DON'T DROP THE WEAK PAIRS: ALIGNMENT-AWARE DISTILLATION OF VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale vision-language pretraining strongly benefits from data filtering, as poorly aligned image-text pairs provide noisy contrastive supervision and are often better discarded. We show that this intuition does not fully carry over to vision-language distillation. During distillation, weakly aligned pairs may indeed be harmful to the image-text contrastive objective but still provide valuable teacher supervision through feature, relation, or representation distillation. Rather than removing such samples, we propose to adapt how they are used. Specifically, we introduce an alignment-aware soft-gating mechanism that uses the teacher's image-text similarity to modulate, on a per-sample basis, alignment-sensitive objectives while preserving knowledge distillation. Well-aligned pairs receive stronger contrastive supervision, while weakly aligned pairs rely more heavily on the teacher signal. The gate is computed online from teacher embeddings already available for distillation, adding negligible overhead. Across CC12M and DataComp-M datasets, our approach consistently improves CLIP-KD and remains complementary to adaptive data selection with ACID. We further show that alignment weighting improves another distillation recipe, in TIPSv2, demonstrating that its benefits extend across different distillation settings. Our results suggest that, for vision-language distillation, noisy pairs are better reweighted than discarded. Code and pretrained weights will be open-sourced.

1 INTRODUCTION

Contrastive vision-language models such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) have become central components of modern multimodal systems (Liu et al., 2023; Beyer et al., 2024; Singer et al., 2023; Bai et al., 2025; Qwen, 2026). However, as their performance scales strongly with model capacity, the most capable models (Chen et al., 2024; Sun et al., 2024; Bolya et al., 2025) are often too large and computationally expensive for practical deployment. A common strategy is therefore to first train a large vision-language model and subsequently transfer its knowledge to a smaller student through knowledge distillation (Wu et al., 2023; Yang et al., 2024). This strategy makes it possible to retain much of the representational strength of large-scale models while producing compact models that are substantially cheaper to deploy, serve, and adapt. This makes effective and efficient vision-language distillation increasingly important.

At the same time, large-scale vision-language pretraining has shown that the quality of image-text pairs is critical. Web-scale datasets contain many weakly related, noisy, or mismatched image-caption pairs. For example, as illustrated in Fig. 1, a caption may be semantically related to the depicted content while describing concepts only weakly grounded in the image itself. Such pairs can provide misleading supervision to the contrastive objective. A widely adopted remedy is to filter the training pool, leaving out misaligned samples according to an image-text alignment score produced by a pretrained model. Data curation studies have shown that retaining only strongly aligned pairs can outperform training on substantially larger unfiltered datasets (Gadre et al., 2023). This follows naturally in standard contrastive pretraining: image-text correspondence itself constitutes the main source of supervision, and poorly aligned pairs therefore directly corrupt the learning signal.

Whether the same principle should apply during vision-language distillation has been largely underexplored. Beyond the image-text contrastive objective, the student also receives complementary teacher supervision through feature distillation (Romero et al., 2015), relation distillation (Yang

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

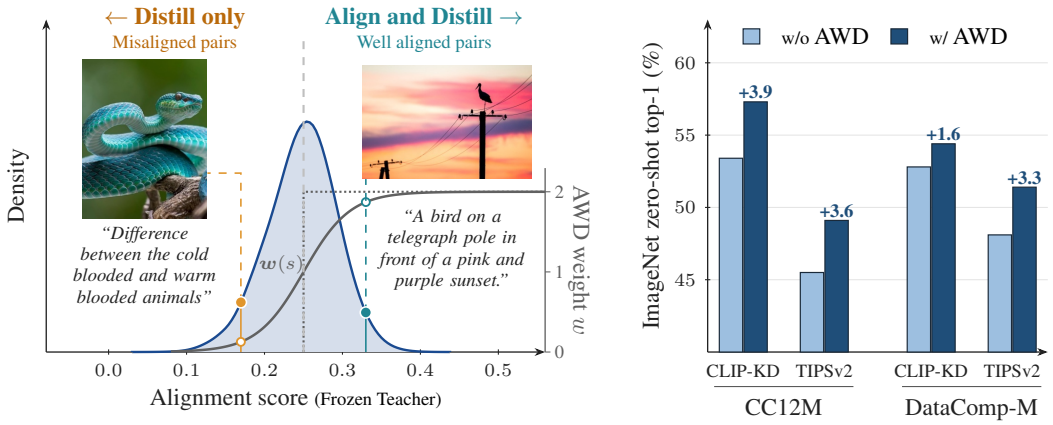


Figure 1: **Motivation for Alignment-Aware Distillation (AWD).** *Left:* CLIP distillation typically combines *alignment-sensitive* objectives (e.g., image-text contrastive loss) with *teacher distillation* objectives (e.g., feature distillation). AWD modulates supervision with a soft-gating weight (grey) derived from teacher alignment scores (blue): weak pairs retain distillation but receive less alignment-sensitive supervision, while well-aligned pairs receive both. *Right:* AWD improves performance for both CLIP-KD and TIPSv2 distillation on CC12M and DataComp-Medium.

et al., 2024), or prototype-based representation distillation (Cao et al., 2026). Importantly, some of these signals do not depend on the quality of the image-text correspondence. For instance, in visual feature distillation, a strong visual teacher supervises the student from the image alone, irrespective of whether its associated caption is well aligned. This changes how a weakly aligned pair should be viewed: while poor image-text correspondence can make the pair unreliable for contrastive learning, each modality can still independently provide useful teacher representations for knowledge transfer.

An alternative to static hard filtering is adaptive data selection. ACID (Udandarao et al., 2025) improves distillation by constructing batches according to teacher and student losses, favoring samples that help reduce the teacher-student gap. However, it remains a hard selection mechanism: a sample is either discarded, or used with the same contrastive and distillation objectives as any other selected sample. This mechanism therefore does not account for the fact that the same sample can provide poor supervision for image-text alignment while remaining informative for knowledge transfer. This motivates a different question from sample selection: rather than deciding *which* samples to use, can we adapt *how* each sample contributes to the different training objectives?

In this work, we address this question by introducing ***A**lignment-a**W**are **D**istillation* (AWD), a simple soft-gating mechanism derived from the teacher's image-text similarity. Strongly aligned pairs receive stronger contrastive supervision, while the contrastive contribution of weakly aligned pairs is downweighted, shifting their relative contribution toward the teacher distillation signal. This allows weakly aligned pairs to remain useful for knowledge transfer without forcing the student to learn from unreliable image-text correspondences. AWD incurs negligible overhead, as the alignment score is computed online from teacher representations already available during distillation.

We summarize our contributions as follows:

- We study how image-text alignment affects vision-language distillation and highlight a limitation of the commonly used hard sample filtering: pairs with unreliable contrastive supervision can nevertheless contain a useful distillation signal, which is ignored.
- We introduce *Alignment-aWare Distillation* (AWD), a soft-gating mechanism using teacher image-text similarity to adapt per-sample contrastive supervision relative to distillation, with negligible overhead.
- Despite its simplicity, AWD consistently improves student models across pretraining datasets and distillation protocols, including CC12M and DataComp-Medium datasets, CLIP-KD (Yang et al., 2024) and TIPSv2 (Cao et al., 2026) distillations, as well as when combined with I-ACID and H-ACID adaptive filtering (Udandarao et al., 2025).

2 Related work

Vision-language distillation. Knowledge distillation is increasingly used to derive practical model families from large-scale vision-language models. Recent high-performing models such as PE-Core (Bolya et al., 2025) and TIPSv2 (Cao et al., 2026) directly pretrain their largest variants at scale and obtain smaller models through distillation. TIPSv2, for instance, distills its smaller variants from a frozen giant teacher using a training objective closely related to its pretraining recipe. Other works have specifically investigated how knowledge should be transferred between vision-language models. TinyCLIP (Wu et al., 2023) introduces affinity mimicking in order to transfer cross-modal relationships from a teacher, together with weight inheritance for efficient student initialization. CLIP-KD (Yang et al., 2024) systematically studies several forms of knowledge transfer for CLIP, including feature, relation, gradient, and contrastive distillation. Beyond the choice of distillation objective, ACID (Udandarao et al., 2025) studies the role of training data and proposes an online selection strategy that constructs batches according to teacher and student losses. These works improve either *what* knowledge is transferred or *which* samples are used for distillation. The role of alignment quality in determining how image-text contrastive and distillation supervision should be combined, however, remains largely unexplored.

Data curation for vision-language pretraining. Large-scale vision-language models are typically trained on image-text pairs collected from the web, where captions can be weakly related, incomplete, or even entirely mismatched with their associated images. A common approach is therefore to curate training data using the similarity produced by a pretrained vision-language model. CLIP-based filtering has been used to construct large-scale datasets such as LAION (Schuhmann et al., 2021; 2022), while DataComp (Gadre et al., 2023) systematically studies data curation under controlled compute budgets and demonstrates the effectiveness of retaining highly aligned image-text pairs. Data Filtering Networks (DFN) (Fang et al., 2024) further show that models can be trained specifically to identify high-quality data for contrastive pretraining. Such filtering is well motivated when image-text correspondence constitutes the main source of supervision. In the distillation setting, however, low-alignment pairs may still provide useful teacher supervision, motivating us to revisit whether such samples should be removed altogether.

Alignment-aware learning from noisy pairs. Several works have explored pretraining using alignment quality as a continuous signal rather than a binary filtering criterion. ALIP (Yang et al., 2023) combines raw and synthetic captions, and dynamically adjusts the weights of the contrastive supervision according to raw and synthetic caption consistency as well as image-text alignment. Noise-Aware Captioning (NoC) (Kang et al., 2023) uses CLIP-derived alignment levels as a conditioning signal to learn from noisy web-crawled image-text data. Related ideas have also appeared beyond representation pretraining: Coherence-Aware Diffusion (CAD) (Dufour et al., 2024) conditions a diffusion model on the coherence of its conditioning signal, allowing samples with unreliable captions to remain useful rather than being discarded. These works show that alignment can be used as a graded training signal, but they use it to modulate a single pretraining or generation objective. We instead consider the distillation setting, where each sample is supervised by both its image-text correspondence and a teacher model. We use teacher-derived alignment to *reallocate* supervision between these two sources, reducing the contribution of unreliable contrastive supervision while preserving the distillation signal carried by the same sample.

3 Alignment-aware vision-language distillation

We first review standard vision-language distillation (Sec. 3.1). We then introduce our teacher-guided alignment weighting, which distinguishes training objectives according to their dependence on the observed image-text correspondence (Sec. 3.2). Finally, we instantiate this principle in two distillation protocols based on CLIP-KD and TIPSv2 (Sec. 3.3).

3.1 Preliminaries: distilling vision-language models

Consider a mini-batch $\mathcal{B} = \{x_i\}_{i=1}^{B}$ of image-text pairs $x_i = (I_i, T_i)$. Let $\mathbf{v}_i^{\mathrm{S}}, \mathbf{t}_i^{\mathrm{S}}$ and $\mathbf{v}_i^{\mathrm{T}}, \mathbf{t}_i^{\mathrm{T}}$ denote normalized post-projection image and text representations (i.e., after the output of both image and text encoders is projected onto the joint, multimodal space), as produced by the student (S) and

teacher (T), respectively. The standard CLIP objective learns from the dataset-provided correspondence by treating (I_i, T_i) as a positive pair.

For the student, the image-to-text $p^{\mathrm{S}}_{\mathrm{v}\to\mathrm{t},i}$ and text-to-image $p^{\mathrm{S}}_{\mathrm{t}\to\mathrm{v},i}$ distributions of the contrastive loss terms over the mini-batch are

$$p^{\mathrm{S}}_{\mathrm{v}\to\mathrm{t},i}(j) = \frac{\exp\left(\mathbf{v}^{\mathrm{S}}_i \cdot \mathbf{t}^{\mathrm{S}}_j/\tau\right)}{\sum_{b=1}^{B}\exp\left(\mathbf{v}^{\mathrm{S}}_i \cdot \mathbf{t}^{\mathrm{S}}_b/\tau\right)}, \quad p^{\mathrm{S}}_{\mathrm{t}\to\mathrm{v},i}(j) = \frac{\exp\left(\mathbf{t}^{\mathrm{S}}_i \cdot \mathbf{v}^{\mathrm{S}}_j/\tau\right)}{\sum_{b=1}^{B}\exp\left(\mathbf{t}^{\mathrm{S}}_i \cdot \mathbf{v}^{\mathrm{S}}_b/\tau\right)}, \tag{1}$$

where τ is the contrastive temperature. The image-text contrastive loss for the i-th pair (sample) is

$$\ell_{\mathrm{CLIP},i} = -\frac{1}{2}\left[\log p^{\mathrm{S}}_{\mathrm{v}\to\mathrm{t},i}(i) + \log p^{\mathrm{S}}_{\mathrm{t}\to\mathrm{v},i}(i)\right]. \tag{2}$$

Vision-language distillation augments this image-text supervision with targets produced by a stronger teacher. For instance, CLIP-KD (Yang et al., 2024), which we use as a representative multi-objective formulation, combines the original CLIP objective with interactive contrastive learning (ICL), contrastive relational distillation (CRD), and feature distillation (FD):

$$\mathcal{L}_{\text{CLIP-KD}} = \frac{1}{B}\sum_{i=1}^{B}\left[\ell_{\mathrm{CLIP},i} + \lambda_{\mathrm{ICL}}\ell_{\mathrm{ICL},i} + \lambda_{\mathrm{CRD}}\ell_{\mathrm{CRD},i} + \lambda_{\mathrm{FD}}\ell_{\mathrm{FD},i}\right]. \tag{3}$$

These additional distillation objectives transfer complementary forms of teacher knowledge. We provide their definitions when introducing our alignment-aware instantiation of CLIP-KD in Sec. 3.3.

3.2 Teacher-guided alignment weighting

Alignment-sensitive supervision. Web-scale data is inherently noisy. Positive correspondences defined by the dataset may thus be misaligned and cannot always be trusted. We call an objective *alignment-sensitive* if it explicitly uses the given image-text pair as a hard target, i.e., treats the paired image and text as a positive match. We denote the set of such objectives by $\mathcal{A}$. In contrast, $\mathcal{T}$ denotes teacher-target objectives whose supervision does not use the image-text pair as a hard target.

This distinction allows us to modulate supervision according to pair reliability without suppressing useful teacher targets. Unlike sample filtering, which fully discards some sample supervision, or sample-level reweighting, which assigns a single, uniform weight to each sample and jointly operates on both types of supervision, we apply the alignment weight only to alignment-sensitive objectives. We define our general alignment-aware distillation objective as

$$\mathcal{L}^{\mathrm{AWD}} = \frac{1}{B}\sum_{i=1}^{B}\left[w_i\sum_{k\in\mathcal{A}}\lambda_k\ell_{k,i} + \sum_{k\in\mathcal{T}}\lambda_k\ell_{k,i}\right]. \tag{4}$$

Thus, alignment determines *how strongly the observed image-text correspondence is trusted*, rather than whether the sample contributes to distillation altogether.

Alignment score and soft-gating weight. We estimate pair alignment using the same vision-language teacher employed for distillation. For each pair, we compute the cosine similarity between its normalized teacher image and text representations $s_i = \mathbf{v}^{\mathrm{T}}_i \cdot \mathbf{t}^{\mathrm{T}}_i$. To implement our soft-gating mechanism, this alignment score is converted via a sigmoid into a weight

$$w_i = 2\sigma\left(\alpha\,\frac{s_i - \mu}{Q}\right), \tag{5}$$

where μ is the median of the alignment scores and Q their quartile deviation (half the interquartile range). We apply this robust standardization, $(s_i - \mu)/Q$, to make the teacher scores comparable across corpora, which may have non-normal distributions, and the sharpness $\alpha \geq 0$ sets how strongly the weight responds to the score. The resulting weight satisfies $w_i \in (0, 2)$, with $w_i = 1$ when $s_i = \mu$, so that the overall balance between alignment-sensitive and teacher-target objectives is preserved. Low-alignment pairs are therefore attenuated while high-alignment pairs are amplified.

The parameter α explicitly controls this transition: $\alpha = 0$ gives uniform weights as then $w_i = 1$, which effectively cancels alignment weighting, whereas $\alpha \to \infty$ approaches a hard mask at μ. We use $\alpha = 1$ in all other experiments and ablate the sensitivity of the method to it in Fig. 3. Since the teacher representations are already required for distillation, computing s_i introduces no additional teacher forward pass.

Online calibration. Teacher similarity distributions can vary across training corpora and teacher models (Fig. 4). Rather than relying on dataset-specific thresholds, we estimate the center and scale of the weighting function online. At training step n, we compute the median $\hat{\mu}_n$ and quartile deviation $\hat{Q}_n$ of the alignment scores in the batch and we maintain exponential moving averages of both quantities,

$$\mu_n = \beta\mu_{n-1} + (1-\beta)\hat{\mu}_n, \qquad Q_n = \beta Q_{n-1} + (1-\beta)\hat{Q}_n, \tag{6}$$

where the momentum is set to $\beta = 0.99$ in all our experiments.

Together with the explicit sharpness α, this separates the scale of the observed teacher scores from the desired sharpness of the weighting function.

3.3 Instantiating alignment-aware distillation

Multi-objective CLIP distillation. We first apply our formulation to CLIP-KD. Its CLIP and ICL objectives both explicitly treat a dataset-provided image-text pair as a positive target and are therefore alignment-sensitive. Specifically, ICL contrasts student image representations against teacher text representations, and vice versa, using a contrastive objective analogous to CLIP. More precisely, the image-to-text $p^{\mathrm{S}\to\mathrm{T}}_{\mathrm{v}\to\mathrm{t},i}$ and text-to-image $p^{\mathrm{S}\to\mathrm{T}}_{\mathrm{t}\to\mathrm{v},i}$ distributions are defined as

$$p^{\mathrm{S}\to\mathrm{T}}_{\mathrm{v}\to\mathrm{t},i}(j) = \frac{\exp\left(\mathbf{v}^{\mathrm{S}}_i \cdot \mathbf{t}^{\mathrm{T}}_j/\tau'\right)}{\sum_{b=1}^{B}\exp\left(\mathbf{v}^{\mathrm{S}}_i \cdot \mathbf{t}^{\mathrm{T}}_b/\tau'\right)}, \quad p^{\mathrm{S}\to\mathrm{T}}_{\mathrm{t}\to\mathrm{v},i}(j) = \frac{\exp\left(\mathbf{t}^{\mathrm{S}}_i \cdot \mathbf{v}^{\mathrm{T}}_j/\tau'\right)}{\sum_{b=1}^{B}\exp\left(\mathbf{t}^{\mathrm{S}}_i \cdot \mathbf{v}^{\mathrm{T}}_b/\tau'\right)}, \tag{7}$$

where τ' is a learnable temperature. Here and below, if the embedding sizes between teacher and student differ, a learnable linear projection head is first applied to student embeddings to match the teacher dimension, as in Yang et al. (2024). The ICL loss for the i-th pair (sample) is

$$\ell_{\mathrm{ICL},i} = -\frac{1}{2}\left[\log p^{\mathrm{S}\to\mathrm{T}}_{\mathrm{v}\to\mathrm{t},i}(i) + \log p^{\mathrm{S}\to\mathrm{T}}_{\mathrm{t}\to\mathrm{v},i}(i)\right]. \tag{8}$$

Hence, despite involving teacher representations, ICL still assumes that the observed pair $j = i$ is a valid positive correspondence.

The contrastive relational distillation objective (CRD) behaves differently. Let $p^{\mathrm{T}}_{\mathrm{v}\to\mathrm{t},i}$ and $p^{\mathrm{T}}_{\mathrm{t}\to\mathrm{v},i}$ denote the teacher cross-modal similarity distributions. Its per-sample objective is

$$\ell_{\mathrm{CRD},i} = \frac{1}{2}\left[D_{\mathrm{KL}}\left(p^{\mathrm{T}}_{\mathrm{v}\to\mathrm{t},i} \| p^{\mathrm{S}}_{\mathrm{v}\to\mathrm{t},i}\right) + D_{\mathrm{KL}}\left(p^{\mathrm{T}}_{\mathrm{t}\to\mathrm{v},i} \| p^{\mathrm{S}}_{\mathrm{t}\to\mathrm{v},i}\right)\right]. \tag{9}$$

Rather than enforcing the dataset-provided pair as a one-hot positive, CRD transfers the teacher's full similarity distribution. A poorly aligned pair can therefore receive low probability from the teacher without invalidating the distillation target.

The feature distillation objective FD similarly provides targets independently of alignment, within each modality. Let $\mathbf{h}^{\mathrm{S}}_i$ and $\mathbf{h}^{\mathrm{T}}_i$ denote visual representations prior to projection onto the shared image–text space. We use

$$\ell_{\mathrm{FD},i} = d_{\mathrm{post}}\left(\mathbf{v}^{\mathrm{S}}_i, \mathbf{v}^{\mathrm{T}}_i,\right) + d_{\mathrm{post}}\left(\mathbf{t}^{\mathrm{S}}_i, \mathbf{t}^{\mathrm{T}}_i,\right) + d_{\mathrm{pre}}\left(\mathbf{h}^{\mathrm{S}}_i, \mathbf{h}^{\mathrm{T}}_i\right), \tag{10}$$

where d_{post} is the cosine distance, and d_{pre} combines cosine and SmoothL1 distances following Sariyildiz et al. (2024).

We consequently partition the CLIP-KD objectives as $\mathcal{A} = \{\mathrm{CLIP}, \mathrm{ICL}\}$ and $\mathcal{T} = \{\mathrm{CRD}, \mathrm{FD}\}$, and applying our alignment weighting thus yields

$$\mathcal{L}^{\mathrm{AWD}}_{\text{CLIP-KD}} = \frac{1}{B}\sum_{i=1}^{B}\left[w_i\left(\ell_{\mathrm{CLIP},i} + \lambda_{\mathrm{ICL}}\ell_{\mathrm{ICL},i}\right) + \left(\lambda_{\mathrm{CRD}}\ell_{\mathrm{CRD},i} + \lambda_{\mathrm{FD}}\ell_{\mathrm{FD},i}\right)\right]. \tag{11}$$

The contribution of poorly aligned pairs is hence reduced in objectives that treat them as positive correspondences, but preserved in those based on teacher-derived distribution and representation targets. We experimentally confirm this intuition in Sec. 4.2, where we validate the effectiveness of the proposed partition.

Hybrid contrastive–self-supervised distillation. Next, we consider hybrid models that combine the CLIP objective with self-supervised, image-only losses (Tschannen et al., 2025; Naeem et al., 2024; Maninis et al., 2025). These models are naturally suited for AWD, since much of their supervision does not depend on image-text alignment. Specifically, we study the recent TIPSv2 (Cao et al., 2026), whose student combines the image-text contrastive objective with DINO (Caron et al., 2021) and iBOT (Zhou et al., 2022a) distillation targets from a large frozen teacher. For TIPSv2, we therefore define $\mathcal{A} = \{\text{CLIP}\}$ and $\mathcal{T} = \{\text{DINO}, \text{iBOT}\}$, and our alignment-aware objective becomes

$$\mathcal{L}^{\text{AWD}}_{\text{TIPSv2}} = \frac{1}{B}\sum_{i=1}^{B}\left[w_i \ell_{\text{CLIP},i} + \lambda_{\text{DINO}}\ell_{\text{DINO},i} + \lambda_{\text{iBOT}}\ell_{\text{iBOT},i}\right]. \tag{12}$$

Alignment weighting therefore modifies only the image-text supervision while preserving the visual teacher signal for every image. This provides another instantiation of our AWD distillation principle.

4 Experiments

Experimental settings and baselines. We use CLIP-KD as our primary distillation baseline. We compare against I-ACID and H-ACID, recent adaptive data-selection methods that select informative training pairs from super-batches using the teacher loss (I-ACID) or both teacher and student losses (H-ACID). We also compare to *hard filtering*, which removes the tertile of pairs with the lowest teacher alignment scores. For each setting, we evaluate the effect of adding AWD. Last, we evaluate AWD on the distilled variants of TIPSv2, providing an additional training setting beyond CLIP-KD. We provide more details on the baselines in Sec. A.1 of the Appendix.

Pretraining data and evaluation. We train on two image-text corpora with complementary scale and alignment quality: 9M downloaded pairs from CC12M (Changpinyo et al., 2021) and 80M downloaded pairs from the 128M-pair uncurated DataComp-Medium (DataComp-M) pool (Gadre et al., 2023). DataComp-M is substantially noisier, with a lower mean teacher alignment score than CC12M (0.20 vs. 0.25). On ImageNet-1k (IN-1k) (Deng et al., 2009), we evaluate zero-shot (ZS) classification, linear probing (LP), and weighted k-NN on frozen representations. We further assess robustness to distribution shift on ImageNet-V2 (IN-V2) (Recht et al., 2019), ImageNet-R (IN-R) (Hendrycks et al., 2021), and ImageNet-Sketch (IN-S) (Wang et al., 2019), image-to-text (I2T) and text-to-image (T2I) retrieval on COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014), and zero-shot transfer to nine fine-grained (FG) classification tasks. Complete evaluation protocols and fine-grained dataset details are provided in Sec. A.2 of the Appendix.

Training details. Across both settings, the student image and text encoders are initialized from scratch. All main experiments are trained for 50k steps at 224×224 resolution on 64 H100 GPUs with a global batch size of 16,384 (256 per GPU), while ablation studies use 20k steps. We use AdamW with a peak learning rate of 5×10^{-4}, weight decay 0.2, and a warmup–stable–drop schedule. For CLIP-KD, we train a ViT-B/16 student from a frozen PE-Core-G14-448 teacher using CLIP, ICL, CRD, and FD; we set $\lambda_{\text{ICL}} = \lambda_{\text{CRD}} = 1$ and $\lambda_{\text{FD}} = 8$. For TIPSv2, we train a ViT-B/14 student using a frozen TIPSv2-g/14 visual teacher with the original CLIP, DINO, and iBOT objectives; the teacher text encoder is used only to compute alignment scores and is not distilled into the student text encoder following the TIPSv2 distillation protocol. We use $\alpha = 1$ and EMA momentum $\beta = 0.99$ for AWD throughout. We give full implementation details in Sec. A.3 of the Appendix.

4.1 Alignment-aware distillation improves downstream performance

Main results. We first evaluate the direct effect of alignment weighting on the standard CLIP-KD objective. As reported in Tab. 1, AWD improves every reported metric on both pretraining corpora. On CC12M, the average gain reaches $+3.5$ points: ImageNet-1k zero-shot accuracy increases by $+3.9$ points, the fine-grained average by $+3.7$, with gains of up to $+4.6$ points on ImageNet-R and $+4.5$ points on Flickr30k text-to-image retrieval. On the larger and noisier DataComp-M corpus, the improvements are smaller but remain consistent, with an average gain of $+1.1$ points, including a rise by $+1.6$ point on ImageNet-1k zero-shot and $+1.2$ on fine-grained recognition. These results suggest that adjusting unreliable alignment-sensitive supervision while preserving alignment-insensitive objectives improves the learned representation across substantially different downstream evaluations.

Table 1: **Results for CLIP-KD distillation.** All students use the same ViT-B/16 architecture. Avg. Δ denotes the average performance difference with respect to CLIP-KD across all reported metrics.

	Method	ImageNet-1k			Distribution Shift			I2T retrieval		T2I retrieval		Fine-grained	Avg. Δ
		ZS	LP	k-NN	IN-V2	IN-R	IN-S	COCO	Flickr	COCO	Flickr		
	Teacher: PE-Core-G14	85.4	89.5	85.5	80.2	96.5	76.5	75.4	96.2	58.1	85.7	85.1	–
CC12M	CLIP-KD	53.4	70.9	66.3	45.6	65.2	41.4	51.0	77.2	35.3	63.1	51.6	–
	+ AWD	**57.3**	**73.1**	**68.5**	**49.5**	**69.8**	**44.7**	**54.7**	**80.6**	**38.8**	**67.6**	**55.3**	**+3.5**
	+ Filtering	**51.4**	**69.1**	**63.5**	**44.0**	64.1	**41.3**	**49.7**	**75.3**	34.1	**62.3**	**50.5**	−1.4
	+ Filtering + AWD	51.0	68.7	63.1	43.6	**64.2**	40.9	49.5	73.6	**34.5**	62.0	50.4	−1.8
	+ I-ACID	55.3	72.0	67.3	47.6	67.6	43.3	54.0	76.8	37.0	65.1	52.3	+1.6
	+ I-ACID + AWD	**57.9**	**73.6**	**69.0**	**50.0**	**71.1**	**45.8**	**55.6**	**80.5**	**39.0**	**67.9**	**55.2**	**+4.1**
	+ H-ACID	53.9	71.2	66.3	47.0	66.3	41.8	51.8	76.9	36.0	63.7	51.2	+0.5
	+ H-ACID + AWD	**55.8**	**72.2**	**67.2**	**48.7**	**68.1**	**43.3**	**53.5**	**80.1**	**37.9**	**66.3**	**53.5**	**+2.3**
DataComp-M	CLIP-KD	52.8	74.0	68.1	45.8	62.6	41.1	46.6	71.5	30.4	55.1	56.0	–
	+ AWD	**54.4**	**74.2**	**68.9**	**47.2**	**64.9**	**42.3**	**47.4**	**71.6**	**31.3**	**57.0**	**57.2**	**+1.1**
	+ Filtering	55.1	74.2	69.0	47.3	65.8	43.2	46.4	**74.2**	30.6	55.7	58.3	+1.4
	+ Filtering + AWD	**55.7**	**74.5**	**69.3**	**48.5**	**66.6**	**44.3**	**47.6**	72.3	**31.2**	**57.1**	**59.1**	**+2.0**
	+ I-ACID	55.1	74.4	69.1	47.9	66.7	43.2	47.1	71.7	30.9	56.1	58.7	+1.5
	+ I-ACID + AWD	**56.5**	**74.7**	**69.5**	**48.9**	**67.5**	**44.6**	**47.9**	**73.5**	**32.6**	**57.1**	**60.1**	**+2.6**
	+ H-ACID	54.9	74.9	**69.7**	47.6	65.2	43.0	**47.5**	72.9	32.2	57.7	58.2	+1.8
	+ H-ACID + AWD	**55.5**	**75.0**	**69.7**	**48.2**	**66.2**	**43.4**	47.4	**76.1**	**32.5**	**58.5**	**58.7**	**+2.5**

The benefit of AWD also remains when combined with adaptive data selection. I-ACID and H-ACID already improve over CLIP-KD by selecting informative training pairs, but AWD provides additional gains on top of both methods. On CC12M, I-ACID improves the baseline by $+1.6$ points and H-ACID by $+0.5$; adding AWD raises these gains to $+4.1$ and $+2.3$, respectively. The same complementarity holds on DataComp-M, where the gains increase from $+1.5$ to $+2.6$ for I-ACID and from $+1.8$ to $+2.5$ for H-ACID. These results are particularly important because the two mechanisms act at different levels: ACID decides which pairs are most useful for the optimization process, whereas AWD modulates how strongly the selected pairs contribute to the alignment-sensitive objectives.

We finally compare AWD with hard filtering, which directly alters the pretraining corpus. We can see that it has a corpus-dependent effect: removing the lowest-alignment pairs degrades CLIP-KD on CC12M (-1.4 points on average) but improves it on DataComp-M ($+1.4$). It illustrates that filtering is not always a safe strategy. The interaction of this curation principle with AWD differs across corpora. On CC12M, adding AWD after filtering slightly decreases the average by 0.4 points, whereas on DataComp-M it provides an additional $+0.6$ points.

Generalization to TIPSv2 distillation. We next test whether AWD transfers beyond the CLIP-KD objective to the substantially different TIPSv2 distillation setting. Unlike CLIP-KD, TIPSv2 combines the CLIP contrastive objective with visual distillation through the DINO and iBOT objectives, which supervise the student visual encoder using the frozen TIPSv2-g/14 visual encoder as teacher. This provides an alternative setting in which AWD acts only on the alignment-sensitive CLIP term, leaving the visual distillation objectives unchanged. As shown in Tab. 2, which averages results for some evaluation datasets for space reasons, AWD yields broad improvements on both corpora. On CC12M, it improves every reported metric and increases the average performance by $+3.2$ points. The gain on ImageNet-1k zero-shot accuracy is $+3.6$ pts, extending to distribution shift ($+3.4$), fine-grained recognition ($+3.6$), image-to-text retrieval ($+3.9$), and text-to-image retrieval ($+4.3$). On DataComp-M, AWD provides a similar $+2.6$-point average

Table 2: **Results for TIPSv2 distillation.** TIPSv2-b/14 students trained with and without AWD. Avg. Δ is computed with respect to TIPSv2-b/14 without AWD.

	Method	IN-1k			Dist. shift	I2T	T2I	FG	Avg. Δ
		ZS	LP	k-NN					
	TIPSv2-g/14	78.0	86.8	83.6	77.2	77.9	63.4	71.4	–
CC12M	TIPSv2-b/14	45.5	76.1	73.6	46.8	59.0	43.1	37.8	–
	+ AWD	**49.1**	**77.8**	**75.3**	**50.2**	**62.9**	**47.4**	**41.4**	**+3.2**
DC-M	TIPSv2-b/14	48.1	**76.1**	68.4	46.7	52.6	36.5	48.9	–
	+ AWD	**51.4**	74.0	**70.4**	**50.7**	**56.1**	**39.6**	**53.6**	**+2.6**

gain, including improvements of $+3.3$ in zero-shot accuracy, $+4.0$ under distribution shift, and $+4.7$ on fine-grained recognition. The only exception is linear probing, which loses -2.1 points while the remaining metrics improve. These results show that the benefit of alignment-aware supervision is not tied to the specific losses or student architecture used by CLIP-KD, but extends to a distinct training recipe combining cross-modal contrastive learning with visual self-distillation.

Recaptioning. We further evaluate AWD on DataComp-M after replacing the original captions with synthetic ones generated by Qwen3-VL-4B-Instruct. As shown in Tab. 3, recaptioning alone substantially improves CLIP-KD, yielding an average gain of $+6.9$ points over the original corpus. Importantly, AWD remains beneficial after this strong data-level intervention: it further improves every reported evaluation group, including distribution shifts ($+1.6$ pts), fine-grained recognition ($+1.4$), and average performance over zero-shot, linear probe and k-NN evaluation on ImageNet-1k ($+0.7$). Overall, adding AWD increases the gain from $+6.9$ to $+7.9$, showing that alignment weighting remains complementary to recaptioning even after image-text alignment has been substantially improved.

Table 3: **Recaptioning on DataComp-M.** Impact of AWD on synthetic captions generated by Qwen3-VL-4B-Instruct. Avg. Δ is the mean difference to CLIP-KD on the original captions.

Method	IN-1k	DS	I2T	T2I	FG	**Avg.** Δ
CLIP-KD	65.0	49.8	59.1	42.8	56.0	–
+ Recap.	67.7	53.5	73.7	57.0	55.2	+6.9
+ Recap. + AWD	**68.5**	**55.1**	**74.2**	**57.7**	**56.6**	**+7.9**

Taken together, these results support using AWD as a practical default for vision-language distillation: it provides strong gains across diverse settings, remains effective alongside stronger selection and data-improvement strategies, and adds negligible computational overhead.

4.2 Ablation Studies

Which objectives for alignment weighting? In Tab. 4, we study which components of the distillation objectives should receive the alignment weight. Applying AWD to the CLIP loss already improves over CLIP-KD, and extending it to ICL (our approach) yields the best performance. In contrast, weighting CRD and FD as well, i.e., applying the alignment weight to the entire objective, reduces performance. This ablation directly motivates our design choice: AWD is applied to all and only the alignment-sensitive terms (CLIP and ICL), whose supervision depends on image-text alignment, while the distillation terms (CRD and FD) remain unweighted and can still transfer useful teacher supervision from poorly aligned pairs.

Table 4: **Ablation of objective weight** (IN-1k).

Loss objective	CLIP	ICL	CRD	FD	ZS	LP
CLIP-KD					48.3	68.2
+ CLIP weight	✓	×	×	×	50.1	69.5
+ AWD weight	✓	✓	×	×	**51.0**	**69.6**
+ uniform weight	✓	✓	✓	✓	49.3	68.7

Online alignment scoring. In our default implementation, alignment scores are computed directly from the teacher forward pass used for distillation, on the 224-resolution training views, rather than from embeddings precomputed at the teacher's native 448 resolution. The corresponding calibration statistics are estimated online (Eq. 6), removing the need for an additional dataset pass. Despite these approximations, online scoring performs on par with the offline reference: IN-1k zero-shot accuracy is identical (51.0), while linear probing is 69.7 vs. 69.6. We further analyze the agreement between online and offline alignment scores in Sec. A.5. These results show that AWD can be implemented entirely within the standard distillation forward pass without sacrificing efficiency.

Effect of corpus alignment quality. Our previous results compare two pretraining datasets with different overall levels of image-text alignment. We now investigate more precisely how corpus alignment quality affects distillation through a controlled experiment. To isolate this effect, we construct controlled subsets of DataComp-Medium by varying the fraction ρ of low-alignment image-text pairs while keeping the dataset size fixed (see Sec. A.6 for the construction). In Fig. 2, we can see that, as the fraction of noisy captions increases, the absolute performance of both CLIP-KD and CLIP-KD + AWD decreases substantially. Nevertheless, using AWD consistently improves over CLIP-KD across the intermediate noise regimes, with gains of roughly $+2$ points, and up to $+2.6$ at $\rho = 75\%$. The gain remains positive at $\rho = 87.5\%$, but nearly vanishes when all captions are low-alignment ($+0.1$ at $\rho = 100\%$). This suggests that alignment weighting is most beneficial when reliable and unreliable correspondences coexist, while its advantage naturally diminishes when little reliable correspondence supervision remains.

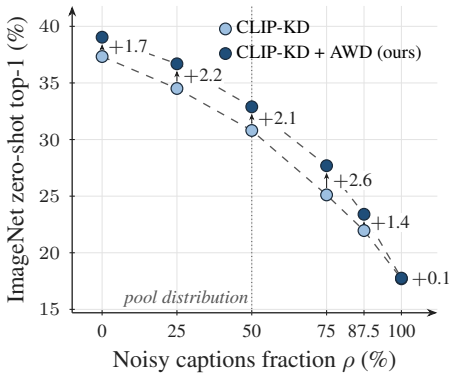


Figure 2: **AWD vs. corpus alignment quality.** IN-1k zero-shot accuracy on DataComp-M subsets with varying fractions of noisy captions.

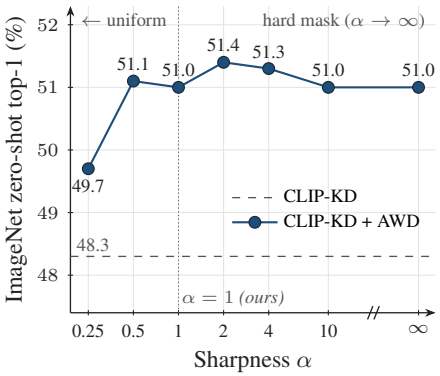


Figure 3: **Sensitivity to the sharpness α.** ImageNet-1k zero-shot accuracy of CLIP-KD + AWD on CC12M against α; dashed line: CLIP-KD without AWD ($\alpha = 0$).

Sensitivity to the sharpness α. We next examine the effect of the sharpness parameter α, which controls the sharpness of the alignment weights. As shown in Fig. 3, performance of IN-1k ZS remains consistently above the unweighted CLIP-KD baseline over a broad range of values, reaching 51.4 at $\alpha = 2$ and 51.0 with our default $\alpha = 1$, compared with 48.3 without AWD. Increasing α sharpens the weighting function, approaching a hard binary mask in which low-alignment pairs receive essentially no alignment-sensitive supervision, while their distillation losses remain active. Conversely, decreasing α makes the weights progressively more uniform, eventually approaching unit weights ($w_i = 1$) and recovering the standard CLIP-KD objective. The broad plateau between these two extremes indicates that AWD does not rely on a narrowly tuned sharpness parameter α.

Training cost. As shown in Tab. 5, alignment weighting introduces virtually no computational overhead: adding AWD to CLIP-KD leaves the analytic FLOPs unchanged and increases the measured time per step by only 1%. In contrast, ACID methods require caching teacher features for the full dataset and introduce additional cost during training. I-ACID loads teacher features for a super-batch $5\times$ larger than the training batch and performs adaptive sampling, resulting in approximately $1.12\times$ the CLIP-KD cost. H-ACID additionally performs a student forward pass over this super-batch, increasing the cost to approximately $1.42\times$. Adding AWD on top of H-ACID incurs no further overhead, confirming that AWD is essentially free once the teacher forward pass required for distillation is available.

Table 5: **Training cost.** Computational overhead of alignment weighting and batch-selection variants.

Method	TFLOPs	s/step	Rel. cost	Time (h)
CLIP-KD	1.16	1.56	1.00×	21.7
+ AWD	1.16	1.57	1.01×	21.8
+ I-ACID	1.16	*1.74*	*1.12*×	*24.2*
+ H-ACID	1.36	*2.21*	*1.42*×	*30.7*
+ H-ACID + AWD	1.36	*2.20*	*1.41*×	*30.6*

5 Conclusion

In this paper, we revisit the common assumption that weakly aligned image-text pairs should be discarded when training vision-language models. While this is well motivated for contrastive pretraining, distillation introduces additional teacher supervision that can remain informative even when the observed image-text correspondence is unreliable. Based on this observation, we introduce alignment-aware distillation, which uses the teacher's image-text similarity to selectively modulate alignment-sensitive objectives while preserving teacher-derived supervision. Across different pretraining corpora and two complementary distillation protocols, CLIP-KD and TIPSv2, this simple strategy consistently improves student models while introducing negligible computational overhead. Our ablations further show that the benefit comes from distinguishing between supervision sources rather than globally reweighting or filtering samples. More broadly, our results suggest that data quality in knowledge distillation should not be treated solely as a property of the sample: the same training example can be unreliable for one objective yet informative for another.

ETHICS STATEMENT

This work studies methods for improving vision-language model distillation and does not involve human subjects or the collection of new personal data. Our experiments use existing large-scale image-text datasets, which may inherit biases, inaccuracies, or other undesirable content from web data. Our method additionally relies on pretrained teacher models to estimate image-text alignment and may therefore inherit biases present in these models. These limitations should be considered when deploying models trained with the proposed approach. We follow the usage conditions of the datasets and pretrained models employed in our experiments.

REPRODUCIBILITY STATEMENT

We provide the full formulation of our alignment-aware distillation, including its online alignment scoring and calibration procedure, in the main paper. The experimental section specifies the datasets, teacher and student architectures, optimization settings, distillation objectives, and evaluation protocols used in our experiments. Additional implementation details, data-processing procedures, ablations, and complete per-dataset results are provided in the Appendix. We will open-source the training and evaluation code together with the pretrained student weights to facilitate reproduction and further study of our results.

AI USE STATEMENT

The authors use a generative AI assistant (Claude, Anthropic) to write and debug code, to give feedback on the experimental methodology, to help analyze results, and to draft and edit the text, the figure code, and the bibliography entries. The assistant is not used to generate data or to formulate mathematical claims. The synthetic captions of the recaptioning experiment (Tab. 3) are produced by Qwen3-VL-4B-Instruct as part of the evaluated setting, not as an aid to writing or analysis. The authors have reviewed all AI-assisted work and take full responsibility for the content of this paper.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 1

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *NeurIPS*, 2025. 1 and 3

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014. 15

Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, Joshua Ainslie, Alex Bewley, Mithun Jacob, René Wagner, Washington Ramos, Krzysztof Choromanski, Mojtaba Seyedhosseini, Howard Zhou, and André Araujo. TIPSv2: Advancing vision-language pretraining with enhanced patch-text alignment. In *CVPR*, 2026. 2, 3, and 6

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 6

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 *CVPR*, 2021. 6

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 1

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 15

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 6 and 15

Nicolas Dufour, Victor Besnier, Vicky Kalogeiton, and David Picard. Don't drop your samples! Coherence-aware training benefits conditional diffusion. In *CVPR*, 2024. 3

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. In *ICLR*, 2024. 3

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *CVIU*, 106(1): 59–70, April 2007. 15

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets. In *NeurIPS*, 2023. 1, 3, and 6

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 *ICCV*, 2021. 6 and 15

Wooyoung Kang, Jonghwan Mun, Sungjun Lee, and Byungseok Roh. Noise-aware learning from web-crawled image-text data for image captioning. In *ICCV*, 2023. 3

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *CVPR*, 2012. 15

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 6 and 15

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 15

Kevis-Kokitsi Maninis, Kaifeng Chen, Soham Ghosh, Arjun Karpur, Koert Chen, Ye Xia, Bingyi Cao, Daniel Salz, Guangxing Han, Jan Dlabal, et al. TIPS: Text-image pretraining with spatial awareness. In *ICLR*, 2025. 6

Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. SILC: Improving vision language pretraining with self-distillation. In *ECCV*, 2024. 6

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 15

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 15

Qwen. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*, 2026. 1

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 *ICML*, 2021. 1

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, 2019. 6 and 15

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: hints for thin deep nets. In *ICLR*, 2015. 1

Mert Bulent Sariyildiz, Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. UNIC: Universal classification models via multi-teacher distillation. In *ECCV*, 2024. 5

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. *arXiv preprint arXiv: 2111.02114*, 2021. Data Centric AI NeurIPS Workshop 2021. 3

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 3 and 21

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-video generation without text-video data. In *ICLR*, 2023. 1

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 15

Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. EVA-CLIP-18B: Scaling CLIP to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024. 1

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 6

Vishaal Udandarao, Nikhil Parthasarathy, Muhammad Ferjad Naeem, Talfan Evans, Samuel Albanie, Federico Tombari, Yongqin Xian, Alessio Tonioni, and Olivier J. Hénaff. Active data curation effectively distills large-scale multimodal models. In *CVPR*, 2025. 2, 3, and 15

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 6 and 15

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 *ICCV*, 2023. 1 and 3

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 15

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 *CVPR*, 2024. 1, 2, 3, 4, 5, 15, and 20

Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. ALIP: Adaptive language-image pre-training with synthetic caption. In *ICCV*, 2023. 3

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*, 2014. 6 and 15

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 1

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT pre-training with online tokenizer. In *ICLR*, 2022a. 6

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022b. 15

A APPENDIX

This appendix provides implementation details and additional analyses. Sec. A.1 describes the baselines, Sec. A.2 the evaluation benchmarks, and Sec. A.3 the training hyperparameters of both distillation settings. Sec. A.4 compares the alignment-score distributions of CC12M and DataComp-M, Sec. A.5 measures the agreement between online and offline teacher scores, and Sec. A.6 explains how the controlled alignment-quality subsets of DataComp-M are built. Sec. A.7 reports the per-dataset fine-grained zero-shot results, Sec. A.8 applies AWD to a simpler CLIP+FD objective and varies the FD weight, and Sec. A.9 scales the full recipe to LAION.

A.1 Baselines 15
A.2 Details on evaluation benchmarks 15
A.3 Training details 16
A.4 Alignment-score distributions of the training corpora 17
A.5 Online vs. offline alignment scores 17
A.6 Controlled alignment-quality subsets of DataComp-M 18
A.7 Detailed fine-grained zero-shot classification results 19
A.8 Alignment weighting under a feature-distillation-only objective 20
A.9 Scaling alignment weighting to LAION 21

A.1 BASELINES

CLIP-KD. CLIP-KD (Yang et al., 2024) distills a CLIP teacher into a smaller student through a combination of objectives. We use the standard CLIP contrastive loss together with interactive contrastive learning (ICL), contrastive relational distillation (CRD) and cosine feature distillation (FD), with the weights given in Tab. 6. We note that, for FD, we did not use the MSE distance because its range depends on the teacher embedding dimension, which make the loss weight ($\lambda_{\mathrm{FD}} = 2000$) provided in Yang et al. (2024) not transferable across teachers with different embedding dimensions (see Sec. A.8 for more details). CLIP-KD trains on uniformly sampled batches, and serves as our primary baseline.

Active data curation (I-ACID and H-ACID). ACID (Udandarao et al., 2025) treats distillation as an online batch-selection problem: rather than changing the objective, a large frozen teacher model guides which examples are presented to the student. At each training step, a super-batch $\mathcal{S}$ of candidate image–text pairs is sampled uniformly, from which a training batch $\mathcal{B} \subset \mathcal{S}$ is selected. For filtering ratio $f \in [0, 1)$, its size is $|\mathcal{B}| = (1 - f)|\mathcal{S}|$.

Because the contrastive loss depends on the other examples in the batch, ACID constructs $\mathcal{B}$ iteratively. Let $\mathcal{B}_r$ denote the subset selected after iteration r, and $\mathcal{L}(\mathcal{C}; \theta)$ the image–text contrastive loss evaluated on a batch $\mathcal{C}$ using the model with weights θ. Each remaining candidate $x_i \in \mathcal{S} \setminus \mathcal{B}_r$ is scored by evaluating the temporary batch $\mathcal{B}_r \cup \{x_i\}$. In our experiments, we use the softmax contrastive loss to match the CLIP-like PE-Core teacher. Since this loss is zero for a singleton batch, the initial subset is sampled uniformly; subsequent iterations score candidates conditionally on the already selected subset, thereby accounting for their contrastive interactions with $\mathcal{B}_r$.

I-ACID uses the frozen teacher model θ_{T} and assigns

$$s_i^{\text{I-ACID}} = -\mathcal{L}(\mathcal{B}_r \cup \{x_i\}; \theta_{\mathrm{T}}), \tag{13}$$

thereby favoring candidates that yield a low contrastive loss under the teacher. H-ACID instead uses the *learnability* score

$$s_i^{\text{H-ACID}} = \mathcal{L}(\mathcal{B}_r \cup \{x_i\}; \theta_{\mathrm{S}}) - \mathcal{L}(\mathcal{B}_r \cup \{x_i\}; \theta_{\mathrm{T}}), \tag{14}$$

where θ_{S} denotes the current student weights. It favors candidates well modeled by the teacher but remain difficult for the student. The resulting scores define the conditional sampling distribution

$$p(x_i \mid \mathcal{B}_r) \propto \exp(s_i), \tag{15}$$

used to extend $\mathcal{B}_r$ until the target training-batch size is reached.

In our experiments, we apply I-ACID and H-ACID selection on top of the full CLIP-KD objective, so that the comparison isolates the effect of data selection. We otherwise use the hyperparameters proposed in the original paper.

Hard filtering. As a static, offline alternative, we compute the teacher's image-text cosine similarity for every training pair once before training and discard the tertile with the lowest scores. The student is then trained on the remaining pairs with the standard CLIP-KD objective for the same number of steps. Unlike ACID, this removes data permanently and does not adapt to the student during training.

A.2 DETAILS ON EVALUATION BENCHMARKS

On ImageNet-1k (Deng et al., 2009), we evaluate zero-shot classification, linear probing, and weighted k-NN on frozen representations. To assess robustness under distribution shift, we report zero-shot accuracy on ImageNet-V2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2021), and ImageNet-Sketch (Wang et al., 2019). We evaluate image-to-text and text-to-image retrieval on COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014) using the standard Karpathy test splits (5k and 1k images, respectively), reporting Recall@1. Finally, we measure zero-shot transfer on nine fine-grained classification datasets: Caltech101 (Fei-Fei et al., 2007), Oxford-Pets (Parkhi et al., 2012), StanfordCars (Krause et al., 2012), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), FGVC-Aircraft (Maji et al., 2013), SUN397 (Xiao et al., 2010), DTD (Cimpoi et al., 2014), and UCF101 (Soomro et al., 2012). For these, we follow the train/validation/test splits proposed by Zhou et al. (2022b), which are standard for evaluating vision-language models, and report top-1 accuracy on the test split.

A.3 TRAINING DETAILS

We summarize all the training hyperparameters of AWD in Tab. 6. We provide details for both of our distillation settings.

Table 6: Training hyperparameters for both distillation settings.

Hyperparameter	**CLIP-KD**	**TIPSv2**
Model		
Teacher (frozen)	PE-Core-G14-448	TIPSv2-g/14
Student image encoder	ViT-B/16	ViT-B/14
Student text encoder	12 layers, width 512, 8 heads	12 layers, width 512, 8 heads
Text context length	77	77
Tokenizer	CLIP BPE	CLIP BPE
Embedding dimension	512	512
Student initialization	from scratch	from scratch
Drop path rate	0	0
Objectives		
Losses	CLIP, ICL, CRD, FD	CLIP, DINO, iBOT
Loss weights	λ_{FD}=8, λ_{ICL}=λ_{CRD}=1	λ_{CLIP}=λ_{DINO}=1, λ_{iBOT}=2
Initial CLIP temperature	0.07	0.07
Initial ICL temperature	0.07	–
Learned temperatures	CLIP, ICL	CLIP
Maximum logit scale	100	100
CRD student temperature	shared with CLIP	–
CRD teacher temperature	teacher's own (frozen)	–
CRD negatives	in-batch	–
Self-supervised heads (TIPSv2)		
Number of global crops	–	1
Number of local crops	–	6
Global crop size	–	224
Local crop size	–	98
iBOT mask ratio (iBOT++)	–	0
Number of prototypes (DINO / iBOT)	–	32,768
Head hidden dimension	–	2048
Head bottleneck dimension	–	256
Student temperature	–	0.1
DINO teacher temperature	–	0.07
iBOT teacher temperature, start	–	0.04
iBOT teacher temperature, end	–	0.07
iBOT teacher temperature warmup steps	–	10k
Optimization		
Optimizer	AdamW	AdamW
β_1, β_2	0.9, 0.98	0.9, 0.98
ϵ	10^{-6}	10^{-6}
Peak learning rate	5×10^{-4}	5×10^{-4}
Weight decay	0.2	0.2
No weight decay on	biases, norms, temperatures, tokens	biases, norms, temperatures, tokens
Gradient clipping	none	none
LR schedule	warmup-stable-drop	warmup-stable-drop
Warmup (fraction of steps)	5%	5%
Linear drop (fraction of steps)	last 15%	last 15%
Final learning rate	5×10^{-5}	5×10^{-5}
Global batch size	16,384	16,384
Training steps (main)	50k	50k
Training steps (ablations)	20k	20k
Precision	bf16 mixed	bf16 mixed
Local batch size	256	256
Hardware	64× H100	64× H100
Soft-gate weighting		
sharpness α	1	1
EMA momentum β	0.99	0.99

A.4 ALIGNMENT-SCORE DISTRIBUTIONS OF THE TRAINING CORPORA

The weighting function of Sec. 3.2 carries no corpus-specific constant: its center and scale are estimated online from the teacher scores of the current batch. Fig. 4 shows why this is needed. The two corpora used in the experiments do not share the same score distribution. CC12M is unimodal and centered at 0.25, with a short left tail. DataComp-Medium is bimodal and centered at 0.20: besides a mode that overlaps the CC12M one, it carries a second mode near 0.15, made of pairs whose caption badly describes the image. A center and a scale fixed once for CC12M would therefore downweight a large part of DataComp-Medium, and conversely, whereas the running median and quartile deviation track each corpus and keep the weights centered on 1 in both cases.

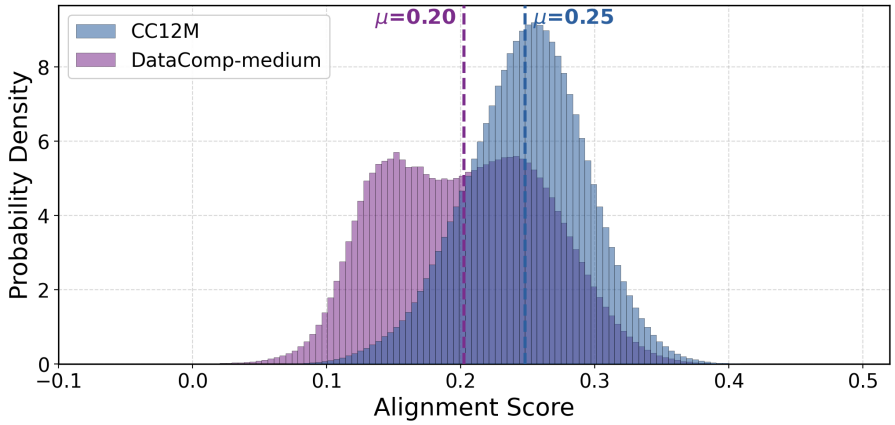


Figure 4: **Alignment-score distributions of the two training corpora.** Density histogram of the teacher image-text alignment score, i.e., the cosine similarity between the PE-Core-G14 image and text embeddings, over the 9M pairs of CC12M (blue) and over a random 9M subset of DataComp-Medium (purple). The dashed lines mark the mean of each corpus.

A.5 ONLINE VS. OFFLINE ALIGNMENT SCORES

By default, AWD reads the alignment score off the teacher forward pass that distillation already requires (Sec. 3.2). This online score departs from an offline reference in two ways: the teacher sees the 224×224 training view instead of the image at its native 448 resolution, and this view is a random crop (scale $[0.9, 1]$) instead of the full image. To separate the two factors, 200k CC12M pairs are scored three times by the PE-Core-G14 teacher: offline at 448 (the reference), offline on the full image resized to 224 (A), and online on the 224 training view (B).

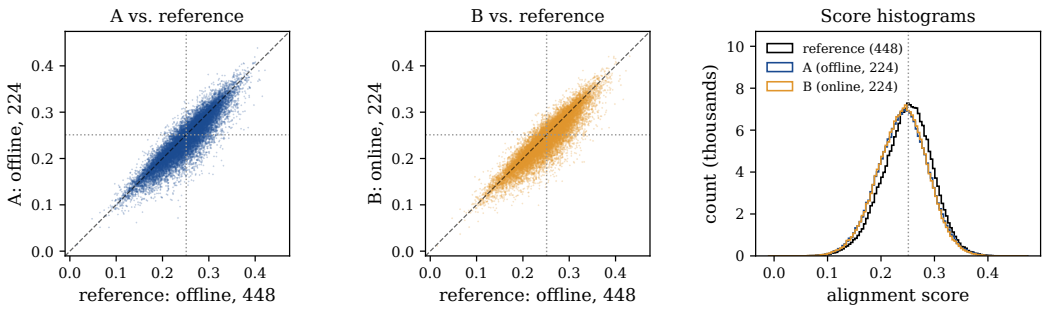


Figure 5: **Online and offline teacher alignment scores.** PE-Core-G14 alignment score of 200k CC12M pairs. Left: offline score of the full image resized to 224 (A) against the offline reference at the teacher's native 448 resolution. Middle: online score of the 224 training view (B) against the same reference; the dashed line is the identity. Right: histograms of the three scores. The dotted lines mark 0.25, the median of the reference scores.

Tab. 7 shows that both 224 scores correlate with the reference at a Spearman ρ of 0.90, while A and B agree with each other at $\rho = 0.96$. The disagreement with the reference therefore comes from the lower input resolution, and computing the score on the training crop inside the forward pass adds little to it. Fig. 5 shows the second effect of resolution: the whole distribution shifts down by about 0.01 (median 0.251 at 448, 0.242 at 224) with an unchanged spread (quartile deviation 0.031 vs. 0.031–0.032). Measured against the fixed reference median 0.25, this offset alone moves the share of pairs below the center from 49% to 57%, and caps the same-side agreement with the reference at 86%. The online calibration of Eq. 6 does not use a fixed center: it tracks the median and quartile deviation of the scores that actually enter the weights, so the offset is absorbed and only the rank disagreement measured by ρ remains. This is consistent with the ablation of Sec. 4.2, where online and offline scoring reach the same IN-1k zero-shot accuracy (51.0).

Table 7: **Agreement between online and offline teacher alignment scores.** 200k CC12M pairs scored by PE-Core-G14. Reference: offline score at 448; A: offline score at 224; B: online score of the 224 training view. Pearson and Spearman correlations, same-side agreement (fraction of pairs that fall on the same side of the reference median 0.25 under both scores), and mean score shift.

Comparison	Pearson r	Spearman ρ	Same side (%)	Mean shift
A (offline 224) vs. reference (448)	0.909	0.902	86.3	−0.009
B (online 224) vs. reference (448)	0.908	0.900	86.2	−0.009
B (online 224) vs. A (offline 224)	0.962	0.959	92.3	−0.001

A.6 Controlled alignment-quality subsets of DataComp-M

The experiment in Fig. 2 isolates the effect of pretraining alignment quality while keeping the training setting otherwise unchanged. Alternative constructions make this difficult to achieve. For example, partitioning the data into disjoint score ranges, such as low-, medium-, and high-alignment terciles, also changes the underlying image distribution and potentially the difficulty of downstream transfer. Similarly, progressively filtering to the top-p% of pairs changes the dataset size and, under a fixed step budget, the number of epochs seen during training. Since repeated exposure to the retained pairs is itself one of the effects induced by filtering, this would confound alignment quality with data reuse. We instead construct subsets of identical size from a common pool and vary only the proportion ρ of low-alignment pairs.

We begin from the 80M-pair DataComp-M pool and score all pairs with the PE-Core-G14 teacher using the 224×224 view employed by the weighting mechanism during training (Sec. A.5). The pool is divided at the median teacher score into higher- and lower-alignment halves. From each half, we sample four disjoint chunks of 2.5M pairs, denoted $C_1, \ldots, C_4$ (clean) and $N_1, \ldots, N_4$ (not clean), respectively. Each experimental subset contains four chunks, for a total of 10M pairs, with the clean-to-noisy composition varied to produce the six mixtures reported in Tab. 8. Consecutive mixtures retain three of the same four chunks, reducing variation due to resampling between neighbouring points. As ρ increases, the mean teacher score decreases monotonically, from 0.239 at $\rho = 0$ to 0.190 at $\rho = 50\%$ and 0.142 at $\rho = 100\%$. The $\rho = 50\%$ mixture contains equal proportions from the two halves and therefore preserves the overall score distribution of the original pool; we mark this point accordingly in Fig. 2.

Table 8: **Composition of the DataComp-M alignment-quality analysis.** Each subset holds 10M pairs. C_i and N_i are disjoint 2.5M chunks sampled from the clean (above-median) and noisy (below-median) half of the pool, and ρ is the fraction of pairs taken from the noisy half.

ρ	Chunks
0%	C_1, C_2, C_3, C_4
25%	C_1, C_2, C_3, N_1
50%	C_1, C_2, N_1, N_2
75%	C_1, N_1, N_2, N_3
87.5%	N_1, N_2, N_3, half N_4, half C_1
100%	N_1, N_2, N_3, N_4

For each mixture, we train a CLIP-KD student and its AWD counterpart from scratch using the same ablation setting as in Tab. 6: 20k optimization steps with global batch size 16,384, a 500-step warmup, and the drop phase over the final 3k steps. Thus, all twelve runs process the same

328M samples, corresponding to approximately 33 passes over each 10M-pair subset. The comparison therefore keeps both the optimization budget and the degree of sample reuse fixed across the alignment ladder. Fig. 2 reports ImageNet-1k zero-shot top-1 accuracy at the final checkpoint.

Two details are important for interpreting this experiment. First, the subsets are defined using offline teacher scores, whereas AWD computes its weights online during training. Both use the same 224×224 view, but the online scores are recomputed from the current training view and normalized using the EMA-tracked center described in Sec. A.5. Second, although the $\rho = 50\%$ mixture matches the score distribution of the full pool, it remains a 10M-pair subset trained with the 20k-step ablation budget. Its absolute accuracy should therefore not be compared directly with the full-pool 50k-step results in Tab. 1.

A.7 Detailed fine-grained zero-shot classification results

The detailed results in Tab. 9 show that the gains from alignment weighting are broadly distributed across the fine-grained evaluation datasets. On CC12M, CLIP-KD + AWD improves all nine benchmarks, raising the average from 51.6 to 55.3 (+3.7). AWD also remains complementary to adaptive selection, bringing I-ACID from +0.7 to +3.6 over CLIP-KD, and H-ACID from −0.4 to +1.9.

The same trend holds on DataComp-M, where AWD improves CLIP-KD from 56.0 to 57.2 and further strengthens I-ACID, reaching 60.1 average accuracy (+4.1 over CLIP-KD). By contrast, the effect of explicit data cleaning is more corpus-dependent: filtering helps on DataComp-M but hurts on CC12M, while recaptioning strongly benefits CC12M but slightly degrades DataComp-M. In these settings, AWD generally preserves or improves performance, suggesting that alignment weighting is most consistently useful as a complement to distillation and adaptive sample selection, whereas its benefit after explicit data cleaning depends on the remaining alignment variability.

Table 9: **Fine-grained zero-shot classification results.** Avg. Δ denotes the average performance difference with respect to CLIP-KD across all reported datasets.

	Method	*Caltech*	*Pets*	*Cars*	*Flowers*	*Food*	*Aircraft*	*SUN*	*DTD*	*UCF*	*Average*	**Avg. Δ**
	Teacher: PE-Core-G14	97.2	96.0	94.4	89.2	93.0	70.1	77.0	63.7	85.2	85.1	–
CC12M (9M)	CLIP-KD	86.7	84.1	38.6	46.9	60.1	5.8	59.3	34.7	48.2	51.6	–
	+ AWD	**89.3**	**84.6**	**41.9**	**50.5**	**64.4**	**8.3**	**64.0**	**39.5**	**55.0**	**55.3**	**+3.7**
	+ I-ACID	88.6	81.3	40.3	44.2	61.6	7.3	60.9	34.9	51.6	52.3	+0.7
	+ I-ACID + AWD	**88.9**	**83.9**	**43.6**	**48.8**	**64.8**	**9.0**	**63.9**	**41.0**	**53.1**	**55.2**	**+3.6**
	+ H-ACID	85.8	80.9	38.9	44.4	60.0	6.9	59.9	33.3	50.6	51.2	−0.4
	+ H-ACID + AWD	**88.8**	**83.2**	**40.6**	**47.0**	**62.2**	**7.4**	**62.3**	**36.9**	**53.1**	**53.5**	**+1.9**
	+ Filtering	84.8	**80.4**	**36.5**	**44.6**	**56.8**	**7.4**	**58.6**	36.5	**49.1**	**50.5**	−1.1
	+ Filtering + AWD	**85.7**	80.3	**36.5**	42.5	56.5	7.0	58.5	**39.1**	47.1	50.4	−1.2
	+ Recaption	**91.9**	82.0	45.9	49.5	69.4	**10.5**	63.7	**48.9**	59.7	57.9	+6.3
	+ Recaption + AWD	91.3	**82.4**	**46.8**	**50.9**	**70.7**	9.2	**65.2**	47.1	**60.6**	**58.2**	**+6.6**
DataComp-M (80M)	CLIP-KD	90.5	70.0	65.4	**46.2**	70.4	5.5	60.3	**44.0**	51.6	56.0	–
	+ AWD	**91.4**	**74.9**	**67.3**	44.4	**72.0**	**7.0**	**61.5**	43.1	**52.9**	**57.2**	+1.2
	+ I-ACID	91.9	76.4	70.6	49.5	72.2	**8.3**	61.7	45.7	51.6	58.7	+2.7
	+ I-ACID + AWD	**92.1**	**76.5**	**72.7**	**52.0**	**74.3**	8.0	**62.9**	**47.9**	**54.8**	**60.1**	**+4.1**
	+ H-ACID	91.4	73.0	67.2	49.5	**73.6**	**7.6**	60.5	**46.9**	**54.2**	58.2	+2.2
	+ H-ACID + AWD	**91.8**	**76.5**	**67.8**	**50.8**	72.9	7.4	**61.7**	45.6	53.8	**58.7**	**+2.7**
	+ Filtering	**92.1**	72.2	70.8	**51.4**	72.1	6.0	**61.6**	46.3	**51.8**	58.3	+2.3
	+ Filtering + AWD	**92.1**	**75.3**	**71.8**	51.1	**73.4**	**7.7**	61.5	**46.9**	**51.8**	**59.1**	**+3.1**
	+ Recaption	91.0	67.5	53.3	43.4	67.9	**5.0**	59.9	50.7	58.0	55.2	−0.8
	+ Recaption + AWD	**91.3**	**69.0**	**56.9**	**46.5**	**69.0**	4.5	**62.6**	**51.2**	**58.2**	**56.6**	**+0.6**

Table 10: **Alignment weighting under a feature-distillation-only objective.** All students use the same ViT-B/16 architecture and are trained for 50k steps. CLIP+FD denotes the student trained with the contrastive and feature-distillation terms only; CLIP-KD adds the ICL and CRD terms. We report ImageNet-1k zero-shot, linear probing, weighted k-NN, distribution-shift classification, image-text retrieval, and average fine-grained classification performance. Avg. Δ denotes the average performance difference with respect to the corresponding unweighted baseline across all reported metrics.

	Method	ImageNet-1k			Distrib. Shift			I2T retrieval		T2I retrieval		Fine-grained	Avg. Δ
		ZS	LP	k-NN	IN-V2	IN-R	IN-S	COCO	Flickr	COCO	Flickr		
CC12M	CLIP+FD	49.7	73.0	67.8	42.8	63.1	37.5	48.0	77.9	33.6	63.0	44.4	–
	+ AWD	**54.9**	**75.3**	**70.2**	**47.7**	**68.4**	**42.5**	**53.6**	**85.3**	**38.5**	**67.9**	**48.7**	**+4.7**
	CLIP-KD	53.4	70.9	66.3	45.6	65.2	41.4	51.0	77.2	35.3	63.1	51.6	–
	+ AWD	**57.3**	**73.1**	**68.5**	**49.5**	**69.8**	**44.7**	**54.7**	**80.6**	**38.8**	**67.6**	**55.3**	**+3.5**
DC-M	CLIP+FD	50.0	74.0	67.7	42.5	60.4	38.7	43.9	71.4	28.6	53.2	51.7	–
	+ AWD	**51.6**	**74.3**	**68.1**	**44.8**	**61.5**	**40.4**	**45.5**	**75.5**	**30.3**	**55.5**	**53.6**	**+1.7**
	CLIP-KD	52.8	74.0	68.1	45.8	62.6	41.1	46.6	71.5	30.4	55.1	56.0	–
	+ AWD	**54.4**	**74.2**	**68.9**	**47.2**	**64.9**	**42.3**	**47.4**	**71.6**	**31.3**	**57.0**	**57.2**	**+1.1**

A.8 Alignment weighting under a feature-distillation-only objective

To verify that the benefits of AWD are not specific to the full CLIP-KD recipe, we consider a simpler distillation setting using only the CLIP contrastive loss and feature distillation (FD), without ICL and CRD.

Both this baseline and the full recipe need a weight for the FD term. CLIP-KD (Yang et al., 2024) implements FD as a mean squared error (MSE) between the ℓ_2-normalized projected student embedding $\hat{\mathbf{s}}$ and the normalized teacher embedding $\hat{\mathbf{t}}$, averaged over the D_{T} teacher dimensions, with weight $\lambda_{\mathrm{MSE}} = 2000$. Our FD uses instead the cosine distance between the embeddings $\mathbf{s}$ and $\mathbf{t}$. On unit vectors these two losses are proportional, $\|\hat{\mathbf{s}} - \hat{\mathbf{t}}\|^2 / D_{\mathrm{T}} = 2\,(1 - \cos(\mathbf{s}, \mathbf{t}))/D_{\mathrm{T}}$, so they have the same value and gradient when $\lambda_{\cos} = 2\,\lambda_{\mathrm{MSE}}/D_{\mathrm{T}}$. The MSE weight therefore depends on the teacher embedding dimension, while the cosine weight does not. In the CLIP-KD ablation setting (ViT-B/16 teacher, $D_{\mathrm{T}} = 512$), $\lambda_{\mathrm{MSE}} = 2000$ corresponds to $\lambda_{\cos} \approx 7.8$, which motivates setting $\lambda_{\mathrm{FD}} = 8$. Keeping $\lambda_{\mathrm{MSE}} = 2000$ with the PE-Core-G14 teacher ($D_{\mathrm{T}} = 1280$) would instead give $\lambda_{\cos} \approx 3.1$.

As shown in Tab. 10, AWD consistently improves this CLIP+FD baseline across both corpora and every evaluation family, yielding average gains of $+4.7$ points on CC12M and $+1.7$ on DataComp-M. On CC12M, ImageNet-1k zero-shot accuracy increases from 49.7 to 54.9, while the gains also extend to all three distribution shifts and all four retrieval metrics; for instance, ImageNet-R improves from 63.1 to 68.4 and Flickr30k image-to-text retrieval from 77.9 to 85.3. The same consistency holds on DataComp-M, where AWD improves every reported shift and retrieval metric. Interestingly, CLIP+FD is not uniformly weaker than the full CLIP-KD recipe: on CC12M, it achieves higher linear-probe and k-NN accuracy ($73.0/67.8$ vs. $70.9/66.3$), and this advantage is preserved with AWD ($75.3/70.2$ vs. $73.1/68.5$). This indicates that the additional CLIP-KD objectives change the balance between downstream properties rather than uniformly improve every evaluation criterion.

The fine-grained results in Tab. 11 provide the same picture. With CLIP+FD, AWD improves the average by $+4.3$ points on CC12M and $+1.9$ on DataComp-M, with gains on seven of nine and eight of nine evaluation datasets, respectively. Moreover, the overall gains from AWD under CLIP+FD are larger than those obtained with the full CLIP-KD objective on both corpora ($+4.7$ vs. $+3.5$ on CC12M and $+1.7$ vs. $+1.1$ on DataComp-M). These results show that the effectiveness of alignment weighting does not depend on ICL or CRD and extends to a simpler contrastive-plus-feature-distillation objective.

Tab. 12 varies λ_{FD} over $\{1, 3, 8\}$ for CLIP+FD and CLIP-KD. Both baselines improve as the weight increases, and $\lambda_{\mathrm{FD}} = 8$ gives the best zero-shot and linear-probing accuracy in both cases. AWD improves its baseline at every weight, by $+2.1$ to $+4.8$ points in zero-shot and $+1.3$ to $+2.3$ in linear probing. The gains reported in the main paper thus do not depend on the choice of λ_{FD}.

Table 11: **Fine-grained zero-shot classification under a feature-distillation-only objective.** Per-dataset top-1 accuracy for the students of Tab. 10. Avg. Δ denotes the average performance difference with respect to the corresponding unweighted baseline across all reported datasets.

	Method	*Caltech*	*Pets*	*Cars*	*Flowers*	*Food*	*Aircraft*	*SUN*	*DTD*	*UCF*	*Average*	**Avg. Δ**
CC12M	CLIP+FD	78.4	72.9	28.8	38.3	54.2	**4.5**	50.6	**28.3**	43.3	44.4	–
	+ AWD	**87.0**	**75.8**	**35.5**	**42.7**	**58.7**	4.3	**56.2**	28.2	**49.7**	**48.7**	**+4.3**
	CLIP-KD	86.7	84.1	38.6	46.9	60.1	5.8	59.3	34.7	48.2	51.6	–
	+ AWD	**89.3**	**84.6**	**41.9**	**50.5**	**64.4**	**8.3**	**64.0**	**39.5**	**55.0**	**55.3**	**+3.7**
DC-M	CLIP+FD	87.4	64.6	58.1	38.6	67.3	3.9	57.9	38.5	**48.7**	51.7	–
	+ AWD	**90.0**	**66.1**	**59.7**	**43.0**	**69.9**	**5.6**	**59.7**	**41.4**	47.1	**53.6**	**+1.9**
	CLIP-KD	90.5	70.0	65.4	**46.2**	70.4	5.5	60.3	**44.0**	51.6	56.0	–
	+ AWD	**91.4**	**74.9**	**67.3**	44.4	**72.0**	**7.0**	**61.5**	43.1	**52.9**	**57.2**	**+1.2**

Table 12: **Feature-distillation weight ablation.** ImageNet-1k zero-shot (ZS) and linear-probing (LP) accuracy of CLIP+FD and CLIP-KD, with and without AWD, trained for 20k steps on CC12M. All FD terms share the weight $\lambda_{\text{FD}} \in \{1, 3, 8\}$; $\lambda_{\text{FD}} = 8$ is the setting used in the rest of the paper.

	$\lambda_{\text{FD}} = 1$		$\lambda_{\text{FD}} = 3$		$\lambda_{\text{FD}} = 8$	
Method	ZS	LP	ZS	LP	ZS	LP
CLIP+FD	30.4	58.2	36.1	64.4	43.0	68.9
+ AWD	**33.2**	**60.1**	**39.7**	**66.7**	**47.8**	**70.7**
CLIP-KD	41.9	62.3	44.5	62.9	48.3	68.2
+ AWD	**44.6**	**64.0**	**46.6**	**64.2**	**51.0**	**69.6**

A.9 Scaling alignment weighting to LAION

To test whether alignment weighting carries over to a corpus of much larger scale, we train the full CLIP-KD recipe with and without AWD on 402M image-text pairs drawn from LAION-2B-en, the English subset of LAION-5B (Schuhmann et al., 2022), using the original web captions, for the same 50k steps as on CC12M and DataComp-M. As shown in Tab. 13, AWD improves the average over all reported metrics by $+0.9$ points. ImageNet-1k zero-shot accuracy increases from 61.6 to 62.3, and all four retrieval metrics improve, with the largest gains on image-to-text retrieval ($56.7 \rightarrow 58.5$ on COCO and $82.4 \rightarrow 84.6$ on Flickr30k). Under distribution shift, ImageNet-R and ImageNet-Sketch improve by $+1.8$ and $+1.2$ points, while ImageNet-V2 stays level (53.8 vs. 53.7). Weighted k-NN accuracy improves from 72.0 to 72.6, while linear probing is unchanged at 76.6.

Table 13: **Scaling alignment weighting to LAION.** All students use the same ViT-B/16 architecture and are trained for 50k steps on the 402M-pair LAION corpus with its original web captions. We report ImageNet-1k zero-shot, linear probing, weighted k-NN, distribution-shift classification, image-text retrieval, and average fine-grained classification performance. Avg. Δ denotes the average performance difference with respect to CLIP-KD across all reported metrics.

	Method	ImageNet-1k			Distrib. Shift			I2T retrieval		T2I retrieval		Fine-grained	**Avg. Δ**
		ZS	LP	k-NN	IN-V2	IN-R	IN-S	COCO	Flickr	COCO	Flickr		
	Teacher: PE-Core-G14	85.4	89.5	85.5	80.2	96.5	76.5	75.4	96.2	58.1	85.7	85.1	–
LAION	CLIP-KD	61.6	**76.6**	72.0	**53.8**	74.3	49.5	56.7	82.4	38.8	67.4	65.0	–
	+ AWD	**62.3**	**76.6**	**72.6**	53.7	**76.1**	**50.7**	**58.5**	**84.6**	**39.9**	**68.1**	**65.1**	**+0.9**

The fine-grained results in Tab. 14 are more balanced: AWD improves Pets ($+2.0$), Flowers ($+1.9$), DTD ($+0.7$) and Aircraft ($+0.4$), matches CLIP-KD on SUN397, and trails it on the remaining four evaluation datasets, leaving the average nearly unchanged ($65.0 \rightarrow 65.1$). The overall gain on LAION is thus smaller than on CC12M ($+3.5$) and DataComp-M ($+1.1$), but it remains positive, driven by the alignment-sensitive evaluations: zero-shot classification and retrieval. LAION is itself

collected by filtering on a CLIP image-text similarity threshold, which leaves less alignment variability for AWD to exploit. This is consistent with the filtering and recaptioning results, where the benefit of AWD depends on how much misalignment remains in the training corpus.

Table 14: **Fine-grained zero-shot classification on LAION.** Per-dataset top-1 accuracy for the students of Tab. 13. Avg. Δ denotes the average performance difference with respect to CLIP-KD across all reported datasets.

	Method	*Caltech*	*Pets*	*Cars*	*Flowers*	*Food*	*Aircraft*	*SUN*	*DTD*	*UCF*	*Average*	**Avg. Δ**
	Teacher: PE-Core-G14	97.2	96.0	94.4	89.2	93.0	70.1	77.0	63.7	85.2	85.1	–
LAION	CLIP-KD	**94.4**	84.0	**80.8**	56.4	**72.7**	12.3	**66.1**	54.6	**63.8**	65.0	–
LAION	+ AWD	94.0	**86.0**	79.7	**58.3**	71.5	**12.7**	**66.1**	**55.3**	62.7	**65.1**	**+0.1**