation loss

$$\mathcal{L}_{\text{base}} \leftarrow \mathcal{L}_{\text{CL}} + \alpha\mathcal{L}_{\text{KL}} + \beta\mathcal{L}_{\text{MSE}} + \delta\mathcal{L}_{\text{ICL}} + \eta\mathcal{L}_{\text{MI}} + \lambda\mathcal{L}_{\text{MSE-}\Delta}$$

`// Within-batch finite differences using adjacent pairs after permutation`
for $i = 1$ **to** $B - 1$ **do**
  $\Delta\mathbf{v}_i^{(S)} \leftarrow \mathbf{v}_{i+1}^{(S)} - \mathbf{v}_i^{(S)}$;
  $\Delta\mathbf{u}_i^{(S)} \leftarrow \mathbf{u}_{i+1}^{(S)} - \mathbf{u}_i^{(S)}$;
  $\Delta\mathbf{v}_i^{(T)} \leftarrow \mathbf{v}_{i+1}^{(T)} - \mathbf{v}_i^{(T)}$;
  $\Delta\mathbf{u}_i^{(T)} \leftarrow \mathbf{u}_{i+1}^{(T)} - \mathbf{u}_i^{(T)}$;
`// TE1: per-modality alignment`
$\text{TE}_{\text{img}} \leftarrow \frac{1}{B-1}\sum_{i=1}^{B-1} \cos^2(\Delta\mathbf{v}_i^{(S)}, \Delta\mathbf{v}_i^{(T)})$;
$\text{TE}_{\text{txt}} \leftarrow \frac{1}{B-1}\sum_{i=1}^{B-1} \cos^2(\Delta\mathbf{u}_i^{(S)}, \Delta\mathbf{u}_i^{(T)})$;
$\text{TE1} \leftarrow \frac{1}{2}\big(\text{TE}_{\text{img}} + \text{TE}_{\text{txt}}\big)$;
`// TE2: joint multimodal alignment`
for $i = 1$ **to** $B - 1$ **do**
  $\Delta\mathbf{c}_i^{(S)} \leftarrow [\Delta\mathbf{v}_i^{(S)} \,\|\, \Delta\mathbf{u}_i^{(S)}]$;
  $\Delta\mathbf{c}_i^{(T)} \leftarrow [\Delta\mathbf{v}_i^{(T)} \,\|\, \Delta\mathbf{u}_i^{(T)}]$;
$\text{TE2} \leftarrow \frac{1}{B-1}\sum_{i=1}^{B-1} \cos^2(\Delta\mathbf{c}_i^{(S)}, \Delta\mathbf{c}_i^{(T)})$;
Set total loss: $\mathcal{L} \leftarrow \mathcal{L}_{\text{base}} - \gamma(\text{TE1} + \text{TE2})$;
Backpropagate $\mathcal{L}$ and update θ_S using Adam with learning rate ρ;

Crucially, TE1/TE2 do not require any additional teacher evaluation beyond standard distillation. We run the teacher forward once, without gradients, on the current batch and detach the resulting embeddings $(\mathbf{v}^{(T)}, \mathbf{u}^{(T)})$. We also run the student forward once on the same permuted batch to obtain $(\mathbf{v}^{(S)}, \mathbf{u}^{(S)})$. To construct finite-difference proxies for local secant geometry, we sample a random permutation π of the batch indices and reorder the batch as $\mathcal{B}_\pi$. This permutation is used only to define a diverse set of within-batch finite-difference pairs; it does not impose any temporal or semantic ordering on the samples.

We then form within-batch differences using adjacent pairs in the permuted order: $\Delta\mathbf{v}_i = \mathbf{v}_{i+1} - \mathbf{v}_i$ and $\Delta\mathbf{u}_i = \mathbf{u}_{i+1} - \mathbf{u}_i$ for both teacher and student. These differences should be interpreted as sampled local secants in embedding space. The index i enumerates permuted samples and has no temporal meaning; accordingly, TE1/TE2 are not temporal transfer-entropy estimators computed from shuffled data. Instead, TE1 measures per-modality orientation-invariant alignment through squared cosine similarity, while TE2 applies the same criterion after concatenating image and text finite differences. Together, these surrogate rewards encourage the student to match the teacher's local secant-line geometry on each batch.

Algorithm 1 presents the most general form in which both TE1 and TE2 are included in the loss. TE1-only and TE2-only ablations are obtained by setting the other surrogate term to zero.

Compute footprint. Per iteration, we run the teacher forward once without gradients and the student forward once with gradients, as in standard distillation. TE1/TE2 add only lightweight vector-difference and cosine-similarity computations on already-computed embeddings, and therefore introduce negligible overhead: no extra teacher forward passes, no Jacobian construction, and no cache-and-replay procedure.

We confirm that all reported results in this paper were obtained using the implementation described in Algorithm 1. The corresponding code is provided in the supplementary material to support reproducibility and to eliminate ambiguity about the training procedure.

F Brief Introduction to the Teacher CLIP Models

We briefly summarize the CLIP architectures used in our experiments.

CLIP ResNet-50 (RN50). CLIP RN50 adopts a ResNet-50 backbone as the visual encoder and a Transformer-based text encoder (Radford et al., 2021). The image encoder processes 224×224 RGB inputs and produces a 1024-dimensional global image embedding, while the text encoder is a 12-layer Transformer (Vaswani et al., 2017) with hidden dimension 512. In the original CLIP implementation, the image encoder has roughly 38M parameters and the text encoder about 63M parameters, for a total of approximately 102M parameters (Radford et al., 2021). In our experiments, CLIP RN50 serves as a moderately sized vision–language teacher model that is well suited for distillation into smaller students.

CLIP ResNet-50$\times$16 (RN50$\times$16). RN50$\times$16 is a higher-capacity variant of CLIP's ResNet family (Radford et al., 2021). It scales up RN50 using a ResNeXt-style design with 16$\times$ wider convolutional blocks, substantially increasing the number of visual parameters and intermediate feature capacity. The released RN50$\times$16 checkpoint uses the same CLIP dual-encoder training paradigm but a higher-capacity model configuration than RN50, yielding stronger zero-shot performance and a richer multimodal embedding space.

CLIP ViT-B/16. CLIP ViT-B/16 replaces the convolutional backbone with a Vision Transformer (ViT-B/16) (Dosovitskiy et al., 2020; Radford et al., 2021). Images are divided into non-overlapping 16×16 patches, which are linearly projected and fed into a 12-layer Transformer encoder with hidden dimension 768 before the final CLIP projection head, producing a global image representation. The text side uses the same 12-layer Transformer encoder with hidden dimension 512 as in RN50. The resulting dual-encoder model has approximately 151M parameters (about 86M in the visual encoder and 63M in the text encoder), and typically achieves stronger zero-shot performance than RN50 due to improved global reasoning and larger effective receptive fields.

Shared CLIP training paradigm. Across all three architectures, CLIP jointly trains the vision and text encoders with a contrastive image–text objective so that paired image–text embeddings are close in cosine-similarity space while unpaired pairs are pushed apart (Radford et al., 2021). This unified training scheme enables zero-shot recognition by comparing image embeddings to text-encoded label prompts at inference time, and provides a natural teacher for distilling smaller vision–language students.

G Additional Experimental Results

This appendix collects implementation details, the full hyperparameter sweeps, training dynamics, representation-level visualizations, the complete versions of the ablations summarized in the main text, and additional teacher–student–dataset configurations.

G.1 Implementation Details and Hyperparameter Selection

Student architecture. The student VLM consists of an RN34 image encoder and a lightweight Transformer text encoder. For the RN50$\rightarrow$RN34 configuration, the RN34 backbone contains approximately 21.8 million parameters and its final projection outputs 1024-dimensional features (He

et al., 2016). The text encoder uses a vocabulary of size 49,408 and hidden dimension 1024, contributing approximately 25.3 million parameters (Mehta et al., 2020); it has 2 encoder layers with 8-head attention (Vaswani et al., 2017). For cross-architecture teacher configurations, the teacher and student outputs are represented in a common comparison dimension before MSE, ICL, MI, MSE-Δ, TE1, and TE2 are computed. Combining the image and text encoders, the total parameter count of the RN34-based student is approximately 55–60 million, substantially smaller than the teacher while retaining sufficient capacity for multimodal distillation. The RN18 student is described in Appendix G.9.

Datasets. MSCOCO 2014 (Lin et al., 2014) contains 82,783 training images and 40,504 validation images, each paired with multiple textual descriptions, and is widely used in vision–language research because of its large scale and diverse image–caption pairs. For Flickr8k (Marco et al., 2023) we use 85% of the data for training and 15% for testing. Food-101 (Bossard et al., 2014) contains 101 categories with 101,000 images, split into 75,750 training and 25,250 test images.

Training and compute. The student is trained with combinations of CL, KL, MSE, ICL, MI, MSE-Δ, and the TE-inspired proxies according to Eq. (16); we study multiple combinations to evaluate how the proxy interacts with standard distillation losses. The training data is shuffled at each epoch. Unless otherwise specified, we use Adam with learning rate $\rho = 10^{-4}$, $\alpha = 1.0$, $\beta = 50$, $\delta = 1.0$, $\tau = 0.07$, $|B| = 64$, and 10 epochs per configuration. The RN50$\rightarrow$RN34 MSCOCO experiments were run on Google Colab Pro with a T4 GPU (High-RAM); each training run plus evaluation took approximately 14 hours. Compute for the other configurations is reported in the corresponding subsections below.

Loss-weight heuristic. The weights $\alpha, \beta, \delta, \eta, \lambda$, and γ control the relative influence of the KL, MSE (feature regression), ICL, MI, MSE-Δ, and TE surrogate terms with respect to the base contrastive objective. By "assigning larger weights to losses with smaller magnitudes," we use the following heuristic: during a short warm-up run (approximately the first epoch) with a fixed reference setting, we log the *batch-averaged* value of each unweighted component $\{\mathcal{L}_{\mathrm{CL}}, \mathcal{L}_{\mathrm{KL}}, \mathcal{L}_{\mathrm{MSE}}, \mathcal{L}_{\mathrm{ICL}}, \mathcal{L}_{\mathrm{MI}}, \mathcal{L}_{\mathrm{MSE}\Delta}, \mathrm{TE}_1, \mathrm{TE}_2\}$, and choose weights so that the corresponding *weighted* components have roughly comparable scale:

$$\alpha\,\mathbb{E}[\mathcal{L}_{\mathrm{KL}}] \approx \beta\,\mathbb{E}[\mathcal{L}_{\mathrm{MSE}}] \approx \delta\,\mathbb{E}[\mathcal{L}_{\mathrm{ICL}}] \approx \eta\,\mathbb{E}[\mathcal{L}_{\mathrm{MI}}] \approx \lambda\,\mathbb{E}[\mathcal{L}_{\mathrm{MSE}\Delta}] \approx \gamma\,\mathbb{E}[\mathrm{TE}] \approx \mathbb{E}[\mathcal{L}_{\mathrm{CL}}], \quad (42)$$

where TE denotes the bounded cosine-based surrogate (TE_1/TE_2). Intuitively, when a term is naturally small, such as the MSE between normalized embeddings, a larger coefficient is needed for it to meaningfully affect optimization; conversely, terms already on the same scale as $\mathcal{L}_{\mathrm{CL}}$ require smaller coefficients.

Grid search and full sweep. After setting a reasonable scale using Eq. (42), we perform a small grid search around these values and select the final configuration by validation Recall@K on MSCOCO. Table 5 summarizes this sweep. In the main RN50$\rightarrow$RN34 MSCOCO setting, we fix $\alpha = 1$ and $\delta = 1$, use $\beta \in \{50, 100\}$ depending on the experiment, and tune γ over $\{1, 2.5, 5, 7.5, 10\}$, with $\gamma = 7.5$ giving the strongest I2T performance. For MI baselines, we tune η analogously, typically over $\{1, 2.5, 5, 7.5\}$, and we set $\lambda = 0$ unless explicitly evaluating the MSE-Δ ablation.

Table 5 reports the sensitivity of retrieval performance to different hyperparameter settings. Overall, the method is reasonably stable under moderate changes in loss weights, while the TE reward weight γ has the clearest effect on retrieval quality. Starting from the baseline without TE ($\gamma = 0$), introducing a nonzero TE weight improves retrieval in both directions, and Image-to-Text performance generally improves as γ increases from 1 to 7.5. In particular, $\gamma = 7.5$ achieves the best I2T result, reaching R@1/R@5/R@10 of $10.27/26.36/36.30$. Increasing γ further to 10 yields slightly weaker performance, suggesting diminishing returns and mild over-regularization when the TE reward is too strong.

By comparison, varying α (KL), β (MSE), and δ (ICL) within the tested ranges leads to relatively smaller fluctuations, indicating that the objective is not overly sensitive to these terms once they are set to reasonable values. Adding the MI loss also improves retrieval, especially for T2I: setting $\eta = 2.5$ with $\gamma = 0$ yields the strongest T2I performance in Table 5, reaching $7.66/20.69/29.25$.

Table 5: Comparison of retrieval performance (Recall@k) in percentage of an RN34-based VLM student distilled from CLIP RN50 on MSCOCO.

α	β	δ	η	λ	γ	I2T R@1	I2T R@5	I2T R@10	T2I R@1	T2I R@5	T2I R@10
1	50	1	0	0	0	6.08%	18.14%	26.93%	5.92%	17.06%	24.89%
1	50	1	0	0	1	8.11%	22.05%	31.57%	7.18%	19.75%	28.14%
1	50	1	0	0	2.5	9.48%	24.68%	34.54%	7.57%	20.55%	29.03%
1	50	1	0	0	5	10.02%	25.80%	35.74%	7.32%	20.04%	28.50%
1	50	1	0	0	7.5	**10.27%**	**26.36%**	**36.30%**	6.98%	18.95%	26.95%
1	50	1	0	0	10	9.77%	25.25%	35.04%	6.94%	18.78%	26.67%
5	50	1	0	0	5	10.19%	26.15%	36.09%	6.81%	18.84%	26.88%
1	100	1	0	0	5	10.25%	25.76%	35.69%	7.30%	20.55%	29.03%
1	50	5	0	0	5	8.20%	22.14%	31.63%	7.14%	19.61%	28.05%
1	50	1	0	50	0	6.67%	18.48%	27.30%	6.01%	17.27%	25.24%
1	50	1	0	100	0	6.52%	18.90%	27.57%	6.29%	17.71%	25.67%
1	50	1	1	1	0	8.35%	22.62%	31.70%	7.43%	20.23%	28.87%
1	50	1	2.5	0	0	8.56%	23.01%	32.41%	**7.66%**	**20.69%**	**29.25%**
1	50	1	5	0	0	8.65%	23.29%	33.05%	7.65%	20.63%	29.05%
1	50	1	7.5	0	0	8.25%	22.41%	32.04%	7.44%	20.27%	28.61%
1	50	1	2.5	50	2.5	9.12%	24.25%	33.88%	7.31%	19.98%	28.40%
1	50	1	2.5	0	5	9.94%	25.73%	35.54%	6.87%	18.92%	27.09%
1	50	1	2.5	50	5	9.26%	24.26%	34.44%	7.05%	19.47%	27.78%
1	50	1	1	0	7.5	9.67%	24.85%	34.78%	6.91%	18.95%	27.07%
1	50	1	1	50	7.5	10.23%	25.93%	35.81%	6.87%	18.91%	27.04%
1	50	1	5	50	7.5	9.03%	24.15%	33.87%	6.85%	18.80%	26.77%

In contrast, using the MSE-Δ loss alone ($\lambda \in \{50, 100\}$ with $\gamma = 0$) provides only marginal gains over the baseline, and combining multiple auxiliary terms does not always produce additive improvements. Overall, these results suggest that moderate TE weighting is the most effective way to strengthen I2T retrieval in this setting, while MI-based supervision remains competitive for T2I.

G.2 Training Dynamics

Figure 3 shows the training losses and TE proxy values for representative loss combinations in RN50→RN34 distillation on MSCOCO at $\tau = 0.07$. The total training loss generally decreases over epochs, while the TE surrogate rewards tend to increase. This indicates that the optimization process is successfully reducing the overall objective while increasing teacher–student geometric alignment under the proxy terms. For experiments that include both TE1 and TE2, such as Fig. 3h and Fig. 3i, the two surrogate rewards increase monotonically with similar but not identical trajectories. At the same time, KL and MSE do not always decrease monotonically, suggesting that the different objective components interact in nontrivial ways during training. We therefore interpret these curves as training-dynamics diagnostics rather than as evidence that each individual component should decrease monotonically in isolation.

The interactions among CL, KL, ICL, MSE, and the TE-inspired proxy terms are also reflected in the final surrogate values. For example, in `CL - TE1` (Fig. 3d), TE1 reaches an average value of 0.7242 at epoch 10; in `CL + KL - TE1` (Fig. 3a), it reaches 0.7865; while in `CL + KL + ICL - TE1` (Fig. 3e), it reaches 0.4611. These comparisons suggest that adding KL is associated with higher TE1 values in this setting, whereas adding ICL on top of CL+KL is associated with lower TE1 values. We emphasize that these observations are empirical properties of the training objective in this experimental regime, rather than general causal claims about the losses themselves.

Effect of the contrastive temperature. The behavior of KL and MSE is strongly recipe-dependent, especially with respect to the contrastive temperature. In particular, the behavior of KL and MSE is strongly recipe-dependent, especially with respect to the contrastive temperature. In the original $\tau = 0.07$ setting (e.g., Fig. 3b), the KL term decreases only slightly, from 0.4721 to 0.4662 (about 1.25%), whereas the MSE term decreases from 5.93×10^{-4} to 3.19×10^{-4} (about 46.2%). This previously suggested that direct feature-level alignment was easier to optimize than similarity-distribution matching (e.g., KL). However, the lower-temperature results in Fig. 4 show that this interpretation is not universal: at $\tau = 0.03$, both TE- and MI-based objectives exhibit substantial KL reduction. In particular, under TE-based training ($\gamma = 7.5$), KL decreases from 0.5800 to 0.3407 (a 41.3% reduction), while MSE decreases from 5.48×10^{-4} to 1.78×10^{-4} (a 67.5% reduction). Under MI-based training ($\eta = 5$), KL decreases from 0.6148 to 0.3403 (a 44.7% re-

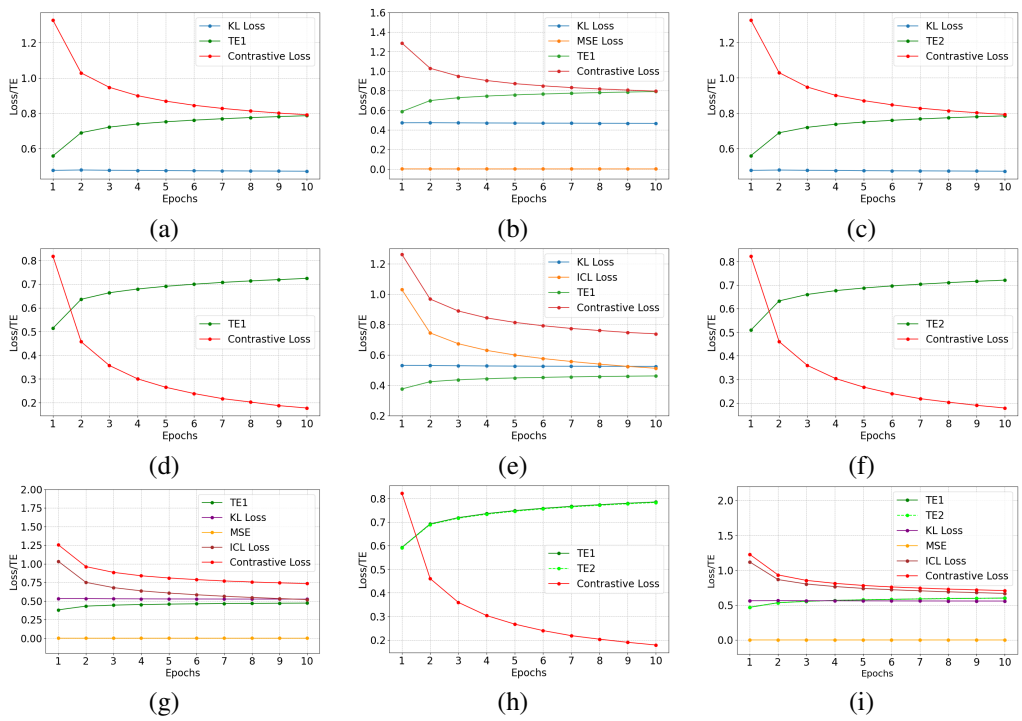


Figure 3: The training losses and TE for different loss functions in the distillation training of RN34-based VLM student with teacher CLIP RN50 using MSCOCO dataset. (a) Contrastive + KL - TE1, (b) Contrastive + KL + MSE - TE2, (c) Contrastive +KL - TE2, (d) Contrastive - TE1, (e) Contrastive + KL + ICL - TE1, (f) Contrastive - TE2, (g) Contrastive + KL +MSE + ICL - TE1, (h) Contrastive - TE1 - TE2, (i) Contrastive + KL + ICL + MSE -TE1 - TE2.

duction), although the MSE term decreases more modestly, from 1.132×10^{-3} to 9.37×10^{-4} (a 17.2% reduction). Thus, the weaker KL trend at $\tau = 0.07$ should not be interpreted as an inherent limitation of the objectives themselves, but rather as a consequence of recipe choice.

These results suggest that the contrastive temperature plays a major role in determining how strongly teacher-structure matching is optimized. A lower temperature sharpens the induced similarity distribution, making mismatches among positive pairs and hard negatives more salient and thereby providing a stronger signal for reducing KL. At the same time, the two objectives still differ in feature-level alignment: TE yields much stronger MSE reduction than MI at $\tau = 0.03$, indicating that TE more effectively preserves direct teacher–student feature correspondence even when both objectives achieve comparable final KL values. We therefore interpret the training dynamics as follows: KL reduction is highly temperature-sensitive, whereas the stronger MSE reduction under TE reflects a more persistent advantage in feature-level teacher alignment. This distinction helps explain why the proposed TE regularizer can produce more teacher-faithful geometry and stronger image-to-text retrieval, even when KL alone does not fully separate the two objectives.

Effect of the batch size. For the batch-size ablation in Table 3, the KL and MSE trajectories remain qualitatively similar across batch sizes. For example, under the TE-inspired setting ($\gamma = 7.5$), KL decreases from 0.6508 to 0.6422 and MSE decreases from 0.000632 to 0.000338 at $|B| = 64$, while at $|B| = 128$ KL decreases from 0.7556 to 0.7484 and MSE decreases from 0.000699 to 0.000404. Under the MI-based setting ($\eta = 5$), KL decreases from 0.6597 to 0.6047 and MSE decreases from 0.001004 to 0.000904 at $|B| = 64$, while at $|B| = 128$ KL decreases from 0.4657 to 0.4351 and MSE decreases from 0.000901 to 0.000690. Thus, changing the batch size does not substantially alter the overall KL/MSE behavior, suggesting that the main TE retrieval gains are not explained by a special batch-size-dependent training dynamic.

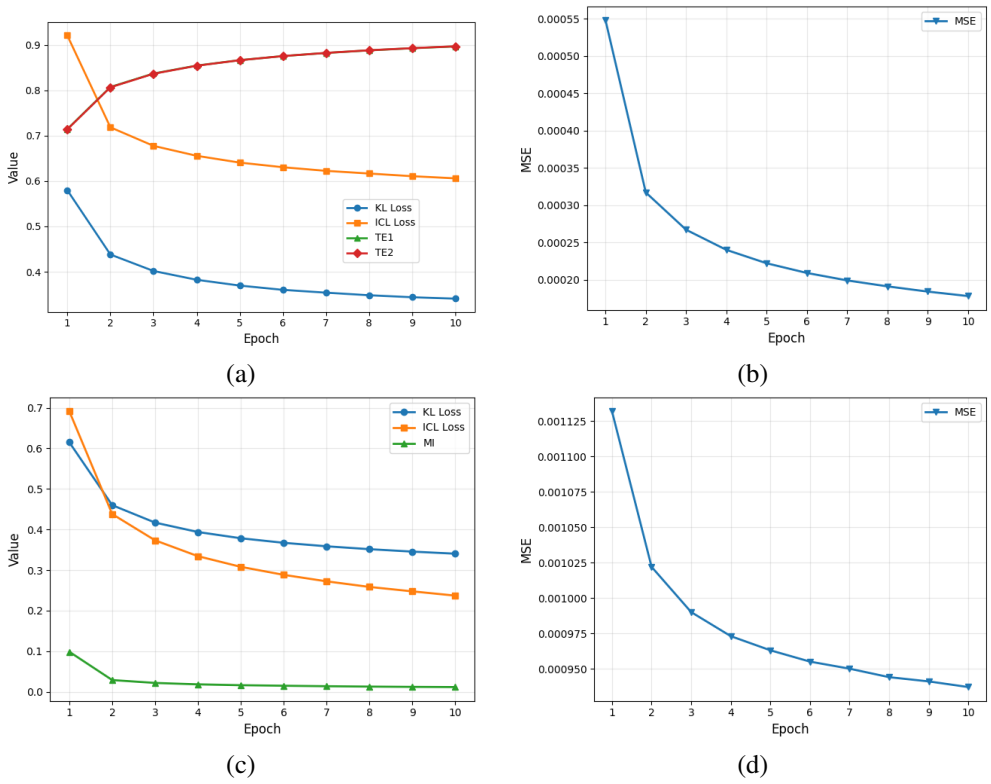


Figure 4: Training dynamics at $\tau = 0.03$ for TE-based ($\alpha = 1, \beta = 50, \delta = 1, \gamma = 7.5$) and MI-based ($\alpha = 1, \beta = 50, \delta = 1, \eta = 5$) distillation. (a) KL, ICL, TE1, and TE2 in TE-based distillation. (b) MSE in TE-based distillation. (c) KL, ICL, and MI in MI-based distillation. (d) MSE in MI-based distillation. At the lower temperature, both objectives exhibit substantial KL reduction, while TE-based yields markedly stronger MSE reduction than MI-based.

G.3 Representation-Level Diagnostics

This section provides the visualizations and per-temperature retrieval results behind the agreement metrics summarized in Section 5.2 of the main text. It studies how the implemented TE-inspired proxy regularizer changes the learned embedding geometry, rather than attempting to directly validate the ideal optimization-step TE quantity in Eq. (5). We compare how TE- and MI-based distillation shape the learned embedding geometry on MSCOCO after 10 epochs of training. We evaluate teacher (CLIP RN50) and student (RN34) embeddings on the same MSCOCO validation samples and use a teacher-only PCA protocol. For each modality, PCA is fit only on the teacher validation embeddings, and both teacher and student embeddings are projected into this shared 2D basis. Because PCA is fit separately for the TE and MI experiments, absolute coordinates are not directly comparable across figures; accordingly, we treat these visualizations as qualitative evidence and rely primarily on quantitative teacher–student agreement metrics.

G.3.1 Temperature=0.07

Modality-specific teacher–student geometry. Figure 1 in the main text shows teacher–student embedding geometry for image and text modalities under the shared teacher-defined basis. In both modalities, the TE-distilled student appears visually closer to the teacher than the MI-distilled student. Since visual inspection alone is insufficient, we additionally quantify teacher–student agreement using local-neighborhood and per-sample similarity metrics.

kNN neighborhood agreement and cosine similarity. To measure local-geometry preservation, we compute kNN overlap@10 between teacher and student *image* embeddings on MSCOCO val.

Table 6: Retrieval and embedding-alignment metrics on MSCOCO val after 10 epochs of distillation on MSCOCO train (RN50 teacher, RN34 student) for $\tau = 0.07$. Retrieval metrics are in %, higher is better. kNN overlap@10 measures local neighborhood agreement between teacher and student image embeddings; cosine similarities are per-sample teacher–student agreement in each modality.

	Image→Text				Text→Image			
Method	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
TE ($\gamma = 7.5$)	10.28	25.88	35.83	18.66	6.88	18.94	27.07	13.67
MI ($\eta = 5$)	8.74	23.19	33.18	16.71	7.62	20.65	29.16	14.82
	kNN overlap@10 (img)		$\cos(T_{\text{img}}, S_{\text{img}})$		$\cos(T_{\text{txt}}, S_{\text{txt}})$			
Method	mean	std	mean	std	mean	std		
TE ($\gamma = 7.5$)	0.605	0.202	0.726	0.078	0.840	0.044		
MI ($\eta = 5$)	0.546	0.200	0.487	0.054	0.528	0.048		

For each sample i, let $\mathcal{N}_{10}^{T}(i)$ and $\mathcal{N}_{10}^{S}(i)$ denote its 10 nearest neighbors under the teacher and student embeddings, respectively. We report

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\left|\mathcal{N}_{10}^{T}(i)\cap\mathcal{N}_{10}^{S}(i)\right|}{10}, \tag{43}$$

along with the standard deviation across samples. We also report per-sample cosine similarity between matched teacher and student embeddings for both modalities. As summarized in Table 6, TE achieves higher image-embedding kNN overlap@10 with the teacher (0.605 ± 0.202 vs. 0.546 ± 0.200), indicating stronger preservation of local neighborhood structure. TE also yields substantially higher per-sample cosine agreement for both image embeddings (0.726 ± 0.078 vs. 0.487 ± 0.054) and text embeddings (0.840 ± 0.044 vs. 0.528 ± 0.048). Beyond Recall@K, we also evaluate MRR (Schütze et al., 2008) for both I2T and T2I retrieval in Table 6. Taken together, these results show that TE more faithfully preserves both local geometry and per-sample teacher–student correspondence than MI.

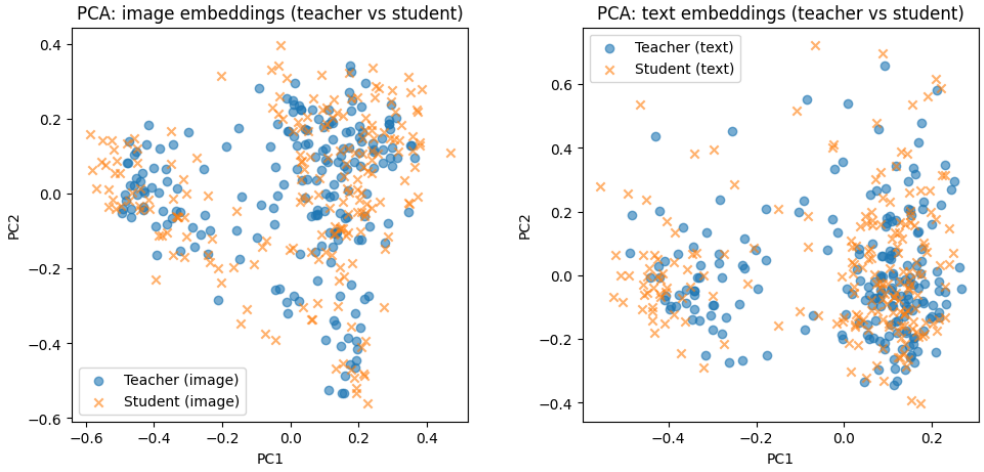
Joint image–text structure. We next visualize the joint image+text embedding space by pooling image and text embeddings and projecting them to 2D with teacher-only PCA fit separately per panel; circles denote image embeddings and crosses denote text embeddings. Figure 2 in the main text shows that the teacher space exhibits a clear bimodal organization, with image and text embeddings occupying distinct regions. The TE-distilled student largely preserves this global structure, maintaining a teacher-like separation between modalities. In contrast, the MI-distilled student shows heavier cross-modal mixing and weaker separation, indicating a greater departure from the teacher's joint embedding geometry. As with the modality-specific PCA plots, we interpret relative separation and overlap within each panel rather than comparing absolute coordinates across panels.

Connection to retrieval. The geometric differences above are also reflected in retrieval performance. Table 6 shows that TE improves Image→Text retrieval relative to MI (R@1 10.28% vs. 8.74%, R@5 25.88% vs. 23.19%, R@10 35.83% vs. 33.18%, MRR 18.66% vs. 16.71%). MI remains slightly stronger on Text→Image (R@1 7.62% vs. 6.88%, R@5 20.65% vs. 18.94%, R@10 29.16% vs. 27.07%, MRR 14.82% vs. 13.67%), indicating that improved teacher-faithfulness does not translate uniformly into gains in both retrieval directions. Overall, however, these diagnostics suggest that the proposed TE regularizer changes not only scalar training dynamics, but also the learned embedding geometry, making the student representation more teacher-faithful in both local and global structure.

G.3.2 TEMPERATURE=0.03

Modality-specific teacher–student geometry. Figures 5 and 6 show teacher–student embedding geometry for image and text modalities under the shared teacher-defined basis at the lower temperature $\tau = 0.03$. As in the $\tau = 0.07$ case, the TE-distilled student appears visually closer to

the teacher than the MI-distilled student in both modalities. For image embeddings, the TE student better preserves the teacher's broad cluster arrangement, including the denser central-right region and the more diffuse left-side structure, whereas the MI student exhibits a more noticeable shift and compression relative to the teacher. For text embeddings, the difference is even clearer: the TE student remains largely co-located with the teacher's dominant clusters, while the MI student shows a systematic upward shift and stronger separation from the teacher cloud, suggesting weaker preservation of the teacher-defined text geometry. Although these PCA plots are qualitative, they are consistent with the view that TE better maintains teacher-faithful embedding structure under the lower-temperature regime.



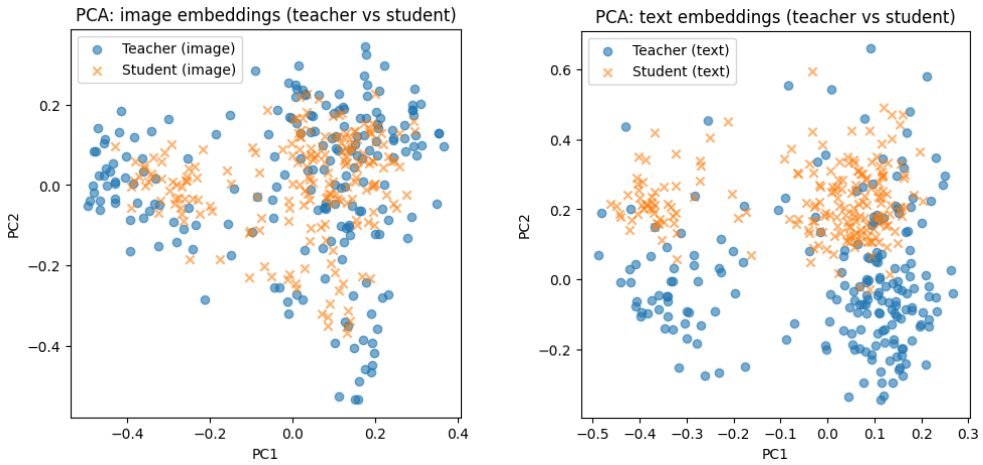
(a) (b)

Figure 5: For $\tau = 0.03$, teacher-only PCA visualization of teacher–student embedding geometry on MSCOCO val after 10 epochs of distillation on MSCOCO train (TE-based; $\gamma = 7.5$, with $\alpha = 1$, $\beta = 50$, $\delta = 1$, $\eta = 0$, $\lambda = 0$). For each modality, PCA is fit on the teacher validation embeddings and both teacher and student embeddings are projected using this shared 2D basis. (a) Image embeddings. (b) Text embeddings.

kNN neighborhood agreement and cosine similarity. We next quantify teacher–student agreement using the same local-neighborhood and per-sample similarity diagnostics as above. Table 7 shows that TE again achieves stronger alignment with the teacher than MI across all geometry-based metrics. In particular, TE yields higher image-embedding kNN overlap@10 (0.487 ± 0.174 vs. 0.429 ± 0.173), indicating better preservation of local neighborhood structure. The gap is even larger for per-sample cosine agreement: for image embeddings, TE reaches 0.809 ± 0.058 versus 0.469 ± 0.042 for MI, and for text embeddings, TE achieves 0.882 ± 0.048 versus 0.522 ± 0.048. These improvements are substantial and suggest that, at $\tau = 0.03$, the TE-distilled student remains much more closely aligned with the teacher in both modalities, especially at the level of pointwise correspondence.

Joint image–text structure. Figure 7 further compares the joint image+text embedding space by pooling both modalities and projecting them to 2D under teacher-only PCA. The teacher space again exhibits a clear bimodal organization, with image and text embeddings occupying distinct regions. The TE-distilled student preserves this overall geometry more faithfully: the image and text clouds remain clearly separated, and the student layout broadly mirrors the teacher's two-cluster structure. In contrast, the MI-distilled student shows a less teacher-like organization, with more distortion in the placement of the text cluster relative to the image cluster and weaker preservation of the teacher's global arrangement. Thus, beyond modality-specific alignment, TE also appears to better preserve the teacher's cross-modal global geometry at $\tau = 0.03$.

Connection to retrieval. These geometric differences are reflected in retrieval performance. Table 7 shows that TE outperforms MI in both retrieval directions at $\tau = 0.03$. For Image→Text,



(a) (b)

Figure 6: For $\tau = 0.03$, teacher-only PCA visualization of teacher–student embedding geometry on MSCOCO val after 10 epochs of distillation on MSCOCO train (MI-based; $\eta = 5$, with $\alpha = 1$, $\beta = 50$, $\delta = 1$, $\lambda = 0$, $\gamma = 0$). For each modality, PCA is fit on the teacher validation embeddings and both teacher and student embeddings are projected using this shared 2D basis. (a) Image embeddings. (b) Text embeddings.

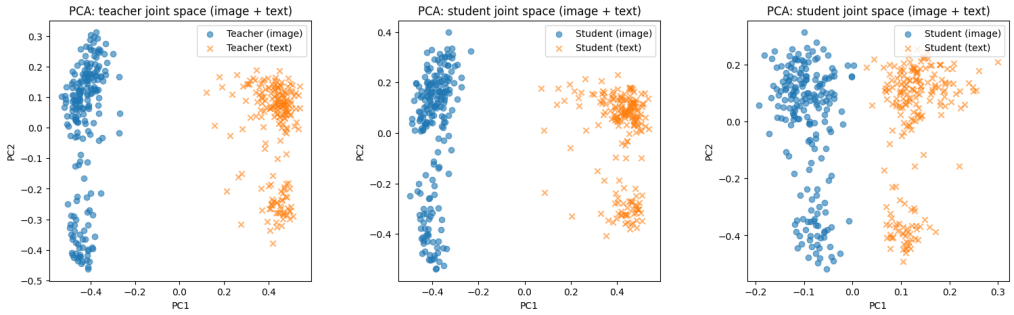
Table 7: For $\tau = 0.03$, retrieval and embedding-alignment metrics on MSCOCO val after 10 epochs of distillation on MSCOCO train (RN50 teacher, RN34 student). Retrieval metrics are in %, higher is better. kNN overlap@10 measures local neighborhood agreement between teacher and student image embeddings; cosine similarities are per-sample teacher–student agreement in each modality.

	Image→Text				Text→Image			
Method	R@1	R@5	R@10	MRR	R@1	R@5	R@10	MRR
TE ($\gamma = 7.5$)	10.85	26.47	36.29	19.21	6.84	18.76	26.94	13.59
MI ($\eta = 5$)	9.14	23.67	33.14	17.09	6.49	18.23	26.44	13.18
	kNN overlap@10 (img)		$\cos(T_{\text{img}}, S_{\text{img}})$		$\cos(T_{\text{txt}}, S_{\text{txt}})$			
Method	mean	std	mean	std	mean	std		
TE ($\gamma = 7.5$)	0.487	0.174	0.809	0.058	0.882	0.048		
MI ($\eta = 5$)	0.429	0.173	0.469	0.042	0.522	0.048		

TE improves over MI from $9.14/23.67/33.14$ with MRR 17.09 to $10.85/26.47/36.29$ with MRR 19.21. For Text→Image, TE also remains slightly better, improving from $6.49/18.23/26.44$ with MRR 13.18 to $6.84/18.76/26.94$ with MRR 13.59. Compared with the $\tau = 0.07$ setting, where MI was slightly stronger on T2I retrieval, the lower-temperature regime yields a cleaner advantage for TE across both directions. Taken together, the PCA visualizations, neighborhood-overlap statistics, cosine similarities, and retrieval metrics all support the same conclusion: at $\tau = 0.03$, TE produces a more teacher-faithful student representation than MI, both locally and globally, and this stronger geometric fidelity is accompanied by more consistent retrieval gains.

G.4 Comparison with Relational Geometry Matching

Relational structure is an important aspect of knowledge distillation. In addition to matching individual teacher and student embeddings, a student model can also be encouraged to preserve the teacher's pairwise representation geometry within a mini-batch. This appendix studies a relational Gram-matching baseline that aligns the within-batch cosine similarity matrices of the teacher and



(a) (b) (c)

Figure 7: For $\tau = 0.03$, joint (image+text) embedding visualization on MSCOCO val after 10 epochs of distillation on MSCOCO train. We pool image (circles) and text (crosses) embeddings and project them to 2D with teacher-only PCA fit separately per panel. (a) Teacher CLIP RN50. (b) TE-distilled RN34 student ($\gamma = 7.5$; $\alpha = 1$, $\beta = 50$, $\delta = 1$, $\eta = 0$, $\lambda = 0$). (c) MI-distilled RN34 student ($\eta = 5$; $\alpha = 1$, $\beta = 50$, $\delta = 1$, $\gamma = 0$, $\lambda = 0$).

student. This baseline represents geometry-matching and relational-distillation objectives because it transfers sample-sample relationships rather than only pointwise feature targets.

Relational Gram-Matching Objective. Let $Z_S \in \mathbb{R}^{B\times d}$ and $Z_T \in \mathbb{R}^{B\times d}$ denote the student and teacher features for a mini-batch of size B, after row-wise ℓ_2 normalization. The student Gram similarity matrix is defined as

$$G_S = Z_S Z_S^\top. \tag{44}$$

The teacher Gram similarity matrix is defined as

$$G_T = Z_T Z_T^\top. \tag{45}$$

For normalized features, the entries of G_S and G_T correspond to within-batch cosine similarities. The diagonal entries are close to one and therefore do not provide meaningful relational information. We exclude the diagonal and define the Gram-matching loss as

$$\mathcal{L}_{\text{Gram}}(Z_S, Z_T) = \frac{1}{B(B-1)} \sum_{i=1}^{B} \sum_{\substack{j=1 \\ j\neq i}}^{B} (G_{S,ij} - G_{T,ij})^2. \tag{46}$$

This objective is closely related to similarity-preserving knowledge distillation (Tung & Mori, 2019) and relational knowledge distillation (Park et al., 2019), which transfer pairwise or higher-order relations among examples rather than only pointwise feature targets. More broadly, Gram-style feature relationships have also been used in knowledge distillation through flow-of-solution-procedure matrices (Yim et al., 2017).

In the vision-language setting, we apply this objective to image embeddings, text embeddings, and joint image-text embeddings. Let v_S and u_S denote the student image and text embeddings, respectively. Let v_T and u_T denote the corresponding teacher image and text embeddings. The image-level relational loss is

$$\mathcal{L}_{\text{Gram}}^{\text{img}} = \mathcal{L}_{\text{Gram}}(v_S, v_T). \tag{47}$$

The text-level relational loss is

$$\mathcal{L}_{\text{Gram}}^{\text{txt}} = \mathcal{L}_{\text{Gram}}(u_S, u_T). \tag{48}$$

For joint multimodal geometry matching, we concatenate the image and text embeddings for each sample and then normalize the resulting joint representation. For the i-th student sample, the unnormalized joint representation is

$$\widetilde{z}_{S,i}^{\text{joint}} = \text{concat}(v_{S,i}, u_{S,i}). \tag{49}$$

The normalized student joint representation is

$$z_{S,i}^{\text{joint}} = \frac{\widetilde{z}_{S,i}^{\text{joint}}}{\left\| \widetilde{z}_{S,i}^{\text{joint}} \right\|_2 + \epsilon}. \tag{50}$$

Similarly, for the teacher, the unnormalized joint representation is

$$\widetilde{z}_{T,i}^{\text{joint}} = \text{concat}(v_{T,i}, u_{T,i}). \tag{51}$$

The normalized teacher joint representation is

$$z_{T,i}^{\text{joint}} = \frac{\widetilde{z}_{T,i}^{\text{joint}}}{\left\| \widetilde{z}_{T,i}^{\text{joint}} \right\|_2 + \epsilon}. \tag{52}$$

Let Z_S^{joint} and Z_T^{joint} denote the matrices obtained by stacking $z_{S,i}^{\text{joint}}$ and $z_{T,i}^{\text{joint}}$ over the mini-batch. The joint Gram loss is

$$\mathcal{L}_{\text{Gram}}^{\text{joint}} = \mathcal{L}_{\text{Gram}} \left(Z_S^{\text{joint}}, Z_T^{\text{joint}} \right). \tag{53}$$

The final relational geometry-matching loss is

$$\mathcal{L}_{\text{Gram}}^{\text{total}} = \frac{1}{3} \left(\mathcal{L}_{\text{Gram}}^{\text{img}} + \mathcal{L}_{\text{Gram}}^{\text{txt}} + \mathcal{L}_{\text{Gram}}^{\text{joint}} \right). \tag{54}$$

Let the base distillation objective be

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{contrastive}} + \alpha \mathcal{L}_{\text{KL}} + \beta \mathcal{L}_{\text{MSE}} + \delta \mathcal{L}_{\text{ICL}}. \tag{55}$$

The full objective for the Gram-matching baseline is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \mu \mathcal{L}_{\text{Gram}}^{\text{total}}. \tag{56}$$

Here, μ controls the strength of relational geometry matching.

Relation to TE-Inspired Finite-Difference Geometry. The Gram-matching objective transfers static pairwise similarity structure from the teacher to the student. For normalized embeddings, matching the Gram matrix corresponds to matching the within-batch angular geometry among samples. Thus, this baseline captures an important class of relational and representation-geometry distillation methods.

The proposed TE-inspired objectives provide a different geometric signal. Rather than directly matching static pairwise similarities, TE_1 and TE_2 match the one-dimensional subspaces spanned by teacher and student finite differences. TE_1 applies this orientation-invariant criterion per modality, while TE_2 applies it jointly after concatenating image and text finite differences. The full TE objective is

$$\text{TE} = \text{TE}_1 + \text{TE}_2. \tag{57}$$

The corresponding training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} - \gamma \left(\text{TE}_1 + \text{TE}_2 \right). \tag{58}$$

Thus, Gram matching asks whether the student preserves the teacher's within-batch similarity matrix, whereas the TE-inspired objective asks whether teacher and student local representation changes span the same secant lines, up to orientation. These two forms of supervision are related but not equivalent. For sufficiently local pairs, the squared-cosine criterion can be interpreted as orientation-invariant Jacobian-action subspace matching, since finite differences approximate local representation changes induced by input perturbations.

Results. Table 8 compares the base objective, the relational Gram-matching baseline, and the TE-inspired variants. All experiments use $\alpha = 1$, $\beta = 50$, and $\delta = 1$. For Gram-matching experiments, we set $\gamma = 0$ and sweep μ. For TE experiments, we set $\mu = 0$ and use $\gamma = 7.5$.

The results show that relational Gram matching provides modest improvements over the base objective. The best Gram setting improves Image-to-Text R@1 from 6.08% to 7.01%, R@5 from 18.14%

Method	Image-to-Text Retrieval			Text-to-Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Base: CL+KL+MSE+ICL	6.08	18.14	26.93	5.92	17.06	24.89
+ Gram, $\mu = 1$	6.26	18.17	26.72	5.77	17.13	25.07
+ Gram, $\mu = 5$	6.45	18.42	27.10	5.70	16.93	24.84
+ Gram, $\mu = 50$	7.01	19.35	28.03	5.92	17.13	24.94
+ Gram, $\mu = 100$	6.91	19.50	28.16	6.05	17.37	25.30
+ TE_1 only	10.10	25.97	35.76	**7.33**	**20.22**	**28.59**
+ TE_2 only	**10.32**	**26.44**	**36.43**	**7.33**	20.03	28.46
+ $\mathrm{TE}_1 + \mathrm{TE}_2$	10.27	26.36	36.30	6.98	18.95	26.95

Table 8: Comparison with relational Gram geometry matching. The Gram baseline aligns within-batch teacher-student cosine similarity matrices for image, text, and joint image-text embeddings. Although Gram matching improves over the base CL+KL+MSE+ICL objective, the TE-inspired finite-difference geometry objectives achieve stronger retrieval performance, especially for Image-to-Text retrieval.

to 19.50%, and R@10 from 26.93% to 28.16%. However, the TE-inspired objectives achieve substantially stronger performance. The full TE objective obtains 10.27%, 26.36%, and 36.30% on Image-to-Text R@1/R@5/R@10, respectively. TE_1 and TE_2 individually also outperform Gram matching, with TE_2 achieving the best Image-to-Text retrieval performance. A condensed version of this comparison appears in Table 2 of the main text.

For Text-to-Image retrieval, the best Gram setting achieves 6.05%, 17.37%, and 25.30% on R@1/R@5/R@10. The full TE objective improves these results to 6.98%, 18.95%, and 26.95%, while the TE_1-only objective achieves the strongest Text-to-Image performance at 7.33%, 20.22%, and 28.59%. These results indicate that preserving static pairwise similarity geometry is helpful but does not fully account for the gains obtained by TE-inspired finite-difference matching. The orientation-invariant secant geometry captured by TE_1 and TE_2 provides a complementary and more effective teacher-student supervision signal for multimodal distillation.

G.5 Isolating the Effect of the TE-Inspired Proxy

The full distillation objective combines several complementary losses. To isolate the contribution of the TE-inspired proxy, we conduct a controlled ablation in which the base objective $\mathcal{L}_{\mathrm{base}}$ of Eq. (55) and its hyperparameters are fixed at $\alpha = 1$, $\beta = 50$, $\delta = 1$, and only the TE-related component is changed: the baseline uses $\gamma = 0$, while the TE variants add $-\gamma\,\mathrm{TE}_1$, $-\gamma\,\mathrm{TE}_2$, or $-\gamma(\mathrm{TE}_1 + \mathrm{TE}_2)$ to the same base objective with $\gamma = 7.5$. We additionally report modality-restricted variants in which the proxy is computed on only the image-side or only the text-side finite differences. The comparison therefore isolates the effect of the TE-inspired proxy while keeping the remaining loss terms and hyperparameters unchanged; Table 2 in the main text places these rows alongside the MSE-Δ, Gram-matching, and MI regularizers under the same base objective.

Table 9 shows that adding the TE-inspired proxy substantially improves retrieval performance over the fixed base objective. The base CL+KL+MSE+ICL objective obtains 6.08%, 18.14%, and 26.93% on Image-to-Text R@1/R@5/R@10, respectively. Adding TE_1 improves these results to 10.10%, 25.97%, and 35.76%, while adding TE_2 improves them to 10.32%, 26.44%, and 36.43%. The full $\mathrm{TE}_1 + \mathrm{TE}_2$ objective achieves 10.27%, 26.36%, and 36.30%.

For Text-to-Image retrieval, the base objective obtains 5.92%, 17.06%, and 24.89% on R@1/R@5/R@10. Adding TE_1 improves these results to 7.33%, 20.22%, and 28.59%, and adding TE_2 improves them to 7.33%, 20.03%, and 28.46%. The full $\mathrm{TE}_1 + \mathrm{TE}_2$ objective obtains 6.98%, 18.95%, and 26.95%.

The modality-specific ablations further show that the image-side and text-side components of TE_1 contribute differently. The image-side TE term improves Image-to-Text retrieval, while the text-side TE term provides stronger gains for Text-to-Image retrieval. These trends are consistent with the interpretation that the TE-inspired proxy transfers orientation-invariant finite-difference geometry in representation space, and that image and text modalities may affect the two retrieval directions

Method	Image-to-Text Retrieval			Text-to-Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Base: CL+KL+MSE+ICL	6.08	18.14	26.93	5.92	17.06	24.89
+ TE_1 only	10.10	25.97	35.76	**7.33**	**20.22**	**28.59**
+ TE_2 only	**10.32**	**26.44**	**36.43**	**7.33**	20.03	28.46
+ $\mathrm{TE}_1 + \mathrm{TE}_2$	10.27	26.36	36.30	6.98	18.95	26.95
+ image-side TE_1 only	7.46	20.92	30.36	5.69	16.67	24.38
+ text-side TE_1 only	7.49	20.27	28.96	7.03	19.40	27.60

Table 9: Ablation isolating the effect of the TE-inspired proxy. The base objective and its hyperparameters are fixed across all rows, with $\alpha = 1$, $\beta = 50$, and $\delta = 1$. The only change is whether the TE-inspired proxy is disabled, added as TE_1 only, added as TE_2 only, or added as the full $\mathrm{TE}_1 + \mathrm{TE}_2$ objective. The results show that the TE-inspired proxy provides substantial gains over the same base objective without changing the other loss weights.

Table 10: Ablation of finite-difference pairing strategies. Results are mean $\pm$ standard deviation over five random seeds. Random adjacent-pairing is the default strategy used in the main experiments. NN-pairing forms finite differences using nearest neighbors in the teacher embedding space.

Pairing Strategy	Image-to-Text Retrieval			Text-to-Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Random adjacent-pairing	10.08 ± 0.17	26.10 ± 0.19	36.22 ± 0.21	7.12 ± 0.19	19.35 ± 0.27	27.85 ± 0.23
NN-pairing	10.35 ± 0.16	26.26 ± 0.22	36.41 ± 0.27	6.98 ± 0.30	19.16 ± 0.26	27.52 ± 0.39

differently. Overall, this ablation isolates the TE-inspired proxy from the other terms in the objective and shows that it is a major contributor to the observed retrieval gains.

G.6 Pairing Strategy Ablation: Random Adjacent Pairs vs. Nearest-Neighbor Pairs

The proposed TE-inspired proxy uses finite differences between paired examples in a mini-batch. In the main implementation, we form these pairs by first randomly permuting the mini-batch and then taking adjacent examples in the permuted order. Random adjacent pairs need not satisfy the small-perturbation motivation of Section 4.2 and may be disconnected from the local neighborhood structure analyzed by our kNN-based representation diagnostics. We therefore compare the default random adjacent-pairing strategy with a nearest-neighbor (NN) pairing strategy.

For random adjacent-pairing, finite differences are formed as

$$\Delta f_i = f(x_{\pi(i+1)}) - f(x_{\pi(i)}), \tag{59}$$

where π is a random permutation of the mini-batch. For NN-pairing, each example is paired with its nearest neighbor in the teacher embedding space:

$$j(i) = \arg\max_{j \neq i} \cos\big(z_T(x_i), z_T(x_j)\big), \tag{60}$$

and the finite difference is formed as

$$\Delta f_i = f(x_{j(i)}) - f(x_i). \tag{61}$$

This NN-pairing variant provides a more local finite-difference construction and is more directly aligned with the small-perturbation motivation of Section 4.2 and the kNN-based representation analysis.

Table 10 reports five-seed results for the two pairing strategies under the same training objective and hyperparameter setting.

The two pairing strategies yield very similar performance. NN-pairing slightly improves Image-to-Text retrieval, while random adjacent-pairing slightly improves Text-to-Image retrieval. This suggests that the proposed TE-inspired proxy is not overly sensitive to the particular finite-difference

pairing construction. We therefore interpret random adjacent-pairing as a lightweight stochastic finite-difference regularizer, rather than as a strict estimator of infinitesimal perturbations. NN-pairing provides a more local variant that better matches the small-perturbation intuition, but the comparable results indicate that the main benefit comes from teacher–student secant-line alignment in representation space rather than from a specific pairing heuristic.

G.7 TEACHER: VIT-B/16, STUDENT: RN34

We distill an RN34-based VLM student from a CLIP ViT-B/16 teacher with $\alpha = 1.0$, $\beta = 100$, $\delta = 1.0$, $\gamma = 5.0$, and $\tau = 0.07$, chosen with the heuristic of Appendix G.1. On MSCOCO, given the scale of the model and dataset, the student is trained for 6 epochs; each row of Table 11 required approximately 10 hours on a Google Colab T4 GPU (High-RAM). On Flickr8k, we used a Google Colab A100 GPU and trained for 10 epochs; each row of Table 13 took around 30 minutes.

Tables 11 and 13 show that adding the TE-inspired proxies delivers the single largest boost among all tested loss terms, leading to 3–4 percentage-point (pp) I2T R@1 gains on MSCOCO and 8–12 pp gains on the low-resource Flickr8k benchmark over the CL-only student.

Table 11: Comparison of retrieval performance (Recall@k) of an RN34-based VLM student distilled from CLIP ViT-B/16 on MSCOCO in VLM distillation with different loss functions. All Loss Function: CL + KL + MSE + ICL - TE1 - TE2. The hyperparameter settings: $\alpha = 1.0$, $\beta = 100$, $\delta = 1.0$, $\gamma = 5.0$.

Model and Loss Function	**I2T Retrieval (R)**			**T2I Retrieval (R)**		
	R@1	R@5	R@10	R@1	R@5	R@10
Teacher Model (ViT-B/16)	29.60%	50.95%	60.35%	15.41%	31.54%	39.96%
Student Models (RN34)						
CL Only (Oord et al., 2018)	4.66%	14.10%	21.28%	3.78%	11.95%	18.40%
CL + MSE (Yang et al., 2024)	4.55%	14.27%	21.36%	3.79%	11.99%	18.44%
CL + KL (Li et al., 2024a)	4.70%	14.46%	22.21%	4.58%	14.15%	21.32%
CL + ICL (Yang et al., 2024)	5.52%	16.25%	23.92%	4.55%	13.86%	20.79%
CL - TE1	7.24%	19.88%	28.55%	5.68%	16.22%	23.71%
CL - TE2	7.02%	20.26%	29.46%	5.83%	16.54%	24.27%
CL - TE1 - TE2	7.44%	20.24%	29.01%	5.78%	16.35%	23.90%
CL + KL + MSE - TE1	7.74%	21.05%	30.14%	5.86%	16.85%	24.64%
CL + KL + ICL - TE1	7.64%	21.07%	30.16%	**6.20%**	**17.64%**	**25.66%**
CL + KL + MSE + ICL - TE1	7.90%	21.49%	30.42%	6.15%	17.48%	25.35%
ALL Loss Function	**8.48%**	**22.02%**	**30.99%**	6.00%	17.01%	24.80%

Table 12 summarizes the hyperparameter sensitivity of retrieval when distilling a student RN34 from a ViT-B/16 teacher on MSCOCO. Overall, introducing the TE reward is consistently beneficial, especially for T2I retrieval: compared to the baseline without TE ($\gamma = 0$), a small TE weight ($\gamma = 1$) yields the strongest T2I scores in the table (R@1/5/10 of $6.75/18.74/26.91$), indicating that TE-based regularization can effectively strengthen bidirectional alignment under this teacher-student pairing. Increasing γ beyond this regime provides diminishing returns and can reduce performance (e.g., $\gamma \geq 5$), suggesting that overly strong TE weighting may over-regularize the student. We also observe that increasing the MSE weight β can improve I2T retrieval: the best I2T performance is achieved at $\beta = 100$ with $\gamma = 5$ (R@1/5/10 of $8.48/22.02/30.99$), highlighting the importance of feature-level matching for transferring visual semantics from ViT-B/16 to RN34. In contrast, varying the MI weight η in isolation ($\gamma = 0$) produces moderate gains over the baseline but does not surpass the best TE-regularized setting for T2I, and combining MI, MSE-Δ ($\lambda > 0$), and TE does not yield additive improvements in this sweep. Taken together, these results indicate that (i) modest TE regularization is most helpful for T2I, while (ii) stronger feature distillation (larger β) primarily benefits I2T, and careful balancing is required to avoid over-regularization when multiple auxiliary terms are enabled.

Flickr8k has a small number of training/evaluation samples which can amplify variance and make performance more sensitive to hyperparameter choices. Motivated by this, we additionally evaluate on Flickr30k (Young et al., 2014), a larger and more diverse captioned dataset, to provide a more

Table 12: Comparison of retrieval performance (Recall@k) of an RN34-based VLM student distilled from CLIP ViT-B/16 on MSCOCO.

α	β	δ	η	λ	γ	I2T R@1	I2T R@5	I2T R@10	T2I R@1	T2I R@5	T2I R@10
1	50	1	0	0	0	5.81%	17.30%	25.89%	5.60%	16.64%	24.46%
1	50	1	0	0	1	7.32%	20.41%	29.58%	**6.75%**	**18.74%**	**26.91%**
1	50	1	0	0	2.5	7.68%	21.16%	30.20%	6.70%	**18.74%**	26.78%
1	50	1	0	0	5	7.87%	21.47%	30.74%	5.98%	17.21%	24.96%
1	100	1	0	0	5	**8.48%**	**22.02%**	**30.99%**	6.00%	17.01%	24.80%
1	50	1	0	0	7.5	7.90%	21.27%	30.22%	5.92%	16.69%	24.42%
1	50	1	0	0	10	7.49%	20.56%	29.51%	5.47%	16.01%	23.38%
1	50	1	1	0	0	7.08%	19.72%	28.44%	6.40%	17.94%	25.86%
1	50	1	2.5	0	0	7.09%	20.00%	29.04%	6.31%	17.97%	25.96%
1	50	1	5	0	0	7.28%	20.05%	29.24%	6.08%	17.37%	25.30%
1	100	1	5	0	0	7.04%	20.17%	29.08%	6.21%	17.43%	25.37%
1	50	1	7.5	0	0	6.69%	19.53%	28.25%	6.01%	16.96%	24.70%
1	50	1	2.5	0	5	7.29%	20.45%	29.37%	5.68%	16.56%	24.13%
1	50	1	2.5	50	2.5	7.33%	20.31%	29.40%	6.00%	17.12%	24.91%
1	50	1	1	0	7.5	7.43%	20.43%	29.56%	5.75%	16.48%	24.00%
1	50	1	1	50	7.5	7.42%	20.51%	29.25%	5.65%	16.33%	24.02%

Table 13: Retrieval performance (Recall@k) on Flickr8k. The RN34-based VLM student is distilled from the teacher model (CLIP ViT-B/16). All Loss Function: CL + MSE + KL + ICL - TE1 - TE2. The weighting factors for loss functions: $\alpha = 1.0$, $\beta = 100$, $\delta = 1.0$, $\gamma = 5.0$.

Model and Loss Function	**I2T Retrieval (R)**			**T2I Retrieval (R)**		
	R@1	R@5	R@10	R@1	R@5	R@10
Teacher Model (ViT-B/16)	57.41%	82.70%	90.61%	55.02%	81.63%	87.64%
Student Models (RN34)						
CL (Oord et al., 2018)	21.09%	46.95%	59.47%	17.53%	42.59%	55.26%
CL + MSE (Yang et al., 2024)	21.17%	46.46%	58.98%	16.26%	42.83%	55.37%
CL + KL (Li et al., 2024a)	24.38%	51.32%	63.84%	19.59%	46.97%	60.03%
CL + ICL (Yang et al., 2024)	26.44%	52.14%	65.32%	20.44%	47.69%	61.61%
CL - TE1	28.91%	57.17%	69.19%	22.98%	50.69%	63.79%
CL - TE2	30.07%	57.41%	68.45%	23.67%	52.04%	65.44%
CL - TE1 - TE2	28.42%	58.90%	70.02%	22.59%	51.10%	64.79%
CL + KL + MSE - TE1	29.90%	59.97%	71.83%	22.34%	51.73%	65.44%
CL + KL + ICL - TE1	28.42%	56.84%	71.17%	23.05%	51.53%	65.09%
CL + KL + MSE + ICL - TE1	29.82%	60.71%	71.75%	23.53%	51.47%	64.99%
All Loss Function	**33.28%**	**64.33%**	**73.97%**	**26.36%**	**56.18%**	**69.64%**

robust assessment of our distillation objective and to verify that the observed benefits of TE-based regularization persist under a less data-limited setting. Table 14 summarizes the retrieval results on Flickr30k when distilling an RN34 student from a ViT-B/16 teacher under different loss weightings. Several consistent trends emerge. First, introducing the TE reward ($\gamma > 0$) generally improves retrieval over the baseline without TE, particularly for moderate TE weights. Increasing γ from 0 to 2.5–5 yields clear gains in both I2T and T2I Recall@K, with I2T R@1 improving from 34.71% to over 40%, and T2I R@1 rising from 29.35% to above 31%. However, very large TE weights (e.g., $\gamma = 7.5$) begin to slightly degrade performance, indicating that excessive regularization may oversuppress the contrastive signal. Second, a stronger KL term ($\alpha = 5$) in combination with moderate TE ($\gamma = 5$) leads to the strongest overall I2T performance (R@1 $= 42.31\%$, R@10 $= 80.37\%$), suggesting complementary effects between KL alignment and TE-based guidance. In contrast, increasing δ or λ yields smaller and less consistent gains, whereas introducing MI ($\eta > 0$) can modestly improve T2I performance. Overall, these results reinforce the pattern observed on Flickr8k: moderate TE regularization enhances the semantic structure preserved in the distilled model, while careful balancing with KL produces the most robust improvements.

G.8 Teacher: RN50, Student: RN34, Dataset: Flickr8k

We further evaluate our approach on the Flickr8k dataset (Marco et al., 2023) (85%/15% train/test split). Performance results for various loss functions are summarized in Table 15. The loss function

Table 14: Comparison of retrieval performance (Recall@k) of an RN34-based VLM student distilled from CLIP ViT-B/16 on **Flickr30k**.

α	β	δ	η	λ	γ	I2T R@1	I2T R@5	I2T R@10	T2I R@1	T2I R@5	T2I R@10
1	50	1	0	0	0	34.71%	62.43%	74.85%	29.35%	57.91%	70.02%
1	50	1	0	0	1	38.36%	66.67%	76.43%	31.40%	60.95%	71.93%
1	50	1	0	0	2.5	40.34%	67.16%	77.51%	31.79%	60.89%	72.01%
1	50	1	0	0	5	40.24%	68.54%	78.80%	31.01%	60.37%	72.19%
1	50	1	0	0	7.5	37.28%	66.86%	78.11%	30.02%	58.93%	71.28%
1	50	1	0	0	10	39.35%	66.67%	76.73%	30.20%	58.64%	70.43%
5	50	1	0	0	5	**42.31%**	**69.92%**	**80.37%**	31.24%	61.08%	72.19%
1	100	1	0	0	5	37.87%	67.36%	78.90%	31.12%	61.24%	72.31%
1	50	5	0	0	5	37.38%	67.26%	77.61%	31.36%	60.37%	71.05%
1	50	1	0	50	0	36.49%	64.20%	75.54%	31.10%	58.78%	70.26%
1	50	1	0	100	0	36.59%	64.99%	76.63%	30.08%	59.90%	71.76%
1	50	1	1	1	0	36.59%	66.77%	78.30%	32.29%	**61.40%**	**73.06%**
1	50	1	2.5	0	0	37.77%	66.47%	77.22%	**32.96%**	60.79%	71.64%
1	50	1	5	0	0	39.25%	65.78%	76.73%	31.52%	60.02%	71.91%
1	50	1	7.5	0	0	39.55%	68.05%	78.40%	31.95%	60.00%	71.48%
5	50	1	7.5	0	0	37.77%	67.36%	77.91%	30.14%	58.21%	69.92%
1	50	1	2.5	50	2.5	40.43%	67.06%	78.11%	32.64%	60.65%	72.21%
1	50	1	2.5	0	5	38.66%	66.77%	78.80%	30.51%	59.47%	71.95%
1	50	1	2.5	50	5	39.94%	68.84%	78.40%	31.32%	59.68%	71.34%
1	50	1	1	0	7.5	36.39%	67.65%	77.91%	30.73%	59.39%	71.48%
1	50	1	1	50	7.5	38.17%	65.78%	77.02%	30.08%	58.09%	70.91%
1	50	1	5	50	7.5	36.79%	65.48%	76.82%	30.41%	58.86%	70.49%
Teacher (ViT-B/16)						81.76%	97.14%	98.92%	62.66%	86.23%	91.36%

employs weighting factors $\alpha = 1.0$, $\beta = 100$, $\delta = 1.0$, $\gamma = 5.0$, and a temperature parameter $\tau = 0.07$. These parameters were selected based on the relative contribution of each loss term to the total loss during training, ensuring balanced optimization. Given the modest size of Flickr8k, all experiments were conducted on a Google Colab instance equipped with an A100 GPU and limited system RAM. Each experiment (i.e., each row in Table 15) required approximately 20 minutes of training time. Notably, incorporating TE1 or TE2 into the loss function consistently improves both image-to-text (I2T) and text-to-image (T2I) retrieval performance compared to baselines that rely solely on standard distillation losses such as CL + KL or CL + MSE. These results underscore the effectiveness of the TE-inspired proxies in guiding student updates during distillation.

Table 15: Retrieval performance (Recall@k) on **Flickr8k** of an RN34-based VLM student distilled from CLIP RN50 under different loss functions. All Loss Function: CL + KL + MSE + ICL - TE1 - TE2. The weighting factors for loss functions: $\alpha = 1.0$, $\beta = 100$, $\delta = 1.0$, $\gamma = 5.0$.

Model and Loss Function	**I2T Retrieval (R)**			**T2I Retrieval (R)**		
	R@1	R@5	R@10	R@1	R@5	R@10
Teacher Model (RN50)	51.65%	78.17%	87.73%	47.28%	75.21%	84.60%
Student Models (RN34)						
CL Only (Oord et al., 2018)	22.73%	48.19%	60.87%	18.47%	43.76%	56.77%
CL + MSE (Yang et al., 2024)	22.98%	49.92%	62.52%	17.84%	44.71%	57.99%
CL + KL (Li et al., 2024a)	27.51%	56.51%	69.19%	23.20%	50.12%	62.82%
CL + ICL (Yang et al., 2024)	24.55%	52.06%	64.50%	19.87%	47.97%	61.24%
CL - TE1	30.48%	62.52%	74.05%	24.42%	54.25%	68.39%
CL - TE2	31.80%	61.37%	72.90%	**25.19%**	54.66%	**68.70%**
CL - TE1 - TE2	32.29%	62.36%	**75.29%**	24.50%	54.56%	68.14%
CL + KL + MSE - TE1	31.22%	59.31%	71.99%	23.64%	53.20%	67.27%
CL + KL + ICL - TE1	31.22%	59.88%	72.08%	24.66%	53.05%	66.34%
CL + KL + MSE + ICL - TE1	30.23%	58.81%	72.24%	24.18%	53.16%	66.64%
All Loss Function	**34.76%**	**63.43%**	74.14%	24.50%	**55.14%**	68.29%

Table 16 reports the hyperparameter sensitivity of retrieval when distilling a student RN34 from an RN50 teacher on Flickr8k. Overall, incorporating the TE reward has a clear and consistent impact: relative to the baseline without TE ($\gamma = 0$), increasing γ improves both I2T and T2I performance, with the best overall results achieved at a moderate TE weight. In particular, $\gamma = 5$ yields the strongest bidirectional retrieval, attaining the highest I2T R@5/R@10 ($65.65/77.84$) and

the best T2I R@1/R@5/R@10 ($27.00/57.20/70.43$), while slightly larger γ (e.g., 7.5) maintains similar I2T R@1 but reduces T2I, indicating diminishing returns and mild over-regularization at high TE weights. Adjusting the other loss weights has comparatively smaller effects: increasing α to 5 or β to 100 under $\gamma = 5$ produces performance close to, but not exceeding, the best setting, whereas increasing the ICL weight ($\delta = 5$) noticeably degrades retrieval, suggesting that overly strong interaction-level supervision can be harmful on this smaller dataset. Adding the MI term alone ($\eta \in \{1, 2.5, 5, 7.5\}$ with $\gamma = 0$) yields improvements over the baseline and can approach the best TE-regularized performance (e.g., $\eta = 5$ achieves I2T R@10 of 77.10 and T2I R@5 of 57.03), but combining MI and MSE-Δ with TE does not provide consistent additive gains within this sweep. Taken together, these results indicate that Flickr8k benefits most from moderate TE regularization (around $\gamma = 5$), while the overall objective remains relatively robust to moderate variations in the remaining hyperparameters.

Table 16: Comparison of retrieval performance (Recall@k) of an RN34-based VLM student distilled from CLIP RN50 on Flickr8k.

α	β	δ	η	λ	γ	I2T R@1	I2T R@5	I2T R@10	T2I R@1	T2I R@5	T2I R@10
1	50	1	0	0	0	26.61%	58.15%	69.93%	22.49%	51.48%	64.74%
1	50	1	0	0	1	31.55%	61.70%	71.99%	24.38%	53.86%	67.84%
1	50	1	0	0	2.5	35.34%	61.94%	73.72%	25.62%	53.11%	67.17%
1	50	1	0	0	5	35.42%	**65.65%**	**77.84%**	**27.00%**	**57.20%**	**70.43%**
1	50	1	0	0	7.5	**35.50%**	63.67%	75.45%	25.55%	55.42%	68.62%
1	50	1	0	0	10	33.20%	62.27%	73.72%	23.90%	54.94%	68.95%
5	50	1	0	0	5	34.10%	63.26%	75.29%	25.17%	54.00%	68.27%
1	100	1	0	0	5	33.53%	64.50%	75.86%	25.67%	55.30%	69.79%
1	50	5	0	0	5	29.65%	59.47%	73.72%	24.32%	52.39%	65.98%
1	50	1	0	50	0	29.00%	58.48%	70.10%	23.25%	51.66%	65.55%
1	50	1	0	100	0	29.08%	58.24%	70.68%	24.18%	52.83%	66.39%
1	50	1	1	1	0	32.70%	62.60%	74.14%	26.00%	54.71%	67.64%
1	50	1	2.5	0	0	31.30%	60.96%	73.39%	24.18%	53.33%	67.23%
1	50	1	5	0	0	34.02%	65.16%	77.10%	26.24%	57.03%	69.67%
1	50	1	7.5	0	0	33.20%	64.66%	75.95%	25.27%	53.99%	67.78%
1	50	1	2.5	50	2.5	32.13%	62.77%	75.37%	25.02%	53.99%	67.51%
1	50	1	2.5	0	5	31.14%	64.99%	76.28%	24.68%	55.90%	69.09%
1	50	1	2.5	50	5	33.61%	62.19%	74.79%	24.84%	55.58%	68.86%
1	50	1	1	0	7.5	31.14%	62.44%	74.63%	24.81%	55.14%	68.55%
1	50	1	1	50	7.5	32.95%	62.44%	74.14%	26.33%	55.67%	68.65%
1	50	1	5	50	7.5	30.15%	60.87%	74.55%	23.84%	54.55%	68.01%

G.9 TEACHER: RN50, STUDENT MODEL: RN18

In addition to using RN34 as the student model, we also conduct experiments with RN18 as the student image encoder. The RN18 architecture is a more compact variant, containing approximately 11.7 million parameters (He et al., 2016). Similar to RN34, the final fully connected layer is modified to output 1024-dimensional features, keeping the overall parameter count stable. Given that the text encoder remains unchanged, the total number of parameters for the RN18-based student model is approximately 45-50 million. This reduction in model size compared to the RN34-based student allows for a more lightweight design while still leveraging the benefits of contrastive learning and effective knowledge transfer from the teacher model.

We applied the same loss components and loss weights as in Section 5, with $\alpha = 1.0$, $\beta = 50$, $\delta = 1.0$, and $\gamma = 1.0$, but used a temperature of $\tau = 0.05$ for this RN18 experiment. Figure 8 presents the training losses and TE rewards over epochs for various configurations. Compared to RN34, RN18 exhibits a similar trend where the total training loss steadily decreases, and TE rewards increase over epochs, indicating effective optimization and knowledge transfer. However, due to the smaller capacity of RN18, the absolute TE rewards remain slightly lower than those observed for RN34, suggesting a less expressive feature alignment between teacher and student. Furthermore, the KL loss and MSE components show even less significant reductions over training epochs, likely due to the more limited representational capacity of RN18. This highlights that while TE-based regularization remains effective in guiding knowledge distillation, the overall learning dynamics are constrained by the smaller network size, making RN34 a more effective student model in terms of retaining structured alignment with the teacher.

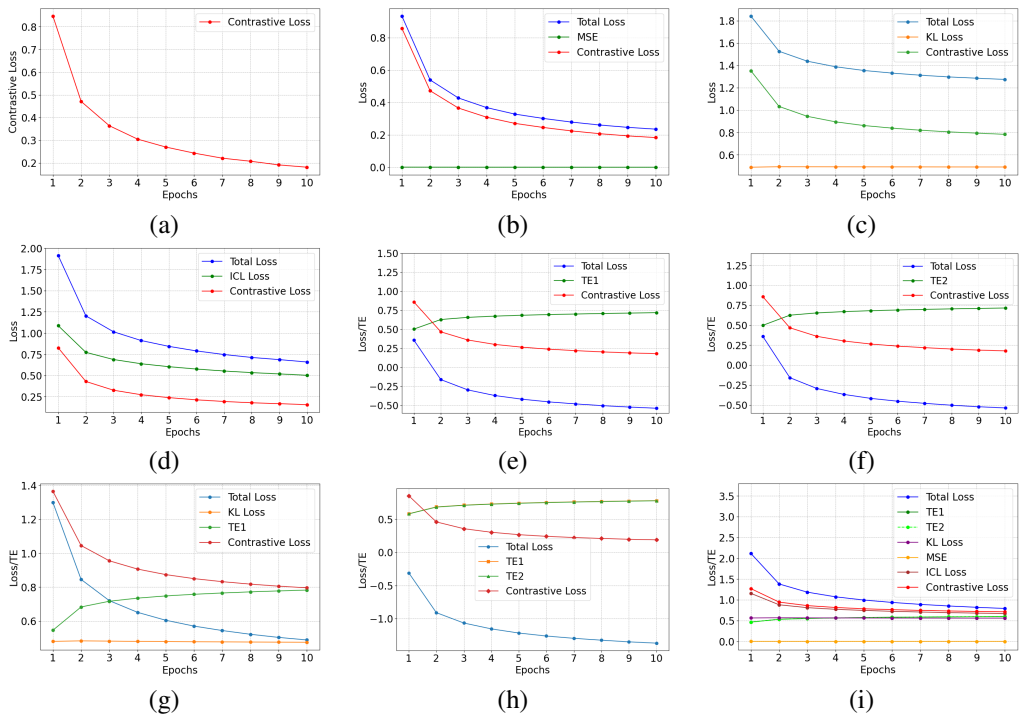


Figure 8: The training losses and TE for different loss functions in the training of RN18-based VLM Student with teacher CLIP RN50. (a) Contrastive only, (b) Contrastive + MSE, (c) Contrastive + KL, (d) Contrastive + ICL, (e) Contrastive - TE1, (f) Contrastive - TE2, (g) Contrastive + KL - TE1, (h) Contrastive - TE1 - TE2, (i) Contrastive + KL + ICL + MSE - TE1 - TE2.

We used Google Colab Pro with a T4 GPU and High-RAM for training and evaluating RN18. Due to its significantly fewer parameters compared to RN34, the student model RN18 required less training time. We trained it for 10 epochs in each loss function combination scenario, with the training and evaluation process taking approximately 11 hours per experimental setup.

We summarize the retrieval performance for the trained RN18 student model in Table 17. Similar observations we can make that the experiment with loss function (Contrastive - TE1 -TE2) achieved the best performance for Image-to-Text retrieval, while the experiment with loss function (Contrastive + KL + MSE + ICL - TE1 - TE2) achieved the best performance in Text-to-Image Retrieval. Comparing Table 17 with Table 1, the results indicate that while RN18 achieves competitive performance across different loss function combinations, it underperforms compared to RN34 for all loss configurations, with RN34 consistently yielding higher Recall@k values. However, the best-performing RN18 model (Contrastive - TE1 - TE2) achieves Recall@1 of 6.65% for image-to-text retrieval, which is not far behind RN34's highest Recall@1 values of 8.24% under the same loss formulation. This suggests that while RN18 is a lighter-weight alternative, RN34 remains a better choice for preserving retrieval performance during distillation. The trade-off between model complexity and retrieval accuracy highlights the importance of selecting an appropriate student architecture based on deployment constraints and performance requirements.

G.10 Teacher: RN50x16, Student: RN34

Table 18 reports retrieval results on MSCOCO at $\tau = 0.07$ when distilling an RN34 student from a substantially stronger teacher, CLIP RN50×16. Across the explored settings, the student consistently outperforms the corresponding RN50/RN34 and ViT-B/16/RN34 configurations, suggesting that a more competitive teacher provides richer supervisory signal under the same distillation framework. As in the earlier experiments, moderate TE regularization remains beneficial. In particular, with $\gamma \in \{2.5, 5\}$, the student achieves the best overall trade-offs: $\gamma = 2.5$ yields the strongest

Table 17: Comparison of retrieval performance (Recall@k) of an RN18-based VLM student distilled from CLIP RN50 under different loss function combinations on MSCOCO. All Loss Function: CL + KL + MSE + ICL - TE1 - TE2.

Model and Loss Function	I2T Retrieval (R)			T2I Retrieval (R)		
	R@1	R@5	R@10	R@1	R@5	R@10
Student Models (RN18)						
CL Only (Oord et al., 2018)	4.38%	13.28%	20.40%	3.39%	11.07%	17.22%
CL + MSE (Yang et al., 2024)	4.27%	13.29%	20.15%	3.47%	11.17%	17.28%
CL + KL (Li et al., 2024a)	4.89%	15.23%	22.90%	4.58%	13.99%	21.05%
CL + ICL (Yang et al., 2024)	5.39%	15.48%	22.95%	4.32%	13.23%	19.96%
CL - TE1	5.48%	16.43%	24.59%	4.60%	13.86%	20.80%
CL - TE2	5.57%	16.67%	24.78%	4.67%	14.08%	20.97%
CL - TE1 - TE2	**6.65%**	**18.75%**	**27.33%**	5.18%	15.09%	22.35%
CL + KL - TE1	6.49%	18.37%	26.83%	5.18%	15.17%	22.52%
All Loss Function	6.52%	18.60%	27.16%	**5.79%**	**16.78%**	**24.47%**

T2I performance, reaching R@1/R@5/R@10 of $8.40/22.45/31.50$, while increasing the feature distillation weight to $\beta = 100$ with $\gamma = 5$ gives the best I2T result, reaching $10.75/27.20/37.41$.

Table 18: Comparison of retrieval performance (Recall@k) in percentage of an RN34-based VLM student distilled from CLIP RN50×16 on MSCOCO.

α	β	δ	η	λ	γ	I2T R@1	I2T R@5	I2T R@10	T2I R@1	T2I R@5	T2I R@10
1	50	1	0	0	2.5	10.29%	26.18%	36.56%	**8.40%**	**22.45%**	**31.50%**
1	50	1	0	0	5	10.52%	26.88%	37.19%	7.85%	21.11%	29.80%
1	100	1	0	0	5	**10.75%**	**27.20%**	**37.41%**	8.01%	21.32%	29.86%
1	50	1	0	0	7.5	9.99%	26.47%	36.48%	7.66%	20.51%	29.06%
1	50	1	1	0	0	9.64%	25.14%	35.55%	7.64%	20.97%	29.79%
1	50	1	2.5	0	0	9.93%	25.94%	35.97%	7.84%	21.26%	30.22%
1	50	1	5	0	0	9.81%	25.57%	35.87%	7.79%	20.99%	29.80%
1	50	1	2.5	50	5	10.01%	25.97%	36.29%	7.77%	20.88%	29.58%
1	50	1	1	50	7.5	10.08%	26.00%	36.21%	7.49%	20.21%	28.73%
Teacher (RN50×16)						32.14%	53.62%	62.67%	17.08%	33.54%	42.00%

Larger TE weights, such as $\gamma = 7.5$, slightly reduce both I2T and T2I performance, again suggesting diminishing returns and mild over-regularization when the reward term becomes too strong. By comparison, MI-only settings ($\eta \in \{1, 2.5, 5\}$ with $\gamma = 0$) remain competitive but are consistently slightly weaker than the best TE-regularized configurations. Adding further auxiliary terms, such as MI together with MSE-Δ, does not yield clear additional gains within this sweep. Finally, despite the strong teacher performance (I2T R@1 of 32.14% and T2I R@1 of 17.08%), a substantial teacher–student gap remains, highlighting the difficulty of compressing RN50×16 into RN34. Even so, these results show that carefully tuned TE regularization and feature distillation can still produce meaningful retrieval improvements under a substantially stronger teacher.

We emphasize that this experiment should be interpreted as a capacity-gap stress test rather than a full scaling study. Although the RN50×16 teacher is substantially stronger than the RN34 student, a large absolute teacher-student performance gap remains. Thus, the result does not show that TE-based distillation closes the gap to a much larger teacher. Instead, it shows that the TE-inspired regularizer can still provide student-side improvements over comparable objectives under a larger teacher. Establishing true scaling behavior will require experiments with larger students, stronger VLM teachers, and larger-scale training data, which we leave for future work.

G.11 Distillation for Classification

G.11.1 Within-Dataset Label-Free Classification on Food-101

We evaluate our TE-based distillation on Food-101 (Bossard et al., 2014), which contains 101 categories and 101,000 images, split into 75,750 training and 25,250 test images. The dataset provides

a complementary setting for evaluating whether teacher knowledge can be transferred to a smaller student without using ground-truth labels.

In our setup, the teacher is CLIP RN50 and the student is RN34. During distillation, the student uses Food-101 images but does not access their ground-truth labels, learning only through the distillation objectives derived from the teacher. We therefore refer to this setting as *label-free classification*, rather than dataset-level zero-shot classification.

Table 19 reports classification accuracy under different loss weights. The baseline with $\gamma = 0$ achieves 80.23% Top-1 and 95.22% Top-5 accuracy, compared with 79.80% and 96.17% for the teacher. Adding TE generally improves the student: $\gamma = 2.5$ increases Top-1 accuracy to 82.46%, while $\gamma = 7.5$ gives the best Top-5 accuracy of 96.62%. The highest Top-1 accuracy, 82.91%, is obtained with $\alpha = 5$ and $\gamma = 2.5$. These results suggest that the TE-inspired objective provides complementary supervision beyond the conventional distillation losses. Each experiment in Table 19 takes approximately 45 minutes on a Colab A100 GPU.

Table 19: Label-free classification accuracy (%) of the RN34 student and CLIP RN50 teacher on Food-101. The student is distilled using Food-101 images without access to ground-truth labels.

α	β	δ	η	λ	γ	Top-1 Acc.	Top-5 Acc.
1	50	1	0	0	0	80.23%	95.22%
1	50	1	0	0	2.5	82.46%	96.23%
1	50	1	0	0	5	82.37%	96.44%
1	50	1	0	0	7.5	82.27%	**96.62%**
1	50	1	0	0	10	82.01%	96.38%
5	50	1	0	0	2.5	**82.91%**	96.47%
1	100	1	0	0	2.5	82.54%	96.10%
1	50	5	0	0	2.5	81.07%	95.30%
1	50	1	2.5	0	0	82.13%	96.10%
1	50	1	5	0	0	82.37%	96.50%
1	50	1	7.5	0	0	81.39%	96.24%
1	50	1	5	50	7.5	81.63%	96.29%
Teacher (RN50)						79.80%	96.17%

Overall, the RN34 student matches or exceeds the RN50 teacher under several configurations despite receiving no ground-truth labels during distillation. The consistent gains obtained by adding TE indicate that the proposed regularizer can improve the transfer of teacher representations in this label-free classification setting.

G.11.2 Out-of-Dataset Zero-Shot Classification on ImageNet-1k

To assess how much semantic structure the student retains beyond Food-101, we evaluate zero-shot classification on the ImageNet-1k benchmark (Deng et al., 2009; Russakovsky et al., 2015). We distill an RN34 student from an RN50 teacher using MSCOCO and Food-101, and then perform zero-shot classification on the ImageNet-1k validation set (50K images) following the standard CLIP-style protocol. Importantly, the student is *never* trained on ImageNet images; thus, this experiment measures out-of-domain semantic transfer under a substantial domain shift.

Table 20: Zero-shot classification accuracy (%) of distilled RN34-based VLM student on ImageNet-1k.

α	β	δ	η	λ	γ	Top-1 Acc.	Top-5 Acc.
1	50	1	0	0	5	10.23%	24.08%
1	50	1	0	0	7.5	10.62%	24.46%
1	50	1	2.5	0	0	8.09%	19.92%
1	50	1	5	0	0	8.61%	21.24%

Table 20 reports Top-1 and Top-5 accuracies for representative hyperparameter settings. Overall, incorporating the TE reward yields the strongest ImageNet performance among the tested configurations: using $\gamma \in \{5, 7.5\}$ achieves Top-1 accuracy of 10.23%–10.62% and Top-5 accuracy of 24.08%–24.46%, outperforming MI-only variants (e.g., $\eta \in \{2.5, 5\}$ with $\gamma = 0$), which attain Top-1 accuracy of 8.09%–8.61% and Top-5 accuracy of 19.92%–21.24%. This trend is consistent with our retrieval ablations, where moderate TE regularization improves alignment and yields better generalization. While absolute ImageNet accuracy remains well below the RN50 teacher, as expected given the student's smaller capacity and the fact that distillation is performed only on MSCOCO and Food-101, the improvement from TE indicates that TE-guided distillation helps preserve transferable semantic information under a challenging out-of-domain evaluation.

G.12 Computational Overhead, Statistical Robustness, and Metric Definitions

Computational overhead. Appendix C explains why computing exact TE is intractable, but it does not quantify the additional cost of our TE surrogates (TE1 and TE2) relative to a standard distillation run. To make this overhead explicit, we measured wall-clock time and FLOPS for two settings on Google Colab A100 GPUs: training with the baseline loss (CL + MSE + KL + ICL) and training with the objective including TE (CL + MSE + KL + ICL - TE). We report the total time for 10 training epochs plus evaluation, as well as the total FLOPS and the portion attributable to the TE terms, in Table 21.

Table 21: Computation and time cost comparison for training (10 epochs) and evaluation on MSCOCO.

Teacher	Student	CL+KL+MSE+ICL–TE	CL+KL+MSE+ICL	FLOPS	FLOPS from TE
RN50	RN34	6h 19m	6h 18m	3.44×10^{16}	1.79×10^{10}
ViT-B/16	RN34	5h 33m	5h 32m	2.60×10^{16}	1.49×10^{10}

The reported FLOPS correspond to the entire training and evaluation run with the loss CL+KL+MSE+ICL-TE. The FLOPS contributed by the TE surrogates are six orders of magnitude smaller than the total, and the wall-clock times with and without TE only have 1 minute difference, indicating that the additional computational cost of TE1/TE2 is negligible in practice. The ViT-B/16 and RN50 runs use different teacher architectures, so their wall-clock times are not directly attributable to a single representation dimension; Table 21 reports the measured end-to-end cost for each configuration.

Seed variability. To assess the robustness of our empirical findings, we run each configuration with five independent random seeds and report mean $\pm$ standard deviation for all retrieval metrics in Table 22. On MSCOCO, the standard deviations are very small (typically on the order of 0.1–0.3), indicating that both I2T and T2I performance are highly stable under different initializations and data shuffling. Flickr30k exhibits slightly larger but still modest variability, suggesting mild sensitivity while remaining well-behaved. In contrast, Flickr8k shows noticeably larger standard deviations (around 1.0), which is expected given its much smaller size and higher sampling noise. Overall, these trends confirm that our conclusions are robust on larger benchmarks, with variability naturally increasing on smaller datasets such as Flickr8k.

Mean reciprocal rank. Beyond Recall@K, we evaluate Mean Reciprocal Rank (MRR) (Schütze et al., 2008) for both I2T and T2I retrieval. MRR is a ranking metric that evaluates how early the first relevant item appears in a retrieved list. For each query, the reciprocal rank is the inverse of the rank of the first correct result. Averaging across N queries gives

$$\mathrm{MRR} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\mathrm{rank}_i}, \tag{62}$$

where rank_i denotes the position of the first relevant item for query i. Higher MRR indicates that correct results tend to appear earlier in the ranking. For I2T, each image is associated with multiple ground-truth captions: we compute similarities to all caption embeddings, determine the 1-indexed

Table 22: Retrieval (mean ± std over five seeds, %) of RN34 students distilled from different CLIP teachers on different datasets ($\delta = 1$, $\lambda = 0$). TE rows use $\eta = 0$; MI rows use $\gamma = 0$. Best within each teacher–dataset pair and metric in bold.

Teacher	Dataset	α	β	η	γ	Image-to-Text			Text-to-Image		
						R@1	R@5	R@10	R@1	R@5	R@10
RN50	MSCOCO	1	50	0	7.5	**10.08±0.17**	**26.10±0.19**	**36.22±0.21**	7.12±0.19	19.35±0.27	27.85±0.23
RN50	MSCOCO	1	50	5	0	8.35±0.22	23.34±0.16	33.25±0.24	**7.35±0.28**	**20.57±0.18**	**28.95±0.26**
RN50	Flickr8k	1	50	0	5	**35.81±1.12**	**65.67±1.13**	**77.47±1.19**	**26.74±1.11**	56.02±1.03	69.22±1.02
RN50	Flickr8k	1	50	5	0	33.22±1.15	64.18±1.09	76.90±1.08	26.32±1.21	**57.12±1.12**	**69.31±1.06**
ViT-B/16	MSCOCO	1	100	0	5	**8.01±0.21**	**21.71±0.18**	**30.75±0.16**	6.03±0.23	17.27±0.15	25.03±0.18
ViT-B/16	MSCOCO	1	50	5	0	7.07±0.16	19.98±0.22	28.97±0.17	**6.13±0.26**	**17.46±0.16**	**25.30±0.21**
ViT-B/16	Flickr30k	5	50	0	5	**42.24±0.79**	**69.42±0.84**	**79.84±0.89**	**31.46±0.82**	**60.56±0.69**	**72.11±0.62**
ViT-B/16	Flickr30k	1	50	5	0	38.75±0.81	65.92±0.78	76.23±0.83	31.21±0.72	60.14±0.64	71.70±0.68

rank of each correct caption, and use the best (minimum) rank to accumulate the reciprocal rank $1/\mathrm{rank}$. For T2I, each caption has a single correct image; we rank all images by similarity to the caption and again use $1/\mathrm{rank}$ to compute MRR. All metrics are computed over the full validation split so that MRR follows the same protocol as Recall@K while providing a more fine-grained assessment of ranking quality across the entire list.

Table 23: Comparison of zero-shot MRR (mean ± std) with different settings ($\delta = 1$, $\lambda = 0$).

Teacher	Student	Dataset	α	β	η	γ	I2T MRR (%)	T2I MRR (%)
RN50	RN34	MSCOCO	1	50	0	7.5	18.32±0.12	13.93±0.19
RN50	RN34	MSCOCO	1	50	5	0	16.46±0.14	14.34±0.23
RN50	RN34	Flickr8k	1	50	0	5	49.03±1.02	39.09±1.15
RN50	RN34	Flickr8k	1	50	5	0	44.62±1.17	38.47±1.12
ViT-B/16	RN34	MSCOCO	1	100	0	5	15.52±0.15	12.46±0.11
ViT-B/16	RN34	MSCOCO	1	50	5	0	14.23±0.22	12.63±0.09
ViT-B/16	RN34	Flickr30k	5	50	0	5	54.79±0.74	45.38±0.65
ViT-B/16	RN34	Flickr30k	1	50	5	0	52.21±0.62	45.02±0.71

Table 23 reports five-seed MRR for the same models as Table 22. Together, the two tables compare the TE-based distillation term (settings with $\eta = 0$, $\gamma > 0$) against the MI-based baseline ($\eta > 0$, $\gamma = 0$). Across all four teacher–student–dataset combinations, TE consistently achieves higher I2T performance than MI, both in Recall@K and in MRR. For T2I, TE is competitive with or better than MI on most metrics, with MI only occasionally matching or slightly exceeding TE on a single Recall@K value. On the larger MSCOCO and Flickr30k benchmarks, the gains from TE are especially clear, indicating that the TE surrogate provides a stronger and more reliable signal for shaping the multimodal embedding space. On the smaller Flickr8k dataset, TE still yields higher I2T MRR and Recall@K than MI despite the higher variance, suggesting that matching teacher–student finite-difference geometry is more effective than MI-based matching under capacity mismatch. Overall, TE outperforms MI in our settings, improving the ranking quality of the student model while maintaining stable training dynamics.

H LIMITATIONS

Despite the demonstrated effectiveness of our TE-inspired proxy regularization for VLM distillation, the approach has several limitations. First, the method does not directly estimate optimization-step transfer entropy; instead, it relies on tractable geometric surrogates based on squared-cosine alignment of the one-dimensional subspaces spanned by local finite differences. While this makes the method computationally practical in high-dimensional multimodal embedding spaces, it also means that the surrogate may not fully capture the complex nonlinear dependencies that motivate the original TE perspective. Accordingly, the empirical results in this paper should be interpreted as validating the usefulness of the final proxy objective, rather than as a direct empirical verification of the ideal TE quantity.

Second, while TE1/TE2 consistently improve retrieval, the student can still remain far from the teacher because these surrogates capture only a *restricted* geometric signal: alignment of *local* representation variations between teacher and student on a batch. Several important aspects of the teacher geometry are not directly constrained. TE1/TE2 do not enforce *global* isometry of the embedding space, such as preservation of long-range pairwise distances, full-distribution neighborhood structure, or higher-order manifold curvature, so a student may match local secant-line geometry while remaining globally mis-scaled or warped. Because the squared-cosine score is orientation-invariant, it also does not distinguish parallel from anti-parallel finite differences or directly control the *magnitude* or anisotropy spectrum of variation, such as principal directions or singular-value structure, which can still affect retrieval quality.

Third, TE1/TE2 operate only on final embeddings and therefore do not explicitly constrain *intermediate* representations, cross-modal interaction mechanisms, or capacity-dependent inductive biases that may contribute to the teacher's stronger semantic organization. In addition, because the finite-difference directions are formed from within-batch pairings, the surrogate samples only a limited subset of directions in the input space. Directions that are especially important for hard negatives, rare concepts, or fine-grained attributes may be underrepresented, leaving residual teacher–student mismatch. The current framework also assumes access to paired teacher–student representations for each sample, which may be less suitable in settings with partial supervision, noisy teacher signals, or missing modalities.

Fourth, although the proposed TE-inspired proxy is evaluated across multiple teacher backbones and datasets, the current experiments are primarily conducted in CLIP-style dual-encoder distillation settings with compact student models. This setting is appropriate for studying global image–text representation alignment, retrieval, and zero-shot transfer, but it does not directly cover multimodal reasoning benchmarks such as VQA, NLVR2, or visual commonsense reasoning. These tasks typically require additional cross-modal fusion, task-specific prediction heads, or generative reasoning architectures, whereas our method operates on pooled image and text embeddings. Extending the proposed teacher–student orientation-invariant finite-difference objective to fusion-based or generative VLMs is an important direction for future work. In addition, while our experiments include stronger teacher backbones, further evaluation with larger students and larger-scale VLMs would provide a more complete understanding of the scaling behavior of the proposed proxy regularizer.

These limitations suggest that TE1/TE2 are best viewed as efficient geometric regularizers that complement, rather than replace, standard distributional matching losses such as contrastive, KL, MSE, or ICL-based objectives. Closing the remaining teacher–student gap will likely require richer direction sampling, stronger global-structure constraints, improved intermediate-level alignment, or higher-capacity student models. Exploring such extensions is an important direction for future work.

I Statement of Broader Impact

The development of efficient distillation frameworks for VLMs, such as our proposed TE-inspired proxy regularization approach, has the potential to yield several positive societal impacts. By enabling the deployment of compact yet high-performing multimodal models on resource-constrained devices, this work may broaden access to advanced AI technologies for individuals and communities with limited computational resources. Such efficiency gains could support wider adoption in applications such as assistive technologies for people with disabilities, low-cost language translation tools, educational platforms, and on-device multimodal interfaces. In addition, parameter-efficient VLM distillation can reduce energy consumption and the environmental footprint associated with large-scale model training and inference, thereby contributing to more sustainable AI deployment.

At the same time, broader deployment of efficient VLMs raises important risks. Increased accessibility to vision–language technology may amplify concerns related to privacy, surveillance, and misuse, including unauthorized content analysis, large-scale monitoring, or the generation of misleading multimodal content. Moreover, distillation can propagate or even amplify biases already present in the teacher model, potentially leading to unfair or discriminatory outcomes for underrepresented groups. Compression may also remove subtle but important information, which could degrade robustness, calibration, or fairness in real-world settings, especially when models are deployed under distribution shift.

To help mitigate these risks, we advocate several practical safeguards. First, bias and fairness audits should be incorporated throughout both teacher evaluation and student distillation, particularly for sensitive or high-impact applications. Second, strong data governance and privacy-preserving practices are important when training or deploying VLMs on user-facing or sensitive data. Third, transparency should be maintained through model cards, clear reporting of evaluation settings and limitations, and documentation of the distillation pipeline. Finally, interdisciplinary collaboration with ethicists, domain experts, and affected stakeholders can help identify harms early and guide more responsible deployment. By proactively addressing these concerns, TE-inspired VLM distillation may be developed and applied in a safer, more equitable, and more socially beneficial manner.