CONTRASTIVE GEOMETRY DISTILLATION FOR VISION-LANGUAGE DUAL ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose contrastive geometry distillation (CGD) for vision–language dual encoders. A decomposition of the teacher–student mutual information gap motivates two variants that combine feature distillation with KL matching of caption distributions for each image. CGD-N also matches distributions over unmatched images for each caption, while CGD-IW corrects selected unmatched pairs using teacher-to-student probability ratios. For CGD-N, exact matching recovers the teacher's similarity matrix up to a global additive constant. Experiments show that CGD lowers similarities for negative pairs more than for positive pairs, widening retrieval margins while retaining alignment of image features with the teacher. The added losses supply small gradient corrections in directions distinct from feature distillation. The resulting students preserve the teacher's candidate ordering more closely than CLIP-KD and CLIP-RD, with improved average retrieval and comparable classification accuracy.

1 INTRODUCTION

Modern generative models process language, vision, audio, and action (OpenAI, 2024; Zitkovich et al., 2023). Shared representations provide one way to connect these modalities (Girdhar et al., 2023). Contrastive Language–Image Pretraining (CLIP, Radford et al., 2021) learns such representations from 400 million image–text pairs collected from the web. Its contrastive objective increases the similarity of matched images and captions relative to mismatched pairs, using naturally occurring text instead of manually assigned class labels. The resulting representations support zero-shot classification and image–text retrieval. For a given caption, the relative scores describe which images compete and how strongly. For a given image, they describe relations among captions. We refer to this relative cross-modal score structure as *contrastive geometry*.

Deploying these models under limited compute motivates transferring their capabilities to smaller encoders through knowledge distillation (Hinton et al., 2015). TinyCLIP (Wu et al., 2023), CLIP-KD (Yang et al., 2024a), and CLIP-RD (Chung et al., 2026) provide supervision in the embedding and score spaces. These methods already transfer relations between examples. We study how feature distillation and matching the distributions of candidate scores jointly constrain the student's image–text comparisons. Small feature errors can change the ranking of closely scored candidates, so feature error alone gives an incomplete description of these comparisons. Unmatched pairs are not interchangeable: a teacher may assign appreciable affinity to a semantically related alternative even when it is not the annotated match. Preserving its contrastive geometry therefore differs from maximizing matched-pair similarity or pushing every unmatched pair farther away. The question is how supervision should allocate corrections to preserve the teacher's relative scores in the student's own score matrix.

We develop *contrastive geometry distillation* (CGD) to transfer these candidate relations, using an analysis of the teacher–student information gap over batch score distributions. Decomposing this gap relates feature error to discrepancies between conditional distributions. Both CGD variants combine feature distillation and matching caption distributions for each image. CGD-IW corrects selected unmatched similarities using the ratio of teacher to student probabilities for image candidates given a caption, while CGD-N matches the distribution over unmatched images for each caption. We show that exact matching of caption distributions and distributions over unmatched images identifies the teacher's score matrix up to a constant. Five-run comparisons and component studies examine how these objectives affect classification and retrieval. Five-seed representation

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.

diagnostics also test agreement with the teacher: both variants have higher full-list negative rank agreement than CLIP-KD and CLIP-RD, although this does not extend to every local neighborhood.

2 Problem Formulation and Preliminaries

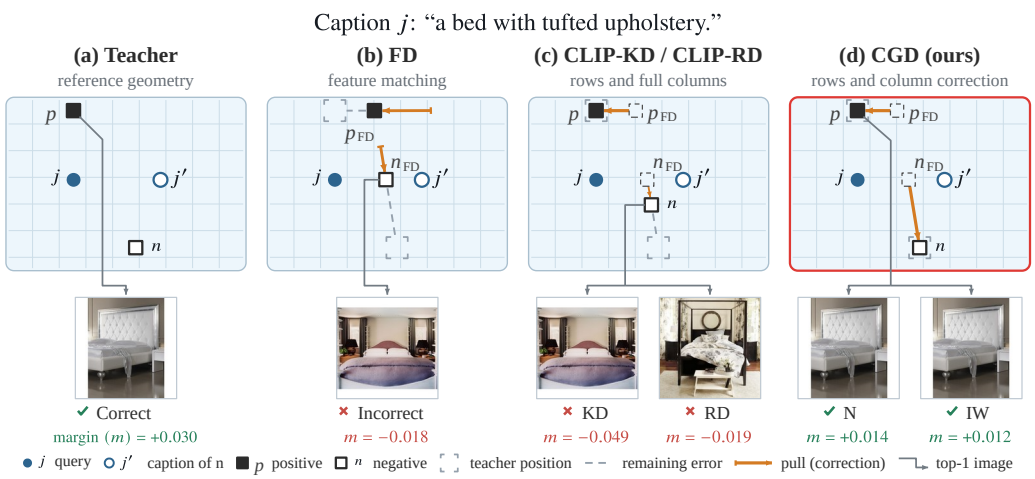


Figure 1: **Preserving the teacher's contrastive geometry. Top:** (a) for caption j, the teacher ranks matched image p above negative image n. (b) FD aligns features without explicitly comparing p and n. (c) The added correction moves n away from j but leaves n closer than p. Full-column KL weights $D_{\mathrm{KL}}(c_j^{T,-} \| c_j^{S,-})$ by $1 - c_j^T(p)$, weakening this correction when the teacher is confident in p. (d) CGD-N weights $\mathcal{L}_N$ independently. CGD-IW uses ratios $c_j^T(i)/c_j^S(i)$. The illustrated correction restores the order. Panels (c,d) start at dashed FD outputs. **Bottom:** selected CC3M-val retrievals (epoch 32, run 0), with $m = S_{pj}^M - \max_{i \neq p} S_{ij}^M$. Positions and arrows are illustrative. j' is hypothetical (Section 3.2).

2.1 Cosine Similarity Matrix

We consider a batch $\mathcal{B} = \{(x_i, t_i)\}_{i=1}^B$ of matched image–caption pairs. For the teacher T and student S, let $u_i^M, v_i^M \in \mathbb{R}^D$, $M \in \{T, S\}$, denote unit-normalized image and text embeddings. The cosine similarity matrix $S^M \in \mathbb{R}^{B \times B}$ contains the candidate cosine similarities $S_{ij}^M = \langle u_i^M, v_j^M \rangle$. Diagonal entries correspond to matched pairs. Off-diagonal entries are treated as negative pairs within the batch.

Conditional distributions over candidates. At temperature $\tau > 0$, $r_i^M(\cdot;\tau)$ is the caption distribution conditioned on image i, and $c_j^M(\cdot;\tau)$ is the image distribution conditioned on caption j:

$$r_i^M(j;\tau) = \frac{e^{S_{ij}^M/\tau}}{\sum_{\ell=1}^B e^{S_{i\ell}^M/\tau}}, \qquad c_j^M(i;\tau) = \frac{e^{S_{ij}^M/\tau}}{\sum_{\ell=1}^B e^{S_{\ell j}^M/\tau}}. \tag{1}$$

At a common temperature, matching a conditional preserves cosine similarity differences within its row or column, since $\tau \log[r_i^M(j;\tau)/r_i^M(\ell;\tau)] = S_{ij}^M - S_{i\ell}^M$.

Joint distributions over candidate pairs. Both teacher and student score the same B^2 image–caption pairs. To analyze the two conditional directions together, we normalize each similarity matrix over all pairs at temperature $\tau_G > 0$:

$$P(i,j) = \frac{e^{S_{ij}^T/\tau_G}}{\sum_{a,b} e^{S_{ab}^T/\tau_G}}, \qquad Q(i,j) = \frac{e^{S_{ij}^S/\tau_G}}{\sum_{a,b} e^{S_{ab}^S/\tau_G}}. \tag{2}$$

Here U, V index images and captions, with marginals P_U, Q_U and P_V, Q_V. Conditioning recovers the candidate distributions: $P(V = j \mid U = i) = r_i^T(j;\tau_G)$ and $P(U = i \mid V = j) = c_j^T(i;\tau_G)$, with the same relations for Q.

2.2 Distillation Objective

With the teacher fixed, we train both student encoders using

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{distill}}, \tag{3}$$

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log r_i^S(i;\tau_S) + \log c_i^S(i;\tau_S)\right]. \tag{4}$$

The task loss $\mathcal{L}_{\text{CLIP}}$ is CLIP's symmetric InfoNCE loss (Radford et al., 2021; van den Oord et al., 2018) at a learned τ_S. The distillation term $\mathcal{L}_{\text{distill}}$ supplies teacher supervision, using common, fixed temperatures when matching candidate distributions. Under i.i.d. sampling with marginal negatives, InfoNCE lower-bounds the mutual information (MI) of encoded variables (Poole et al., 2019). The MI analyzed below is that of the batch-level joints P, Q (Appendix B).

2.3 Related Work on CLIP Distillation

TinyCLIP (Wu et al., 2023) combines weight inheritance with affinity mimicking of row distributions r_i^M and column distributions c_j^M. CLIP-KD (Yang et al., 2024a) studies feature and relational distillation. Its feature distillation aligns the two models' image and text embeddings:

$$\mathcal{L}_{\text{FD}} = \frac{1}{BD}\sum_{i=1}^{B}\left(\|u_i^T - u_i^S\|_2^2 + \|v_i^T - v_i^S\|_2^2\right). \tag{5}$$

Equation (5) averages over the D coordinates. CLIP-KD's contrastive relational distillation (CRD) matches the row distributions r_i^M and column distributions c_j^M, including matched pairs:

$$\mathcal{L}_{\text{CRD}} = \frac{1}{B}\sum_i D_{\text{KL}}(r_i^T \| r_i^S) + \frac{1}{B}\sum_j D_{\text{KL}}(c_j^T \| c_j^S). \tag{6}$$

CLIP-KD applies a contrastive loss between student image and teacher text embeddings, and between student text and teacher image embeddings, termed Interactive Contrastive Learning (ICL). CLIP-RD (Chung et al., 2026) further constrains similarities formed across the two models. Appendix A compares their objectives and gradients with CGD.

3 Methodology

3.1 Information Gap between Teacher and Student in Multi-Modal Alignment

To relate feature distillation to matching candidate distributions, we decompose the MI difference between the teacher and student joints P, Q. Their mutual information is

$$I_P = \mathbb{E}_{(i,j)\sim P}\left[\log \frac{P(i,j)}{P_U(i)P_V(j)}\right], \qquad I_Q = \mathbb{E}_{(i,j)\sim Q}\left[\log \frac{Q(i,j)}{Q_U(i)Q_V(j)}\right]. \tag{7}$$

Equal MI need not preserve candidate relationships (Appendix B.1). Our target is therefore the relations in P, Q, not equality of their scalar MI values.

Comparing I_P and I_Q changes both the pointwise mutual information and the distribution used to average it. We first keep the student's pointwise mutual information fixed and average it under P:

$$I_Q^{\text{IW}} = \mathbb{E}_{(i,j)\sim P}\left[\log \frac{Q(i,j)}{Q_U(i)Q_V(j)}\right]. \tag{8}$$

Importance weighting by P/Q gives the same expectation under Q. Adding and subtracting I_Q^{IW} splits the gap into $(I_P - I_Q^{\text{IW}}) + (I_Q^{\text{IW}} - I_Q)$.

The KL chain rule expresses $I_P - I_Q^{\text{IW}}$ using the row distributions r_i^T, r_i^S at τ_G. Define

$$\Delta_{\text{dist}} = \mathbb{E}_{i\sim P_U}\left[D_{\text{KL}}(r_i^T \| r_i^S)\right], \qquad \Delta_{\text{info}} = I_Q^{\text{IW}} - I_Q. \tag{9}$$

Δ_{dist} measures disagreement between caption distributions for the same image. Δ_{info} keeps $\log[Q(i,j)/(Q_U(i)Q_V(j))]$ fixed and changes its averaging distribution.

Proposition 1 (Information gap decomposition)**.** *The KL chain rule gives*

$$I_P - I_Q = \Delta_{\text{dist}} - D_{\text{KL}}(P_V \| Q_V) + \Delta_{\text{info}} \le \Delta_{\text{dist}} + \Delta_{\text{info}}. \tag{10}$$

Dropping the nonpositive term $-D_{\mathrm{KL}}(P_V\|Q_V)$ gives the upper bound (Appendix B.3). Here $\Delta_{\mathrm{dist}} \geq 0$, whereas Δ_{info} can have either sign. The KL chain rule alone also gives Δ_{dist}. Comparing MI additionally identifies Δ_{info}.

Exact row matching can leave $P \neq Q$: adding a different constant to each image's row preserves its caption distribution but can change comparisons among images. If $\Delta_{\mathrm{dist}} = 0$, the remaining gap is $\Delta_{\mathrm{info}} - D_{\mathrm{KL}}(P_V\|Q_V)$.

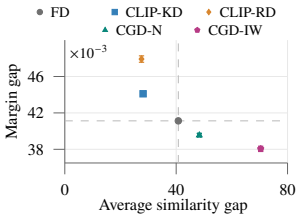


Figure 2: CGD has smaller margin gaps but larger average similarity gaps from the teacher than FD. Dashed lines mark FD (Appendix E.8).

These image comparisons depend on the margin $S^M_{jj} - S^M_{nj}$ between matched image j and competitor n for caption j. Equal changes in their similarities leave this margin unchanged. Figure 2 uses the teacher's most similar unmatched image as n. Before rescaling, CGD has a larger gap from the teacher in average similarity but a smaller margin gap than FD. CLIP-KD/RD show the opposite pattern (Appendix E.8). Matching average similarity more closely therefore does not ensure better preservation of the teacher's margins. To supervise these comparisons, CGD-IW uses a surrogate based on Δ_{info}, while CGD-N uses the column expression of Δ_{dist}.

Variational representation of Δ_{info}. To relate Δ_{info} to feature distillation, let $f(i,j)$ be a scalar critic and $\rho(i,j) = P(i,j)/Q(i,j)$. The NWJ variational representation (Nguyen et al., 2010), using one critic in the ordinary bound $\mathcal{J}_Q(f) \leq I_Q$ and the importance-weighted bound $\mathcal{J}^{\mathrm{IW}}_Q(f) \leq I^{\mathrm{IW}}_Q$, gives

$$\Delta_{\mathrm{info}} = \underbrace{\big(\mathbb{E}_P[f(i,j)] - \mathbb{E}_Q[f(i,j)]\big)}_{\mathcal{J}_+(f)} + \underbrace{\mathbb{E}_{Q_U \otimes Q_V}[(1-\rho(i,j))e^{f(i,j)-1}]}_{\mathcal{J}_-(f)} + \mathcal{R}(f). \tag{11}$$

Here $\mathcal{R}(f) = (I^{\mathrm{IW}}_Q - \mathcal{J}^{\mathrm{IW}}_Q(f)) - (I_Q - \mathcal{J}_Q(f))$ is the difference of the variational slacks. Appendix B.5 defines $\mathcal{J}_Q, \mathcal{J}^{\mathrm{IW}}_Q$ and derives the identity. For training, we use the affine cosine critic $f^S(i,j) = aS^S_{ij}/\tau_{\mathrm{IW}} + b$ with fixed a, b, τ_{IW}. For fixed τ_G, Proposition 5 in Appendix B.5 gives $|\mathcal{R}(f^S)| \leq C\ \mathrm{TV}(P,Q)$ for a uniform $C < \infty$, where $\mathrm{TV}(P,Q) = \frac{1}{2}\|P-Q\|_1$ is total variation. Matching P, Q therefore tightens the bound on $|\mathcal{R}(f^S)|$, and $\mathcal{R}(f^S) \to 0$ as $\mathrm{TV}(P,Q) \to 0$.

3.2 Objectives for Candidate Comparisons

Feature and row matching. For unit-normalized embeddings, Proposition 3 bounds $|\mathcal{J}_+(f^S)|$ in Equation (11) in terms of $\mathcal{L}_{\mathrm{FD}}$:

$$|\mathcal{J}_+(f^S)| \leq \frac{|a|}{\tau_{\mathrm{IW}}\tau_G}\sqrt{2BD\mathcal{L}_{\mathrm{FD}}}. \tag{12}$$

Reducing $\mathcal{L}_{\mathrm{FD}}$ tightens this bound on $|\mathcal{J}_+(f^S)|$. The distribution term Δ_{dist} averages $D_{\mathrm{KL}}(r^T_i\|r^S_i)$ with weights $P_U(i)$. Replacing $P_U(i)$ by $1/B$ at $\tau_{\mathrm{KL}} = \tau_G$ gives the row loss $\mathcal{L}_{\mathrm{KL}}$. Combining $\mathcal{L}_{\mathrm{KL}}$ with $\mathcal{L}_{\mathrm{FD}}$ and the task loss $\mathcal{L}_{\mathrm{CLIP}}$ gives the common objective:

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{B}\sum_i D_{\mathrm{KL}}(r^T_i\|r^S_i), \qquad \mathcal{L}_{\mathrm{base}} = \mathcal{L}_{\mathrm{CLIP}} + \lambda_F\mathcal{L}_{\mathrm{FD}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}. \tag{13}$$

The bound on $|\mathcal{J}_+(f^S)|$ can remain large at attained $\mathcal{L}_{\mathrm{FD}}$ (Appendix B.6). We add $\mathcal{L}_{\mathrm{IW}}$ or $\mathcal{L}_N$ to $\mathcal{L}_{\mathrm{base}}$ to directly supervise image candidates for each caption.

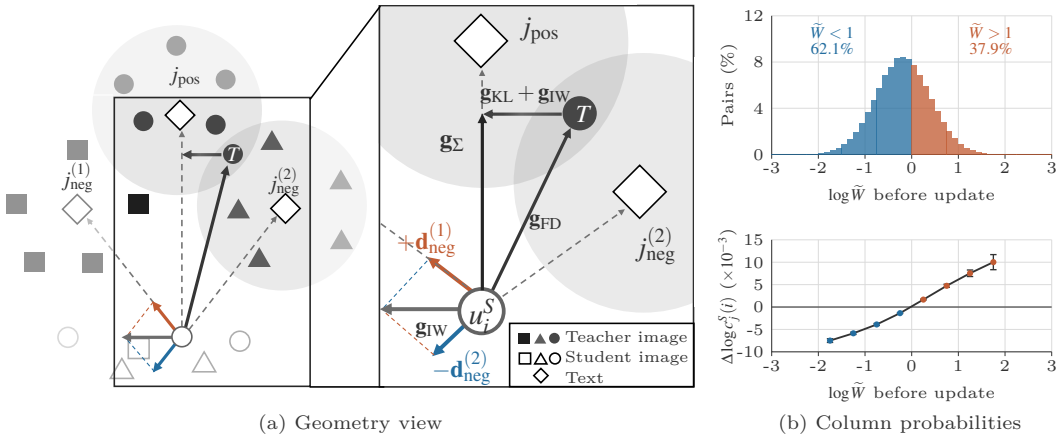
Column correction with importance weights ($\mathcal{L}_{\mathrm{IW}}$). We construct $\mathcal{L}_{\mathrm{IW}}$ from the product term $\mathcal{J}_-$ in Equation (11), using the ratio of column probabilities $c^T_j(i)/c^S_j(i)$ in $\rho(i,j) = [c^T_j(i)/c^S_j(i)][P_V(j)/Q_V(j)]$. Let H contain the teacher's k_h most similar unmatched captions for each image. At temperature τ_w, define

$$W_{ij} = \mathrm{sg}\left[\frac{c^T_j(i)}{c^S_j(i)}\right], \quad Z_H = \mathrm{sg}\left[\frac{1}{|H|}\sum_{(i,j)\in H} W_{ij}\right], \quad \widetilde{W}_{ij} = \frac{W_{ij}}{Z_H}. \tag{14}$$

The weights $\widetilde{W}_{ij}$ have mean one on H. The IW loss is

$$\mathcal{L}_{\mathrm{IW}} = \frac{1}{|H|} \sum_{(i,j)\in H} (1 - \widetilde{W}_{ij}) e^{f^S(i,j)-1}. \tag{15}$$

We recompute H and the weights each batch and detach them during backpropagation. For $a, \tau_{\mathrm{IW}} > 0$, the direct correction raises S^S_{ij} when $\widetilde{W}_{ij} > 1$ and lowers it when $\widetilde{W}_{ij} < 1$ (Figure 3(a)). It vanishes at $W_{ij} = Z_H$. The critic f^S and loss coefficient λ_{IW} determine its magnitude (Equation (20)).



(a) Geometry view (b) Column probabilities

Figure 3: **Feature alignment with corrections from negative captions. (a)** FD pulls the student image embedding u^S_i toward its corresponding teacher embedding u^T_i. In this schematic, row and IW corrections add attraction toward $j^{(1)}_{\mathrm{neg}}$ and repulsion from $j^{(2)}_{\mathrm{neg}}$. The combined direction $\mathbf{g}_\Sigma$ redirects feature alignment, opposing movement toward the second negative caption. **(b)** Initial importance weights (top) and changes in column probabilities from adding $\lambda_{\mathrm{IW}}\mathcal{L}_{\mathrm{IW}}$ for one update (bottom). The probability changes follow the sign of $\widetilde{W} - 1$ (Appendix E.5).

Column correction with negative KL ($\mathcal{L}_N$). CGD-N instead starts from the distribution term Δ_{dist} in Equation (10). The KL chain rule rewrites its row divergences as column divergences with marginal terms $D_{\mathrm{KL}}(P_U\|Q_U)$ and $D_{\mathrm{KL}}(P_V\|Q_V)$:

$$\Delta_{\mathrm{dist}} - D_{\mathrm{KL}}(P_V\|Q_V) = \mathbb{E}_{j\sim P_V}\left[D_{\mathrm{KL}}(c^T_j\|c^S_j)\right] - D_{\mathrm{KL}}(P_U\|Q_U). \tag{16}$$

For caption j, remove the matched image and renormalize: $c^{M,-}_j(i) = c^M_j(i)/(1-c^M_j(j))$ for $i \neq j$. The target/non-target decomposition of Zhao et al. (2022) writes the full column KL $D_{\mathrm{KL}}(c^T_j\|c^S_j)$ as $D_{\mathrm{KL}}(\mathrm{Ber}(c^T_j(j))\|\,\mathrm{Ber}(c^S_j(j)))$ plus $(1-c^T_j(j))D_{\mathrm{KL}}(c^{T,-}_j\|c^{S,-}_j)$ (Appendix B.4). CGD-N weights $D_{\mathrm{KL}}(c^{T,-}_j\|c^{S,-}_j)$ independently of $1 - c^T_j(j)$, retaining supervision over unmatched images even when the teacher is confident in the match. Averaging over captions at temperature τ_N gives

$$\mathcal{L}_N = \frac{1}{B}\sum_j D_{\mathrm{KL}}(c^{T,-}_j\|c^{S,-}_j). \tag{17}$$

$\mathcal{L}_N$ complements $\mathcal{L}_{\mathrm{KL}}$ by constraining the constants that exact row matching leaves undetermined. Proposition 2 characterizes their joint zero set.

Proposition 2 (Joint identification by conditional matching)**.** *For $B \geq 3$, finite cosine similarity matrices, $\tau_{\mathrm{KL}}, \tau_N, \tau_G > 0$, and any $\omega > 0$,*

$$\mathcal{L}_{\mathrm{KL}} + \omega\mathcal{L}_N = 0 \iff S^S = S^T + c\mathbf{1}^\top \text{ for some } c \in \mathbb{R} \iff P = Q. \tag{18}$$

CRD's row and full-column KL share the zero set $P = Q$ (Appendix B.4). Away from it, CGD-N weights $D_{\mathrm{KL}}(c^{T,-}_j\|c^{S,-}_j)$ independently of $1 - c^T_j(j)$.

Recovering the matched margin. $\mathcal{L}_N$ excludes the matched image from each column, yet text-to-image retrieval for caption j depends on the margin $S_{jj} - S_{ij}$ over a competing image $i \neq j$. For any $k \notin \{i, j\}$ and $M \in \{T, S\}$, two row comparisons and one negative comparison in column k determine this margin (Figure 4):

$$S^M_{jj} - S^M_{ij} = (S^M_{jj} - S^M_{jk}) + (S^M_{jk} - S^M_{ik}) + (S^M_{ik} - S^M_{ij}). \tag{19}$$

Matching these comparisons matches the margin, making Proposition 2 explicit. Appendix B.4 gives the error identity. Corollary 2 bounds margin errors by feature errors.

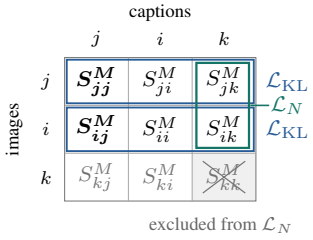


Figure 4: $\mathcal{L}_{\mathrm{KL}}$ and $\mathcal{L}_N$ jointly recover the matched margin.

Gradients with respect to cosine similarities. To compare how the losses change candidate similarities, define the descent coefficient $g^m_{ij} = -\partial \mathcal{L}_m / \partial S^S_{ij}$ for $m \in \{\mathrm{KL}, N, \mathrm{IW}\}$, before applying loss coefficients. Fixing the teacher and detaching selected pairs and importance weights gives

$$\begin{aligned} g^{\mathrm{KL}}_{ij} &= \frac{r^T_i(j) - r^S_i(j)}{B\tau_{\mathrm{KL}}}, \qquad g^N_{ij} = \frac{c^{T,-}_j(i) - c^{S,-}_j(i)}{B\tau_N} \quad (i \neq j), \\ g^{\mathrm{IW}}_{ij} &= \big[a/(\tau_{\mathrm{IW}}|H|)\big](\widetilde{W}_{ij} - 1)e^{f^S(i,j)-1} \quad ((i,j) \in H). \end{aligned} \tag{20}$$

Here $g^N_{jj} = 0$ and $g^{\mathrm{IW}}_{ij} = 0$ outside H. The sums $\sum_j g^{\mathrm{KL}}_{ij} = 0$ and $\sum_{i \neq j} g^N_{ij} = 0$ reflect invariance to adding a constant to a row or to the unmatched entries of a column, respectively. Through the shared encoders, an update can also change row and column means.

At the common temperature $\tau_{\mathrm{IW}} = \tau_w = \tau$, CRD's column coefficient is $c^S_j(i)(W_{ij} - 1)/(B\tau)$, with W_{ij} evaluated at τ. For selected pairs in one column with equal W_{ij} and nonzero corrections, their magnitudes scale as $c^S_j(i)$ for CRD but $[c^S_j(i)]^a$ for IW. With $a = 0.05$, IW allocates more of its correction to candidates with lower student probability (Appendix A).

3.3 Training Objective and Settings

The full training objective $\mathcal{L}_{\mathrm{train}}$ is $\mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{IW}}(t)\mathcal{L}_{\mathrm{IW}}$ for CGD-IW and $\mathcal{L}_{\mathrm{base}} + \lambda_N(t)\mathcal{L}_N$ for CGD-N, where t indexes training updates. A controller scales $\mathcal{L}_{\mathrm{IW}}$ or $\mathcal{L}_N$ relative to $\mathcal{L}_{\mathrm{base}}$ using student-feature gradient norms. CGD-IW activates it for the last quarter of training after using a fixed coefficient. CGD-N uses it throughout. Appendix D.1 gives the gradient measurements, controller settings, and schedules.

4 Experiments

4.1 Experimental Setup

We distill a frozen LAION-pretrained ViT-B/16 dual encoder into a student with a ViT-T/16 image encoder (5.6M parameters) and a text encoder (40.5M). Both are trained on CC3M+CC12M (Sharma et al., 2018; Changpinyo et al., 2021). Training uses 32 epochs, batch size 1,024, and AdamW (Loshchilov & Hutter, 2019). CLIP-KD (Yang et al., 2024a) and CLIP-RD (Chung et al., 2026) use the authors' released losses with the same model, data, and optimizer settings. All methods include $\mathcal{L}_{\mathrm{CLIP}} + \lambda_F \mathcal{L}_{\mathrm{FD}}$. The baseline labeled FD uses only this objective (Appendix D.1).

We average zero-shot top-1 and top-5 accuracy across ImageNet, ImageNet-V2, ImageNet-R, and ImageNet-Sketch (Deng et al., 2009; Recht et al., 2019; Hendrycks et al., 2021; Wang et al., 2019). Retrieval recall is averaged over MSCOCO, Flickr30K, and CC3M-val (Lin et al., 2014; Young et al., 2014) for image-to-text (I2T) and text-to-image (T2I) retrieval. We report means and sample standard deviations over five training runs. Bold and underlining rank student values. Ties at the displayed precision share a mark.

4.2 Main Results

Both CGD variants exceed FD in every classification and retrieval setting (Table 1). On these benchmarks, the clearest benefit over CLIP-KD/RD is retrieval: CGD-N and CGD-IW reach mean

Table 1: ViT-T/16 results after 32 epochs (%): means and sample standard deviations over five runs. Classification reports top-1. Retrieval reports R@1. Teacher results are excluded from ranking. Evaluation details are in Appendix D.1.

(a) Zero-shot classification

Method	ImageNet	ImageNet-V2	ImageNet-R	ImageNet-Sketch
FD	42.22 ± 0.36	36.48 ± 0.27	46.54 ± 0.37	25.72 ± 0.22
CLIP-KD	43.08 ± 0.30	37.11 ± 0.53	47.73 ± 0.51	26.93 ± 0.51
CLIP-RD	43.05 ± 0.19	37.06 ± 0.17	48.53 ± 0.48	**27.66** ± 0.46
CGD-N (Ours)	**43.18** ± 0.31	**37.37** ± 0.45	48.54 ± 0.18	27.35 ± 0.27
CGD-IW (Ours)	43.08 ± 0.42	37.10 ± 0.31	**48.63** ± 0.95	27.31 ± 0.57
Teacher	67.05	59.66	77.92	52.37

(b) Image–text retrieval

	MSCOCO		Flickr30K		CC3M-val	
Method	I2T	T2I	I2T	T2I	I2T	T2I
FD	40.38 ± 0.45	27.57 ± 0.13	65.28 ± 1.16	51.39 ± 0.80	33.82 ± 0.44	33.43 ± 0.21
CLIP-KD	41.98 ± 0.70	27.70 ± 0.37	67.32 ± 1.07	51.88 ± 0.73	34.08 ± 0.24	32.60 ± 0.09
CLIP-RD	42.18 ± 0.36	27.89 ± 0.29	66.84 ± 1.01	51.73 ± 0.57	34.22 ± 0.16	32.97 ± 0.30
CGD-N (Ours)	**42.61** ± 0.24	**28.15** ± 0.18	67.34 ± 0.39	**52.44** ± 0.29	34.36 ± 0.27	33.74 ± 0.26
CGD-IW (Ours)	42.32 ± 0.54	28.13 ± 0.42	**67.58** ± 1.63	52.39 ± 0.61	**34.86** ± 0.40	**33.75** ± 0.27
Teacher	55.44	38.33	83.30	65.52	41.48	39.78

R@1 of 43.11 ± 0.13 and 43.17 ± 0.48, versus 42.64 ± 0.29 for CLIP-RD, with comparable mean classification accuracy (Table 7). For text-to-image retrieval on CC3M-val, CLIP-KD and CLIP-RD trail FD across architectures. CGD stays closer to or above FD (Table 13).

	Image-to-text			Text-to-image		
Method	R@1	R@5	R@10	R@1	R@5	R@10
COCO standard						
FD	40.82	66.94	76.64	27.48	52.88	64.63
CLIP-KD	41.44	68.04	78.38	27.70	53.20	64.79
CLIP-RD	42.44	68.38	78.34	27.72	53.05	64.80
CGD-N (Ours)	42.16	67.78	77.90	28.20	**54.01**	65.58
CGD-IW (Ours)	**42.88**	**68.86**	**78.44**	**28.52**	53.76	**65.92**
CxC multi-positive						
FD	42.30	69.32	79.16	29.27	55.92	67.95
CLIP-KD	42.76	70.42	80.82	29.51	56.36	68.09
CLIP-RD	44.06	70.76	80.94	29.57	56.37	68.28
CGD-N (Ours)	43.28	70.12	80.38	29.97	**57.12**	69.01
CGD-IW (Ours)	**44.12**	**71.08**	**80.98**	**30.28**	56.90	**69.31**

Table 2: Retrieval with COCO and CxC labels (%, seed 0). Both use the same models, cosine similarities, and 25,000 captions matched to CxC. Only the labels change. Table 10 uses 25,010 captions.

	Without $\mathcal{L}_{\mathrm{KL}}$			With $\mathcal{L}_{\mathrm{KL}}$		
$\mathcal{L}_{\mathrm{IW}}$	×	✓	×	×	✓	×
$\mathcal{L}_{N}$	×	×	✓	×	×	✓
Classification top-1						
ImageNet	32.48	32.40	**32.66**	33.13	33.04	**33.28**
V2	27.58	27.54	**27.65**	28.10	28.11	**28.28**
R	36.56	36.81	**37.20**	37.25	**37.43**	37.24
Sketch	17.97	17.94	**18.16**	18.22	18.50	**18.62**
Mean retrieval recall						
R@1	33.31	33.69	**33.72**	33.62	34.07	**34.83**
R@5	58.33	**58.53**	58.41	58.89	58.64	**59.10**
R@10	68.47	68.50	**68.71**	68.92	68.86	**69.17**

Table 3: Loss components (%) after four further epochs from the common model in Section 4.4. All retain $\mathcal{L}_{\mathrm{CLIP}} + \lambda_F \mathcal{L}_{\mathrm{FD}}$. V2, R, and Sketch are ImageNet variants. Bold and underlining rank values within each group.

Additional relevant matches. To test whether the retrieval comparison depends on incomplete positive labels, Table 2 reevaluates the same cosine similarities with CxC's additional relevance annotations (Parekh et al., 2021). For text-to-image retrieval, CGD-N and CGD-IW remain ahead of CLIP-RD by 0.40 and 0.71 R@1 points, respectively. Image-to-text retrieval is more sensitive: CGD-IW's advantage narrows from 0.44 to 0.06 points, while CGD-N's deficit increases from 0.28 to 0.78. Additional valid matches thus change the I2T comparison more (Appendix D.3).

4.3 Student Architectures

We extend CGD to CNN and other students using the ViT-T/16 training and loss hyperparameters, without further search or intermediate-layer matching (Table 4). Both variants improve mean top-1 and R@1 over FD and CLIP-KD on all four architectures. Both exceed CLIP-RD on

Table 4: Results for four student architectures at epoch 32 (%, seed 0). Top-1 accuracy averages the four classification datasets. R@1 averages the six retrieval settings. Parameter counts include both encoders and projections (Table 11). Per-dataset results and other recall cutoffs: Appendix E.2.

	ResNet-18 51.93M (He et al., 2016)		EfficientNet-B0 45.16M (Tan & Le, 2019)		MobileViT-S 45.76M (Mehta & Rastegari, 2022)		Swin-Tiny 68.41M (Liu et al., 2021)	
Method	Top-1	R@1	Top-1	R@1	Top-1	R@1	Top-1	R@1
FD	34.97	37.71	42.64	44.72	42.37	45.64	53.43	52.90
CLIP-KD	35.71	38.12	43.07	45.02	43.04	46.38	53.43	52.94
CLIP-RD	35.84	38.75	43.48	45.24	**43.41**	**47.14**	54.24	53.81
CGD-N	**35.94**	38.25	43.60	**45.59**	43.20	47.02	54.29	53.85
CGD-IW	35.74	**38.76**	**43.70**	45.38	43.22	46.82	**54.38**	**53.97**

EfficientNet-B0 and Swin-Tiny and are comparable on ResNet-18 and MobileViT-S, within 0.50 points (Appendix E.2).

4.4 Loss Components and Training Dynamics

For Table 3, we train with $\mathcal{L}_{\text{CLIP}} + \lambda_F \mathcal{L}_{\text{FD}}$ for 16 epochs, then resume the same student and optimizer state under each objective for four epochs, using identical training examples in the same order. Adding $\lambda_{\text{IW}}\mathcal{L}_{\text{IW}}$ or $\lambda_N \mathcal{L}_N$ to $\mathcal{L}_{\text{base}}$ improves mean R@1. The addition of $\mathcal{L}_N$ also improves R@5 and R@10, whereas $\mathcal{L}_{\text{IW}}$ does not (Appendices D.2 and E.3). Here IW and N use the same gradient-norm target. The 32-epoch runs use different coefficient schedules (Section 3.3).

How do competing candidates change during training? We resume the student and optimizer state used in Table 3, saved after 16 epochs of FD training. Using the training hyperparameters of Table 1, we continue FD, $\mathcal{L}_{\text{base}}$, CLIP-KD, CLIP-RD, CGD-IW, and CGD-N for 2,000 updates with five seeds (Figure 5). Each seed uses the same data order across objectives. For COCO queries the teacher answers correctly but the student initially misses, (a,b) track similarity with the initially most similar positive and negative. Panel (c) reselects these candidates, and (d) measures negative-column KL (Appendix E.6).

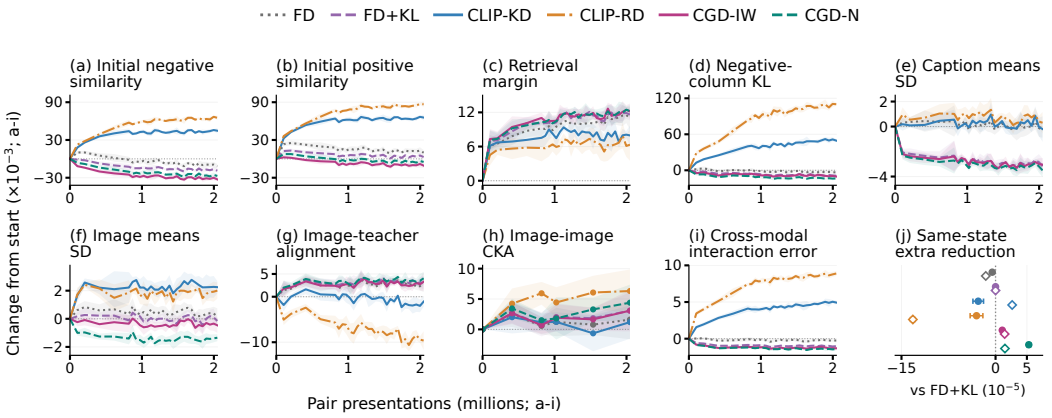


Figure 5: **Training trajectories and updates from the same states.** (a–i): mean changes ± 1 sample SD over five continuations from the FD model at epoch 16. FD+KL denotes $\mathcal{L}_{\text{base}}$. (e,f) average $S^S_{ij} - S^T_{ij}$ over images and captions, respectively, before taking the SD. (j): extra reductions in (f) (circles) and (i) (diamonds) versus FD+KL, from identical saved states, in legend order from top (Appendix E.10).

CGD lowers cosine similarity with the negative more than with the positive (a,b), widening their margin through the competitor correction illustrated in Figure 1. CLIP-KD and CLIP-RD increase both similarities. CGD has larger retrieval margin gains than CLIP-KD/RD (c), while its gain over $\mathcal{L}_{\text{base}}$ is within seed variation (Appendix E.6). CGD-N has lower negative-column KL than $\mathcal{L}_{\text{base}}$ in each continuation (d), showing that $\mathcal{L}_N$ also improves distribution matching on COCO.

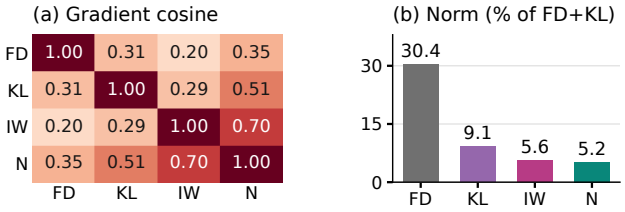


Figure 6: Encoder gradients at shared FD states. (a) Mean cosine. (b) Gradient norm relative to $\|\nabla\mathcal{L}_{\text{base}}\|_2$. We average batches and seeds, then the three update counts (Appendix E.9).

How do the two matching directions complement each other? We decompose $D_{ij} = S^S_{ij} - S^T_{ij}$ as $\mu + a_i + b_j + E_{ij}$, where μ is its overall mean. For image i, $a_i = \text{mean}_j\, D_{ij} - \mu$ is its mean similarity gap across captions minus μ. For caption j, $b_j = \text{mean}_i\, D_{ij} - \mu$ averages over images instead. The residual E_{ij} contains the remaining pairwise differences (Appendix E.7). For caption j, positive p, and competitor i, the margin error is $(a_p - a_i) + (E_{pj} - E_{ij})$. The term b_j is shared by both images and cancels. Adding $\mathcal{L}_{\text{KL}}$ to FD reduces $\text{SD}_j(b_j)$ (Figure 5e), while adding $\mathcal{L}_N$ to $\mathcal{L}_{\text{base}}$ reduces $\text{SD}_i(a_i)$ (f), in all five continuations. The latter reduces differences in a_i across images, which row matching leaves unconstrained (Proposition 2). CGD-IW's reduction is smaller and less consistent.

CGD retains image alignment with the teacher (g) while reducing interaction error $\sqrt{\text{mean}_{ij}\, E^2_{ij}}$ (i). Together, (f,i) show reductions in both sources of T2I margin error. CLIP-RD combines higher image CKA (h) with larger interaction error. At the same states, CGD reduces $\text{SD}_i(a_i)$ and interaction error more than $\mathcal{L}_{\text{base}}$ (j). CLIP-KD has a larger immediate reduction in interaction error, while that error increases along its own continuation (i).

At the same FD states, gradients of $\mathcal{L}_{\text{IW}}$ and $\mathcal{L}_N$ align more with each other than with $\mathcal{L}_{\text{FD}}$ (Figure 6a). Their weighted norms are small relative to $\|\nabla\mathcal{L}_{\text{base}}\|_2$ (b), while CLIP-KD/RD's combined auxiliary gradients are comparable to it (Table 22). These small corrections accompany maintained alignment and lower interaction error (Figure 5g,i). Derivatives of fixed T2I margins are positive initially and slightly negative later (Appendix E.9).

4.5 Representation Analysis

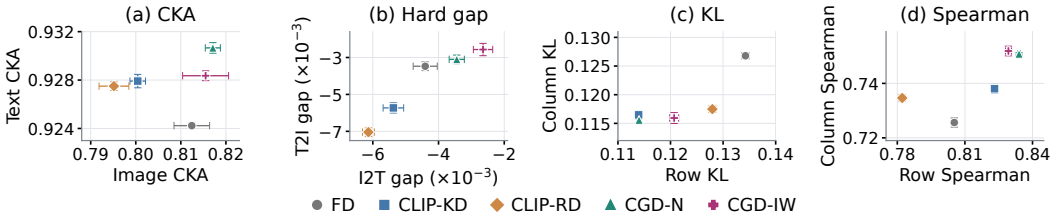


Figure 7: Representation diagnostics after 32 epochs on three CC3M-val folds: five-run means and ± 1 sample SD on both axes (Appendix D.4).

Feature alignment alone does not determine retrieval: FD has higher image CKA than CLIP-KD/RD but lower R@1 (Figure 7a). CGD has larger hard gaps, the matched similarity minus the largest unmatched similarity (b), despite lower matched similarities than CLIP-RD. Higher Spearman correlations (d) also show closer agreement with the teacher's ordering across all candidates. CGD has lower top-64 overlap, and CGD-IW has higher row KL than CLIP-KD (c, Table 15). Retrieval depends on the matched margin. Top-64 overlap instead measures which unmatched candidates the two models rank highest.

5 Conclusion

CGD combines feature and row matching with importance weights (CGD-IW) or negative-column KL (CGD-N). Both improve five-run mean retrieval over CLIP-KD/RD with comparable classification. In controlled continuations, CGD reduces similarity with incorrect candidates more than with positives, whereas CLIP-KD and CLIP-RD raise both. The evidence covers one teacher and training corpus, with mixed results against CLIP-RD across student architectures.

Reproducibility statement. Appendices D–E document training, evaluation, per-seed results, student architectures, loss components, and representation metrics. Training and evaluation code will be released.

AI use statement. Generative AI assisted with writing, LaTeX, mathematical formulation, control discussions, implementation, figure preparation, and result interpretation. It was not used to generate benchmark images, captions, or relevance labels. Experimental values were obtained from trained models and checked against stored evaluation outputs. The authors retain responsibility for all methods, claims, and results.

REFERENCES

Youva Addad, Alexis Lechervy, and Frédéric Jurie. Intra-modal divergence-weighted distillation for vision–language models. In *British Machine Vision Conference*, 2025.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

Jeannie Chung, Hanna Jang, Ingyeong Yang, Uiwon Hwang, and Jaehyeong Sim. CLIP-RD: Relative distillation for efficient CLIP knowledge distillation. *arXiv preprint arXiv:2603.25383*, 2026.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

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. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Sachin Mehta and Mohammad Rastegari. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=vh-0sUt8HlG.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010. doi: 10.1109/TIT.2010.2068870.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. URL https://arxiv.org/abs/2410.21276.

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021.

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

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5171–5180, 2019.

Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision–language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6967–6977, 2023.

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

Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. AM-RADIO: Agglomerative vision foundation model – reduce all domains into one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5389–5400, 2019.

Yashas Malur Saidutta, Rakshith S. Srinivasa, Jaejin Cho, Ching-Hua Lee, Chouchang Yang, Yilin Shen, and Hongxia Jin. CIFD: Controlled information flow to enhance knowledge distillation. In *Advances in Neural Information Processing Systems*, 2024.

Jinghuan Shang, Karl Schmeckpeper, Brandon B. May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. In *Conference on Robot Learning*, 2024.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2556–2565, 2018.

Ximeng Sun, Pengchuan Zhang, Peizhao Zhang, Hardik Shah, Kate Saenko, and Xide Xia. DIME-FM: Distilling multimodal and efficient foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15521–15533, 2023.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 6105–6114, 2019. URL `https://proceedings.mlr.press/v97/tan19a.html`.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019.

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

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

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Chen, Xinggang Wang, Hongyang Chao, and Han Hu. TinyCLIP: CLIP distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21970–21980, 2023.

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*, pp. 15952–15962, 2024a.

Kaicheng Yang, Tiancheng Gu, Xiang An, Haiqiang Jiang, Xiangzi Dai, Ziyong Feng, Weidong Cai, and Jiankang Deng. CLIP-CID: Efficient CLIP distillation via cluster-instance discrimination. *arXiv preprint arXiv:2408.09441*, 2024b.

Wenzhe Yin, Zehao Xiao, Pan Zhou, Shujian Yu, Jiayi Shen, Jan-Jakob Sonke, and Efstratios Gavves. Distributional vision language alignment by Cauchy–Schwarz divergence. In *International Conference on Learning Representations*, 2026.

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.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11953–11962, 2022.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229, pp. 2165–2183, 2023. URL `https://proceedings.mlr.press/v229/zitkovich23a.html`.

A Related Work and Objective Comparison

Distillation for CLIP dual encoders. Smaller CLIP dual encoders are obtained through architecture design, training data, or knowledge transfer. DIME-FM distills from unpaired data (Sun et al., 2023), TinyCLIP combines affinity mimicking with weight inheritance (Wu et al., 2023), and MobileCLIP uses reinforced data (Vasu et al., 2024). CLIP-CID adds cluster–instance discrimination (Yang et al., 2024b). CLIP-KD studies feature and relational distillation (Yang et al., 2024a). CIFD addresses the capacity gap with rate–distortion modules (Saidutta et al., 2024), while intra-modal divergence weighting reweights examples by teacher–student disagreement (Addad et al., 2025). AM-RADIO combines several vision foundation models into one student (Ranzinger et al., 2024), and Theia distills them for robot learning (Shang et al., 2024).

Relational knowledge distillation. Beyond softened predictions (Hinton et al., 2015), relational methods preserve distances, angles, similarities, or contrastive relations (Park et al., 2019; Tung & Mori, 2019; Tian et al., 2020). In image–text pretraining, affinity mimicking couples positive and negative pairs (Wu et al., 2023), hard-negative mining sharpens the contrastive signal (Radenovic et al., 2023), and CLIP-RD adds vertical and cross relational objectives that enforce consistency between distributions (Chung et al., 2026). DKD separates target and non-target knowledge and independently weights their contributions (Zhao et al., 2022). CGD-N applies DKD's non-target KL $D_{\mathrm{KL}}(c_j^{T,-} \| c_j^{S,-})$ to image candidates for each caption j, matching $c_j^{T,-}$ and $c_j^{S,-}$.

Mutual information and distribution alignment. CS-Aligner (Yin et al., 2026) combines mutual information with a Cauchy–Schwarz divergence between image and text feature distributions. It compares the two modality distributions. CGD compares teacher and student joints over the same batch indices. Equal MI permits different joints and matched decisions (Appendix B.1).

Variational bounds on mutual information. Contrastive learning is commonly analyzed through variational MI bounds (van den Oord et al., 2018; Poole et al., 2019). The NWJ functional $\mathcal{J}_Q(f)$ lower-bounds $I_Q = D_{\mathrm{KL}}(Q \| Q_U \otimes Q_V)$ and is tight at $f_Q^* = 1 + \log[Q/(Q_U Q_V)]$ (Nguyen et al., 2010). We use the NWJ representation to decompose Δ_{info} under a shared critic.

Table 5: Direct supervision signals with teacher quantities fixed. Coefficients are negative gradients with respect to the corresponding cosine similarities, at a common temperature and before loss coefficients. FD acts in feature space. ICL uses a cross-model cosine similarity. $\Delta r_{ij} = r_i^T(j) - r_i^S(j)$ and $\pi_k(j)$ is the softmax over similarities between student image k and teacher texts.

Term	Directly supervises	Descent coefficient
FD	image/text features	no independent coefficient on S^S
ICL	student–teacher contrast	$(\delta_{kj} - \pi_k(j))/(2B\tau)$
Row KL (CRD; $\mathcal{L}_{\mathrm{KL}}$)	row distributions r_i^M	$\Delta r_{ij}/(B\tau)$
Column KL (CRD)	column distributions c_j^M	$c_j^S(i)(W_{ij} - 1)/(B\tau)$
$\mathcal{L}_N$ (CGD-N)	unmatched images: $c_j^{M,-}$	$(c_j^{T,-}(i) - c_j^{S,-}(i))/(B\tau_N)$, $i \neq j$
$\mathcal{L}_{\mathrm{IW}}$ (CGD-IW)	selected pairs: $\widetilde{W}_{ij}$	$\frac{a}{\tau_{\mathrm{IW}}\lvert H\rvert}(\widetilde{W}_{ij} - 1)e^{f^S(i,j)-1}$ on H

Objective comparison. The equations below use a common temperature and unit term weights to compare gradient structure, with u for image and v for text embeddings. The released CRD averages its two directions. The scales used in the implementation are listed in Appendix D.1. Each coefficient is the negative gradient with respect to its cosine similarity, with teacher quantities fixed. CLIP-KD (Yang et al., 2024a) trains the task loss with feature distillation, interactive contrastive learning (ICL), and contrastive relational distillation (CRD).

The KL divergence between column distributions c_j^T, c_j^S and its negative gradient are

$$\mathcal{L}_{\mathrm{col}} = \frac{1}{B}\sum_j D_{\mathrm{KL}}(c_j^T \| c_j^S), \qquad -\frac{\partial \mathcal{L}_{\mathrm{col}}}{\partial S_{ij}^S} = \frac{c_j^S(i)(W_{ij} - 1)}{B\tau}. \tag{21}$$

Here $W_{ij} = c_j^T(i)/c_j^S(i)$ is evaluated at the same temperature. For example, probabilities 2×10^{-4} and 10^{-4} have a small absolute difference but a ratio of two.

CRD matches teacher and student row distributions r_i^T, r_i^S and column distributions c_j^T, c_j^S:

$$\begin{aligned} \mathcal{L}_{\mathrm{CRD}} &= \frac{1}{B}\sum_i D_{\mathrm{KL}}(r_i^T \| r_i^S) + \frac{1}{B}\sum_j D_{\mathrm{KL}}(c_j^T \| c_j^S), \\ -\frac{\partial \mathcal{L}_{\mathrm{CRD}}}{\partial S_{ij}^S} &= \frac{\Delta r_{ij} + \Delta c_{ij}}{B\tau}. \end{aligned} \tag{22}$$

with $\Delta r_{ij} = r_i^T(j) - r_i^S(j)$ and $\Delta c_{ij} = c_j^T(i) - c_j^S(i)$, so its descent coefficient for S_{ij}^S sums the differences in $r_i^M(j)$ and $c_j^M(i)$. Its column loss $\mathcal{L}_{\mathrm{col}}$ in Equation (21) is the only CLIP-KD term that compares the teacher's and the student's probability of the same pair along the column axis, where images compete for a caption. ICL is an InfoNCE loss between a student embedding and the teacher embeddings of the other modality,

$$\begin{aligned} \mathcal{L}_{\mathrm{ICL}} &= -\frac{1}{2B}\sum_k \left[\log \frac{e^{\langle u_k^S, v_k^T\rangle/\tau}}{\sum_b e^{\langle u_k^S, v_b^T\rangle/\tau}} + \log \frac{e^{\langle v_k^S, u_k^T\rangle/\tau}}{\sum_b e^{\langle v_k^S, u_b^T\rangle/\tau}} \right], \\ -\frac{\partial \mathcal{L}_{\mathrm{ICL}}}{\partial \langle u_k^S, v_j^T\rangle} &= \frac{\delta_{kj} - \pi_k(j)}{2B\tau}. \end{aligned} \tag{23}$$

where $\pi_k(j)$ is the softmax probability of teacher text j given student image k. For each unmatched pair, ICL's direct descent coefficient is negative and proportional to $\pi_k(j)$, regardless of the probability assigned by the teacher. CLIP-RD's $\mathcal{L}_{\mathrm{VRD}}$ and $\mathcal{L}_{\mathrm{XRD}}$ (Chung et al., 2026) compare distributions formed from teacher–student similarities across modalities or directions. Their target probabilities are held fixed during each update but depend on the student through the cross-model similarities (Appendix A.1). The authors' released code departs from the preprint's equations in the direction of the KL terms, the treatment of the target distributions, and the handling of the relational scales. Our implementation follows the released code, and Appendix D.1 lists the differences.

The released CLIP-RD objective also includes CRD. Thus both baseline recipes contain $\mathcal{L}_{\mathrm{CRD}}$, whose exact zero at a common temperature requires $P = Q$. CGD changes which column pairs are compared and how their corrections are weighted. Proposition 2 characterizes when $\mathcal{L}_{\mathrm{KL}} + \omega\mathcal{L}_N$ vanishes.

How student probabilities weight the corrections. Fix a caption j and write $p_i = c_j^T(i)$, $q_i = c_j^S(i) = \mathrm{softmax}_i(z)$, and $z_i = S_{ij}^S/\tau$, holding p fixed. For a differentiable convex generator φ with $\varphi(1) = 0$, define

$$D_\varphi(p\|q) = \sum_i q_i \varphi(W_i), \qquad W_i = \frac{p_i}{q_i}, \qquad h_\varphi(w) = \varphi(w) - w\varphi'(w).$$

Since $\partial D_\varphi/\partial q_i = h_\varphi(W_i)$ and $\partial q_k/\partial z_i = q_k(\delta_{ki} - q_i)$, the chain rule gives

$$-\frac{\partial D_\varphi}{\partial S_{ij}^S} = -\frac{q_i}{\tau}\left[h_\varphi(W_i) - \sum_k q_k h_\varphi(W_k) \right].$$

Thus two entries in the same column with equal W_i have equal $h_\varphi(W_i) - \sum_k q_k h_\varphi(W_k)$. If their corrections are nonzero, the ratio of their magnitudes is q_i/q_k. For KL, $\varphi(w) = w\log w$ gives $h_\varphi(w) = -w$, recovering $(p_i - q_i)/\tau$.

For IW, set $\tau_{\mathrm{IW}} = \tau_w = \tau$ and let $Z_j^S = \sum_\ell \exp(S_{\ell j}^S/\tau)$. The affine critic gives

$$e^{f^S(i,j)-1} = e^{b-1}(Z_j^S)^a q_i^a.$$

The factor $e^{b-1}(Z_j^S)^a$ is common within column j. For $(i,j),(k,j) \in H$ with $W_i = W_k$, the normalized ratios are also equal. When both corrections are nonzero, Equation (20) therefore gives

$$\frac{|g_{ij}^{\mathrm{IW}}|}{|g_{kj}^{\mathrm{IW}}|} = \left(\frac{q_i}{q_k}\right)^a.$$

For example, $q_i/q_k = 100$ gives a correction ratio of 100 for CRD but only $100^{0.05} \simeq 1.26$ for IW under these conditions. IW therefore assigns a larger share of its correction to candidates with lower student probability. This comparison holds W_i equal across candidates. When W_i varies, both q_i and $h_\varphi(W_i) - \sum_k q_k h_\varphi(W_k)$ determine the correction from D_φ. IW additionally differs from column KL in its selected pairs H and normalization by Z_H.

A.1 Gradients of the CLIP-RD Components

Equation (6) defines $\mathcal{L}_{\mathrm{CRD}}$ as the sum of row and column KL shared by CLIP-KD and CLIP-RD, where it is called Horizontal Relational Distillation (HRD). We analyze $\mathcal{L}_{\mathrm{VRD}}$ and $\mathcal{L}_{\mathrm{XRD}}$ using the released implementation described in Appendix D.1. Teacher embeddings and the target of each KL term are held fixed during differentiation.

Cross-entropy and KL derivatives. For scaled logits z, let $p_i(k) = \mathrm{softmax}_k(z_{ik})$ be the row probability and $q_i(k)$ its target. With matched index i,

$$-\frac{\partial\,\mathrm{CE}(p_i, i)}{\partial z_{ik}} = \delta_{ik} - p_i(k), \qquad -\frac{\partial\, D_{\mathrm{KL}}(\mathrm{sg}[q_i] \| p_i)}{\partial z_{ik}} = q_i(k) - p_i(k). \tag{24}$$

Here $\mathrm{CE}(p_i, i) = -\log p_i(i)$. The following expressions include batch averaging and the released component weights. To obtain derivatives with respect to cosine similarities, multiply by the corresponding auxiliary scale.

VRD. Let U_T, V_T, U_S, V_S contain the unit-normalized image and text embeddings as rows, after any student projection. At the current auxiliary scales, the image and text logits are

$$A_{\mathrm{img}} = s_{\mathrm{img}} U_T U_S^\top, \qquad A_{\mathrm{txt}} = s_{\mathrm{txt}} V_T V_S^\top.$$

Write p^r_{ij}, p^c_{ij} for the row and column softmax probabilities of A_{img}, and q^r_{ij}, q^c_{ij} for those of A_{txt}. VRD applies cross-entropy with the matched index as its target in both query directions for each modality, and matches the image probabilities to the detached text probabilities. Each of these four cross-entropy terms and two KL terms has coefficient $1/2$. Including both query directions gives

$$\begin{aligned} -\frac{\partial \mathcal{L}_{\mathrm{VRD}}}{\partial (A_{\mathrm{img}})_{ij}} &= \frac{2\delta_{ij} + q^r_{ij} + q^c_{ij} - 2p^r_{ij} - 2p^c_{ij}}{2B}, \\ -\frac{\partial \mathcal{L}_{\mathrm{VRD}}}{\partial (A_{\mathrm{txt}})_{ij}} &= \frac{2\delta_{ij} - q^r_{ij} - q^c_{ij}}{2B}. \end{aligned} \tag{25}$$

For an unmatched entry in one image-side direction, cross-entropy contributes $-p/(2B)$ and KL contributes $(q-p)/(2B)$. Their sum is $(q-2p)/(2B)$: it raises the scaled logit when $q > 2p$. The text-side KL target is detached, so its text gradient comes from cross-entropy.

XRD. The cross-modal teacher–student logits are

$$A_{\mathrm{it}} = s_{\mathrm{x}} U_T V_S^\top, \qquad A_{\mathrm{ti}} = s_{\mathrm{x}} V_T U_S^\top.$$

Write e^r_{ij}, e^c_{ij} and f^r_{ij}, f^c_{ij} for their respective row and column softmax probabilities. XRD averages four directed KL terms, each with a detached target. Thus

$$\begin{aligned} -\frac{\partial \mathcal{L}_{\mathrm{XRD}}}{\partial (A_{\mathrm{it}})_{ij}} &= \frac{f^r_{ij} + f^c_{ij} - e^r_{ij} - e^c_{ij}}{4B}, \\ -\frac{\partial \mathcal{L}_{\mathrm{XRD}}}{\partial (A_{\mathrm{ti}})_{ij}} &= \frac{e^r_{ij} + e^c_{ij} - f^r_{ij} - f^c_{ij}}{4B}. \end{aligned} \tag{26}$$

Each directed term contributes $(q-p)/(4B)$ and can raise an unmatched scaled logit when its target probability exceeds its prediction. The reverse directed terms update the other student encoder.

Comparison with CGD. The KL corrections in CRD, VRD, and XRD depend on absolute probability differences. When both compared probabilities are small, a correction can be small even if their ratio differs substantially from one. For VRD and XRD, these probabilities come from

cross-model similarities, not the student column distribution c_j^S. Their feature gradients also change multiple entries of $S^S = U_S V_S^\top$ through both student encoders. A small $c_j^S(i)$ alone therefore does not determine the combined CLIP-RD update to S_{ij}^S.

CGD-N compares the distributions $c_j^{T,-}, c_j^{S,-}$ over unmatched images and weights $D_{\mathrm{KL}}(c_j^{T,-} \| c_j^{S,-})$ independently of $c_j^T(j)$. CGD-IW instead uses normalized ratios $\widetilde{W}_{ij}$ on the selected pairs H. Its coefficient in Equation (20) scales as $[c_j^S(i)]^a$ for selected pairs in one column with equal W_{ij} and common temperatures, compared with $c_j^S(i)$ for CRD (Appendix A). Table 5 compares the CRD and CGD corrections.

B PROOFS

For the proofs, abbreviate the ratios of each joint to its marginal product as $\kappa_P = P/(P_U P_V)$ and $\kappa_Q = Q/(Q_U Q_V)$.

InfoNCE and distributions over candidates. Under i.i.d. sampling of paired data with marginal negatives, $\log B - \mathbb{E}_{\mathcal{B}}[\mathcal{L}_{\mathrm{CLIP}}]$ lower-bounds the MI between the encoded image and text variables (Poole et al., 2019). The finite joints P, Q describe batch indices and recover r_i^M, c_j^M in $\mathcal{L}_{\mathrm{CLIP}}$ at a common temperature. Their MI values I_P, I_Q differ from the encoded-variable MI in this bound.

B.1 EQUAL MUTUAL INFORMATION DOES NOT IDENTIFY THE JOINT

Consider strictly positive joint distributions over three fixed image and caption indices, with the diagonal denoting matched pairs:

$$P = \frac{1}{15}\begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix}, \qquad Q = \frac{1}{15}\begin{pmatrix} 1 & 3 & 1 \\ 1 & 1 & 3 \\ 3 & 1 & 1 \end{pmatrix}.$$

The second joint permutes the columns of the first without changing the meaning of the caption indices. Both have uniform marginals and

$$I_P = I_Q = \frac{3}{5}\log\frac{9}{5} + \frac{2}{5}\log\frac{3}{5}, \qquad D_{\mathrm{KL}}(P\|Q) = \frac{2}{5}\log 3 > 0.$$

Every teacher row ranks the matched caption first, whereas every student row ranks a different caption first. Equal MI therefore permits complete disagreement on the matched decisions. These joint distributions are compatible with the setup with normalized embeddings: take teacher image and text embeddings as the three coordinate basis vectors, permute the student text vectors, and set $\tau_G = 1/\log 3$.

Because the marginals agree, Equation (10) gives $\Delta_{\mathrm{dist}} = D_{\mathrm{KL}}(P\|Q) > 0$ and $\Delta_{\mathrm{info}} = -\Delta_{\mathrm{dist}}$ in this example. $I_P - I_Q = 0$ although the joint distributions and matched decisions differ. Proposition 2 instead concerns the joint zero set of $\mathcal{L}_{\mathrm{KL}}$ and $\mathcal{L}_N$, which identifies $P = Q$.

B.2 FEATURE-ERROR BOUNDS

Proposition 3 (FD bounds $|\mathcal{J}_+(f^S)|$)**.** *For unit-normalized embeddings, the joint distributions P, Q in Equation* (2)*, and $f^S(i,j) = aS_{ij}^S/\tau_{\mathrm{IW}} + b$, $|\mathcal{J}_+(f^S)| \le |a|\sqrt{2BD\mathcal{L}_{\mathrm{FD}}}/(\tau_{\mathrm{IW}}\tau_G)$.*

Corollary 1 (FD bounds $|\Delta_{\mathrm{info}}|$)**.** *With the same unit-normalized embeddings and strictly positive joint probabilities, $|\Delta_{\mathrm{info}}| \le \|P - Q\|_1 \|\log \kappa_Q\|_\infty \le 4\sqrt{2BD\mathcal{L}_{\mathrm{FD}}}/\tau_G^2$. Consequently,*

$$I_P - I_Q \le \Delta_{\mathrm{dist}} + \frac{4}{\tau_G^2}\sqrt{2BD\mathcal{L}_{\mathrm{FD}}}. \tag{27}$$

Appendix B.6 proves both statements.

B.3 Distribution and Information Gaps

Recall I_Q^{IW} from Equation (8) and $\Delta_{\mathrm{info}} = I_Q^{\mathrm{IW}} - I_Q$.

Proof of Proposition 1. Using the conditional factorization of κ_P and κ_Q,

$$\begin{aligned} I_P - I_Q^{\mathrm{IW}} &= \mathbb{E}_P\left[\log \frac{P(V \mid U)}{Q(V \mid U)}\right] + \mathbb{E}_P\left[\log \frac{Q_V(V)}{P_V(V)}\right] \\ &= \Delta_{\mathrm{dist}} - D_{\mathrm{KL}}(P_V \| Q_V). \end{aligned} \tag{28}$$

Adding $I_Q^{\mathrm{IW}} - I_Q$ proves the identity in Equation (10). Since $D_{\mathrm{KL}}(P_V \| Q_V) \geq 0$, dropping $-D_{\mathrm{KL}}(P_V \| Q_V)$ gives the upper bound. $D_{\mathrm{KL}}(P_V \| Q_V) = 0$ exactly when $P_V = Q_V$. Change of measure also gives

$$\Delta_{\mathrm{info}} = I_Q^{\mathrm{IW}} - I_Q = \mathbb{E}_Q[(\rho - 1) \log \kappa_Q]. \tag{29}$$

□

By the chain rule, $\Delta_{\mathrm{dist}} - D_{\mathrm{KL}}(P_V \| Q_V) = D_{\mathrm{KL}}(P \| Q) - D_{\mathrm{KL}}(P_U \| Q_U) - D_{\mathrm{KL}}(P_V \| Q_V)$. The distribution term Δ_{dist} is nonnegative, whereas the information term Δ_{info} can have either sign.

B.4 Column Matching and Joint Identification

Column expression. The two KL chain rules give

$$D_{\mathrm{KL}}(P \| Q) = D_{\mathrm{KL}}(P_U \| Q_U) + \Delta_{\mathrm{dist}} = D_{\mathrm{KL}}(P_V \| Q_V) + \mathbb{E}_{j \sim P_V}[D_{\mathrm{KL}}(c_j^T \| c_j^S)].$$

Subtracting $D_{\mathrm{KL}}(P_U \| Q_U)$ and $D_{\mathrm{KL}}(P_V \| Q_V)$ proves Equation (16).

Target/non-target split. For $i \neq j$, write $c_j^M(i) = (1 - c_j^M(j)) c_j^{M,-}(i)$ for either model M. Separating $c_j^T(j) \log[c_j^T(j)/c_j^S(j)]$ in $D_{\mathrm{KL}}(c_j^T \| c_j^S)$ gives

$$\begin{aligned} D_{\mathrm{KL}}(c_j^T \| c_j^S) = c_j^T(j) \log \frac{c_j^T(j)}{c_j^S(j)} + (1 - c_j^T(j)) \log \frac{1 - c_j^T(j)}{1 - c_j^S(j)} \\ + (1 - c_j^T(j)) \sum_{i \neq j} c_j^{T,-}(i) \log \frac{c_j^{T,-}(i)}{c_j^{S,-}(i)}. \end{aligned} \tag{30}$$

The first line equals $D_{\mathrm{KL}}(\mathrm{Ber}(c_j^T(j)) \| \mathrm{Ber}(c_j^S(j)))$, where $\mathrm{Ber}(p)$ is the Bernoulli distribution with parameter p. This is the established DKD decomposition (Zhao et al., 2022), applied to image candidates for a caption. In the expected column KL of Equation (16), the KL divergence $D_{\mathrm{KL}}(c_j^{T,-} \| c_j^{S,-})$ has coefficient $P_V(j)(1 - c_j^T(j))$. $\mathcal{L}_N$ averages this KL divergence over captions with an independent coefficient λ_N. Differentiating $\mathcal{L}_N$ gives Equation (20). The positive diagonal has zero direct gradient from $\mathcal{L}_N$. Shared encoder parameters can nevertheless change the cosine similarities of matched pairs during an update.

Proof of Proposition 2. Since $\mathcal{L}_{\mathrm{KL}}, \mathcal{L}_N \geq 0$ and $\omega > 0$, $\mathcal{L}_{\mathrm{KL}} + \omega \mathcal{L}_N$ vanishes precisely when every $D_{\mathrm{KL}}(r_i^T \| r_i^S)$ and $D_{\mathrm{KL}}(c_j^{T,-} \| c_j^{S,-})$ vanishes. Equality of two positive softmax distributions is equivalent to their logits differing by a constant. Each loss uses the same temperature for teacher and student. The temperatures may differ between losses. Thus $\mathcal{L}_{\mathrm{KL}} = 0$ gives $S_{ij}^S - S_{ij}^T = a_i$ for every i, j, while $\mathcal{L}_N = 0$ gives $S_{ij}^S - S_{ij}^T = b_j$ for every $i \neq j$. For any two rows r, s, choose $j \notin \{r, s\}$, possible because $B \geq 3$. Then $a_r = b_j = a_s$. Hence all a_i equal one constant c, including their diagonal entries, and $S^S = S^T + c\mathbf{1}\mathbf{1}^\top$. Conversely, adding the same constant to all cosine similarities preserves row distributions r_i^M and column distributions c_j^M, so both losses vanish. The joint Gibbs softmax has the same invariance, and $P = Q$ implies a constant difference between its logits. This proves both equivalences. □

Scope of the characterization. For $B = 2$, each negative column contains only one image and $\mathcal{L}_N = 0$ regardless of the cosine similarities: different constants added to the two rows remain unidentified.

Matched comparisons and finite errors. For $i \neq j$, choose $k \notin \{i, j\}$. Equation (19) follows by cancellation of S^M_{jk} and S^M_{ik}. At their respective temperatures, its three terms equal

$$\tau_{\mathrm{KL}} \log \frac{r^M_j(j)}{r^M_j(k)}, \qquad \tau_N \log \frac{c^{M,-}_k(j)}{c^{M,-}_k(i)}, \qquad \tau_{\mathrm{KL}} \log \frac{r^M_i(k)}{r^M_i(j)}.$$

The ratio $c^{M,-}_k(j)/c^{M,-}_k(i)$ is defined because both images are unmatched in column k. Thus row and negative column matching also determine the matched comparison in column j. This argument does not require equal temperatures between losses. For $D_{ab} = S^S_{ab} - S^T_{ab}$ and $m^M_{ij} = S^M_{jj} - S^M_{ij}$, subtracting the two identities gives

$$m^S_{ij} - m^T_{ij} = (D_{jj} - D_{jk}) + (D_{jk} - D_{ik}) + (D_{ik} - D_{ij}). \tag{31}$$

Consequently, $|m^S_{ij} - m^T_{ij}| \leq |D_{jj} - D_{jk}| + |D_{jk} - D_{ik}| + |D_{ik} - D_{ij}|$. If this upper bound is smaller than a positive teacher margin m^T_{ij}, the student retains the teacher's preference for the matched image over image i. This condition concerns the individual comparison errors, rather than the batch-averaged KL values.

B.5 Two NWJ Functionals for the Information Gap

The full joint distributions P and Q are defined in Equation (2). $Q_U \otimes Q_V$ is the product of the student's image and caption marginals. Recall the density ratio $\rho(i, j) = P(i, j)/Q(i, j)$ and the proof notation $\kappa_Q(i, j) = Q(i, j)/(Q_U(i)Q_V(j))$. The NWJ functional and its importance-weighted counterpart are

$$\begin{aligned} \mathcal{J}_Q(f) &= \mathbb{E}_Q[f(i, j)] - \mathbb{E}_{Q_U \otimes Q_V}[e^{f(i,j)-1}], \\ \mathcal{J}^{\mathrm{IW}}_Q(f) &= \mathbb{E}_P[f(i, j)] - \mathbb{E}_{Q_U \otimes Q_V}[\rho(i, j)e^{f(i,j)-1}]. \end{aligned} \tag{32}$$

Their variational representations share the same maximizer:

$$I_Q = \sup_f \mathcal{J}_Q(f), \qquad I^{\mathrm{IW}}_Q = \sup_f \mathcal{J}^{\mathrm{IW}}_Q(f), \qquad f^*_Q(i, j) = 1 + \log \frac{Q(i, j)}{Q_U(i)Q_V(j)}. \tag{33}$$

We derive both bounds below. Expectations under P or Q sample (i, j) from the corresponding joint distribution. $Q_U \otimes Q_V$ samples the two indices independently.

Proposition 4 (Variational bounds and joint–product decomposition)**.** *For every real-valued critic f on the batch pairs, $\mathcal{J}_Q(f) \leq I_Q$ and $\mathcal{J}^{\mathrm{IW}}_Q(f) \leq I^{\mathrm{IW}}_Q$. Both bounds are tight at $f^*_Q(i, j) = 1 + \log \kappa_Q(i, j)$, which proves Equation* (33)*. The difference $\mathcal{J}^{\mathrm{IW}}_Q(f) - \mathcal{J}_Q(f)$ equals $\mathcal{J}_+(f) + \mathcal{J}_-(f)$. At $f = f^*_Q$, it equals Δ_{info}.*

Proof. For $x > 0$ and $t \in \mathbb{R}$, the Fenchel inequality $x \log x \geq xt - e^{t-1}$ is tight if and only if $t = 1 + \log x$. Apply it at each pair with $x = \kappa_Q(i, j)$ and $t = f(i, j)$. Taking the expectation under the product of student marginals gives

$$\begin{aligned} I_Q &= \mathbb{E}_{Q_U \otimes Q_V}[\kappa_Q(i, j) \log \kappa_Q(i, j)] \\ &\geq \mathbb{E}_Q[f(i, j)] - \mathbb{E}_{Q_U \otimes Q_V}[e^{f(i,j)-1}] = \mathcal{J}_Q(f). \end{aligned} \tag{34}$$

For the importance-weighted bound, multiply the same pointwise inequality by $\rho(i, j) > 0$. Since $\rho(i, j)\kappa_Q(i, j)Q_U(i)Q_V(j) = P(i, j)$, integration yields

$$\begin{aligned} I^{\mathrm{IW}}_Q &= \mathbb{E}_{Q_U \otimes Q_V}[\rho(i, j)\kappa_Q(i, j) \log \kappa_Q(i, j)] \\ &\geq \mathbb{E}_P[f(i, j)] - \mathbb{E}_{Q_U \otimes Q_V}[\rho(i, j)e^{f(i,j)-1}] = \mathcal{J}^{\mathrm{IW}}_Q(f). \end{aligned} \tag{35}$$

This argument does not require the measure with masses $\rho(i, j)Q_U(i)Q_V(j)$ to be normalized. Both inequalities become equalities at every pair when $f = f^*_Q$, proving that their suprema equal I_Q and I^{IW}_Q. Subtracting the functional definitions gives

$$\begin{aligned} \mathcal{J}^{\mathrm{IW}}_Q(f) - \mathcal{J}_Q(f) &= \mathbb{E}_P[f(i, j)] - \mathbb{E}_Q[f(i, j)] \\ &\quad + \mathbb{E}_{Q_U \otimes Q_V}[(1 - \rho(i, j))e^{f(i,j)-1}] \\ &= \mathcal{J}_+(f) + \mathcal{J}_-(f). \end{aligned} \tag{36}$$

At the common maximizer, this difference is $I^{\mathrm{IW}}_Q - I_Q = \Delta_{\mathrm{info}}$. □

Define the nonnegative slacks $\eta_{\mathrm{IW}}(f) = I_Q^{\mathrm{IW}} - \mathcal{J}_Q^{\mathrm{IW}}(f)$ and $\eta_Q(f) = I_Q - \mathcal{J}_Q(f)$. Then Δ_{info} is

$$\Delta_{\mathrm{info}} = \mathcal{J}_+(f) + \mathcal{J}_-(f) + \eta_{\mathrm{IW}}(f) - \eta_Q(f). \tag{37}$$

The difference $\mathcal{J}_Q^{\mathrm{IW}}(f) - \mathcal{J}_Q(f)$ alone has no fixed inequality direction relative to Δ_{info}.

Proposition 5 (Shared-critic slack cancellation)**.** *For $h(t) = e^t - 1 - t$ and $\delta_f = f - f_Q^*$, $\eta_Q = \mathbb{E}_Q[h(\delta_f)]$ and $\eta_{\mathrm{IW}} = \mathbb{E}_P[h(\delta_f)]$. Hence*

$$|\eta_{\mathrm{IW}} - \eta_Q| \le \|P - Q\|_1 \|h(\delta_f)\|_\infty. \tag{38}$$

At $P = Q$, $\eta_{\mathrm{IW}}(f) - \eta_Q(f) = 0$ for every shared critic, although $\eta_{\mathrm{IW}}(f) = \eta_Q(f) = 0$ only at $f = f_Q^*$.

Proof of Proposition 5. Write $r = \kappa_Q$, $f_Q^* = 1 + \log r$, and $\delta_f = f - f_Q^*$. The pointwise Fenchel residual is

$$r \log r - r f + e^{f-1} = r\big(e^{\delta_f} - 1 - \delta_f\big) = r\, h(\delta_f). \tag{39}$$

Integrating under $Q_U \otimes Q_V$ gives $\eta_Q(f) = \mathbb{E}_Q[h(\delta_f)]$. Multiplying by $\rho = P/Q$ before integrating gives $\eta_{\mathrm{IW}}(f) = \mathbb{E}_P[h(\delta_f)]$. Hence

$$\eta_{\mathrm{IW}}(f) - \eta_Q(f) = \sum_{i,j} (P(i,j) - Q(i,j)) h(\delta_f(i,j)). \tag{40}$$

Taking absolute values and applying the triangle inequality proves Equation (38). When $P = Q$, $\eta_{\mathrm{IW}}(f) - \eta_Q(f) = 0$ for any common critic, even if $\eta_{\mathrm{IW}}(f), \eta_Q(f) > 0$. Finally, $h(t) \ge 0$ with equality exactly at $t = 0$. Since every mass of P and Q is positive, either $\eta_{\mathrm{IW}}(f)$ or $\eta_Q(f)$ is zero if and only if $\delta_f(i,j) = 0$ for every pair. □

Constant critic shifts. For $f = f_Q^* + c$, $\eta_{\mathrm{IW}}(f) = \eta_Q(f) = h(c) = e^c - 1 - c$ for any P, Q. Thus $\eta_{\mathrm{IW}}(f) - \eta_Q(f) = 0$ even for $c \neq 0$ and $P \neq Q$. At $P = Q$, both $\mathcal{J}_+(f)$ and $\mathcal{J}_-(f)$ are zero for any common critic, as is Δ_{info}.

Fixed cosine critics and $\mathcal{L}_{\mathrm{FD}}$. For $f^S(i,j) = a S_{ij}^S/\tau_{\mathrm{IW}} + b$, cosine similarities in $[-1,1]$, fixed finite a, b, and fixed positive $\tau_{\mathrm{IW}}, \tau_G$, $\|h(\delta_f)\|_\infty$ is uniformly bounded. Indeed, writing $x_{ij} = S_{ij}^S/\tau_G$ gives

$$\kappa_Q(i,j) = \frac{e^{x_{ij}} \sum_{k,\ell} e^{x_{k\ell}}}{(\sum_\ell e^{x_{i\ell}})(\sum_k e^{x_{kj}})}.$$

Since $|x_{ij}| \le 1/\tau_G$, we have $|\log \kappa_Q(i,j)| \le 4/\tau_G$, and therefore

$$|\delta_f(i,j)| \le |a|/\tau_{\mathrm{IW}} + |b-1| + 4/\tau_G.$$

Continuity of h on this bounded interval bounds $\|h(\delta_f)\|_\infty$. Consequently, Equation (38) gives $|\mathcal{R}(f^S)| \le C\ \mathrm{TV}(P,Q)$, where $\mathrm{TV}(P,Q) = \frac{1}{2}\|P - Q\|_1$ and the finite constant C depends only on $a, b, \tau_{\mathrm{IW}}, \tau_G$.

The Gibbs continuity argument in Equation (43) and the bound on similarity errors in terms of feature errors in its proof give

$$|\eta_{\mathrm{IW}}(f^S) - \eta_Q(f^S)| \le \frac{\|h(f^S - f_Q^*)\|_\infty}{\tau_G} \sqrt{2BD\,\mathcal{L}_{\mathrm{FD}}}. \tag{41}$$

Thus $\mathcal{L}_{\mathrm{FD}} \to 0$ implies $\mathcal{R}(f^S) = \eta_{\mathrm{IW}}(f^S) - \eta_Q(f^S) \to 0$ under these fixed settings, without requiring $f^S = f_Q^*$.

B.6 Feature Distillation Bounds $\mathcal{J}_+$ and Δ_{info}

For reference, the two bounds in terms of feature error are

$$|\mathcal{J}_+(f^S)| \le \frac{|a|}{\tau_{\mathrm{IW}} \tau_G} \sqrt{2BD\mathcal{L}_{\mathrm{FD}}}, \qquad |\Delta_{\mathrm{info}}| \le \frac{4}{\tau_G^2} \sqrt{2BD\mathcal{L}_{\mathrm{FD}}}. \tag{42}$$

Proof of Proposition 3. Flatten the two cosine similarity matrices into vectors $x = \mathrm{vec}(S^T)/\tau_G$ and $y = \mathrm{vec}(S^S)/\tau_G$, so $P = \mathrm{softmax}(x)$ and $Q = \mathrm{softmax}(y)$. In the following bounds, the matrix ∞ norm denotes the largest absolute entry. For $p_t = \mathrm{softmax}(y + t(x - y))$ and $h = x - y$,

$$\frac{d(p_t)_k}{dt} = (p_t)_k(h_k - \mathbb{E}_{p_t} h).$$

Cauchy–Schwarz and the variance bound for a variable in $[-\|h\|_\infty, \|h\|_\infty]$ give $\sum_k |d(p_t)_k/dt| \le \sqrt{\mathrm{Var}_{p_t}(h)} \le \|h\|_\infty$. Integrating along the segment yields

$$\|P - Q\|_1 \le \|x - y\|_\infty = \frac{\|S^T - S^S\|_\infty}{\tau_G}. \tag{43}$$

Because $\sum_{i,j}(P - Q)(i, j) = 0$, the critic bias b cancels. Since $|S^S_{ij}| \le 1$,

$$|\mathcal{J}_+(f^S)| = \left| \sum_{i,j} (P - Q)(i, j) \frac{a S^S_{ij}}{\tau_{\mathrm{IW}}} \right| \le \frac{|a|}{\tau_{\mathrm{IW}} \tau_G} \|S^T - S^S\|_\infty. \tag{44}$$

For $\epsilon^u_i = \|u^T_i - u^S_i\|_2$ and $\epsilon^v_j = \|v^T_j - v^S_j\|_2$, unit normalization gives $|S^T_{ij} - S^S_{ij}| \le \epsilon^u_i + \epsilon^v_j$. Thus

$$\|S^T - S^S\|_\infty \le \sqrt{2 \sum_i [(\epsilon^u_i)^2 + (\epsilon^v_i)^2]} = \sqrt{2BD\mathcal{L}_{\mathrm{FD}}},$$

which proves the bound on $|\mathcal{J}_+(f^S)|$ in Equation (42). □

Proof of Corollary 1. By the definition of importance weighting,

$$\Delta_{\mathrm{info}} = \sum_{i,j} (P - Q)_{ij} \log \kappa_Q(i, j).$$

Taking absolute values yields $|\Delta_{\mathrm{info}}| \le \|P - Q\|_1 \|\log \kappa_Q\|_\infty$. Equation (43) and the estimate of similarity errors above give $\|P - Q\|_1 \le \sqrt{2BD\mathcal{L}_{\mathrm{FD}}}/\tau_G$. For completeness, let $x_{ij} = S^S_{ij}/\tau_G$ and write

$$\log \kappa_Q(i, j) = x_{ij} + \log \sum_{k,\ell} e^{x_{k\ell}} - \log \sum_{\ell} e^{x_{i\ell}} - \log \sum_{k} e^{x_{kj}}.$$

Since $|x_{ij}| \le 1/\tau_G$, the two $\log B$ terms cancel the $\log B^2$ term and the four remaining deviations have absolute sum at most $4/\tau_G$. Multiplying the bounds on $\|P - Q\|_1$ and $\|\log \kappa_Q\|_\infty$ proves the bound on $|\Delta_{\mathrm{info}}|$ in Equation (42). Combining it with Equation (10) proves Equation (27). □

The bound applies to the entire information term Δ_{info}, including $\mathcal{J}_-$ and $\mathcal{R}$ in Equation (11). Its coefficient $4\sqrt{2BD}/\tau^2_G$ can make the bound large at attained feature errors.

Corollary 2 (Sufficient condition for ranking preservation)**.** *Write $\epsilon^u_i = \|u^T_i - u^S_i\|_2$ and $\epsilon^v_i = \|v^T_i - v^S_i\|_2$. For T2I, let $m^M_{ij} = S^M_{jj} - S^M_{ij}$ as in Appendix B.4. For I2T, let $\widetilde{m}^M_{ij} = S^M_{ii} - S^M_{ij}$. Unit normalization gives*

$$|m^T_{ij} - m^S_{ij}| \le \epsilon^u_j + \epsilon^u_i + 2\epsilon^v_j,$$
$$|\widetilde{m}^T_{ij} - \widetilde{m}^S_{ij}| \le 2\epsilon^u_i + \epsilon^v_i + \epsilon^v_j.$$

In either direction, a teacher margin larger than the corresponding bound implies a positive student margin.

Proof. Apply $|S^T_{ij} - S^S_{ij}| \le \epsilon^u_i + \epsilon^v_j$ to the matched and unmatched cosine similarities and use the triangle inequality. □

C Variational Functionals and Surrogate Objectives

C.1 Variational Envelopes

Retaining $-D_{\mathrm{KL}}(P_V\|Q_V)$ gives an exact characterization without assuming equal text marginals. Write $G(P,Q) = I_P - I_Q$. Proposition 1 gives $G = \Delta_{\mathrm{dist}} - D_{\mathrm{KL}}(P_V\|Q_V) + I_Q^{\mathrm{IW}} - I_Q$. Define

$$\overline{G}(f) = \Delta_{\mathrm{dist}} - D_{\mathrm{KL}}(P_V\|Q_V) + I_Q^{\mathrm{IW}} - \mathcal{J}_Q(f), \tag{45}$$

$$\underline{G}(f) = \Delta_{\mathrm{dist}} - D_{\mathrm{KL}}(P_V\|Q_V) + \mathcal{J}_Q^{\mathrm{IW}}(f) - I_Q. \tag{46}$$

Proposition 4 gives $\underline{G}(f) \le G(P,Q) \le \overline{G}(f)$ for every critic. Both bounds are tight at $f_Q^* = 1 + \log \kappa_Q$, so

$$\sup_f \underline{G}(f) = G(P,Q) = \inf_f \overline{G}(f).$$

The same sandwich gives $|G| \le \max\{|\underline{G}(f)|, |\overline{G}(f)|\}$, with equality at f_Q^*. Hence $\inf_f \max\{|\underline{G}(f)|, |\overline{G}(f)|\} = |I_P - I_Q|$.

The critic slacks. Let $D(f) = \Delta_{\mathrm{dist}} - D_{\mathrm{KL}}(P_V\|Q_V) + \mathcal{J}_Q^{\mathrm{IW}}(f) - \mathcal{J}_Q(f)$, with $\eta_{\mathrm{IW}}(f), \eta_Q(f)$ as in Equation (37). Then

$$\begin{aligned} G &= D(f) + \eta_{\mathrm{IW}}(f) - \eta_Q(f), \\ \underline{G}(f) &= D(f) - \eta_Q(f), \qquad \overline{G}(f) = D(f) + \eta_{\mathrm{IW}}(f). \end{aligned} \tag{47}$$

The difference $D(f)$ alone is not a bound: dropping $\eta_{\mathrm{IW}}(f) - \eta_Q(f)$ has no fixed inequality direction. If $P_V = Q_V$, $-D_{\mathrm{KL}}(P_V\|Q_V) = 0$.

C.2 The Tight Critic and $\mathcal{J}_-$

Lemma 1 (Cancellation of $\mathcal{J}_-$ at the tight critic)**.** *At $f_Q^* = 1 + \log \kappa_Q$,*

$$\mathcal{J}_-(f_Q^*) = 0, \qquad \mathcal{J}_+(f_Q^*) = \Delta_{\mathrm{info}}. \tag{48}$$

Proof. Since $e^{f_Q^*(i,j)-1} = \kappa_Q(i,j) = Q(i,j)/(Q_U(i)Q_V(j))$, $\mathcal{J}_-(f_Q^*) = \sum_{i,j}(Q(i,j) - P(i,j)) = 0$. The other equality follows from Proposition 4. □

With ρ held fixed, $\partial\mathcal{J}_-(f)/\partial f(i,j)$ at $f = f_Q^*$ is still $Q(i,j) - P(i,j)$. We use the fixed affine critic $f^S(i,j) = aS_{ij}^S/\tau_{\mathrm{IW}} + b$ defined in Section 3.1.

C.3 Relation to the Implemented Objective

In $\mathcal{L}_{\mathrm{base}}$ (Section 3.2), $\mathcal{L}_{\mathrm{KL}}$ replaces the Gibbs marginal weights $P_U(i)$ in Δ_{dist} by $1/B$. For CGD-IW, $\mathcal{L}_{\mathrm{IW}}$ replaces $\mathcal{J}_-$ using the fixed critic f^S, the conditional ratio $c_j^T(i)/c_j^S(i)$ in place of $\rho(i,j) = P(i,j)/Q(i,j)$, support H, and normalization by Z_H in a uniform average. W_{ij}, H, and Z_H are detached. For CGD-N, the target/non-target split gives $D_{\mathrm{KL}}(c_j^{T,-}\|c_j^{S,-})$, averaged over captions with an independent coefficient λ_N. These substitutions define $\mathcal{L}_{\mathrm{IW}}$ and $\mathcal{L}_N$. The variational bounds above apply to $\mathcal{J}_Q$ and $\mathcal{J}_Q^{\mathrm{IW}}$. Proposition 2 characterizes when $\mathcal{L}_{\mathrm{KL}} + \omega\mathcal{L}_N$ vanishes, independently of the critic envelopes.

D Experimental Protocol

D.1 Training and Evaluation

Baseline implementations. All methods use the model, data, and optimization settings in Table 6. CLIP-KD follows the released implementation: the two directions in $\mathcal{L}_{\mathrm{CRD}}$ are averaged, and $\mathcal{L}_{\mathrm{FD}}$ is an element-wise mean, and ICL multiplies similarities directly by $\log(1/0.07) \approx 2.66$, without exponentiation. We retain its coefficient on feature loss of 2000. All methods train on 13,875,035 image–caption pairs with a batch of 1,024 in a single process.

CLIP-RD. We use the released code (commit `af71ec0`): $\mathcal{L}_{\text{CLIP}} + \lambda_F \mathcal{L}_{\text{FD}}$, CLIP-KD's relational and interactive contrastive terms, and VRD and XRD at coefficient 1. The released relational KLs reverse the preprint's direction and detach their target distributions. VRD averages two KL terms and weights each cross-entropy component by $1/2$. XRD averages four directions. Auxiliary similarity scales are learnable raw multipliers initialized at $\log(1/0.07) \approx 2.66$ and used without exponentiation. XRD and ICL share one cross scale. Relational terms are computed in FP32 outside the mixed-precision context.

Table 6: Configuration for 32-epoch training. Exact scalar values are recorded in the repository configuration files.

Item	Value
Teacher	ViT-B/16, LAION-400M checkpoint, frozen
Student	ViT-T/16; image and text encoders trainable
Data	CC3M + CC12M; 13,875,035 image–caption pairs per epoch
Optimization	32 epochs; batch size 1,024; automatic mixed precision; AdamW
AdamW	Learning rate 10^{-3}; $\beta = (0.9, 0.98)$; $\epsilon = 10^{-6}$; weight decay 0.1
Schedule	Learning-rate warmup for 10,000 updates; decay over the full 32 epochs
$\mathcal{L}_{\text{FD}}$	Squared differences between normalized teacher and student embeddings, for both encoders; averaged over the batch and D=512 dimensions; coefficient 2,000
$\mathcal{L}_{\text{KL}}$	$D_{\text{KL}}(r_i^T \| r_i^S)$ averaged over images; coefficient 1; $\tau_{\text{KL}} = 0.07$
$\mathcal{L}_{\text{IW}}$ support	The teacher's 64 most similar unmatched captions for each image ($k_h = 64$)
$\mathcal{L}_{\text{IW}}$ ratio	$W_{ij} = \text{sg}[c_j^T(i)/c_j^S(i)]$; $\tau_w = 0.07$; Z_H is detached
$\mathcal{L}_{\text{IW}}$ critic	$f = 0.05\, S^S/0.07$
$\mathcal{L}_{\text{IW}}$ schedule	$\lambda_{\text{IW}} \approx 20.92$ through epoch 24; target ratio of gradient norms 0.10 thereafter
CGD-N component	$\mathcal{L}_N$; $\tau_N = 0.07$; no temperature-squared multiplier
CGD-N schedule	gradient-norm ratio target 0.0382 from the first update; Equation (50)
Training seeds	0–4 for all five methods; seed 0 for the loss component study
Inference	student only

Setting loss coefficients from gradient norms. The controller measures gradients with respect to the image and text feature tensors entering the loss. Let F^S concatenate the student image and text encoder outputs before unit normalization. For $m \in \{\text{IW}, N\}$, the controller targets

$$\frac{\|\lambda_m(t)\nabla_{F^S}\mathcal{L}_m\|_2}{\|\nabla_{F^S}\mathcal{L}_{\text{base}}\|_2} \approx \gamma_m. \tag{49}$$

At step t,

$$n_x^{(t)} = \|\nabla_{F^S}\mathcal{L}_x\|_2, \qquad x \in \{\text{base}, m\}, \quad m \in \{\text{IW}, N\}.$$

At a controller update, exponential moving averages $\bar{n}_x^{(t)} = 0.95\bar{n}_x^{(t-1)} + 0.05 n_x^{(t)}$ determine

$$\lambda_m^{(t)} = \text{clip}_{[0,300]}\left(\frac{\gamma\, \bar{n}_{\text{base}}^{(t)}}{\bar{n}_m^{(t)} + \epsilon}\right), \qquad \epsilon = 10^{-12}. \tag{50}$$

The first measurement initializes each EMA directly. Updates occur every 50 steps. The coefficient is cached between updates. CGD-IW activates this controller after epoch 24 with $\gamma = 0.10$. CGD-N uses it from the first update with $\gamma \approx 0.0382$. The current ratio $\lambda_m^{(t)} n_m^{(t)}/n_{\text{base}}^{(t)}$ can differ from γ because the controller uses moving averages and reuses coefficients between updates.

Training-gradient calibration of the initial IW coefficient. The fixed coefficient $\lambda_{\text{IW}} = 20.92$ was computed from a preliminary model trained for one epoch with $\mathcal{L}_{\text{base}}$. On four training batches of 128 pairs, we measured parameter-gradient norms and averaged the coefficients that made $\|\lambda_{\text{IW}}\nabla\mathcal{L}_{\text{IW}}\|_2$ equal to 10% of $\|\nabla(\lambda_F\mathcal{L}_{\text{FD}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}})\|_2$. The 10% budget is a design choice. This calibration excludes $\nabla\mathcal{L}_{\text{CLIP}}$ and uses model parameters. The controller above instead uses student features and the full $\mathcal{L}_{\text{base}}$. Measurements at eight checkpoints of two preliminary runs found that this fixed coefficient contributed 3.50–3.85% of the feature-gradient norm of $\mathcal{L}_{\text{base}}$, motivating direct control in feature space.

Transfer of the measured budget to CGD-N. On a separate $\mathcal{L}_{\text{base}}$ model at epoch 16 (seed 0), four fixed training batches of 256 pairs gave a mean feature-gradient ratio of 0.0382 for the fixed λ_{IW} relative to $\mathcal{L}_{\text{base}}$. We reused this measured value as the common target for $\mathcal{L}_{\text{IW}}$ and $\mathcal{L}_N$ in the component study, then retained it from the first to the last update of CGD-N full training.

CGD-IW schedule. The gradient measurements above motivated controlling the IW contribution in feature space. We retained the calibrated fixed coefficient through epoch 24 and the target $\gamma = 0.10$ thereafter. Pilot continuations compared targets 0.10 and 0.20, retaining the fixed coefficient, removing $\mathcal{L}_{\text{IW}}$, and activating a target of 0.05 earlier. These comparisons used ImageNet zero-shot validation and the same COCO, Flickr30K, and CC3M-val splits reported here. The earlier 0.05 schedule improved mean retrieval but reduced ImageNet accuracy. The retained schedule was fixed before the subsequent full-run replications.

Negative-column implementation. CGD-N masks the diagonal before computing the column softmax at $\tau_N = 0.07$. Teacher probabilities are detached. The loss sums $c_j^{T,-}(i) \log[c_j^{T,-}(i)/c_j^{S,-}(i)]$ over unmatched images $i \neq j$, then averages over captions j. Softmax and KL arithmetic use FP32, without a τ_N^2 multiplier. The controller sets λ_N from the first update. The coefficients on $\mathcal{L}_{\text{CLIP}}$, $\mathcal{L}_{\text{FD}}$, and $\mathcal{L}_{\text{KL}}$ are 1, 2,000, and 1, respectively.

Evaluation and aggregation. Main comparisons use models trained for 32 epochs in five runs per method (seeds 0–4). We report means and sample standard deviations with denominator $n - 1$. Classification accuracy and retrieval recall are averaged over evaluation settings within each run before aggregation across runs. Runs with the same seed index are not paired across methods. Comparisons between student architectures use seed 0. All 25 models were reevaluated after 32 epochs with the same evaluation code and settings, using FP32 and unit-normalized embeddings. COCO retrieval uses 5,000 images and 25,010 captions, and Flickr30K uses 1,000 images and 5,000 captions. Each recall cutoff is averaged across the three retrieval datasets and both directions. The four classification datasets are averaged separately at each top-k cutoff.

D.2 Loss Components

We first train one model with $\mathcal{L}_{\text{CLIP}} + \lambda_F \mathcal{L}_{\text{FD}}$ and seed 0 for 16 epochs, halfway through the full 32-epoch schedule. We then copy the model, optimizer, and learning-rate scheduler states into six branches. Each branch receives the same training examples in the same order for four further epochs. This compares what would change if we added a different loss at the same point in training. All branches retain $\mathcal{L}_{\text{CLIP}} + \lambda_F \mathcal{L}_{\text{FD}}$ and the original learning-rate schedule. The additional losses are none, $\mathcal{L}_{\text{KL}}$, $\lambda_{\text{IW}} \mathcal{L}_{\text{IW}}$, $\lambda_N \mathcal{L}_N$, $\mathcal{L}_{\text{KL}} + \lambda_{\text{IW}} \mathcal{L}_{\text{IW}}$, or $\mathcal{L}_{\text{KL}} + \lambda_N \mathcal{L}_N$. The coefficients λ_{IW} and λ_N are set by Equation (50) with target $\gamma = 0.0382$.

D.3 Additional Relevant Matches

The comparison between COCO and CxC labels uses the same models trained for 32 epochs with seed 0 and the same stored FP32 similarities. It maps the COCO Karpathy data to 5,000 image and 25,000 caption identifiers. Standard COCO relevance uses annotated image–caption matches. CxC uses the additional relevance annotations of Parekh et al. (2021). Both evaluations retain the same candidate gallery and change only which candidates count as positives.

D.4 Representation Metrics

All metrics of Section 4.5 are computed on FP32, unit-normalized embeddings of CC3M-val image–caption pairs, over the same three disjoint folds of 4,096 pairs each. We first average the three fold values within each seed, then report the mean and sample standard deviation over seeds 0–4. Row and column KL use $\tau = 0.07$. Hard gaps and rank statistics use cosine similarities before temperature scaling.

Image and text CKA are linear centered kernel alignment values between the student's and the teacher's embedding matrices on the same fold, computed separately per modality. We also report the mean cosine similarity between each student image embedding and its teacher counterpart.

Positive cosine similarity is the mean cosine between the student's matched image and caption embeddings.

For row KL, we use $r_i^{M,-}(j) = r_i^M(j)/(1 - r_i^M(i))$ for $j \neq i$ and average $D_{\mathrm{KL}}(r_i^{T,-} \| r_i^{S,-})$ over images. Column KL averages $D_{\mathrm{KL}}(c_j^{T,-} \| c_j^{S,-})$ over captions. The I2T hard gap averages $S_{ii}^S - \max_{j \neq i} S_{ij}^S$ over images. The T2I hard gap averages $S_{jj}^S - \max_{i \neq j} S_{ij}^S$ over captions. Spearman measures the rank correlation between teacher and student cosine similarities over all unmatched entries of a row (column). For each fold and direction, we select the same 512 query indices for all methods and compare all 4,095 unmatched candidates for each query. Tied cosine similarities receive average ranks. We average across queries within each fold, then across the three folds. The two KL values depend on the scale of each method's cosine similarities, whereas the Spearman and overlap quantities do not (see Calibration below). Top-64 overlap is the fraction of the teacher's 64 unmatched candidates with the highest cosine similarities that also appear in the student's 64 most similar unmatched candidates.

Calibration. For each method and seed, we fit one scalar $a_{M,f} \in [0.05, 5]$ by minimizing the sum of negative-only teacher-to-student row and column KL at $\tau = 0.07$ on fold $(f - 1) \bmod 3$. We multiply the student's similarities on evaluation fold f by this scalar and recompute both KLs. Spearman and top-64 overlap are invariant to this positive rescaling.

E Additional Results

E.1 Main Results

Table 7 gives classification and retrieval averages over five runs. Table 8 lists classification by seed. Table 10 gives retrieval at all three cutoffs for seed 0. The pretrained teacher's results are in Table 9.

Table 7: ViT-T/16 results averaged over evaluation settings after 32 epochs (%): means and sample standard deviations over five runs. Classification averages four datasets. Retrieval averages three datasets and both directions within each run. Dataset-level results are in Table 1.

	Classification		Retrieval		
Method	Top-1	Top-5	R@1	R@5	R@10
FD	37.74 ± 0.22	65.54 ± 0.20	41.98 ± 0.24	67.22 ± 0.25	76.28 ± 0.25
CLIP-KD	38.71 ± 0.43	66.68 ± 0.38	42.59 ± 0.43	67.46 ± 0.29	76.39 ± 0.24
CLIP-RD	<u>39.07</u> ± 0.26	**67.25** ± 0.38	42.64 ± 0.29	67.48 ± 0.28	76.42 ± 0.13
CGD-N (Ours)	**39.11** ± 0.12	67.02 ± 0.23	<u>43.11</u> ± 0.13	**68.06** ± 0.21	<u>76.84</u> ± 0.06
CGD-IW (Ours)	39.03 ± 0.54	<u>67.19</u> ± 0.54	**43.17** ± 0.48	<u>68.02</u> ± 0.37	**76.96** ± 0.32
Teacher (ViT-B/16)	64.25	86.96	53.97	75.51	82.06

Table 8: Zero-shot classification after 32 epochs, by training seed (%). Each entry gives top-1 and top-5 accuracy, separated by a slash. Bold and underlining mark the best and second-best values for each metric.

Seed	Dataset	FD	CLIP-KD	CLIP-RD	CGD-N (Ours)	CGD-IW (Ours)
0	ImageNet	41.80 / 72.06	43.05 / 72.93	43.30 / 73.34	43.10 / 73.14	**43.45 / 73.75**
	ImageNet-V2	36.19 / 65.51	37.21 / 67.15	37.19 / **67.23**	**37.69** / 66.41	37.30 / 66.80
	ImageNet-R	46.37 / 72.97	47.67 / 73.98	49.07 / 75.00	48.45 / 75.10	**49.98 / 75.79**
	ImageNet-Sketch	25.71 / 51.20	26.47 / 52.72	27.70 / 54.22	27.39 / 53.84	**28.02 / 55.23**
1	ImageNet	41.99 / 72.12	42.94 / 72.78	42.96 / 73.21	43.43 / 73.19	**43.50 / 73.33**
	ImageNet-V2	36.56 / 65.39	36.79 / 66.31	37.20 / **67.13**	**37.56** / 66.97	37.38 / 66.60
	ImageNet-R	46.55 / 72.90	47.59 / 73.96	48.52 / 74.86	48.43 / 75.07	**49.05 / 75.47**
	ImageNet-Sketch	25.84 / 51.82	26.70 / 52.76	**27.89** / 54.65	27.60 / 54.31	27.80 / **54.72**
2	ImageNet	42.69 / 72.17	**43.28** / 73.25	42.91 / **73.40**	42.68 / 73.09	42.79 / 72.97
	ImageNet-V2	36.63 / 65.73	**37.31** / 66.82	36.82 / **66.83**	36.67 / 66.16	36.75 / 66.66
	ImageNet-R	47.17 / 73.17	48.33 / 74.51	48.80 / **75.12**	**48.85** / 74.63	47.57 / 73.79
	ImageNet-Sketch	25.87 / 51.92	27.56 / 53.40	**28.09 / 54.34**	27.52 / 53.51	26.88 / 53.27
3	ImageNet	42.45 / 72.13	42.68 / 72.42	42.88 / 72.84	**43.22 / 73.13**	42.54 / 72.83
	ImageNet-V2	36.80 / 65.25	36.42 / **66.45**	36.92 / 66.24	**37.16** / 66.28	36.76 / 66.41
	ImageNet-R	46.22 / 72.81	46.99 / 73.47	47.79 / 73.92	**48.54 / 74.48**	47.94 / 74.39
	ImageNet-Sketch	25.34 / 50.88	26.51 / 52.39	26.89 / 53.28	**27.33 / 53.48**	26.73 / 52.94
4	ImageNet	42.14 / 72.26	43.44 / 73.25	43.21 / **73.42**	**43.45** / 73.14	43.12 / 73.00
	ImageNet-V2	36.20 / 65.91	**37.80** / 66.73	37.15 / 66.71	37.75 / 66.41	37.29 / **66.97**
	ImageNet-R	46.37 / 72.97	48.08 / 74.66	48.45 / 75.02	48.45 / 74.51	**48.59 / 75.18**
	ImageNet-Sketch	25.81 / 51.67	27.40 / 53.74	**27.72 / 54.33**	26.91 / 53.59	27.11 / 53.68
Mean	ImageNet	42.22 / 72.15	43.08 / 72.93	43.05 / **73.24**	**43.18** / 73.14	43.08 / 73.18
	ImageNet-V2	36.48 / 65.56	37.11 / 66.69	37.06 / **66.83**	**37.37** / 66.45	37.10 / 66.69
	ImageNet-R	46.54 / 72.96	47.73 / 74.12	48.53 / 74.78	48.54 / 74.76	**48.63 / 74.92**
	ImageNet-Sketch	25.72 / 51.50	26.93 / 53.00	**27.66 / 54.16**	27.35 / 53.74	27.31 / 53.97

Table 9: Frozen pretrained ViT-B/16 teacher results (%). Classification reports top-1 and top-5. Retrieval reports recall at three cutoffs. The 4-set and retrieval means average four datasets and six dataset–direction combinations, respectively.

Dataset	Top-1	Top-5	R@1	R@5	R@10
ImageNet	67.05	90.43	-	-	-
ImageNet-V2	59.66	85.40	-	-	-
ImageNet-R	77.92	92.91	-	-	-
ImageNet-Sketch	52.37	79.10	-	-	-
4-set	64.25	86.96	-	-	-
MSCOCO I2T	-	-	55.44	79.72	86.90
MSCOCO T2I	-	-	38.33	63.61	73.85
Flickr30K I2T	-	-	83.30	96.90	98.50
Flickr30K T2I	-	-	65.52	88.28	92.98
CC3M-val I2T	-	-	41.48	63.27	70.80
CC3M-val T2I	-	-	39.78	61.29	69.34
Retrieval mean	-	-	53.97	75.51	82.06

Table 10: Retrieval results for each dataset (%, seed 0). Each entry gives R@1, R@5, and R@10, separated by slashes. Bold and underlining mark the best and second-best student values per metric.

Dataset	Direction	FD	CLIP-KD	CLIP-RD	CGD-N (Ours)	CGD-IW (Ours)
MSCOCO	I2T	40.88 / 66.94 / 76.60	41.44 / 68.00 / 78.38	42.44 / 68.36 / 78.30	42.20 / 67.80 / 77.92	**42.94 / 68.86 / 78.46**
	T2I	27.48 / 52.87 / 64.62	27.69 / 53.19 / 64.78	27.71 / 53.04 / 64.80	28.20 / **54.00** / 65.57	**28.51** / 53.75 / **65.92**
Flickr30K	I2T	65.70 / 86.60 / 92.00	**68.30 / 89.40** / 93.70	67.00 / 89.30 / **94.10**	66.90 / 88.40 / 93.20	65.50 / 87.50 / 92.80
	T2I	50.64 / 77.76 / 85.28	51.92 / 77.88 / 86.02	51.78 / 78.60 / 86.46	52.58 / 78.98 / 86.60	**52.84 / 79.46 / 87.02**
CC3M-val	I2T	33.53 / 58.55 / 68.76	34.35 / 58.86 / 68.09	34.23 / 58.68 / 68.33	34.36 / 59.70 / 69.17	**35.30 / 59.96 / 69.50**
	T2I	33.67 / 58.98 / 68.73	32.74 / 57.90 / 67.31	33.21 / 57.78 / 67.48	33.96 / 58.98 / 68.55	**34.09 / 59.35 / 68.94**

E.2 STUDENT ARCHITECTURES

Tables 12 and 13 give the classification and retrieval results behind Table 4. Each model is trained for 32 epochs with seed 0. Table 11 lists image and text encoder parameter counts.

Table 11: Model parameter counts, including projection layers and excluding loss-only modules. The total includes shared scalar parameters. For ViT-T/16, image/text encoder sizes decrease from 86.2M/63.4M to 5.6M/40.5M: compression is larger in the image encoder.

Model	Image encoder	Text encoder	Native total
ViT-T/16	5,622,912	40,493,184	46,116,097
ResNet-18	11,438,656	40,493,184	51,931,841
EfficientNet-B0	4,662,908	40,493,184	45,156,093
MobileViT-S	5,265,312	40,493,184	45,758,497
Swin-Tiny	27,912,570	40,493,184	68,405,755
Teacher ViT-B/16	86,192,640	63,428,096	149,620,737

Table 12: Classification for each student architecture and dataset (%). Each entry gives top-1 and top-5 accuracy, separated by a slash. Best and second-best values are marked for each metric within each row.

Dataset	FD	CLIP-KD	CLIP-RD	CGD-N (Ours)	CGD-IW (Ours)
Student: ResNet-18					
ImageNet	38.42 / 68.34	39.26 / **69.18**	**39.48** / 69.13	39.15 / 68.66	39.01 / 68.63
ImageNet-V2	32.34 / 61.57	33.14 / 62.35	33.08 / 62.89	**33.63** / **63.20**	33.35 / 62.01
ImageNet-R	44.11 / 70.12	44.97 / 70.90	45.14 / **71.26**	**45.16** / **71.26**	44.90 / 71.14
ImageNet-Sketch	25.03 / 50.18	25.48 / 50.96	25.66 / **52.10**	**25.80** / 51.13	25.70 / 51.47
Student: EfficientNet-B0					
ImageNet	46.48 / 75.63	46.17 / 76.13	46.63 / **76.69**	46.70 / 76.49	**46.77** / 76.26
ImageNet-V2	39.17 / 68.65	39.61 / 69.55	40.07 / 70.02	40.28 / 69.48	**40.32** / **70.14**
ImageNet-R	53.30 / 78.51	54.34 / 78.82	54.41 / **79.28**	54.84 / 79.04	**54.90** / 79.13
ImageNet-Sketch	31.62 / 59.15	32.18 / 59.67	**32.82** / 60.44	32.59 / 60.23	32.81 / **60.47**
Student: MobileViT-S					
ImageNet	46.16 / 76.54	**47.08** / 76.98	47.01 / **77.30**	47.01 / 76.88	46.62 / 76.64
ImageNet-V2	39.73 / 70.05	39.94 / 70.72	40.58 / **71.04**	**40.69** / 70.53	40.38 / 69.88
ImageNet-R	53.00 / 78.24	53.83 / 79.07	54.44 / **79.38**	53.90 / 78.94	**54.64** / 79.14
ImageNet-Sketch	30.58 / 58.05	31.34 / 59.10	**31.60** / **59.33**	31.21 / 59.08	31.23 / 58.97
Student: Swin-Tiny					
ImageNet	55.29 / 83.02	55.43 / 83.50	56.20 / **84.15**	56.20 / 83.81	**56.53** / 83.74
ImageNet-V2	48.53 / 77.19	48.03 / 77.56	49.06 / **78.44**	49.51 / 78.27	**49.53** / 78.24
ImageNet-R	67.82 / 87.47	67.93 / 87.51	**68.74** / **88.05**	68.42 / 87.95	68.60 / 87.85
ImageNet-Sketch	42.09 / 70.16	42.32 / 70.36	42.95 / **71.23**	**43.02** / 70.91	42.86 / 70.70

Table 13: Retrieval for each student architecture and dataset (%). Each entry gives R@1, R@5, and R@10, separated by slashes. Best and second-best values are marked for each metric within each row.

Dataset	Direction	FD	CLIP-KD	CLIP-RD	CGD-N (Ours)	CGD-IW (Ours)
Student: ResNet-18						
MSCOCO	I2T	37.12 / 63.48 / 74.82	37.86 / 64.60 / 75.32	**39.04 / 65.24 / 76.02**	38.78 / 63.90 / 74.16	38.90 / 63.74 / 74.96
MSCOCO	T2I	23.42 / **48.34 / 60.58**	23.11 / 47.43 / 59.32	**23.69** / 48.00 / 60.49	23.52 / 48.22 / 60.48	22.60 / 47.23 / 59.43
Flickr30K	I2T	61.20 / 83.40 / 91.10	63.10 / 86.50 / **92.80**	63.00 / 85.50 / 91.90	61.50 / 85.60 / 91.80	**65.30 / 86.60** / 91.70
Flickr30K	T2I	45.72 / 74.20 / 83.40	45.74 / 75.08 / 83.70	**47.20** / 74.78 / 83.28	46.22 / **75.44 / 84.22**	46.60 / 75.16 / 83.26
CC3M-val	I2T	29.59 / 54.04 / 64.06	30.37 / 54.89 / 64.20	**30.79 / 55.18 / 65.24**	30.37 / 54.65 / 64.61	30.32 / 54.38 / 64.23
CC3M-val	T2I	**29.23** / 53.95 / 64.14	28.51 / 53.08 / 62.93	28.80 / 53.29 / 63.45	29.09 / **54.06 / 64.17**	28.82 / 53.52 / 63.74
Student: EfficientNet-B0						
MSCOCO	I2T	43.78 / 69.64 / 79.26	44.32 / 70.70 / 80.04	45.22 / **71.00** / 80.76	**45.38** / 70.94 / 80.14	44.26 / 70.30 / **80.84**
MSCOCO	T2I	29.16 / 55.12 / 66.86	**29.54** / 55.30 / 66.55	29.39 / 55.20 / 66.79	29.10 / **55.41 / 67.07**	28.65 / 54.73 / 66.24
Flickr30K	I2T	70.30 / 90.30 / 94.40	71.40 / 90.20 / 94.40	71.40 / **91.70** / 94.80	**72.40** / 90.60 / **95.20**	**72.40** / 90.30 / 94.40
Flickr30K	T2I	55.04 / 80.40 / 88.14	55.24 / 80.14 / 87.94	55.50 / 80.98 / 88.02	**55.86** / 81.28 / **88.50**	55.62 / **81.34** / 88.26
CC3M-val	I2T	35.30 / 60.66 / 70.16	35.53 / 60.31 / 69.63	35.71 / 60.63 / 69.38	36.00 / 60.59 / 70.11	**36.61 / 61.36 / 70.68**
CC3M-val	T2I	34.75 / **60.80 / 70.26**	34.07 / 59.03 / 68.30	34.24 / 58.73 / 68.03	**34.81** / 60.19 / 69.80	34.73 / 60.54 / 69.53
Student: MobileViT-S						
MSCOCO	I2T	44.44 / 71.34 / 80.36	46.38 / 72.24 / 80.98	**47.04 / 73.08 / 82.08**	46.72 / 72.20 / 81.00	45.86 / 72.00 / 81.06
MSCOCO	T2I	30.20 / 56.25 / 67.92	30.62 / 56.73 / 67.93	**30.70** / 56.71 / **68.19**	30.50 / **57.02** / 68.17	30.58 / 56.55 / 68.04
Flickr30K	I2T	70.90 / 90.50 / 94.40	73.30 / 91.40 / 94.50	**75.60** / 90.50 / 94.80	75.10 / **92.20 / 95.60**	74.30 / 91.00 / **95.60**
Flickr30K	T2I	55.80 / 81.18 / 88.36	56.18 / 81.72 / 88.80	57.06 / 82.02 / **89.42**	**57.52** / 81.74 / 88.92	57.22 / **82.04** / 88.70
CC3M-val	I2T	36.44 / 61.16 / 70.92	36.64 / 61.24 / 70.68	**37.34** / 62.16 / 71.07	36.74 / 61.73 / 71.18	37.31 / **62.66 / 72.17**
CC3M-val	T2I	**36.05** / 61.27 / **71.07**	35.16 / 60.08 / 69.56	35.09 / 60.81 / 69.91	35.54 / 61.16 / 70.67	35.65 / **61.63 / 71.07**
Student: Swin-Tiny						
MSCOCO	I2T	51.70 / 76.66 / 85.36	52.40 / 77.26 / 85.28	**54.08 / 78.26 / 86.08**	52.82 / 77.14 / 85.88	53.52 / 78.20 / 85.88
MSCOCO	T2I	35.97 / 62.32 / 72.74	35.83 / 61.87 / 72.42	36.51 / 62.63 / 72.61	36.50 / 62.79 / **73.57**	**36.86 / 62.86** / 73.32
Flickr30K	I2T	78.00 / 94.60 / 96.90	80.30 / 95.20 / 97.20	**81.70** / 95.00 / 97.20	80.80 / **95.70 / 97.40**	80.30 / 95.10 / 97.30
Flickr30K	T2I	63.38 / 86.76 / 92.18	64.04 / 87.06 / 92.14	64.22 / **87.30** / 92.60	64.34 / 87.02 / **92.70**	**64.48** / 86.84 / 92.34
CC3M-val	I2T	44.98 / 69.55 / 77.81	43.46 / 67.82 / 76.42	44.22 / 68.51 / 77.03	45.08 / 69.98 / **78.17**	**45.24 / 70.11** / 78.12
CC3M-val	T2I	43.38 / 68.67 / **77.01**	41.63 / 65.93 / 74.56	42.16 / 67.02 / 75.15	**43.55** / 68.90 / 76.83	43.41 / **68.94** / 76.84

E.3 Loss Components

Table 14 gives per-dataset R@1 for the six loss combinations in Table 3. All branches retain $\mathcal{L}_{\mathrm{CLIP}} + \lambda_F \mathcal{L}_{\mathrm{FD}}$ and follow Appendix D.2.

Table 14: Retrieval R@1 (%) for each loss combination, by dataset and direction. All branches retain $\mathcal{L}_{\mathrm{CLIP}} + \lambda_F \mathcal{L}_{\mathrm{FD}}$ and train for four further epochs from the common halfway point described in Appendix D.2. Bold and underlining rank values separately within the groups with and without $\mathcal{L}_{\mathrm{KL}}$.

		Without $\mathcal{L}_{\mathrm{KL}}$			With $\mathcal{L}_{\mathrm{KL}}$		
Dataset	Direction	None	$\mathcal{L}_{\mathrm{IW}}$	$\mathcal{L}_N$	$\mathcal{L}_{\mathrm{KL}}$	$\mathcal{L}_{\mathrm{KL}} + \mathcal{L}_{\mathrm{IW}}$	$\mathcal{L}_{\mathrm{KL}} + \mathcal{L}_N$
MSCOCO	I2T	31.12	**32.16**	31.42	31.48	32.12	**32.78**
	T2I	20.98	21.25	**21.48**	21.48	21.49	**21.87**
Flickr30K	I2T	53.90	54.60	**55.60**	54.50	56.10	**58.40**
	T2I	**42.42**	42.08	41.68	42.34	42.56	**43.26**
CC3M-val	I2T	25.84	26.07	**26.16**	25.72	26.18	**26.59**
	T2I	25.59	**25.99**	25.98	**26.21**	25.97	26.09

E.4 Representation Diagnostics

Table 15 gives the values behind Figure 7, together with the cosine similarity between corresponding teacher and student image embeddings, matched image–caption cosine similarity, and top-64 overlap. It reports means and sample standard deviations across five runs, after averaging the three CC3M folds within each run. Both CGD variants have higher Spearman correlation than CLIP-KD and CLIP-RD in both directions, but lower top-64 overlap. These measure agreement over the full candidate list and within its highest-ranked subset, respectively.

Table 15: Representation measures after 32 epochs. We average three CC3M-val subsets within each run, then report the mean and sample standard deviation over five runs. Before computing KL, we calibrate student cosine similarities on a separate subset. Definitions are in Appendix D.4.

Metric	FD	CLIP-KD	CLIP-RD	CGD-N	CGD-IW
Image CKA	0.8124 ± 0.0040	0.8004 ± 0.0017	0.7952 ± 0.0033	0.8171 ± 0.0017	0.8155 ± 0.0051
Text CKA	0.9242 ± 0.0002	0.9279 ± 0.0006	0.9275 ± 0.0004	0.9307 ± 0.0004	0.9284 ± 0.0004
Cosine between image features (T vs. S)	0.7682 ± 0.0016	0.7570 ± 0.0009	0.7546 ± 0.0010	0.7711 ± 0.0005	0.7702 ± 0.0014
Positive cosine similarity	0.2728 ± 0.0005	0.2894 ± 0.0004	0.3046 ± 0.0005	0.2622 ± 0.0005	0.2387 ± 0.0006
I2T hard gap	-0.0044 ± 0.0004	-0.0054 ± 0.0003	-0.0061 ± 0.0002	-0.0034 ± 0.0002	-0.0026 ± 0.0003
T2I hard gap	-0.0035 ± 0.0002	-0.0057 ± 0.0003	-0.0070 ± 0.0002	-0.0031 ± 0.0002	-0.0026 ± 0.0003
Row KL	0.1343 ± 0.0007	0.1139 ± 0.0005	0.1280 ± 0.0007	0.1140 ± 0.0004	0.1207 ± 0.0009
Column KL	0.1268 ± 0.0005	0.1165 ± 0.0007	0.1175 ± 0.0006	0.1156 ± 0.0004	0.1159 ± 0.0010
Row Spearman	0.8053 ± 0.0011	0.8232 ± 0.0008	0.7822 ± 0.0014	0.8340 ± 0.0005	0.8293 ± 0.0013
Column Spearman	0.7256 ± 0.0017	0.7380 ± 0.0016	0.7346 ± 0.0013	0.7508 ± 0.0004	0.7519 ± 0.0018
I2T top-64 overlap	0.4415 ± 0.0020	0.5103 ± 0.0011	0.5109 ± 0.0024	0.4887 ± 0.0021	0.4669 ± 0.0019
T2I top-64 overlap	0.4767 ± 0.0019	0.5131 ± 0.0018	0.5150 ± 0.0015	0.5004 ± 0.0006	0.5004 ± 0.0025

E.5 One-step response to importance weights

Figure 3(b) uses a ViT-T/16 student trained with $\mathcal{L}_{\mathrm{CLIP}} + \lambda_F \mathcal{L}_{\mathrm{FD}}$ and seed 0 for 16 epochs, half of the full 32-epoch schedule, and the LAION ViT-B/16 teacher. Each of eight training batches contains 1,024 image–caption pairs. For each batch, we restore identical model, optimizer, and scaler states and use the same inputs and random-number state for two updates: $\mathcal{L}_{\mathrm{base}}$ and $\mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{IW}} \mathcal{L}_{\mathrm{IW}}$, with $\lambda_{\mathrm{IW}} \approx 20.92$. Each branch takes one optimizer step.

The support H contains the teacher's 64 most similar unmatched captions per image, giving 65,536 selected pairs per batch. We calculate $W_{ij} = c_j^T(i)/c_j^S(i)$ using full column distributions at temperature 0.07, then divide by their mean on H. The value of $\log \widetilde{W}_{ij}$ before the update determines the bins. After each update, we recompute full column distributions and measure $\log c_j^{S,\mathrm{base+IW}}(i) - \log c_j^{S,\mathrm{base}}(i)$. For each bin, we first average within each batch, then report the mean and sample standard deviation of the eight batch means. Response bins have width 0.5 and at least 128 selected pairs per batch. Histogram bins have width 0.125. Both cover over 99.8% of selected pairs, with histogram percentages computed over the full support H.

The ratio $\widetilde{W}_{ij}$ is below one for $62.08 \pm 0.15\%$ of selected pairs and above one for $37.92 \pm 0.15\%$ (mean $\pm$ sample standard deviation across batches). The corresponding differences in $\log c_j^S(i)$ are -0.002874 ± 0.000169 and $+0.002887 \pm 0.000229$. For each selected pair, we measure the absolute difference $|\log c_j^T(i) - \log c_j^S(i)|$ between the teacher's and student's log probabilities, then average over H. Relative to the base update, adding $\mathcal{L}_{\mathrm{IW}}$ reduces this error by 0.002190 ± 0.000143. It also reduces $\mathcal{L}_N$ by 0.001053 ± 0.000056 and increases their Spearman correlation by 0.000994 ± 0.000067.

E.6 Controlled Updates and Short Continuations

Starting state and training batches. We first train one student with $\mathcal{L}_{\mathrm{CLIP}} + \lambda_F \mathcal{L}_{\mathrm{FD}}$ and seed 0 for 16 epochs, half of the full 32-epoch schedule. The five continuations share this model and optimizer state and differ in training order and randomness. For each continuation, we save the states reached with $\mathcal{L}_{\mathrm{base}}$ at 0, 50, …, 1,000 successful updates. From each saved state, we independently apply each objective to four training batches, restoring the model, optimizer, scaler, inputs, learning rate, and random-number state before each comparison. We compare $\mathcal{L}_{\mathrm{CLIP}} + \lambda_F \mathcal{L}_{\mathrm{FD}}$ (FD), $\mathcal{L}_{\mathrm{base}}$, and the objectives of CLIP-KD, CLIP-RD, CGD-IW, and CGD-N. The five seeds, 21 states, four batches, and six objectives give 2,520 separate diagnostic updates. Each branch is discarded after measurement, and training with $\mathcal{L}_{\mathrm{base}}$ continues from the restored state. The batch size is 1,024. CGD-IW uses the fixed $\lambda_{\mathrm{IW}} \approx 20.92$. For CGD-N, the diagnostic coefficient is recomputed each batch to set the ratio of gradient norms with respect to student features to approximately 0.03821 relative to $\mathcal{L}_{\mathrm{base}}$. This uses norms from the current batch. The controller used throughout 32-epoch training uses moving averages. Each objective retains its own coefficient, without rescaling gradient norms to match the other objectives. The fixed IW coefficient comes from Appendix D.1. The separate gradient-norm measurements in Appendix E.9 use FD states.

Measurements on COCO validation data. We reuse the same 500 COCO images and their 2,500 captions. For cosine similarity and margin measurements, queries are selected before each update: the teacher retrieves an annotated positive and the student does not. All objectives are evaluated on these same queries. For each query, p_0 and n_0 are the student's most similar annotated positive and negative before the update. Their identities remain fixed when measuring changes in their cosine similarities. The retrieval margin instead reselects the most similar positive and negative:

$$m = \max_{p \in \mathcal{P}} s(p) - \max_{n \in \mathcal{N}} s(n).$$

Here $s(i)$ is the cosine similarity between the query and candidate i. We average over queries in each direction, then give I2T and T2I equal weight. Negative-column KL is averaged over all 2,500 caption queries, excluding each caption's matched image and renormalizing the teacher and student distributions at temperature 0.07.

For any measurement x, let $\Delta x^{(m)} = x^{(m)}_{\text{after}} - x_{\text{before}}$. When comparing updates from the same state, we use paired differences $\delta x^{(m)} = \Delta x^{(m)} - \Delta x^{(\text{base})}$. For KL reduction, the sign is reversed so positive values indicate additional improvement. Differences are computed within each matched state and batch before averaging. At each retained state, we first average four batches within a seed, then compute the mean and sample standard deviation across five seeds. Table 16 additionally averages all 21 states within each seed. Its standard deviation is computed from the five means of the paired differences. States and batches are repeated measurements within each continuation.

Table 16: Difference from an update with $\mathcal{L}_{\text{base}}$ when the starting model and optimizer state are identical. We report the mean and sample standard deviation over five continuations. Within each continuation, we average the paired differences over four batches at each of 21 states.

Measurement	CGD-IW minus $\mathcal{L}_{\text{base}}$	CGD-N minus $\mathcal{L}_{\text{base}}$
Negative cosine similarity ($\times 10^{-4}$)	-1.050 ± 0.032	-1.298 ± 0.042
Positive cosine similarity ($\times 10^{-4}$)	-0.764 ± 0.023	-1.008 ± 0.053
Retrieval margin ($\times 10^{-4}$)	$+0.278 \pm 0.014$	$+0.280 \pm 0.016$
Negative-column KL reduction ($\times 10^{-3}$)	$+0.0754 \pm 0.0072$	$+0.2167 \pm 0.0113$
Queries with lower negative similarity (pp)	$+1.594 \pm 0.056$	$+1.782 \pm 0.044$

Both variants reduce similarity with the selected negative more, increase the retrieval margin more, and reduce negative-column KL more than $\mathcal{L}_{\text{base}}$ at all 105 combinations of seed and model state after averaging the four batches. The observations are correlated within each of the five continuations. Similarity with the positive still increases, but less than with $\mathcal{L}_{\text{base}}$ (Table 17). The last row reports the fraction of queries for which similarity with the selected negative decreases, using the threshold $\Delta s(n_0) < -10^{-7}$.

Table 17: Changes on the COCO subset after one update with each objective from identical states saved during training with $\mathcal{L}_{\text{base}}$. Values are means and sample standard deviations over five continuations. The first three columns use units 10^{-4}. The last uses 10^{-3}. Changes are relative to the values before the update.

Objective	Cosine similarity with negative	Cosine similarity with positive	Retrieval margin	Negative-column KL reduction
FD	-2.279 ± 0.824	0.636 ± 0.810	1.680 ± 0.255	-0.071 ± 0.146
$\mathcal{L}_{\text{base}}$	-2.056 ± 0.892	1.430 ± 0.845	2.237 ± 0.239	0.023 ± 0.148
CLIP-KD	-0.418 ± 1.246	5.978 ± 1.251	4.591 ± 0.179	0.375 ± 0.140
CLIP-RD	5.313 ± 1.555	8.485 ± 1.599	1.255 ± 0.183	0.056 ± 0.165
CGD-IW	-3.105 ± 0.867	0.665 ± 0.831	2.515 ± 0.234	0.098 ± 0.149
CGD-N	-3.354 ± 0.901	0.422 ± 0.848	2.517 ± 0.228	0.240 ± 0.151

Table 17 compares the changes produced by all six objectives. CLIP-KD produces larger increases in the retrieval margin and larger reductions in negative-column KL at these states. CLIP-KD and CLIP-RD also increase similarity with positives more, whereas CGD reduces similarity with the selected negative more.

Training for 2,000 updates. Separately, we train each objective for 2,000 successful updates from the common model trained halfway with $\mathcal{L}_{\text{CLIP}} + \lambda_F \mathcal{L}_{\text{FD}}$. Each of the six objectives is repeated

with five seeds, for 30 runs. Each run processes 2,048,000 image–caption pairs, or 0.148 epoch. The trajectories follow a fixed set of queries that the teacher answers correctly and the student gets wrong at the start (119 I2T and 638 T2I), equally weighting the two directions.

At update 2,000, paired margin-gain differences from $\mathcal{L}_{\text{base}}$ are 0.571 ± 1.438 for CGD-IW and 0.594 ± 0.926 for CGD-N (units 10^{-3}). Each is positive in four of five seeds. Their two-sided 95% t intervals are $[-1.215, 2.357] \times 10^{-3}$ and $[-0.555, 1.743] \times 10^{-3}$, both including zero ($t_4 = 0.888$, $p = 0.425$ for CGD-IW and $t_4 = 1.435$, $p = 0.224$ for CGD-N). Table 18 reports all six objectives.

Table 18: Results after 2,000 updates: means and sample standard deviations over five continuations. Margin gain uses the queries the teacher answers correctly and the student gets wrong at the start. R@1 uses all queries, and negative-column KL uses all captions. Margin and R@1 weight the two retrieval directions equally. Margin gain is measured from the common starting model.

Objective	Margin gain ($\times 10^{-3}$)	R@1 (%)	Negative-column KL ($\times 10^{-3}$)
FD	10.970 ± 2.016	54.772 ± 0.639	179.799 ± 2.254
$\mathcal{L}_{\text{base}}$	11.616 ± 1.030	55.764 ± 0.531	174.277 ± 1.297
CLIP-KD	7.954 ± 1.727	55.660 ± 0.574	232.709 ± 2.378
CLIP-RD	6.328 ± 2.830	56.096 ± 0.727	293.389 ± 2.517
CGD-IW	12.187 ± 1.375	55.768 ± 0.686	172.844 ± 2.191
CGD-N	12.210 ± 1.037	56.424 ± 0.939	169.628 ± 2.208

CGD-N has lower negative-column KL than $\mathcal{L}_{\text{base}}$ in all five seeds at update 2,000. The paired KL difference is -4.649×10^{-3} with a 95% interval of $[-6.537, -2.760] \times 10^{-3}$. The paired R@1 differences are $+0.004$ points for CGD-IW (95% confidence interval $[-0.631, 0.639]$) and $+0.660$ for CGD-N ($[-0.655, 1.975]$). Both include zero. The intervals are unadjusted for multiple comparisons. Over the full continuation, the time-averaged R@1 differences are -0.159 points for CGD-IW and $+0.138$ points for CGD-N, computed by trapezoidal integration over saved observations within each seed. Figure 5 reports changes in cosine similarity with positives and negatives separately. At update 2,000, CGD-IW reduces similarity with the initial negative by 32.634×10^{-3} and with the initial positive by 10.736×10^{-3}. CGD-N lowers them by 27.136×10^{-3} and 4.904×10^{-3}, respectively. The mean cosine similarity over negatives changes by -5.460×10^{-3} for CGD-IW and -1.134×10^{-3} for CGD-N. All these changes are measured from the starting cosine similarities before subtracting any mean. Thus similarity with positives decreases along the CGD trajectories, whereas they increase in the individual updates from restored $\mathcal{L}_{\text{base}}$ states. CLIP-KD and CLIP-RD increase similarity with both candidates during these runs.

Comparing individual updates and training trajectories. The diagnostic updates restore states saved during training with $\mathcal{L}_{\text{base}}$ and select queries before each update. The continuations instead follow each method's own model and optimizer states while evaluating a fixed set of queries.

E.7 Complementary Effects of Row and Column Matching

What is measured. We reuse the complete cosine similarity matrices from the continuations and diagnostic updates in Appendix E.6. Let $D_{ij} = S^S_{ij} - S^T_{ij}$ for $n = 500$ images and $q = 2{,}500$ captions. We separate the common mean, the mean for each image, the mean for each caption, and the remaining difference for each pair:

$$\mu = \frac{1}{nq}\sum_{i,j} D_{ij}, \qquad a_i = \frac{1}{q}\sum_{j} D_{ij} - \mu, \qquad b_j = \frac{1}{n}\sum_{i} D_{ij} - \mu,$$
$$D_{ij} = \mu + a_i + b_j + E_{ij}. \tag{51}$$

The means of a_i and b_j are zero, as are the row and column means of E. Figure 5(e,f) uses

$$\sigma_{\text{cap}} = \sqrt{\frac{1}{q}\sum_{j} b_j^2}, \qquad \sigma_{\text{img}} = \sqrt{\frac{1}{n}\sum_{i} a_i^2}.$$

These measure how unevenly the student differs from the teacher across captions and images, after averaging over the opposite modality. All entries, including matched pairs, enter both averages.

Subtracting μ does not change either standard deviation. Excluding each image's five matched captions changes the final decrease in σ_{img} from $\mathcal{L}_{\text{base}}$ to CGD-N from 1.195 to 1.200 (units 10^{-3}).

Connection to the matching directions. A constant added to one row does not change the caption distribution in that row, so $\mathcal{L}_{\text{KL}}$ alone cannot identify the differences among a_i. Those differences can change the competition between images in a column. Conversely, adding a constant to a column leaves its image distribution unchanged, while differences among b_j can change caption comparisons within rows. For caption j, positive p, and competitor i, Equation (51) gives

$$(S^S_{pj} - S^S_{ij}) - (S^T_{pj} - S^T_{ij}) = (a_p - a_i) + (E_{pj} - E_{ij}).$$

The global mean μ and the term b_j shared by both images cancel, leaving differences in their mean similarity gaps $a_p - a_i$ and pairwise residuals $E_{pj} - E_{ij}$. Proposition 2 establishes identification when the conditional distributions match exactly. Through the shared encoders, either loss can change both a_i and b_j.

Table 19: Variation in the mean difference from the teacher after 2,000 updates, in units of 10^{-3}. Values are means and sample standard deviations over five continuations from one shared model. All six objectives use the same evaluation images and captions.

Objective	Across captions, σ_{cap}	Across images, σ_{img}
FD	16.021 ± 0.630	17.208 ± 0.584
$\mathcal{L}_{\text{base}}$	13.107 ± 0.624	16.896 ± 0.218
CLIP-KD	16.086 ± 1.019	19.296 ± 0.439
CLIP-RD	16.587 ± 0.864	19.053 ± 0.583
CGD-IW	13.239 ± 0.520	16.583 ± 0.538
CGD-N	12.779 ± 0.539	15.701 ± 0.250

Effects of adding each term. Table 20 compares objectives within each continuation seed. Adding $\mathcal{L}_{\text{KL}}$ to $\mathcal{L}_{\text{CLIP}} + \lambda_F \mathcal{L}_{\text{FD}}$ reduces σ_{cap} from 16.021 to 13.107 (18.2%), while adding $\mathcal{L}_N$ to $\mathcal{L}_{\text{base}}$ reduces σ_{img} from 16.896 to 15.701 (7.1%), each in all five seeds. The percentages compare the means across seeds. CGD-IW reduces σ_{img} by 1.9%, in four of five seeds. The paired interval includes zero.

Table 20: Additional decrease after adding a loss term. Final differences (units 10^{-3}) give means and paired 95% t intervals over five seeds. Intervals are not adjusted for multiple comparisons. For one update (units 10^{-5}), values are means and sample standard deviations after averaging four batches and 21 saved states within each seed. Positive values mean a greater reduction.

	After 2,000 updates ($\times 10^{-3}$)		One update ($\times 10^{-5}$)	
Added term	Captions	Images	Captions	Images
$\mathcal{L}_{\text{KL}}$ to FD	$2.914\ [2.454, 3.375]$	$0.312\ [-0.428, 1.053]$	5.856 ± 0.388	0.529 ± 0.102
$\mathcal{L}_N$ to $\mathcal{L}_{\text{base}}$	$0.328\ [-0.146, 0.801]$	$1.195\ [0.777, 1.613]$	0.789 ± 0.084	5.300 ± 0.300
$\mathcal{L}_{\text{IW}}$ to $\mathcal{L}_{\text{base}}$	$-0.132\ [-0.589, 0.326]$	$0.313\ [-0.448, 1.073]$	-0.084 ± 0.034	1.072 ± 0.083

The diagnostic updates restore the same model and optimizer state before each objective is applied. The additional reductions in σ_{cap} from $\mathcal{L}_{\text{KL}}$ and in σ_{img} from $\mathcal{L}_N$ are positive at all 105 combinations of seed and saved state, after averaging the four batches. CGD-IW reduces σ_{img} more than $\mathcal{L}_{\text{base}}$ at 99 of these 105 combinations, and in each of the five averages over states. These measurements use states through 1,000 updates. The training trajectories retain each objective's own state through 2,000 updates.

Rescaling changes the values of σ_{cap} and σ_{img} while preserving candidate order.

E.8 Similarity Values and Candidate Margins

We use the three disjoint CC3M-val folds of 4,096 matched pairs in Appendix D.4. For each caption j retrieved correctly by the teacher, we fix its matched image $+$ and the teacher's most similar unmatched image n. This gives 6,383 captions, shared by all methods and five training runs. All checkpoints are from epoch 32.

Define $d_{ij} = S^S_{ij} - S^T_{ij}$. The matched image + has index j, so write $d_+ = d_{jj}$ and $d_n = d_{nj}$. Then

$$\bar{d}_j = \frac{d_+ + d_n}{2}, \qquad \delta_j = d_+ - d_n, \qquad d_+ = \bar{d}_j + \frac{\delta_j}{2}, \quad d_n = \bar{d}_j - \frac{\delta_j}{2}.$$

Here $\bar{d}_j$ is the difference between the student and teacher in the average similarity $(S^M_{jj} + S^M_{nj})/2$. The difference δ_j is the student's margin $S^S_{jj} - S^S_{nj}$ minus the teacher's margin $S^T_{jj} - S^T_{nj}$. Equal changes to both similarities cancel in the margin. Their order reverses when $\delta_j < -(S^T_{jj} - S^T_{nj})$. Equality gives a tie.

For each run, we pool the selected captions. The average similarity gap is $\mathbb{E}|\bar{d}_j|$, and the margin gap is $\mathbb{E}|\delta_j|$. Both are mean absolute differences from the teacher. Figure 2 plots their five-run means and sample SD, in units of 10^{-3}. Dashed lines mark the FD means. The plotted values use the original cosine similarities before rescaling.

Effect of rescaling. We reuse the positive scalars from Appendix D.4: for each method and run, fold $(f-1) \bmod 3$ supplies $a_{M,f} > 0$ for evaluation fold f. We replace S^S by $a_{M,f}S^S$, keeping the teacher and selected pairs fixed. This preserves every student's ordering but can change the absolute differences from the teacher (Table 21).

Table 21: Mean absolute differences from the teacher ($\times 10^{-3}$). The two values in each cell are the mean and sample SD over five runs.

	Before rescaling		After rescaling	
Method	$\mathbb{E}\lvert\bar{d}_j\rvert$	$\mathbb{E}\lvert\delta_j\rvert$	$\mathbb{E}\lvert\bar{d}_j\rvert$	$\mathbb{E}\lvert\delta_j\rvert$
FD	40.775 ± 0.495	41.127 ± 0.198	89.025 ± 0.464	37.680 ± 0.173
CLIP-KD	28.118 ± 0.097	44.101 ± 0.239	82.025 ± 0.248	37.651 ± 0.211
CLIP-RD	27.550 ± 0.119	47.916 ± 0.354	80.433 ± 0.364	38.479 ± 0.144
CGD-N	48.317 ± 0.547	39.552 ± 0.193	74.878 ± 0.599	37.585 ± 0.129
CGD-IW	70.355 ± 0.585	38.051 ± 0.289	83.097 ± 0.698	37.188 ± 0.264

After rescaling, CGD-IW and CGD-N still have the lowest mean margin differences. CLIP-KD and FD exchange positions and are nearly tied. CGD-N versus CLIP-KD changes from 4.55 to 0.066, and CGD-IW versus CLIP-RD from 9.87 to 1.29 (all $\times 10^{-3}$). The separation before rescaling therefore depends partly on similarity scale.

E.9 Gradient Directions and Magnitudes at Saved Training States

States and inputs. We examine the FD continuation from the common epoch-16 ViT-T/16 student and ViT-B/16 teacher used in Appendix E.6. Its initial state and the states after 1,000 and 2,000 updates for five continuation seeds give 11 distinct states. The initial state is shared. Two fixed batches of 1,024 pairs use the same forward pass in deterministic FP32 evaluation mode for all losses, without optimizer updates.

Directions and cross terms. For encoder parameters θ, let $h_m = \nabla_\theta(\lambda_m \mathcal{L}_m)$. For the task, FD, row KL, IW, N, and CLIP-KD/RD components, define

$$G_{ab} = h_a^\top h_b, \qquad C_{ab} = \frac{G_{ab}}{\sqrt{G_{aa}G_{bb}}}, \qquad \Big\|\sum_m h_m\Big\|_2^2 = \sum_m G_{mm} + 2\sum_{a<b} G_{ab}.$$

C_{ab} uses uncentered gradient vectors. The final identity separates individual squared norms G_{mm} from cross terms $2\sum_{a<b} G_{ab}$. Figure 6 averages the batch and seed means equally over the three update counts, giving the shared initial state the same weight as each later count.

IW and N are more aligned with each other than with FD or row KL, but their directions differ: their mean cosine is 0.737, 0.690, and 0.665 at the three states. FD and row KL have cosines 0.277, 0.299, and 0.340.

Gradient norms. Table 22 reports $\|h_m\|_2/\|\nabla_\theta \mathcal{L}_{\text{base}}\|_2$ for both encoders together. The coefficients are $\lambda_F = 2{,}000$ and $\lambda_{\text{KL}} = 1$. IW uses the continuation coefficient 20.92. For N, each fixed



Figure 8: Cosines between encoder gradients at identical FD states and inputs. KL denotes the row term $\mathcal{L}_{\mathrm{KL}}$. At 1,000 and 2,000 updates, we average the two batches within each continuation seed, then average the five seeds. At zero, only the two batches are averaged.

Table 22: Encoder-gradient norm as a percentage of the norm of $\nabla_\theta \mathcal{L}_{\mathrm{base}}$. Values at 1,000 and 2,000 updates are means and sample SD over five continuation seeds after averaging two batches. The initial value averages two batches from one state.

Component or sum	0	1,000	2,000
$\lambda_F \mathcal{L}_{\mathrm{FD}}$	29.20	32.66 ± 4.81	29.26 ± 1.56
$\lambda_{\mathrm{KL}} \mathcal{L}_{\mathrm{KL}}$	9.33	9.43 ± 0.62	8.69 ± 0.31
$\lambda_{\mathrm{IW}} \mathcal{L}_{\mathrm{IW}}$	6.04	5.62 ± 0.40	5.13 ± 0.15
$\lambda_N \mathcal{L}_N$	5.48	5.05 ± 0.24	4.98 ± 0.15
CLIP-KD relational + ICL	93.98	96.76 ± 4.01	93.34 ± 3.35
CLIP-RD relational + ICL + VRD + XRD	101.69	102.30 ± 3.71	99.45 ± 3.51

state and batch uses the feature-gradient budget 0.0382. The coefficient is recomputed at each FD state and batch. CLIP-KD/RD retain their released component coefficients.

$\mathcal{L}_{\mathrm{IW}}$ and $\mathcal{L}_N$ have weighted gradient norms approximately 5–6% of $\|\nabla_\theta \mathcal{L}_{\mathrm{base}}\|_2$.

Projection onto fixed candidate comparisons. At the common initial state, we select the most similar unmatched candidate for each query and retain its identity at all later states. For its margin $m = S_{\mathrm{pos}} - S_{\mathrm{neg}}$, the directional derivative under the component's Euclidean descent direction is

$$p_m = -\nabla_\theta m^\top h_m = -\nabla_\theta S_{\mathrm{pos}}^\top h_m + \nabla_\theta S_{\mathrm{neg}}^\top h_m.$$

We average across queries, keeping the positive and negative terms separate. Both components lower the two similarities on average. They lower the competitor more initially, giving positive mean margin derivatives. At 1,000 and 2,000 updates the derivatives are small and negative (Figure 9). I2T shows the same sign change.

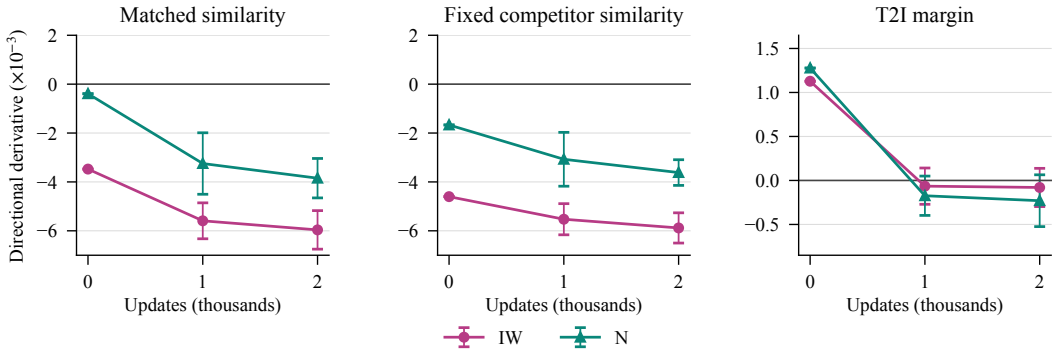


Figure 9: Mean T2I score and margin derivatives along the applied IW and N encoder-gradient directions. Candidates are fixed at the initial state. Error bars at 1,000 and 2,000 updates are sample SD over five continuation seeds. The initial point averages two batches from one state.

Scope. These Euclidean derivatives omit Adam momentum and preconditioning. Appendix E.6 measures optimizer steps.

E.10 Representation and Gradient Measurements

These results supplement Figure 5, using the same five continuations, common starting model, and COCO subset as Appendix E.6. FD+KL denotes $\mathcal{L}_{\text{base}}$, including the shared task loss $\mathcal{L}_{\text{CLIP}}$.

Representation measurements. Alignment and centered linear CKA use the definitions in Appendix D.4. CKA and feature centroids are measured at six saved states. The other curves use 30. Equation (51) averages the similarity gap over captions for each image. The SD of these means is $\text{SD}_i(a_i)$, and interaction error is $\sqrt{\text{mean}_{ij}\, E_{ij}^2}$, including matched pairs.

Table 23: Additional error reduction relative to $\mathcal{L}_{\text{base}}$ after one optimizer update ($\times 10^{-5}$), corresponding to Figure 5(j). Values are means and sample SD over five continuation means. Positive values indicate a greater reduction.

Objective	SD of image means, $\text{SD}_i(a_i)$	Interaction RMS
FD	-0.529 ± 0.102	-1.558 ± 0.040
$\mathcal{L}_{\text{base}}$	0.000 ± 0.000	0.000 ± 0.000
CLIP-KD	-2.762 ± 0.846	2.649 ± 0.114
CLIP-RD	-3.035 ± 1.016	-13.187 ± 0.375
CGD-IW	1.072 ± 0.083	1.453 ± 0.053
CGD-N	5.300 ± 0.300	1.496 ± 0.058

Same-state objective contributions. Table 23 applies these errors to the 2,520 one-update branches in Appendix E.6, averaging 84 paired comparisons per continuation.

Results. CGD preserves pointwise teacher alignment while reducing interaction error along these continuations. RD has higher mean image CKA despite lower pointwise cosine. In matched-state updates, CGD-N has positive extra reductions in both $\text{SD}_i(a_i)$ and interaction RMS. CLIP-KD reduces interaction error more in one update and also has a larger one-update retrieval-margin gain.

Recorded gradients along the learned trajectories. Along their own trajectories, IW and N have feature-gradient norms of $3.351 \pm 0.011\%$ and $3.850 \pm 0.001\%$ relative to $\mathcal{L}_{\text{CLIP}} + \lambda_F \mathcal{L}_{\text{FD}}$, and cosines of -0.009 ± 0.003 and 0.202 ± 0.001 with $\mathcal{L}_{\text{KL}}$. We average common recorded updates within each continuation, then report the mean and sample SD across five continuations. N targets a 3.82% norm ratio relative to $\mathcal{L}_{\text{base}}$, so its stable magnitude reflects coefficient control. Figure 11 averages trajectory curves within 100-update bins. Gradients are with respect to student output features.

F Limitations

The experiments cover the evaluated teacher, training data, and image–text tasks. Main results compare five runs per training objective. Student architectures and the four-epoch component study have one run per setting. The individual updates and continuations share one starting model trained with $\mathcal{L}_{\text{CLIP}} + \lambda_F \mathcal{L}_{\text{FD}}$ for half of the 32-epoch schedule. Their five seeds measure variation in the continuation, with the starting model held fixed. Linking these responses to the gains after 32 epochs requires further study. Method ordering varies across architectures and datasets, and the two CGD variants use different schedules for their additional loss coefficients.

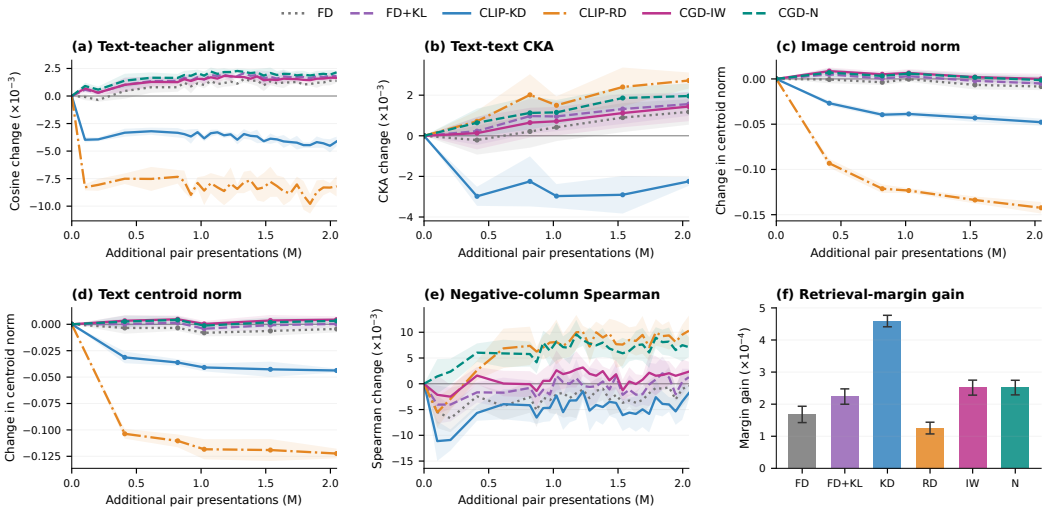


Figure 10: **Text representations and retrieval margins.** (a–e) Changes from the initial state. (f) One-update margin gains on queries where the teacher retrieves a correct match but the student does not, with equal I2T/T2I weight. Lines and bars show means and sample SD across five continuations.

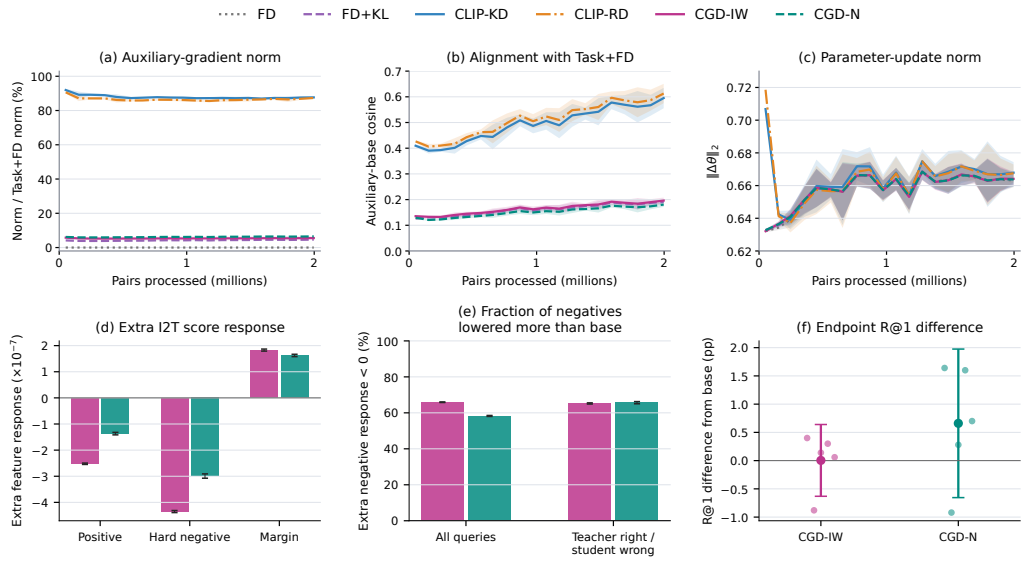


Figure 11: **Gradient magnitudes, local responses, and endpoint recall.** (a,b) Auxiliary feature-gradient norms and cosines with $\mathcal{L}_{\mathrm{CLIP}} + \lambda_F \mathcal{L}_{\mathrm{FD}}$. (c) Parameter-update norms. (d,e) Extra infinitesimal output-feature responses relative to $\mathcal{L}_{\mathrm{base}}$ on training I2T batches at updates 801–2,000. (f) Endpoint R@1 differences from $\mathcal{L}_{\mathrm{base}}$ with paired 95% t intervals. Other bands and bars show sample SD over five continuations.