DARK: DIAGONAL-ANCHORED REPULSIVE KNOWLEDGE DISTILLATION FOR VISION-LANGUAGE MODELS UNDER EXTREME COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models provide useful representations for fetal ultrasound, but their computational cost makes deployment on portable devices difficult. Knowledge distillation offers a route to smaller models, although continued imitation of a large teacher may constrain a much smaller student. We propose Diagonal-Anchored Repulsive Knowledge Distillation (DARK), which preserves supervision on matched image–caption pairs while gradually changing supervision on other pairs from imitation to repulsion. This lets the student first learn the teacher's relationships and then depart from them without abandoning matched-pair alignment. We use DARK to compress FetalCLIP into MobileFetalCLIP, whose visual encoder has approximately $26\times$ fewer parameters and runs in 1.6 ms on an iPhone 16 Pro. Compared with static distillation, DARK improves brain sub-plane macro-F1 from 0.701 to 0.775. On an external cohort of 403 images from five African countries, anatomical macro-F1 rises from 0.770 to 0.895, with gains in all four classes. Ablations show that the sign change is essential: fixed decoupled weights, as in decoupled KD, and stopping imitation at zero both fall short. Linear probes on the compact model's frozen features retain 97–98% of the teacher's downstream performance. These results support selective departure from teacher imitation as a way to learn compact fetal-ultrasound representations for both zero-shot recognition and supervised downstream use.

1 INTRODUCTION

Vision-language models (VLMs) learn visual representations from images and text, supporting both text-based recognition and downstream classifiers. In fetal ultrasound, FetalCLIP demonstrates the value of domain-specific pretraining across several assessment tasks (Maani et al., 2026). Bringing such models to portable ultrasound could extend prenatal assessment to settings with limited resources (Stewart et al., 2020; Pokaprakarn et al., 2022; Gomes et al., 2022), but their computational cost remains a barrier. Our goal is to retain useful fetal-ultrasound representations in a visual encoder more than an order of magnitude smaller. We assess zero-shot recognition, linear classification with frozen features, and mobile runtime.

Knowledge distillation (KD) trains a smaller student using supervision from a larger teacher (Hinton et al., 2015). For CLIP-style models, TinyCLIP combines affinity mimicking with weight inheritance (Wu et al., 2023), while CLIP-KD investigates several knowledge-transfer objectives (Yang et al., 2024). Distillation can nevertheless become less effective as the capacity gap grows (Cho & Hariharan, 2019; Mirzadeh et al., 2020). This raises a question: should a much smaller student continue reproducing all of the teacher's image–text relationships throughout training?

In a training batch, an image has a paired caption but also has teacher-assigned similarities to other captions. These non-target relationships can be semantically meaningful without all being useful constraints for a compact student. We hypothesize that the student benefits from learning them early in training, then departing from them while retaining supervision on matched pairs.

To test this hypothesis, we introduce **Diagonal-Anchored Repulsive Knowledge Distillation (DARK)**. DARK separates distillation into matched-pair and non-target terms, corresponding to diagonal and off-diagonal entries in the image–text similarity matrix. The non-target coefficient de-

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.

creases linearly from positive to negative, turning imitation into repulsion. Matched-pair supervision and the student's contrastive objective continue to encourage alignment throughout training.

We compress FetalCLIP into **MobileFetalCLIP**, a student with a FastViT visual encoder (Vasu et al., 2023). The visual encoder has approximately $26\times$ fewer parameters than FetalCLIP's and runs approximately $24\times$ faster on an iPhone 16 Pro. Across repeated runs, DARK improves brain sub-plane recognition and the consistency of age estimates with head circumference over static distillation, with little change in five-plane recognition. It exceeds the teacher on the first two tasks but remains below it on five-plane recognition. On an external cohort from five African countries, anatomical macro-F1 rises from 0.770 to 0.895 over static distillation. Linear classifiers fitted to frozen features achieve near-teacher scores on anatomical classification and congenital heart disease detection.

Ablations show that DARK's gain needs both of its parts. Fixed decoupled weights, as in DKD, and stopping non-target imitation at zero both fall short, so the sign change matters; reversing matched-pair supervision also falls short, so the anchor matters. An exploratory size study finds that DARK's benefit is concentrated at extreme compression, the regime that motivates this work. Our contributions are:

- **Selective repulsion for contrastive distillation.** DARK anchors matched-pair supervision and schedules non-target supervision from attraction to repulsion.
- **MobileFetalCLIP.** A compact fetal-ultrasound VLM supporting zero-shot recognition and downstream classification with fast mobile visual encoding.
- **Evidence for the distillation strategy.** Controlled comparisons and ablations support DARK's design, and external evaluation demonstrates improved anatomical recognition over static distillation across five countries.

2 Related Work

Knowledge Distillation under Capacity Constraints. Distillation can transfer output probabilities (Hinton et al., 2015), intermediate features (Romero et al., 2015), or distances and angles between representations (Park et al., 2019). Contrastive Representation Distillation instead uses a contrastive objective (Tian et al., 2020). For vision-language models, CLIP-KD studies feature and image–text similarity matching (Yang et al., 2024), while TinyCLIP combines affinity mimicking with weight inheritance (Wu et al., 2023). DARK builds on similarity distillation and changes its supervision during training.

Teacher accuracy alone does not determine distillation quality: large capacity gaps can reduce transfer (Cho & Hariharan, 2019; Mirzadeh et al., 2020), and closer prediction matching does not always improve student generalization (Stanton et al., 2021). Teacher assistants address this gap through intermediate models (Mirzadeh et al., 2020); other approaches relax prediction matching through correlation-based objectives (Huang et al., 2022) or logit standardization (Sun et al., 2024). Stopping distillation before student training ends can also improve performance (Cho & Hariharan, 2019). Our stopping control separates the effect of non-target repulsion from the benefit of withdrawing imitation. Students have also surpassed their teachers with the same parameter count (Furlanello et al., 2018) and approximately tenfold fewer parameters (Romero et al., 2015).

Decoupled Knowledge Distillation (DKD) (Zhao et al., 2022) is closely related in its separate treatment of target and non-target supervision. DKD independently weights a binary target-versus-rest KL divergence and a KL divergence between renormalized non-target distributions. DARK partitions symmetric image–text cross-entropy into matched and non-target entries, with probabilities normalized over the full set of candidates in each direction. The central difference is that DARK's non-target coefficient changes sign during training, turning imitation into repulsion while matched-pair supervision remains attractive. Our experiments test this difference directly: DKD-style fixed decoupled weights fall well short of DARK (Section 4.3).

Regularization and Representation Geometry. Barlow Twins reduces redundancy between embedding dimensions (Zbontar et al., 2021), while Wang & Isola (2020) characterize contrastive learning through positive-pair alignment and feature uniformity. These perspectives inform our representation analysis. Unlike confidence regularization, which encourages higher-entropy pre-

dictions (Pereyra et al., 2017), DARK weights repulsion by teacher probabilities. Our confidence-penalty baseline tests this distinction.

Vision-Language Pretraining and Medical Specialization. CLIP established transferable zero-shot recognition through image–text pretraining (Radford et al., 2021). MobileCLIP and MobileCLIP2 pursue efficient models through lightweight architectures and multimodal reinforced training (Vasu et al., 2024; Faghri et al., 2025). In medical imaging, MedCLIP learns from unpaired images and text (Wang et al., 2022), while BiomedCLIP and UniMed-CLIP extend pretraining across diverse biomedical image types (Zhang et al., 2025; Khattak et al., 2024). CheXzero studies zero-shot chest-radiograph classification (Tiu et al., 2022), and FetalCLIP specializes pretraining to fetal ultrasound (Maani et al., 2026). We distill FetalCLIP into a MobileCLIP-based FastViT student (Vasu et al., 2023), testing whether a compact architecture retains useful domain-specific representations.

3 METHOD

DARK separates teacher supervision on matched image–caption pairs from similarities among other batch pairs. It gradually reverses non-target imitation while continuing to reinforce matched pairs. Figure 1 summarizes the framework.

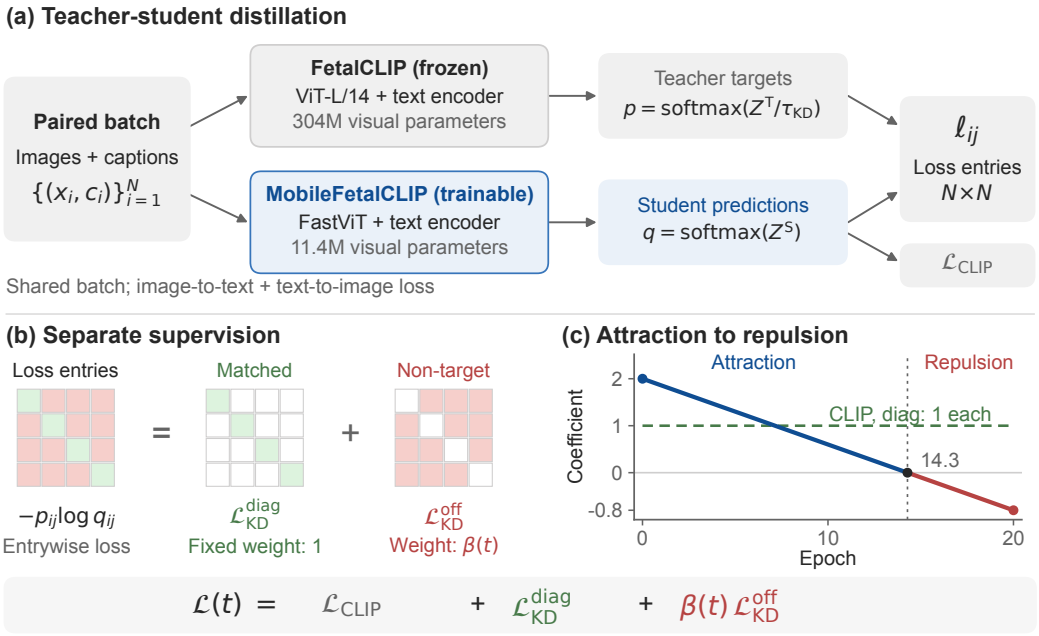

Figure 1: **DARK retains matched-pair supervision while reversing non-target imitation.** **(a)** A frozen FetalCLIP teacher and trainable MobileFetalCLIP student process the same batch of image–caption pairs. The student also learns through the CLIP contrastive loss. **(b)** Cross-entropy between teacher probabilities p and student probabilities q separates into matched (diagonal) and non-target (off-diagonal) contributions. The loss averages image-to-text and text-to-image directions; direction indices are omitted here. Colors indicate index roles, not measured values. **(c)** The non-target coefficient decreases from positive to negative, while matched-pair and CLIP coefficients remain one. Parameter counts refer to visual encoders.

3.1 PRELIMINARIES

A training batch contains N image–caption pairs, with image x_i paired with caption c_i. Each model encodes the images and captions separately, then compares every image with every caption in the batch. Let $m = \mathrm{T}$ denote the teacher and $m = \mathrm{S}$ the student. Their image and text encoders, f^m_{img} and f^m_{txt}, produce feature vectors normalized to length one. The score for image i and caption j is

their cosine similarity multiplied by a positive scale a_m:

$$Z_{ij}^m = a_m f_{\text{img}}^m(x_i)^\top f_{\text{txt}}^m(c_j). \tag{1}$$

These scores, or logits, form an $N \times N$ matrix Z^m whose diagonal contains the matched pairs. Teacher and student use the same pair ordering, so their score matrices can be compared even when their embedding dimensions differ.

For an image query, a softmax converts its scores into probabilities over the batch captions. The reverse direction uses each caption as a query and the images as candidates. We write these directions as $d \in \mathcal{D} = \{\text{img} \to \text{txt}, \text{txt} \to \text{img}\}$, with $Z^{m,d} = Z^m$ for image queries and $Z^{m,d} = (Z^m)^\top$ for text queries. The teacher temperature $\tau_{\text{KD}} > 0$ controls how concentrated its probabilities are; larger values give a softer distribution. Normalizing each row gives teacher probabilities p_i^d and student probabilities q_i^d:

$$p_i^d = \text{softmax}(Z_{i,:}^{\text{T},d}/\tau_{\text{KD}}), \qquad q_i^d = \text{softmax}(Z_{i,:}^{\text{S},d}). \tag{2}$$

The temperature is applied only to the teacher logits. An entry p_{ij}^d or q_{ij}^d is the probability assigned to candidate j for query i; candidate i is the matched pair in either direction. The symmetric CLIP loss encourages the student to identify this matched candidate (Radford et al., 2021):

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N} \sum_{d \in \mathcal{D}} \sum_i \log q_{ii}^d. \tag{3}$$

3.2 Separating Matched and Non-target Supervision

Standard distillation trains the student to imitate the teacher's full probability distribution, following CLIP-KD (Yang et al., 2024). We separate its cross-entropy loss into matched-pair and non-target contributions:

$$\begin{aligned} \mathcal{L}_{\text{KD}}^{\text{diag}} &= -\frac{1}{2N} \sum_{d \in \mathcal{D}} \sum_i p_{ii}^d \log q_{ii}^d, \\ \mathcal{L}_{\text{KD}}^{\text{off}} &= -\frac{1}{2N} \sum_{d \in \mathcal{D}} \sum_i \sum_{j \neq i} p_{ij}^d \log q_{ij}^d, \end{aligned} \tag{4}$$

with $\mathcal{L}_{\text{KD}} = \mathcal{L}_{\text{KD}}^{\text{diag}} + \mathcal{L}_{\text{KD}}^{\text{off}}$. Both the CLIP loss and diagonal distillation reinforce matched pairs. CLIP weights each pair equally, whereas diagonal distillation weights it by the teacher's confidence p_{ii}^d. The off-diagonal term transfers the teacher's similarities among non-matched pairs. Such pairs can be semantically related, so non-target supervision is not necessarily erroneous. Both terms retain probabilities normalized over all candidates in the batch. This differs from DKD's binary target-versus-rest and renormalized non-target KL terms (Zhao et al., 2022); Appendix B.2 gives the algebraic relationship.

3.3 DARK Training Objective

Our hypothesis is that early imitation gives a compact student useful image–text relationships, while later departure lets it adapt those relationships to its limited capacity. Matched pairs provide a consistent alignment objective during this transition. DARK therefore keeps their supervision positive and changes only the weight $\beta(t)$ on non-target distillation:

$$\mathcal{L}(t) = \mathcal{L}_{\text{CLIP}} + \mathcal{L}_{\text{KD}}^{\text{diag}} + \beta(t)\mathcal{L}_{\text{KD}}^{\text{off}}. \tag{5}$$

At $\beta = 1$, the objective recovers standard full-distribution KD alongside the CLIP loss. At $\beta = 0$, it removes non-target distillation while retaining both sources of matched-pair supervision. A negative β reverses the non-target distillation gradient; we call this phase *repulsion*. The teacher still guides this phase: its probabilities weight the non-target relationships that the student is encouraged to move away from. The matched-pair and CLIP terms retain unit weight throughout training. These terms act together, so repulsion does not require every non-target similarity to decrease.

3.4 Annealed Attraction-to-Repulsion Schedule

To give the student time to learn from the teacher before repulsion begins, we decrease β gradually from a positive starting value β_{start} to a negative final value β_{stop}. Let E be the total number of training epochs and $t \in [0, E]$ the elapsed training time in epoch units. We update the coefficient at each optimization step using a linear schedule:

$$\beta(t) = \beta_{\text{start}} + \frac{t}{E}(\beta_{\text{stop}} - \beta_{\text{start}}). \tag{6}$$

The default schedule decreases from 2 to -0.8 over 20 epochs. We train the student encoders and similarity scale while keeping the teacher frozen. Constraining the student scale to $1 \leq a_{\text{S}} \leq 100$ bounds the objective below even during repulsion (Appendix B.3). Appendix A gives the remaining implementation settings.

4 Experiments and Results

4.1 Experimental Setup

Models and training. We distill FetalCLIP (Maani et al., 2026) into MobileFetalCLIP, a student with an 11.4M-parameter FastViT visual encoder and a 12-layer Transformer text encoder (Vasu et al., 2023; 2024). The teacher remains frozen, while both student encoders are trained. Training uses hospital and textbook image–caption pairs; Appendix A details the 227,521 source pairs, sampling, and optimization settings.

Comparisons. Our principal comparison tests DARK against static knowledge distillation with the same student architecture, corpus, and optimization settings. Both teacher and student use 224×224 images. We train for 20 epochs with AdamW (Loshchilov & Hutter, 2019) and report the mean across four runs, evaluating each model at the end of training. Static KD uses the image–text cross-entropy objective studied in CLIP-KD (Yang et al., 2024). We also compare with training without distillation, test DARK's two design choices separately, and examine alternative objectives, schedules, and student sizes. Table S3 lists the teacher and published VLMs for reference.

Tasks and metrics. We evaluate zero-shot recognition by comparing images with text descriptions of candidate classes. On Planes DB (Burgos-Artizzu et al., 2020), we evaluate five anatomical planes using 8,187 images and three brain sub-planes using 2,949 images. Macro-F1 gives equal weight to each scored class. These evaluations use the full labeled pools, with no additional train/test split.

HC18 (van den Heuvel et al., 2018) tests whether predicted age is consistent with measured head circumference. We compare each head image with age-specific text prompts and predict the median of the 15 highest-scoring candidate ages. A prediction is counted as valid when the circumference lies within the WHO 2.5th–97.5th percentile range for that age (Kiserud et al., 2017). The resulting proportion, called the HC18 validity rate, measures circumference–age consistency rather than age-estimation accuracy, because ground-truth ages are unavailable. We evaluate 814 images after circumference filtering; Appendix C gives the prompts and evaluation procedure.

All task scores range from zero to one; higher is better. The Composite Score averages HC18 validity and Combined F1, the class-count-weighted average of the two classification scores (Appendix C.3).

Reporting. We set the default schedule once, by Composite Score on the evaluation benchmarks, and keep it fixed across seeds, input sizes and the external cohort, which played no part in any selection. Our main claims rest on final-epoch models; ablations marked as Composite-selected compare objectives under the same checkpoint rule (Appendix E).

4.2 Main Results

DARK improves on static distillation. DARK improves brain sub-plane recognition and HC18 validity, while five-plane recognition changes little (Table 1). The brain-F1 gain is 0.074, with a paired patient-bootstrap 95% interval of $[0.062, 0.086]$. For five-plane F1, the difference is 0.001 with an interval of $[-0.005, 0.007]$. These intervals measure patient-sampling uncertainty for the trained models; the standard deviations in the table describe variation across training runs. All four

runs favor DARK on brain F1, HC18 validity, and both aggregate scores. The gains raise mean Composite Score from 0.811 to 0.881.

Table 1: **DARK improves brain recognition and HC18 validity over static KD.** Both methods are trained and evaluated at 224×224. Scores are measured at the end of training and show mean $\pm$ sample SD across four runs. Five-plane F1 changes little. p-values are from unadjusted paired t-tests.

Metric	Static KD	**DARK**	Δ	p
Five-plane F1	0.943 ± 0.008	0.944 ± 0.006	$+0.001$	–
Brain F1	0.701 ± 0.029	0.775 ± 0.014	$+0.074$	0.0176
Combined F1	0.852 ± 0.009	0.880 ± 0.007	$+0.028$	0.0124
HC18 validity	0.770 ± 0.020	0.881 ± 0.008	$+0.111$	0.0018
Composite	0.811 ± 0.009	0.881 ± 0.006	$+0.070$	0.0005

All scores are on $[0, 1]$; Δ is DARK minus Static KD. A dash marks a metric without a paired test; Section 4.2 gives its patient-bootstrap interval. Individual runs: Table S14.

Comparison with the teacher. The compact student also exceeds FetalCLIP on brain recognition and circumference–age consistency at the same input resolution (Table S11). Five-plane F1 remains lower than the teacher's. Together with the controlled static-KD comparison, these results show that changing only the distillation objective lets a $26\times$ smaller student exceed its teacher on two of the three zero-shot tasks.

4.3 Ablation Studies

The ablations test the roles of matched-pair supervision and non-target repulsion. These component studies use one training run with 256×256 student inputs and 224×224 teacher inputs. Table 2 separates results selected by Composite Score from those measured at the end of training.

Retaining matched-pair supervision. Coupled repulsion applies the attraction-to-repulsion schedule to the entire distillation loss, including its matched-pair term. DARK applies the same schedule only to non-target supervision. Keeping the matched-pair term positive improves all three task scores (Table 2(a)), supporting continued image–caption supervision during repulsion.

Repelling rather than stopping imitation. To test whether simply removing non-target supervision is sufficient, we use a control that follows DARK's positive schedule and then holds its coefficient at zero. Matched-pair supervision stays positive throughout; we call this control *anchored zero-hold*. DARK and the control are trained independently with identical settings. At the end of training, DARK improves all three task scores and raises Composite from 0.759 to 0.884 (Table 2(b), Figure 2). Selecting checkpoints by Composite also favors DARK, although the selected zero-hold model has higher five-plane F1. Continuing into repulsion therefore improves on simply withdrawing non-target imitation.

Alternative objectives. These comparisons test whether reducing distillation, matching teacher features, or adding teacher-independent confidence regularization can recover the same gains. DARK also outperforms full-KD decay, static KD with feature matching, and a confidence penalty in the component study (Table S4). Full-KD decay reduces both matched and non-target supervision and lowers HC18 validity relative to static KD. Adding feature matching to static KD lowers HC18 validity and brain F1; the confidence penalty (Pereyra et al., 2017) has a similar Composite Score to training without distillation. A four-run comparison with the same FastViT architecture at 256×256 raises mean Composite from 0.739 without KD to 0.859 with DARK (Table S10).

Repulsion versus fixed decoupled weights. We compare seven non-target schedules with the same decoupled loss (Table S12). A constant positive weight ($\beta = 2$) corresponds to DKD-style decoupling with fixed weights; it reaches Composite 0.803, whereas DARK's attraction-to-repulsion schedule reaches 0.869. Constant repulsion fails (0.412), so the attractive phase is also needed. A half-cosine schedule and a late step give similar Composite Scores to the linear schedule, so its exact shape is not essential to the aggregate gain. An earlier step gives a lower Composite Score despite sharing the linear schedule's endpoints and mean coefficient. Timing therefore matters, although

Table 2: **Testing the matched-pair anchor and non-target repulsion.** DARK outperforms coupled repulsion and, at the end of training, anchored zero-hold on all three tasks. Each student setting uses one run with 256×256 student inputs and 224×224 teacher inputs. (a) Checkpoints selected by Composite Score, compared under the same rule; (b) completed training schedules, with no checkpoint selection. Bold marks the best student score in each panel; the teacher is a separate reference. All scores are on $[0, 1]$.

(a) Composite-selected checkpoints

Model / objective	Five-plane F1 ↑	Brain F1 ↑	HC18 validity ↑	Composite ↑
CLIP only	0.889	0.712	0.713	0.768
Static KD	0.946	0.715	0.794	0.826
Coupled repulsive KD	0.933	0.763	0.844	0.857
Anchored zero-hold	**0.956**	0.617	0.784	0.806
DARK	0.946	**0.784**	**0.886**	**0.886**
FetalCLIP teacher	0.973	0.702	0.835	0.853

(b) Final checkpoints

Model / objective	Five-plane F1 ↑	Brain F1 ↑	HC18 validity ↑	Composite ↑
Anchored zero-hold	0.844	0.740	0.713	0.759
DARK	**0.949**	**0.786**	**0.880**	**0.884**

Zero-hold has the highest selected five-plane F1; that checkpoint precedes its hold phase. Coupled repulsion and DARK share the $2 \rightarrow -0.8$ non-target schedule. Run details: Appendix E; reference models: Table S3.

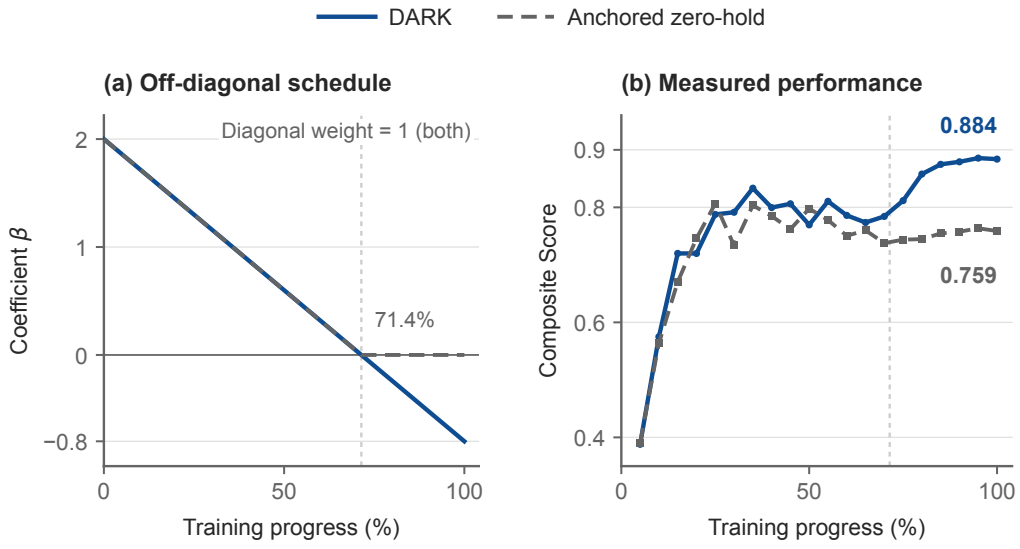

Figure 2: **Repulsion versus stopping non-target imitation.** (a) Both schedules retain matched-pair supervision and share the positive phase. DARK then enters repulsion, while anchored zero-hold remains at zero. (b) DARK achieves the higher final Composite Score. The curves show two independently trained models with identical settings, not a shared training trajectory. The dashed line marks the schedule's zero crossing; endpoint labels give final scores. Table 2(b) reports task scores and Appendix E.4 gives the protocol.

timing and cumulative weighting vary together. The endpoint study likewise finds that several, but not all, attraction-to-repulsion settings improve on static KD (Table S4).

How the benefit varies with student size. We compare MobileCLIP2-S0, MobileCLIP2-S2, and ViT-B students at a shared 224×224 resolution. DARK improves Composite only for the smallest, S0 (Appendix E.5). The larger students gain slightly on brain F1 but lose on five-plane recognition and HC18 validity. DARK's benefit is therefore concentrated in the extreme-compression regime that motivates this work; these single-run results do not fix an exact capacity threshold.

4.4 External Cohort Evaluation

We evaluate generalization on an external fetal-ultrasound cohort from Algeria, Egypt, Ghana, Malawi, and Uganda (Sendra-Balcells et al., 2023). The cohort was excluded from DARK's training, prompt adjustment, and model selection. After removing exact duplicate images from the 450-image release, the primary analysis includes 403 images spanning abdomen, brain, femur, and thorax views. We evaluate the primary models without adaptation, retaining the same preprocessing and nine candidate labels (Appendix F).

DARK raises mean macro-F1 from 0.770 with static KD to 0.895 (Figure 3). The paired gain is 0.125, with a 95% bootstrap interval of $[0.102, 0.150]$. Mean F1 improves in every country and all four anatomical classes; abdomen shows the largest gain. On this unseen cohort, the $26\times$ smaller student performs on par with its teacher (0.895 vs. 0.894), although the teacher retains higher accuracy and leads in two countries (Tables S15–S16).

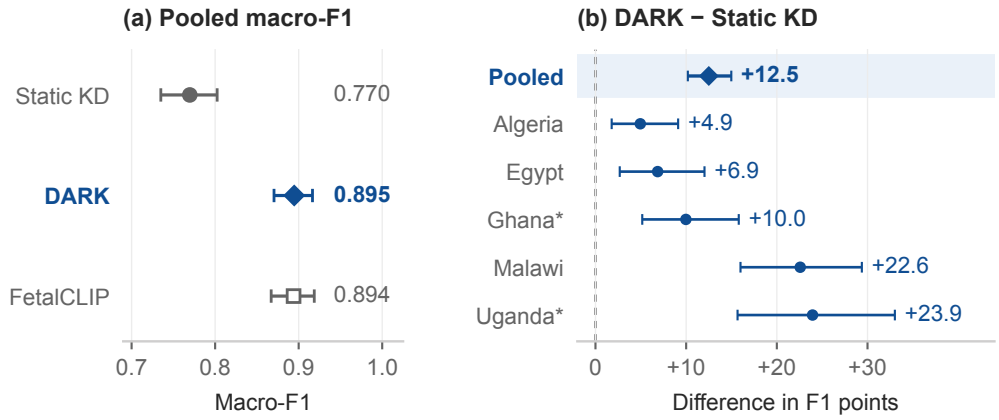

Figure 3: **DARK improves recognition on an external five-country cohort.** (a) Macro-F1 on 403 unique images. (b) DARK's gain over static KD, shown in F1 points (100 times the score difference). Student results average scores from four separately evaluated models. Whiskers show 95% paired bootstrap intervals from resampling linked record groups within countries. Asterisks mark countries with three annotated classes; all other countries and the pooled result score four. Appendix F gives the protocol and full results.

4.5 Downstream Classification and Efficiency

Transfer to task-specific classifiers. To assess the usefulness of the visual representations beyond zero-shot prediction, we freeze the image encoder and train a linear classifier for each downstream task. On six-view recognition, MobileFetalCLIP achieves F1 0.930 versus 0.947 for FetalCLIP; for congenital heart disease detection, AUROC is 0.769 versus 0.787. Across the three tasks, the student retains 97–98% of the teacher's linear-probe performance (Table S8). Appendix D.5 describes the patient-level splits and classifier training.

Inference on mobile devices. MobileFetalCLIP reduces the visual encoder from 304M to 11.4M parameters and its computation from 38.9 to 1.2 billion multiply–accumulate operations (Table 3).

On an iPhone 16 Pro, encoding one image takes 1.6 ms, compared with 37.6 ms for FetalCLIP. The iPhone 17 Pro shows a similar advantage; these timings cover visual encoding rather than a complete clinical pipeline.

Table 3: **MobileFetalCLIP reduces encoder size, computation, and latency.** Measurements use CoreML with FP16 precision and one image per batch. All counts and timings refer to the visual encoder. GMACs denotes billions of multiply–accumulate operations.

Model	Visual params.	GMACs	iPhone 16 Pro latency (ms)	iPhone 17 Pro latency (ms)
FetalCLIP	304M	38.9	37.6	31.9
MobileFetalCLIP	11.4M	1.2	1.6	1.4

4.6 Representation Analysis

Class separation and feature use. On the full five-plane set, DARK improves class separation over static KD, raising the silhouette score from 0.375 to 0.525. It also increases participation-ratio dimensionality, indicating broader use of feature dimensions. Lower inter-class cosine similarity and the uniformity measure of Wang & Isola (2020) support this pattern (Table S7).

Agreement with the teacher. On 500 images, DARK's class-probability distributions are closer to the teacher's, while their image–image similarity rankings agree less closely (Figure S3). We call this pattern *structured decorrelation*: similarity relationships change while probability agreement improves. Appendix D.2 defines the measures and reports their values.

5 Discussion and Limitations

DARK shows that a compact student can use teacher relationships without reproducing all of them throughout training. The ablations show that both parts of this design matter. Keeping matched-pair supervision positive outperforms reversing the full distillation loss, and reversing non-target supervision outperforms both stopping it at zero and holding it at a fixed positive weight, as DKD-style decoupling does. The student therefore benefits from learning the teacher's relationships early and then actively moving away from them while keeping image–text alignment.

Closer teacher–student class-probability agreement despite altered similarity rankings is consistent with this interpretation. The student thus departs from the teacher's geometry while retaining task-relevant agreement. These measurements characterize the learned representations; they do not identify which individual relationships cause the gains. Gains that hold across all four seeds and transfer to an unseen five-country cohort support the practical value of this strategy for point-of-care deployment.

Scope and limitations. We evaluate DARK with FetalCLIP in fetal ultrasound, motivated by the computational constraints of point-of-care imaging. The external African cohort demonstrates improved recognition across countries within this application. Extending the evaluation to other teachers and imaging domains is a direction for future work.

6 Conclusion

DARK separates two roles of teacher supervision: reinforcing matched image–text pairs and shaping relationships among non-target pairs. Keeping the first role fixed while changing the second from attraction to repulsion improves compact fetal-ultrasound distillation. Repeated comparisons and external evaluation establish gains over static imitation, and the ablations show that both the matched-pair anchor and the sign change are needed; fixed decoupled weights are not enough. The resulting MobileFetalCLIP encodes an image in 1.6 ms on a phone, exceeds its $26\times$ larger teacher on two zero-shot tasks, and retains 97–98% of its linear-probe performance. These findings support selective, scheduled teacher supervision for compressing fetal-ultrasound VLMs.

AI USE STATEMENT

GPT-4o was used to generate caption templates for the training data (Appendix A.3). AI tools were also used for language polishing and coding assistance. The authors retain responsibility for the manuscript, code, and reported results.

ETHICS STATEMENT

This study uses retrospective fetal ultrasound data and evaluates compact models for potential assistive applications. Use of the anonymized hospital data was approved by the regional health authority. Hospital image–caption data are not publicly released, and textbook material is subject to copyright restrictions. The reported benchmarks do not establish safety or clinical effectiveness, and deployment requires prospective assessment across patient populations, imaging devices, and operators.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix B specify the objective and schedule. Appendices A and C detail training, evaluation, and seeds; Appendix D contains additional analyses. Appendix F documents the public external cohort, frozen evaluation, duplicate handling, and uncertainty estimation. Anonymized training and evaluation code is provided in the supplementary material; we will release it with the MobileFetalCLIP weights upon publication. Restricted access to the training corpus limits independent retraining.

REFERENCES

Christian F. Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P. Fletcher, Sandra Smith, Lisa M. Koch, Bernhard Kainz, and Daniel Rueckert. SonoNet: Real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE Transactions on Medical Imaging*, 36(11):2204–2215, 2017. doi: 10.1109/tmi.2017.2712367.

Thomas Breuel. WebDataset, n.d. URL `https://github.com/webdataset/webdataset`. Software repository. Accessed 25 September 2026.

Xavier P. Burgos-Artizzu, David Coronado-Gutiérrez, Brenda Valenzuela-Alcaraz, Elisenda Bonet-Carne, Elisenda Eixarch, Fatima Crispi, and Eduard Gratacós. Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Scientific Reports*, 10(1):10200, 2020. doi: 10.1038/s41598-020-67076-5.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4794–4802, 2019. URL `https://openaccess.thecvf.com/content_ICCV_2019/html/Cho_On_the_Efficacy_of_Knowledge_Distillation_ICCV_2019_paper.html`.

Fartash Faghri, Pavan Kumar Anasosalu Vasu, Cem Koc, Vaishaal Shankar, Alexander T. Toshev, Oncel Tuzel, and Hadi Pouransari. MobileCLIP2: Improving multi-modal reinforced training. *Transactions on Machine Learning Research*, 2025. URL `https://openreview.net/forum?id=WeF9zolng8`.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1607–1616. PMLR, 2018. URL `https://proceedings.mlr.press/v80/furlanello18a.html`.

Ryan G. Gomes, Bellington Vwalika, Chace Lee, Angelica Willis, Marcin Sieniek, Joan T. Price, Christina Chen, Margaret P. Kasaro, James A. Taylor, Elizabeth M. Stringer, Scott Mayer McKinney, Ntazana Sindano, George E. Dahl, William Goodnight, III, Justin Gilmer, Benjamin H. Chi, Charles Lau, Terry Spitz, T. Saensuksopa, Kris Liu, Tiya Tiyasirichokchai, Jonny Wong, Rory Pilgrim, Akib Uddin, Greg Corrado, Lily Peng, Katherine Chou, Daniel Tse, Jeffrey S. A. Stringer, and Shravya Shetty. A mobile-optimized artificial intelligence system for gestational

age and fetal malpresentation assessment. *Communications Medicine*, 2(1):128, 2022. doi: 10.1038/s43856-022-00194-5.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable, 2022. URL https://github.com/huggingface/accelerate.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33716–33727. Curran Associates, Inc., 2022. doi: 10.52202/068431-2443.

Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. UniMed-CLIP: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. *arXiv preprint arXiv:2412.10372*, 2024. URL https://arxiv.org/abs/2412.10372.

Torvid Kiserud, Gilda Piaggio, Guillermo Carroli, Mariana Widmer, José Carvalho, Lisa Neerup Jensen, Daniel Giordano, José Guilherme Cecatti, Hany Abdel Aleem, Sameera A. Talegawkar, Alexandra Benachi, Anke Diemert, Antoinette Tshefu Kitoto, Jadsada Thinkhamrop, Pisake Lumbiganon, Ann Tabor, Alka Kriplani, Rogelio Gonzalez Perez, Kurt Hecher, Mark A. Hanson, A. Metin Gülmezoglu, and Lawrence D. Platt. The World Health Organization fetal growth charts: A multinational longitudinal study of ultrasound biometric measurements and estimated fetal weight. *PLOS Medicine*, 14(1):e1002220, 2017. doi: 10.1371/journal.pmed.1002220.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Fadillah Maani, Numan Saeed, Tausifa Jan Saleem, Zaid Farooq, Hussain Alasmawi, Werner Diehl, Ameera Mohammad, Gareth Waring, Saudabi Valappil, Leanne Bricker, and Mohammad Yaqub. FetalCLIP: a visual-language foundation model for fetal ultrasound image analysis. *npj Digital Medicine*, 2026. doi: 10.1038/s41746-026-02907-9.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5191–5198, 2020. doi: 10.1609/aaai.v34i04.5963.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *International Conference on Learning Representations Workshop*, 2017. URL https://openreview.net/forum?id=HyhbYrGYe.

Teeranan Pokaprakarn, Juan C. Prieto, Joan T. Price, Margaret P. Kasaro, Ntazana Sindano, Hina R. Shah, Marc Peterson, Mutinta M. Akapelwa, Filson M. Kapilya, Yuri V. Sebastião, William Goodnight, Elizabeth M. Stringer, Bethany L. Freeman, Lina M. Montoya, Benjamin H. Chi, Dwight J. Rouse, Stephen R. Cole, Bellington Vwalika, Michael R. Kosorok, and Jeffrey S. A. Stringer. AI estimation of gestational age from blind ultrasound sweeps in low-resource settings. *NEJM Evidence*, 1(5), 2022. doi: 10.1056/evidoa2100058.

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

Machine Learning Research, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6550.

Carla Sendra-Balcells, Víctor M. Campello, Jordina Torrents-Barrena, Yahya Ali Ahmed, Mustafa Elattar, Benard Ohene-Botwe, Pempho Nyangulu, William Stones, Mohammed Ammar, Lamya Nawal Benamer, Harriet Nalubega Kisembo, Senai Goitom Sereke, Sikolia Z. Wanyonyi, Marleen Temmerman, Eduard Gratacós, Elisenda Bonet, Elisenda Eixarch, Kamil Mikolaj, Martin Grønnebæk Tolsgaard, and Karim Lekadir. Generalisability of fetal ultrasound deep learning models to low-resource imaging settings in five African countries. *Scientific Reports*, 13(1):2728, 2023. doi: 10.1038/s41598-023-29490-3.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? In *Advances in Neural Information Processing Systems*, volume 34, pp. 6906–6919. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/376c6b9ff3bedbbea56751a84fffc10c-Abstract.html.

Kelsey A. Stewart, Sergio M. Navarro, Sriharsha Kambala, Gail Tan, Revanth Poondla, Sara Lederman, Kelli Barbour, and Chris Lavy. Trends in ultrasound use in low and middle income countries: A systematic review. *International Journal of Maternal and Child Health and AIDS (IJMA)*, 9 (1):103–120, 2020. doi: 10.21106/ijma.294.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15731–15740, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Sun_Logit_Standardization_in_Knowledge_Distillation_CVPR_2024_paper.html.

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

Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022. doi: 10.1038/s41551-022-00936-9.

Thomas L. A. van den Heuvel, Dagmar de Bruijn, Chris L. de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2D ultrasound images. *PLOS ONE*, 13(8):e0200412, 2018. doi: 10.1371/journal.pone.0200412.

Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. FastViT: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5785–5795, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Vasu_FastViT_A_Fast_Hybrid_Vision_Transformer_Using_Structural_Reparameterization_ICCV_2023_paper.html.

Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. MobileCLIP: Fast image-text models through multi-modal reinforced training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15963–15974, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Vasu_MobileCLIP_Fast_Image-Text_Models_through_Multi-Modal_Reinforced_Training_CVPR_2024_paper.html.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9929–9939. PMLR, 2020. URL https://proceedings.mlr.press/v119/wang20k.html.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.256.

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*, pp. 21970–21980, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wu_TinyCLIP_CLIP_Distillation_via_Affinity_Mimicking_and_Weight_Inheritance_ICCV_2023_paper.html.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. CLIP-KD: An empirical study of CLIP model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15952–15962, 2024. doi: 10.1109/CVPR52733.2024.01510.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2025. doi: 10.1056/aioa2400640.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11953–11962, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhao_Decoupled_Knowledge_Distillation_CVPR_2022_paper.html.

A Implementation Details

A.1 Model Architecture

The teacher is FetalCLIP with a ViT-L/14 visual encoder, 768-dimensional output embeddings, approximately 304M visual parameters, and 427M total parameters (Maani et al., 2026). It uses a 224-pixel input size and a 117-token text context, and remains frozen throughout distillation. The student uses the FastViT-based MCi0 image encoder of MobileCLIP (Vasu et al., 2023; 2024), with 512-dimensional output and 11.4M visual parameters. The primary comparison uses 224×224 student inputs; the component studies use 256×256 inputs, as detailed in Appendix E. Its CLIP-style text Transformer has 12 layers, eight attention heads, a hidden width of 512, a 49,408-token vocabulary, and 512-dimensional output embeddings. The text weights are initialized from MobileCLIP2-S0 (Faghri et al., 2025), with positional embeddings linearly interpolated from 77 to 117 tokens. The total student parameter count is approximately 75M. The student encoders and native logit scale are trainable. We parameterize the scale as $a_{\mathrm{S}} = \exp(s_{\mathrm{S}})$ and clamp s_{S} to $[0, \log 100]$ after each optimizer update. The log-scale parameter is excluded from weight decay; the explicit clamp enforces its range independently of gradient clipping.

A.2 Training Configuration

Table S1 records the base configuration. The component-study runs of DARK, Static KD and CLIP-only training use one GPU with 512 image–text pairs per forward pass. Gradients accumulate over

two forward passes, giving an effective optimization batch of 1,024 pairs. Each contrastive or distillation loss uses a 512×512 similarity matrix; embeddings are not combined across accumulation steps. The learning rate warms up over 500 optimizer updates before cosine decay.

Objective-specific settings appear in Table S4: static KD keeps its coefficient at one; full-KD annealing varies the coefficient on both components; DARK fixes the diagonal coefficient and schedules only the off-diagonal term. The feature-KD ablation follows CLIP-KD (Yang et al., 2024): learned linear heads project the student's 512-dimensional image and text embeddings to the teacher's 768 dimensions, and the mean-squared error between the normalized projections and the teacher embeddings is added with weight 2,000; the confidence-penalty ablation uses coefficient 0.1 (Pereyra et al., 2017). The component studies use seed 42, and the default DARK configuration is repeated with seeds 7 and 123. The multi-seed comparisons use seeds 1, 7, 42 and 123; Appendix E gives their resolutions and checkpoint rules.

Table S1: **Base training configuration.** Batch, accumulation, and warmup entries describe the component-study runs of DARK, Static KD and CLIP-only training. Appendix E describes the other comparisons.

Setting	Value
Optimizer	AdamW; $(\beta_1, \beta_2) = (0.9, 0.98)$, $\epsilon = 10^{-6}$
Peak learning rate	10^{-5}
Weight decay	0.1
Learning-rate schedule	Cosine decay after linear warmup
Warmup	500 optimizer updates
Pairs per forward pass	512 on one GPU
Gradient accumulation	Two forward passes per optimizer update
Effective optimization batch	1,024 image–text pairs
Candidates per query	$N = 512$; 512×512 similarity matrix
Training length	20 epochs
Distillation temperature	Teacher-only $\tau_{\mathrm{KD}} = 5$; no τ_{KD}^2 prefactor
Student logit scale	$a_{\mathrm{S}} \in [1, 100]$; log-space clamp after optimizer updates
DARK default	$\beta : 2 \rightarrow -0.8$; diagonal and CLIP coefficients each one
Numerical precision	FP16
Gradient clipping	Global ℓ_2 norm one
Training framework	HuggingFace Accelerate (Gugger et al., 2022)
Training GPU	NVIDIA RTX PRO 5000 Blackwell
Component-study seeds	42; DARK repeated with 7 and 123
Multi-seed comparisons	1, 7, 42, 123

A.3 Data Curation and Augmentation

The hospital source contains 225,429 fetal ultrasound images from 6,493 patients with mean gestational age 148 ± 16 days. All images include gestational-age and pixel-spacing information; 161,033 have clinician-provided anatomical labels and 64,396 have anatomical pseudo-labels. Clinician labels come from text annotations on the scans. Pseudo-labels follow the approach of Maani et al. (2026): view classifiers trained with five-fold cross-validation on clinician-labeled images assign anatomical views, including brain and heart sub-views, to images without annotations. Captions are also built as in Maani et al. (2026): GPT-4o generates caption templates for each label, and each template has placeholders for gestational age and pixel spacing, which are filled with image-specific metadata. An abdomen caption, for example, describes gestational age in weeks and days, the anatomical view, and pixel spacing in mm/pixel. These captions provide the paired text used for training.

An additional 2,092 image–caption pairs come from a fetal-ultrasound textbook. Repeating each textbook pair ten times yields $225{,}429 + 10 \times 2{,}092 = 246{,}349$ training entries from 227,521 source pairs. The dataset is served in 25 WebDataset tar shards (Breuel, n.d.). The hospital data are not publicly released because of privacy and access restrictions; the textbook material is also subject to copyright restrictions.

Training augmentation applies a random affine transformation with rotation in $[-7°, 7°]$, translation up to 0.05 along each axis, and fixed scale one, followed by brightness, contrast, and saturation jitter of 0.15. Teacher and student branches share the random transformation parameters for each image. In the component studies, the student then applies `Resize(256)`, `CenterCrop(256)`, and ImageNet normalization; the teacher applies `Resize(224)` without a center crop and uses OpenAI CLIP normalization. The views therefore share augmented content while retaining branch-specific preprocessing. The primary matched-resolution comparison instead uses 224×224 student input.

B Mathematical Details

B.1 Full-Distribution Distillation

For query i and direction d, the teacher and student distributions are defined in Equation 2. The symmetric loss is

$$\mathcal{L}_{\mathrm{KD}} = \frac{1}{2N} \sum_{d \in \mathcal{D}} \sum_{i} \mathcal{H}(p_i^d, q_i^d), \qquad \mathcal{H}(p, q) = -\sum_j p_j \log q_j. \tag{7}$$

Cross-entropy satisfies

$$\mathcal{H}(p, q) = \mathrm{KL}(p \| q) + \mathcal{H}(p) \tag{8}$$

Here $\mathcal{H}(p)$ is fixed teacher entropy, so both objectives have the same student gradients. We use natural logarithms. Teacher-only softening, the student's learned native logit scale, and omission of a temperature-squared prefactor are conventions of our implementation.

For a paired index i, the entrywise identity is

$$\mathcal{H}(p_i^d, q_i^d) = -p_{ii}^d \log q_{ii}^d - \sum_{j \neq i} p_{ij}^d \log q_{ij}^d. \tag{9}$$

Averaging yields

$$\mathcal{L}_{\mathrm{KD}} = \mathcal{L}_{\mathrm{KD}}^{\mathrm{diag}} + \mathcal{L}_{\mathrm{KD}}^{\mathrm{off}}. \tag{10}$$

The diagonal term encourages increased student probability on the matched pair, weighted by teacher confidence; it does not independently match the student diagonal probability to the teacher's value. At a zero off-diagonal coefficient, the objective remains $\mathcal{L}_{\mathrm{CLIP}} + \mathcal{L}_{\mathrm{KD}}^{\mathrm{diag}}$. Solving $\beta(t_0) = 0$ gives

$$t_0 = \frac{\beta_{\mathrm{start}}}{\beta_{\mathrm{start}} - \beta_{\mathrm{stop}}} E. \tag{11}$$

The default therefore crosses at $100/7 \approx 14.3$ epochs, leaving $2/7 \approx 28.6\%$ of the run for repulsion.

B.2 Relationship to Decoupled Knowledge Distillation

DARK and DKD (Zhao et al., 2022) both distinguish target and non-target supervision, but their decompositions differ algebraically. Fix query i and direction d. Write $p_+ = p_{ii}^d$ and $q_+ = q_{ii}^d$ for the matched probabilities, with remaining masses $P_- = 1 - p_+$ and $Q_- = 1 - q_+$. Define conditional non-target distributions $\widetilde{p}_j = p_{ij}^d / P_-$ and $\widetilde{q}_j = q_{ij}^d / Q_-$ for $j \neq i$. Then DARK's off-diagonal cross-entropy is

$$\begin{aligned} -\sum_{j \neq i} p_{ij}^d \log q_{ij}^d &= -P_- \log Q_- + P_- \mathcal{H}(\widetilde{p}, \widetilde{q}) \\ &= -P_- \log Q_- + P_- \mathrm{KL}(\widetilde{p} \| \widetilde{q}) + P_- \mathcal{H}(\widetilde{p}). \end{aligned} \tag{12}$$

It therefore retains the teacher's non-target mass and the student's target-versus-rest allocation. By contrast, DKD independently weights a binary target-versus-rest KL term and a renormalized non-target KL term. DARK's matched cross-entropy alone is not DKD's binary KL, and its off-diagonal term is not an independently normalized non-target KL. Table S2 summarizes the distinction.

Section 4.3 tests the practical consequence of this difference: in DARK's partition, a fixed positive non-target weight ($\beta = 2$), the DKD-style choice, reaches Composite 0.803, whereas the sign-changing schedule reaches 0.869 (Table S12). Negative β reverses the gradient of the off-diagonal

Table S2: **DARK and DKD use different loss partitions.** The sign-changing schedule is DARK's additional design choice; the terms are not interchangeable algebraic definitions.

Aspect	DKD	DARK
Candidate set	Classification labels	Paired image–text batch candidates
Target term	Binary target-versus-rest KL	Matched-entry cross-entropy
Non-target term	KL on conditional non-target distributions	Off-diagonal cross-entropy under full normalization
Weighting	Independently weighted attractive terms	Fixed positive diagonal; scheduled off-diagonal coefficient crossing zero
Temperature	Teacher and student softening in the standard formulation	Teacher-only softening in our implementation

component. This does not establish monotonic pairwise separation under the combined objective or guarantee convergence for every schedule. The empirical schedule comparisons assess the resulting optimization behavior in our setting.

B.3 Bounded Objective under the Scale Constraint

The implementation trains $s_{\mathrm{S}} = \log a_{\mathrm{S}}$ and projects it onto $[0, \log A]$ after each optimizer update, with $A = 100$. Unit-normalized embeddings and $a_{\mathrm{S}} \le A$ imply $-A \le Z_{ij}^{\mathrm{S},d} \le A$. For each of the N candidates in a query distribution,

$$q_{ij}^{d} \ge \frac{e^{-A}}{Ne^{A}} = \frac{e^{-2A}}{N}, \qquad 0 \le -\log q_{ij}^{d} \le \log N + 2A. \tag{13}$$

Since $\sum_{j \ne i} p_{ij}^{d} \le 1$, averaging over queries and directions gives $0 \le \mathcal{L}_{\mathrm{KD}}^{\mathrm{off}} \le \log N + 2A$. The CLIP and diagonal losses are nonnegative, so

$$\mathcal{L}(t) \ge \min\{0, \beta(t)\}\big(\log N + 2A\big). \tag{14}$$

For the default schedule, this yields the uniform lower bound $\mathcal{L}(t) \ge -0.8(\log N + 200)$. The scale constraint therefore prevents the negative off-diagonal term from driving the objective to negative infinity through unrestricted scale growth. This bound does not guarantee convergence of the nonconvex training dynamics or successful optimization for every schedule.

C Evaluation Protocols and Seed Robustness

C.1 Zero-Shot Classification

Planes DB contains 12,400 images from 1,792 patients across two hospitals (Burgos-Artizzu et al., 2020). The five-plane labeled pool contains 8,187 abdomen, brain, femur, thorax, and cervix images after excluding "Other" images. The brain sub-plane pool contains 2,949 transthalamic, transcerebellum, and transventricular images. These zero-shot evaluations use the full labeled pools, without a further train/test partition. Images are zero-padded on the shorter axis to form a square before resizing.

Classification compares image embeddings with ensembles of class-prompt embeddings, following the FetalCLIP evaluation setup (Maani et al., 2026). Text embeddings are averaged within each prompt set, and the class with the highest image similarity is selected. Five-plane evaluation uses nine candidate labels: abdomen, brain, femur, heart, kidney, lips/nose, profile, spine, and cervix. The dataset's thorax label is mapped to heart. Macro-F1 is computed over the five ground-truth classes, while predictions to the other four candidates remain errors and are grouped in a separate confusion-matrix column (Figure S1).

C.2 HC18 Validity

HC18's public annotated partition contains 999 head-circumference images (van den Heuvel et al., 2018). Filtering to HC values in $[100, 342]$ mm retains 814 images, following FetalCLIP (Maani

et al., 2026). We estimate gestational age using 175 candidate ages at daily intervals, spanning 14 weeks 0 days to 38 weeks 6 days. For each candidate age, we average the cosine similarities between the image embedding and five age-specific text-prompt embeddings. The predicted age is the median of the 15 candidate ages with the highest averaged similarities. Since true GA is unavailable in the released evaluation labels, validity uses observed HC as a proxy.

For a predicted age $\widehat{g}_i$ in days, let $B_{2.5}(\widehat{g}_i)$ and $B_{97.5}(\widehat{g}_i)$ denote the WHO chart's HC bounds (Kiserud et al., 2017). Then

$$\mathrm{VR}_{\mathrm{HC18}} = \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}[B_{2.5}(\widehat{g}_i) \le \mathrm{HC}_i \le B_{97.5}(\widehat{g}_i)], \quad M = 814. \tag{15}$$

The indicator $\mathbb{I}$ equals one when its condition holds and zero otherwise. Let g denote gestational age in days and $w = g/7$ the corresponding age in weeks. The implementation computes the head-circumference bound for percentile r as

$$B_r(g) = \exp\left(b_{r,0} + b_{r,1}w + b_{r,2}w^2 + b_{r,3}w^3 + b_{r,4}w^4\right), \tag{16}$$

where r is 2.5 or 97.5 and the coefficients are percentile-specific. This metric quantifies whether the predicted age is consistent with the measured circumference interval; it does not directly measure GA error against an observed age. HC18 prompts encode pixel spacing scaled to the input size. We therefore compare classification, but not HC18 validity, across input sizes, and we compare students with the teacher on HC18 only within 224×224, using resolution-specific handling (Appendix E.2).

C.3 Classification Metrics and Aggregation

For each scored class k, precision, recall, and F1 are

$$\mathrm{Prec}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}, \quad \mathrm{Rec}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}, \quad \mathrm{F1}_k = \frac{2\mathrm{Prec}_k\mathrm{Rec}_k}{\mathrm{Prec}_k + \mathrm{Rec}_k}. \tag{17}$$

Here TP, FP, and FN denote true positives, false positives, and false negatives; undefined ratios are set to zero. Task macro-F1 averages class F1 values uniformly over the five planes or three brain sub-planes. The combined classification summary is

$$\mathrm{F1}_{\mathrm{Pl_Br}} = \frac{5\mathrm{F1}_{\mathrm{5Plane}} + 3\mathrm{F1}_{\mathrm{3Brain}}}{8}. \tag{18}$$

It averages eight class-specific scores from two tasks; it is not an eight-way classifier evaluated on a single label space. The Composite Score gives equal weight to this classification summary and HC18 validity:

$$C = \frac{\mathrm{F1}_{\mathrm{Pl_Br}} + \mathrm{VR}_{\mathrm{HC18}}}{2}. \tag{19}$$

Table S3 lists the scores of all reference models, and Table S4 gives the complete ablation. Direct TinyCLIP-style weight inheritance (Wu et al., 2023) is incompatible with the FastViT–ViT-L/14 architecture pair used here.

C.4 Seed Robustness

The component studies use seed 42, 256×256 student inputs and Composite-based checkpoint selection. Table S5 repeats the default DARK configuration with seeds 123 and 7, also with Composite-based checkpoint selection. Mean Composite Score is 0.867 ± 0.016, and all three runs exceed Static KD (0.826) and the teacher (0.853). Brain F1 varies more (0.735 ± 0.069) than five-plane F1 or HC18 validity: seed 123 reaches 0.656, below the teacher's 0.702, whereas seeds 42 and 7 exceed it. These repeats show aggregate consistency for DARK, but they do not include matched baseline runs.

For seed 123, Composite-based selection chose epoch 15, where HC18 validity peaks while brain F1 is still improving; the same run reaches brain F1 0.743 at epoch 20. Composite-based selection can therefore pick a checkpoint before brain F1 stops improving. The multi-seed comparisons train DARK and the baselines with seeds $1, 7, 42, 123$ and evaluate the final epoch. With 256×256 training and evaluation, the four final-epoch DARK brain F1 scores are 0.786, 0.743, 0.789 and 0.747. With 224×224 training and evaluation, DARK's mean brain F1 is 0.775 ± 0.014 (Table S14). Table S9 compares DARK and Static KD under both checkpoint rules.

Table S3: **Component-study results and external references.** Student rows use seed 42, 256×256 inputs and Composite-selected checkpoints. Bold marks the best MobileFetalCLIP student score in each column, including ties. Student rows compare training objectives for MobileFetalCLIP; external reference results are taken from Maani et al. (2026). VLM predictions use image–text similarity, while SonoNet (Baumgartner et al., 2017) is a supervised classifier. Parameter counts cover the full model. All scores lie in $[0, 1]$, with higher values better; a dash denotes a metric not reported. The two summary scores are defined in Equations 18 and 19.

Model / objective	Params.	Five-plane F1 ↑	Brain F1 ↑	Combined F1 ↑	HC18 validity ↑	Composite ↑
MobileFetalCLIP student						
CLIP only	75M	0.889	0.712	0.823	0.713	0.768
Static logit KD	75M	**0.946**	0.715	0.859	0.794	0.826
Coupled repulsive KD	75M	0.933	0.763	0.869	0.844	0.857
DARK	75M	**0.946**	**0.784**	**0.886**	**0.886**	**0.886**
Teacher						
FetalCLIP (ViT-L/14)	427M	0.973	0.702	0.871	0.835	0.853
Other VLM references						
CLIP (ViT-L/14)	427M	0.308	0.206	0.270	0.110	0.190
BiomedCLIP (ViT-B/16)	150M	0.603	0.236	0.466	0.240	0.353
UniMed-CLIP (ViT-B/16)	150M	0.679	0.187	0.495	0.090	0.293
Supervised reference						
SonoNet-32	3.7M	0.827	0.485	0.699	–	–

Table S4: **Selected-checkpoint objective and schedule ablations.** Bold marks the best student score per column; the teacher is a separate reference. Blue shading identifies default DARK. The anchored zero-hold control uses the same Composite-based checkpoint selection; Appendix E.4 gives its protocol and Table 2(b) separates the end-of-training comparison.

Configuration	Five-plane F1 ↑	Brain F1 ↑	Combined F1 ↑	HC18 validity ↑	Composite ↑
FetalCLIP teacher	0.973	0.702	0.871	0.835	0.853
CLIP only	0.889	0.712	0.823	0.713	0.768
Static KD	0.946	0.715	0.859	0.794	0.826
Full-KD decay $1 \to 0.1$	0.902	0.713	0.831	0.746	0.788
Full-KD decay $1 \to 0$	0.842	0.742	0.805	0.731	0.768
Static KD + feature KD	0.946	0.664	0.840	0.759	0.800
Confidence penalty	0.854	0.680	0.789	0.749	0.769
Coupled repulsion $2 \to -0.8$	0.933	0.763	0.869	0.844	0.857
Anchored zero-hold	**0.956**	0.617	0.829	0.784	0.806
DARK $2 \to -0.8$	0.946	**0.784**	**0.886**	**0.886**	**0.886**
DARK $4 \to -0.8$	0.950	0.709	0.860	0.854	0.857
DARK $8 \to -0.8$	0.943	0.714	0.857	0.786	0.822
DARK $2 \to -0.4$	0.938	0.724	0.858	0.784	0.821
DARK $2 \to -0.5$	0.941	0.725	0.860	0.846	0.853

C.5 Patient-Sampling Uncertainty

We estimate sampling uncertainty for the eight fixed final-checkpoint models in the matched 224×224 comparison. Their saved predictions reproduce the final-checkpoint F1 scores to within 10^{-6}. The brain and five-plane pools contain 1,060 and 1,747 eligible patient identifiers, respectively. For each task, we draw its observed number of patients with replacement and include all eligible images from each selected patient, preserving multiplicity. The same draw is applied to all eight models; we calculate pooled-image macro-F1 separately for each model, average the four scores per method,

Table S5: **DARK repeated across three seeds.** Checkpoints are selected by Composite Score. The default schedule is $2 \rightarrow -0.8$. The final row gives the mean and standard deviation across the three seeds.

Seed	Five-plane F1 ↑	Brain F1 ↑	HC18 validity ↑	Composite ↑
42	0.946	0.784	0.886	0.886
123	0.945	0.656	0.882	0.860
7	0.933	0.764	0.845	0.857
Mean ± std.	0.941 ± 0.007	0.735 ± 0.069	0.871 ± 0.022	0.867 ± 0.016

and subtract Static KD from DARK. Predictions are not ensembled, and per-patient F1 values are not averaged.

Table S6 reports the observed differences and the 2.5th–97.5th percentiles of 10,000 paired bootstrap replicates. Scored classes remain fixed at three brain sub-planes or five anatomical planes; the five-plane classifier retains all nine candidate predictions. These marginal intervals condition on the fitted models and the provided patient identifiers, without multiplicity adjustment. They describe patient-sampling variability for the fitted models; Table 1 reports variation across training runs. We do not compute patient-bootstrap intervals for HC18 or Composite because the available HC18 metadata does not establish patient groups.

Table S6: **Patient-bootstrap uncertainty on Planes DB classification.** Scores average the same four final 224×224 models per method used in the matched-seed comparison. Δ is DARK minus Static KD; paired 95% intervals use 10,000 patient-cluster resamples with the fitted models fixed.

Task	Patients	Images	Static KD	**DARK**	Δ	95% interval
Brain sub-planes	1,060	2,949	0.7008	0.7748	$+0.0739$	$[0.0616, 0.0865]$
Five planes	1,747	8,187	0.9430	0.9439	$+0.0009$	$[-0.0050, 0.0066]$

Intervals describe patient-sampling uncertainty, not training-seed variation. They are task-specific marginal intervals. HC18 and Composite are excluded because HC18 patient grouping is unavailable.

D Additional Analyses

D.1 Confusion Matrices and Per-Class Scores

Figure S1 shows row-normalized five-plane confusion matrices. DARK raises brain recall from 95% to 99% and abdomen recall from 82% to 89%, while cervix, femur, and thorax recall fall from 96%, 98%, and 93% to 92%, 93%, and 86%, respectively. Thorax-to-brain predictions increase from 9 to 113 cases. These changes show redistribution of errors rather than uniform improvement across classes; they do not establish the clinical consequences of individual misclassifications. Figure S2 gives per-class F1 at each illustrated run's best Composite-Score epoch, rounded to two decimal places.

D.2 Image-Similarity and Probability Analysis

Figure S3 compares static KD and DARK on the first 500 five-plane samples, excluding ground-truth "Other" images. Mean student softmax entropy decreases from 0.666 to 0.044, indicating sharper predictions. Spearman rank correlation ρ between teacher and student image–image similarities decreases from 0.851 to 0.822, retaining substantial rank agreement. We use the strict upper triangles of matrices containing 100 times the cosine similarity between normalized image embeddings. Mean per-image agreement, $1 - d_{\mathrm{JS}}$, increases from 0.762 to 0.925. Here d_{JS} denotes Jensen–Shannon distance (the square root of Jensen–Shannon divergence, using natural logarithms) between teacher and student class-probability vectors. These use softmax over the nine candidates in Appendix C.1, normalized class-prompt embeddings, and fixed logit scale 100.

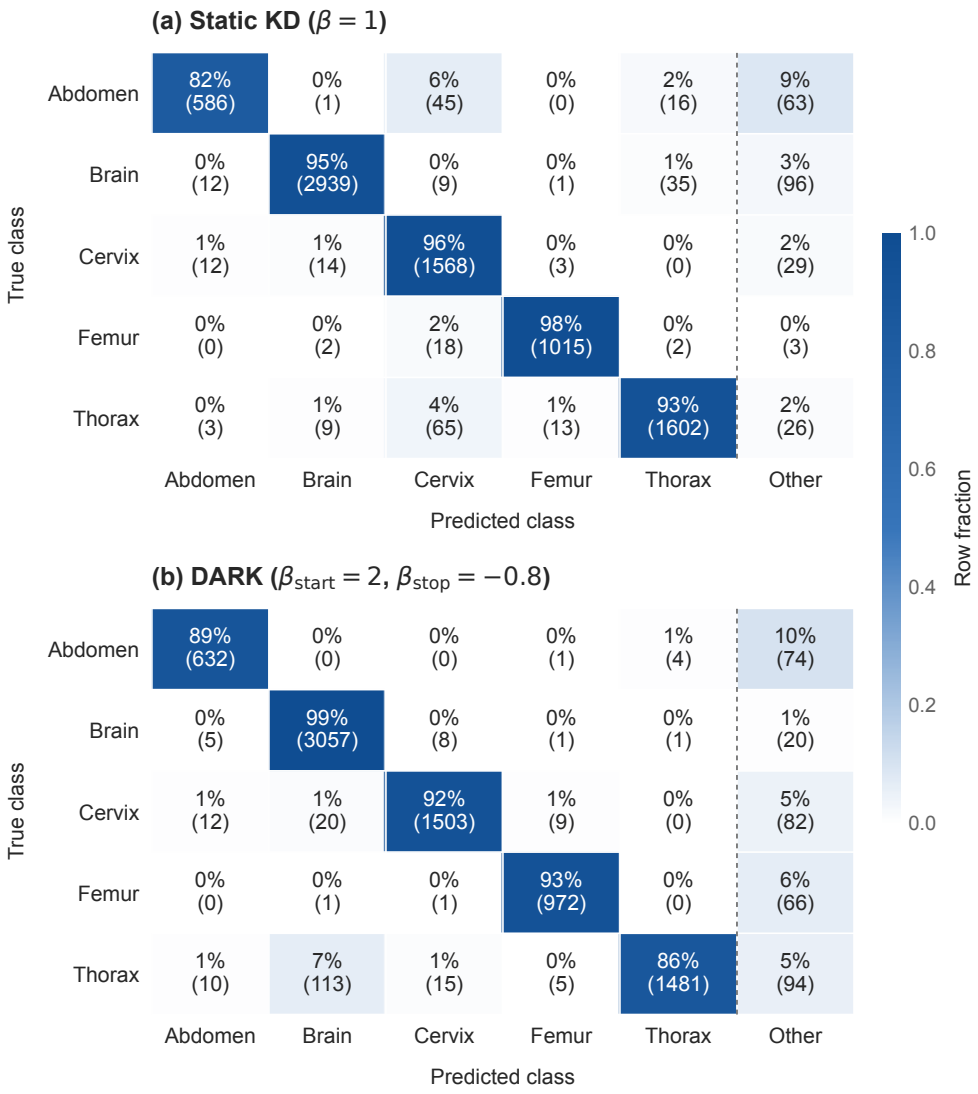

Figure S1: **Five-plane confusion matrices.** Static KD ($\beta = 1$, top) and DARK ($\beta_{\text{start}} = 2$, $\beta_{\text{stop}} = -0.8$, bottom) are shown on a shared row-fraction scale. Cells display row-normalized percentages with counts in parentheses; each panel contains 8,187 images. The "Other predictions" column includes labels outside the five scored classes; its meaning differs from the excluded ground-truth "Other" pool.

This pattern supports structured decorrelation: probability vectors become closer to the teacher while similarity rankings diverge modestly. Greater confidence is not itself evidence of correct classification or calibration, and per-image class-probability agreement differs from the matched-pair probabilities in the training loss. Diagonal supervision is a plausible contributor to retained image–text associations; these observational statistics do not prove that it causes the measured agreement change. This analysis uses the first 500 samples, whereas Appendix D.4 uses all 8,187 images.

D.3 Training Dynamics

Figure S4 shows the logged training trajectories. The panels include static KD, positive decay, coupled repulsion, confidence regularization, and DARK schedules. They show that several repulsive

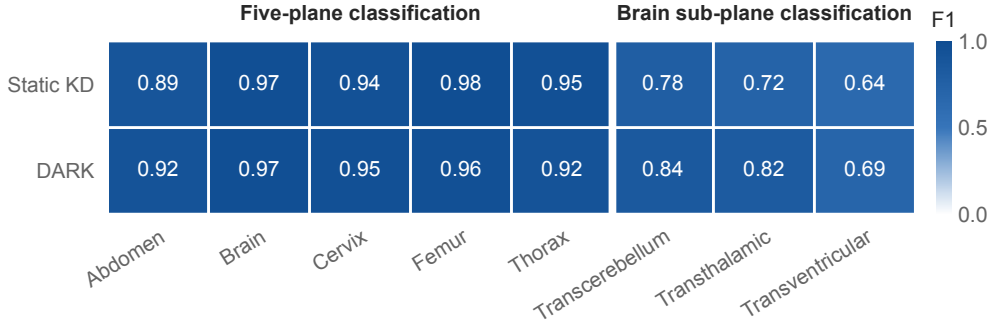

Figure S2: **Per-class F1 at the illustrated runs' best Composite-Score checkpoints.** Scores are shown to two decimal places on a shared $[0, 1]$ color scale. Columns are grouped by task. These are selected-checkpoint values, not the four-seed final-epoch results.

runs continue improving late in training, whereas static KD reaches its strongest region earlier. Each curve is one logged run, not a replicate of the corresponding table row. Figure 2 provides the focused DARK–zero-hold comparison described in Appendix E.4.

In an exploratory run, coupled repulsion with $\beta_{\text{stop}} = -1.6$ collapsed. That variant reverses both matched-pair and non-target distillation; DARK retains positive matched-pair supervision and uses $\beta_{\text{stop}} = -0.8$ by default. This failure therefore concerns a different objective and repulsion strength.

The auxiliary traces show the logged distillation-loss scalar and learned logit scale. Because it includes the signed weight, this logged value can be negative, unlike the unweighted $\mathcal{L}_{\text{KD}}$ in Equation 7. A loss component's scalar magnitude is not its gradient norm, and a rising logit scale can affect confidence independently of changes in class geometry. These traces therefore describe optimization behavior rather than isolate a single cause of generalization gains.

D.4 Cluster and Spectral Metrics

The main geometry analysis uses the $8{,}187 \times 512$ student embedding matrix from the five-plane evaluation pool. Table S7 reports effective dimensionality, rank needed for 95% variance, silhouette, intra-class and inter-class cosine similarity, and uniformity. We center the embedding matrix by subtracting its mean feature vector. For its singular values σ_k, we use spectral weights $\lambda_k = \sigma_k^2$ and compute participation-ratio dimensionality as

$$d_{\text{eff}} = \frac{(\sum_k \lambda_k)^2}{\sum_k \lambda_k^2}. \tag{20}$$

This statistic summarizes how evenly spectral weight is distributed across components. DARK has greater participation-ratio dimensionality than static KD (10.0 vs. 8.0), while its 95%-variance rank is lower (74 vs. 77); these metrics summarize different aspects of the spectrum. Intra-class cosine also decreases (0.712 to 0.623), so the silhouette gain does not reflect improvement in every geometry statistic.

Dimensionality alone does not determine task score: confidence regularization has higher dimensionality than Static KD but lower Composite, whereas coupled repulsion has lower dimensionality but higher Composite (Tables S7 and S4).

D.5 Linear Probing

Linear probing evaluates frozen visual representations with a learned task-specific classifier. We extract ℓ_2-normalized features from the frozen image encoder and fit a single linear layer.

The six-view and three-class brain tasks use patient-level Planes DB training/test partitions of 7,129/5,271 and 1,543/1,406 images, respectively. For these tasks, five folds rotate the training

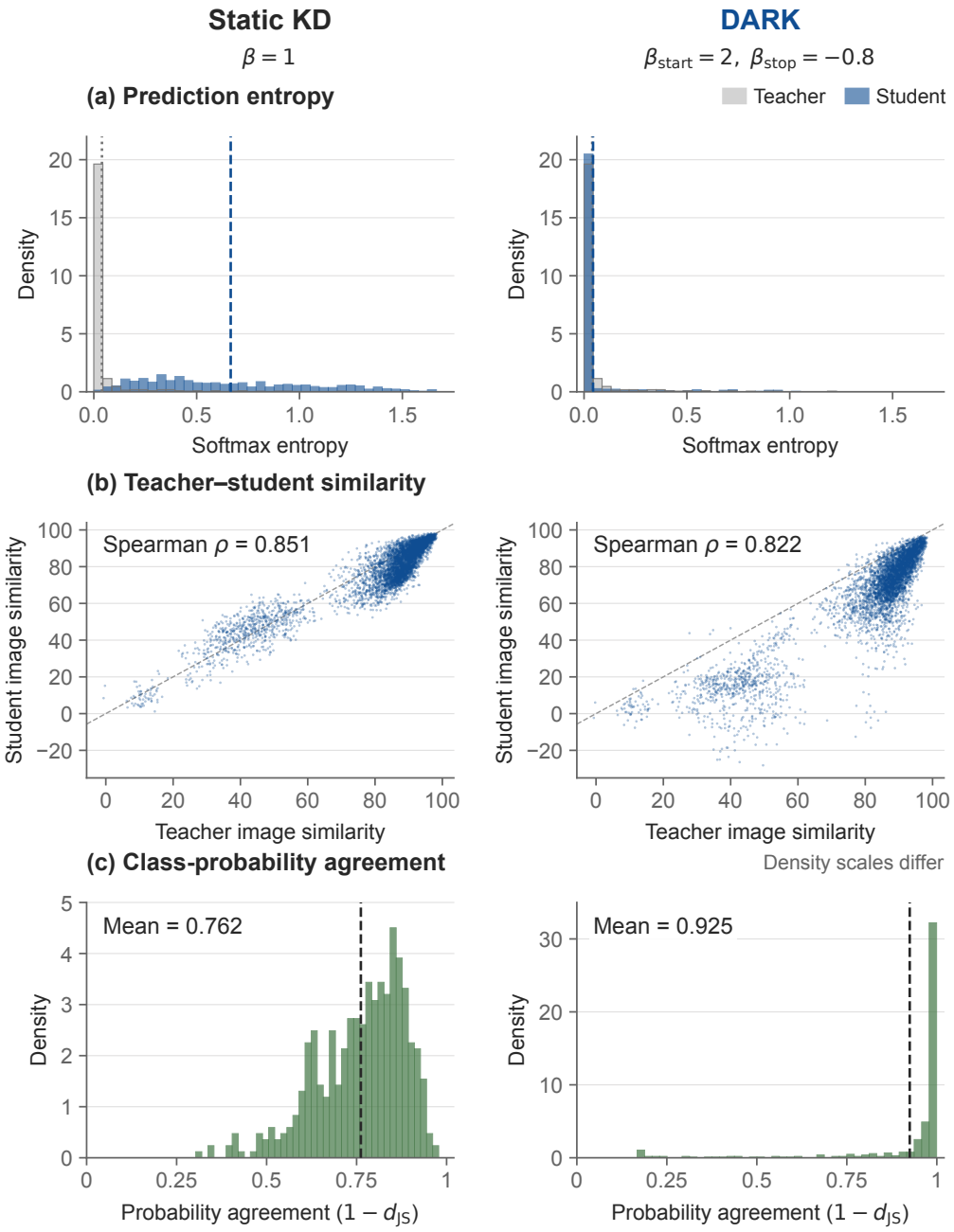

Figure S3: **Teacher–student image-similarity and probability analysis.** Columns compare static KD and DARK. (a) Entropy distributions; dotted gray and dashed blue lines mark teacher and student means. (b) Image–image similarities (100 times cosine similarity) with Spearman correlations; gray dashed lines denote equal similarities ($y = x$). (c) Per-image probability agreement ($1 - d_{\mathrm{JS}}$), with dashed mean lines. Corresponding panels share horizontal limits; entropy and similarity also share vertical limits. The agreement panels use different density scales to keep both distributions readable.

and validation assignments within the training partition, while the held-out test partition remains fixed. We fit five randomly initialized heads per fold and select each head by its lowest validation loss before evaluating it on the fixed test set, yielding 25 probe fits. These patient-level partitions differ from the full-pool zero-shot evaluations.

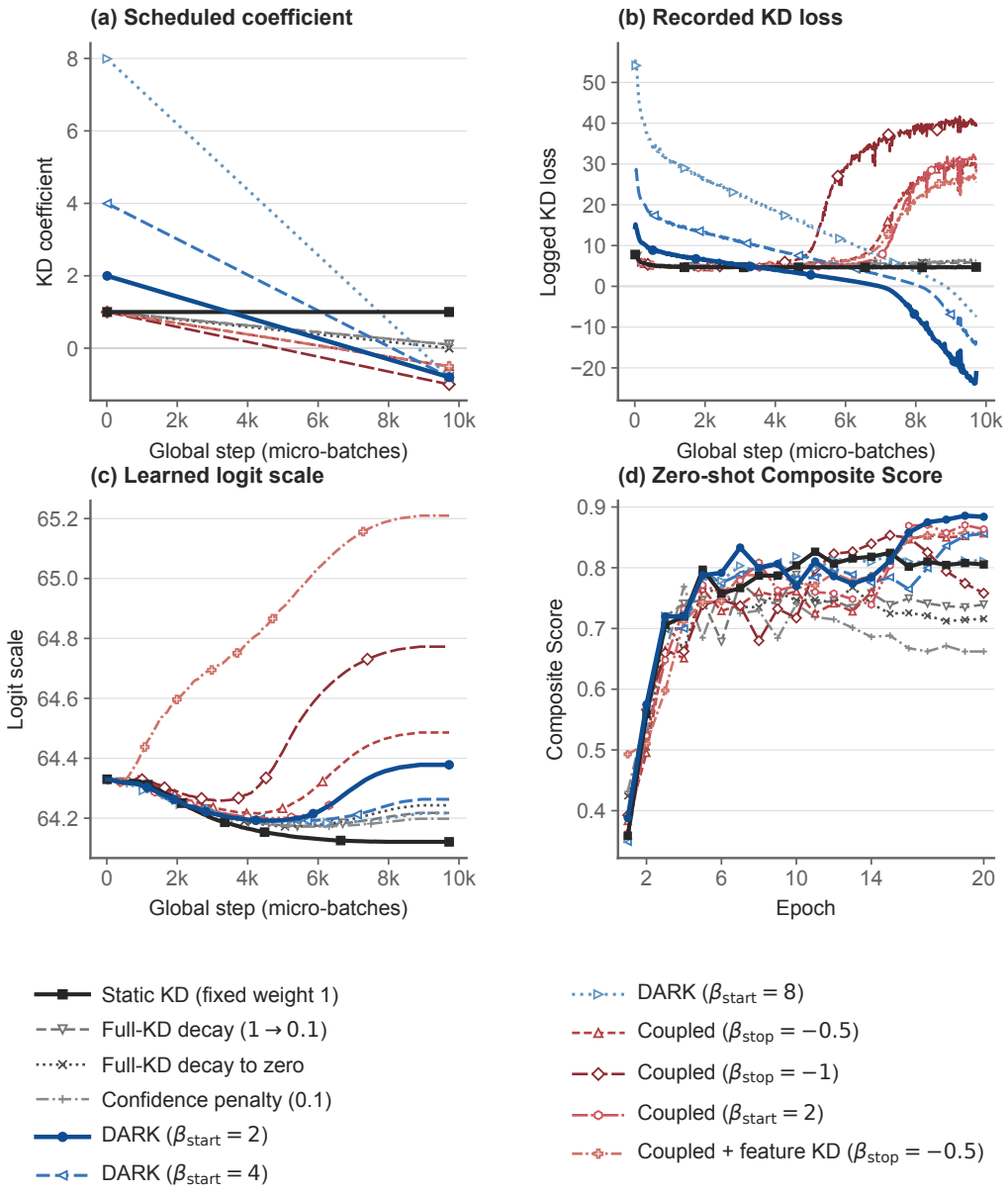

Figure S4: **Full logged training dynamics.** Panels show the scheduled coefficient, recorded distillation loss, learned logit scale, and Composite Score. The horizontal axes use micro-batch steps for the first three panels and epochs for the fourth. Line patterns and markers identify configurations. Full-KD decay to zero weakens both terms, unlike the anchored zero-hold control in Figure 2.

Congenital heart disease (CHD) detection uses 418 four-chamber fetal-heart videos (161 normal and 257 abnormal), with a patient-stratified training/test split of 333/85 videos. We encode 16 frames per video and concatenate their features for classification. CHD performance uses the area under the receiver operating characteristic curve (AUROC). Table S8 reports point estimates and 95% intervals across 25 probe fits per task. These intervals summarize variation across probe fits; the fits do not constitute 25 independent patient cohorts.

MobileFetalCLIP reaches F1 0.930 on six-view classification, F1 0.799 on brain sub-planes, and AUROC 0.769 on CHD detection. Relative to FetalCLIP's 0.947, 0.820, and 0.787, these corre-

Table S7: **Embedding geometry on the five-plane set.** Rank_{95} is the number of leading components accounting for 95% of the summed squared singular values of the centered embeddings. Intra- and inter-class values are cosine-similarity summaries.

Method	d_{eff}	Rank_{95}	Silhouette	Intra	Inter	Uniformity
Static KD	8.0	77	0.375	0.712	0.445	−1.662
Confidence penalty	9.0	84	0.406	0.693	0.389	−1.811
Coupled repulsion	6.4	50	0.509	0.645	0.010	−2.231
DARK	10.0	74	0.525	0.623	0.076	−2.308

spond to score ratios of 98.2%, 97.4%, and 97.7%. The ratios describe retained metric values, not equivalent relative error reductions. The student exceeds the three general/biomedical VLM references by 0.063–0.074 F1 on six-view, 0.165–0.217 F1 on brain, and 0.051–0.126 AUROC on CHD. The near-teacher probe results coexist with the two zero-shot gains, consistent with the two protocols testing different aspects of the learned representation.

Table S8: **Frozen-feature linear probes.** Point estimates with 95% intervals summarizing variation across probe fits. Retention is the student-to-teacher ratio of each point estimate. Bold marks the largest point estimate in each column.

Model	Six-view F1	Brain F1	CHD AUROC
CLIP	0.867 (0.866–0.869)	0.634 (0.632–0.637)	0.679 (0.650–0.708)
BiomedCLIP	0.856 (0.855–0.858)	0.582 (0.577–0.588)	0.643 (0.622–0.665)
UniMed-CLIP	0.860 (0.858–0.861)	0.607 (0.603–0.610)	0.718 (0.702–0.734)
FetalCLIP	**0.947** (0.947–0.948)	**0.820** (0.818–0.822)	**0.787** (0.770–0.804)
MobileFetalCLIP	0.930 (0.930–0.931)	0.799 (0.797–0.800)	0.769 (0.758–0.779)
Retention	98.2%	97.4%	97.7%

E Controlled Comparisons

E.1 Reporting Conventions and Checkpoint Selection

The component studies in Section 4.3 use one run (seed 42) and select checkpoints by Composite Score. The multi-seed studies use seeds $1, 7, 42, 123$ and evaluate the final epoch unless stated otherwise. Student–student comparisons use paired t-tests, and student–teacher comparisons use one-sample t-tests against the fixed teacher score. The tests are unadjusted, and the correlated task summaries do not turn four training seeds into 20 independent replicates. These tests describe variation across training runs, not patient-sampling uncertainty (Appendix C.5).

Schedule selection and robustness. We chose the default endpoints, $\beta_{\text{start}} = 2$ and $\beta_{\text{stop}} = -0.8$, by Composite Score on Planes DB and HC18, without a separate validation split, so scores at this setting may be optimistic. Three results show that DARK's gains do not rely on this choice. First, the setting was then used unchanged across all multi-seed and resolution studies, with no per-run adjustment. Second, the external cohort, which played no part in any selection, shows a gain of 0.125 macro-F1 over static KD, with improvements in every country and class (Section 4.4). Third, without checkpoint selection, DARK exceeds static KD in all four seeds on brain F1, HC18 validity and both aggregate scores (Table S14), and two other schedule shapes with the same endpoints reach similar Composite Scores (Table S12).

Table S9 compares the two checkpoint rules. At the final epoch, DARK improves all five metrics, and all four seeds favor DARK on each metric. This comparison is separate from the 224×224 comparison in Table 1.

Table S4 compares schedule endpoints, and Table S12 compares schedule shapes.

Table S9: **Checkpoint-rule comparison across four matched seeds.** Differences are DARK minus Static KD over seeds 1, 7, 42, and 123, for Composite-selected and final-epoch checkpoints. p-values are from unadjusted paired t-tests. This comparison is separate from the 224×224 comparison in Table 1.

	Selected checkpoint		Final epoch	
Metric	Δ (DARK $-$ Static)	p	Δ (DARK $-$ Static)	p
Five-plane F1	$+0.023$	0.1900	$+0.037$	0.0310
Brain F1	$+0.040$	0.3200	$+0.090$	0.0120
Combined F1	$+0.029$	0.1400	$+0.057$	0.0070
HC18 validity	$+0.096$	0.0002	$+0.094$	0.0250
Composite	$+0.063$	0.0020	$+0.075$	0.0130

Selected and final checkpoints are evaluated separately; the final-epoch comparison evaluates each completed training run.

E.2 Resolution and Architecture Controls

All students in the component studies use 256×256 inputs, so comparisons among them already match input resolution. The teacher uses 224×224. Training DARK and Static KD at a shared 224×224 retains gains in Composite, HC18 validity, brain F1 and combined classification F1 (Table 1). The same-architecture comparison at 256×256 (Table S10) adds DARK to a fixed FastViT student, separating the objective's contribution from the architecture's.

Table S10: **Final-epoch same-architecture comparison at 256×256.** DARK and no KD use the same FastViT architecture. Scores are means over four matched seeds (1, 7, 42, and 123); differences are DARK minus no KD on the $[0, 1]$ scale. p-values are from paired t-tests over these seeds.

Metric	No KD	**DARK**	Δ	p
Combined F1	0.818	0.879	$+0.060$	0.0230
HC18 validity	0.659	0.839	$+0.179$	0.0280
Composite	0.739	0.859	$+0.120$	0.0120

Means and differences are rounded independently.

Evaluating the four 256×256-trained final checkpoints at 224×224, without retraining, improves five-plane F1 by 0.013 ($p = 0.0029$) and combined classification F1 by 0.009 ($p = 0.0126$), with brain F1 unchanged. All four seeds favor the 224×224 input for both gains. We exclude HC18 from this across-resolution comparison because its prompts encode pixel spacing scaled to the input size. The student–teacher comparisons in Table S11 are made within 224×224, using resolution-specific HC18 handling. The teacher remains ahead on five-plane F1.

Table S11: **Final-epoch comparisons with the fixed teacher at 224×224 evaluation.** Columns separate DARK trained at 224×224 from DARK trained at 256×256. Each comparison uses four student seeds against the fixed teacher score, with one-sample t-test p-values. Positive differences favor DARK. A dash marks a comparison that was not computed.

	DARK trained at 224×224		DARK trained at 256×256	
Metric	Δ vs. teacher	p	Δ vs. teacher	p
Brain F1	$+0.073$	0.0018	$+0.068$	0.0020
Combined F1	–	–	$+0.017$	0.0036
HC18 validity	$+0.046$	0.0013	$+0.038$	0.0118
Composite	$+0.028$	0.0029	–	–

Both student and teacher are evaluated at 224×224 within each comparison. These are distinct training settings, not a direct HC18 comparison across evaluation resolutions.

E.3 Seven Off-Diagonal Schedules

The diagonal coefficient is fixed at one and only the off-diagonal schedule varies in this comparison. For normalized progress $u = t/E \in [0, 1]$, write $\beta_h(u) = 2 - 2.8h(u)$. Linear uses $h(u) = u$, recovering Equation 6 at $u = t/E$; half cosine uses $h(u) = (1 - \cos(\pi u))/2$. A step at v uses $h(u) = \mathbb{I}[u \geq v]$, with $v = 0.5$ or 0.714. Delayed onset uses $h(u) = 0$ before $u = 0.5$ and $h(u) = 2u - 1$ thereafter. The constant controls use $h \equiv 0$ or $h \equiv 1$; the endpoint conditions $h(0) = 0$, $h(1) = 1$ apply only to the transitioning schedules. Figure S5 plots the schedules, and Table S12 gives their final-epoch scores.

Table S12: **Final-epoch schedule controls.** Diagonal supervision stays at weight 1 while $\beta_h(u) = 2 - 2.8h(u)$ changes, with normalized training progress $u = t/E$. Bold marks the largest score in each column; shading identifies the default linear schedule. Each row is a single training run.

Off-diagonal schedule	Five-plane F1 ↑	Brain F1 ↑	HC18 validity ↑	Composite ↑
Linear (DARK)	0.944	0.771	0.859	**0.869**
Half cosine	0.912	0.776	**0.872**	0.867
Step at $u = 0.5$	0.700	0.793	0.419	0.577
Step at $u = 0.714$	0.902	**0.799**	0.859	0.861
Delayed onset at $u = 0.5$	0.950	0.675	0.789	0.818
Constant repulsion ($\beta = -0.8$)	0.487	0.589	0.299	0.412
Constant attraction ($\beta = 2$)	**0.953**	0.728	0.737	0.803

Constant attraction uses $\beta = 2$; Static KD uses $\beta = 1$. Each row reports a final-epoch point estimate for the stated coefficient function with diagonal supervision fixed at one.

The continuous mean coefficient is 0.60 for linear, half cosine and the midpoint step; their negative-phase durations are 28.6%, 35.9% and 50.0%. The late step approximately matches linear's negative-phase duration but has mean coefficient 1.20. For a step at exactly $5/7$, integrated positive weighting and integrated negative-weight magnitude are both twice their linear-schedule values. The comparison therefore does not isolate smoothness or duration alone. The constant-positive control uses $\beta = 2$, whereas the fixed-diagonal formulation recovers Static KD at $\beta = 1$. Each schedule is a single training run, so we do not report confidence intervals.

E.4 Zero-Hold Control

The zero-hold control tests whether stopping off-diagonal imitation is sufficient while retaining the diagonal anchor. For normalized training progress $u = t/E \in [0, 1]$, its coefficient is

$$\beta_{\text{zero}}(u) = \max\{0, 2 - 2.8u\}. \tag{21}$$

The diagonal coefficient remains one. The control shares DARK's positive coefficient trajectory and reaches zero at $u = 5/7$; DARK then continues toward -0.8, whereas the control stays at zero. This differs from full-KD decay, which also weakens diagonal supervision.

We train the control with seed 42 and the same configuration as the default DARK run: the same student and teacher initialization, corpus, preprocessing, optimization settings, and evaluation protocol (Appendices A and C). Both use 256×256 student inputs, 224×224 teacher inputs, and the same 20-epoch training budget. The sole objective change in the control clamps the computed off-diagonal coefficient at zero. Data shuffling is not fully seeded, so the two runs follow separate training trajectories; they do not share the positive phase as a common prefix. Figure 2 plots both schedules and their Composite scores over training.

Table 2(b) compares the completed schedules at the end of training. DARK exceeds zero-hold by 0.125 Composite, 0.105 five-plane F1, 0.046 brain F1, and 16.7 percentage points of HC18 validity. Combined classification F1 is 0.888 for DARK and 0.805 for zero-hold. Under Composite-based checkpoint selection (Table S4), zero-hold instead reaches 0.806, compared with DARK's 0.886. That selected control checkpoint precedes its zero-hold phase; it has higher five-plane F1 (0.956 vs. 0.946), but lower brain F1 and HC18 validity. The completed-schedule comparison is therefore the relevant test, and it supports the repulsive phase.

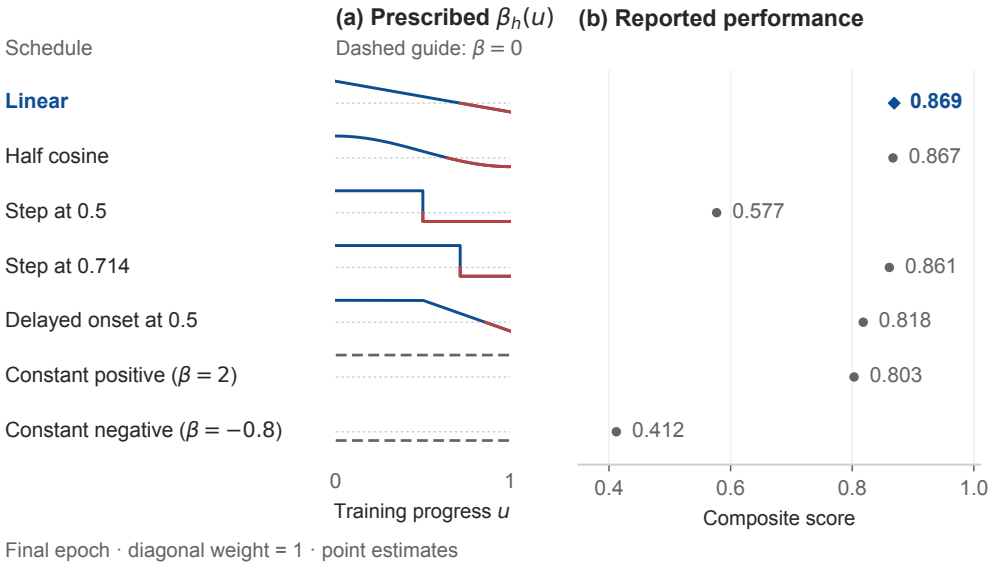

Figure S5: **Changing the supervision schedule changes the outcome.** Prescribed off-diagonal schedules and final-epoch Composite point estimates, with diagonal weight fixed at one. Constant attraction uses $\beta = 2$, whereas Static KD uses $\beta = 1$. Full task scores are in Table S12; each schedule is a single run.

E.5 Student-Size Study

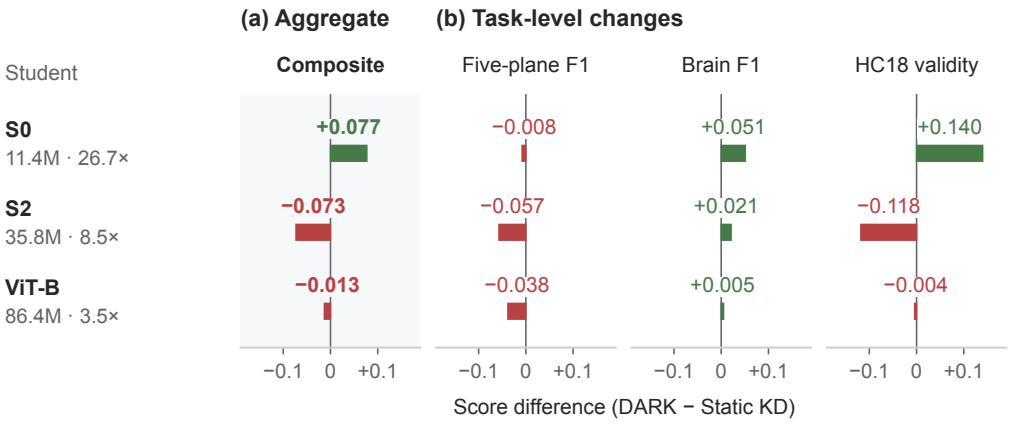

Figure S6: **The largest tested compression gap yields the aggregate gain.** DARK minus Static KD at 224×224, final epoch after 20 epochs, seed 42. Sizes and compression ratios concern the visual encoder. Larger configurations lose Composite despite small brain-F1 gains. Full scores are in Table S13.

The size study uses a frozen ViT-L/14 teacher with 304M visual parameters, shared 224×224 input resolution, learning rate 10^{-5}, teacher KD temperature 5, 20 epochs and seed 42. Every row uses the final epoch. Table S13 reports all task scores; Figure S6 shows the Composite and task differences. The S0 student gains 0.077 Composite, whereas S2 and ViT-B lose 0.073 and 0.013. Brain F1 increases in all three configurations, but larger students lose five-plane F1 and HC18 validity. One explanation is that a student with spare capacity can push non-target probabilities much closer to

the lower bound in Appendix B.3 once repulsion begins, whereas the capacity-limited S0 student cannot. On this view, the capacity gap that limits imitation also keeps repulsion stable; we do not test this explanation directly. DARK's aggregate gain is therefore concentrated at the largest tested compression gap; three configurations do not pin down an exact threshold. S0 and S2 belong to the same MobileCLIP2 family; ViT-B changes architecture as well as capacity.

Table S13: **DARK across three student configurations.** All students are trained and evaluated at 224×224 for 20 epochs (seed 42), using a frozen teacher with 304M visual parameters, learning rate 10^{-5}, and KD temperature 5. Results use the final epoch. Bold marks the larger score within each student pair; parameter counts and compression ratios refer to visual encoders.

Objective	Five-plane F1 ↑	Brain F1 ↑	Combined F1 ↑	HC18 validity ↑	Composite ↑
S0: 11.4M visual parameters; $26.7\times$ *compression*					
Static KD	**0.943**	0.729	0.863	0.747	0.805
DARK	0.936	**0.780**	**0.877**	**0.887**	**0.882**
S2: 35.8M visual parameters; $8.5\times$ *compression*					
Static KD	**0.952**	0.753	**0.877**	**0.808**	**0.843**
DARK	0.895	**0.774**	0.850	0.690	0.770
ViT-B: 86.4M visual parameters; $3.5\times$ *compression*					
Static KD	**0.932**	0.783	**0.876**	**0.802**	**0.839**
DARK	0.893	**0.788**	0.854	0.799	0.826

These are single-seed, final-epoch results (seed 42). Changes in architecture and configuration accompany the changes in visual-encoder size.

The size study uses one seed per configuration, so it is not a replicated capacity sweep. Some differences deviate by 0.001 from the difference of the displayed values, consistent with rounding scores and differences separately.

E.6 Absolute Results across Matched Seeds

Table S14 gives the individual final-checkpoint scores behind the primary 224×224 comparison in Table 1. Means and sample standard deviations across the four seeds describe variation between training runs. DARK improves brain F1, HC18 validity and both aggregate scores in every pair. Five-plane F1 is similar in mean, with DARK higher in two of the four pairs.

F External Validation on African Fetal Planes

An external cohort tests transfer across five African countries. We evaluate the public African fetal-plane release of Sendra-Balcells et al. (2023) (Zenodo record 7540448), containing 450 images from Algeria, Egypt, Ghana, Malawi and Uganda. The cohort had not been used for training, prompt adjustment, or tuning DARK. The cohort is not part of our hospital or textbook training data; we did not check it against the web-scale data used to pretrain the student's initial weights. Models, scoring rules, duplicate handling and the primary comparison were fixed before inspecting predictions. We combine the release's original training and test partitions for frozen evaluation, with no training or adaptation on either partition. The results therefore concern this complete external cohort, rather than the release's designated test partition.

Duplicate handling and sampling units. Exact comparison of decoded RGB pixels and image dimensions identifies 47 duplicate pairs, leaving 403 unique images. The primary analysis retains the lexicographically first image identifier in each duplicate group, without filtering by appearance, difficulty or model performance. The retained class counts are 112 abdomen, 115 brain, 113 femur and 63 thorax images. We first group records by country and the patient number provided with the release, then merge groups connected by exact duplicate images, including transitive links. This yields 112 conservative dependence groups: 19, 23, 25, 20 and 25 in Algeria, Egypt, Ghana, Malawi and Uganda, respectively. These groups are not 112 verified distinct patients; additional unidentified

Table S14: **Individual final-checkpoint results for the primary matched-resolution comparison.** Both methods are trained and evaluated at 224×224 for 20 epochs. All scores lie in $[0, 1]$; higher is better. Summary rows give the mean and sample SD across four training seeds.

Seed	Five-plane F1	Brain F1	Combined F1	HC18 validity	Composite
Static KD					
1	0.934	0.722	0.855	0.795	0.825
7	0.954	0.670	0.848	0.765	0.806
42	0.943	0.729	0.863	0.747	0.805
123	0.941	0.682	0.844	0.771	0.808
Mean $\pm$ SD	0.943 ± 0.008	0.701 ± 0.029	0.852 ± 0.009	0.770 ± 0.020	0.811 ± 0.009
DARK					
1	0.950	0.775	0.884	0.887	0.886
7	0.947	0.788	0.887	0.878	0.883
42	0.936	0.780	0.877	0.887	0.882
123	0.943	0.756	0.873	0.871	0.872
Mean $\pm$ SD	0.944 ± 0.006	0.775 ± 0.014	0.880 ± 0.007	0.881 ± 0.008	0.881 ± 0.006

Means and SDs are computed before rounding; SD measures variation across training seeds, not uncertainty across patients.

patient overlap remains possible. No duplicate group crosses countries or contains conflicting class labels. The prespecified sensitivity analysis retains all 450 original image records.

Frozen models and scoring. We use the eight final-epoch student checkpoints from the 224×224 comparison: DARK and Static KD at seeds $1, 7, 42, 123$. The reference is the frozen FetalCLIP teacher with its original architecture and 117-token context. Evaluation preserves centered zero-padding, model-specific preprocessing and normalization, and the prompt ensemble and nine classification candidates of Appendix C.1. Thorax maps to the heart candidate; predictions outside the cohort's ground-truth classes remain errors. The primary metric is pooled-image macro-F1 over abdomen, brain, femur and thorax, computed separately for each model and then averaged across the four students per method. We do not ensemble predictions. Country macro-F1 uses the same four classes except in Ghana and Uganda, which have no thorax images and use the other three; comparisons across countries therefore involve different class sets.

Paired uncertainty conditional on the fitted models. We generate 10,000 bootstrap replicates by sampling dependence groups with replacement within each country, keeping each country's observed group count fixed. All retained images from a sampled group enter with its sampling multiplicity, and every model uses the same draw. For each replicate, we recompute each model's pooled-image macro-F1 and the difference between the two four-model means. The 2.5th and 97.5th percentiles give the 95% intervals. Class sets remain fixed, and no replicate omits a scored class. These intervals describe cohort-sampling uncertainty conditional on the fitted models, observed countries and grouping assumptions; they do not include retraining uncertainty or variability across new sites or countries. The pooled DARK–Static KD difference is primary; teacher comparisons and country intervals are secondary and are not adjusted for multiplicity.

DARK improves on Static KD and matches the teacher. Mean macro-F1 rises from 0.7697 to 0.8947, a difference of 0.1250 with 95% interval $[0.1019, 0.1498]$ (Table S15). All four matched student pairs favor DARK; sample SD across the four scores is 0.0068 for DARK and 0.0230 for Static KD, reflecting training-run variation rather than cohort sampling. Mean DARK macro-F1 exceeds Static KD in each country, and retaining all duplicate records gives a similar pooled difference, 0.1240. The teacher scores 0.8939; DARK minus teacher is 0.0008 with interval $[-0.0249, 0.0271]$, so the $26\times$ smaller student performs on par with its teacher, although this interval does not establish formal equivalence. The teacher remains ahead of DARK in Ghana and Uganda. The release labels broad brain anatomy, so this evaluation does not test three-way brain sub-plane classification or gestational-age prediction.

Table S15: **External macro-F1 on African fetal planes.** Student scores average four separately evaluated models. The primary cohort contains 403 unique images; the final row retains the full release as a sensitivity analysis. Δ is DARK minus Static KD, with paired 95% cohort-bootstrap intervals.

Cohort	Images	Static KD	**DARK**	Teacher	Δ	95% interval
Pooled (primary)	403	0.7697	0.8947	0.8939	+0.1250	[0.1019, 0.1498]
Algeria	76	0.9369	0.9863	0.9304	+0.0494	[0.0179, 0.0912]
Egypt	97	0.8010	0.8696	0.8678	+0.0686	[0.0268, 0.1202]
Ghana	75	0.7682	0.8679	0.8911	+0.0997	[0.0516, 0.1580]
Malawi	80	0.6178	0.8436	0.8210	+0.2258	[0.1599, 0.2938]
Uganda	75	0.6072	0.8466	0.9340	+0.2394	[0.1567, 0.3302]
Full release	450	0.7774	0.9014	0.8962	+0.1240	[0.1008, 0.1482]

Ghana and Uganda average three classes; all other rows average four. Country intervals are descriptive and not multiplicity-adjusted. The pooled score is computed across images, not by averaging country scores.

Table S16: **Class-level and overall external performance.** The primary cohort contains 403 unique images. Class rows report F1; student entries average four separately evaluated models. Macro-F1 is the primary metric. Micro-F1 is a secondary descriptive summary over all nine prediction candidates and equals accuracy.

Class or metric	Images	Static KD	**DARK**	Teacher
Abdomen	112	0.5572	0.8253	0.8103
Brain	115	0.9677	0.9913	0.9515
Femur	113	0.7141	0.8214	0.8932
Thorax	63	0.8398	0.9407	0.9206
Macro-F1 (four classes)	403	0.7697	0.8947	0.8939
Micro-F1 / accuracy	403	0.6663	0.8207	0.8362

Thorax ground truth maps to the heart candidate. Predictions to the other five candidates remain errors. Class and micro-F1 summaries are descriptive; no uncertainty intervals are estimated for these rows.

Class-level and reproducibility checks. Table S16 gives per-class F1 and secondary micro-F1, equal to accuracy over all nine candidates. We save all predictions and class probabilities, and an independent implementation reproduces all nine model scores.