Intra-Modal Divergence-Weighted Distillation for Vision-Language Models

Youva Addad
youva.addad@unicaen.fr
Alexis Lechervy
alexis.lechervy@unicaen.fr
Frederic Jurie
frederic.jurie@unicaen.fr

GREYC, Normandy University,
UNICAEN, ENSICAEN,
UMR CNRS 6072, France

Abstract

Large vision-language models like CLIP offer strong zero-shot capabilities but are computationally demanding. Knowledge distillation is crucial for creating efficient student models; however, effectively transferring the teacher's nuanced understanding of within-modality relationships, especially among negative examples, remains challenging. We introduce a novel distillation method focused on capturing the teacher's intra-modal relational knowledge. Our approach employs Kullback-Leibler divergence to measure the disagreement between student and teacher pairwise similarity distributions within each modality. This disagreement score then dynamically weights the distillation loss, compelling the student to prioritize learning from samples exhibiting the most significant relational discrepancies. This strategy encourages closer alignment of the student's internal representation space with the teacher's. Experiments demonstrate our method produces performant and efficient student models by effectively transferring this vital relational information. The source code will be made publicly available.

1 Introduction

Vision-Language Models (VLMs) like CLIP [21] have revolutionized visual representation learning, demonstrating remarkable zero-shot capabilities across diverse tasks such as classification [7, 18, 23], object detection [1, 29], and visual question answering [3, 4]. By learning from vast quantities of web-sourced image-text pairs [13], these models align visual and textual modalities in a shared embedding space, reducing reliance on costly manual annotations [5, 27]. However, the impressive performance of VLMs often comes at a significant computational cost due to their large scale and reliance on massive datasets [21], limiting their accessibility and practical deployment, particularly in resource-constrained settings.

Knowledge Distillation (KD) [11] has emerged as a promising technique to mitigate these challenges. By transferring knowledge from a large, pre-trained teacher model (e.g., CLIP) to a smaller, more efficient student model, KD enables the creation of lightweight yet powerful models [28, 33, 34]. Existing KD methods for VLMs have explored various strategies, including feature mimicking [34], cross-modal affinity alignment [33], and distilling contrastive objectives [20].

Despite these advances, a critical aspect often underexplored is the comprehensive transfer of the teacher's nuanced understanding of *within-modality* relationships. Powerful VLMs like CLIP not only align images and text but also develop a sophisticated internal representation of semantic similarities and dissimilarities among samples within the same modality (e.g., image-to-image or text-to-text). This relational knowledge, particularly concerning the subtle distinctions between negative examples or among related positive examples, is encoded in the teacher's internal similarity structures but is not fully captured by distillation methods focusing primarily on cross-modal alignment or direct feature matching.

This paper introduces a novel knowledge distillation method designed to specifically address this gap by transferring the teacher's rich within-modality relational knowledge. Our approach leverages the Kullback-Leibler (KL) divergence to quantify the discrepancy between the student's and teacher's pairwise similarity distributions within each modality (image-to-image and text-to-text) for samples within a batch. This discrepancy measure then dynamically weights the distillation loss, compelling the student to prioritize learning from samples where its understanding of these internal relationships significantly deviates from the teacher's. By focusing on these challenging relational structures, especially among negative examples, our method encourages the student to develop a more teacher-aligned and well-structured embedding space.

Our contributions are: (1) A novel knowledge distillation strategy that emphasizes the transfer of the teacher's within-modality relational knowledge by matching pairwise similarity distributions. (2) A dynamic weighting mechanism for the distillation loss, guided by the KL divergence between student and teacher intra-modal similarity distributions, to focus learning on challenging relational patterns. (3) The empirical validation demonstrating that our method trains performant and efficient student models that learn a richer, more teacher-aligned representation.

2 Related Work

2.1 Vision-Language Models

The landscape of visual understanding has been profoundly reshaped by Vision-Language Models (VLMs) [39]. Seminal works like CLIP [21] and ALIGN [13] established the efficacy of contrastive pre-training on web-scale image-text data, enabling powerful zero-shot generalization by aligning visual and textual representations in a common embedding space. Subsequent research has explored diverse VLM architectures and training paradigms. For instance, Florence [37] and FLAVA [26] aimed to build more comprehensive foundational models. Others, like DeCLIP [14], investigated data-efficient contrastive learning, while UniCL [35] proposed unifying contrastive learning across image, text, and label spaces. Techniques such as LiT [38] focused on efficient transfer learning strategies for large pre-trained VLMs. While these models exhibit remarkable capabilities, their sheer scale and computational demands [21] necessitate methods for creating more efficient variants, prominently including knowledge distillation [8, 11].

2.2 Knowledge Distillation for VLMs

Knowledge Distillation (KD), first introduced by Hinton et al. [11], aims to transfer the "dark knowledge" from a large teacher model to a smaller student model. This is typically achieved by training the student to mimic the teacher's softened softmax outputs (logits) or intermediate feature representations [8]. Adapting KD to the multimodal and complex nature of VLMs like CLIP [21] presents unique challenges and has spurred specialized techniques.

Existing VLM distillation approaches can be broadly categorized by the type of knowledge they aim to transfer: **Cross-Modal Alignment Distillation:** Many methods focus on ensuring the student learns the teacher's ability to align images and text. TinyCLIP [33], for example, employs affinity mimicking for cross-modal distillation and weight inheritance, though the latter often requires architectural similarity. Other works distill the cross-modal contrastive loss itself, encouraging the student's image-text similarity scores to match the teacher's [20]. **Feature-Level Mimicry:** CLIP-KD [34] provides an empirical study demonstrating the effectiveness of making the student's feature embeddings (for both image and text encoders) directly match those of the teacher, typically using L2 loss. **Logit-Based Distillation:** Standard KD techniques applied to the final similarity scores (logits) between image-text pairs can also be employed, where the student learns to replicate the teacher's distribution over potential matches. **Specialized Distillation Strategies:** Other works address specific scenarios. DIME-FM [28] focuses on distilling VLMs with limited unpaired data, while PromptKD [15] introduces an unsupervised framework using prompts for domain-specific distillation.

While these methods have achieved considerable success in compressing VLMs, they often prioritize the transfer of cross-modal alignment signals or direct feature/logit matching. The rich, nuanced relational knowledge *within* each modality (e.g., how similar one image is to other images, or one text to other texts, according to the teacher) is often not explicitly targeted. This intra-modal relational understanding, especially regarding subtle differences among negative samples or fine-grained similarities among positive ones, is crucial for robust representation learning. Our work distinguishes itself by directly focusing on distilling these intra-modal similarity distributions, compelling the student to learn a more comprehensive and teacher-aligned internal representational structure. By dynamically weighting the distillation based on discrepancies in these intra-modal relationships, we specifically target areas where the student's understanding is weakest compared to the teacher's sophisticated internal landscape.

3 Method

3.1 Overview of CLIP

CLIP [21] is a vision-language model that learns visual representations through natural language supervision by utilizing contrastive learning. It operates on a dataset of image-text pairs, denoted as $D = \{(I_k, T_k)\}_{k=1}^{|D|}$, where I_k represents an image and T_k is its corresponding textual description. The model employs two distinct encoders: an image encoder $f(\cdot)$ and a text encoder $g(\cdot)$. These encoders independently map images and texts into a shared multimodal embedding space. For an image I_k and text T_k, their normalized embeddings are:

$$\mathbf{v}_k = \frac{f(I_k)}{\|f(I_k)\|_2}, \quad \mathbf{u}_k = \frac{g(T_k)}{\|g(T_k)\|_2}$$

Here, $\mathbf{v}_k$ and $\mathbf{u}_k$ represent the normalized embeddings of the image and text, respectively.

The model is optimized using two separate contrastive losses. For a batch of N image-text pairs, the image-to-text loss, $\mathcal{L}_{\text{image-to-text}}$, encouraging image embeddings to align with

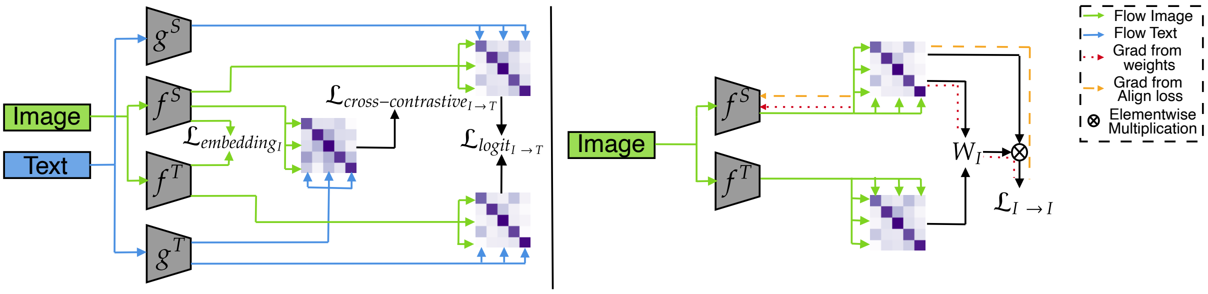

Figure 1: Flow diagram for the Image modality; a symmetric design applies to Text. Here, f and g are the encoders for Image and Text, respectively, and S and T denote the Student and Teacher components. The left side illustrates the standard loss flow, while the right side shows the intra-KD loss flow. Gradient paths from weights and alignment loss are indicated, with elementwise multiplication representing sample weighting.

their corresponding text embeddings, is defined as:

$$\mathcal{L}_{\text{image-to-text}} = \frac{1}{N}\sum_{k=1}^{N}\left[-\log\frac{\exp(\text{sim}(\mathbf{v}_k,\mathbf{u}_k)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{v}_k,\mathbf{u}_j)/\tau)}\right] \tag{1}$$

Similarly, the text-to-image loss, $\mathcal{L}_{\text{text-to-image}}$, ensures text embeddings align with their corresponding image embeddings:

$$\mathcal{L}_{\text{text-to-image}} = \frac{1}{N}\sum_{k=1}^{N}\left[-\log\frac{\exp(\text{sim}(\mathbf{u}_k,\mathbf{v}_k)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{u}_k,\mathbf{v}_j)/\tau)}\right] \tag{2}$$

In these formulations, $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity between vectors, and τ is a learnable temperature parameter controlling the concentration of the similarity distribution. The total CLIP loss [21] is the average of these two:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}\left(\mathcal{L}_{\text{image-to-text}} + \mathcal{L}_{\text{text-to-image}}\right) \tag{3}$$

Optimizing $\mathcal{L}_{\text{CLIP}}$ encourages both image-to-text ($I \to T$) and text-to-image ($T \to I$) alignment, enabling the learning of a shared multimodal embedding space where matching pairs are pulled together and mismatched pairs are pushed apart.

3.2 Background: Common Knowledge Distillation Losses for CLIP

To create smaller, efficient student models from large teacher CLIP models, various knowledge distillation (KD) losses are often employed. These aim to transfer different facets of the teacher's knowledge. Let $(\mathbf{v}_k^T,\mathbf{u}_k^T)$ be embeddings from the teacher model and $(\mathbf{v}_k^S,\mathbf{u}_k^S)$ be from the student model.

Embedding Alignment Loss: Directly aligns student embeddings with teacher embeddings:

$$\mathcal{L}_{\text{embedding}_I} = \frac{1}{N}\sum_{k=1}^{N}||\mathbf{v}_k^T - \mathbf{v}_k^S||_2^2 \tag{4}$$

$$\mathcal{L}_{\text{embedding}_T} = \frac{1}{N}\sum_{k=1}^{N}||\mathbf{u}_k^T - \mathbf{u}_k^S||_2^2 \tag{5}$$

$$\mathcal{L}_{\text{embedding}} = \mathcal{L}_{\text{embedding}_I} + \mathcal{L}_{\text{embedding}_T} \tag{6}$$

Logit Distillation Loss: Matches the student's output probability distributions (derived from image-text similarity logits) to the teacher's. Let $P^M_{I\to T,k}(\cdot)$ be the probability distribution for image $\mathbf{v}^M_k$ over all text embeddings $\{\mathbf{u}^M_j\}_{j=1}^N$ in the batch for model $M \in \{S,T\}$, and $P^M_{T\to I,k}(\cdot)$ be for text $\mathbf{u}^M_k$ over image embeddings. Specifically, $P^M_{I\to T,k}(j) = \frac{\exp(\text{sim}(\mathbf{v}^M_k,\mathbf{u}^M_j)/\tau_M)}{\sum_{l=1}^N \exp(\text{sim}(\mathbf{v}^M_k,\mathbf{u}^M_l)/\tau_M)}$, and similarly for $P^M_{T\to I,k}(j)$. The loss is:

$$\mathcal{L}_{\text{logit}_{I\to T}} = \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N} \text{KL}(P^T_{I\to T,k}(j)||P^S_{I\to T,k}(j)) \tag{7}$$

$$\mathcal{L}_{\text{logit}_{T\to I}} = \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{N} \text{KL}(P^T_{T\to I,k}(j)||P^S_{T\to I,k}(j)) \tag{8}$$

$$\mathcal{L}_{\text{logit}} = \frac{1}{2}(\mathcal{L}_{\text{logit}_{I\to T}} + \mathcal{L}_{\text{logit}_{T\to I}}) \tag{9}$$

where KL denotes Kullback-Leibler divergence, defined by the formula in Equation (10), and τ_S, τ_T are student/teacher temperatures.

$$\text{KL}(P^T_{I\to T,k}(j)||P^S_{I\to T,k}(j)) = P^T_{I\to T,k}(j)\log\frac{P^T_{I\to T,k}(j)}{P^S_{I\to T,k}(j)} \tag{10}$$

Cross-Contrastive Distillation Loss: Guides the student to learn cross-modal relationships by contrasting student embeddings of one modality against teacher embeddings of the other:

$$\mathcal{L}_{\text{cross-contrastive}_{I\to T}} = \frac{1}{N}\sum_{k=1}^{N}\left[-\log\frac{\exp(\text{sim}(\mathbf{v}^S_k,\mathbf{u}^T_k)/\tau_D)}{\sum_{l=1}^N \exp(\text{sim}(\mathbf{v}^S_k,\mathbf{u}^T_l)/\tau_D)}\right]. \tag{11}$$

$$\mathcal{L}_{\text{cross-contrastive}_{T\to I}} = \frac{1}{N}\sum_{k=1}^{N}\left[-\log\frac{\exp(\text{sim}(\mathbf{u}^S_k,\mathbf{v}^T_k)/\tau_D)}{\sum_{l=1}^N \exp(\text{sim}(\mathbf{u}^S_k,\mathbf{v}^T_l)/\tau_D)}\right] \tag{12}$$

$$\mathcal{L}_{\text{cross-contrastive}} = \frac{1}{2}(\mathcal{L}_{\text{cross-contrastive}_{I\to T}} + \mathcal{L}_{\text{cross-contrastive}_{T\to I}}) \tag{13}$$

where τ_D is a distillation temperature. These losses, often combined with $\mathcal{L}_{\text{CLIP}}$ for the student, form the basis of many CLIP distillation strategies.

3.3 Proposed Method: Distilling Intra-Modal Relational Knowledge

Our core idea as depicted in Figure 1 is to explicitly transfer the teacher's understanding of the relational structure within each modality. The student network is encouraged to align its intra-modal pairwise cosine similarities with those of the teacher. This alignment is encouraged by first defining a student self-consistency objective within each modality (e.g., Equation 17). We then introduce an adaptive weighting mechanism (Equation 21) where the weights are derived from the KL divergence between the student's and teacher's full intra-modal similarity distributions. This directs the student to prioritize learning its self-consistency for samples where its overall intra-modal view, including off-diagonal relationships, significantly deviates from the teacher's "relational map", ensuring efficient and targeted knowledge transfer.

3.3.1 Intra-Modal Similarity Distributions

For a batch of N images with embeddings $\{\mathbf{v}_k^M\}_{k=1}^N$ and N texts with embeddings $\{\mathbf{u}_k^M\}_{k=1}^N$ from model $M \in \{S,T\}$, we define intra-modal similarity distributions.

The image-to-image similarity distribution for the k-th image in model M is a probability distribution $P_{I\to I,k}^M(\cdot)$ over all other images j in the batch:

$$P_{I\to I,k}^M(j) = \frac{\exp(\mathrm{sim}(\mathbf{v}_k^M,\mathbf{v}_j^M)/\tau_{\text{intra}})}{\sum_{l=1}^N \exp(\mathrm{sim}(\mathbf{v}_k^M,\mathbf{v}_l^M)/\tau_{\text{intra}})} \tag{14}$$

Similarly, the text-to-text similarity distribution for the k-th text in model M is $P_{T\to T,k}^M(\cdot)$:

$$P_{T\to T,k}^M(j) = \frac{\exp(\mathrm{sim}(\mathbf{u}_k^M,\mathbf{u}_j^M)/\tau_{\text{intra}})}{\sum_{l=1}^N \exp(\mathrm{sim}(\mathbf{u}_k^M,\mathbf{u}_l^M)/\tau_{\text{intra}})} \tag{15}$$

τ_{intra} is a learnable temperature parameter specific to intra-modal distillation, which can be different for teacher and student, or shared. These distributions capture how model M perceives the similarity of sample k to all other samples of the same modality within the batch. The fact that $(\mathrm{sim}(\mathbf{v}_k^M,\mathbf{v}_k^M) = 1)$ and $(\mathrm{sim}(\mathbf{u}_k^M,\mathbf{u}_k^M) = 1)$ signifies that each sample exhibits maximum similarity to itself. This principle is pertinent to the discussion in the subsequent section.

3.3.2 Intra-Modal Relational Distillation Loss ($\mathcal{L}_{\text{intra-KD}}$)

We propose a new distillation loss, $\mathcal{L}_{\text{intra-KD}}$, to align the student's intra-modal similarity distributions with the teacher's. Our approach is designed to encourage the student's image-to-image and text-to-text similarity distribution to better match those of the teacher by minimizing the KL divergence between their respective similarity matrices. By applying a temperature-scaled softmax and using these divergences to compute adaptive weights, the model gives more importance to samples with lower discrepancy, helping the student focus on accurately capturing the relational structure within each modality.

$$\mathcal{L}'_{I\to I} = \frac{1}{N}\sum_{k=1}^N \left[-\log P_{I\to I,k}^S(k)\right] \tag{16}$$

$$= \frac{1}{N}\sum_{k=1}^N \left[-\log \frac{\exp\left(\mathrm{sim}(\mathbf{v}_k^S,\mathbf{v}_k^S)/\tau_{\text{intra}}\right)}{\exp\left(\mathrm{sim}(\mathbf{v}_k^S,\mathbf{v}_k^S)/\tau_{\text{intra}}\right) + \sum_{l\neq k}\exp\left(\mathrm{sim}(\mathbf{v}_k^S,\mathbf{v}_l^S)/\tau_{\text{intra}}\right)}\right] \tag{17}$$

Equation (17) defines a loss that encourages the predicted probability $P_{I\to I,k}^S(k)$ to be close to 1, while driving the probabilities $P_{I\to I,k}^S(j)$ for $j \neq k$ towards 0. Since $\mathrm{sim}(\mathbf{v}_k^S,\mathbf{v}_k^S) = 1$, the numerator of the softmax expression—i.e., the logit for the positive pair—is constant across steps. Therefore, increasing $P_{I\to I,k}^S(k)$ requires minimizing the denominator:

$$\min \left(\exp\left(\mathrm{sim}(\mathbf{v}_k^S,\mathbf{v}_k^S)/\tau_{\text{intra}}\right) + \sum_{l\neq k}\exp\left(\mathrm{sim}(\mathbf{v}_k^S,\mathbf{v}_l^S)/\tau_{\text{intra}}\right)\right) \tag{18}$$

Since the first term is fixed, minimizing the denominator effectively reduces to minimizing the sum of exponentiated similarities between non-matching pairs:

$$\min \sum_{l \neq k} \exp\left(\mathrm{sim}(\mathbf{v}_k^S, \mathbf{v}_l^S)/\tau_{\text{intra}}\right) \tag{19}$$

Optimizing Equation (19) drives the off-diagonal terms $P^S_{I\to I,k}(j)$ for $j \neq k$ toward zero. This is achieved by pushing the cosine similarities of non-matching pairs toward -1 (i.e., maximum dissimilarity). Thus, the cross-entropy loss implicitly encourages large negative similarity values for all non-corresponding pairs.

$$\mathcal{L}'_{T\to T} = \frac{1}{N}\sum_{k=1}^{N}\left[-\log P^S_{T\to T,k}(k)\right] \tag{20}$$

Initially, the intra-modality losses defined in Equations (17) and (20) are formulated based on the student's self-similarity within a modality. In this form, the student is not yet directly guided by the teacher's intra-modal understanding for these specific loss components. To incorporate the teacher's relational knowledge and focus the student's learning, we introduce an adaptive weighting mechanism based on the KL divergence between the student's and teacher's intra-modal similarity distributions. This can be viewed as a form of curriculum learning, where the student prioritizes learning relational patterns that it has not yet mastered relative to the teacher.

$$\begin{aligned} K_I &= \left[\sum_{j=1}^{N} \mathrm{KL}\left(P^T_{I\to I,k}(j) \,\|\, P^S_{I\to I,k}(j)\right)\right]_{k=1}^{N}, \quad \text{where } K_I \in \mathbb{R}^N \\ W_I &= Softmax(K_I/c), \quad \text{where } W_I \in \mathbb{R}^N \end{aligned} \tag{21}$$

Equation (21) defines the calculation of sample-wise importance weights W_I (and similarly W_T). First, K_I quantifies the total KL divergence between the student's P^S and teacher's P^T similarity-derived distributions for each sample k within the batch. W_I is then computed by applying a softmax to K_I scaled by $1/c$, where c is a hyperparameter controlling the smoothness of the softmax output. These weights (W_I, W_T) reflect sample importance. Because the diagonal elements of teacher and student similarity matrices are fixed pre-softmax, differences in their distributions arise only from off-diagonal entries (similarities between distinct samples). The KL divergence thus aligns these off-diagonal similarities. Higher weights (W_I or W_T) are assigned to samples where the student's relational structure deviates most from the teacher's (i.e., high K_I), directing focus to these "difficult" examples ("most wrong" areas), akin to hard negative mining. Crucially, W_I and W_T remain differentiable and part of the gradient flow during training.

$$\begin{aligned} \mathcal{L}_{I\to I} &= \sum_{k=1}^{N} W_{I,k}\cdot\left[-\log P^S_{I\to I,k}(k)\right] \\ \mathcal{L}_{T\to T} &= \sum_{k=1}^{N} W_{T,k}\cdot\left[-\log P^S_{T\to T,k}(k)\right] \end{aligned} \tag{22}$$

$$\mathcal{L}_{\text{intra-KD}} = \mathcal{L}_{I\to I} + \mathcal{L}_{T\to T} \tag{23}$$

Table 1: Ablation analysis of intra-modal distillation components using a ViT-T/16 student and a LAION-400M pretrained ViT-B/16 teacher, with CC12M+CC3M as the distillation corpus. Results are presented for classification (ImageNet variants) and cross-modal retrieval (CC3M, MSCOCO, Flickr).

Method	**IN**	**INV2**	**IN-R**	**IN-S**	**CC3M Val**		**MSCOCO**		**Flickr**	
	Acc	Acc	Acc	Acc	I2T	T2I	I2T	T2I	I2T	T2I
T: ViT-B/16	67.1	59.6	77.9	52.37	43.8	42.3	39.5	36.5	76.9	75.5
S: ViT-T/16	30.5	25.6	35.7	17.3	33.3	33.5	20.7	20.3	46.4	47.7
+ $\mathcal{L}_{\text{standard}}$	41.8	36.0	46.9	26.1	37.2	36.0	26.7	25.8	59.3	57.7
+ $\mathcal{L}_{\text{intra-KD}}$	**43.3**	**37.1**	**49.6**	**27.8**	38.2	36.4	28.2	**26.3**	60.4	**60.1**
w/ Uniform Weights	42.1	36.4	48.3	27.1	37.9	36.5	28.1	26.2	60.3	57.1
Weights w/ No Grad	42.9	37.0	49.5	27.1	38.2	36.4	27.9	25.7	60.2	58.9
Weighted w/ $\mathcal{L}_{\text{CLIP}}$	42.6	36.7	48.9	27.6	**38.6**	**36.7**	27.9	25.6	59.9	58.1
w/o $\mathcal{L}_{I \to I}$	42.4	36.6	47.8	26.8	37.6	36.4	27.9	**26.3**	59.2	59.2
w/o $\mathcal{L}_{T \to T}$	42.7	36.7	48.8	27.2	38.0	35.9	**28.4**	25.4	**61.1**	58.8

3.3.3 Overall Training Objective

The student model is trained by minimizing a combined loss function. This includes the standard CLIP contrastive loss $\mathcal{L}^S_{\text{CLIP}}$ (Eq. 3) applied to student model outputs and ground-truth pairs), our proposed intra-modal relational distillation loss $\mathcal{L}_{\text{intra-KD}}$, and the standard distillation losses, $\mathcal{L}_{\text{embedding}}(Eq. 6)$, $\mathcal{L}_{\text{logit}}$ (Eq. 9), or $\mathcal{L}_{\text{cross-contrastive}}$ (Eq. 13)).

$$\mathcal{L}_{\text{standard}} = \alpha \cdot \mathcal{L}_{\text{embedding}} + \beta \cdot \mathcal{L}_{\text{logit}} + \gamma \cdot \mathcal{L}_{\text{cross-contrastive}} \tag{24}$$

The overall objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}^S_{\text{CLIP}} + \mathcal{L}_{\text{standard}} + \delta \cdot \mathcal{L}_{\text{intra-KD}} \tag{25}$$

where α, β γ, and δ are hyperparameters balancing the contributions of the different loss terms. Through $\mathcal{L}_{\text{intra-KD}}$, our method transfers crucial within-modality relational knowledge, leading to more robust and teacher-aligned student models.

4 Experiments

We evaluate our intra-modal KD approach on classification (ImageNet (IN) [5], IN-V2 [22], IN-R [10], IN-S [32]) and retrieval (CC3M val, MSCOCO [16], Flickr30K [36]). We analyze loss contributions, training configurations, and teacher-student pairings. Appendices A, B provide full experimental details. The teacher is a LAION-400M [24] pretrained encoder; distillation uses CC12M [2] + CC3M [25]. Further results are presented in Appendix C.

4.1 Ablation Study

Our ablation study (Table 1) analyzes each component's impact on the ViT-T/16 student. The baseline student's 30.5% ImageNet top-1 accuracy is lifted by 11.3% (to 41.8%) with standard distillation ($\mathcal{L}_{\text{standard}}$). This also substantially boosts R@1 retrieval scores by 3.9% on CC3M, 6.0% on MSCOCO, and 12.9% on Flickr for I2T tasks.

Crucially, our proposed intra-modal loss ($\mathcal{L}_{\text{intra-KD}}$), incorporating adaptive KL-weighting, further improves ImageNet top-1 accuracy by an additional 1.5% (to 43.3%) and enhances

Table 2: Zero-shot performance comparison of the proposed method ('Our') against baselines (TinyCLIP, CLIP-KD) across different student architectures (ViT-T/16, ViT-B/16, ResNet-50) and teacher models on ImageNet-1K classification and MSCOCO/Flickr retrieval. Teacher models were pretrained on LAION-400M, and distillation utilized the CC12M+CC3M dataset.

Method	**IN-1K**	**MSCOCO**		**Flickr**	
	Acc	I2T	T2I	I2T	T2I
T_1: ViT-L/14	72.8	42.7	40.9	80.5	79.5
T_2: ViT-B/16	67.1	39.5	36.5	76.9	75.5
S: ViT-T/16	30.6	20.7	20.3	46.4	47.7
+TinyCLIP (from T_1)	39.3	26.4	24.1	57.6	57.4
+TinyCLIP (from T_2)	40.8	26.8	24.7	58.6	58.5
+CLIP-KD (from T_1)	40.9	27.2	25.5	59.7	59.7
+CLIP-KD (from T_2)	42.6	28.1	26.0	**60.4**	59.9
+Our (from T_1)	40.9	27.3	25.8	59.8	60.0
+Our (from T_2)	**43.3**	**28.2**	**26.3**	**60.4**	**60.1**
S: ViT-B/16	37.0	25.0	24.7	54.6	56.6
+TinyCLIP (from T_1)	55.4	35.9	33.6	73.2	72.8
+CLIP-KD (from T_1)	57.5	37.6	35.6	75.3	74.5
+Our (from T_1)	**59.3**	**38.7**	**37.1**	**76.3**	**75.8**
S: ResNet-50	35.3	23.5	24.7	55.1	55.0
+CLIP-KD (from T_2)	55.4	**36.3**	33.4	**73.0**	**72.2**
+Our (from T_2)	**56.1**	36.1	**33.9**	**73.0**	70.8

R@1 retrieval by 1.0% on CC3M (I2T), 1.5% on MSCOCO (I2T), and up to 2.4% on Flickr (T2I: 60.1 vs 57.7). These results underscore the efficacy of adaptively modeling intra-modal consistency alongside cross-modal alignment.

We explored different weighting strategies for $\mathcal{L}_{\text{intra-KD}}$. Our proposed adaptive KL-weighting (row "+ $\mathcal{L}_{\text{intra-KD}}$") generally outperforms using uniform weights or weights detached from the gradient graph ("Weights w/ No grad"), particularly on classification tasks and Flickr retrieval. For instance, adaptive KL-weighting yields 43.3% IN Acc, compared to 42.1% for uniform weights and 42.9% for no-gradient weights. While weighting sample contributions with $\mathcal{L}_{\text{CLIP}}$ did lead to a 0.4% gain in CC3M I2T R@1, it negatively impacted performance on other datasets. This suggests a benefit to dynamically and differentiably balancing loss contributions based on student-teacher intra-modal disagreement.

Finally, experiments removing either the image-specific intra-modal loss component $\mathcal{L}_{I\to I}$ or the text-specific one $\mathcal{L}_{T\to T}$ from our full $\mathcal{L}_{\text{intra-KD}}$ (with adaptive KL-weights) individually showed that the complete method generally achieves better or competitive results. For example, removing $\mathcal{L}_{I\to I}$ slightly drops IN Acc to 42.4%, while removing $\mathcal{L}_{T\to T}$ drops CC3M T2I R@1 from 36.4% to 35.9%. While minor variations exist (e.g., MSCOCO T2I for w/o $\mathcal{L}_{I\to I}$), the overall trend supports the inclusion of both components.

4.2 Zero-Shot Evaluation

Table 2 details the zero-shot evaluation of our method against baselines (TinyCLIP, CLIP-KD) across ViT-T/16, ViT-B/16, and ResNet-50 students. These were distilled from LAION-400M pretrained teachers using the CC12M+CC3M dataset.

Our approach consistently demonstrates superior performance. For the ViT-T/16 student (distilled from T_2), our method achieves 43.3% ImageNet top-1 accuracy, outperforming TinyCLIP by 2.5% and CLIP-KD by 0.7%. It also leads in retrieval, with R@1 gains such as +1.4% on MSCOCO I2T, while achieving state-of-the-art Flickr30K performance (e.g., +0.2% T2I R@1 compared to CLIP-KD). Consistent with CLIP-KD, ViT-B/16 is a more effective teacher than ViT-L/14 for ViT-T/16.

With a ViT-B/16 student (from T_1), our method reaches 59.3% on ImageNet (+1.8% vs. CLIP-KD, +3.9% vs. TinyCLIP), and improves R@1 scores by up to 1.1% on MSCOCO and 1.0% on Flickr over baselines.

This efficacy generally extends to the ResNet-50 backbone (distilled from T_2), where our method increases ImageNet accuracy by 0.7% over CLIP-KD (56.1% vs 55.4%). On retrieval tasks, it delivers competitive performance, for example, improving MSCOCO T2I R@1 by +0.5% (33.9% vs 33.4%) over CLIP-KD, though CLIP-KD maintains a slight edge on Flickr T2I for this specific student. Such trends of improved or competitive performance across diverse architectures highlight our framework's potential robustness and broad applicability.

5 Conclusion

We proposed an intra-modal knowledge distillation approach that efficiently transfers knowledge from large vision-language models to compact student networks. By integrating standard and intra-modal losses with tailored weighting strategies, our method consistently boosts student performance in both ablation and zero-shot evaluations, surpassing previous techniques in classification and retrieval benchmarks. Future research will extend this framework by incorporating diverse CLIP variants alongside other multimodal and unimodal models to leverage their architectural and training differences, thereby enriching the knowledge source for distillation. Furthermore, we will conduct rigorous experiments to determine the optimal selection of the hyperparameters α, β, γ, and δ in our total loss function (Eq. 25).

Acknowledgments. Research reported in this paper was supported by the Normandy Region, the ANR under award number ANR-19-CHIA-0017 and was performed using computing resources of CRIANN and Jean-Zay.

References

[1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[2] 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 (CVPR)*, pages 3558–3568, June 2021.

[3] Kang Chen and Xiangqian Wu. Vtqa: Visual text question answering via entity alignment and cross-media reasoning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27208–27217, 2024. doi: 10.1109/CVPR52733.2024.02570.

[4] Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. Improving selective visual question answering by learning from your peers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24049–24059, June 2023.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[7] Yunhao Ge, Jie Ren, Andrew Gallagher, Yuxiao Wang, Ming-Hsuan Yang, Hartwig Adam, Laurent Itti, Balaji Lakshminarayanan, and Jiaping Zhao. Improving zero-shot generalization and robustness of multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11093–11101, June 2023.

[8] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, Jun 2021. ISSN 1573-1405. doi: 10.1007/s11263-021-01453-z. URL https://doi.org/10.1007/s11263-021-01453-z.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[10] 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 (ICCV)*, pages 8340–8349, October 2021.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, March 2015. URL http://arxiv.org/abs/1503.02531. arXiv:1503.02531.

[12] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume

139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jia21b.html.

[14] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zq1iJkNk3uN.

[15] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26617–26626, June 2024.

[16] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

[17] 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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. doi: 10.1109/ICCV48922.2021.00986.

[18] Ségolène Martin, Yunshi Huang, Fereshteh Shakeri, Jean-Christophe Pesquet, and Ismail Ben Ayed. Transductive zero-shot and few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28816–28826, June 2024.

[19] Sachin Mehta and Mohammad Rastegari. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer, 2022. URL https://arxiv.org/abs/2110.02178.

[20] Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18983–18992, June 2023.

[21] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[22] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine*

Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

[23] Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17542–17552, June 2024.

[24] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs, November 2021. URL http://arxiv.org/abs/2111.02114. arXiv:2111.02114.

[25] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238/.

[26] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650, June 2022.

[27] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[28] 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 (ICCV)*, pages 15521–15533, October 2023.

[29] Chufeng Tan, Xing Xu, and Fumin Shen. A survey of zero shot detection: Methods and applications. *Cognitive Robotics*, 1:159–167, 2021. ISSN 2667-2413. doi: https://doi.org/10.1016/j.cogr.2021.08.001. URL https://www.sciencedirect.com/science/article/pii/S2667241321000124.

[30] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/tan19a.html.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.,

2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[32] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3eefceb8087e964f89c2d59e8a249915-Paper.pdf.

[33] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi (Stephen) 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 (ICCV)*, pages 21970–21980, October 2023.

[34] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[35] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19163–19173, June 2022.

[36] 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. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

[37] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A New Foundation Model for Computer Vision, November 2021. URL http://arxiv.org/abs/2111.11432. arXiv:2111.11432.

[38] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18123–18133, June 2022.

[39] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Intra-Modal Divergence-Weighted Distillation for Vision-Language Models Supplementary Materials

A Training Settings

Visual encoder			Text encoder: Transformer [31]			
Model	Type	Params	Layer	Width	Head	Params
ViT-L/14 [6]	ViT	304.0M	12	768	12	85.1M
ViT-B/16 [6]		86.2M	12	512	8	37.8M
ViT-T/16 [6]		5.6M	12	384	6	21.3M
MobileViT-S [19]		5.3M				
Swin-T [17]		27.9M				
ResNet-50 [9]	CNN	38.3M	12	512	8	37.8M
ResNet-18 [9]		11.4M	12	384	6	21.3M
MobileNetV3 [12]		2.0M				
EfficientNet-B0 [30]		4.7M				

Table 3: Comparison of visual and text encoder configurations.

We perform knowledge distillation by training our models on the combined Conceptual Captions 3M (CC3M) and Conceptual 12M (CC12M) datasets. The teacher models are typically pretrained on LAION-400M, except for those referenced in Table 5, which are both pretrained and distilled using the CC3M+CC12M dataset. Comprehensive configuration details for both the training and pretrained models are provided in Table 3.

Training is optimized using the AdamW optimizer with an initial learning rate of 0.001 and a weight decay of 0.1. We adopt a cosine learning rate schedule with a linear warm-up phase over the first 10,000 iterations, spanning a total of 32 epochs. All experiments are conducted on 8 NVIDIA A100 GPUs with a total batch size of 1024, distributed as 128 samples per GPU.

The distillation loss components are weighted as follows: $\alpha = 2000$, $\beta = 1.0$, and $\gamma = 1.0$, where γ is selected based on best performance from the range $\{0.5, 1, 1.5, 2\}$, following the CLIP-KD setup [34]. We also set $\delta = 1$. Learnable temperature parameters τ, τ_M, τ_D, and τ_{intra} are all initialized to 0.07. All other training hyperparameters are aligned with those used in the original CLIP model [21].

In addition, the hyperparameter c for the sample weighting strategy is set to 0.006, selected from the candidate values $\{0.006, 0.001, 0.01, 0.1, 1\}$ based on empirical performance.

B Testing Settings

We evaluated the zero-shot classification performance of our models on several benchmark datasets, including ImageNet (IN) [5], ImageNet-V2 (IN-V2) [22], ImageNet-Rendition (IN-R) [10], and ImageNet-Sketch (IN-S) [32]. For retrieval tasks, we assessed performance on the CC3M validation set [16], MSCOCO [16], and Flickr30K [36].

Consistent with standard evaluation protocols, we used Recall@K (R@K) to measure retrieval accuracy among the top-K nearest neighbors. Our primary metrics were top-1 accuracy (Acc) for image classification, and Recall@1 (R@1) for both Image-to-Text (I2T) and Text-to-Image (T2I) retrieval tasks. Details on dataset sizes and splits are provided in Table 4.

Dataset	Split	Number of Samples
Image Classification		
ImageNet (IN)	Val	50,000
ImageNet-V2 (IN-V2)	Test	10,000
ImageNet-Rendition (IN-R)	Test	30,000
ImageNet-Sketch (IN-S)	Test	50,000
Cross-Modal Retrieval		
CC3M	Val	13,000
MSCOCO	Test	5000
Flickr30K	Test	1000

Table 4: Dataset Sample Counts

C Additional Experiments

Table 5: Performance of various architectures with and without knowledge distillation on ImageNet classification and cross-modal retrieval tasks. The distillation dataset and pre-training data for all models is CC12M+CC3M.

Method	**IN**	**INV2**	**IN-R**	**IN-S**	**CC3M Val**		**MSCOCO**		**Flickr**	
	Acc	Acc	Acc	Acc	I2T	T2I	I2T	T2I	I2T	T2I
T: ViT-B/16	37.0	32.1	48.4	26.0	40.2	39.5	25.0	24.7	54.6	56.6
S: Mobile ViT-S	32.6	27.6	39.5	20.1	36.0	35.6	22.3	22.9	50.1	53.0
+CLIP-KD	36.0	31.1	44.5	23.5	39.4	38.6	26.1	24.9	55.0	56.2
+Our	**36.9**	**31.8**	**46.0**	**24.3**	**39.5**	**38.8**	**26.4**	**25.0**	**56.3**	**57.6**
S: Swin-T	36.4	31.1	45.9	24.4	39.8	39.2	24.7	25.3	53.4	54.4
+CLIP-KD	40.2	34.9	51.4	28.2	43.7	42.5	28.5	28.6	**62.2**	60.9
+Our	**40.7**	**35.2**	**53.3**	**29.2**	**43.9**	**42.7**	**29.0**	**28.7**	60.2	**64.3**
S: MobileNetV3	25.1	20.7	29.2	13.4	28.1	27.5	15.3	15.0	36.9	38.0
+CLIP-KD	**27.0**	**23.0**	**30.6**	14.1	**30.1**	**28.6**	**17.9**	**16.0**	**42.4**	**42.3**
+Our	25.6	22.3	30.4	**14.4**	29.7	28.2	17.5	15.8	40.3	39.5
S: EfficientNet-B0	32.6	27.8	40.9	20.7	35.4	34.9	21.7	21.1	48.3	50.1
+CLIP-KD	35.4	**30.6**	44.7	23.7	39.0	38.0	26.0	23.9	55.5	54.2
+Our	**35.5**	30.3	**45.8**	**24.3**	**39.2**	**38.1**	**26.0**	**24.0**	**55.6**	**56.6**
S: ResNet-18	28.6	24.0	35.3	18.1	31.1	30.4	19.2	18.6	41.0	43.3
+CLIP-KD	**31.4**	26.9	**39.2**	20.0	**34.2**	**33.0**	21.3	**19.8**	47.8	47.1
+Our	30.9	**27.1**	38.9	**20.6**	33.8	32.4	**21.5**	19.6	**49.4**	**47.3**

Table 5 presents a comprehensive evaluation of our proposed intra-modal knowledge distillation method ("+Our") against the "+CLIP-KD" technique and baseline student models across five diverse architectures: Mobile ViT-S, Swin-T, MobileNetV3, EfficientNet-B0, and ResNet-18. The evaluation covers ImageNet classification variants and cross-modal retrieval on CC3M Val, MSCOCO, and Flickr30K, with a ViT-B/16 teacher and CC12M+CC3M for distillation. Overall, our proposed method demonstrates superior or highly competitive performance. For instance, with the Mobile ViT-S student, our approach consistently achieves the best results, improving IN accuracy by 0.9% over CLIP-KD to 36.9% and enhancing

Flickr I2T R@1 by 1.3%. Similarly, when applied to the Swin-T student, our method generally leads, increasing IN accuracy by 0.5% to 40.7% and MSCOCO I2T R@1 by 0.5% compared to CLIP-KD, although CLIP-KD shows a stronger result on Flickr I2T, our method excels on Flickr T2I. With CNN-based students like EfficientNet-B0, our method often provides slight advantages or matches CLIP-KD, such as a 0.1% gain in IN accuracy and better performance on IN-R, IN-S, and Flickr T2I. For ResNet-18, the results are more nuanced: our method shows strengths on IN-S and Flickr retrieval tasks, while CLIP-KD leads on IN accuracy and IN-R. A notable exception is the MobileNetV3 student, where CLIP-KD generally outperforms our method across most metrics, though our approach does secure a slight edge on the IN-S benchmark. Despite this, the experiments broadly indicate that our distillation strategy is effective across varied student architectures, frequently offering an advantage over CLIP-KD, particularly for Transformer-style students and on several retrieval tasks, and consistently uplifts performance significantly from the student baselines. The relatively smaller gains for CNN-based students may be attributed to suboptimal hyperparameter choices, which were initially tuned for Transformer architectures and may require adjustment for CNNs.