INTRA-MODAL DIVERGENCE-WEIGHTED DISTILLATION FOR VISION-LANGUAGE MODELS

Youva Addad[1], Alexis Lechervy[1], Frederic Jurie[1]

[1]Université Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR 6072, F-14000 Caen, France

youva.addad@unicaen.fr, alexis.lechervy@unicaen.fr, frederic.jurie@unicaen.fr

GREYC Electronics and Computer Science Laboratory

BMVC 2025

1. Problem & Motivation

- **Context** : Large VLMs like CLIP are powerful and show strong zero-shot capabilities , but they are computationally expensive and large.
- **Goal** : Knowledge Distillation is used to transfer knowledge from a large "teacher" model to a smaller, more efficient "student" model.
- **The Gap** : Existing KD methods for VLMs often focus on transferring cross-modal (image-text) alignment or direct feature/logit matching.
- **The Challenge** : These methods often fail to capture the teacher's nuanced understanding of intra-modal relationships. This is the sophisticated internal knowledge of how, for example, one image relates to other images or one text relates to other texts, especially the subtle distinctions between negative examples.

2. Proposed Method

Our method introduces a dynamic loss, $\mathcal{L}_{\text{intra-KD}}$, that aligns the student with the teacher's within-modality relational structure.

Standard Distillation Losses

These are common KD losses we use as a baseline :

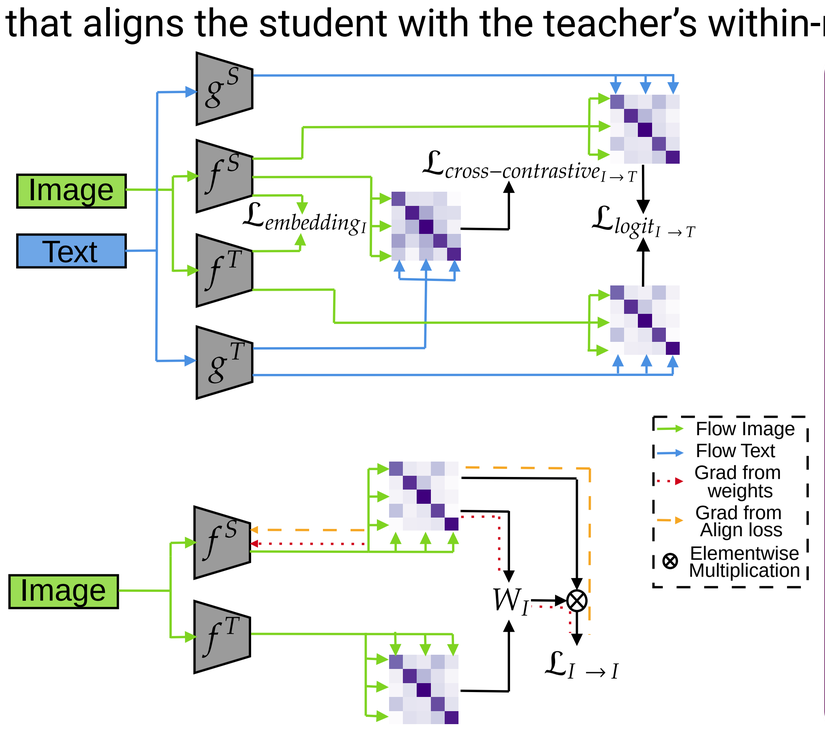
$$\mathcal{L}_{\text{embedding}_I} = ||\mathbf{v}_k^T - \mathbf{v}_k^S||_2^2$$

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

$$\mathcal{L}_{\text{cross-contrastive}_{I\to T}} = -\log \frac{\exp(\text{sim}(\mathbf{v}_k^S,\mathbf{u}_k^T)/\tau_D)}{\sum_{l=1}^N \exp(\text{sim}(\mathbf{v}_k^S,\mathbf{u}_l^T)/\tau_D)}$$

Combine Losses : The total training objective combines the student's own CLIP loss, the standard KD losses, and our new intra-modal loss.

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



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

This is our core contribution, which captures the teacher's relational knowledge within each modality (e.g., Image-to-Image, Text-to-Text). $\mathcal{L}_{\text{intra-KD}} = \mathcal{L}_{I\to I} + \mathcal{L}_{T\to T}$

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

$$K_{I,k} = \sum_{j=1}^N \text{KL}\left(P^T_{I\to I,k}(j) \,\|\, P^S_{I\to I,k}(j)\right)$$

$$W_I = Softmax(K_I/c)$$

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

3. Experimental Results

Ablation Study

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

Zero-Shot Evaluation vs. Baselines

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

The 36th British Machine Vision Conference 24th - 27th November 2025, Sheffield, UK