WHEN DOES ATTRIBUTION STABILITY IMPROVE EXPLANATION-GUIDED LEARNING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Teacher attribution maps provide spatial guidance without region annotations, but the same teacher can highlight different regions across label-preserving views. Uniform alignment gives these inconsistent targets the same weight as repeatable ones. We propose Perturbation-Consistent Gated Attribution Alignment (PCGA), which averages class-matched teacher maps into a spatial target and uses their cross-view agreement to weight student alignment. Matched comparisons against uniform and prediction-derived weights test whether agreement improves learning. On Chest CT, gated cosine alignment improves AUC over conventionally tuned uniform alignment, although matching the auxiliary-gradient strength removes that advantage. Prediction-confidence weighting leads on RADGEN. Stability weighting has the highest mean AUC on CIFAR-10, with gains that vary across seeds. Gradient measurements and spatial interventions clarify the distinction: map agreement chiefly controls alignment strength in the Chest CT comparison and does not consistently identify regions that affect predictions. Cross-view agreement can therefore set the strength of attribution supervision, but its value as an image-ranking signal depends on the task.

1 INTRODUCTION

In image-level CT classification, a label identifies the class but not the image regions that support the decision. A frozen teacher can supply class-specific attribution maps for student alignment. Yet the same teacher may emphasize different regions after appearance changes that preserve the label. Averaging those maps gives the student a target while concealing disagreement among views. Uniform alignment then treats a variable target like one on which the views agree. This raises a practical question: should disagreement reduce the target's influence on student learning?

Existing guidance methods obtain a spatial signal in several ways. Expert-informed gradient constraints use annotated relevant or irrelevant regions (Ross et al., 2017; Rieger et al., 2020), which image-label-only training lacks. Explanation-enhanced distillation transfers both teacher predictions and attributions (Parchami-Araghi et al., 2024), while sample-level meta-weighting uses held-out task performance to tune auxiliary supervision (Grégoire et al., 2024). These approaches address the content or downstream value of guidance. They leave open a more local decision available before student training: whether variation among a frozen teacher's views should change the weight of an otherwise fixed attribution target.

Agreement across mild appearance changes measures how repeatably the teacher identifies a region. It does not tell us whether that region supports the prediction. Saliency maps can remain similar after model randomization (Adebayo et al., 2018) or change under prediction-preserving perturbations (Ghorbani et al., 2019). Medical-image studies likewise find that predictive accuracy and explanation localization can diverge (Arun et al., 2021; Saporta et al., 2022). We therefore evaluate the weight through student learning and test the selected regions separately through spatial interventions.

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

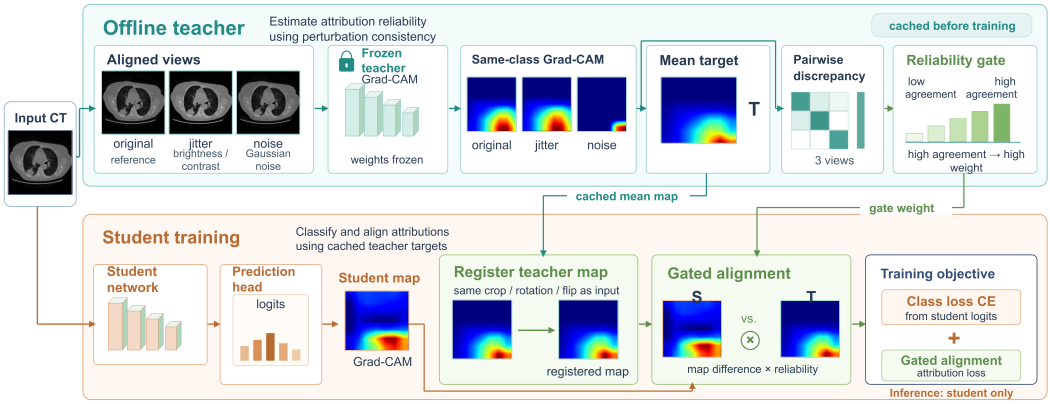

Figure 1: PCGA training pipeline. A frozen teacher produces class-matched maps under aligned appearance views. Their mean supplies the spatial target, while pairwise disagreement determines its alignment weight. Both are obtained before student training. Inference uses only the student classifier.

In this paper, we propose PCGA to turn cross-view teacher-map agreement into a sample-level alignment weight. The views produce a mean spatial target, while their disagreement controls the strength of differentiable student alignment, as illustrated in Figure 1. We match attribution class and spatial coordinates so that comparisons change the weight rather than the target. A gate can then affect learning in two distinct ways: its mean changes the auxiliary-gradient magnitude, and its variation across images can change the update direction. Matched-state gradient measurements and a scale-only training control distinguish these effects.

We compare uniform, stability, and prediction-derived weights using matched targets and validation selection. We then vary the alignment objective and measure how deletion and retention of attributed regions affect predictions. The main contributions are:

- We formulate cross-view attribution agreement as an alignment weight and show that gated cosine alignment improves Chest CT AUC over conventionally tuned uniform alignment.
- Matched comparisons establish that the preferred weight changes with the task and loss: prediction confidence leads on RADGEN, and the stability gate does not consistently lead under L1 alignment.
- A gradient decomposition and scale-only control identify aggregate auxiliary strength as the principal source of the Chest CT gain. Spatial interventions show that map repeatability and prediction response are distinct properties.

2 Related Work

Explanation-guided supervision. Explanation-guided learning specifies where a model should respond. Expert masks, interaction, or contextual feedback define relevant and irrelevant evidence (Ross et al., 2017; Rieger et al., 2020; Schramowski et al., 2020; Zhang et al., 2023). Other objectives promote attribution energy, class contrast, or medically plausible alignment (Rao et al., 2023; Mahapatra et al., 2022; Guo et al., 2025; Shu et al., 2026; Liu et al., 2025). Recent work adapts guidance from sparse expert interactions or augments distillation data with Integrated Gradients (Moazen et al., 2026; Hernandez et al., 2026). These methods shape the supervision itself. When a frozen teacher supplies one target per image, however, its changing attribution across label-preserving views raises a different question: how strongly should that target guide the student? We address this question by holding the target fixed and varying its weight.

Teacher explanations and sample weighting. Teacher explanations provide supervision without expert region annotations. Explanation-enhanced distillation jointly aligns student predictions and attributions with a teacher (Parchami-Araghi et al., 2024). SLGrad instead learns auxiliary-task

sample weights from a held-out objective (Grégoire et al., 2024). Prediction transfer and validation-based weighting capture different sources of guidance, but neither directly measures the repeatability of a fixed teacher target. We compare this repeatability with uniform and prediction-derived weights while matching the targets, losses, and validation-selection budgets.

Stability and explanation quality. The distinction between repeatability and useful supervision also appears in attribution evaluation. Grad-CAM produces class-specific maps from feature activations and gradients (Selvaraju et al., 2017), while randomization and prediction-preserving perturbations expose different ways such maps can fail (Adebayo et al., 2018; Ghorbani et al., 2019). Geometry, infidelity, and sensitivity probe other aspects of explanation behavior (Dombrowski et al., 2019; Yeh et al., 2019). Medical-image studies further separate classification from localization (Arun et al., 2021; Saporta et al., 2022), and MetaQuantus assesses quality estimators through noise resilience and randomization reactivity (Hedström et al., 2023). Such diagnostics characterize maps and their evaluators. To learn whether repeatable teacher maps deserve more weight, we compare student training outcomes and examine both optimization gradients and spatial response.

3 Attribution Weights and Their Optimization Effect

3.1 Perturbation-consistent teacher targets

We construct a teacher target and measure its variation across views. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ contain N labeled training images, where x_i is an image and y_i its class label. A frozen teacher f_t supplies targets for a student f_θ with parameters θ. Both models return class logits. The Grad-CAM operator $E(f, x, q)$ returns the attribution of model f on image x for class q, resized to an $H \times W$ grid of height H and width W. For a raw map A, E applies min–max normalization $\mathrm{MM}(A) = (A - \min A)/(\max A - \min A + 10^{-8})$ after resizing. The target class q_i is either y_i or the teacher's predicted class on x_i, according to the evaluation condition. Let a_m be appearance transformation m among V views. The maps and their mean target are

$$T_i^{(m)} = E(f_t, a_m(x_i), q_i), \qquad T_i = \frac{1}{V}\sum_{m=1}^{V} T_i^{(m)}. \tag{1}$$

We use $V = 3$ spatially aligned views: the original image, brightness/contrast jitter in $[-0.15, 0.15]$, and additive Gaussian noise with standard deviation 0.05 in denormalized image space. With gate scale $\tau = 0.5$, the mean pairwise pixelwise L_1 discrepancy $\bar{D}_i$ and gate g_i are

$$\bar{D}_i = \frac{2}{V(V-1)} \sum_{m<n} \frac{\|T_i^{(m)} - T_i^{(n)}\|_1}{HW}, \tag{2}$$

$$g_i = \exp(-\bar{D}_i/\tau). \tag{3}$$

Here $\|\cdot\|_1$ sums absolute pixel differences and $m < n$ ranges over distinct view pairs. The target and gate are obtained before student training, as illustrated in Figure 1. A larger gate indicates closer agreement under the chosen perturbations. Full-grid averaging can nevertheless give shared background more influence than disagreement over salient regions. Two disjoint binary regions each covering 1% of the grid, for example, have distance 0.02 and weight $\exp(-0.02/0.5) \approx 0.961$.

For a salient-region control, let R_i be the union of the top-25% pixels in each view, with p indexing a spatial pixel and $|R_i|$ the number of selected pixels. We replace the spatial mean in Equation 2 by $|R_i|^{-1} \sum_{p \in R_i} |T_i^{(m)}(p) - T_i^{(n)}(p)|$, leaving the target, loss, data, and selection rule unchanged.

3.2 Attribution alignment and spatial registration

In the main setting, $q_i = y_i$, so teacher and student maps target the same class. The teacher-predicted-class control changes only the teacher target. The two classes may then differ. Let b_i denote a sampled crop, rotation, and flip. The student map is $S_i = E(f_\theta, b_i(x_i), y_i)$, and $\widehat{T}_i = b_i(T_i)$ applies the same geometry and resampling to the teacher target. For minibatch size B, let $d_i(\theta)$ be the selected map distance, $w_i \geq 0$ its weight, λ the alignment coefficient, and CE cross-entropy. We write $\mathcal{L}_c$ and $\mathcal{L}_a$

for batch-mean classification and alignment losses. With vec denoting map flattening and cos cosine similarity, the distances and shared objective are

$$
\begin{aligned}
d_i^{(1)}(\theta) &= \frac{1}{HW}\|S_i - \widehat{T}_i\|_1, \\
d_i^{(c)}(\theta) &= 1 - \cos\big(\mathrm{vec}(S_i), \mathrm{vec}(\widehat{T}_i)\big), \\
\mathcal{L}_c(\theta) &= \frac{1}{B}\sum_i \mathrm{CE}(f_\theta(b_i(x_i)), y_i), \\
\mathcal{L}_a(\theta) &= \frac{1}{B}\sum_i w_i d_i(\theta), \\
\mathcal{L}(\theta) &= \mathcal{L}_c(\theta) + \lambda \mathcal{L}_a(\theta).
\end{aligned}
\tag{4}
$$

(5)

Each experiment selects $d_i = d_i^{(1)}$ or $d_i^{(c)}$, while the gate discrepancy remains pixelwise L_1. Uniform and stability weighting set $w_i = 1$ and $w_i = g_i$, respectively. Student maps remain differentiable through Grad-CAM. The unregistered control resizes teacher maps without applying b_i. All four factorial conditions pad images to an approximately square canvas and resize to 256×256. Teacher maps cover the full canvas, while student augmentation produces 224×224 inputs.

To compare map repeatability with prediction-derived reliability while holding the teacher target fixed, let $p_t(k \mid x)$ be the teacher's probability of class k for input x, and let C be the number of classes. Write $\bar{p}_i$ for the mean class-probability distribution across views and $\mathcal{H}$ for Shannon entropy. The mean distribution, confidence weight c_i, and entropy-based weight h_i are

$$
\bar{p}_i = \frac{1}{V}\sum_{m=1}^{V} p_t(\cdot \mid a_m(x_i)), \qquad c_i = \max_k p_t(k \mid x_i), \qquad h_i = 1 - \frac{\mathcal{H}(\bar{p}_i)}{\log C}. \tag{6}
$$

Confidence c_i uses the identity view, while entropy-based weight h_i uses all views. Both weights are computed on training images and replace w_i in Equation 5. Let N_k be the number of training images in class k and μ_k their mean gate value. To control differences in total weight across classes, we also evaluate the class-normalized gate $\widetilde{g}_i$:

$$
\mu_k = \frac{1}{N_k}\sum_{j:y_j=k} g_j, \qquad \widetilde{g}_i = \frac{g_i}{\mu_{y_i}}. \tag{7}
$$

Class normalization preserves within-class ordering and sets each class's mean weight to one. It uses training labels, whereas the raw gate requires only the maps for the chosen attribution class. Teacher maps and weights are used during training only. Inference requires the student classifier.

3.3 Separating scale from direction

A weighted loss changes both the strength of alignment and the relative influence of examples. To distinguish the two, let $v_i = \nabla_\theta d_i$ be the per-image alignment gradient and $\bar{w} = B^{-1}\sum_i w_i$ the minibatch-mean weight. Before multiplication by λ, the weighted auxiliary gradient decomposes as

$$
\frac{1}{B}\sum_i w_i v_i = \bar{w}\,\frac{1}{B}\sum_i v_i + \frac{1}{B}\sum_i (w_i - \bar{w}) v_i. \tag{8}
$$

The first term rescales uniform alignment. The residual

$$
r_w = \frac{1}{B}\sum_i (w_i - \bar{w})(v_i - v), \qquad v = \frac{1}{B}\sum_i v_i, \tag{9}
$$

captures changes in direction. Define the minibatch dispersions $\sigma_w^2 = B^{-1}\sum_i (w_i - \bar{w})^2$ and $\sigma_v^2 = B^{-1}\sum_i \|v_i - v\|_2^2$. Cauchy–Schwarz gives

$$
\|r_w\|_2 \le \sigma_w \sigma_v. \tag{10}
$$

Thus, image weighting can substantially redirect the auxiliary update only when the weights and per-image gradients are both dispersed. If $\bar{w}\|v\|_2 > \|r_w\|_2$, then

$$
\cos(v, v_w) \ge \frac{\bar{w}\|v\|_2 - \|r_w\|_2}{\bar{w}\|v\|_2 + \|r_w\|_2}, \qquad v_w = \bar{w}v + r_w. \tag{11}
$$

The bound predicts nearly collinear updates when the residual is small relative to the rescaled uniform gradient. Appendix F provides the derivation.

Let $u = \nabla_\theta \mathcal{L}_c$ be the classification gradient. We compare the norms and cosine similarities of u, v, and v_w on the same augmented batch and model state. To control scale during training, we use a batchwise coefficient

$$\alpha = \min\left(10,\ 0.3\frac{\|u\|_2}{\|v_w\|_2}\right), \qquad \mathcal{L} = \mathcal{L}_c + \alpha\mathcal{L}_a, \tag{12}$$

and disable the auxiliary term when $\|v_w\|_2 \le 10^{-12}$. The coefficient is held fixed when differentiating the training loss.

To test whether image-specific weights matter beyond total auxiliary strength, we also construct a scale-only cosine control. At each training batch and model state, let v and v_w be the uniform and stability-weighted gradients of the cosine alignment loss. We compute $\beta = \|v_w\|_2/\|v\|_2$ and train with

$$\mathcal{L}_s = \mathcal{L}_c + \lambda\beta\frac{1}{B}\sum_i d_i^{(c)}(\theta). \tag{13}$$

We hold β fixed for the parameter update and set it to zero if both gradient norms vanish. The control matches PCGA's auxiliary-gradient norm on every batch while removing the image-specific weighting from the update direction. Its λ is the validation-selected PCGA coefficient, with no new search. We also scale the confidence-weighted cosine gradient to the same PCGA gradient norm at each batch, using the same λ and no additional search. This retains confidence-based sample ranking while matching total auxiliary strength.

3.4 Spatial-response diagnostic

We separately measure whether highly attributed regions affect predictions under deletion and retention. Let M select the highly attributed region and let Q be an equal-area random mask. For $Z \in \{M, Q\}$, $x_{d,Z}$ replaces pixels inside Z and $x_{r,Z}$ replaces pixels outside Z. If p_q is the evaluator probability for attribution class q, the two advantages are

$$A_d = \mathbb{E}_Q[p_q(x_{d,Q})] - p_q(x_{d,M}), \tag{14}$$

$$A_r = p_q(x_{r,M}) - \mathbb{E}_Q[p_q(x_{r,Q})]. \tag{15}$$

Positive values indicate a stronger response than equal-area random masks. The intervention is evaluated separately from student training.

4 Experimental Design

Tasks and models. The two CT tasks differ in cohort size and teacher quality. RADGEN uses one supplied CT viewer export per patient from the NSCLC Radiogenomics collection (Bakr et al., 2018). Its binary EGFR task has 119 training, 17 validation, and 35 test patients. Chest CT contains 833 unique-content images from four classes, partitioned 484/65/284 (Hany, 2020). Both tasks use ImageNet-initialized ResNet-50 teachers and students (He et al., 2016; Deng et al., 2009). The frozen RADGEN teacher reaches 57.69% validation AUC and predicts only the majority class, whereas the Chest CT teacher reaches 99.52% AUC. This contrast tests weighting under weak and strong teacher evidence. Final-block Grad-CAM supplies the maps. Appendix A reports image provenance, preprocessing, and teacher diagnostics.

Training and selection. Students train for 50 epochs with Adam (Kingma & Ba, 2015), batch size 16, learning rate 10^{-4}, and weight decay 10^{-4}. L1 alignment searches $\lambda \in \{0.1, 0.3, 0.6\}$ and cosine alignment searches $\{0.3, 1, 3\}$ using seeds 42, 123, and 456. Mean validation AUC across seeds selects one coefficient per method before test evaluation. Fixed-configuration replications use seeds 17, 43, and 2024. Appendix A gives the complete coefficient grids.

Independent student task. We also train students from scratch on CIFAR-10 (Krizhevsky, 2009), using disjoint subsets of 10,000 training, 1,000 validation, and 1,000 diagnostic images from the

official training split. The official test split is reserved for evaluation after validation selection. A frozen ResNet-18 teacher supplies the same class-matched target to uniform, stability-weighted, and confidence-weighted cosine alignment. Classification-only training and an auxiliary-gradient-norm-matched uniform control complete the comparison. All students use the same architecture, 30-epoch schedule, and three seeds. Appendix I gives the selection rule and full results.

Comparison objectives. The matched L1 comparison changes only w_i in Equation 5. Uniform, stability, prediction-confidence, and predictive-entropy weights share the teacher target and validation budget. We also include IGIB-style class contrast (Mahapatra et al., 2022), RRR-style outside-gradient suppression (Ross et al., 2017), attribution-energy guidance (Rao et al., 2023), prediction distillation, and adapted e^2KD (Parchami-Araghi et al., 2024). All use registered ground-truth-class targets when the objective permits. Appendix A specifies the adapted objectives and coefficient grids.

Metrics and uncertainty. Binary ROC AUC is primary on RADGEN and macro one-vs-rest ROC AUC on Chest CT and CIFAR-10. Balanced accuracy and macro-F1 measure class-balanced prediction on CT, while CIFAR-10 also reports accuracy. We report sample standard deviations across seeds and paired AUC differences. Seed bootstrap holds the cohort fixed. Class-stratified cohort bootstrap resamples the same cases for both methods, using patients on RADGEN and images on Chest CT. Appendix A specifies the CT resampling budgets; Appendix I gives the CIFAR-10 comparison.

Response and optimization diagnostics. The gradient probe compares weighting rules on matched augmented batches at initialization and at the selected uniform-alignment state, with model parameters fixed during measurement. Spatial response uses the CT validation sets and a disjoint CIFAR-10 diagnostic partition (Krizhevsky, 2009). Separate teacher and evaluator models prevent the source model from defining its own success criterion. Appendix F and Appendix E contain the diagnostic protocols.

5 Results

5.1 Weighting benefits depend on the task and alignment objective

We first ask whether cross-view stability improves student prediction when the spatial target is held fixed. Uniform, stability, and prediction-derived weights use the same registered ground-truth-class teacher maps and three-seed validation rule. Table 1 also includes alternative explanation-guidance and distillation objectives, placing the matched weight comparison in a broader performance context.

For cosine alignment on Chest CT, stability weighting exceeds conventionally tuned uniform alignment by 0.29 AUC percentage points in Table 1. Balanced accuracy follows the same ordering. The two variants share the target, loss type, candidate grid, and validation-selection rule, but select different coefficients. Stability weighting also exceeds adapted e^2KD by 0.41 AUC points; that method additionally distills predictions and uses online teacher maps. The paired image-bootstrap intervals for both differences exclude zero. Once auxiliary-gradient strength is matched batchwise, uniform alignment reaches a slightly higher mean AUC than stability weighting. Norm-matched confidence weighting trails PCGA on Chest CT by 0.37 AUC points (paired image-bootstrap interval [0.12, 0.70]); its RADGEN mean is higher, although the patient interval spans zero and balanced accuracy is near chance. The Chest CT comparison therefore favors stability over confidence weighting, while the uniform control accounts for its gain over conventional alignment.

Full-map agreement can also reflect shared low-attribution background, so we recompute the gate over salient regions. This raises RADGEN mean AUC slightly relative to full-map PCGA but lowers Chest CT AUC. The RADGEN paired cohort interval is broad, whereas the Chest CT interval is negative. Focusing the gate on salient pixels changes which images receive emphasis, but it does not improve both tasks.

The RADGEN comparison gives a different answer. Prediction-confidence weighting has the highest mean AUC and balanced accuracy in Table 1. Adapted e^2KD also exceeds cosine PCGA, and uniform cosine exceeds its stability-weighted counterpart. On the training images, confidence is more closely associated with teacher correctness than map stability is. That association helps explain the RADGEN

Table 1: Test performance (%; mean ± sample SD over seeds 42, 123, and 456). Matched alignment variants share registered ground-truth-class maps; adapted e^{2}KD uses online predicted-class maps. Scale-only uniform and norm-matched confidence match PCGA's auxiliary-gradient norm batchwise at its selected coefficient. Remaining coefficients are validation-selected.

	RADGEN		Chest CT	
Method	AUC	BAcc	AUC	BAcc
ERM	67.38 ± 8.43	54.13 ± 5.74	98.50 ± 0.71	95.25 ± 1.62
Uniform alignment (L1)	69.80 ± 10.20	56.62 ± 7.22	99.13 ± 0.20	95.73 ± 1.03
Stability PCGA (L1)	71.23 ± 2.15	60.97 ± 3.27	98.51 ± 0.18	93.18 ± 1.53
Class-normalized PCGA	61.68 ± 3.21	56.13 ± 5.30	98.81 ± 0.28	94.87 ± 2.18
Prediction-confidence weight	74.50 ± 6.11	64.74 ± 5.56	98.72 ± 0.60	93.73 ± 3.50
Predictive-entropy weight	59.54 ± 5.58	56.13 ± 6.43	98.94 ± 0.27	94.65 ± 2.99
IGIB-style (2022)	62.96 ± 6.05	57.91 ± 1.11	99.16 ± 0.18	95.93 ± 1.35
RRR-style (2017)	63.53 ± 10.13	52.99 ± 3.80	99.09 ± 0.25	94.48 ± 1.62
Energy guidance (2023)	62.39 ± 6.58	57.91 ± 4.63	98.78 ± 0.47	94.81 ± 0.74
Prediction KD	68.09 ± 11.33	51.78 ± 3.27	98.78 ± 0.19	95.81 ± 1.86
Supervised KD	65.38 ± 15.46	60.33 ± 7.33	98.81 ± 0.53	95.06 ± 2.81
e^{2}KD (2024), adapted	72.65 ± 1.54	54.27 ± 3.80	98.78 ± 0.14	94.79 ± 0.56
Uniform alignment (cosine)	67.95 ± 9.77	51.21 ± 2.10	98.90 ± 0.10	93.60 ± 0.46
Scale-only uniform (cosine)	66.81 ± 3.88	56.05 ± 5.55	99.25 ± 0.32	96.06 ± 0.76
Norm-matched confidence (cosine)	67.38 ± 5.69	49.93 ± 0.12	98.83 ± 0.59	94.10 ± 2.31
Stability PCGA (cosine)	64.39 ± 9.38	53.49 ± 3.02	99.19 ± 0.27	95.66 ± 1.34
Salient-region PCGA (cosine)	65.67 ± 7.31	54.84 ± 4.61	99.01 ± 0.19	95.15 ± 2.79

ranking. Patient-bootstrap intervals for confidence versus adapted e^{2}KD and L1 PCGA span zero, leaving those mean differences uncertain at the cohort level.

The independent CIFAR-10 student comparison extends the cosine test beyond CT. Stability weighting reaches 97.67% mean test macro AUC, compared with 97.52% for uniform alignment, 97.59% for the norm-matched uniform control, and 97.37% for confidence weighting. The advantage over both uniform controls is small and appears in only one of three paired seeds. On this task, the higher mean does not establish a consistent benefit from stability-based image ranking. Appendix Table 19 reports accuracy and seed variation alongside AUC.

Changing the alignment objective alters the ranking. With L1 distance and registered ground-truth-class targets, stability weighting exceeds uniform alignment on RADGEN but falls below it on Chest CT. The RADGEN paired seed interval spans zero, whereas every Chest CT seed favors uniform L1 alignment. L1 penalizes pixelwise amplitude differences; cosine alignment emphasizes the pattern of the two flattened maps after normalization. The gate therefore acts on different per-image gradients even when the teacher target is unchanged. Its Chest CT benefit under cosine alignment does not transfer to L1.

A coefficient-matched comparison separates this interaction from validation selection. At $\lambda = 0.3$, raw stability weighting trails uniform L1 alignment on both validation sets. At $\lambda = 0.6$, it leads on both, as reported in Appendix Table 5. The target and gate definitions stay fixed across these candidates. Alignment strength therefore changes the ranking even before the loss type or teacher target changes.

Attribution class and spatial registration. The teacher target class and spatial transform can change which regions the student is asked to match. We compare predicted- and ground-truth-class targets with and without registration before interpreting the weight itself.

In Table 2(a), registering predicted-class teacher maps changes a large negative PCGA–uniform gap on RADGEN into a small positive one. Ground-truth-class targeting with registration also gives a positive mean gap. Chest CT moves in the opposite direction between predicted-class and ground-truth-class targets. We therefore use registered ground-truth-class targets for the main diagnostics, ensuring that the target class and geometry do not change with the weight.

Table 2: Controls for target construction and cohort uncertainty. Values are PCGA minus matched uniform AUC in percentage points. Panel (a) reports seed means. Panel (b) reports paired class-stratified cohort-bootstrap intervals.

(a) Target class and geometry

Class	Geometry	RADGEN	Chest CT
Prediction	Unregistered	-8.97	$+0.55$
	Registered	$+1.99$	-0.05
Ground truth	Unregistered	-0.43	-0.25
	Registered	$+1.42$	-0.62

(b) Cohort bootstrap

Dataset	Gate	ΔAUC [95% interval]
RADGEN	Raw	$+1.42$ [-14.68, 18.66]
	Class-norm.	-8.12 [-20.52, 3.56]
Chest CT	Raw	-0.62 [-1.11, -0.17]
	Class-norm.	-0.32 [-0.81, 0.09]

Weight normalization and replication. The raw gate gives the RADGEN classes different total attribution weights, which could obscure an image-level effect. Class normalization restores a mean weight of one within each class without changing the ranking of its images. AUC nevertheless remains below uniform alignment on both tasks. The within-class ordering alone does not recover a gain in this comparison.

To see whether the initial seeds determine these rankings, we repeat the same configurations with three additional seeds. With frozen ground-truth-class/registered targets, RADGEN has a positive mean PCGA–uniform difference but still includes a negative seed; Chest CT remains slightly negative. With predicted-class/unregistered targets, the differences are positive on RADGEN and near zero on Chest CT. The direction depends on target construction as well as initialization.

Seed variation does not capture uncertainty in the evaluation cohort. Table 2(b) resamples cases while retaining the same trained models. RADGEN patient-bootstrap intervals span substantial gains and losses for both gates. Chest CT retains a negative image-bootstrap interval for raw L1 stability weighting. The observed RADGEN differences are therefore much less precise than the Chest CT result.

5.2 Stability weighting largely preserves the auxiliary-gradient direction

Student AUC alone cannot distinguish a change in auxiliary-gradient magnitude from a change in direction. We compare uniform and stability-weighted gradients on identical batches and model states to separate these effects.

Table 3: Frozen-state gradient medians over 24 matched batches. R/C: RADGEN/Chest CT; I/S: initial/uniform-selected state. $c_1 = \cos(u, v)$, $c_2 = \cos(v, v_g)$, $r = \|v_g\|/\|v\|$, and $r_\lambda = \|\lambda_g v_g\|/\|\lambda_u v\|$.

Task	Obj./state	c_1	c_2	r	r_λ
R	L1/I	-0.102	0.9989	0.912	1.823
	L1/S	0.109	0.9988	0.891	1.782
	Cos./I	0.123	0.9983	0.895	0.298
	Cos./S	0.034	0.9991	0.896	0.299
C	L1/I	0.077	0.9974	0.778	1.556
	L1/S	0.488	0.9979	0.751	1.503
	Cos./I	0.204	0.9975	0.783	0.235
	Cos./S	0.198	0.9983	0.745	0.224

Table 4: Deletion and retention advantages for ground-truth-class maps under an independent evaluator. D/R are probability-point gains over random masks; ρ_D and ρ_R are image-level correlations with the gate.

Dataset	Fill	D	R	ρ_D	ρ_R
Chest CT	blur	11.10	27.84	-0.042	-0.080
	mean	7.14	28.12	0.239	0.014
CIFAR-10	blur	10.53	10.41	0.057	0.112
	mean	1.21	17.78	0.094	0.129
RADGEN	blur	-6.91	5.10	-0.363	0.380
	mean	-5.90	19.00	-0.559	0.743

At the measured states, stability-weighted and uniform auxiliary gradients are nearly collinear (Table 3). The raw gate reduces their norm, especially on Chest CT. Selected L1 coefficients reverse this attenuation, while cosine coefficients retain it. The measured change is mainly one of scale.

Before gating, L1 classification and auxiliary gradients conflict in 83.3% of RADGEN batches versus 8.3% on Chest CT at initialization. At the selected state, the rates are 37.5% and 29.2% (Appendix

Table 15). Gating changes these rates little; most of the difference is present before the weights are applied.

Equation 8 shows that a common multiplier removes the centered-weight term and makes the gradients collinear. The near-one measured cosines motivate the norm-matched training control.

Scale-only cosine control. The uniform control in Equation 13 matches PCGA's auxiliary-gradient norm batchwise and uses the same validation checkpoint rule. On Chest CT, its 99.25% AUC slightly exceeds PCGA's 99.19%; the paired difference is -0.05 percentage points with a 95% image-bootstrap interval of $[-0.20, 0.10]$. On RADGEN, the paired difference is -2.42 points with a patient-bootstrap interval of $[-11.68, 7.26]$, and only one seed favors PCGA. Norm matching removes the Chest CT gain over conventional uniform alignment, as expected from the nearly collinear gradients.

5.3 Response relevance depends on the evaluator

At 10%, 20%, and 40% area, we compare deletion and retention against random masks under mean and blurred fills. The teacher and an independent classifier evaluate ground-truth- and predicted-class maps; Appendix E reports the full combinations.

On RADGEN, deleting selected regions weakens predictions less than random deletion, whereas retaining them strengthens predictions more (Table 4). Under the independent evaluator, stability associates negatively with deletion advantage and positively with retention advantage. These interventions measure necessity and sufficiency, respectively; averaging them would hide the opposing associations.

Chest CT has positive advantages under both fills, but its stability associations change with the fill rule. CIFAR-10 advantages are positive while its correlations remain weak. RADGEN's within-class retention correlations also differ in sign. Region effect and map repeatability are distinct: gate magnitude does not consistently rank useful teacher targets.

6 Conclusion

PCGA uses cross-view map agreement to weight teacher attribution alignment from image-level labels. Gated cosine alignment improves Chest CT AUC over conventional uniform alignment, while prediction confidence leads on RADGEN. CIFAR-10 gives stability weighting a small mean advantage that varies across seeds. Gradient and norm-matched controls attribute the Chest CT gain chiefly to auxiliary strength. Spatial interventions show that repeatability does not uniformly track prediction response. Together, these results define the conditions under which agreement can set alignment strength without treating it as a universal sample-ranking signal. Independent cohorts with reliable teachers provide a further test of its interaction with prediction confidence.

AI Assistance Disclosure

Generative AI assisted with English translation, language polishing, manuscript formatting, code editing, and reviewing recorded experimental outputs. The research questions, experimental design, and final conclusions were determined by the authors.

Ethics Statement

The experiments use public, de-identified datasets and involve no new participant recruitment or clinical intervention. Image-level results on a dataset without patient identifiers should not be used to infer patient-level diagnostic performance. These experiments study attribution supervision and do not establish clinical readiness.

Reproducibility Statement

The method and optimization rules are specified in Section 3; experimental budgets, model selection, and uncertainty definitions appear in Section 4. The appendices report full metrics, coefficient grids, replications, and diagnostic settings. Resampling intervals use paired per-example predictions from the reported test evaluations.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NeurIPS*, volume 31, pp. 9525–9536, 2018. URL https://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps.

Nishanth Arun, Nathan Gaw, Praveer Singh, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol. Artif. Intell.*, 3(6):e200267, 2021. doi: 10.1148/ryai.2021200267. URL https://pubmed.ncbi.nlm.nih.gov/34870212/.

Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, et al. A radiogenomic dataset of non-small cell lung cancer. *Sci. Data*, 5:180202, 2018. doi: 10.1038/sdata.2018.202. URL https://www.nature.com/articles/sdata2018202.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://image-net.org/static_files/papers/imagenet_cvpr09.pdf.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *NeurIPS*, volume 32, pp. 13567–13578, 2019. URL https://papers.nips.cc/paper/9511-explanations-can-be-manipulated-and-geometry-is-to-blame.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *AAAI*, volume 33, pp. 3681–3688, 2019. doi: 10.1609/aaai.v33i01.33013681. URL https://ojs.aaai.org/index.php/AAAI/article/view/4252.

Emilie Grégoire, Muhammad Hafeez Chaudhary, and Sam Verboven. Sample-level weighting for multi-task learning with auxiliary tasks. *Appl. Intell.*, 54(4):3482–3501, 2024. doi: 10.1007/s10489-024-05300-9. URL https://link.springer.com/article/10.1007/s10489-024-05300-9.

Hanning Guo, Farah Abdellatif, Yu Fu, Jon N. Shah, Abigail Morrison, and Jürgen Dammers. XAIguiFormer: Explainable artificial intelligence guided transformer for brain disorder identification. In *ICLR*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/35023b871910f21b03c379dc90764f0f-Paper-Conference.pdf.

Mohamed Hany. Chest CT-Scan images dataset. Kaggle dataset, 2020. URL https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images. Accessed: 2026-03-22.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Anna Hedström, Philine Bommer, Kristoffer K. Wickstrøm, Wojciech Samek, Sebastian Lapuschkin, and Marina M.-C. Höhne. The meta-evaluation problem in explainable AI: Identifying reliable estimators with MetaQuantus. *TMLR*, 2023. URL https://openreview.net/forum?id=j3FK00HyfU.

David E. Hernandez, Jose Ramon Chang, and Torbj"orn E. M. Nordling. Model compression using knowledge distillation with integrated gradients. *Neural Comput. Appl.*, 38:434, 2026. doi: 10.1007/s00521-026-12137-5. URL https://doi.org/10.1007/s00521-026-12137-5.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. URL https://arxiv.org/abs/1412.6980.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Mingzhou Liu, Ching-Wen Lee, Xinwei Sun, Xueqing Yu, Yu Qiao, and Yizhou Wang. Learning causal alignment for reliable disease diagnosis. In *ICLR*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/6ce9d51dded7dac82b3b4d3dbb1d73bc-Paper-Conference.pdf.

Dwarikanath Mahapatra, Alexander Poellinger, and Mauricio Reyes. Interpretability-guided inductive bias for deep learning based medical image. *Med. Image Anal.*, 81:102551, 2022. doi: 10.1016/j.media.2022.102551. URL https://www.sciencedirect.com/science/article/pii/S1361841522001980.

Hadi Moazen, Flavie Lavoie-Cardinal, and Audrey Durand. Interactive learning from explanations with adaptive guidance. In *CanAI*, volume 318 of *PMLR*, pp. 392–407, 2026. URL https://proceedings.mlr.press/v318/moazen26a.html.

Amin Parchami-Araghi, Moritz Böhle, Sukrut Rao, and Bernt Schiele. Good teachers explain: Explanation-enhanced knowledge distillation. In *ECCV*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/09286.pdf.

Sukrut Rao, Moritz Böhle, Amin Parchami-Araghi, and Bernt Schiele. Studying how to efficiently and effectively guide models with explanations. In *ICCV*, pp. 1922–1933, 2023. doi: 10.1109/ICCV51070.2023.00184. URL https://openaccess.thecvf.com/content/ICCV2023/html/Rao_Studying_How_to_Efficiently_and_Effectively_Guide_Models_with_Explanations_ICCV_2023_paper.html.

Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In *ICML*, volume 119 of *PMLR*, pp. 8116–8126, 2020. URL https://proceedings.mlr.press/v119/rieger20a.html.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *IJCAI*, pp. 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. URL https://www.ijcai.org/proceedings/2017/371.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven Q. H. Truong, Chanh D. T. Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G. Blankenberg, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. Benchmarking saliency methods for chest x-ray interpretation. *Nat. Mach. Intell.*, 4(10):867–878, 2022. doi: 10.1038/s42256-022-00536-x. URL https://www.nature.com/articles/s42256-022-00536-x.

Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.*, 2(8):476–486, 2020. doi: 10.1038/s42256-020-0212-3. URL https://www.nature.com/articles/s42256-020-0212-3.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74. URL https://openaccess.thecvf.com/content_ICCV_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html.

Shelley Zixin Shu, Haozhe Luo, Alexander Poellinger, and Mauricio Reyes. Hybrid explanation-guided learning for transformer-based chest x-ray diagnosis. In *iMIMIC 2025*, volume 16464 of *LNCS*, pp. 33–42, 2026. doi: 10.1007/978-3-032-17611-0_4. URL https://link.springer.com/chapter/10.1007/978-3-032-17611-0_4.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I. Inouye, and Pradeep K. Ravikumar. On the (in)fidelity and sensitivity of explanations. In *NeurIPS*, volume 32, pp. 10965–10976, 2019. URL https://papers.nips.cc/paper/9278-on-the-infidelity-and-sensitivity-of-explanations.

Yifei Zhang, Siyi Gu, Yuyang Gao, Bo Pan, Xiaofeng Yang, and Liang Zhao. MAGI: Multi-annotated explanation-guided learning. In *ICCV*, pp. 1977–1987, 2023. doi: 10.1109/ICCV51070.2023.00189. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_MAGI_Multi-Annotated_Explanation-Guided_Learning_ICCV_2023_paper.html.

A Experimental Protocols and Objective Details

The CT ResNet-50 classifier uses a 256-unit sigmoid hidden layer with dropout 0.4 before its output layer. Teacher targets are final-block Grad-CAM maps, and student attribution gradients remain differentiable during alignment. Both maps use min–max normalization with 10^{-8} stabilization before the mean pixelwise absolute error is evaluated on the common spatial grid.

RADGEN training and validation use 136 supplied 1578×903 viewer-export images without lesion cropping. A training-only audit matches 115 of 119 training images to public CT series and identifies 14 displayed window settings. In a six-patient pilot, source-image reconstruction recovers the displayed slices, with masked correlations of 0.9888–0.9971. The supplied images do not establish a general slice-selection rule. All comparisons preserve the original image content and windowing, and RADGEN uncertainty is evaluated by patient-level resampling.

For registered supervision, the same sampled spatial transform is applied to the image and teacher map. Images are padded to an approximately square canvas and resized to 256×256; both images and maps are then cropped to 224×224, rotated within $[-15, 15]$ degrees, and horizontally flipped with probability 0.5. Image rotation uses nearest-neighbor interpolation and map rotation uses bilinear interpolation. The unregistered control resizes maps without matching the image transform. The earlier center-crop target construction is reported separately because it changes the teacher target.

Uniform, raw-stability, class-normalized, confidence, and predictive-entropy matched alignment search coefficients $\{0.1, 0.3, 0.6\}$. On both CT datasets, selected coefficients are 0.3 for uniform and 0.6 for raw stability. Class-normalized coefficients are 0.3 on RADGEN and 0.6 on Chest CT. Confidence selects 0.3 on RADGEN and 0.1 on Chest CT; predictive entropy selects 0.1 and 0.6, respectively. IGIB-style searches $\{0.003, 0.01, 0.03\}$ and selects 0.01 on RADGEN and 0.003 on Chest CT. RRR-style searches $\{0.1, 0.3, 0.6\}$ and selects 0.6 and 0.3. Energy guidance searches $\{0.01, 0.03, 0.1\}$ and selects 0.01 and 0.1. Each objective in this block uses the same three-seed, three-candidate validation budget.

RRR-style penalizes the fraction of absolute ground-truth-logit input-gradient magnitude outside the teacher's top-25% support. Energy guidance penalizes the fraction of positive student Grad-CAM mass outside that support. IGIB-style contrasts ground-truth and strongest-alternative student maps with a total-variation term, using a fixed TV-to-contrast coefficient ratio of $0.015/0.12$.

Baseline objectives. Let $P = HW$ be the number of map pixels, $\epsilon = 10^{-8}$ a numerical stabilizer, and $M_i(p)$ the binary mask selecting the largest $\max(1, \lfloor 0.25P \rfloor)$ pixels of the registered teacher map. For RRR-style training, $G_i(p)$ is the channel mean of $|\partial f_\theta(x_i)_{y_i} / \partial x_i(p)|$. Its auxiliary penalty $\mathcal{R}_R$ is

$$\mathcal{R}_R = \frac{1}{B} \sum_i \frac{\sum_p (1 - M_i(p)) G_i(p)}{\max(\sum_p G_i(p), \epsilon)}. \tag{16}$$

Energy guidance uses the min–max-normalized positive student Grad-CAM S_i and minimizes the objective $\mathcal{L}_E$:

$$\mathcal{L}_E = \mathcal{L}_c - \frac{\lambda}{B} \sum_i \frac{\sum_p M_i(p) S_i(p)}{\sum_p S_i(p) + \epsilon}. \tag{17}$$

For IGIB-style training, the alternative class is $r_i = \arg\max_{k \neq y_i} f_\theta(x_i)_k$. Let C_i^y and C_i^r be the positive, spatially normalized and resized CAMs for the ground-truth and alternative classes, and let

TV denote spatial total variation. The auxiliary penalty $\mathcal{R}_I$ is

$$\mathcal{R}_I = \frac{P}{B} \sum_i \left[\sum_p C_i^y(p) C_i^r(p) + \frac{0.015}{0.12} \operatorname{TV}(C_i^y) \right], \tag{18}$$

$$\operatorname{TV}(C) = \frac{\sum_{h=1}^{H-1} \sum_w |C_{h+1,w} - C_{h,w}|}{(H-1)W} + \frac{\sum_h \sum_{w=1}^{W-1} |C_{h,w+1} - C_{h,w}|}{H(W-1)}. \tag{19}$$

RRR-style minimizes $\mathcal{L}_c + \lambda \mathcal{R}_R$, and IGIB-style minimizes $\mathcal{L}_c + \lambda \mathcal{R}_I$. The TV term uses horizontal and vertical absolute differences.

Table 5: Validation AUC (%) at a shared alignment coefficient, averaged over three best-validation checkpoints per setting.

RADGEN				Chest CT			
λ	Uniform	Stability	Class-norm.	λ	Uniform	Stability	Class-norm.
0.1	64.74 ± 2.22	68.59 ± 12.36	64.10 ± 10.59	0.1	99.56 ± 0.18	99.56 ± 0.16	99.35 ± 0.11
0.3	73.72 ± 2.22	64.74 ± 9.49	68.59 ± 12.80	0.3	99.70 ± 0.22	99.61 ± 0.12	99.55 ± 0.24
0.6	71.79 ± 2.22	73.08 ± 1.92	64.10 ± 10.59	0.6	99.50 ± 0.27	99.85 ± 0.06	99.76 ± 0.15

The common-coefficient comparison in Table 5 holds loss scaling fixed across weighting rules, using the existing validation candidates. These values reflect best-checkpoint selection, whose variability is pronounced on the 17-patient RADGEN validation partition. The gradient-norm-matched comparison uses a shared normalization rule and reports both selected-checkpoint and final-epoch AUC.

Inference boundary. Teacher maps and gates are used only during training. Selected RADGEN students have 24.03 million parameters and median batch-one latency between 5.80 and 5.87 ms on an NVIDIA A40. The corresponding Chest CT range is 4.80–4.86 ms. Peak allocated memory is approximately 117 MB. Attribution alignment affects training cost, while inference uses the same student classifier.

B Full Matched Results

Table 6: RADGEN full matched metrics in percent, three-seed mean $\pm$ sample standard deviation.

Method	AUC	BAcc	Macro-F1
Uniform alignment	69.80 ± 10.20	56.62 ± 7.22	52.85 ± 9.99
Stability PCGA	71.23 ± 2.15	60.97 ± 3.27	62.07 ± 4.22
Class-normalized PCGA	61.68 ± 3.21	56.13 ± 5.30	54.05 ± 9.90
IGIB-style (2022)	62.96 ± 6.05	57.91 ± 1.11	58.25 ± 1.32
RRR-style (2017)	63.53 ± 10.13	52.99 ± 3.80	49.94 ± 7.44
Energy guidance (2023)	62.39 ± 6.58	57.91 ± 4.63	57.75 ± 6.41

Table 7: Chest CT full matched metrics in percent, three-seed mean $\pm$ sample standard deviation.

Method	AUC	BAcc	Macro-F1
Uniform alignment	99.13 ± 0.20	95.73 ± 1.03	94.81 ± 0.74
Stability PCGA	98.51 ± 0.18	93.18 ± 1.53	91.64 ± 0.42
Class-normalized PCGA	98.81 ± 0.28	94.87 ± 2.18	94.74 ± 1.99
IGIB-style (2022)	99.16 ± 0.18	95.93 ± 1.35	95.13 ± 0.78
RRR-style (2017)	99.09 ± 0.25	94.48 ± 1.62	92.98 ± 3.38
Energy guidance (2023)	98.78 ± 0.47	94.81 ± 0.74	93.55 ± 0.87

C Original Gate Comparison

The original center-crop target comparison includes confidence, stability, product, average, shuffled, inverted, and correctness-based weights. Shuffled, inverted, and correctness-based weights serve as controls. The teacher disagrees with 11 of 119 RADGEN training labels and none of the 484 Chest CT labels. Under the alternative resize-only predicted-class targets, the corresponding counts are seven and four, with gate–correctness Spearman correlations of -0.119 and 0.069. The two target constructions are reported separately.

Table 8: Original center-crop target comparison, three-seed test means in percent. Target construction differs from Table 1.

	RADGEN			Chest CT		
Method	AUC	BAcc	F1	AUC	BAcc	F1
ERM	67.38	54.13	50.87	98.50	95.25	94.34
Uniform alignment	70.51	61.04	62.23	98.82	93.55	92.44
Confidence	73.36	59.83	60.01	98.56	92.59	91.25
Stability	56.55	54.20	51.82	98.98	94.52	92.87
Product	70.66	58.55	59.08	98.90	95.98	95.05
Average	68.23	51.85	46.19	98.76	95.12	93.75
Shuffled	63.68	52.85	50.70	99.08	94.85	93.83
Inverted	63.11	54.77	54.19	98.94	94.30	93.26
Correctness	65.38	56.13	54.05	99.08	96.28	95.50

D Fixed-Configuration Replication

Additional seeds vary initialization, minibatch order, and augmentation on the same cohort. Data splits, teacher targets, architecture, and coefficients selected from seeds 42, 123, and 456 remain fixed. Three additional seeds compare uniform and stability weighting under predicted-class/unregistered and ground-truth-class/registered targets.

Table 9: Additional-seed paired test AUC differences, stability minus uniform, in percentage points. Coefficients are inherited rather than reselected.

Dataset	Condition	Seed 17	Seed 43	Seed 2024	Mean
RADGEN	Pred./unreg.	$+6.84$	-11.11	$+17.09$	$+4.27$
	GT/reg.	$+2.14$	$+16.24$	-8.12	$+3.42$
Chest CT	Pred./unreg.	-0.77	$+0.01$	$+0.66$	-0.03
	GT/reg.	-0.32	-0.11	-0.13	-0.18

E Response Diagnostic Details

Scores are averaged within nonoverlapping 16×16-pixel CT blocks or 2×2-pixel CIFAR-10 blocks before ranking. Both grids contain 256 blocks; area fractions of 0.1, 0.2, and 0.4 select 26, 51, and 102 blocks. Ties are randomized with a fixed seed. Mean fill uses each image's per-channel spatial mean. Gaussian blur uses a 31×31 kernel with $\sigma = 8$ pixels for CT and a 7×7 kernel with $\sigma = 2$ for CIFAR-10. Deletion and retention compare top-ranked, bottom-ranked, and three random masks at each area fraction. The CT diagnostic uses 136 RADGEN and 549 Chest CT training/validation images. Bootstrap resampling is performed over images after averaging fractions and random masks. CT validation results are plotted separately from training results.

The CIFAR-10 response diagnostic draws disjoint, class-stratified subsets of 10,000 training, 1,000 validation, and 1,000 diagnostic images from the official training portion. Two ResNet-18 models with a 3×3 first convolution and no initial max pooling are trained from scratch, with seeds 42 and 123. Training uses SGD for 30 epochs, batch size 128, initial learning rate 0.1, momentum 0.9,

weight decay 5×10^{-4}, and cosine learning-rate decay. Best validation accuracies are 78.0% and 82.3%. This response diagnostic does not use the official test set.

Ground-truth and teacher-predicted classes are held fixed during each intervention comparison. The combined response score is the mean of deletion and retention advantages after averaging area fractions. Correlations are computed across images, with 2,000 descriptive bootstrap resamples. Each dataset has eight target/evaluator/fill cells per evaluated split. The complete analysis has 40 cells because the CT datasets each include training and validation, while CIFAR-10 includes one diagnostic partition. The combined-score analysis is summarized separately in Figure 2.

Table 10: Separate intervention means and correlations in all 24 held-out conditions. D and R are probability-point advantages. ρ_D and ρ_R are image-level Spearman correlations with the stability gate. Other denotes the independent evaluator.

Dataset	Target	Evaluator	Fill	D	R	ρ_D	ρ_R
Chest CT	GT	Other	blur	11.10	27.84	-0.042	-0.080
			mean	7.14	28.12	0.239	0.014
		Teacher	blur	12.13	27.85	0.352	-0.278
			mean	-1.10	41.93	0.532	-0.112
	Pred	Other	blur	11.42	24.41	-0.072	-0.163
			mean	9.56	25.22	0.278	-0.113
		Teacher	blur	12.60	29.57	0.360	-0.251
			mean	-2.00	44.69	0.479	-0.003
CIFAR-10	GT	Other	blur	10.53	10.41	0.057	0.112
			mean	1.21	17.78	0.094	0.129
		Teacher	blur	10.15	7.29	0.030	0.125
			mean	0.81	16.13	0.087	0.174
	Pred	Other	blur	11.24	10.61	0.029	0.078
			mean	1.54	19.09	0.078	0.109
		Teacher	blur	12.51	8.75	0.016	0.086
			mean	0.60	18.21	0.102	0.118
RADGEN	GT	Other	blur	-6.91	5.10	-0.363	0.380
			mean	-5.90	19.00	-0.559	0.743
		Teacher	blur	-8.49	14.05	-0.556	0.716
			mean	-8.34	24.63	-0.667	0.777
	Pred	Other	blur	-9.11	8.62	0.277	-0.034
			mean	-9.90	25.08	-0.154	0.723
		Teacher	blur	-9.80	19.17	-0.488	0.556
			mean	-13.32	36.16	-0.627	0.723

Table 11: Within-class validation correlations for ground-truth-class maps under the independent evaluator. Class indices are dataset-encoded labels. n is the number of validation examples in each class.

Dataset	Fill	Class	n	$\rho(g, D)$	$\rho(g, R)$
Chest CT	blur	0	23	-0.041	-0.012
		1	21	-0.142	-0.205
		2	6	0.314	-0.143
		3	15	0.007	-0.089
	mean	0	23	0.232	-0.017
		1	21	-0.043	0.005
		2	6	-0.314	-0.371
		3	15	-0.168	-0.075
RADGEN	blur	0	13	0.335	-0.126
		1	4	-0.800	-0.400
	mean	0	13	-0.082	0.577
		1	4	0.200	-0.800

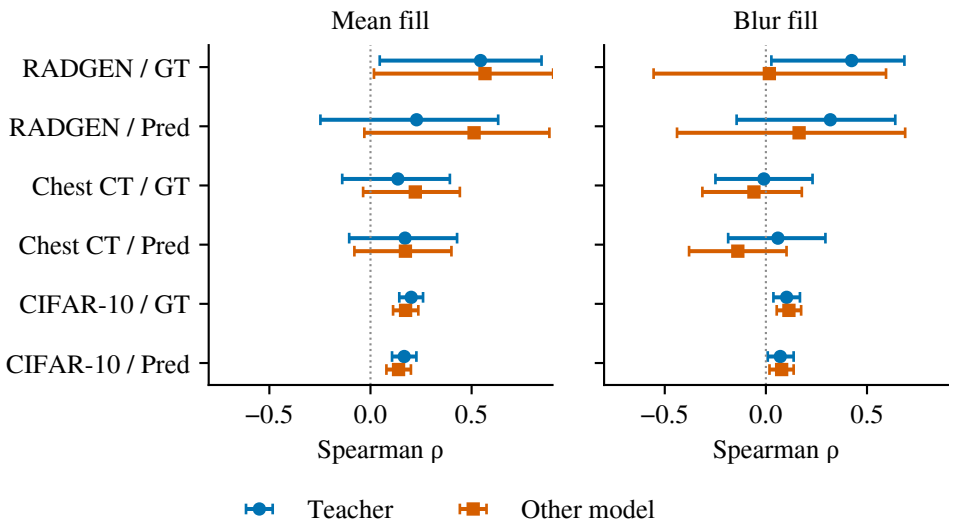

Figure 2: Correlation between stability and the combined deletion/retention advantage. Points denote Spearman correlations and lines denote descriptive 95% sample-bootstrap intervals. CT uses validation images, while CIFAR-10 uses its disjoint diagnostic partition. GT and Pred indicate attribution targets. Each panel changes the fill rule, and each marker changes the response evaluator.

Table 12: RADGEN held-out response diagnostics. D and R are deletion and retention advantages in probability points. Correlations compare stability with the combined advantage. Brackets give descriptive 95% intervals.

Target	Evaluator	Fill	D	R	ρ [95% interval]
GT	Other model	blur	-6.91	5.10	0.017 [-0.555, 0.594]
		mean	-5.90	19.00	0.566 [0.017, 0.903]
	Teacher	blur	-8.49	14.05	0.424 [0.027, 0.685]
		mean	-8.34	24.63	0.544 [0.046, 0.846]
Pred	Other model	blur	-9.11	8.62	0.164 [-0.439, 0.689]
		mean	-9.90	25.08	0.512 [-0.030, 0.885]
	Teacher	blur	-9.80	19.17	0.319 [-0.145, 0.640]
		mean	-13.32	36.16	0.228 [-0.248, 0.631]

Table 13: Chest CT held-out response diagnostics. D and R are deletion and retention advantages in probability points. Correlations compare stability with the combined advantage. Brackets give descriptive 95% intervals.

Target	Evaluator	Fill	D	R	ρ [95% interval]
GT	Other model	blur	11.10	27.84	-0.059 [-0.314, 0.178]
		mean	7.14	28.12	0.222 [-0.036, 0.442]
	Teacher	blur	12.13	27.85	-0.010 [-0.249, 0.231]
		mean	-1.10	41.93	0.135 [-0.140, 0.393]
Pred	Other model	blur	11.42	24.41	-0.139 [-0.380, 0.102]
		mean	9.56	25.22	0.173 [-0.080, 0.400]
	Teacher	blur	12.60	29.57	0.059 [-0.186, 0.294]
		mean	-2.00	44.69	0.171 [-0.105, 0.428]

Table 14: CIFAR-10 held-out response diagnostics. D and R are deletion and retention advantages in probability points. Correlations compare stability with the combined advantage. Brackets give descriptive 95% intervals.

Target	Evaluator	Fill	D	R	ρ [95% interval]
GT	Other model	blur	10.53	10.41	0.114 [0.054, 0.174]
		mean	1.21	17.78	0.173 [0.112, 0.237]
	Teacher	blur	10.15	7.29	0.103 [0.037, 0.168]
		mean	0.81	16.13	0.201 [0.142, 0.260]
Pred	Other model	blur	11.24	10.61	0.078 [0.018, 0.137]
		mean	1.54	19.09	0.139 [0.079, 0.200]
	Teacher	blur	12.51	8.75	0.071 [0.009, 0.137]
		mean	0.60	18.21	0.167 [0.107, 0.227]

F Optimization Diagnostic Details

Derivation of the directional bound. Because $\sum_i (w_i - \bar{w}) = 0$, the centered term in Equation 8 satisfies

$$\frac{1}{B}\sum_i (w_i - \bar{w})v_i = \frac{1}{B}\sum_i (w_i - \bar{w})(v_i - v) = r_w. \tag{20}$$

The vector form of Cauchy–Schwarz yields

$$\|r_w\|_2 \le \left(\frac{1}{B}\sum_i (w_i - \bar{w})^2\right)^{1/2} \left(\frac{1}{B}\sum_i \|v_i - v\|_2^2\right)^{1/2} = \sigma_w \sigma_v. \tag{21}$$

Furthermore, $v_w = \bar{w}v + r_w$. The numerator of $\cos(v, v_w)$ is at least $\|v\|_2(\bar{w}\|v\|_2 - \|r_w\|_2)$, while its denominator is at most $\|v\|_2(\bar{w}\|v\|_2 + \|r_w\|_2)$. Their ratio gives Equation 11 whenever its numerator is positive. The bound holds for any minibatch, including correlated per-image gradients.

The frozen-state probes compare L1 and cosine alignment on matched augmented batches at initialization and at validation-selected uniform-alignment checkpoints. They cover both CT datasets and all three seeds. Model parameters and batch-normalization statistics are held fixed during measurement, so the gradients reflect the compared weighting rules rather than training updates.

For RADGEN selected checkpoints, 14 of 192 probed sample appearances produce constant student maps, a rate of 7.29%. Teacher targets remain nonconstant, and every batch still has a nonzero aggregate auxiliary gradient. The corresponding student-map rate is zero at initialization and in both Chest CT states. These counts concern repeated training-sample appearances across seeds, not 192 independent patients.

Table 15: Frozen-state diagnostics, medians over 24 batches per row. Auxiliary norms are before multiplication by an alignment coefficient. Conflict denotes the percentage of batches with negative classification/auxiliary gradient cosine.

Dataset	State	Weight	$\|v_w\|/\|u\|$	$\cos(u, v_w)$	Conflict (%)
RADGEN	Initial	Uniform	2.766	-0.102	83.3
		Stability	2.529	-0.103	87.5
		Class-norm.	2.785	-0.100	83.3
	Selected	Uniform	0.367	0.109	37.5
		Stability	0.314	0.111	37.5
		Class-norm.	0.352	0.107	37.5
Chest CT	Initial	Uniform	1.775	0.077	8.3
		Stability	1.385	0.075	8.3
		Class-norm.	1.819	0.073	8.3
	Selected	Uniform	2.811	0.488	29.2
		Stability	2.192	0.490	29.2
		Class-norm.	2.897	0.489	29.2

L1 norm-matched validation follow-up. We train both L1 weighting branches with Equation 12, using the same three seeds and 50-epoch budget. Stability weighting has a higher mean best-checkpoint validation AUC on both tasks, but RADGEN seeds disagree in direction and its final-epoch mean reverses the ordering. Selection and reporting use validation data only.

Table 16: Auxiliary-gradient-norm-matched follow-up on validation data only. AUC values are percentages. Best-checkpoint columns report mean ± sample standard deviation across seeds. Differences are stability minus uniform. The final-epoch difference is computed without checkpoint selection.

Dataset	Uniform best	Stability best	Best difference	Final difference
RADGEN	62.82 ± 8.01	64.74 ± 13.91	$+1.92$	-3.21
Chest CT	99.49 ± 0.17	99.70 ± 0.10	$+0.21$	$+1.16$

Table 17: Seed-level validation AUC under auxiliary-gradient norm matching. Best AUC selects among 50 epochs; final AUC is measured at epoch 50. Values are percentages.

Dataset	Weight	Seed	Best epoch	Best AUC	Final AUC
RADGEN	Uniform	42	9	69.23	26.92
		123	8	65.38	59.62
		456	22	53.85	32.69
	Stability	42	30	55.77	50.00
		123	20	80.77	25.00
		456	5	57.69	34.62
Chest CT	Uniform	42	42	99.31	98.89
		123	22	99.49	97.52
		456	19	99.65	98.57
	Stability	42	50	99.81	99.81
		123	48	99.63	99.11
		456	46	99.65	99.56

Adam's coordinate-wise moment estimates can produce different parameter updates even when raw auxiliary-gradient norms match. The comparison isolates the stated norm-control rule rather than exact parameter-update equality.

G Cosine Alignment Follow-up

We replace the spatial L1 distance with $1 - \cos(\mathrm{vec}(S_i), \mathrm{vec}(\widehat{T}_i))$, retaining the classification loss, stability gate, registered teacher maps for the ground-truth class, and training protocol. The validation grid evaluates $\lambda \in \{0.3, 1, 3\}$ for both gated and uniform alignment using seeds 42, 123, and 456. Mean validation AUC selects PCGA $\lambda = 1$ and uniform $\lambda = 3$ on RADGEN, and PCGA $\lambda = 0.3$ and uniform $\lambda = 1$ on Chest CT. RADGEN validation AUC is 76.92% for PCGA and 76.28% for uniform alignment; the corresponding Chest CT values are 99.75% and 99.65%. Each model trains for 50 epochs, with its checkpoint selected by validation AUC. This changes the teacher–student alignment distance; the cross-perturbation stability distance defining g_i remains L1.

The salient-region gate uses the same cosine objective and searches the same coefficient grid. Mean validation AUC selects $\lambda = 3$ on RADGEN (75.00±5.09%) and $\lambda = 0.3$ on Chest CT (99.73±0.05%). Across training samples, the full-map and salient-region gates have Pearson correlations of 0.826 and 0.900, respectively; their means change from 0.883 to 0.780 on RADGEN and from 0.774 to 0.658 on Chest CT. The teacher targets are unchanged, and each selected checkpoint is evaluated once on the test split.

H Comparison with Prediction and Explanation Distillation

We compare cosine PCGA with prediction-only KD, supervised KD, and the explanation-distillation objective of e^{2}KD (Parchami-Araghi et al., 2024). All use the same CT splits, frozen teacher checkpoints, ImageNet-initialized ResNet-50 students, batch size 16, Adam with learning rate 10^{-4} and weight decay 10^{-4}, and 50 training epochs. This comparison adapts e^{2}KD to the CT architecture and training recipe. Each method receives three candidate configurations, selected by mean validation AUC over seeds 42, 123, and 456.

For distillation, let $p_t(x;T)$ and $p_s(x;T)$ be teacher and student class distributions at temperature T, and let KL be Kullback–Leibler divergence. Define the prediction-distillation loss $\mathcal{L}_K = T^2\,\mathrm{KL}(p_t(x;T)\|p_s(x;T))$. Prediction KD minimizes $\mathcal{L}_K$, whereas supervised KD minimizes $\mathcal{L}_c + 0.25\mathcal{L}_K$, both with $T \in \{1,2,5\}$. Let J_t and J_s be vectorized positive Grad-CAM maps for the teacher's top-predicted class, computed online from the same augmented input and resized without min–max normalization. The e^{2}KD objective adds $\beta(1-\cos(J_s,J_t))$ to $\mathcal{L}_K$, with $(T,\beta) \in \{(1,1),(5,1),(5,5)\}$. For cosine stability weighting, we instead retain classification supervision and registered, ground-truth-class teacher maps, searching $\lambda \in \{0.3,1,3\}$ as in Appendix G.

Table 18: Seed-42 validation AUC (%, higher is better) for the initial three-configuration screen. Bold marks the largest AUC in each column. All Table 1 configurations are selected by three-seed mean validation AUC.

Method	RADGEN	Chest CT
Prediction KD	84.62	99.66
Supervised KD	61.54	**99.84**
e^{2}KD (ECCV 2024), adapted	84.62	99.68
Uniform cosine alignment	86.54	99.76
PCGA, cosine alignment	**92.31**	99.71

The initial seed-42 screen gives cosine PCGA 92.31% validation AUC on RADGEN and supervised KD 99.84% on Chest CT, as recorded in Table 18. Final selection uses the complete three-seed grid. RADGEN selects PCGA $\lambda = 1$, uniform $\lambda = 3$, e^{2}KD $(T,\beta) = (5,1)$, prediction KD $T = 5$, and supervised KD $T = 1$. Chest CT selects PCGA $\lambda = 0.3$, uniform $\lambda = 1$, e^{2}KD $(T,\beta) = (5,5)$, prediction KD $T = 2$, and supervised KD $T = 5$. Gated and uniform cosine alignment share the same target and objective. KD comparisons additionally vary label supervision and attribution targets.

At the seed-42 screening checkpoints, RADGEN balanced accuracy/macro-F1 are 50.00/43.33% for prediction KD and e^{2}KD, and 54.81/55.03% for supervised KD. Chest CT values are 85.92/87.38%, 90.58/92.78%, and 91.28/91.82%, respectively. The three-seed test results in Table 1 provide the primary comparison.

I Independent CIFAR-10 Student Comparison

The CIFAR-10 student comparison uses a frozen ResNet-18 teacher trained from scratch on the 10,000-image training subset. Its validation accuracy is 78.0%. The teacher supplies ground-truth-class Grad-CAM targets from three spatially aligned views, and the target is registered to each student's crop and flip. Students are randomly initialized ResNet-18 models trained for 30 epochs with SGD, batch size 128, momentum 0.9, initial learning rate 0.1 with cosine decay, and weight decay 5×10^{-4}. For each aligned method, $\lambda \in \{0.1, 0.3, 1\}$ is selected by mean validation macro one-vs-rest AUC over seeds 42, 123, and 456. All three methods select $\lambda = 1$. The norm-matched uniform control uses the selected PCGA coefficient without another search. Each selected checkpoint is evaluated once on the official test set.

The paired AUC differences for stability weighting minus uniform alignment are -0.022, $+0.475$, and -0.027 percentage points across the three seeds. Against norm-matched uniform alignment they are -0.004, $+0.668$, and -0.434 points. The highest mean therefore does not correspond to a consistent paired-seed advantage.

Table 19: Independent CIFAR-10 student test performance (%, mean ± sample standard deviation over three seeds). Macro one-vs-rest AUC is primary. Alignment methods share the frozen teacher target; the norm-matched control uses PCGA's selected coefficient.

Method	AUC	Accuracy
ERM	97.45 ± 0.44	78.38 ± 1.92
Uniform cosine alignment	97.52 ± 0.21	78.77 ± 1.16
PCGA (cosine alignment)	$\mathbf{97.67 \pm 0.35}$	$\mathbf{79.32 \pm 1.92}$
Confidence-weighted cosine alignment	97.37 ± 0.35	77.92 ± 1.92
Norm-matched uniform cosine alignment	97.59 ± 0.23	78.79 ± 0.92

J Classification-First Timing Control

The timing intervention uses $\lambda_e = 0$ for epochs 1–5, $\lambda_e = 0.3(e-5)/5$ for epochs 6–10, and $\lambda_e = 0.3$ thereafter. Uniform and stability alignment share the same ground-truth targets, registered maps, seeds, splits, optimizer, and 50-epoch budget. Immediate-alignment references use the same final coefficient. Selection and reporting use validation data only.

Table 20: Classification-first timing control. Best validation AUC (%); Δ is delayed minus immediate in points. Both schedules use $\lambda = 0.3$ after alignment begins.

RADGEN				Chest CT			
Weight	Immediate	Delayed	Δ	Weight	Immediate	Delayed	Δ
Uniform	73.72 ± 2.22	61.54 ± 12.01	-12.18	Uniform	99.70 ± 0.22	99.60 ± 0.14	-0.09
Stability	64.74 ± 9.49	64.74 ± 11.27	$+0.00$	Stability	99.61 ± 0.12	99.41 ± 0.40	-0.20

Table 21: Per-seed timing-control validation results. Immediate and delayed columns are best AUC; the epoch and final AUC belong to the delayed schedule. All selected delayed checkpoints occur after alignment begins.

Dataset	Weight	Seed	Immediate	Delayed	Epoch	Final
RADGEN	Uniform	42	75.00	71.15	27	50.00
		123	75.00	65.38	21	25.00
		456	71.15	48.08	21	34.62
	Stability	42	53.85	69.23	10	36.54
		123	69.23	73.08	10	36.54
		456	71.15	51.92	9	21.15
Chest CT	Uniform	42	99.58	99.60	48	99.54
		123	99.57	99.46	20	98.54
		456	99.95	99.75	10	98.33
	Stability	42	99.50	99.00	14	98.74
		123	99.60	99.79	19	99.19
		456	99.74	99.45	13	98.46

RADGEN stability selects epochs 10, 10, and 9. Its paired delayed-minus-immediate best-AUC differences are $+15.38$, $+3.85$, and -19.23 points, giving a zero mean change. By epoch 50, classification losses have fallen to 0.043, 0.054, and 0.073, while validation AUC falls to 36.54%, 36.54%, and 21.15%. This pattern is consistent with deteriorating late-training generalization.

K Normalization and Frozen-Feature Diagnostics

Running statistics. A fixed-weight diagnostic tests whether recalibrating batch-normalization statistics improves validation performance. It compares uniform and stability weighting under

immediate and delayed alignment on both datasets, across three seeds at $\lambda = 0.3$. Batch-normalization statistics are re-estimated from one pass over augmented training images while model parameters remain fixed.

On RADGEN, recalibration changes immediate-alignment mean validation AUC from 73.72% to 64.10% for uniform weighting and from 64.74% to 60.26% for PCGA. With delayed alignment, the corresponding changes are 61.54% to 51.28% and 64.74% to 50.64%. Chest CT changes are small: -0.03, -0.11, $+0.09$, and -0.01 points in the same order. No group improves across all three seeds.

Representation and input field of view. A separate RADGEN probe uses frozen ImageNet ResNet-50 pooled features and a fixed L2 logistic classifier ($C = 1$, balanced class weights). Feature standardization is fitted separately inside each training fold. Five-fold patient-level cross-validation is repeated with split seeds 42, 123, and 456 on the 119 training patients. These are repeated partitions of the same cohort, not independent optimization replications.

The 136 training and validation PNGs are all 1578×903 viewer exports. The original preprocessing pads to a square before resizing. A single label-independent alternative crops the central 903×903 field before applying the same resize and center-crop operations. This increases central-field sampling resolution while also changing field of view; it is not a lesion crop.

The original view yields out-of-fold AUCs of 59.66%, 55.39%, and 60.41%. The square view yields 49.06%, 43.03%, and 56.07%. Fitting each fixed classifier on all training patients gives validation AUCs of 26.92% and 46.15%, respectively. The increase on the 17-patient validation set accompanies lower internal cross-validation AUC in every repeat. These diagnostics use training and validation inputs only.

L Patient-Separated Classifier Diagnostics

Protocol. To examine the weak RADGEN classification signal, we compare four fine-tuning strategies on the 119 training patients. Five stratified outer folds each reserve 20% of their training patients for validation-based checkpoint selection. The held-out outer fold provides the AUC estimate; the official validation and test partitions are unused.

All strategies share the original image transforms, ImageNet initialization, 256-unit classification head, and 50-epoch Adam budget (10^{-4} learning rate and weight decay; batch size 16). Full fine-tuning updates all parameters and batch-normalization (BN) statistics; frozen-BN fine-tuning updates model parameters with fixed BN statistics. Last-stage and head-only tuning update the final residual stage plus classifier and the classifier alone, respectively, with fixed BN statistics. Initialization and sample order are matched within each seed.

Table 22: Training-only RADGEN classifier diagnosis. AUC is in percent, averaged over five outer patient folds. The $\pm$ values are fold standard deviations for inner-selected checkpoints. Seeds 123 and 456 repeat the screened last-stage strategy and full fine-tuning on the same patient partitions.

Strategy	Seed	Inner-selected AUC	Final-epoch AUC
Full fine-tuning	42	54.36 ± 15.58	57.19
Frozen-BN fine-tuning	42	59.36 ± 10.72	57.32
Last-stage tuning	42	63.79 ± 4.91	52.60
Head-only tuning	42	39.62 ± 5.99	46.61
Full fine-tuning	123	64.68 ± 17.99	62.75
Last-stage tuning	123	53.91 ± 3.41	63.30
Full fine-tuning	456	54.42 ± 17.27	58.57
Last-stage tuning	456	56.70 ± 7.60	62.44

Replication. Last-stage tuning improves mean outer AUC by +9.43 points in the seed-42 screen. The screening rule requires a gain of at least two points and improvement in at least four folds. The selected strategy is then compared with full fine-tuning using seeds 123 and 456, giving paired mean

changes of -10.77 and +2.28 points. It fails the prespecified confirmation rule: a positive gain for each new seed and an average gain of at least two points. These repetitions characterize optimization variability on the same patients. The original teacher and attribution comparisons are retained.