PERFORMANCE WITHOUT EVIDENCE: A PLACEBO-CONTROLLED AUDIT OF MEDICAL FOUNDATION MODEL ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In few-label adaptation of medical foundation models, AUC gain is routinely read as evidence that the model has learned the target pathology. We show that this inference is unsafe and argue that standard adapters frequently achieve high performance by covertly exploiting spurious artifacts rather than genuine lesions. Because purely observational metrics cannot expose this shortcut, we introduce a ***Placebo-Controlled Auditing*** framework to isolate true grounding by verifying that predictions degrade when the actual lesion is occluded, yet remain stable otherwise. The protocol first ***measures*** causal reliance with a paired counterfactual intervention that occludes the annotated lesion against a shape- and texture-matched control. It then ***validates*** that measurement with a lesion-free placebo: the identical comparison with no lesion present, which must stay silent for the intervention score to mean lesion reliance. Across eight public task beds and six frozen encoders per bed, standard adaptation raises AUC substantially, yet the trained adapter stays statistically indistinguishable from a prior-matched, gradient-free null in 44 of 48 cells. Because the capacity for grounded transfer cannot be reliably predicted by encoder provenance, it must be explicitly measured and enforced. To address this, we ***act*** via ***Counterfactual-Consistency Adaptation (CCA)***, which translates the validated causal contrast directly into a training constraint. CCA successfully forces reliance on genuine lesions, raising both causal evidence and external AUC on every CXR-domain encoder while outperforming six mask-matched baselines. Causal reliance is measurable, unpredictable, yet trainable once validated—positioning our *measure–validate–act* loop as the reporting standard for label-scarce clinical adaptation. The code is available: ***here***.

1 INTRODUCTION

Medical foundation models transfer effectively to common findings via self-supervised encoders (Pérez-García et al., 2025), report-supervised contrastive towers (Huang et al., 2021; Wu et al., 2023; Tiu et al., 2022), and grounded corpora (Deng et al., 2025). Clinical findings, however, follow a long-tailed distribution in which annotations for the rare tail remain scarce (Lin et al., 2025a). The standard recourse is a few-shot adaptation: fit a small head on $k \sim 5$–100 labels and report the AUC improvement. The improvement is then read as evidence of a learned pathology representation (Shakeri et al., 2024; Huang et al., 2024; Dalsania, 2025).

We argue that this inference is fundamentally unsafe. In the label-scarce regime, spurious correlates—such as site marks, support devices, and co-findings—are cheaper to learn than genuine pathology, and models exploiting them fail under distribution shift (Zech et al., 2018; DeGrave et al., 2021; Glocker et al., 2023). Consequently, standard adapters achieve high performance by covertly exploiting these artifacts. Because purely observational metrics cannot expose this shortcut, true grounding must be isolated through active perturbation. While grounding audits attempt to validate this interpretation, current approaches leave the inference untested in two ways. First, existing evaluations target fixed checkpoints (Saporta et al.; Luo et al., 2026; Gosai et al., 2025; Lian et al., 2026; Zain et al., 2026) rather than the adaptation trajectory that actually produces the reported gain. Second, and more consequentially, the audits themselves remain unvalidated. Saliency alignment is known not to follow out-of-distribution generalization (Viviano et al., 2019), but that

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.

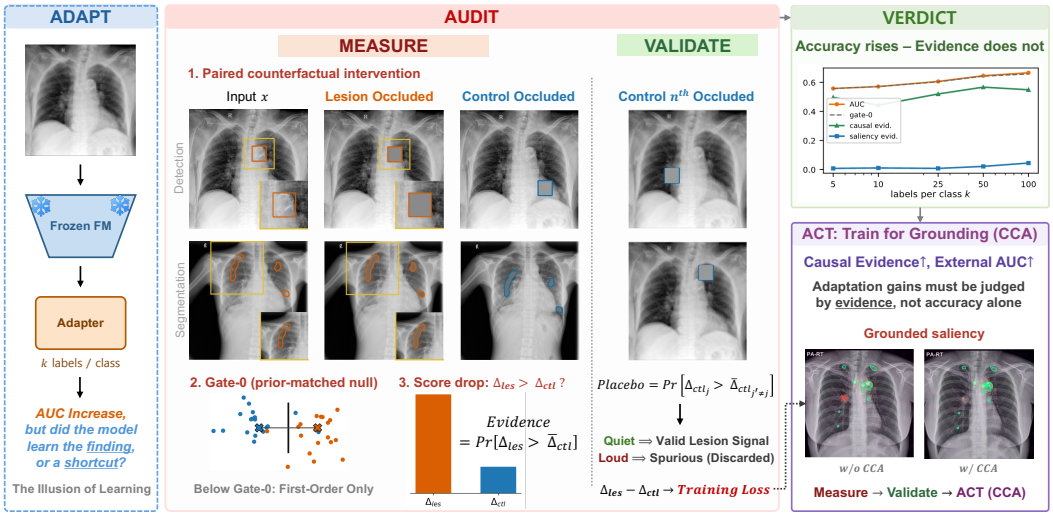


Figure 1: **MEASURE–VALIDATE–ACT. ADAPT**. Few-label adaptation raises AUC but leaves the model's reliance on the pathology unchecked. **AUDIT**. **Measure**: a paired counterfactual intervention occludes the lesion (orange) and a matched control (blue) to quantify causal reliance, and gate-0 supplies a prior-matched null. **Validate**: a placebo runs the identical occlusions with no lesion, filtering spurious artifacts. **VERDICT**. Accuracy improves alongside gate-0, yet grounded evidence stays at chance. **ACT**. CCA embeds the validated contrast into the *training loss*, enforcing grounding and improving external transfer.

result concerns saliency, not causal reliance on the finding. While occlusion-based causal grounding rates exist (Lotfinia et al., 2026), none have been validated against a negative control. Such a sanity check is already standard practice for saliency (Adebayo et al., 2018). Without this control, an occlusion score cannot distinguish genuine reliance on the lesion from a generalized sensitivity to the occlusion operator.

To address this, we introduce a ***Placebo-Controlled Auditing*** framework that isolates true grounding by verifying that predictions degrade when the actual lesion is occluded, yet remain stable otherwise. This intuition drives our proposed *measure–validate–act* loop (Figure 1): First, the protocol **measures** causal reliance on two axes. To claim an adapter learned anything beyond first-order class statistics, it must first surpass a prior-matched, gradient-free null (gate-0). A paired counterfactual intervention then tests if this learned information isolates the actual finding by occluding it against a morphologically and texturally matched control region. Differencing these two performance drops analytically cancels baseline occlusion artifacts. Second, a lesion-free *placebo* **validates** this measurement. By applying the identical comparison where no lesion is present, we test whether the model remains silent. A silent placebo confirms true lesion reliance, while a reactive one exposes spurious spatial confounders. Third, because grounded transfer remains unpredictable, it must be structurally enforced. We **act** via ***Counterfactual-Consistency Adaptation (CCA)***, integrating a margin on the validated contrast into the training loss. This forces the model to depend on the lesion more than its matched neighbor, imposing a strict constraint on counterfactual behavior under intervention. Crucially, the entire audit requires only four cached encoder passes per positive sample. The experimental evaluation spans eight public task beds (including chest radiography and fundus photography), utilizing six frozen encoders per bed.

We bring the following **contributions:**

- ***A Placebo-Controlled Audit framework.*** The protocol *measures* causal reliance via paired counterfactual occlusion and *validates* it against a lesion-free placebo. Extending the rigor of negative controls (Adebayo et al., 2018) to causal grounding, our lesion-free placebo filters out intervention-induced artifacts, keeping only the cells whose score genuinely reflects lesion reliance at zero computational overhead.
- ***Performance without evidence.*** Across eight task beds and six encoders, the inference from classification AUC to causal grounding consistently fails. Despite substantial AUC gains, stan-

dard adapters remain statistically indistinguishable from a prior-matched, gradient-free null in 44 of 48 cells. The placebo explicitly exposes spurious occlusion artifacts that otherwise masquerade as genuine grounding, establishing that grounded transfer cannot be assumed but must be explicitly measured.

- ***Counterfactual-Consistency Adaptation (CCA).*** The audited quantity is inherently actionable. CCA translates the validated contrast directly into a training constraint, successfully forcing reliance on genuine lesions. This simultaneously raises causal evidence and external AUC on every CXR-domain model, outperforming six mask-matched baselines on average.

Positioning. Existing audits evaluate static checkpoints (Saporta et al.; Luo et al., 2026; Zain et al., 2026); we audit the adaptation trajectory itself. Prior causal occlusion methods (Lotfinia et al., 2026) lack our lesion-free placebo validation. Unlike region-supervised methods requiring corpus-scale masks (Deng et al., 2025), penalizing static saliency (Ross et al., 2017; Li et al., 2022; Rieger et al., 2020; Erion et al., 2021), or relying on generative editors (Cooke et al., 2026), CCA directly constrains counterfactual output against a matched control, requiring masks only for the k training positives. See Appendix J for a full comparison.

2 The audit protocol: measure and validate

This section formalizes the AUDIT panel in Figure 1: ***Measure*** (Section 2.1) and ***Validate*** (Section 2.2). While prior work introduces causal grounding (Lotfinia et al., 2026), the proposed placebo validation provides a novel negative control. This ensures downstream findings (Section 3) reflect true lesion dependence rather than generic occlusion artifacts.

Task bed. The primary test bed evaluates VinDr-CXR calcification classification (Nguyen et al., 2022). We utilize a frozen data split containing $130^+/400^-$ test and $200^+/602^-$ train images, drawing $k \in \{5, 10, 25, 50, 100\}$ labels per class across 20 seeds. Lesion masks are the union of three independent radiologists' annotations on the consensus test set. Negatives contain other findings rather than being healthy, preventing an any-abnormality shortcut. Transfer evaluation spans PadChest-GR (Bustos et al., 2020; De Castro et al., 2025) ($64^+/3147^-$) and ChestX-Det10 (Lian et al., 2021) ($201^+/3342^-$), with label provenance detailed in Appendix B.3.

Encoders and statistics. The evaluation spans six frozen encoders: RAD-DINO (Pérez-García et al., 2025), MedSigLIP (Sellergren et al., 2025), and XrayCLIP-B/16 and -L/14 (Chen et al., 2024) as CXR-domain models, with DINOv2-S (Oquab et al., 2023) and an ImageNet ResNet-50 (He et al., 2016) as natural-image controls, at three adaptation depths (linear probe, attentive pooling, LoRA). Because shared data pools across seeds inflate naive run-level p-values, all reported significances use dependence-aware seed-block permutations or hierarchical bootstraps (Appendix F.1).

2.1 Measure

Prior-matched control (gate-0). For class centroids μ_k^+ and μ_k^- of labeled features $f(x)$, gate-0 scores inputs via cosine similarity $x \mapsto \cos(f(x), \mu_k^+)$, while gate-0b uses the centroid difference $f(x)^\top(\mu_k^+ - \mu_k^-)$. Though nearest-class-mean rules serve as standard few-shot baselines (Snell et al., 2017; Wang et al., 2019; Zhang et al., 2022; Huang et al., 2024; Shakeri et al., 2024), this protocol reframes them as a strict audit precondition. Under isotropic Gaussian assumptions, gate-0b constitutes the population Bayes rule (Proposition 2, Appendix B.1). Outperforming it is therefore a mathematical necessity, not a mere target. Failing to beat gate-0 implies the adapter extracts no novel features beyond first-order class statistics, regardless of its absolute AUC.

Paired counterfactual intervention. The protocol measures causal reliance by occluding a target region R with a local ring-mean fill and calculating the resulting score drop:

$$x_{\setminus R}(u) = \begin{cases} x(u), & u \notin R \\ \bar{x}_{\partial R}, & u \in R \end{cases} \qquad \Delta(R) \;=\; s(x) - s(x_{\setminus R}) \;=\; w^\top \underbrace{\big(f(x) - f(x_{\setminus R})\big)}_{\delta(R)}, \tag{1}$$

where $\partial R = \mathrm{dil}_{15}(R) \setminus R$ defines the dilation ring and $\bar{x}_{\partial R}$ represents its mean intensity. The right-hand identity holds exactly for a linear head $s(x) = w^\top f(x) + b$ because the bias mathematically cancels. Consequently, the feature difference $\delta(R)$ is cached once per region, reducing every subsequent audit to a computationally trivial dot product. For each annotated test positive, the protocol

Table 1: **Placebo control on representative public cells** (complete classification in Table 10). Values reflect partial Spearman correlations of each signal against external AUC, controlling for internal AUC and $\log k$, accompanied by seed-block permutation p-values. A cell qualifies as *grounded* when the placebo is not significantly positive but the gap is; *quiet* when neither clears the significance bar; and *loud* when the placebo itself predicts transfer, mandating the cell's discard.

Bed / encoder	Site	evidence	placebo (p)	gap (p)	verdict
Nodule / MedSigLIP	CheXpert	$+0.390$	-0.192 (0.015)	$+\mathbf{0.430}$ (0.000)	**grounded**
	NIH	$+0.379$	-0.089 (0.225)	$+\mathbf{0.422}$ (0.006)	**grounded**
	VinDr	$+0.503$	-0.259 (0.000)	$+\mathbf{0.552}$ (0.000)	**grounded**
Calcification / XrayCLIP-B/16	ChestX-Det10	-0.149	$+0.170$ (0.082)	$-\mathbf{0.191}$ (0.001)	quiet
	PadChest-GR	$+0.371$	-0.116 (0.261)	$+\mathbf{0.432}$ (0.000)	**grounded**
Calcification / RAD-DINO	ChestX-Det10	$+0.092$	-0.051 (0.793)	$+\mathbf{0.100}$ (0.254)	quiet
	PadChest-GR	$+0.226$	$+0.438$ (0.000)	$+\mathbf{0.140}$ (0.061)	*loud, discarded*
Calcification / DINOv2-S	ChestX-Det10	-0.404	$+0.515$ (0.000)	$-\mathbf{0.453}$ (0.000)	*loud, discarded*
	PadChest-GR	$+0.023$	-0.237 (0.004)	$+\mathbf{0.048}$ (0.490)	quiet

applies Eq. (1) to the lesion and to three disjoint, shape- and intensity-matched control regions (sampling details in Appendix B.2). The grounded evidence score represents the empirical win-rate of $\Delta_{les} > \overline{\Delta}_{ctl}$, treating 0.5 as a descriptive reference rather than an exact null (Appendix D.4). Without such a paired control, generalized sensitivity to any arbitrary occlusion would artificially inflate the evidence score. This paired design prevents such inflation by analytically canceling region-generic artifacts.

Proposition 1 (Cancellation of region-generic artifacts)**.** *Suppose the score change from occluding a region R decomposes as $\Delta(R) = \beta\,\mathbf{1}[R \supseteq \textit{lesion}] + a(R)$, where β denotes the per-image lesion-specific effect and the artifact term $a(R)$ depends exclusively on region shape and local texture statistics. If the controls match the shape and texture statistics of the lesion region, and the disjointness constraint (Eq. (4)) ensures $\mathbf{1}[R_{ctl} \supseteq \textit{lesion}] = 0$, then $\mathbb{E}[\Delta_{les} - \overline{\Delta}_{ctl}] = \beta$. Thus, the paired contrast provides an unbiased estimate of the lesion-specific effect, exactly canceling any occlusion-generic component.*

While the formal proof is straightforward (Appendix B.1), the underlying assumptions warrant scrutiny. The current implementation matches shape exactly but texture only to the first order. Consequently, the cancellation remains mathematically exact in theory but approximate in practice. Appendix D.5 empirically quantifies this residual, and Appendix D.7 confirms that the audit's verdicts remain robust across three distinct occlusion operators.

Grounded saliency. We evaluate intrinsic evidence maps against annotations using a detection harness, scoring each of the N patch tokens by $w^\top f_p$ (or by its attention weight for attentive heads). Metrics include sensitivity at a fixed false-positive count per image (FPPI) and peak-in-GT, with the latter's chance level tracking the lesion area fraction (≈ 0.006). For a linear head on mean-pooled features, $\partial s/\partial f_p = w/N$ is constant in p. Grad-CAM and zero-baseline integrated gradients therefore reduce to $w^\top f_p$ up to a positive factor and rank patches exactly as the intrinsic map does. We report the intrinsic map for linear adapters, and post-hoc attributions only for the attentive and LoRA tiers, where the maps diverge (Appendix C.6). Appendix B.2 extends the instrument across three annotation formats and a second modality where the target area is more than an order of magnitude smaller (Figure 5).

2.2 VALIDATE

To guarantee specificity, the protocol introduces a placebo control that isolates spurious regional effects by comparing regions without a true lesion.

Placebo control. Designating control j as the pseudo-treatment yields:

$$\text{evidence} = \mathbb{E}_i\big[\mathbf{1}[\Delta^{(i)}_{les} > \overline{\Delta}^{(i)}_{ctl}]\big], \qquad \text{placebo}_j = \mathbb{E}_i\big[\mathbf{1}[\Delta^{(i)}_{ctl_j} > \overline{\Delta}^{(i)}_{ctl_{j'\neq j}}]\big],$$
$$\text{gap} = \text{evidence} - \overline{\text{placebo}} \tag{2}$$

where $\mathbb{E}_i$ averages over annotated test positives and $\overline{\text{placebo}} = \frac{1}{2}(\text{placebo}_1 + \text{placebo}_2)$. Because the best-matched region (control 0) is non-exchangeable, the protocol averages only $j \in \{1, 2\}$

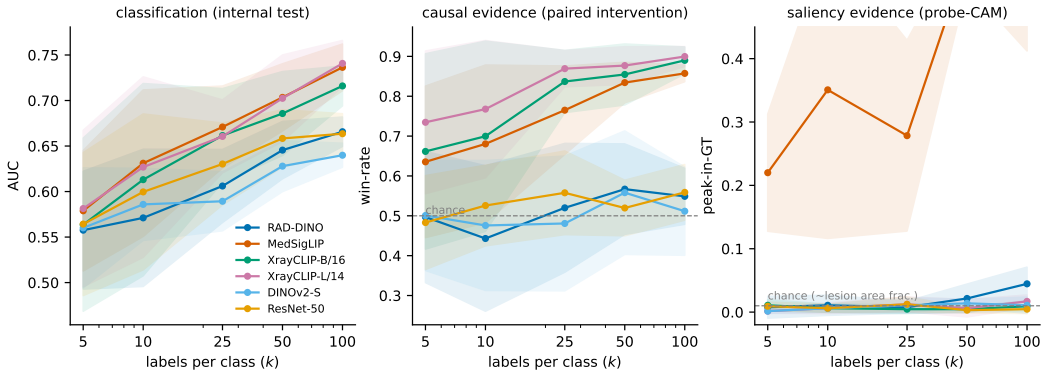


Figure 2: **Dissociation of performance and grounded evidence.** As the label budget (k) increases on the primary bed, AUC consistently improves across all encoders (left). However, grounded evidence diverges unpredictably. RAD-DINO improves AUC without relying on the target lesion (flat causal and saliency channels). Furthermore, the evidence channels conflict: XrayCLIP exhibits strong causal reliance but chance-level saliency, whereas MedSigLIP localizes well but shows moderate causal reliance. This severe dissociation underscores the necessity of a multi-channel audit.

(Appendix D.1). Lacking a lesion, the identically structured placebo tests only whether the model reacts to arbitrary occlusion. A placebo predicting transfer as effectively as the evidence score indicates a failure to isolate genuine lesion reliance. Table 1 reports partial Spearman correlations against external AUC; the gap column reflects the differenced signal's correlation, not the arithmetic difference of individual correlations. Comparing Δ_{les} against three controls precludes an analytical 0.5 baseline (Appendix D.4). The placebo instead supplies the empirical null, reaching 0.533 (RAD-DINO), 0.506 (MedSigLIP), and 0.504 (XrayCLIP-L/14) on the primary bed. On adapters trained with permuted labels, evidence and placebo scores perfectly coincide across all encoders (maximum $|\text{gap}| = 0.019$; Appendix D.2), confirming the instrument fabricates no artificial grounding.

Interpretability rule. The protocol replaces subjective visual inspection with a formal decision rule.

Definition 1 (Reportable cell)**.** *Let c be a (bed, encoder, site) cell carrying partial correlations against external AUC with statistics (r_{plc}, p_{plc}) for the placebo and (r_{gap}, p_{gap}) for the gap, computed via seed-block permutation at level $\alpha = 0.05$. Then*

$$\textsc{Quiet}(c) \iff \big(p_{plc} \geq \alpha\big) \vee \big(r_{plc} \leq 0\big), \qquad \textsc{Report}(c) \iff \textsc{Quiet}(c) \wedge \big(p_{gap} < \alpha\big) \wedge \big(r_{gap} > 0\big).$$

A cell satisfying $\neg\textsc{Quiet}(c)$ is deemed placebo-loud *and discarded.*

Across 36 evaluated cells, this rule identifies 30 quiet and 6 loud cells. Exactly 6 quiet cells exhibit a significant positive gap, qualifying as reportable. Applying a Benjamini–Hochberg correction (Benjamini & Hochberg, 1995) preserves all six reportable cells while reducing the loud count to four. The evaluation conservatively discards all six uncorrected loud cells. Consequently, multiple-testing correction rescues cells from discard without artificially inflating positive evidence. Table 1 details representative cells under Definition 1, with the complete classification deferred to Table 10.

Texture-match diagnostic. Calcification lesions carry higher edge energy than their paired controls ($d = +0.80$; Appendix D.5), so absolute win-rates are instrument-relative; every comparison we draw differences identical fixed regions across adapters, which neutralizes that baseline. Redrawing the controls under varying pool sizes, matching criteria, and placement constraints preserves the partition of trustworthy cells (Appendix D.6), and Appendix B.2 details the cached-region protocol used on the seven further beds of Section 3.

3 Performance without evidence

Accuracy rises, but grounded evidence does not. Figure 2 reads the three channels against the label budget on the primary calcification bed. RAD-DINO improves from an AUC of 0.558 at $k = 5$ to 0.666 at $k = 100$. However, neither evidence channel follows. The intrinsic map's peak-in-GT

Table 2: **Encoder axis at** k=100 on the primary bed (VinDr-CXR calcification), averaged over 20 seeds. Rows below the dividing line represent natural-image controls.

Encoder	AUC	gate-0b	Interv. win-rate	CAM loc	External AUC PadChest-GR	External AUC ChestX-Det10
RAD-DINO	0.666	0.657	0.549	0.045	0.606	0.530
MedSigLIP	0.736	0.736	0.857	0.495	0.607	0.650
XrayCLIP-B/16	0.716	0.714	0.890	0.009	0.605	0.673
XrayCLIP-L/14	0.741	0.730	0.900	0.017	0.675	0.696
DINOv2-S	0.640	0.639	0.512	0.011	0.588	0.525
ResNet-50	0.664	0.663	0.559	0.005	0.593	0.531

reaches only 0.045 — above the 0.006 area-fraction line, yet an order of magnitude below the one map that does localize (0.495, below). The intervention win-rate shifts from 0.50 to 0.55, tracking the empirical null of 0.533. An adapter yielding a $+0.11$ AUC increase demonstrates no measurable reliance on the targeted finding (Table 2). This dissociation holds across three beds spanning box annotations, pixel masks, and a second modality (Figure 19, Appendix H.5). The two channels also disagree with each other: XrayCLIP-L/14 reaches a win-rate of 0.900 while its saliency stays at chance (0.017), and MedSigLIP is the only encoder whose map localizes the lesion (0.495). Static saliency and causal reliance are therefore not interchangeable, and any grounding claim must say which one it means.

Standard-recipe adapters are gate-0-equivalent. On the primary bed, the trained probe exceeds the centroid-difference rule by merely 0.000 to $+0.011$ AUC across all six encoders, with MedSigLIP matching the baseline to four decimal places (Table 2; per-cell intervals in Appendix H.5, Table 32). Pooled across eight public beds (8×6 cells; substituting RETFound for fundus imaging), statistical equivalence at a 0.03 margin holds in 44 cells and persists across all label budgets up to $k = 200$ (Appendix C.2). Stronger prior-matched nulls preserve this outcome: a closed-form shrinkage LDA rule outperforms the probe on the two strongest encoders, while multi-prototype variants trail the single centroid. Only four cells break equivalence. Two occur on chest beds for RAD-DINO, and the audit flags both surpluses as ungrounded. On effusion a $+0.029$ gain comes with a chance win-rate of 0.53; on MS-CXR pneumothorax a $+0.049$ gain comes with a win-rate of 0.45, below the encoder's own placebo null. The final two exceptions involve RAD-DINO and DINOv2-S on the fundus bed, representing extreme out-of-domain evaluation.

Qualifications and architectural limits. The equivalence is a property of the default literature recipe (`sklearn` logistic regression, $C = 1$), which gate-0 matches in supervision. A cross-validated probe does extract up to $+0.054$ AUC over gate-0b, but the audit reads that surplus as incoherent, with win-rate shifts from -0.105 to $+0.205$ (Appendix C.1). Depth does not close the gap either: at $k = 100$ the linear probe (0.666 AUC) beats attentive pooling (0.593) and LoRA (Hu et al., 2021) (0.647), which also localizes worst (0.009 against 0.045), consistent with frozen embeddings discarding small-lesion signal (Muthyala et al., 2026). Neither partial unfreezing nor higher resolution, a shifted operating point, or an in-context segmenter moves the deficit, so the bottleneck is the frozen representation itself (Appendices C.4, C.3, C.5, E.2, E.3, E.4).

The instrument verifies its own lesion-specificity. When the evidence signal is genuine, the placebo stays silent. On the nodule bed with MedSigLIP, evidence predicts external AUC ($+0.379$ to $+0.503$) while the placebo remains negative at every site, so the gap exceeds the raw score (Table 1). Conversely, the placebo exposes spurious correlations, such as DINOv2-S on calcification ($r_{\text{plc}} = +0.515$), where features predicting transfer activate without a lesion. The protocol discards such placebo-loud cells outright. Subtracting an uninformative placebo degrades the signal; the gap improves upon raw evidence only when the placebo correlates non-positively with transfer (Proposition 3, Appendix D.3). This condition holds unconditionally for the 20 surviving cells, whereas the six discarded cells are precisely those where the gap fails. Extending standard negative controls (Adebayo et al., 2018), this protocol evaluates causal grounding at zero computational overhead by reusing cached occlusion deltas.

Domain pretraining does not dictate scope conditions. The applicability of the evidence–transfer link (the *scope condition*) must be predictable. A prevalent hypothesis assumes this requires domain-appropriate representations (Azizi et al., 2023; Moor et al., 2023; Sowrirajan et al., 2021). Evalu-

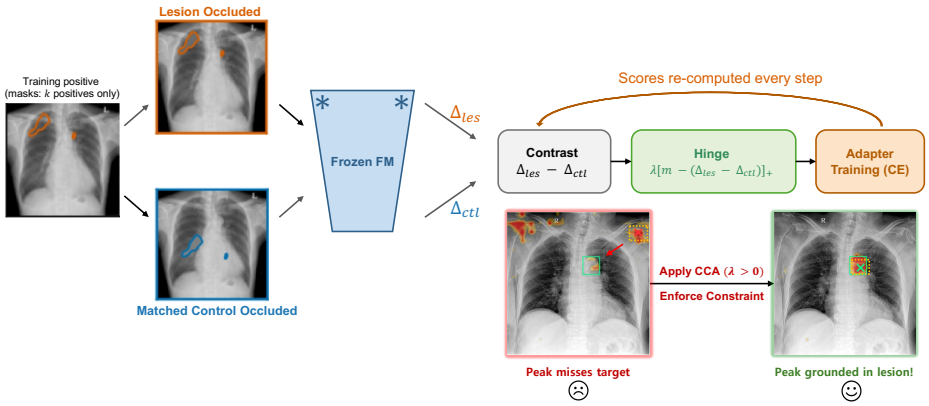


Figure 3: **Counterfactual-consistency adaptation (CCA). Top.** The paired intervention is computed once per training positive and cached, so training adds no new encoder passes; the lesion drop Δ_{les} and matched-control drop Δ_{ctl} are differenced and the contrast enters a hinge added to the classification loss. Only the k training positives require masks. **Bottom.** A MedSigLIP test positive at k=25: without the constraint (λ=0) the evidence peak misses the target (×); under CCA (λ>0) it moves inside the annotation (✓). Overlays shade the top 20% of evidence; the gt box is green, and the evidence peak is marked by an orange dotted box.

ating 102 cells (Table 20, Appendix F.1) does not support this: domain encoders yield a significant positive link in 35% of cells (mean $r = +0.129$), statistically indistinguishable from natural-image controls (38%, mean $r = +0.146$). This parity persists under Benjamini–Hochberg correction (31% vs 32%) and encoder-level aggregation ($p = 0.93$; Appendix F.1). The confidence interval ($[-0.228, +0.170]$) excludes any effect large enough to justify encoder provenance as a scope condition; indeed, the strongest correlations ($+0.65$, $+0.58$) belong to ResNet-50. Alternative hypotheses similarly fail: lesion focality lacks significance ($p = 0.21$), and taxonomy alignment produces sign reversals (Appendix F.4). Because no universal rule emerges, applicability must be empirically verified for each (finding, encoder) pair. This heterogeneity perfectly aligns with the protocol's core claims, which never pool directional effects across cells.

The dissociation replicates across eight public task beds. The observed dissociation is not an artifact of the primary bed. Beyond the budget curves shown for calcification (Figure 2) and two additional annotation formats (Figure 19), the remaining five beds behave the same way across all encoders (Tables 31, 33. The two pneumothorax beds are evaluated on CheXpert (Irvin et al., 2019) and NIH ChestX-ray14 (Wang et al., 2017), and the fundus bed substitutes RETFound (Zhou et al., 2023) for XrayCLIP-L/14 (Appendix H.1). Gate-0 equivalence and evidence dissociation hold on every evaluated bed. The only nuanced variation emerges around lesion scale: MedSigLIP successfully localizes larger findings (peak-in-GT 0.45–0.64), whereas RAD-DINO remains near chance (0.02–0.13) despite reaching AUCs up to 0.83. On effusion, RAD-DINO's causal evidence also stagnates at chance (0.53 at 0.81 AUC). Comprehensive per-budget metrics and qualitative decoupling cases appear in Appendices H.4 and H.3.

4 From Audit to Training Constraint

While Sections 3 and 2.2 establish that grounding and accuracy dissociate, they do not demonstrate whether enforcing grounding directly improves transfer. Counterfactual-consistency adaptation (CCA) evaluates this intervention directly. Because the scope condition lacks *a priori* predictability, CCA serves as an empirical test of whether the audited quantity is causally actionable, rather than a universal performance booster.

4.1 Counterfactual-consistency adaptation (CCA)

For each of the k annotated positives, CCA reuses the audit's cached deltas (Eq. (1)) and imposes a margin on the target contrast:

$$\mathcal{L}(w) = \mathrm{CE}(w) + \frac{\lambda}{k}\sum_{i=1}^{k}\Big[m - w^{\top}\big(\delta_{les}^{(i)} - \bar{\delta}_{ctl}^{(i)}\big)\Big]_{+}, \qquad \bar{\delta}_{ctl}^{(i)} = \frac{1}{|\mathcal{J}_i|}\sum_{j\in\mathcal{J}_i}\delta_{ctl_j}^{(i)}, \tag{3}$$

Table 3: **CCA at a single fixed** λ=1 on the primary VinDr-CXR calcification bed, compared to an optimizer-matched baseline (λ=0). Values report the mean across budgets and seeds, with external AUC averaged over ChestX-Det10 and PadChest-GR. Δext is the paired difference (95% seed-block bootstrap interval); ** indicates $p < 0.001$ (sign-permutation test). † sites denotes the number of significantly positive external sites out of the total evaluated for each encoder (see Table 23 for the full breakdown).

	internal AUC		interv. win-rate		external AUC			
Encoder	λ=0	λ=1	λ=0	λ=1	λ=0	λ=1	Δext [95% CI]	sites†
RAD-DINO	0.613	0.622	0.543	0.747	0.554	0.570	$+\mathbf{0.016}^{**}$ [+0.013, +0.019]	5/22
MedSigLIP	0.664	0.686	0.753	0.883	0.594	0.640	$+\mathbf{0.046}^{**}$ [+0.040, +0.052]	15/22
XrayCLIP-B/16	0.647	0.658	0.780	0.889	0.595	0.625	$+\mathbf{0.030}^{**}$ [+0.023, +0.036]	3/5
XrayCLIP-L/14	0.677	0.702	0.836	0.912	0.641	0.675	$+\mathbf{0.034}^{**}$ [+0.029, +0.038]	2/2
DINOv2-S	0.600	0.607	0.543	0.739	0.537	0.530	$-\mathbf{0.008}^{**}$ [-0.010, -0.005]	3/5
ResNet-50	0.624	0.615	0.558	0.691	0.546	0.539	$-\mathbf{0.007}^{**}$ [-0.010, -0.004]	3/5

where $\bar{\delta}_{ctl}^{(i)}$ averages all controls for positive i, and the margin remains fixed at m=0.2. Sweeping m across a 20-fold range leaves all reported signs unchanged (Appendix G.4; λ regimes in Appendix G.5). Figure 3 illustrates the pipeline and its effect on individual cases. For frozen linear tiers, cached δ values eliminate per-step encoder passes, requiring only four passes per annotated training positive upfront (Table 5). The hinge preserves convexity in w.

With $c_i = \delta_{les}^{(i)} - \bar{\delta}_{ctl}^{(i)}$, each annotated positive contributes a linear inequality $w^\top c_i \geq m$. The audit consequently carves a convex polyhedron out of adapter space, pulling the cross-entropy objective toward it (Eq. (3)). As λ grows, the hinge approaches a hard constraint. Cells positioned far from the cross-entropy optimum must trade accuracy for grounding. This mechanism aligns with CCA's gains peaking at λ=1 without collapsing at λ=3, whereas the strongest baselines degrade monotonically (Appendix G.2). The core of this mechanism lies in the subtraction of Δ_{ctl}. Constraining Δ_{les} alone permits global sensitivity to arbitrary occlusion. The matched control cancels this generic artifact, forcing the loss to decrease only when predictions depend more heavily on the lesion than on a shape-matched neighbor. This constraint formally enforces the exact cancellation property validated by the placebo in Section 2.2.

CCA differs structurally from prior region-supervised training (Appendix J). First, unlike grounded-VLM pretraining (Deng et al., 2025; Lin et al., 2025b), these masks do not act as dense localization targets and apply only to the k training positives. Second, unlike explanation-constrained training (Ross et al., 2017; Li et al., 2022), CCA constrains counterfactual *behavior* rather than a static saliency *proxy*. Under identical configurations, an RRR/SGAT-style head penalizing positive evidence outside the k masks barely moves the win-rate or external AUC, whereas CCA moves both on all four CXR-domain encoders (Table 35). Aligning a saliency map does not guarantee causal reliance; constraining the counterfactual output enforces it. Third, unlike counterfactual augmentation (Cooke et al., 2026), CCA requires no generative editor, as the paired control cancels the artifact in expectation (Proposition 1; Appendix B.1).

Supervision cost and mask-matched baselines. CCA requires pixel- or box-level annotations for the k training positives, demanding a larger budget than image-level labels alone. The evaluation benchmarks CCA against six mask-matched baselines utilizing identical annotations (Table 25). At $\lambda = 1$, CCA achieves the highest mean external gain and the most stable worst-case performance ($+0.018$ versus -0.009 to $+0.003$). While a variant dropping the matched control beats CCA on XrayCLIP-L/14, it fluctuates by 0.085 across encoders. The matched control buys stability rather than a universal advantage. The strongest explanation-constrained baselines, CDEP (Rieger et al., 2020) and the attribution prior (Erion et al., 2021), remain net negative at $\lambda = 1$. Even with an oracle λ selected on evaluation sites, they trail CCA's fixed setting and win only on RAD-DINO, an encoder operating at chance grounding (Appendix G.2). Because several baselines artificially inflate the win-rate while degrading external AUC, the grounding score serves primarily as a diagnostic instrument rather than an optimization target (Appendices G.1, G.3).

Results and site generalization. The audit accurately predicts where the constraint succeeds: encoders possessing measurable latent grounding capacity. Evaluated at a single fixed $\lambda = 1$ applied uniformly across all cells (without per-site selection), the constraint raises the causal win-rate across every encoder and boosts external AUC on all four CXR-domain encoders (Table 3). Figure 4 details

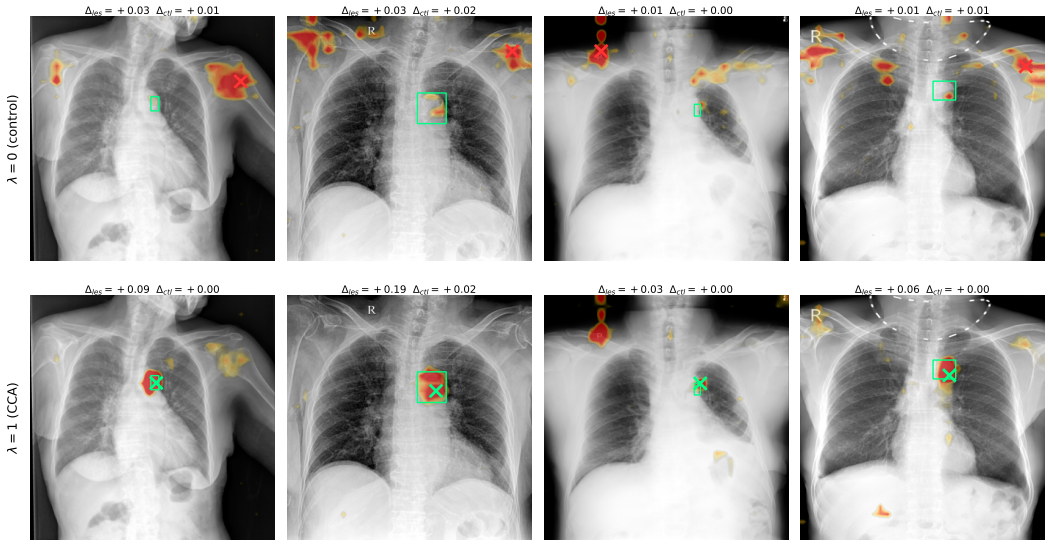


Figure 4: **CCA forces the evidence peak into the lesion** (MedSigLIP, VinDr-CXR calcification, k=25; matched seed and optimizer, differing only by λ). **Top.** Under the λ=0 control the peak misses the annotation (×). **Bottom.** CCA (λ=1) shifts the peak inside (✓), with measured Δ_{les} and $\overline{\Delta}_{ctl}$ inset per panel. These selected columns illustrate the flipping mechanism; globally across all 130 test positives, localization succeeds in 29 cases at λ=0 versus 78 at λ=1.

this shift: on MedSigLIP at k=25, enforcing CCA drives the evidence peak inside the annotation for 78 of 130 test positives (versus 29 under the λ=0 control) and raises the intervention win-rate from 0.838 to 0.908. MedSigLIP's two-site external mean rises from 0.594 to 0.640. RAD-DINO, operating at chance-level baseline grounding, predictably gains the least external AUC. These gains generalize broadly. Splitting performance by evaluation site (Table 3, Table 23) reveals that 34 of 64 public cells at $\lambda = 1$ are significantly positive, spanning all eight external sites and seven of eight beds.

Divergence, failures, and limitations. Failures concentrate by encoder rather than by site, with RAD-DINO absorbing 13 of the 20 significant negatives (Appendix F.6), underscoring that the CCA effect sign remains unpredictable *a priori* without the scope-condition audit. Internal and external metrics also frequently diverge. For MedSigLIP, CCA holds internal AUC stable (0.664 to 0.686) while driving a fourfold larger gain on PadChest-GR, exposing the inadequacy of internal validation for selecting transferable adapters (Appendix H.2). While the audit implies a training-free selection rule, utilizing it requires external labels to establish the evidence–transfer link per pair, defeating the method's annotation efficiency. CCA evades this circularity by measuring efficacy directly against an optimizer-matched control within the same cell (Appendices F.5, F.7).

5 Conclusion

Few-label adaptation yields AUC gains that do not track causal grounding. Across eight task beds, standard adapters fail to surpass a prior-matched, gradient-free null in 44 of 48 cells. These metrics fundamentally change how a reported gain must be read. Because the capacity for grounded transfer cannot be predicted *a priori*, we secure our audit with a placebo control to strictly discard spurious cells. Rather than leaving this grounding to chance, counterfactual-consistency adaptation (CCA) enforces it directly during training. While specific methodological constraints bound these findings (Appendix I), our ultimate contribution remains the loop: measure grounding, validate the measurement, then train for it.

Ethics Statement

This study uses only publicly available, de-identified datasets, each obtained through its official release channel and used under its own license (Table 4). Because no new patient data were col-

lected and no human subjects recruited, the work is exempt from institutional review board approval. External label provenance varies by site—report-mined, expert-localized, and proxy-derived—and Appendix B.3 details the rule family and cohort size for every cell so that label noise remains attributable rather than hidden.

The adapters studied here are retrospective research probes, not clinical tools. None has been prospectively validated, regulator-cleared, or tested for subgroup fairness, and none should inform patient care. Our central finding is itself the primary ethical caution: accuracy alone does not evidence clinical validity, as adapters can post sizeable AUC gains without measurable reliance on the targeted finding. Furthermore, grounding scores must not be indiscriminately optimized as training targets; several mask-using baselines artificially inflate the audit's win-rate while actually *lowering* external AUC (Appendix G.1). We thus release the audit strictly as a diagnostic instrument. Remaining limitations bearing on responsible interpretation are enumerated in Appendix I.

Reproducibility statement

All datasets are public. We release frozen splits for all eight task beds (Table 4), all cached encoder features, and the complete analysis scripts required to reproduce every number and figure. Crucially, the audit requires no gradients and no GPU once features are cached: gate-0, the paired intervention, and the placebo are computed purely from these caches. Training recipes, seeds, and λ grids are fully specified (Appendices B.2, C.3; Section 4.1). All significance statements utilize the seed-block permutation and hierarchical bootstrap procedures of Appendix F.1, with full implementations provided in the released codebase.

Use of AI

We used AI assistants, specifically large language models, as a general-purpose writing aid: grammar correction, phrasing, and copy-editing; drafting non-technical text; and surfacing candidate related work, which the authors located and read in original form before citation. AI was not used to generate research ideas, design the protocol, select or implement baselines, run experiments, or produce any numerical result, figure, or claim in this paper. Every statement and number was derived by the authors from the released code and verified against it, and the authors take full responsibility for all content, including AI-assisted text.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756–779, 2023.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66: 101797, 2020.

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 21381–21394. IEEE, 2023.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S Chaudhari, and Curtis Langlotz. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024. URL https://arxiv.org/abs/2401.12208.

Geon Choi, Hangyul Yoon, Hyunju Shin, Hyunki Park, Sang Hoon Seo, Eunho Yang, and Edward Choi. Instruction-guided lesion segmentation for chest x-rays with automatically generated large-scale dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1482–1492, 2026.

Lauren H Cooke, Matthias Jung, Jan M Brendel, Nora M Kerkovits, Borek Foldyna, Michael T Lu, and Vineet K Raghu. Roentmod: a synthetic chest x-ray modification model to identify and correct image interpretation model shortcuts. *npj Digital Medicine*, 9(1):324, 2026.

Heet Nitinkumar Dalsania. Label-efficient chest x-ray diagnosis via partial clip adaptation. *arXiv preprint arXiv:2507.07254*, 2025.

Daniel Coelho De Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, et al. Padchest-gr: A bilingual chest x-ray dataset for grounded radiology report generation. *NEJM AI*, 2(7):AIdbp2401120, 2025.

Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, John-Richard Ordóñez-Varela, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, pp. 231–234, 2014.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Ziye Deng, Ruihan He, Jiaxiang Liu, Yuan Wang, Zijie Meng, Songtao Jiang, Yong Xie, and Zuozhu Liu. Med-glip: Advancing medical language-image pre-training with large-scale grounded dataset. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 865–871. IEEE, 2025.

Sihao Ding, Santosh Vasa, and Aditi Ramadwar. Explanation-driven counterfactual testing for faithfulness in vision-language model explanations. *arXiv preprint arXiv:2510.00047*, 2025.

Kabilan Elangovan and Daniel Ting. When fine-tuning changes the evidence: Architecture-dependent semantic drift in chest x-ray explanations. *arXiv preprint arXiv:2604.08513*, 2026.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631, 2021.

David Freedman and David Lane. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298, 1983.

Ben Glocker, Charles Jones, Mélanie Roschewitz, and Stefan Winzeck. Risk of bias in chest radiography deep learning foundation models. *Radiology: Artificial Intelligence*, 5(6):e230060, 2023.

Advait Gosai, Arun Kavishwar, Stephanie L McNamara, Soujanya Samineni, Renato Umeton, Alexander Chowdhury, and William Lotter. Beyond diagnosis: Evaluating multimodal llms for pathology localization in chest radiographs. *Proceedings of machine learning research*, 297:572, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Gabriel Huang, Hugo Larochelle, and Simon Lacoste-Julien. Are few-shot learning benchmarks too simple? solving them without task supervision at test-time. *arXiv preprint arXiv:1902.08605*, 2019.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3922–3931. IEEE, 2021.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. Lp++: A surprisingly strong linear probe for few-shot clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23773–23782, 2024.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Karthik, Maggie, and Sohier Dane. Aptos 2019 blindness detection. `https://kaggle.com/competitions/aptos2019-blindness-detection`, 2019. Kaggle.

Younghyun Kim, Jongheon Jeong, Sangkyung Kwak, Kyungmin Lee, Juho Lee, and Jinwoo Shin. Starft: Robust fine-tuning of zero-shot models via spuriosity alignment. *arXiv preprint arXiv:2505.13232*, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Aditya V Kulkarni, John R Zech, and Paul H Yi. Sharpness matters: Higher image resolution improves generalization and explainability in chest x-ray classification. *Scientific Reports*, 16(1): 21856, 2026.

Bhavesh Kumar, Dylan Feng, and Leonard Tang. Mj1: Multimodal judgment via grounded verification. *arXiv preprint arXiv:2603.07990*, 2026.

Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511–522, 2019.

Xin Li, Yao Qiang, Chengyin Li, Sijia Liu, and Dongxiao Zhu. Saliency guided adversarial training for learning generalizable features with applications to medical imaging classification system. *arXiv preprint arXiv:2209.04326*, 2022.

Chenyu Lian, Hong-Yu Zhou, Yizhou Yu, and Liansheng Wang. Less could be better: Parameter-efficient fine-tuning advances medical vision foundation models. *arXiv preprint arXiv:2401.12215*, 2024.

Chenyu Lian, Hong-Yu Zhou, Chun-Ka Wong, and Jing Qin. Concept-guided noisy negative suppression for zero-shot classification and grounding of chest x-ray findings. *arXiv preprint arXiv:2605.19374*, 2026.

Jie Lian, Jingyu Liu, Shu Zhang, Kai Gao, Xiaoqing Liu, Dingwen Zhang, and Yizhou Yu. A structure-aware relation network for thoracic diseases detection and segmentation. *IEEE Transactions on Medical Imaging*, 40(8):2042–2052, 2021.

Mingquan Lin, Gregory Holste, Song Wang, Yiliang Zhou, Yishu Wei, Imon Banerjee, Pengyi Chen, Tianjie Dai, Yuexi Du, Nicha C Dvornek, et al. Cxr-lt 2024: A miccai challenge on long-tailed, multi-label, and zero-shot disease classification from chest x-ray. *Medical Image Analysis*, pp. 103739, 2025a.

Qika Lin, Yifan Zhu, Bin Pu, Ling Huang, Haoran Luo, Jingying Ma, Zhen Peng, Tianzhe Zhao, Fangzhi Xu, Jian Zhang, et al. A foundation model for chest x-ray interpretation with grounded reasoning via online reinforcement learning. *arXiv e-prints*, pp. arXiv–2509, 2025b.

Mahshad Lotfinia, Sebastian Ziegelmayer, Lisa Adams, Daniel Truhn, Andreas Maier, and Soroosh Tayebi Arasteh. Vision-language models for chest radiography do not always need the image. *arXiv preprint arXiv:2606.17710*, 2026.

Haozhe Luo, Shelley Zixin Shu, Ziyu Zhou, Sebastian Otálora, and Mauricio Reyes. Xbench: A comprehensive benchmark for visual-language explanations in chest radiography. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2026.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Raajitha Muthyala, Zhenan Yin, Alekhya Jilla, Frank Li, Theo Dapamede, Bardia Khosravi, Mohammadreza Chavoshi, Judy Gichoya, and Saptarshi Purkayastha. Frozen foundation-model embeddings discard small-lesion signal in chest radiography: Implications for pre-deployment evaluation. *arXiv preprint arXiv:2606.11606*, 2026.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, 2025.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pp. 8116–8126. PMLR, 2020.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017.

Binesh Sadanandan and Vahid Behzadan. Attention without grounding: Causal evaluation of visual explanations in medical vlms. *arXiv preprint arXiv:2607.18577*, 2026.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

A Saporta, X Gui, A Agrawal, A Pareek, SQ Truong, CD Nguyen, VD Ngo, J Seekins, FG Blankenberg, AY Ng, et al. Benchmarking saliency methods for chest x-ray interpretation., 2022, 4. *DOI: https://doi. org/10.1038/s42256-022-00536-x*, pp. 867–878.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Fereshteh Shakeri, Yunshi Huang, Julio Silva-Rodríguez, Houda Bahig, An Tang, Jose Dolz, and Ismail Ben Ayed. Few-shot adaptation of medical vision-language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 553–563. Springer, 2024.

Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. Gradmask: Reduce overfitting by regularizing saliency. *arXiv preprint arXiv:1904.07478*, 2019.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Théo Sourget, Michelle Hestbek-Møller, Amelia Jiménez-Sánchez, Jack Junchi Xu, and Veronika Cheplygina. Mask of truth: model sensitivity to unexpected regions of medical images. *Journal of imaging informatics in medicine*, 39(1):767–784, 2026.

Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, pp. 728–744. PMLR, 2021.

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.

Joseph D Viviano, Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. Saliency is a possible red herring when diagnosing poor generalization. *arXiv preprint arXiv:1910.00199*, 2019.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens Van Der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7949–7961. IEEE, 2022.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 21315–21326. IEEE, 2023.

Tianrun Xu, Haoda Jing, Ye Li, Yuquan Wei, Jun Feng, Guanyu Chen, Haichuan Gao, Tianren Zhang, and Feng Chen. Defacto: Counterfactual thinking with images for enforcing evidence-grounded and faithful reasoning. *arXiv preprint arXiv:2509.20912*, 2025.

SM Zain, Eram Mahamud, Md Assaduzzaman, Nafiz Fahad, and Tze Hui Liew. Do deep chest x-ray classifiers explain where they look? a multi-architecture faithfulness benchmark with false-positive attribution audit on vindr-cxr. *Intelligence-Based Medicine*, pp. 100431, 2026.

Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, ParasLakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation. `https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation`, 2019. Kaggle.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. Cf-vlm: counterfactual vision-language fine-tuning. *arXiv preprint arXiv:2506.17267*, 2025.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.

CONTENTS

A Discussion **17**

B Theory and protocol **17**

B.1 Formal justification of the audit 17

B.2 Task bed construction and protocol details 18

B.3 External label provenance and alignment 20

B.4 Reporting checklist for few-label adaptation audits 20

C Prior-matched nulls and adapter capacity **21**

C.1 Is the probe under-tuned? Gate-0 versus the tuned probe 21

C.2 Stronger prior-matched nulls 22

C.3 Adaptation depth 23

C.4 Partially unfrozen encoders: a preliminary probe 23

C.5 Is the attentive head under-optimized? 24

C.6 Post-hoc attribution 24

D Validating the intervention **24**

D.1 Placebo control: per-designation values 24

D.2 Label-shuffled negative control 25

D.3 Why the gap improves on the raw evidence . . . 27
D.4 Why we do not test the win-rate against 0.5 . . . 28
D.5 How well are control regions texture-matched? . . . 28
D.6 Sensitivity to control selection . . . 28
D.7 Sensitivity to the occlusion strategy . . . 29

E What the frozen representation supports **29**
E.1 Three frozen routes to localizing the finding . . . 29
E.2 Recall is representation-bound: resolution and operating point. . . . 30
E.3 Universal in-context segmentation is organ-tuned . . . 31
E.4 Substrate inversion . . . 31

F Statistics, power, and selection **32**
F.1 Dependence-aware statistics . . . 32
F.2 How much external data does the applicability check need? . . . 33
F.3 Scope-condition audit by bed . . . 34
F.4 A second attempt at the scope condition: taxonomy-controlled acquisition shift . . . 34
F.5 Evidence-gated selection: two signals, both negative . . . 35
F.6 CCA per external site . . . 36
F.7 Why we do not propose training-free selection . . . 36

G Training for grounding **36**
G.1 Supervision-matched mask baselines . . . 36
G.2 The two strongest mask baselines . . . 38
G.3 CCA under noisy training masks . . . 39
G.4 Sensitivity to the hinge margin . . . 39
G.5 λ regimes: when grounding starts to cost accuracy . . . 40

H Replication and additional results **40**
H.1 Modality transfer: a fundus bed . . . 40
H.2 Calibration and decision thresholds out of domain . . . 41
H.3 Qualitative case studies . . . 42
H.4 Per-budget and per-cell detail . . . 44
H.5 Full result tables and figures . . . 45
H.6 External calibration under CCA . . . 46

I Limitations in full **46**

J Extended related work **48**

A Discussion

What the audit is for. The proposed framework requires minimal computational overhead, requiring only four one-time cached encoder passes per annotated positive and a subsequent dot product per adapter (Table 5). Despite this efficiency, the method fundamentally reframes the interpretation of label-scarce adaptation. An AUC of 0.67 achieved without measurable reliance on the target finding is clinically and causally distinct from the same AUC achieved with an intervention win-rate of 0.90. The audit explicitly separates these states. Consequently, we advocate integrating gate-0, the paired counterfactual intervention, and the lesion-free placebo as a minimum reporting standard for few-shot adaptation claims (Appendix B.4). However, this audit serves as a diagnostic instrument rather than an absolute deployment gate. Verifying its applicability to novel (finding, encoder) pairs requires empirical validation. Furthermore, the statistical power for such verification remains constrained at realistic external sample sizes (Appendix F.2).

Heterogeneity and Criteria for CCA Deployment. Evidence-transfer correlations exhibit significant variance across experimental cells and range from -0.34 to $+0.65$. This heterogeneity is fully compatible with our core findings because no conclusions rely on aggregating directional effects across disparate domains. Instead, the unpredictability of these activation conditions empirically justifies the necessity of per-cell auditing. This intra-cell evaluation directly informs a principled deployment strategy that requires no external validation labels. The first step is to train a standard adapter (λ=0) and audit its placebo response. If the cell is spuriously reactive or "loud" according to Definition 1, the underlying mechanism of CCA cannot be validated. In such cases, practitioners should exclusively report accuracy metrics. If the cell is "quiet," one can train the constrained model (λ=1) and accept it if the paired contrast increases without degrading internal AUC. Applied to our test beds, this strictly measurement-driven protocol correctly adopted CCA for all CXR-domain encoders while rejecting it for natural-image controls.

Generalization to Volumetric Modalities. While the primary evaluation targets projection radiography, the underlying logic of the "measure-validate-act" loop extends to volumetric imaging such as CT or ultrasound. The foundational principles of paired contrast, placebo validation, and optimizer-matched training transfer directly. However, the practical implementation demands modality-specific adaptations. Occlusion operators must respect local noise distributions, such as avoiding unnatural smoothing in ultrasound speckle. Additionally, control selection requires morphological and intensity matching across three dimensions, which exponentially expands the candidate search space. Developing these 3D implementations represents a concrete direction for future research.

B Theory and protocol

B.1 Formal justification of the audit

Definition 2 (gate-0)**.** *Given k labeled positives (and optionally negatives) with frozen embeddings, gate-0 scores $s_0(x) = \cos\big(f(x), \mu_k^+\big)$, or $s_{0b}(x) = f(x)^\top(\mu_k^+ - \mu_k^-)/\|\mu_k^+ - \mu_k^-\|$; only a decision threshold is fit.*

Proposition 2 (gate-0 is Bayes-optimal under first-order class structure)**.** *Assume class conditionals are homoscedastic isotropic Gaussians, $f(x) \mid y \sim \mathcal{N}(\mu_y, \sigma^2 I)$. The Bayes-optimal decision rule is a threshold on $f(x)^\top(\mu_+ - \mu_-)$. This corresponds to the population version of s_{0b}.*

Proof. The log-likelihood ratio is $\log \frac{p(f|+)}{p(f|-)} = \frac{1}{\sigma^2} f^\top(\mu_+ - \mu_-) + \kappa$, where κ is constant with respect to f. The Bayes rule for 0–1 loss thresholds the log-likelihood ratio at $\log(\pi_-/\pi_+)$ for class priors $\pi_\pm$. Since that ratio is strictly increasing in $f^\top(\mu_+ - \mu_-)$, the decision boundary simplifies to a threshold on $f^\top(\mu_+ - \mu_-)$. □

Consequently, under these theoretical assumptions, an adapter that fails to outperform gate-0b demonstrates nothing beyond first-order class statistics and centroid estimation noise. The converse, however, does not hold. Because gate-0b is theoretically optimal under this generative model, failing to surpass it provides no conclusive information regarding the adapter's internal representations. These assumptions serve as a theoretical idealization rather than an empirical description of

the actual embeddings. Training pool class-conditional covariances exhibit strong anisotropy and heteroscedasticity. We observe top-to-mean eigenvalue ratios ranging from 7 to 17. Relative covariance differences, calculated as $\|\Sigma^+ - \Sigma^-\|_F / \frac{1}{2}(\|\Sigma^+\|_F + \|\Sigma^-\|_F)$, span 0.38 to 0.96 based on the top-50 principal components across all evaluated cells.

While Proposition 2 theoretically motivates the structure of gate-0, the equivalence finding remains empirical and its interpretation strictly directional. Standard-recipe adapters yield performance consistent with mere first-order class statistics on these test sets. Although regularization-tuned probes can technically extract signals beyond this first-order structure (Appendix C.1), the audit confirms that such extracted signals remain largely ungrounded.

Proposition 1 (from Section 2.1)**.** *Suppose the score change from occluding a region R decomposes as $\Delta(R) = \beta\,\mathbf{1}[R \supseteq \textit{lesion}] + a(R)$, where β is the per-image lesion-specific effect and the artefact term a depends only on region shape and local texture statistics. If control regions are sampled to match the exact shape and texture statistics of the lesion region, and the disjointness constraint in Eq.* (4) *guarantees $\mathbf{1}[R_{ctl} \supseteq \textit{lesion}] = 0$, then $\mathbb{E}[\Delta_{les} - \Delta_{ctl}] = \beta$. Consequently, the paired contrast serves as an unbiased estimate of the lesion-specific effect, mathematically canceling any occlusion-generic components common to both branches.*

Proof. By definition, $\mathbb{E}[\Delta_{les}] = \beta + \mathbb{E}[a(R_{les})]$ and, by the disjointness constraint, $\mathbb{E}[\Delta_{ctl}] = \mathbb{E}[a(R_{ctl})]$. The matching condition assumes $a(R_{ctl}) \overset{d}{=} a(R_{les})$, which yields $\mathbb{E}[a(R_{les})] = \mathbb{E}[a(R_{ctl})]$ and isolates the difference as exactly β. □

Two practical constraints separate this implementation from the theoretical idealization. First, although shape is matched exactly, texture is matched only to the first order. Control candidates are screened by local mean intensity (Eq. (4)), leaving second-order statistics partially misaligned. Appendix D.5 quantifies this residual mismatch, providing a cross-encoder argument that it does not drive the win-rate. Second, the artifact term $a(\cdot)$ may encapsulate long-range interactions extending beyond the occluded region itself, rendering the theoretical cancellation approximate. Appendix D.7 bounds this effect, demonstrating that the intervention's outcomes remain stable across three distinct occlusion methodologies. For a linear head, the score change reduces exactly to $\Delta(R) = w^\top \big(f(x) - f(x_{\setminus R})\big)$, allowing the paired contrast to be computed directly from cached occlusion features. Consequently, the CCA constraint in Eq. (3) applies a hinge loss on $w^\top \big(\delta_{les} - \delta_{ctl}\big)$. Trained jointly with the standard logistic loss, this formulation renders the linear tier of CCA a strictly convex optimization problem.

B.2 Task bed construction and protocol details

Headline bed. The core evaluation utilizes VinDr-CXR (Nguyen et al., 2022) calcification data. Section 2 details the split sizes and label budgets. Positive samples contain a consensus bounding box for the calcification family, and the k training labels are drawn exclusively from the training pool.

Replication evaluation beds. The replication framework incorporates four PadChest-GR findings (effusion, nodule, atelectasis, consolidation) utilizing grounded-report bounding boxes. We also evaluate two pneumothorax datasets with independent provenance: SIIM-ACR (Zawacki et al., 2019) pixel masks and MIMIC/MS-CXR (Johnson et al., 2019; Boecking et al., 2022) expert boxes. Additionally, we assess cross-modality transfer using fundus microaneurysm data from DDR (Li et al., 2019), transferring to IDRiD (Porwal et al., 2018), Messidor-2 (Decencière et al., 2014), and APTOS (Karthik et al., 2019). Each chest radiograph bed is evaluated across all six encoders, while the fundus bed substitutes RETFound for XrayCLIP-L/14. Table 4 details the frozen split sizes, median lesion scales, and external-site positive counts for all eight public datasets.

Compute accounting. Table 5 delineates the one-time caching overhead from recurring computational costs. All reported quantities represent a single encoder forward pass at the working resolution.

Evaluation axes. The audit requires four encoder passes per annotated positive sample, whereas the placebo control requires zero. The evaluation utilizes six frozen encoders: RAD-DINO (Pérez-García et al., 2025) (CXR SSL), MedSigLIP (Sellergren et al., 2025) (biomedical VLM), two Xray-

Table 4: **All task beds are public.** Split sizes are frozen. The k-sweep draws exclusively from the train pool. "Lesion" denotes the median annotated area as an image fraction, spanning over two orders of magnitude across beds (0.02% to 7.55%). External-site positive counts appear in parentheses. Sites capped at 500^+ are subsampled from a larger pool.

Finding	Source	Annotation	Train / test ($^+/^-$)	Lesion	External sites
Calcification	VinDr-CXR (Nguyen et al., 2022)	box union (3 readers)	200/602 · 130/400	0.32%	ChestX-Det10 (Lian et al., 2021) (201^+), PadChest-GR (64^+)
Nodule	PadChest-GR (Bustos et al., 2020; De Castro et al., 2025)	grounded-report box	105/2576 · 130/400	0.77%	CheXpert (Irvin et al., 2019) (500^+), NIH (Wang et al., 2017) (500^+), VinDr-CXR (Nguyen et al., 2022) (500^+)
Effusion	PadChest-GR (Bustos et al., 2020; De Castro et al., 2025)	grounded-report box	244/2437 · 130/400	3.85%	CheXpert (Irvin et al., 2019) (500^+), NIH (Wang et al., 2017) (500^+), VinDr-CXR (Nguyen et al., 2022) (500^+)
Atelectasis	PadChest-GR (Bustos et al., 2020; De Castro et al., 2025)	grounded-report box	139/2542 · 130/400	2.56%	CheXpert (Irvin et al., 2019) (500^+), NIH (Wang et al., 2017) (500^+), VinDr-CXR (Nguyen et al., 2022) (187^+)
Consolidation	PadChest-GR (Bustos et al., 2020; De Castro et al., 2025)	grounded-report box	100/2638 · 73/400	7.55%	CheXpert (Irvin et al., 2019) (500^+), NIH (Wang et al., 2017) (500^+), VinDr-CXR (Nguyen et al., 2022) (353^+)
Pneumothorax	SIIM-ACR (Zawacki et al., 2019)	pixel mask	200/602 · 130/400	0.95%	NIH (Wang et al., 2017) (500^+), VinDr-CXR (Nguyen et al., 2022) (96^+)
Pneumothorax	MIMIC / MS-CXR (Johnson et al., 2019; Boecking et al., 2022)	expert box	153/500 · 80/300	3.51%	CheXpert (Irvin et al., 2019) (500^+), NIH (Wang et al., 2017) (500^+), VinDr-CXR (Nguyen et al., 2022) (96^+)
Microaneurysms	DDR (Li et al., 2019)	pixel mask (fundus)	192/602 · 130/400	0.02%	APTOS (Karthik et al., 2019) (1857^+), IDRiD (Porwal et al., 2018) (348^+), Messidor-2 (Decencière et al., 2014) (727^+)

Table 5: **Encoder passes by stage.** The cache is built once per (bed, encoder), rendering all subsequent steps free of encoder forward passes. On the primary bed (1,332 images, 458 of them annotated positives), the full cache requires approximately 20 GPU-minutes per encoder on a single worker—dominated by image decoding, not the forward passes—after which one adapter audit is a handful of dot products and one 500-step linear CCA run takes seconds on CPU.

Stage	Encoder passes	Frequency	Notes
Clean feature cache	1 per image	once	all splits and extras
Audit cache	$+4$ per annotated positive	once	1 lesion-occ $+$ 3 control-occ (Eq. (4))
CCA constraint cache	$+4$ per annotated *train* positive	once	same regions, train split
Placebo	0	—	reuses the cached control deltas
Adapter audit	0	per adapter	dot products with cached δ
CCA training	0	per step	hinge on cached contrasts (Eq. (3))

CLIP models (-B/16, -L/14) (Chen et al., 2024), DINOv2-S (Oquab et al., 2023), and an ImageNet ResNet-50 (He et al., 2016) as a natural-image baseline. Adaptation is evaluated across three architectural depths: a logistic probe on mean-pooled features, an attentive pooling head, and LoRA (Hu et al., 2021) applied to encoder blocks. All adaptation depths utilize identical cached control regions to guarantee strictly paired evidence comparisons. Finally, we deploy this identical audit framework across seven additional public beds (Section 3) to empirically establish the method's scope conditions.

Detection harness. The per-patch activation map $w^\top f_p$ is bilinearly upsampled from a 37×37 grid to a 512×512 resolution and min-max normalized per image. Nineteen global thresholds ranging from 0.05 to 0.95 binarize this map. Connected components smaller than 16 pixels are discarded. A ground-truth lesion is considered successfully detected if a predicted component achieves an Intersection over Union (IoU) ≥ 0.1. Sensitivity at FPPI ≤ 1 represents the maximum pooled lesion-level sensitivity achieved at thresholds that yield an average of one or fewer false-positive components per image. Peak-in-GT measures the fraction of test positives where the single argmax pixel falls strictly inside the annotation. The chance level for this metric tracks the bed's mean annotated-area fraction (0.0061 on the primary bed), representing the mathematical probability of

a uniformly random argmax landing inside the target. Confidence intervals are derived via image-index bootstrapping (B=100).

Occlusion and control regions. The occlusion operator and the resulting score drop are Eq. (1). Control regions translate the lesion's exact shape to non-lesion positions, avoiding lesion overlap. Writing π for a translation, F for the central 10–90% of the frame, and $\bar{x}_R$ for the mean intensity inside R, we draw a candidate pool — at most 60 translation draws, keeping up to 18 disjoint candidates — and select its three best texture matches:

$$\begin{aligned} &\mathcal{C} \text{ is a 60-element sample of } \big\{\pi(R_{les}) : \pi(R_{les}) \cap R_{les} = \emptyset,\ \pi(R_{les}) \subseteq F\big\}, \\ &\{R_{ctl_0}, R_{ctl_1}, R_{ctl_2}\} = \underset{R' \in \mathcal{C}}{\arg\min}^{(3)} \left|\bar{x}_{R'} - \bar{x}_{R_{les}}\right|, \end{aligned} \tag{4}$$

where $\arg\min^{(3)}$ denotes the three smallest values, indexed in increasing order of mismatch. Shape is therefore matched exactly and texture to first order only; R_{ctl_0} is by construction the best match, which is why the placebo of Eq. (2) excludes it (Appendix D.1). These regions are cached once per image and reused across all adapter evaluations. Consequently, evidence comparisons remain paired across methods and adaptation tiers by construction.

B.3 External label provenance and alignment

External datasets diverge not only in image acquisition protocols but also in their underlying label generation methodologies, fundamentally bounding the interpretation of cross-site AUC metrics. Table 6 details the label source, binary conversion rule, and resulting class prevalence for every evaluated configuration. These labels fall into three distinct categories. *Report-mined labels* (e.g., NIH ChestX-ray14, CheXpert) rely on natural language processing extractions from clinical radiology reports. While highly scalable, these labels exhibit inherent noise and draw negative samples from a heterogeneous pool of healthy and alternately abnormal studies. *Expert-localized labels* (e.g., ChestX-Det10, PadChest-GR) provide explicit spatial bounding for the target finding. An image is classified as positive if at least one annotator marks a target-class region. For VinDr-CXR positives, we utilize the per-reader release rather than the consensus annotations used for the primary test masks, preserving the dataset's native asymmetry. Finally, *Proxy labels* (e.g., fundus imaging sites) derive from international disease severity grading systems. Because a diabetic retinopathy grade of ≥ 1 structurally implies the presence of microaneurysms, the site-level label serves as a proxy rather than a direct lesion annotation.

Two specific alignment disparities require careful consideration. First, on CheXpert the explicit-negative pool is exhausted for two findings, so negatives are topped up with `No Finding` exams (nodule: 1,916 of 2,000; atelectasis: 1,895 of 2,000). This conflates the specific target-versus-rest classification with a generalized anomaly detection task, introducing a systemic shortcut opportunity absent in other datasets. Uncertain labels (-1) are strictly excluded across all evaluations. Second, negative pools differ in hardness: the SIIM pneumothorax bed's VinDr negatives are all abnormal images without pneumothorax boxes (hard negatives), whereas the PadChest beds' VinDr negatives are predominantly no-finding exams. Configurations with severely limited external positive samples, such as PadChest-GR calcification (64^+) and VinDr pneumothorax (96^+), invariably yield wider AUC confidence intervals, placing strict limits on available statistical power (Appendix F.2).

B.4 Reporting checklist for few-label adaptation audits

While the proposed audit is computationally inexpensive, incomplete reporting can easily obscure critical methodological details. We recommend that any few-label adaptation claim concerning grounded features reports the following minimum specifications:

1. **Prior-matched null.** Report the adapter AUC alongside the gate-0/gate-0b null at every label budget, specifying the equivalence margin and statistical test (Eq. (5)). A probe failing to outperform a threshold-recalibrated fixed rule has not demonstrated learning beyond first-order statistics.
2. **Paired intervention.** The occlusion win-rate with the control-selection rule (Eq. (4)), the number of controls, and the caching policy that keeps comparisons paired across methods.

Table 6: **External label provenance for every (bed, site) cell.** Rule families: NLP = report-mined label; box/polygon = expert-localized, positive iff ≥ 1 target-class region; proxy = DR grade ≥ 1 implies the lesion. n^+/n^- are the evaluated cohort sizes (positives capped at 500 where the pool allows).

Bed	Site	Source	Label rule	n^+	n^-	pos %
Calcification (VinDr)	ChestX-Det10	expert boxes	box: `Calcification`	201	3342	5.7
	PadChest-GR	report-grounded polygons	polygon: calcification family	64	3147	2.0
Effusion (PadChest-GR)	NIH	report NLP	NLP: `Effusion`	500	2000	20.0
	CheXpert	labeler NLP	NLP: `Pleural Effusion`= 1	500	2000	20.0
	VinDr-CXR	reader boxes ($\geq$1 of 3)	box: `Pleural effusion`	500	2000	20.0
Nodule (PadChest-GR)	NIH	report NLP	NLP: `Nodule`	500	2000	20.0
	CheXpert	labeler NLP	NLP: `Lung Lesion`= 1 (nodule/mass merged)	500	2000	20.0
	VinDr-CXR	reader boxes	box: `Nodule/Mass` (merged)	500	2000	20.0
Atelectasis (PadChest-GR)	NIH	report NLP	NLP: `Atelectasis`	500	2000	20.0
	CheXpert	labeler NLP	NLP: `Atelectasis`= 1	500	2000	20.0
	VinDr-CXR	reader boxes	box: `Atelectasis` (full pool)	187	2000	8.6
Consolidation (PadChest-GR)	NIH	report NLP	NLP: `Consolidation`	500	2000	20.0
	CheXpert	labeler NLP	NLP: `Consolidation`= 1	500	2000	20.0
	VinDr-CXR	reader boxes	box: `Consolidation` (full pool)	353	2000	15.0
Pneumothorax (SIIM)	NIH	report NLP	NLP: `Pneumothorax`	500	2000	20.0
	VinDr-CXR	reader boxes	box: `Pneumothorax`; hard negatives	96	2000	4.6
Pneumothorax (MS-CXR)	CheXpert	labeler NLP	NLP: `Pneumothorax`= 1 (bed-matched taxonomy)	500	2000	20.0
	NIH	report NLP	NLP: `Pneumothorax`	500	2000	20.0
	VinDr-CXR	reader boxes	box: `Pneumothorax`; hard negatives	96	2000	4.6
Microaneurysms (DDR)	IDRiD-B	adjudicated DR grade	proxy: grade ≥ 1	348	168	67.4
	APTOS 2019	clinical DR grade	proxy: grade ≥ 1	1857	1805	50.7
	Messidor-2	adjudicated DR grade	proxy: grade ≥ 1	727	1017	41.7

3. **Instrument placebo.** Disclose the placebo win-rate and the specific per-cell verdict (Definition 1). Spuriously reactive cells must be discarded outright rather than mathematically corrected.
4. **Optimizer-matched training control.** Any proposed train-for-grounding strategy must be reported alongside its unconstrained λ=0 baseline utilizing identical recipes, seeds, and training budgets. Optimizing hyperparameters disparately effectively masquerades standard training gains as grounding improvements.
5. **Dependence-aware statistics.** Because experimental seeds share evaluation pools and cached controls, standard run-level p-values artificially inflate the effective sample size. Evaluations must employ block-permutation or hierarchical bootstrap alternatives (Appendix F.1).
6. **Denominators.** Explicitly state every cell count with its underlying decomposition and document the exclusion rationale for discarded cells.

C Prior-matched nulls and adapter capacity

C.1 Is the probe under-tuned? Gate-0 versus the tuned probe

The equivalence in Section 3 compares gate-0 against the standard few-shot linear-probe recipe utilizing logistic regression with default regularization (C=1). This specific recipe is widely reported in the adaptation literature (Radford et al., 2021; Oquab et al., 2023; Zhang et al., 2023; Tiu et al., 2022). Optimizing C via 5-fold cross-validation within the $2k$ training labels represents a practical deployment protocol. While this hyperparameter tuning extracts additional internal AUC for specific encoders, yielding up to a $+0.054$ surplus over gate-0b for XrayCLIP-L/14 at k=100, the surplus remains marginal ($\leq +0.010$) for RAD-DINO, ResNet-50, DINOv2-S, and XrayCLIP-B/16.

The audit reveals that hyperparameter tuning yields highly inconsistent grounding behavior. The intervention win-rate moves independently of internal AUC gains. On XrayCLIP-B/16, internal AUC rises while the win-rate collapses from 0.890 to 0.785. On DINOv2-S, internal AUC barely moves, yet the win-rate jumps from 0.512 to 0.717. For XrayCLIP-L/14, the extracted surplus is genuinely grounded (win-rate stable at 0.900) and successfully transfers ($+0.038$ external). For MedSigLIP, the surplus is grounded but transfers at less than half its internal face value ($+0.049$ internal, $+0.018$ external). Hyperparameter tuning therefore acquires an inseparable mixture of grounded and ungrounded signals that internal validation simply cannot untangle. Exposing this exact failure mode remains the primary function of the audit.

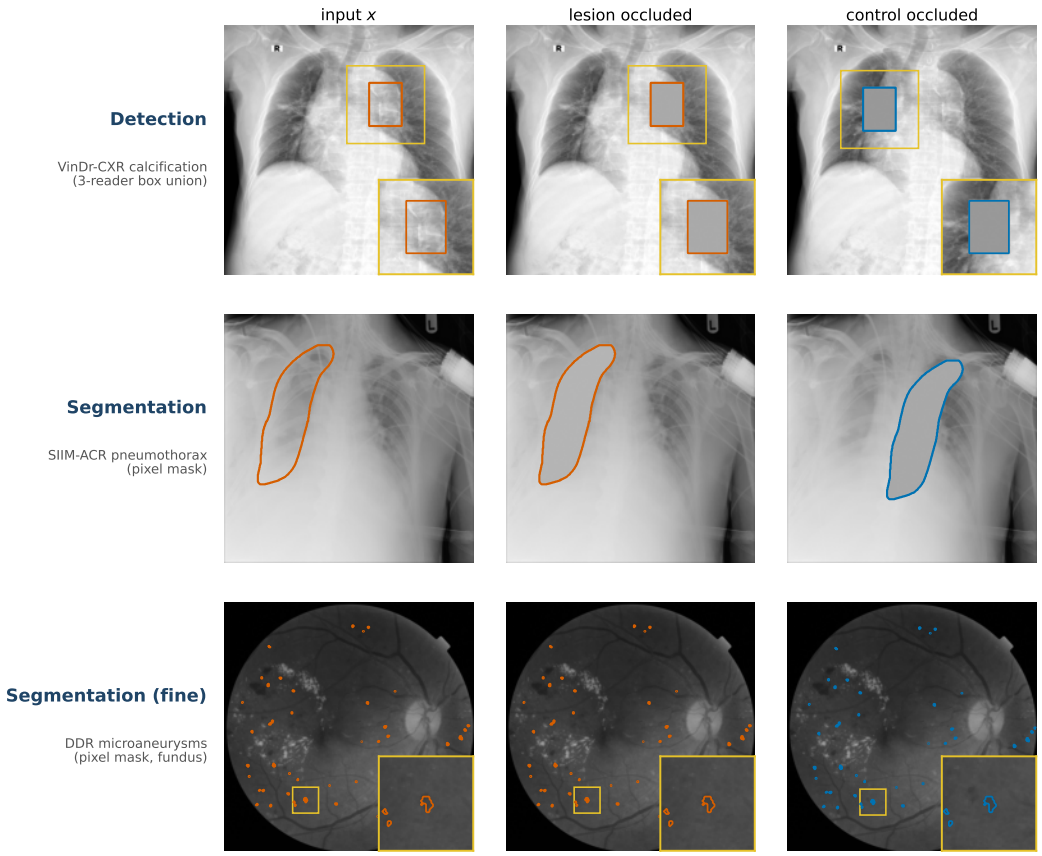


Figure 5: **The paired intervention across three annotation formats.** Examples include detection-style boxes (top: VinDr-CXR calcification), coarse pixel masks (middle: SIIM-ACR pneumothorax), and fine pixel masks in a distinct modality (bottom: DDR microaneurysms). Insets magnify the occluded regions, excluding the middle row where the lesion is too large. Despite the lesion area spanning nearly a factor of 50 across these beds (0.02% to 0.95% of the image), the protocol construction remains identical.

C.2 Stronger prior-matched nulls

Because gate-0 relies on a single positive centroid, the observed equivalence could theoretically result from an under-powered null, particularly for findings with multiple visual subtypes. To address this, we compare the trained probe against three more expressive, gradient-free nulls utilizing the identical k labels. These include M-prototype k-means ($M \in \{2, 4\}$), a 1-nearest-neighbor rule, and shrinkage LDA incorporating second-order statistics (Table 7).

These results strengthen the equivalence claim in two distinct directions. First, the multi-prototype and nearest-neighbor nulls perform worse than gate-0b across every encoder. Every variant trails the single centroid, so the observed equivalence is not an artifact of summarizing positives with too few prototypes. Splitting the prototypes actively degrades the null, as theoretically expected at these constrained budgets. Second, shrinkage LDA provides the only competitive null. As a closed-form rule requiring no gradient steps, it actually beats the trained probe on the two strongest encoders: MedSigLIP (0.747 vs. 0.736) and XrayCLIP-L/14 (0.754 vs. 0.741). On the remaining four encoders, it trails by a mere 0.017 (DINOv2-S) to 0.035 (RAD-DINO). A second-order rule that costs nothing to fit operates within a 0.035 margin everywhere and outright leads where the underlying representation is strongest. This rules out an under-powered null as the source of the equivalence. It also isolates the sole condition where a trained probe justifies its cost: when encoder features remain too weak for a closed-form second-order rule to exploit.

Table 7: **Prior-matched nulls at** k=100 on the primary VinDr-CXR calcification bed. All nulls use the identical k labels and fit no gradients. The proto-M label denotes M-prototype k-means, nn-1 denotes nearest neighbor, and lda denotes shrinkage LDA.

Encoder	probe	gate-0b	proto-2	proto-4	nn-1	LDA
RAD-DINO	0.666	0.657	0.608	0.561	0.565	0.631
MedSigLIP	0.736	0.736	0.627	0.611	0.620	**0.747**
XrayCLIP-B/16	0.716	0.713	0.645	0.600	0.575	0.697
XrayCLIP-L/14	0.741	0.730	0.660	0.625	0.616	**0.754**
DINOv2-S	0.640	0.640	0.530	0.500	0.556	0.623
ResNet-50	0.664	0.663	0.564	0.482	0.548	0.642

Larger budgets. Because the primary VinDr-CXR dataset contains 200 positive training images, we extend the gate-0 evaluation to $k \in \{150, 200\}$. Equivalence at the 0.03 margin holds for all six encoders at all budgets. At k=200 the probe-minus-gate-0b differences are $+0.000$ (MedSigLIP), $+0.003$ (XrayCLIP-B/16), $+0.004$ (DINOv2-S, ResNet-50), $+0.010$ (RAD-DINO), and $+0.021$ (XrayCLIP-L/14, 95% CI $[+0.018, +0.024]$). XrayCLIP-L/14 is the only encoder exhibiting a surplus that scales with k, reaching $+0.021$ at k=200. Therefore, the equivalence claim is strictly bounded by the budgets supported by these public datasets rather than asserted asymptotically.

C.3 Adaptation depth

We further evaluate adaptation depth on the primary bed using identical splits, seeds, and cached control regions: a logistic probe on mean features, an attentive-pooling head, and LoRA (Hu et al., 2021) applied to encoder query and value matrices (q, v) with rank 8, α=16, dropout 0.1, and 200 steps. At k=100, the linear probe achieves an AUC of 0.666. Deeper architectures underperform, with the attentive head reaching 0.593 and LoRA reaching 0.647. Simultaneously, causal evidence remains largely stagnant across models (0.543 to 0.572). LoRA also produces the lowest evidence localization score (0.009 versus the probe's 0.045), indicating that additional adapter capacity primarily absorbs ungrounded signals.

Table 8 summarizes these trends across representative budgets. While performance is comparable at k=5, the linear probe consistently outpaces deeper architectures as the label budget increases. External transfer metrics for LoRA mirror its internal AUC gains without any corresponding improvement in evidence localization, replicating the fundamental dissociation observed in linear probes.

Table 8: **Adaptation depth at three budgets** (headline bed (VinDr-CXR calcification), means over 20 seeds). Ev. loc is probe-CAM peak-in-GT against a chance line of 0.006. Deeper tiers are worse on AUC and no better on either evidence channel.

Method	k	AUC	Interv. win	Ev. loc	ext. PadChest-GR
Linear probe	5	0.554	0.434	0.008	–
	25	0.616	0.462	0.008	–
	100	0.666	0.543	0.045	–
Attentive pool	5	0.503	0.451	0.002	–
	25	0.549	0.531	0.022	–
	100	0.593	0.548	0.023	–
LoRA	5	0.559	0.492	0.009	0.535
	25	0.598	0.429	0.005	0.589
	100	0.647	0.572	0.009	0.618

C.4 Partially unfrozen encoders: a preliminary probe

To verify whether the observed dissociation stems strictly from freezing the encoder, we conduct a preliminary evaluation unfreezing the final two transformer blocks of RAD-DINO. We adapt this model on the effusion dataset (the largest lesions, 3.85% area) at k=25 with the LoRA-tier recipe (AdamW 10^{-4}, 200 steps, batch 8), $\lambda \in \{0, 1\}$, three seeds. Under the standard objective (λ=0), in-

ternal AUC reaches 0.845–0.929 and external AUC scales up to 0.916, yet the paired causal win-rate stagnates at 0.518–0.645. This confirms that performance gains remain decoupled from evidence grounding even when the underlying encoder weights are updated. Applying the CCA constraint (λ=1) successfully trains the audited behavior, elevating the win-rate to 0.873–0.945. However, the external AUC effects remain inconclusive across random seeds. This variability triggers the internal-AUC safeguard of our proposed decision rule (Appendix A) and suggests that full-block updates may require specialized learning rate schedules under the hinge loss.

C.5 Is the attentive head under-optimized?

The poor performance of the attentive head represents a representational deficit rather than an optimization failure. Sweeping the learning rate across $\{10^{-3}, 5{\cdot}10^{-3}, 2{\cdot}10^{-2}\}$ consistently yields training AUCs between 0.95 and 1.00. Despite perfectly fitting the $2k$ training labels, test AUC plateaus at 0.567–0.578 and the intervention win-rate remains at chance level.

C.6 Post-hoc attribution

A potential critique is that our analysis of static saliency evaluates only a single map family. This section refutes that by separating mathematical necessity from empirical measurement.

Where the objection cannot apply. For a linear head on mean-pooled features, the logit gradient with respect to patch feature f_p is w/N, independent of p. Consequently, methods such as Grad-CAM and zero-baseline integrated gradients strictly reduce to $w^\top f_p$ up to a positive scaling factor, rendering their patch rankings identical. We empirically verify that this mathematical degeneracy extends to the final token layer of the LoRA head, where token-layer Grad-CAM perfectly matches the intrinsic map across all evaluated configurations (Table 9).

Where it can apply, and what happens there. Map degeneracy breaks in three specific scenarios. These include softmax pooling in the attentive tier, intermediate-block Grad-CAM in the LoRA tier, and pixel-space integrated gradients. Evaluating these non-degenerate methods using the identical scorer and training recipe from Appendix C.3 confirms they still fail to localize the target lesion. Every evaluated map family remains within a factor of three of the 0.006 chance line, and sensitivity at 1 FPPI never exceeds 0.0022. While intermediate-block Grad-CAM for LoRA slightly separates from the intrinsic map (reaching a peak-in-GT of 0.031 at k=100), it remains an order of magnitude below genuinely localized maps. Furthermore, pixel-space integrated gradients fail to place their peak inside the annotation in any evaluated cell. The saliency failure reported in Section 3 is therefore independent of the attribution methodology (Table 9).

Figure 6 visualizes these two conclusions. As required by the degeneracy argument, the intrinsic map and token-layer Grad-CAM are indistinguishable. The two non-degenerate families produce visibly distinct maps, confirming they extract different information. However, none of the four maps concentrate inside the annotation. Instead, they highlight image borders, laterality markers, and mediastinal edges. These regions represent the exact spurious shortcuts the audit is designed to detect.

D Validating the intervention

D.1 Placebo control: per-designation values

The placebo evaluation in Equation 2 designates one of the three cached controls as the treatment. The candidate extraction process ranks control regions by how closely their mean intensity matches the lesion. Consequently, control 0 represents the optimal texture match and is not structurally exchangeable with the remaining candidates. Designating it as the treatment produces a comparison along the match-quality axis rather than a valid null. This exclusion follows directly from the extraction ordering rather than an empirical result. The empirical data align with this structural rule. Specifically, placebo$_0$ produces a significant signal on our development bed, whereas placebo$_1$ and placebo$_2$ do not. The two exchangeable designations ($j = 1, 2$) demonstrate strong agreement, justifying their use in the reported average.

Table 9: **Post-hoc attribution across map families and tiers.** Metrics are evaluated on the headline VinDr-CXR calcification bed (mean $\pm$ standard deviation over $k \in \{5, 25, 100\} \times 3$ seeds). The peak-in-GT chance line is 0.006. The "degenerate" label marks families provably reducing to a positive rescaling of the intrinsic map, matching to eight decimal places. No evaluated family localizes the lesion. Sensitivity at 1 FPPI remains near zero across all methods.

Tier	Attribution map	peak-in-GT	sens. @ 1 FPPI	
Attentive	attention (intrinsic)	0.0092 ± 0.0195	0.0004	
	Grad-CAM	0.0108 ± 0.0203	0.0022	
	integrated gradients	0.0123 ± 0.0236	0.0009	
LoRA	$w^\top \hat{f}_p$ (intrinsic)	0.0026 ± 0.0038	0.0000	
	Grad-CAM, token layer	0.0026 ± 0.0038	0.0000	degenerate
	Grad-CAM, block $6/12$	0.0162 ± 0.0190	0.0000	
	integrated gradients, pixels	0.0000 ± 0.0000	0.0000	

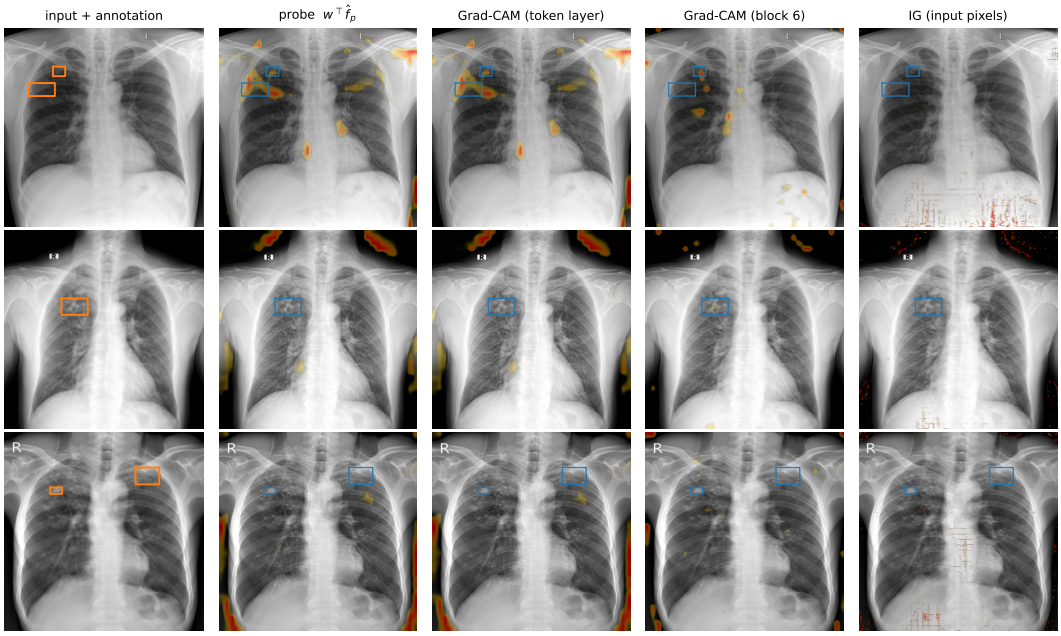


Figure 6: **Four attribution maps on a single LoRA adapter.** The model is evaluated at k=100 (seed 0) on the primary VinDr-CXR calcification bed. Orange outlines the annotated lesion on the input image. Blue repeats this outline across all attribution maps. Normalization between the 80th and 99.5th percentiles isolates the strongest attributions. The second and third columns are provably proportional and visually identical. The fourth and fifth columns are non-degenerate and visibly differ. No map places its mass on the actual lesion.

While randomizing the control ordering during extraction would provide a mathematically cleaner design, the empirical agreement between the exchangeable designations mitigates this limitation. Table 1 in the main text illustrates representative cells under the verdict rule. Table 10 details all 36 (bed, encoder, site) cells, and Table 33 provides the complete replication matrix.

D.2 Label-shuffled negative control

The placebo tests whether the model's response depends on the specific non-lesion region being occluded. A complementary requirement is to verify that the instrument does not artificially report lesion reliance for adapters lacking the relevant finding. We therefore audit adapters trained under a strict s=0 null: the k labels are permuted before fitting, with the recipe, seeds ($k \in \{5, 25, 100\}$, 10 seeds), cached regions, and evaluation otherwise identical. As a second off-target condition we audit

Table 10: **Placebo control across all 36 cells** on the public beds from Table 4. Values represent the partial Spearman correlation (signal → external AUC | internal AUC, $\log k$) evaluated exclusively on public sites.

Task bed	Encoder	Site	evidence	placebo	gap
Calcification (VinDr)	DINOv2-S	ChestX-Det10	-0.404	$+0.515$	$\mathbf{-0.453}$
	DINOv2-S	PadChest-GR	$+0.023$	-0.237	$\mathbf{+0.048}$
	MedSigLIP	ChestX-Det10	-0.034	-0.147	$\mathbf{+0.009}$
	MedSigLIP	PadChest-GR	$+0.261$	$+0.115$	$\mathbf{+0.215}$
	RAD-DINO	ChestX-Det10	$+0.092$	-0.051	$\mathbf{+0.100}$
	RAD-DINO	PadChest-GR	$+0.226$	$+0.438$	$\mathbf{+0.140}$
	ResNet-50	ChestX-Det10	$+0.343$	-0.279	$\mathbf{+0.391}$
	ResNet-50	PadChest-GR	$+0.227$	$+0.211$	$\mathbf{+0.160}$
	XrayCLIP-B/16	ChestX-Det10	-0.149	$+0.170$	$\mathbf{-0.191}$
	XrayCLIP-B/16	PadChest-GR	$+0.371$	-0.116	$\mathbf{+0.432}$
	XrayCLIP-L/14	ChestX-Det10	-0.021	$+0.159$	$\mathbf{-0.050}$
	XrayCLIP-L/14	PadChest-GR	-0.023	-0.209	$\mathbf{+0.025}$
Nodule	MedSigLIP	CheXpert	$+0.390$	-0.192	$\mathbf{+0.430}$
	MedSigLIP	NIH	$+0.379$	-0.089	$\mathbf{+0.422}$
	MedSigLIP	VinDr	$+0.503$	-0.259	$\mathbf{+0.552}$
	RAD-DINO	CheXpert	-0.023	$+0.016$	$\mathbf{-0.036}$
	RAD-DINO	NIH	-0.277	$+0.208$	$\mathbf{-0.308}$
	RAD-DINO	VinDr	$+0.148$	$+0.048$	$\mathbf{+0.134}$
Effusion	MedSigLIP	CheXpert	$+0.106$	$+0.213$	$\mathbf{+0.021}$
	MedSigLIP	NIH	$+0.134$	$+0.146$	$\mathbf{+0.095}$
	MedSigLIP	VinDr	$+0.010$	$+0.360$	$\mathbf{-0.091}$
	RAD-DINO	CheXpert	-0.194	-0.077	$\mathbf{-0.176}$
	RAD-DINO	NIH	-0.259	-0.220	$\mathbf{-0.202}$
	RAD-DINO	VinDr	-0.052	-0.062	$\mathbf{-0.041}$
Atelectasis	MedSigLIP	CheXpert	$+0.129$	-0.190	$\mathbf{+0.204}$
	MedSigLIP	NIH	-0.013	$+0.106$	$\mathbf{-0.031}$
	MedSigLIP	VinDr	-0.073	-0.137	$\mathbf{+0.023}$
	RAD-DINO	CheXpert	-0.145	-0.181	$\mathbf{-0.140}$
	RAD-DINO	NIH	-0.227	-0.252	$\mathbf{-0.218}$
	RAD-DINO	VinDr	$+0.155$	$+0.129$	$\mathbf{+0.139}$
Consolidation	MedSigLIP	CheXpert	$+0.083$	-0.182	$\mathbf{+0.214}$
	MedSigLIP	NIH	$+0.087$	-0.300	$\mathbf{+0.247}$
	MedSigLIP	VinDr	$+0.019$	-0.307	$\mathbf{+0.178}$
	RAD-DINO	CheXpert	-0.110	$+0.147$	$\mathbf{-0.112}$
	RAD-DINO	NIH	-0.080	$+0.033$	$\mathbf{-0.039}$
	RAD-DINO	VinDr	-0.084	-0.107	$\mathbf{+0.002}$

adapters trained on *other* findings (effusion, nodule, atelectasis, consolidation, SIIM pneumothorax) against the calcification regions (Table 11). The reproduction of the trained condition under this harness matches the reported values (MedSigLIP win-rate 0.756 against 0.753 in Table 34), which validates the pipeline.

Table 11: **Label-shuffled negative control on the primary bed.** Gap = evidence − placebo, averaged over budgets and seeds. Under permuted labels the gap is at most 0.019 in magnitude on every encoder, so the instrument fabricates no lesion reliance from texture mismatch or occlusion artifacts alone. Adapters trained on other findings can register non-zero gaps on calcification regions without any classification transfer (calcification AUC 0.35–0.62), and the sign of that response is encoder-specific.

Encoder	trained (gap)	label-shuffled (gap)	other-finding adapters (gap range)
RAD-DINO	-0.007	$+0.009$	$+0.081$ to $+0.159$
MedSigLIP	$+0.250$	$+0.013$	-0.028 to $+0.131$
XrayCLIP-B/16	$+0.270$	$+0.002$	-0.241 to $+0.086$
XrayCLIP-L/14	$+0.345$	-0.019	-0.076 to $+0.260$
DINOv2-S	$+0.034$	-0.002	-0.024 to $+0.054$
ResNet-50	$+0.093$	-0.008	$+0.001$ to $+0.154$

These evaluations yield two primary conclusions. First, the label-shuffled null remains clean across all six encoders. The ordering of the trained gaps successfully reproduces the ordering of the win-rates reported in the main text. Second, off-target adapters can exhibit dependence on the annotated

region's content. The sign of this dependence varies by encoder, confirming it is not a generic procedural artifact. The win-rate therefore measures the causal reliance of a specific model on the annotated region's visual content, rather than semantic reliance on the clinical concept. This supports the assertion that causal reliance and external transfer represent distinct axes. The placebo channel remains completely flat across all evaluated conditions (0.48–0.54).

D.3 WHY THE GAP IMPROVES ON THE RAW EVIDENCE

Fix a specific evaluation cell and let e, q, and y denote the evidence, the placebo, and the external AUC across all runs. Each variable is residualized on internal AUC and $\log k$, as detailed in Appendix F.1. We define $r_e = \mathrm{corr}(e, y)$, $r_q = \mathrm{corr}(q, y)$, $r_g = \mathrm{corr}(e - q,\ y)$, and $\rho = \mathrm{sd}(q)/\mathrm{sd}(e)$. All correlations denote Pearson coefficients computed on the residualized scores.

Proposition 3 (What differencing buys)**.** *Assume* $\mathrm{sd}(e) > 0$ *and* $\mathrm{sd}(q) > 0$*. If* $\mathrm{corr}(e, q) = 0$ *then*

$$r_g \;=\; \frac{r_e - \rho\, r_q}{\sqrt{1 + \rho^2}},$$

and consequently

(i) $r_g > r_e \iff r_q < -\, r_e\, \dfrac{\sqrt{1+\rho^2} - 1}{\rho}$*;*

(ii) if $r_q = 0$ *and* $r_e > 0$ *then* $r_g = r_e/\sqrt{1+\rho^2} < r_e$*;*

(iii) the threshold in (i) is strictly negative for $r_e > 0$ *and tends to* 0 *as* $r_e \to 0$*.*

Proof. $\mathrm{Cov}(e - q, y) = \mathrm{sd}(y)\,\mathrm{sd}(e)\,(r_e - \rho\, r_q)$ and, by the vanishing correlation, $\mathrm{Var}(e - q) = \mathrm{Var}(e) + \mathrm{Var}(q) = \mathrm{sd}(e)^2(1 + \rho^2)$; dividing gives the display. For (i), multiply out and divide by $-\rho < 0$. (ii) is the display at $r_q = 0$, and (iii) is immediate. □

The critical insight lies in condition (ii). Subtracting the placebo inherently inflates the variance denominator. A placebo lacking information regarding transfer will strictly degrade the gap compared to the raw evidence. Therefore, this subtraction constitutes a statistical penalty that only improves the correlation if the placebo itself correlates negatively with transfer.

Empirical measurements validate this vanishing-correlation hypothesis. Across the 36 evaluated public cells, the within-cell rank correlation between e and q exhibits a median of -0.08 and is positive in only 10 cells. The variance-minimizing coefficient $\mathrm{Cov}(e, q)/\mathrm{Var}(q)$ has a median of -0.04. This explicitly contradicts the assumption that the two signals share a positive nuisance term. This independence reflects their construction. The evidence metric contrasts the lesion site against control sites, while the placebo contrasts one control site against the remaining controls. The median ρ is 1.01, inflating the denominator by approximately $\sqrt{2}$. Because the raw correlations r_e remain small (mean $+0.035$), the theoretical threshold from condition (iii) sits at a median of -0.012.

	cells	$r_g > r_e$	mean $r_g - r_e$
$r_q \le 0$	20	20/20	$+0.075$
$r_q > 0$	16	2/16	-0.052

Proposition 3 effectively reproduces the empirically measured r_g with a rank correlation of 0.84 and a mean absolute error of 0.086. The theoretical threshold accurately predicts the sign of $r_g - r_e$ in 92% of the cells. Because this threshold closely approximates zero, the simplified rule $r_q \le 0$ achieves 94% accuracy. The reporting screen defined in Definition 1 strictly evaluates the statistical significance of r_q rather than its raw sign. This screen successfully removes the six cells where the gap performs worst and retains 30 cells. The gap improves the correlation in 22 of these retained cells. While strictly filtering by the sign of r_q would align these partitions perfectly, it requires discarding 10 additional cells based on statistically insignificant fluctuations. We prioritize the significance-based rule to maximize retained data while discarding demonstrably spurious cells.

D.4 WHY WE DO NOT TEST THE WIN-RATE AGAINST 0.5

The win-rate is formally defined as $\mathbb{E}_i[\mathbf{1}[\Delta_{les}^{(i)} > \overline{\Delta}_{ctl}^{(i)}]]$. If Δ_{les} and the three Δ_{ctl_j} were identically exchangeable, each pairwise comparison $\Delta_{les} > \Delta_{ctl_j}$ would yield a probability of exactly 0.5. However, comparing against their mean is a different event. For exchangeable variables $X_1, \ldots, X_4$, the event $\Pr[X_1 > \frac{1}{3}(X_2+X_3+X_4)]$ is mathematically equivalent to $\Pr[X_1 > \bar{X}]$. This probability departs from $\frac{1}{2}$ whenever the underlying delta distribution is skewed. Empirical occlusion deltas exhibit significant skew across every evaluated task bed. We therefore treat 0.5 only as a descriptive reference. The same skew makes the evidence and placebo *levels* structurally incommensurable: the evidence compares one delta against the mean of three, while each placebo designation compares one control against the mean of the remaining two (Eq. (2)). Recomputing the placebo with a three-way comparator matched to the evidence shifts its level by 0.007–0.025 (mean 0.017 across the six encoders) — an order of magnitude below the evidence effects we interpret, and immaterial to Definition 1, which consumes the placebo only through correlations, where a level offset cancels. Every comparative claim in this paper differences the identical fixed regions across two adapters, ensuring this baseline reference mathematically cancels. The placebo from Section 2.2 supplies the rigorous empirical null for all absolute evidence levels.

D.5 HOW WELL ARE CONTROL REGIONS TEXTURE-MATCHED?

Proposition 1 assumes control regions accurately match the lesion's local statistics. While the proposed implementation matches region shape exactly, it only screens candidates based on local mean intensity (Eq. (4)). We quantify the residual mismatch in second-order statistics directly on the source images. Evaluated metrics include within-region mean, standard deviation, Sobel gradient magnitude, and Laplacian energy. Table 30 reports the difference between the lesion and the average of its controls as a paired Cohen's d. On the primary calcification bed, lesion regions systematically carry higher edge energy than their controls ($d = +0.80$). This outcome is structurally expected, as calcifications represent inherently bright and compact abnormalities.

Three structural factors mitigate the impact of this residual mismatch. First, all comparative metrics difference identical fixed regions across adapters, neutralizing any baseline region-intrinsic bias. Second, audit verdicts remain robust across diverse occlusion fills (Appendix D.7). Third, the placebo control directly tests for and isolates this residual artifact. Absolute win-rate levels must nonetheless be interpreted relative to this specific evaluation instrument.

D.6 SENSITIVITY TO CONTROL SELECTION

The core audit relies on fixed hyperparameters governing the candidate pool size, intensity matching criterion, and anatomical placement constraints. We resampled all controls for the 130 annotated test positives of the primary bed under every combination the choices admit: pool size $\in \{18, 20, 60, 120\}$ valid candidates, matching by intensity alone versus intensity plus Sobel edge energy, and placement anywhere in the central frame versus fully inside a body mask. Each configuration re-extracted the control deltas from scratch and underwent evaluation by identical λ=0 probes ($k \in \{5, 25, 100\} \times 10$ seeds). Table 12 reports k=25.

Table 12: **Win-rate under re-drawn controls at k=25 (130 audit images, mean over 10 seeds).** *cached* is the paper's original control set; the remaining columns re-draw controls from scratch with the stated pool size N, matching criterion, and placement constraint.

	cached	N=18	N=20	N=60	N=120	N=60, +edge	N=60, in-body
MedSigLIP	0.792	0.773	0.785	0.780	0.785	0.790	0.776
XrayCLIP-L/14	0.882	0.872	0.874	0.876	0.882	0.885	0.871
XrayCLIP-B/16	0.817	0.680	0.678	0.692	0.688	0.710	0.683
ResNet-50	0.618	0.636	0.646	0.627	0.608	0.628	0.633
RAD-DINO	0.507	0.434	0.468	0.435	0.483	0.494	0.455
DINOv2-S	0.546	0.410	0.389	0.407	0.386	0.395	0.406

These evaluations yield two primary conclusions. First, the specific selection hyperparameters exert minimal influence on the final verdicts. Across the diverse configurations, the within-encoder

spread ranges strictly between 0.015 and 0.060. Expanding the candidate pool sevenfold or enforcing anatomical bounding leaves the core audit conclusions unchanged. Second, fully resampling the control locations successfully reproduces the causal stability of strong-signal cells (MedSigLIP -0.012, XrayCLIP-L/14 -0.006 at the matched configuration) while shifting absolute win-rates in weaker cells. This highlights the instrument-relative nature of absolute win-rates (Appendix I) and confirms that cross-cell comparisons do not rely on exact chance-level magnitudes.

D.7 Sensitivity to the occlusion strategy

Verifying Proposition 1 requires assessing whether the audit verdicts depend on the specific occlusion methodology. We re-evaluated the paired intervention utilizing two alternative localized fills: heavy Gaussian blur (σ=15) and ring-matched Gaussian noise. Table 13 contrasts these results against the default ring-mean fill across the full probe evaluation on the primary bed.

When the causal signal is genuinely strong, the win-rate remains invariant to the fill method. Models such as MedSigLIP and XrayCLIP-L/14 deviate by a maximum of 0.061 across all three occlusions. Conversely, encoders operating at chance exhibit extreme sensitivity to the occlusion strategy. RAD-DINO and DINOv2-S demonstrate deviations exceeding 0.220. Consequently, fill invariance acts as a secondary diagnostic separating grounded representations from spurious ones. Both robustness checks—the lesion-free placebo and the structural change of the occlusion operator—agree on five of the six evaluated encoders. This test cannot exclude long-range contextual artifacts shared across all localized fills, which remains a stated limitation of occlusion-based testing.

Table 13: **Fill invariance at k=100.** Intervention win-rates evaluated under three distinct occlusion fills on the primary VinDr-CXR calcification bed. Encoders possessing strong causal signals remain stable across fill methods. Encoders operating at chance demonstrate extreme sensitivity to the chosen occlusion strategy.

Encoder	ring-mean	blur	noise	max deviation
RAD-DINO	0.549	0.795	0.610	0.247
MedSigLIP	0.857	0.819	0.796	0.061
XrayCLIP-B/16	0.890	0.845	0.897	0.045
XrayCLIP-L/14	0.900	0.858	0.883	0.041
DINOv2-S	0.512	0.423	0.741	0.229
ResNet-50	0.559	0.632	0.625	0.073

E What the frozen representation supports

E.1 Three frozen routes to localizing the finding

Establishing whether the frozen representation inherently localizes the target finding is a prerequisite to auditing adapted probes. We evaluate three zero-shot localization strategies on the primary calcification dataset. None of these strategies utilize the adapted probe (Table 14).

(a) Prototype detection is representation-bound. A 5-shot dense prototype matcher applied to patch features yields near-zero detection performance for five of the six evaluated encoders. Sensitivity at 1 FPPI remains exactly 0.000 for these models. RAD-DINO is the sole encoder producing a measurable signal, achieving a mean sensitivity of 0.216 and a peak-in-GT of 0.228. However, this performance exhibits extreme variance across support draws. One seed reaches a sensitivity of 0.649, while the remaining two collapse to 0.000. This behavior explicitly inverts the classification ranking reported in Section 3, where CLIP-style architectures demonstrated superior transfer. This indicates that dense localization and linear separability depend on fundamentally different structural properties within the frozen feature space.

(b) The bottleneck is localization, not delineation. Supplying a ground-truth bounding box as a prompt enables SAM (Kirillov et al., 2023) to recover the lesion with near-perfect accuracy (sensitivity 0.994, peak-in-GT 0.923). MedSAM (Ma et al., 2024) performs similarly well (0.799 and 0.762, respectively). However, prompting with only a center point causes both models to collapse to 0.000 sensitivity. The primary bottleneck is therefore initial localization rather than downstream

Table 14: **Three frozen routes to localization on VinDr-CXR calcification test positives.** (a) 5-shot dense prototype detection over patch features (mean over three support draws). (b) SAM oracle prompted with a ground-truth box versus a center point. (c) Dense patch-text prompting. Sens@1 denotes sensitivity at 1 false positive per image; the peak-in-GT chance line is 0.006.

(a) 5-shot prototype detection, frozen encoders		
Encoder	sens. @ 1 FPPI	peak-in-GT
RAD-DINO	0.216	0.228
MedSigLIP	0.000	0.110
XrayCLIP-B/16	0.000	0.062
XrayCLIP-L/14	0.000	0.041
DINOv2-S	0.000	0.056
ResNet-50	0.000	0.023
(b) SAM oracle — what the prompt supplies		
Model (prompt)	sens. @ 1 FPPI	peak-in-GT
sam_base (box)	0.994	0.923
sam_base (center)	0.000	0.515
medsam (box)	0.799	0.762
medsam (center)	0.000	0.585
(c) dense patch–text prompting		
Model	sens. @ 1 FPPI	peak-in-GT
xrayclip	0.000	0.000
medsiglip	0.208	0.554

segmentation capacity. The frozen features evaluated in the previous step simply fail to provide this necessary localization prior.

(c) Text prompting does not substitute. Dense patch-text matching yields a score of 0.000 across both metrics for XrayCLIP. While MedSigLIP achieves better performance (sensitivity 0.208, peak-in-GT 0.554), this improvement concentrates exclusively on massive aortic calcifications implicitly favored by the bounding box annotations. This outcome does not provide evidence that text prompting resolves small-lesion localization deficits. The results remain consistent across all three evaluated routes: frozen representations support linear classification long before they support dense localization. CCA explicitly addresses this specific representational gap.

E.2 Recall is representation-bound: resolution and operating point.

Appendix E.1 demonstrates that dense prototype detection yields near-zero performance for five of the six frozen encoders. Confirming this as a structural deficit requires eliminating two potential confounders. The first confounder is tokenization resolution. Because calcifications are small, the standard 37^2 patch grid might lack the spatial capacity to resolve them. The second confounder is the decision operating point. Usable recall might exist exclusively at highly relaxed false-positive rates.

Resolution does not create the recall. RAD-DINO natively interpolates position embeddings. This property allows us to re-encode the full image at higher resolutions of 742 and 1036 (grids of 53^2 and 74^2) while preserving global context. As shown in Table 15, sensitivity at 1 FPPI does not improve with higher resolution. Instead, both sensitivity and peak-in-GT exhibit a downward trend. A 2×2 tiling variant achieves the identical 74^2 token count by sacrificing global context, yet it yields a similarly flat sensitivity of 0.206. Both upsampling strategies fail to improve detection, eliminating tokenization limits as the primary bottleneck.

Relaxing the operating point does not recover recall. Table 16 evaluates the dense prototype detector across a highly relaxed range of false positives per image (FPPI $\in \{0.5, 1, 2, 4, 8\}$). RAD-DINO improves steadily through FPPI 4 before saturating. Conversely, the remaining five encoders stay exactly at 0.000 through FPPI 2 and fail to achieve usable recall even at the clinically impractical threshold of FPPI 8. No downstream verification or threshold adjustment can recover this missing recall because the candidate feature set inherently lacks the necessary signal.

Table 15: **Resolution axis** on the primary VinDr-CXR calcification bed (5-shot dense prototype, mean over three support draws). Two distinct routes to a high-density 74^2 grid—context-preserving re-encoding and tiling—both fail to improve detection beyond the 37^2 baseline.

Encoding	grid	sens. @ 1 FPPI	peak-in-GT
full view, 518 (baseline)	37^2	0.216	0.228
full view, 742	53^2	0.210	0.203
full view, 1036	74^2	0.199	0.190
2×2 tiling, 518/tile	74^2	0.206	0.215

Table 16: **Operating-point axis** on VinDr-CXR calcification: sensitivity of the 5-shot dense prototype detector at increasing false positives per image. RAD-DINO saturates by FPPI 4; the remaining five encoders yield negligible signal at any clinically viable operating point.

Encoder	FPPI 0.5	FPPI 1.0	FPPI 2.0	FPPI 4.0	FPPI 8.0
RAD-DINO	0.000	0.216	0.346	0.370	0.370
MedSigLIP	0.000	0.000	0.000	0.009	0.067
XrayCLIP-B/16	0.000	0.000	0.000	0.000	0.139
XrayCLIP-L/14	0.000	0.000	0.000	0.000	0.087
DINOv2-S	0.000	0.000	0.000	0.000	0.160
ResNet-50	0.000	0.000	0.000	0.011	0.097

Collectively, these experiments show that recall for this finding family is bound by *what the frozen representation encodes*, rather than by resolution, thresholding, or segmentation capacity. This finding verifies the fundamental premise established in the main text.

E.3 Universal in-context segmentation is organ-tuned

Prior evaluations establish that frozen features support classification long before they support spatial localization. A potential counter-argument is that purpose-built universal in-context segmentation models could resolve this limitation. We test this directly, and the empirical results sharpen our diagnosis: the underlying bottleneck is not segmentation capacity, but rather the strict spatial-alignment assumption these models require.

We evaluate UniverSeg (Butoi et al., 2023) against a training-free RAD-DINO single-prototype matcher utilizing an identical 5-shot protocol. This evaluation spans one organ task, Montgomery lung-field segmentation (Jaeger et al., 2014), alongside three lesion tasks (Table 17). The organ task provides fixed anatomical positioning and consistent shape, guaranteeing spatial alignment. In contrast, the lesion tasks feature highly unpredictable spatial distributions across bounding boxes, pixel masks, and distinct imaging modalities.

Spatial alignment drives performance. On the organ task, UniverSeg achieves a Dice score of 0.931, cleanly outperforming the prototype, and proves robust to support selection. However, on the lesion tasks it collapses: 0.135 (SIIM pneumothorax), 0.129 (VinDr calcification), and 0.001 (DDR microaneurysms). Strikingly, the relative performance ranking between the two methods explicitly inverts on both chest lesion datasets, with the prototype achieving 0.242 and 0.333. The alignment-free prototype underperforms in structurally aligned contexts but dominates when spatial predictability breaks down. This isolation of the spatial-alignment assumption identifies it as the operative point of failure. Finally, the DDR dataset serves as a strict performance floor, yielding near-zero scores for both methods.

E.4 Substrate inversion

The encoder exhibiting the strongest classification performance is not necessarily the best localizer. On the primary dataset, CLIP-style models lead in classification. XrayCLIP-L/14 achieves an AUC of 0.741 and MedSigLIP reaches 0.736, while RAD-DINO trails at 0.666. However, the frozen dense-detection diagnostic from Appendix E.1 completely inverts this ranking. RAD-DINO is the only encoder to exhibit a measurable prototype-detection signal, whereas every CLIP tower

Table 17: **Organ versus lesion under identical 5-shot in-context protocols.** The organ task utilizes Montgomery lung fields. Lesion tasks include SIIM-ACR pneumothorax, VinDr-CXR calcification, and DDR fundus microaneurysms. Metrics denote Dice at the optimal threshold and peak-in-GT, evaluated with both area-matched and randomly drawn support sets. The dense in-context model leads on the organ task but loses to an alignment-free prototype on both chest lesion beds.

		matched support		random support	
Task	Model	Dice	peak-in-GT	Dice	peak-in-GT
Lung field (Montgomery)	UniverSeg	0.931	0.983	0.906	0.983
	rad-DINO proto	0.796	1.000	0.776	1.000
Pneumothorax (SIIM)	UniverSeg	0.135	0.183	0.179	0.250
	rad-DINO proto	0.242	0.267	0.201	0.300
Calcification (VinDr)	UniverSeg	0.129	0.083	0.072	0.167
	rad-DINO proto	0.333	0.500	0.212	0.317
Microaneurysm (DDR)	UniverSeg	0.001	0.000	0.001	0.000
	rad-DINO proto	0.001	0.000	0.000	0.000

collapses to exactly 0.000. The linear separability of a pooled representation and the spatial correspondence of its constituent patch tokens rely on distinct structural properties. Consequently, practitioners cannot reliably select foundation models for localization-dependent tasks by simply extrapolating from standard classification benchmarks.

F Statistics, power, and selection

F.1 Dependence-aware statistics

The partial correlations in Section 3 pool $k \times$ seed runs that share the identical training pool, test set, and cached occlusion regions. Consequently, naive run-level p-values artificially inflate the effective sample size. We re-analyzed every public (bed, encoder, site) cell using two dependence-aware procedures evaluated directly on the frozen probes. The first procedure is a seed-block permutation test within the Freedman-Lane family (Freedman & Lane, 1983). We residualize both the evidence and external AUC on an intercept, internal AUC, and $\log k$. The test statistic is the Spearman correlation of these calculated residuals. The null distribution is generated by permuting the evidence residuals in whole-seed blocks while leaving the outcome residuals fixed over 10^4 iterations. The reported p-value counts the observed statistic, $p = (1 + \#\{|r^{(b)}| \geq |r|\})/(B + 1)$, so its smallest attainable value is 10^{-4}. The second procedure employs a hierarchical bootstrap that resamples seed blocks with replacement and refits the residualization per draw over 2000 iterations.

Two additional test statistics appear throughout the evaluation. The gate-0 equivalence evaluates a two one-sided test (TOST) at level α on the population paired AUC difference θ. Given a defined margin ε, statistical equivalence serves as the alternative hypothesis rather than the null:

$$H_0 : |\theta| \geq \varepsilon \qquad \text{vs.} \qquad H_1 : |\theta| < \varepsilon, \qquad \theta = \mathbb{E}\big[\mathrm{AUC}_{\text{probe}} - \mathrm{AUC}_{\text{g0b}}\big], \qquad \varepsilon \in \{0.03, 0.05\}. \tag{5}$$

The scope-condition audit of Section 3 compares the significant-cell *rate* between the domain group $g{=}\mathrm{dom}$ and the natural-image group $g{=}\mathrm{nat}$:

$$\pi_g \;=\; \frac{1}{|g|} \sum_{c \in g} \mathbf{1}\big[\, p_c < \alpha \;\wedge\; r_c > 0 \,\big], \qquad T \;=\; \pi_{\mathrm{dom}} - \pi_{\mathrm{nat}}, \tag{6}$$

with p_c from the seed-block permutation test above. The null distribution of T is obtained by permuting the group label over cells ($B{=}2{\times}10^4$), which preserves the observed dependence within each cell but assumes cells are exchangeable across the two encoder groups; cells sharing a bed or site are not independent, so the test is a reference comparison rather than an exact statistical proof.

Cell counts used in this paper. A cell refers to a unique (bed, encoder, site) triplet unless otherwise specified. All denominators explicitly exclude in-house evaluation sites. The overall cell counts decompose functionally based on the specific evaluation. The gate-0 equivalence test utilizes 48

cells, representing 8 beds evaluated across 6 encoders. The placebo audit evaluates 36 cells. This includes 12 cells from the primary bed and 24 cells from the PadChest-GR beds, restricted to the two encoders with cached per-run placebo values. The scope audit analyzes 102 cells. This subset covers 6 encoders across 17 distinct external site pairs derived from the six replication beds. The primary calcification bed is held out of its own hypothesis test, and the fundus bed is excluded to maintain a strict four-domain versus two-natural encoder grouping. The comprehensive statistical re-analysis in Table 18 covers 132 public cells. Finally, the evidence-gated selection analysis in Appendix F.5 indexes 180 distinct configurations by incorporating the label budget directly into the cell definition.

Applying these dependence-aware procedures leaves the statistical verdicts essentially unchanged. Out of 132 public cells, exactly one cell flips significance between the naive and permutation tests (Table 18). We therefore report strict permutation p-values throughout the text. The scope-condition audit utilizes this identical procedure. This confirms that its negative results stem directly from the underlying data distribution rather than from an underpowered statistical test.

Table 18: **Naive versus dependence-aware significance** on the public external sites (Table 4). The "flipped" column tallies cells where the significance verdict at $\alpha = 0.05$ diverges between the two tests.

Task bed	cells	sig. (naive)	sig. (permutation)	flipped	median $\lvert r \rvert$
Calcification	12	6	6	0	0.188
Nodule	18	11	10	1	0.280
Effusion	18	7	7	0	0.171
Atelectasis	18	9	9	0	0.208
Consolidation	18	5	5	0	0.122
Ptx (SIIM)	12	5	5	0	0.177
Ptx (MS-CXR)	18	9	9	0	0.211
Fundus MA	18	12	12	0	0.229
Pooled	132	64	63	**1**	0.186

Encoder-level exact test for the scope audit. The 102-cell group-label permutation of Section 3 treats cells as exchangeable, but the 17 cells sharing an encoder share seeds and adapters. As a conservative check we aggregate to the encoder level, which absorbs all within-encoder dependence, and compare the domain and natural-image groups on the per-encoder significant-cell fraction and mean r. With two natural-image encoders among six there are $\binom{6}{2} = 15$ assignments, so the permutation is exact. Per-encoder significant-cell fractions are 0.294 (MedSigLIP), 0.000 (RAD-DINO), 0.647 (XrayCLIP-B/16), 0.471 (XrayCLIP-L/14), 0.529 (DINOv2-S), and 0.235 (ResNet-50). The domain-minus-natural difference is -0.029 in significant-cell rate (exact two-sided $p = 0.93$, null range $[-0.338, +0.368]$) and -0.017 in mean r ($p = 0.87$, null range $[-0.128, +0.153]$), with the observed statistic at the center of the null range in both cases. The conclusion of the cell-level test is unchanged.

F.2 How much external data does the applicability check need?

Section 4 establishes that the correlation between evidence and external transfer must be validated empirically for each novel model deployment. This requirement raises a critical practical question regarding the volume of externally labeled data necessary for reliable validation.

To evaluate this requirement, we resample the external validation sets at sample sizes of $n \in \{25, 50, 100, 200\}$ while rigorously preserving the underlying disease prevalence. We subsequently recompute the partial correlations to record the fraction of resamples that successfully recover a statistically significant effect (Table 19). The median re-detection rate across all public cells reaches only 0.38 at a sample size of 100 and 0.41 at a sample size of 200. While this constitutes a resampling diagnostic conditional on the observed effect size rather than a formal power analysis against a specified alternative, the magnitudes indicate a severe practical limitation. Post-hoc evidence-gated model selection is highly data-intensive and cannot be treated as a cost-free operation.

Table 19: **Power of the applicability check.** The table reports the median fraction of resamples successfully recovering a significant effect, computed across all public cells (Table 4) and specifically across cells exhibiting a strong correlation ($|r| \geq 0.45$).

External sample n	median power (all cells)	median power ($|r| \geq 0.45$)
25	0.33	0.61
50	0.35	0.77
100	0.38	0.89
200	0.41	0.99

F.3 Scope-condition audit by bed

Table 20 details the 102-cell scope audit originally presented in Section 3, stratified by task bed. The nodule detection task represents the only dataset where domain-specific encoders demonstrate clear dominance. Across every other evaluated bed, natural-image controls achieve significant-cell rates equal to or exceeding those of the domain-pretrained models.

Table 20: **Scope-condition audit** over six public beds, six encoders, and public external sites (Table 4). Cells with a significant positive evidence–transfer link under seed-block permutation ($B = 10^4$, $\alpha = 0.05$), scored on public sites only. If the correlation required a domain-appropriate representation the right-hand columns would sit near zero; they do not.

Task bed	Domain encoders (4)		Natural-image controls (2)	
	cells	sig. positive	cells	sig. positive
Nodule	12	8 (67%)	6	0 (0%)
Pneumothorax (SIIM)	8	3 (38%)	4	2 (50%)
Atelectasis	12	5 (42%)	6	3 (50%)
Pneumothorax (MS-CXR)	12	5 (42%)	6	4 (67%)
Effusion	12	2 (17%)	6	1 (17%)
Consolidation	12	1 (8%)	6	3 (50%)
Pooled	68	**24 (35%)**	34	**13 (38%)**

F.4 A second attempt at the scope condition: taxonomy-controlled acquisition shift

Section 3 explicitly rejects encoder provenance as a reliable predictor for the evidence-transfer link. Before concluding that these activation conditions defy general characterization, we evaluate one final hypothesis regarding label taxonomies. Mismatched clinical definitions present a major confounding factor across external evaluation sites. Bounding boxes derived from grounded reports, natural language extractions from CheXpert, and reader consensus from VinDr-CXR do not define radiological findings identically. A null correlation result could theoretically imply either a genuine lack of causal linkage or simply that the target labels measure fundamentally different clinical constructs. The PadChest-GR dataset provides a natural experiment to eliminate this semantic confound. The dataset originates from a single hospital utilizing a unified clinical taxonomy, yet it aggregates images from two distinct hardware acquisition devices. By restricting the training data to one manufacturer and reserving the other exclusively as the external validation site, we create an evaluation cell where the label system remains perfectly fixed. This setup successfully isolates hardware acquisition shift from semantic shift. If taxonomy alignment were the primary driver of the evidence-transfer link, the measured correlations should robustly survive this dataset split.

The empirical results completely invalidate this hypothesis (Table 21). Across six evaluated cells, the outcomes fragment unpredictably. MedSigLIP demonstrates a significant positive link for nodule detection ($+0.514$, $p_{\text{perm}} = 10^{-4}$) and a weak positive link for effusion. Conversely, RAD-DINO yields strictly null results for both of these findings. Most crucially, both encoders exhibit significant negative correlations on the atelectasis task. An effect that reverses sign across different findings under an identical label system cannot serve as a reliable scope condition. This taxonomy-controlled evaluation confirms the fundamental unpredictability reported in the main text while explicitly elim-

Table 21: **Taxonomy-controlled acquisition shift.** Partial Spearman correlation (evidence $\rightarrow$ external AUC | internal AUC, $\log k$) where the training bed is restricted to one PadChest-GR manufacturer and the other serves as the external site. This holds the label taxonomy identical across both sides. The 95% CI reflects a hierarchical bootstrap over seeds; p_{perm} derives from the seed-block permutation test (Appendix F.1). The correlation sign is highly unstable across findings.

Finding	Encoder	r_{partial}	95% CI	p_{perm}	verdict
Nodule	MedSigLIP	+0.514	[+0.251, +0.734]	0.0001	positive
	RAD-DINO	+0.145	[-0.098, +0.360]	0.1979	null
Effusion	MedSigLIP	+0.204	[-0.013, +0.348]	0.0282	positive
	RAD-DINO	-0.107	[-0.302, +0.071]	0.2948	null
Atelectasis	MedSigLIP	-0.335	[-0.539, -0.114]	0.0007	*negative*
	RAD-DINO	-0.247	[-0.495, +0.021]	0.0042	*negative*

inating semantic mismatch as a potential justification. We report this secondary failure to reinforce our central methodological argument. The applicability of the causal grounding metric must be actively measured for each specific deployment cell rather than passively predicted from dataset properties.

F.5 Evidence-gated selection: two signals, both negative

Selecting among candidate adapters based on their grounded evidence scores represents an intuitive, training-free application of the proposed audit. However, this strategy completely fails to improve generalization on public datasets. We transparently report this failure as it constitutes the primary practical takeaway of the scope-condition analysis.

The proposed selection heuristic filters the candidate pool to the top half based on internal AUC and subsequently deploys the single model maximizing grounded evidence. The baseline strategy simply deploys the model achieving the maximum internal AUC. We evaluate statistical significance by resampling the candidate pool over 2000 iterations. We test two distinct gating signals: the raw intervention win-rate and the placebo-controlled gap. The latter evaluates the hypothesis that the raw score suffers from contamination in placebo-loud cells. Neither signal succeeds (Table 22). Averaged across 180 configurations on five public beds, the mean external AUC change is exactly -0.007 for both methods. Zero configurations achieve statistical significance under either selection rule. The gap metric beats the raw score in exactly 93 out of 180 cases, representing a statistical coin flip. The measurement instrument itself remains statistically sound. The failure of the selection rule stems entirely from the inherent unpredictability of the observational link between grounding and transfer.

Table 22: **Evidence-gated selection** across five public beds (VinDr-CXR calcification and four PadChest-GR findings). Values indicate the external AUC change of the evidence-deployed model versus the internal-AUC baseline across two selection signals. Both signals fail to outperform the baseline.

		select by evidence		select by gap	
Task bed	cells	mean Δ	sig. $+$	mean Δ	sig. $+$
Calcification (VinDr)	60	-0.0010	0	-0.0039	0
Nodule	30	-0.0042	0	+0.0000	0
Effusion	30	-0.0141	0	-0.0177	0
Atelectasis	30	-0.0122	0	-0.0106	0
Consolidation	30	-0.0076	0	-0.0055	0
Pooled	180	**-0.0067**	0	**-0.0069**	0

F.6 CCA PER EXTERNAL SITE

Table 23 tallies the specific public external sites that respond positively to the CCA constraint. Table 24 details the numerical results for the underlying cells. For every evaluated triplet of task bed, encoder, and public site, we strictly pair the unconstrained baseline and the CCA run utilizing the identical random seed and label budget. We report the mean paired difference alongside a 95% seed-block bootstrap confidence interval. Statistical significance is determined using a seed-block sign-permutation test. In-house evaluation sites are strictly excluded throughout this analysis. This exclusion reduces the available site count per bed to a maximum of three compared to the five originally available for datasets like PadChest-GR.

Table 23: **CCA replicates across external sites.** The table reports significant positive $\lambda{=}0 \rightarrow 1$ cells over total cells per bed and encoder, scored exclusively on individual public external sites rather than site averages (seed-block bootstrap CI and seed-block sign-permutation test, $\alpha = 0.05$). Bold text denotes a bed-encoder pair where every public site improves. Table 24 lists all 64 cells with intervals.

Bed	sites	RAD-DINO	MedSigLIP	XrayCLIP-B/16	XrayCLIP-L/14	DINOv2-S	ResNet-50	RETFound	all
Calcification (VinDr)	2	**2/2**	1/2	1/2	**2/2**	0/2	0/2	—	6/12
Nodule	3	0/3	**3/3**	—	—	—	—	—	3/6
Effusion	3	0/3	2/3	—	—	—	—	—	2/6
Atelectasis	3	0/3	2/3	—	—	—	—	—	2/6
Consolidation	3	0/3	0/3	—	—	—	—	—	0/6
Pneumothorax (SIIM)	2	0/2	1/2	—	—	—	—	—	1/4
Pneumothorax (MS-CXR)	3	1/3	**3/3**	—	—	—	—	—	4/6
Fundus MA (DDR)	3	2/3	**3/3**	2/3	—	**3/3**	**3/3**	**3/3**	16/18
All beds	8	5/22	15/22	3/5	2/2	3/5	3/5	3/3	**34/64**

F.7 WHY WE DO NOT PROPOSE TRAINING-FREE SELECTION

Section 4 explicitly declines to position the audit as a training-free model selection rule. The logical rationale is straightforward. Deploying the most grounded adapter among those clearing an internal accuracy threshold remains sound only if evidence strongly predicts transfer for that specific configuration. The scope-condition audit conclusively demonstrates that this predictive link cannot be reliably inferred from encoder provenance or taxonomy alignment. Validating the link requires direct measurement on externally labeled data. This requirement reintroduces the exact annotation burden the few-shot methodology originally sought to bypass. Substituting the placebo-controlled gap does not resolve this fundamental limitation, as both gating signals fail to consistently outperform a standard internal-AUC selection rule (Appendix F.5). CCA explicitly avoids this vulnerability. The efficacy of the training constraint is measured directly against an optimizer-matched control within the specific deployment cell. This approach enforces grounded behavior without relying on brittle observational correlations established in external datasets.

G TRAINING FOR GROUNDING

G.1 SUPERVISION-MATCHED MASK BASELINES

Table 25 details the per-encoder values summarized in Section 4. Two distinct patterns emerge from this comparison. Saliency suppression achieves the best worst-case performance ($+0.000$) simply because it barely affects any encoder, yielding marginal changes between $+0.000$ and $+0.004$. Among the five baselines that actively modify the adapter, CCA presents the least damaging worst-case bound at -0.008. This significantly outperforms the alternative methods, which degrade performance by -0.017 to -0.041. The lesion-occluded-negative baseline shares the occlusion mechanism of CCA but omits the matched control region. While it outperforms CCA on XrayCLIP-L/14 ($+0.044$ versus $+0.034$), its performance wildly fluctuates between $+0.044$ and -0.041 across different encoders. Without the paired control, this constraint artificially rewards global occlusion sensitivity whenever the inherent lesion signal is weak.

Two baselines further demonstrate that the audit metric can be artificially optimized without any accompanying transfer. Lesion-crop augmentation raises ResNet-50's win-rate from 0.558 to 0.729 while *lowering* its external AUC by 0.025. The lesion-occluded-negative baseline artificially inflates

Table 24: **CCA at λ=1 per public external site.** Bold text highlights cells exhibiting a statistically significant positive effect.

Bed / encoder	site	λ=0	λ=1	Δ	95% CI
Calcification (VinDr) / RAD-DINO	ChestX-Det10	0.530	0.537	$\mathbf{+0.007}^{**}$	[+0.005, +0.008]
	PadChest-GR	0.579	0.604	$\mathbf{+0.025}^{**}$	[+0.020, +0.030]
Calcification (VinDr) / MedSigLIP	ChestX-Det10	0.603	0.601	-0.002	[-0.007, +0.002]
	PadChest-GR	0.584	0.678	$\mathbf{+0.094}^{**}$	[+0.083, +0.104]
Calcification (VinDr) / XrayCLIP-B/16	ChestX-Det10	0.607	0.605	-0.002	[-0.008, +0.003]
	PadChest-GR	0.583	0.645	$\mathbf{+0.062}^{**}$	[+0.051, +0.073]
Calcification (VinDr) / XrayCLIP-L/14	ChestX-Det10	0.630	0.650	$\mathbf{+0.021}^{**}$	[+0.018, +0.023]
	PadChest-GR	0.653	0.700	$\mathbf{+0.047}^{**}$	[+0.039, +0.055]
Calcification (VinDr) / DINOv2-S	ChestX-Det10	0.517	0.518	$+0.001$	[-0.002, +0.004]
	PadChest-GR	0.558	0.542	-0.016^{**}	[-0.021, -0.011]
Calcification (VinDr) / ResNet-50	ChestX-Det10	0.528	0.523	-0.005^{**}	[-0.007, -0.003]
	PadChest-GR	0.565	0.556	-0.009^{*}	[-0.014, -0.004]
Nodule / RAD-DINO	CheXpert	0.628	0.604	-0.024^{**}	[-0.027, -0.020]
	NIH	0.572	0.560	-0.011^{**}	[-0.014, -0.009]
	VinDr	0.642	0.627	-0.015^{**}	[-0.020, -0.011]
Nodule / MedSigLIP	CheXpert	0.664	0.713	$\mathbf{+0.049}^{**}$	[+0.038, +0.059]
	NIH	0.638	0.675	$\mathbf{+0.037}^{**}$	[+0.033, +0.041]
	VinDr	0.753	0.813	$\mathbf{+0.059}^{**}$	[+0.051, +0.068]
Effusion / RAD-DINO	CheXpert	0.777	0.766	-0.012^{**}	[-0.013, -0.009]
	NIH	0.709	0.695	-0.015^{**}	[-0.016, -0.013]
	VinDr	0.768	0.751	-0.017^{**}	[-0.020, -0.015]
Effusion / MedSigLIP	CheXpert	0.899	0.904	$\mathbf{+0.005}^{*}$	[+0.001, +0.009]
	NIH	0.791	0.811	$\mathbf{+0.020}^{**}$	[+0.017, +0.023]
	VinDr	0.842	0.835	-0.007^{*}	[-0.012, -0.003]
Atelectasis / RAD-DINO	CheXpert	0.710	0.697	-0.013^{**}	[-0.015, -0.010]
	NIH	0.635	0.629	-0.006^{**}	[-0.008, -0.005]
	VinDr	0.638	0.623	-0.015^{**}	[-0.019, -0.011]
Atelectasis / MedSigLIP	CheXpert	0.795	0.808	$\mathbf{+0.013}^{**}$	[+0.007, +0.019]
	NIH	0.734	0.743	$\mathbf{+0.008}^{*}$	[+0.004, +0.012]
	VinDr	0.838	0.819	-0.018^{*}	[-0.031, -0.007]
Consolidation / RAD-DINO	CheXpert	0.788	0.779	-0.009^{**}	[-0.011, -0.006]
	NIH	0.708	0.706	-0.002^{*}	[-0.003, -0.001]
	VinDr	0.780	0.761	-0.019^{**}	[-0.023, -0.015]
Consolidation / MedSigLIP	CheXpert	0.906	0.907	$+0.001$	[-0.002, +0.003]
	NIH	0.763	0.762	-0.000	[-0.001, +0.001]
	VinDr	0.931	0.919	-0.012^{**}	[-0.015, -0.008]
Pneumothorax (SIIM) / RAD-DINO	NIH	0.798	0.799	$+0.001$	[-0.002, +0.004]
	VinDr	0.766	0.767	$+0.001$	[-0.003, +0.005]
Pneumothorax (SIIM) / MedSigLIP	NIH	0.813	0.813	$+0.000$	[-0.004, +0.004]
	VinDr	0.790	0.811	$\mathbf{+0.020}^{**}$	[+0.013, +0.027]
Pneumothorax (MS-CXR) / RAD-DINO	CheXpert	0.709	0.710	$+0.001$	[-0.000, +0.003]
	NIH	0.740	0.739	-0.001	[-0.003, +0.001]
	VinDr	0.776	0.781	$\mathbf{+0.006}^{**}$	[+0.004, +0.008]
Pneumothorax (MS-CXR) / MedSigLIP	CheXpert	0.677	0.713	$\mathbf{+0.036}^{**}$	[+0.028, +0.043]
	NIH	0.721	0.780	$\mathbf{+0.059}^{**}$	[+0.048, +0.070]
	VinDr	0.771	0.829	$\mathbf{+0.058}^{**}$	[+0.046, +0.070]
Fundus MA (DDR) / RAD-DINO	APTOS	0.598	0.656	$\mathbf{+0.057}^{**}$	[+0.049, +0.065]
	IDRiD	0.769	0.783	$\mathbf{+0.014}^{**}$	[+0.007, +0.023]
	Messidor-2	0.474	0.468	-0.006^{*}	[-0.011, -0.001]
Fundus MA (DDR) / MedSigLIP	APTOS	0.997	0.998	$\mathbf{+0.001}^{**}$	[+0.000, +0.002]
	IDRiD	0.951	0.953	$\mathbf{+0.002}^{*}$	[+0.000, +0.004]
	Messidor-2	0.818	0.828	$\mathbf{+0.010}^{**}$	[+0.007, +0.013]
Fundus MA (DDR) / XrayCLIP-B/16	APTOS	0.879	0.887	$\mathbf{+0.008}^{*}$	[+0.003, +0.013]
	IDRiD	0.829	0.857	$\mathbf{+0.028}^{**}$	[+0.020, +0.036]
	Messidor-2	0.665	0.660	-0.005^{*}	[-0.007, -0.002]
Fundus MA (DDR) / DINOv2-S	APTOS	0.895	0.918	$\mathbf{+0.023}^{**}$	[+0.017, +0.030]
	IDRiD	0.895	0.913	$\mathbf{+0.019}^{**}$	[+0.013, +0.025]
	Messidor-2	0.684	0.697	$\mathbf{+0.014}^{**}$	[+0.008, +0.020]
Fundus MA (DDR) / ResNet-50	APTOS	0.825	0.863	$\mathbf{+0.038}^{**}$	[+0.032, +0.044]
	IDRiD	0.838	0.862	$\mathbf{+0.023}^{**}$	[+0.018, +0.030]
	Messidor-2	0.610	0.646	$\mathbf{+0.035}^{**}$	[+0.032, +0.039]
Fundus MA (DDR) / RETFound	APTOS	0.625	0.630	$\mathbf{+0.005}^{**}$	[+0.004, +0.007]
	IDRiD	0.632	0.638	$\mathbf{+0.006}^{**}$	[+0.005, +0.008]
	Messidor-2	0.618	0.623	$\mathbf{+0.004}^{**}$	[+0.003, +0.005]

XrayCLIP-L/14's win-rate to 0.941. A grounding score serves strictly as a diagnostic instrument rather than a direct optimization target. Reporting this metric without an external generalization endpoint inevitably invites this exact failure mode.

Table 25: **Same masks: CCA and six baseline utilizations.** The table reports external AUC change at λ=1 against the optimizer-matched λ=0 control per encoder on the primary VinDr-CXR calcification bed. Every row receives identical k lesion annotations, differing only in how the loss function consumes them. Bold text highlights the counterfactual margin (CCA). CCA yields the best mean and the best worst-case bounds, though it is not uniformly superior.

Mask used as	RAD-DINO	MedSigLIP	XrCLIP-B	XrCLIP-L	DINOv2	RN50	mean	worst	best-λ mean
counterfactual margin (CCA)	**+0.016**	**+0.046**	**+0.030**	**+0.034**	-0.008	-0.007	+0.018	-0.008	—
lesion-occluded as negative	-0.036	+0.032	+0.024	+0.044	-0.041	-0.035	-0.002	-0.041	+0.010
lesion-crop augmentation	+0.016	+0.021	-0.012	-0.030	-0.003	-0.025	-0.006	-0.030	-0.001
ROI-pooled features	+0.016	+0.021	-0.008	-0.017	+0.000	-0.019	-0.001	-0.019	+0.002
saliency suppression (RRR)	+0.004	+0.004	+0.004	+0.004	+0.003	+0.000	+0.003	+0.000	+0.004
contribution penalty (CDEP)	+0.022	+0.005	-0.017	-0.011	-0.001	-0.000	-0.000	-0.017	+0.008
attribution prior, region-conditioned	+0.026	-0.034	-0.021	-0.020	-0.006	+0.003	-0.009	-0.034	+0.010

G.2 THE TWO STRONGEST MASK BASELINES

To ensure a rigorous evaluation, we must rule out the possibility that CCA merely outperforms a weak baseline. We therefore extend our comparison beyond the standard RRR (right-for-the-right-reasons) method. We implemented two substantially stronger mechanisms from the explanation-constrained literature. Both baselines are evaluated under identical supervision, optimizers, random seeds, and λ grids.

CDEP (Rieger et al., 2020) penalizes the contextual-decomposition contribution of a designated region rather than its input gradient. For a mean-pooled linear head, this decomposition is mathematically exact. The output logit evaluates to $\frac{1}{N}\sum_p w^\top f_p + b$. The contribution originating outside the target annotation R is explicitly available in closed form. We penalize its share of the total contribution magnitude according to Eq. (7):

$$\Omega_{\mathrm{CDEP}}(w) \;=\; \mathbb{E}_i\left[\frac{\sum_{p\notin R^{(i)}} \left|w^\top f_p^{(i)}\right|}{\sum_p \left|w^\top f_p^{(i)}\right|}\right], \qquad \mathcal{L} = \mathrm{CE} + \lambda\,\Omega_{\mathrm{CDEP}}. \tag{7}$$

The absolute value is essential. A one-sided penalty on the outside contribution, $[\,w^\top f_p]_+$ summed over $p \notin R$, is identically zero at the optimum in this setting and leaves the objective unchanged; we verified this before adopting Eq. (7). This formulation provides a strictly stronger constraint than a standard saliency penalty due to architectural structures. Because $\partial\,\mathrm{logit}/\partial f_p = w/N$ does not depend on p at this tier, a penalty on the input gradient outside the mask carries no positional information: it can rescale w, but it cannot move where the model looks. This structural limitation explains why the RRR baseline in Table 35 remains inert rather than merely under-tuned. It also dictates why our implemented strong baselines explicitly penalize the activation $w^\top f_p$, which correctly varies across patches. While the original medical evaluation of CDEP derived penalized regions programmatically, we supply the algorithm with expert masks. This provides the baseline with strictly more supervision than its original formulation assumed.

Axiomatic attribution priors (Erion et al., 2021) natively regularize specific properties of a path-integrated attribution map, such as smoothness or sparsity, without requiring explicit region annotations. Because there is no published region-conditioned variant of this method, our implementation represents a custom adaptation. This formulation penalizes path-integrated attribution mass falling outside the target annotation. We utilize the mean patch of the k training negatives as the integration reference $\bar{f}^-$, formulated in Eq. (8):

$$a_p^{(i)}(w) = w^\top\big(f_p^{(i)} - \bar{f}^-\big), \qquad \Omega_{\mathrm{prior}}(w) \;=\; \mathbb{E}_i\left[\frac{\sum_{p\notin R^{(i)}} \left|a_p^{(i)}(w)\right|}{\sum_p \left|a_p^{(i)}(w)\right|}\right]. \tag{8}$$

Note that Eq. (7) represents the special case where $\bar{f}^-$=0. For comparison, the standard RRR baseline suppresses positive evidence outside the annotation using $\Omega_{\mathrm{RRR}}(w) = \mathbb{E}_i\big[\,\mathrm{mean}_{p\notin R^{(i)}}[\,w^\top f_p^{(i)}]_+\big]$. Both of these strong baselines yield net non-positive external gains at the protocol-matched λ=1 setting, averaging -0.000 and -0.009. Furthermore, they degrade monotonically with increasing λ across four of the six evaluated encoders. The mechanism is visible in the physical magnitudes: the annotated lesion occupies merely 2% of the patch grid on this

bed (0.6% of pixels, rasterized to whole patches). Forcing the majority of attribution mass into such a tightly constrained area imposes an excessively harsh penalty on a representation inherently unequipped to support dense localization. Permitting an oracle selection of the optimal λ directly on the evaluation sites marginally lifts the mean gains to $+0.008$ for CDEP and $+0.010$ for the attribution prior. Both remain inferior to CCA's $+0.018$ gain achieved at a fixed, a priori setting. To verify that this oracle bound does not stem from an overly aggressive starting grid, we explicitly extended the attribution prior evaluation downward to $\lambda \in \{0.01, 0.03, 0.1, 0.3, 1, 3\}$. The optimum selected the smallest possible value for only two encoders, yielding a net zero gain in both cases.

Only one specific cell inverts this performance ordering. On RAD-DINO, the attribution prior reaches $+0.026$ and CDEP reaches $+0.022$, outperforming CCA's $+0.016$. RAD-DINO is the sole encoder whose absolute grounding on this dataset rests exactly at chance. This identifies the precise cell where the proposed audit flags zero latent grounding capacity. The success of alternative regularization mechanisms in this specific regime directly supports our core argument. CCA explicitly constrains grounded behavior. Consequently, it succeeds exclusively where grounded features are structurally available within the frozen representation. When such features are absent, any observed transfer gains must derive from alternative statistical sources. We explicitly do not position CCA as the optimal intervention for fully ungrounded representations.

G.3 CCA UNDER NOISY TRAINING MASKS

Because CCA consumes region annotations, evaluating its sensitivity to annotation quality is critical. We repeat the λ=1 versus λ=0 comparison while degrading the k training masks in two directions: morphologically dilated (over-inclusive, mimicking a hurried annotator) and eroded (under-inclusive). All other experimental variables remain completely identical. The final evaluation consistently utilizes exact ground-truth masks. This experimental design cleanly isolates the specific impact of training signal degradation.

The performance gains robustly survive this degradation, as detailed in Table 26. Across the four CXR-domain encoders, the external AUC gain remains strictly positive under both dilated masks ($+0.009$ to $+0.052$) and eroded masks ($+0.025$ to $+0.051$). This compares favorably against the $+0.024$ to $+0.055$ range achieved with exact masks. While the lower bound of the dilated condition drops slightly, it never changes sign. Furthermore, the performance of MedSigLIP deviates by less than 0.005 across all three conditions. Both natural-image control encoders elevate their causal win-rates without achieving any external AUC gains in every evaluated condition. This exactly mirrors the dissociation previously observed in the standard mask baselines. Practically, the CCA constraint merely requires the underlying lesion to be roughly positioned where the mask indicates. It does not strictly demand precise pixel-level delineation. This robustness directly validates the underlying mathematical mechanism. CCA optimizes a contrastive comparison between a region and a shape-matched neighbor, completely bypassing the need to penalize a sharp feature boundary.

Table 26: **CCA gain under degraded training masks** on the primary VinDr-CXR calcification bed. Each cell reports the paired difference between λ=1 and the optimizer-matched λ=0 control for both the intervention win-rate and external AUC (averaged over public sites). Evaluation masks remain consistently exact.

	exact		dilated		eroded	
Encoder	Δwin	Δext	Δwin	Δext	Δwin	Δext
RAD-DINO	+0.232	+0.024	+0.201	+0.009	+0.273	+0.025
MedSigLIP	+0.115	+0.055	+0.114	+0.052	+0.095	+0.051
XrayCLIP-B/16	+0.095	+0.034	+0.099	+0.037	+0.068	+0.025
XrayCLIP-L/14	+0.045	+0.045	+0.058	+0.040	+0.054	+0.043
DINOv2-S	+0.207	-0.007	+0.238	-0.016	+0.200	-0.002
ResNet-50	+0.107	-0.008	+0.146	-0.022	+0.056	+0.000

G.4 SENSITIVITY TO THE HINGE MARGIN

The main evaluation fixes the hinge margin at m=0.2. To evaluate sensitivity to this parameter, we execute a grid sweep across $m \in \{0, 0.05, 0.1, 0.5, 1.0\}$ while maintaining λ=1. We evaluate

this sweep under an identical experimental recipe encompassing varying label budgets, 20 random seeds, and all six encoders. Each run is explicitly compared to its corresponding optimizer-matched λ=0 baseline. Across the 36 evaluated configurations, the mean external-AUC gain remains strictly positive at every evaluated margin. The aggregate gains scale smoothly from $+0.003$ at m=0 to $+0.025$ at m=1.0. The reported baseline of m=0.2 therefore sits squarely in the middle of this operational range rather than representing an isolated optimum. The directional sign for each encoder remains remarkably stable over the entire 20-fold parameter span. On the PadChest-GR site at k=25, the four CXR-domain encoders scale their performance monotonically with the margin. Simultaneously, both natural-image control encoders remain within ± 0.01 of zero at every evaluated m. This confirms that the null transfer effect observed on those natural-image models is an inherent representational deficit, not an artifact of a poorly tuned margin. The intervention win-rate reliably increases with m before gracefully saturating. A minor exception occurs on the ChestX-Det10 dataset, where overall gains are structurally smaller. Here, MedSigLIP and XrayCLIP-B/16 exhibit slightly negative returns at the extreme m=1.0 setting. Consequently, we claim that the directional signs and performance orderings are highly robust to m. We explicitly avoid asserting that larger margins yield uniformly superior results across all deployments.

G.5 λ REGIMES: WHEN GROUNDING STARTS TO COST ACCURACY

The evaluation grid explicitly incorporates a λ=3 setting for every encoder. This identifies the specific regime where the training constraint begins to aggressively trade classification accuracy for causal grounding. Table 27 details the performance dynamics at k=25. The intervention win-rate rises monotonically with λ across all six evaluated encoders. The CCA constraint consistently achieves the exact behavioral change it explicitly prices. However, the downstream accuracy consequences heavily depend on the underlying foundation model. On the four CXR-domain encoders, both internal AUC and external transfer metrics remain stable or actively improve through the λ=3 setting. For example, XrayCLIP-L/14 experiences continuous gains, shifting from 0.687 to 0.727 internally and 0.649 to 0.720 externally. Conversely, the expected accuracy trade-off clearly materializes on the natural-image controls. ResNet-50 simultaneously loses internal AUC and external AUC as λ increases. DINOv2-S similarly degrades externally at the λ=3 boundary. These findings practically validate the decision rule established in Appendix A. The harmful optimization regime remains fully detectable without requiring external validation labels. The internal AUC of the constrained model predictably degrades relative to the λ=0 baseline exactly where the external generalization cost first emerges.

Table 27: λ regimes at k=25 (20 seeds): internal AUC, audit win-rate, and PadChest external AUC as λ grows. Bold marks the natural-image encoders, where large λ trades accuracy for grounding.

	Internal AUC			Win-rate			External AUC (PadChest)		
λ	0	1	3	0	1	3	0	1	3
MedSigLIP	0.674	0.690	0.680	0.763	0.901	0.909	0.581	0.688	0.697
XrayCLIP-L/14	0.687	0.716	0.727	0.872	0.923	0.928	0.649	0.701	0.720
XrayCLIP-B/16	0.666	0.669	0.672	0.816	0.913	0.938	0.585	0.645	0.681
RAD-DINO	0.614	0.627	0.631	0.542	0.783	0.787	0.576	0.612	0.615
DINOv2-S	0.593	0.608	0.606	0.526	0.740	0.815	0.549	0.545	0.527
ResNet-50	0.633	0.626	0.602	0.580	0.685	0.728	0.569	0.562	0.538

H REPLICATION AND ADDITIONAL RESULTS

H.1 MODALITY TRANSFER: A FUNDUS BED

To verify that the proposed protocol generalizes beyond chest radiography, we replicate it strictly on microaneurysm detection in retinal fundus photographs. We utilize the DDR dataset (Li et al., 2019) with pixel masks as the internal training bed. We evaluate transfer performance on IDRiD (Porwal et al., 2018), Messidor-2 (Decencière et al., 2014), and APTOS (Karthik et al., 2019). We include RETFound as the domain-specific encoder. The pipeline requires no modifications. We deploy the identical occlusion construction, control sampling, and k-sweep configurations.

The previously observed patterns reproduce consistently. MedSigLIP matches gate-0b exactly ($\Delta = +0.000$) at an AUC of 0.993. This marks the eighth consecutive evaluation bed demonstrating this statistical equivalence. The dissociation between performance and grounding emerges sharply: saliency localization remains at or below 0.03 across all six encoders, including those demonstrating high causal win-rates (0.91 for ResNet-50, 0.73 for RETFound). Furthermore, this dataset provides a clear example of a spurious shortcut. RAD-DINO, a chest-specific encoder processing retinal data, outperforms the centroid rule by $+0.087$. This represents the largest surplus observed in the study and constitutes one of the four cells failing gate-0 equivalence at the 0.03 margin. Simultaneously, its saliency explicitly remains at 0.00. The adapter successfully extracts information beyond first-order statistics, but it fundamentally fails to learn the target finding.

Figure 7 illustrates individual test cases from this dataset. Microaneurysms represent the smallest annotated targets in this study, comprising merely 0.02% of the image frame. They also represent the only targets permitting immediate pixel-verifiable ground truth. Figure 8 replicates the unconstrained-versus-constrained comparison of Figure 4 directly on these pixel-verifiable cases.

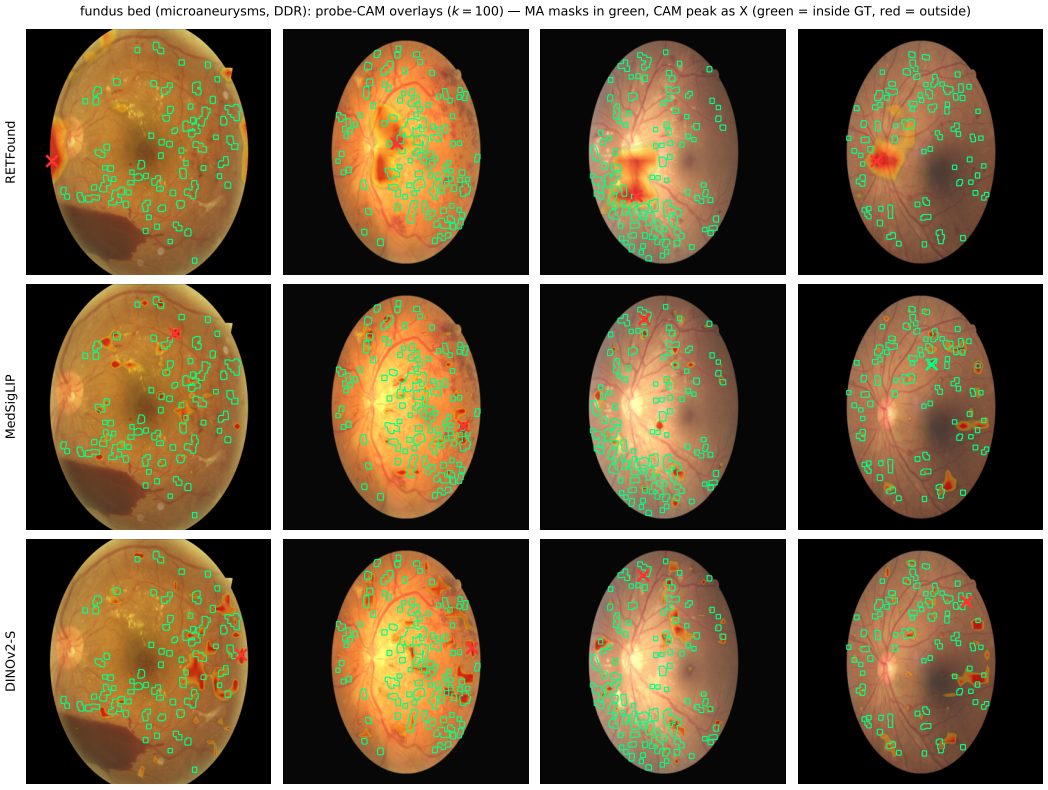


Figure 7: **Fundus bed, qualitative** (DDR retinal fundus photographs, k=100). Green denotes microaneurysm annotations. Each row displays an encoder's probe-CAM with its argmax explicitly marked (✓ inside an annotation, × outside). The targets are one to two orders of magnitude smaller than the chest lesions and number in the dozens per image. Yet, the evidence maps consistently concentrate on the optic disc and illumination gradients rather than any actual lesion.

H.2 Calibration and decision thresholds out of domain

Although AUC evaluates models independent of a threshold, clinical deployment strictly requires a defined operating point. Practitioners must select this threshold internally and transfer it to the external evaluation site. We fit Platt scaling on the internal test set and measure the 10-bin expected calibration error (ECE) both internally and externally. We then compare the F_1 score obtained by deploying the internally optimized threshold against the theoretical upper-bound F_1 achievable via explicit external re-optimization (Table 28).

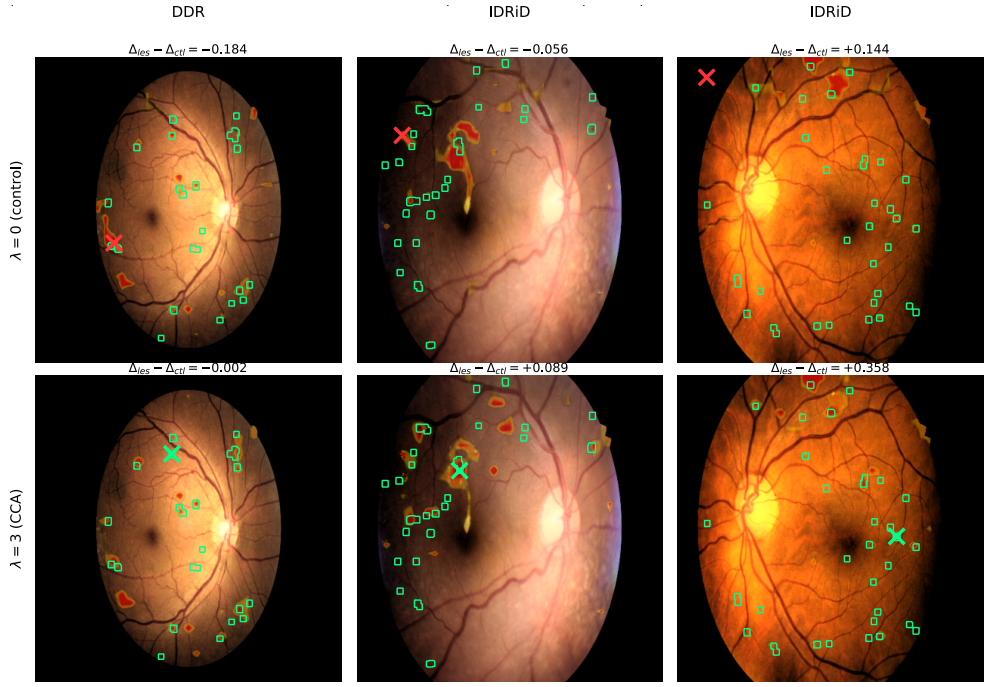


Figure 8: **What CCA changes on the fundus bed** for pixel-verifiable cases (DINOv2-S, k=25, identical seed per column; one DDR case and two IDRiD (Porwal et al., 2018) cases). Top: under the λ=0 control, the evidence peak misses every annotated microaneurysm (×). Bottom: Enforcing CCA (λ=3) shifts the peak inside a valid target (✓). The printed quantity reflects the paired contrast $\Delta_{les} - \overline{\Delta}_{ctl}$, which widens in every column ($-0.184 \rightarrow -0.002$, $-0.056 \rightarrow +0.089$, $+0.144 \rightarrow +0.358$). Notably, both individual deltas are strictly negative on this bed—occluding either region *raises* the score. CCA successfully constrains the difference, not the absolute magnitude of either term. Overlays shade the top 20% of evidence values, restricted exclusively to the retinal field of view.

Internal calibration performs excellently by construction. However, this calibration fails to survive domain shifts. External ECE deteriorates by an order of magnitude, and the internal encoder ranking collapses completely. Blindly transferring the internal threshold degrades the attainable F_1 score by a median of 24%. This performance cost depends heavily on the specific target site in ways that internal validation cannot predict. The loss averages 11% on ChestX-Det10 but increases to 33% on PadChest-GR. The worst-performing case, MedSigLIP, suffers a 41% penalty. Its attainable F_1 score drops from 0.076 to 0.045. Notably, MedSigLIP achieves the best internal calibration with an ECE of 0.000, yet it suffers the largest threshold-transfer loss. Consequently, internal calibration quality fails as a rank-preserving proxy for external threshold safety. This outcome provides a threshold-level confirmation of our central thesis. The internally optimized metric and the actual deployable quantity represent distinct properties. The operational gap between them remains strictly invisible from internal validation alone.

H.3 Qualitative Case Studies

Figure 9 visualizes individual test positives that decouple the two evidence channels. These visual examples reflect the identical taxonomy reported in the aggregate metrics. The first category includes cases that are causally grounded but saliency-blind. In these instances, the occlusion intervention drops the score, yet the attribution map peaks outside the lesion. The second category includes cases where saliency flatters a fundamentally ungrounded adapter. The map correctly targets the lesion, but the occlusion intervention yields no score change. The final category includes cases where both metrics logically agree. Figure 13 illustrates the paired occlusion construction across multiple evaluation beds. Figure 14 plots the per-image distribution of Δ_{les} against $\overline{\Delta}_{ctl}$.

Table 28: **Out-of-domain calibration and threshold transfer** on the primary VinDr-CXR calcification bed. F_1@int applies the internally optimized threshold to external data. F_1@ext re-optimizes the threshold directly on the external site, representing a theoretical upper bound unachievable by a practitioner lacking external labels.

Encoder	Site	ECE int	ECE ext	F_1@int	F_1@ext	loss
DINOv2-S	ChestX-Det10	0.006	0.189	0.107	0.115	+0.007
	PadChest-GR	0.006	0.231	0.041	0.065	+0.024
MedSigLIP	ChestX-Det10	0.000	0.189	0.127	0.163	+0.037
	PadChest-GR	0.000	0.226	0.045	0.076	+0.031
RAD-DINO	ChestX-Det10	0.018	0.192	0.107	0.118	+0.011
	PadChest-GR	0.018	0.229	0.044	0.064	+0.020
ResNet-50	ChestX-Det10	0.001	0.188	0.107	0.117	+0.010
	PadChest-GR	0.001	0.226	0.043	0.058	+0.015
XrayCLIP-B/16	ChestX-Det10	0.023	0.170	0.145	0.159	+0.014
	PadChest-GR	0.023	0.235	0.048	0.065	+0.018
XrayCLIP-L/14	ChestX-Det10	0.023	0.177	0.148	0.167	+0.018
	PadChest-GR	0.023	0.229	0.054	0.084	+0.030

The aggregate win-rate calculates a strict sign test on this paired difference rather than representing a simple difference of means.

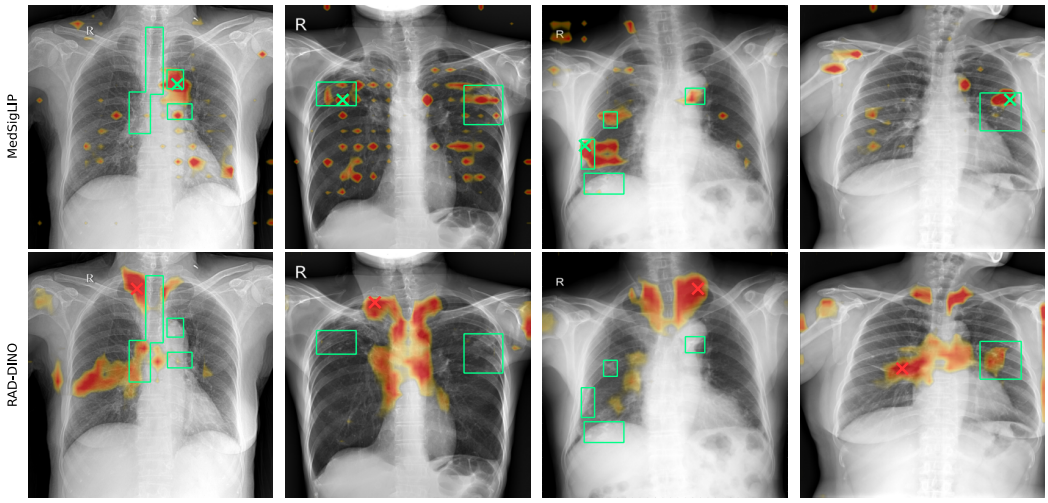


Figure 9: **Cases where the two evidence channels disagree** on the primary VinDr-CXR calcification bed. Ground truth regions are marked in green. The adapter's evidence map is overlaid with its argmax explicitly marked.

Figure 11 extends this visual comparison across all six encoders on identical test positives to isolate differences in saliency behavior. Only MedSigLIP consistently localizes the lesion. Conversely, the two XrayCLIP models consistently peak off-lesion, even when yielding the highest causal win-rates in the entire study.

Figures 4 and 12 demonstrate the localized effect of the CCA constraint. Each column represents a single test positive. The top row displays the optimizer-matched λ=0 control. The bottom row displays the CCA-constrained adapter, with the annotation in green and the evidence peak marked. Under the unconstrained baseline, the evidence peak falls completely outside the annotation. Under the CCA constraint, the peak reliably shifts inside. The printed paired deltas structurally explain this dynamic. Specifically, Δ_{les} significantly increases while Δ_{ctl} remains stable. This visualizes the precise per-case mechanics driving the aggregate win-rate shift. Notably, this mechanism functions successfully both on an encoder with high baseline grounding (MedSigLIP) and one starting exactly at chance (RAD-DINO).

Figure 10 explicitly categorizes the discrepancies between the two evaluation channels, detailing the measured deltas per panel. It clearly contrasts causally grounded but saliency-blind cases against fully ungrounded cases where neither channel detects the target lesion.

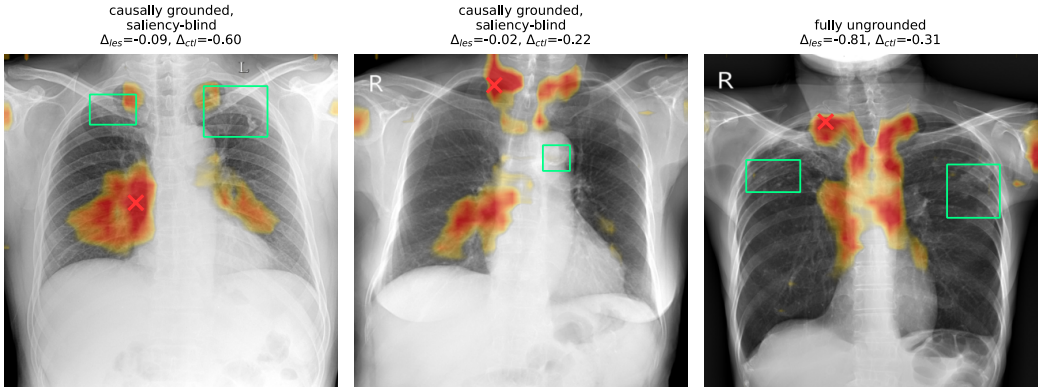


Figure 10: **Instrument-disagreement case studies** on VinDr-CXR calcification (RAD-DINO probe, k=100). The annotation is green, and the probe-CAM argmax is marked ($\times$ denotes a peak outside the annotation). Δ_{les} and Δ_{ctl} indicate the measured score drops when the lesion and its matched control are occluded. These panels directly contrast the behavior of the two evidence channels on a per-case basis.

H.4 Per-budget and per-cell detail

Figure 15 plots the complete per-encoder dissociation curves on the primary evaluation bed across every label budget. Figure 17 visualizes the evidence-transfer scatter for every evaluated configuration. Each point in the scatter plot represents a distinct trained probe. This visualization provides the raw distributions underlying the scope-condition partial correlations. It highlights the rare configurations where the causal relationship remains visible and explicitly contrasts them against the vast majority where the relationship fails.

Table 29 reports the corresponding numerical data for every configuration on the primary bed, extending the k=100 slice of Table 2 across the full budget sweep. It details internal AUC, the gate-0b null value, the intervention win-rate, CAM localization, and mean external AUC. Analyzing each encoder block synthesizes both the dissociation and the gate-0 equivalence into a single view. The gate-0b null consistently tracks internal AUC within a 0.011 margin at every budget across all evaluated encoders. However, the evidence metrics deviate independently from both accuracy measures. CAM localization remains firmly within the noise threshold of its chance line (0.006) for five of the six encoders across the entire budget sweep. MedSigLIP represents the sole exception, although its localization score behaves non-monotonically with respect to the label budget. The causal win-rate reliably increases for MedSigLIP and both XrayCLIP architectures. Conversely, the causal win-rate explicitly stalls at chance levels for RAD-DINO, DINOv2-S, and ResNet-50. This stagnation occurs over the exact budget range where their internal AUC scores increase by 0.10 or more. Classification accuracy, the prior-matched null, causal reliance, and static saliency effectively constitute four disconnected quantities.

Figure 16 collapses this data into a single value per encoder. The shaded background band denotes CXR-domain pretrained models. Domain pretraining fails to isolate positive correlation coefficients from negative ones. ResNet-50 represents a natural-image control, yet it yields positive correlations on both external sites. In contrast, DINOv2-S yields a strongly negative correlation on one site. This analysis strictly evaluates the primary test bed. The comprehensive 102-cell scope audit presented in Table 20 covers the six replication beds and strictly excludes this initial dataset.

Table 29: **Per-budget results across all six encoders** on the primary VinDr-CXR calcification bed. Metrics represent means calculated over 20 random seeds. External AUC is averaged uniformly across the bed's specific public external sites.

Encoder	k	AUC	gate-0b	Interv. win	CAM loc	ext. AUC
RAD-DINO	5	0.558	0.558	0.497	0.008	0.518
	10	0.571	0.570	0.443	0.011	0.531
	25	0.606	0.604	0.520	0.008	0.542
	50	0.646	0.643	0.567	0.022	0.565
	100	0.666	0.657	0.549	0.045	0.568
MedSigLIP	5	0.579	0.579	0.635	0.220	0.552
	10	0.631	0.631	0.680	0.351	0.574
	25	0.671	0.671	0.765	0.278	0.594
	50	0.703	0.703	0.834	0.558	0.621
	100	0.736	0.736	0.857	0.495	0.628
XrayCLIP-B/16	5	0.564	0.564	0.662	0.011	0.549
	10	0.613	0.613	0.700	0.006	0.572
	25	0.662	0.660	0.837	0.005	0.597
	50	0.686	0.685	0.855	0.005	0.624
	100	0.716	0.714	0.890	0.009	0.639
XrayCLIP-L/14	5	0.581	0.581	0.735	0.002	0.568
	10	0.627	0.627	0.768	0.006	0.591
	25	0.661	0.658	0.869	0.012	0.614
	50	0.702	0.696	0.877	0.006	0.657
	100	0.741	0.730	0.900	0.017	0.685
DINOv2-S	5	0.560	0.560	0.501	0.002	0.514
	10	0.586	0.586	0.476	0.006	0.531
	25	0.589	0.589	0.481	0.011	0.535
	50	0.628	0.628	0.558	0.014	0.550
	100	0.640	0.639	0.512	0.011	0.556
ResNet-50	5	0.564	0.564	0.483	0.009	0.517
	10	0.600	0.600	0.526	0.006	0.541
	25	0.630	0.630	0.558	0.012	0.548
	50	0.658	0.658	0.520	0.003	0.573
	100	0.664	0.663	0.559	0.005	0.562

H.5 Full result tables and figures

Figure 18 illustrates the parameter response curve supporting the fixed λ=1 evaluation utilized in Table 3. The performance gain increases monotonically up to λ=1 for CXR-domain encoders. Furthermore, the performance does not reverse or collapse at λ=3. This structural stability justifies the deployment of a single *a priori* λ parameter rather than relying on heavily optimized per-cell selections.

Several tables and figures referenced in the main text are deferred to this section due to space constraints. These include the gate-0 equivalence tests alongside bootstrap intervals (Table 32) and the comprehensive eight-bed replication matrix for RAD-DINO and MedSigLIP (Table 33). We additionally provide the detailed per-parameter and per-site CCA results (Table 34). Finally, Table 35 presents the direct head-to-head comparison evaluating strict saliency alignment against the proposed counterfactual margin optimization.

Figure 20 provides one test positive from each evaluation bed alongside its corresponding annotation and a shape-matched control region. This visualizes the extensive range of lesion scales and diverse annotation formats systematically evaluated in this study.

Figure 21 visualizes the actual spatial inputs consumed by the CCA constraint. It displays the annotated lesion region and its structurally matched control. These regions are explicitly differenced during training for the most successful encoder evaluated on each dataset bed.

Table 30: **Texture-match diagnostic** for the five box-annotated chest beds of Table 4. Values represent the paired lesion-minus-control effect size (Cohen's d, where 0 indicates a perfect match). The mean is matched by construction; Appendix D.5 reads off the consequences.

Task bed	mean	std	gradient	Laplacian
Calcification (VinDr)	-0.16	$+0.40$	$+0.80$	$+0.25$
Nodule	-0.28	-0.40	$+0.02$	-0.04
Effusion	$+0.52$	-0.20	$+0.07$	-0.12
Atelectasis	$+0.18$	-0.61	$+0.32$	-0.05
Consolidation	$+0.28$	-0.75	$+0.13$	$+0.06$

Table 31: **Gate-0 equivalence, summarized by bed.** Aggregate of Table 32. *equiv.* counts cells statistically equivalent to the centroid-difference rule within a 0.03 margin. The trained probe buys a median of $+0.010$ AUC over a rule that fits only a threshold. † The fundus bed's RETFound cell is equivalent and is reported in Appendix H.1; Table 32 lists the other 47 cells.

Task bed	cells	equiv. @ 0.03	median Δ	largest Δ (encoder)
Calcification (VinDr)	6	6	$+0.0015$	$+0.0107$ (XrayCLIP-L/14)
Nodule	6	6	$+0.0082$	$+0.0251$ (RAD-DINO)
Effusion	6	5	$+0.0133$	$+0.0293$ (RAD-DINO)
Atelectasis	6	6	$+0.0121$	$+0.0207$ (XrayCLIP-B/16)
Consolidation	6	6	$+0.0152$	$+0.0216$ (XrayCLIP-B/16)
Pneumothorax (SIIM)	6	6	$+0.0117$	$+0.0255$ (XrayCLIP-B/16)
Pneumothorax (MS-CXR)	6	5	$+0.0066$	$+0.0487$ (RAD-DINO)
Fundus MA (DDR)	$6^{\dagger}$	$4^{\dagger}$	$+0.0193$	$+0.0868$ (RAD-DINO)
Pooled	**48**	**44**	$+0.0100$	$+0.0868$ (RAD-DINO, fundus)

H.6 External calibration under CCA

We evaluate whether the CCA constraint alters probability calibration across different sites. Table 36 reports the 10-bin expected calibration error (ECE) of the raw sigmoid outputs. This evaluation covers both external sites of the primary bed at $k{=}25$ over 20 random seeds. It directly compares the unconstrained $\lambda{=}0$ baseline against the constrained $\lambda{=}1$ counterpart using an identical optimization recipe.

Two distinct patterns emerge from this evaluation. First, the absolute ECE remains large across all configurations, strictly ranging from 0.40 to 0.51. These error magnitudes are nearly identical across varying encoders and constraint weights. The adapters undergo class-balanced training, whereas the actual external disease prevalence ranges from 2.0% to 5.7%. Consequently, the raw output probabilities are entirely dominated by this severe prevalence shift regardless of the applied training objective. This shift accounts for the internal-external calibration drift noted in the main text. It also explains why clinical deployment requires explicit threshold or intercept recalibration under any selected λ. Second, the CCA constraint does not systematically degrade calibration quality. The ECE difference between the constrained model and the baseline stays strictly within a ± 0.03 margin on every evaluated pair. The directional sign of this minor difference also varies randomly. For example, DINOv2-S improves by -0.037 on PadChest, while XrayCLIP-B/16 degrades slightly by $+0.014$ on ChestX-Det10. Therefore, the observed grounding improvements are not purchased at the cost of probability calibration degradation.

I Limitations in full

Absolute win-rates must be interpreted as instrument-relative, because the controls match shape and mean intensity but not second-order texture (Appendix D.5). Furthermore, occlusion inherently constitutes an out-of-distribution operation. The paired control explicitly cancels region-generic artifacts, but not the act of occlusion itself. Consequently, long-range contextual artifacts shared across all local fills escape our fill-invariance check. Regarding supervision, CCA requires spatial region annotations for the k training positives. While our six mask-matched baselines isolate the

Table 32: **Gate-0 equivalence across the eight public beds** from Table 4. Values show the per-seed paired difference between the trained probe and the centroid-difference rule at the largest budget, accompanied by a bootstrap 95% confidence interval (B=10^4). The ✓ symbol marks statistical equivalence within a strict 0.03 margin. Section 3 formally discusses the four exceptions.

Task bed	Encoder	Δ (probe − gate-0b)	95% CI	equiv. @ 0.03
Calcification (VinDr)	RAD-DINO	+0.0083	[+0.0059, +0.0106]	✓
	MedSigLIP	−0.0000	[−0.0000, +0.0000]	✓
	XrayCLIP-B/16	+0.0025	[+0.0013, +0.0036]	✓
	XrayCLIP-L/14	+0.0107	[+0.0086, +0.0129]	✓
	DINOv2-S	+0.0004	[+0.0002, +0.0007]	✓
	ResNet-50	+0.0002	[+0.0001, +0.0004]	✓
Nodule	RAD-DINO	+0.0251	[+0.0218, +0.0285]	✓
	MedSigLIP	+0.0001	[+0.0000, +0.0001]	✓
	XrayCLIP-B/16	+0.0117	[+0.0094, +0.0140]	✓
	XrayCLIP-L/14	+0.0165	[+0.0139, +0.0192]	✓
	DINOv2-S	+0.0046	[+0.0030, +0.0062]	✓
	ResNet-50	+0.0000	[−0.0000, +0.0001]	✓
Effusion	RAD-DINO	+0.0293	[+0.0264, +0.0320]	×
	MedSigLIP	+0.0000	[+0.0000, +0.0000]	✓
	XrayCLIP-B/16	+0.0151	[+0.0144, +0.0158]	✓
	XrayCLIP-L/14	+0.0125	[+0.0119, +0.0130]	✓
	DINOv2-S	+0.0141	[+0.0131, +0.0152]	✓
	ResNet-50	+0.0004	[+0.0002, +0.0006]	✓
Atelectasis	RAD-DINO	+0.0185	[+0.0158, +0.0214]	✓
	MedSigLIP	+0.0000	[+0.0000, +0.0001]	✓
	XrayCLIP-B/16	+0.0207	[+0.0194, +0.0218]	✓
	XrayCLIP-L/14	+0.0181	[+0.0167, +0.0194]	✓
	DINOv2-S	+0.0060	[+0.0051, +0.0069]	✓
	ResNet-50	+0.0000	[+0.0000, +0.0001]	✓
Consolidation	RAD-DINO	+0.0140	[+0.0123, +0.0156]	✓
	MedSigLIP	+0.0000	[+0.0000, +0.0000]	✓
	XrayCLIP-B/16	+0.0216	[+0.0204, +0.0227]	✓
	XrayCLIP-L/14	+0.0209	[+0.0204, +0.0215]	✓
	DINOv2-S	+0.0163	[+0.0153, +0.0172]	✓
	ResNet-50	+0.0042	[+0.0036, +0.0048]	✓
Pneumothorax (SIIM)	RAD-DINO	+0.0134	[+0.0092, +0.0186]	✓
	MedSigLIP	+0.0000	[+0.0000, +0.0001]	✓
	XrayCLIP-B/16	+0.0255	[+0.0243, +0.0267]	✓
	XrayCLIP-L/14	+0.0214	[+0.0186, +0.0243]	✓
	DINOv2-S	+0.0100	[+0.0073, +0.0125]	✓
	ResNet-50	+0.0010	[+0.0004, +0.0018]	✓
Pneumothorax (MS-CXR)	RAD-DINO	+0.0487	[+0.0445, +0.0526]	×
	MedSigLIP	−0.0000	[−0.0000, +0.0000]	✓
	XrayCLIP-B/16	+0.0128	[+0.0119, +0.0136]	✓
	XrayCLIP-L/14	+0.0077	[+0.0069, +0.0085]	✓
	DINOv2-S	+0.0055	[+0.0049, +0.0061]	✓
	ResNet-50	+0.0003	[+0.0001, +0.0006]	✓
Fundus MA (DDR)	RAD-DINO	+0.0868	[+0.0673, +0.1062]	×
	MedSigLIP	+0.0002	[+0.0001, +0.0002]	✓
	XrayCLIP-B/16	+0.0034	[+0.0016, +0.0053]	✓
	DINOv2-S	+0.0415	[+0.0317, +0.0528]	×
	ResNet-50	+0.0193	[+0.0147, +0.0234]	✓

counterfactual margin, an exhaustive separation from every conceivable mask-using strategy remains open. Finally, our external validation cohorts are modest in size, and the evaluation over-indexes on chest radiography, relying on a single fundus bed to verify modality transfer.

Table 33: **Replication across the eight public task beds** from Table 4. Results show the largest budget per bed for RAD-DINO and MedSigLIP . Δg0b represents probe AUC minus gate-0b. CCA Δext denotes the external AUC change evaluated at a fixed a priori λ=1 against the optimizer-matched control, averaged across the bed's specific public sites.

Task bed	Encoder	AUC	Δg0b	Interv.	CAM	CCA Δext (λ=1)
Calcification (VinDr)	RAD-DINO	0.666	+0.008	0.55	0.04	+0.016
	MedSigLIP	0.736	-0.000	0.86	0.50	+0.046
Nodule	RAD-DINO	0.629	+0.025	0.56	0.02	-0.017
	MedSigLIP	0.724	+0.000	0.73	0.45	+0.048
Effusion	RAD-DINO	0.812	+0.029	0.53	0.07	-0.015
	MedSigLIP	0.882	+0.000	0.81	0.54	+0.006
Atelectasis	RAD-DINO	0.667	+0.019	0.81	0.03	-0.011
	MedSigLIP	0.769	+0.000	0.70	0.64	+0.001
Consolidation	RAD-DINO	0.826	+0.014	0.75	0.13	-0.010
	MedSigLIP	0.879	+0.000	0.83	0.57	-0.004
Pneumothorax (SIIM)	RAD-DINO	0.862	+0.013	0.79	0.05	+0.001
	MedSigLIP	0.828	+0.000	0.77	0.08	+0.010
Pneumothorax (MS-CXR)	RAD-DINO	0.816	+0.049	0.45	0.19	+0.002
	MedSigLIP	0.910	-0.000	0.77	0.23	+0.051
Fundus MA (DDR)	RAD-DINO	0.746	+0.087	0.71	0.00	+0.022
	MedSigLIP	0.993	+0.000	0.71	0.03	+0.004

J EXTENDED RELATED WORK

Grounding audits of fixed medical models. CheXlocalize evaluates saliency methods against expert segmentations, reporting that seven attribution methods fall well short of a human benchmark, particularly on small and irregular lesions (Saporta et al.). XBench scales this evaluation to seven medical vision-language models across 36 findings, documenting the static zero-shot version of our core phenomenon: calcification remains among the findings models successfully *classify* without actually *grounding* (Luo et al., 2026). Related audits probe multimodal LLM localization (Gosai et al., 2025), classify-and-ground pipelines (Lian et al., 2026), and attribution faithfulness (Zain et al., 2026). All these works evaluate a *fixed*, static checkpoint. Our study asks a different question. We do not ask "Is this fixed model grounded?" but rather "Do the performance *gains* from few-label adaptation stem from grounded features" The object of our study is the adaptation trajectory along the label budget, not a static snapshot.

Causal and occlusion-based evaluation. Standard deletion/insertion metrics inherently suffer from an out-of-distribution confound: masking pixels creates artificial images unlike any training data, conflating causal reliance with simple network damage. Recent medical literature has shifted toward controlled interventions. These include masking annotated targets against equal-area irrelevant regions to compute causal grounding rates (Lotfinia et al., 2026), comparing attention against occlusion importance across LoRA configurations (Sadanandan & Behzadan, 2026), masking unexpected regions (Sourget et al., 2026), and treating explanations as falsifiable hypotheses under counterfactual edits (Ding et al., 2025). We tighten this control significantly by transitioning from generic equal-area masking to strict *shape- and intensity-matched* regions. This pairing analytically cancels first-order occlusion artifacts (Proposition 1; residual mismatch in Appendix D.5). We deploy this across 6 encoders $\times$ 3 depths $\times$ 5 budgets $\times$ 20 seeds rather than evaluating a single checkpoint. However, that tightening alone remains incremental. Our defining contribution is the introduction of a *validity test for the instrument itself* (Section 2.2). None of the aforementioned works employ a lesion-free placebo control, an omission that directly alters downstream conclusions.

Region-supervised and explanation-constrained training. Utilizing spatial regions at training time takes several established forms. CCA diverges from all of them in both mechanism and regime.

Localization supervision. Grounded medical VLM pretraining treats bounding boxes or pixel masks as dense localization targets at corpus scale (Huang et al., 2021; Wu et al., 2023; Deng et al., 2025;

Table 34: **CCA on the primary VinDr-CXR calcification bed** across encoders and constraint weights λ. Metrics reflect means over seeds and budgets for internal AUC, intervention win-rate, and per-site external AUC. The λ=0 baseline represents the optimizer-matched control. We apply a single a priori λ=1 throughout the main text. We report the full grid here to demonstrate that the method is not brittle to this single hyperparameter.

Encoder	λ	Internal AUC	Interv. win	ChestX-Det10	PadChest-GR
RAD-DINO	0	0.613	0.543	0.530	0.579
	0.3	0.618	0.727	0.533	0.598
	1	0.622	0.747	0.537	0.604
	3	0.626	0.752	0.541	0.606
MedSigLIP	0	0.664	0.753	0.603	0.584
	0.3	0.684	0.869	0.607	0.641
	1	0.686	0.883	0.601	0.678
	3	0.671	0.891	0.594	0.687
XrayCLIP-B/16	0	0.647	0.780	0.607	0.583
	0.3	0.653	0.843	0.607	0.612
	1	0.658	0.889	0.605	0.645
	3	0.659	0.921	0.601	0.675
XrayCLIP-L/14	0	0.677	0.836	0.630	0.653
	0.3	0.692	0.891	0.642	0.682
	1	0.702	0.912	0.650	0.700
	3	0.710	0.921	0.657	0.713
DINOv2-S	0	0.600	0.543	0.517	0.558
	0.3	0.605	0.639	0.518	0.554
	1	0.607	0.739	0.518	0.542
	3	0.598	0.799	0.518	0.522
ResNet-50	0	0.624	0.558	0.528	0.565
	0.3	0.622	0.604	0.526	0.563
	1	0.615	0.691	0.523	0.556
	3	0.591	0.743	0.515	0.530

Table 35: **Counterfactual behavior versus saliency alignment** on VinDr-CXR calcification. Both methods utilize identical supervision (k masks), optimizers, seeds, and λ=1. RRR explicitly aligns the evidence *map*. CCA constrains the counterfactual *contrast*. Bold values indicate cells where CCA outperforms RRR.

	Intervention win-rate			External AUC		
Encoder	λ=0	RRR	CCA	λ=0	RRR	CCA
RAD-DINO	0.543	0.566	**0.747**	0.554	0.559	**0.570**
MedSigLIP	0.753	0.744	**0.883**	0.594	0.598	**0.640**
XrayCLIP-B/16	0.780	0.811	**0.889**	0.595	0.599	**0.625**
XrayCLIP-L/14	0.836	0.850	**0.912**	0.641	0.645	**0.675**
DINOv2-S	0.543	0.543	**0.739**	0.537	0.540	0.530
ResNet-50	0.558	0.562	**0.691**	0.546	0.547	0.539

Choi et al., 2026; Lin et al., 2025b). The explicit objective is the ability to ground, requiring large annotated grounded corpora.

Gradient and saliency penalties. Right-for-the-right-reasons (Ross et al., 2017), GradMask (Simpson et al., 2019), and SGAT (Li et al., 2022) explicitly penalize input-gradient or saliency mass falling outside the annotated region. These methods constrain a first-order *proxy*. The main text demonstrates that saliency and causal reliance dissociate in both directions. Aligning a static saliency map neither guarantees nor is required for actual grounded behavior. Consequently, our head-to-head comparison against RRR under identical masks fails to move either performance axis (Table 35).

Table 36: External expected calibration error (ECE, 10 bins, raw sigmoid outputs) at k=25, mean over 20 seeds.

	ChestX-Det10		PadChest	
	λ=0	λ=1	λ=0	λ=1
MedSigLIP	0.443	0.445	0.488	0.478
XrayCLIP-L/14	0.416	0.406	0.485	0.463
XrayCLIP-B/16	0.402	0.416	0.511	0.513
RAD-DINO	0.444	0.432	0.512	0.492
DINOv2-S	0.432	0.402	0.493	0.456
ResNet-50	0.439	0.434	0.494	0.487

Contribution and attribution penalties. CDEP (Rieger et al., 2020) penalizes the contextual decomposition contribution of a designated region rather than its gradient. This constitutes a strictly stronger constraint than a saliency penalty. Notably, its flagship medical experiment derives penalized regions programmatically rather than from expert annotation. Axiomatic attribution priors (Erion et al., 2021) regularize mathematical properties of the attribution map (smoothness, sparsity) without requiring region annotations at all. We implement both mechanisms as supervision-matched baselines (Appendix G.1). Because the published attribution prior is region-free, our variant constitutes a region-conditioned adaptation of it, explicitly penalizing path-integrated attribution mass outside the annotation.

Representation-consistency penalties. The mechanism closest to CCA is ActDiff (Viviano et al., 2019), which penalizes $\|o_\ell(x_{\text{masked}}) - o_\ell(x)\|_2$, forcing the encoder's activations for masked and unmasked inputs to agree. CCA's margin deploys the opposite sign and targets a different objective: it mathematically requires the *prediction* to change strictly more when the actual lesion is removed than when a shape- and texture-matched control is removed. Agreement under masking artificially forces the representation to become insensitive to the region. Our constraint forces the output to become causally sensitive to it, relative to a strict matched null.

Counterfactual-consistency training in general-domain multimodal models. A concurrent general-domain line enforces evidence-grounding through counterfactual supervision at scale. DeFacto trains multimodal LLMs with GRPO-style reinforcement learning over automatically localized evidence regions and their counterfactual/random-masked variants, rewarding answers that change when the evidence is removed (Xu et al., 2025). CF-VLM fine-tunes VLMs with contrastive objectives over generated counterfactual scenes, sharpening sensitivity to minimal causal edits (Zhang et al., 2025). MJ1 adds a counterfactual-consistency reward to a multimodal judge, requiring verdict invariance under input swaps (Kumar et al., 2026). CCA shares the underlying principle—train the output to move under causal edits and stay put under matched non-causal ones—but instantiates it with none of the machinery: no generator, no RL loop, no auxiliary reward model. The counterfactual is a cached occlusion pair; the objective is one convex hinge; the edit is clinically meaningful (the annotated lesion against its shape- and intensity-matched neighbor); and, unlike all three, the instrument that the constraint optimizes is itself validated by a lesion-free placebo before any conclusion is drawn from it.

Counterfactual augmentation. Generative editing of pathology (Cooke et al., 2026) inherits compounding generator-fidelity risks. The paired matched control in CCA analytically supplies the necessary artifact cancellation without generative modeling. Against this entire body of work, CCA directly constrains *counterfactual output behavior*. It requires no generative model, no static saliency proxy, and no large grounded corpus. It operates exclusively in the few-shot regime, utilizing masks solely for the k training positives.

Fine-tuning changes the evidence. Elangovan & Ting (2026) document "semantic drift," observing that attribution maps of CXR classifiers alter substantially between transfer and fine-tuning checkpoints. However, their measurement remains purely reference-free (lacking lesion ground truth) and simply compares two CNN checkpoints. We measure ground-truth-referenced evidence continuously along the label-budget axis against matched causal controls. This explicitly answers the fundamental question their framing raises but cannot resolve: whether the semantic drift moves *toward* or *away from* the actual clinical finding.

Shortcut learning in chest radiography. The phenomenon of CXR models exploiting cohort shortcuts is robustly established. Site signatures reliably break external generalization (Zech et al., 2018), COVID-19 classifiers notoriously exploit laterality markers and patient positioning over pathology (DeGrave et al., 2021), and foundation-model embeddings structurally encode demographics, causing downstream subgroup disparities (Glocker et al., 2023). Counterfactual image editing can expose and partially correct such shortcuts (Cooke et al., 2026), while robust fine-tuning methods penalize spurious feature reliance in natural-image zero-shot settings (Kim et al., 2025). However, this literature primarily audits models trained on massive datasets post hoc. We demonstrate that the *few-label* adaptation regime is shortcut-dominated from the very first gradient step—standard-recipe adapters fail to beat a centroid rule. Furthermore, we formalize the audit as a rigorous pass/fail gate rather than a passive post hoc diagnosis.

Label-efficient adaptation of medical foundation models. Few-shot adaptation of medical VLMs features structured benchmarks (Shakeri et al., 2024) and comparisons between PEFT and full fine-tuning (Lian et al., 2024). Our gate-0 control acts as a nearest-class-mean rule, for which we claim no algorithmic novelty. Prototype and centroid classifiers serve as long-standing few-shot baselines (Snell et al., 2017; Wang et al., 2019), repeatedly proving competitive with linear probes at extremely low shot counts (Zhang et al., 2022; Huang et al., 2024). Closest in diagnostic spirit, Huang et al. (2019) show that several few-shot benchmarks are solvable by centroid clustering *without test-time labels at all*, implying that reported adaptation can reduce to representation quality; gate-0 turns the same suspicion into a per-cell statistical equivalence test tied to causal grounding rather than a benchmark-level critique. Our core contribution elevates this known baseline into a strict audit precondition enforced via a supervision-matched protocol and a rigorous statistical equivalence test. Recent label-efficient CXR literature reports precisely the curve we start from: mean AUC plotted against shot count (Dalsania, 2025), with RAD-DINO documenting strong transfer capacity (Pérez-García et al., 2025). The rare-finding setting remains highly realistic; even the largest auto-generated CXR lesion-segmentation dataset covers only seven common findings, excluding the calcification family (Choi et al., 2026). Closest to our diagnostic findings, frozen foundation-model embeddings discard small-lesion signals under linear probing (Muthyala et al., 2026). Our companion diagnostic study (Appendix E.1) reproduces this localization ceiling across six encoders and three frozen remedies. None of these prior works report any causal evidence axis alongside the accuracy axis. Providing this dual axis—and proving that the two dissociate—remains the central objective of this paper.

Evidence and generalization. Correlational links between explanation quality and external performance certainly exist. Higher input resolution simultaneously improves both explainability and external CXR generalization (Kulkarni et al., 2026), while saliency-guided training enhances robustness in natural images (Li et al., 2022). Group-robust optimization (e.g., group DRO; Sagawa et al., 2019) removes spurious reliance exclusively when group labels are explicitly available. In our few-shot regime, neither the spurious shortcut variable nor the group structure is known. CCA circumvents this by utilizing the lesion annotation itself as a direct causal handle. WiSE-FT evaluates the robustness cost of fine-tuning against a zero-shot anchor (Wortsman et al., 2022). Gate-0 provides a tighter, supervision-matched null (threshold-only recalibration of a fixed rule, Proposition 2), explicitly formalized as a validity precondition. To our knowledge, no prior work (i) pairs classification gain and ground-truth-referenced evidence gain on an identical label-budget axis for rare-finding adaptation, (ii) validates a causal-grounding instrument against a strict lesion-free placebo before drawing conclusions from it, or (iii) closes the loop by deploying the audit as an explicit training constraint.

Positioning at a glance. Table 37 compresses the comparison along the five properties this paper treats as load-bearing; the caption defines the parenthesized partial marks. Of these, matched pairs supplied by a generative editor carry an attendant fidelity risk.

Table 37: **Comparison with prior families.** Columns: (a) measures grounding *along the label-budget axis* of adaptation rather than on a static checkpoint; (b) paired shape- and intensity-matched counterfactual controls; (c) a lesion-free placebo that tests the instrument itself; (d) trains for counterfactual output behavior; (e) operates in the few-label regime. A parenthesized mark denotes a partial form: equal-area instead of shape- and intensity-matched controls in (b), a saliency proxy instead of a counterfactual output constraint in (d), or matched pairs supplied by a generative editor.

Family	(a) budget axis	(b) matched pair	(c) placebo	(d) trains	(e) few-label
Static grounding audits (Saporta et al.; Luo et al., 2026)	×	×	×	×	×
Occlusion/intervention metrics (Lotfinia et al., 2026; Ding et al., 2025)	×	(✓)	×	×	×
Saliency-penalty training (Ross et al., 2017; Simpson et al., 2019; Li et al., 2022)	×	×	×	(✓)	✓
Contribution/attribution priors (Rieger et al., 2020; Erion et al., 2021)	×	×	×	(✓)	✓
Representation consistency (Viviano et al., 2019)	×	×	×	(✓)	✓
Generative counterfactual editing (Cooke et al., 2026)	×	(✓)	×	×	×
Counterfactual-consistency MLLM training (Xu et al., 2025; Zhang et al., 2025; Kumar et al., 2026)	×	(✓)	×	✓	×
Grounded VLM pretraining (Huang et al., 2021; Wu et al., 2023)	×	×	×	✓	×
This work	✓	✓	✓	✓	✓

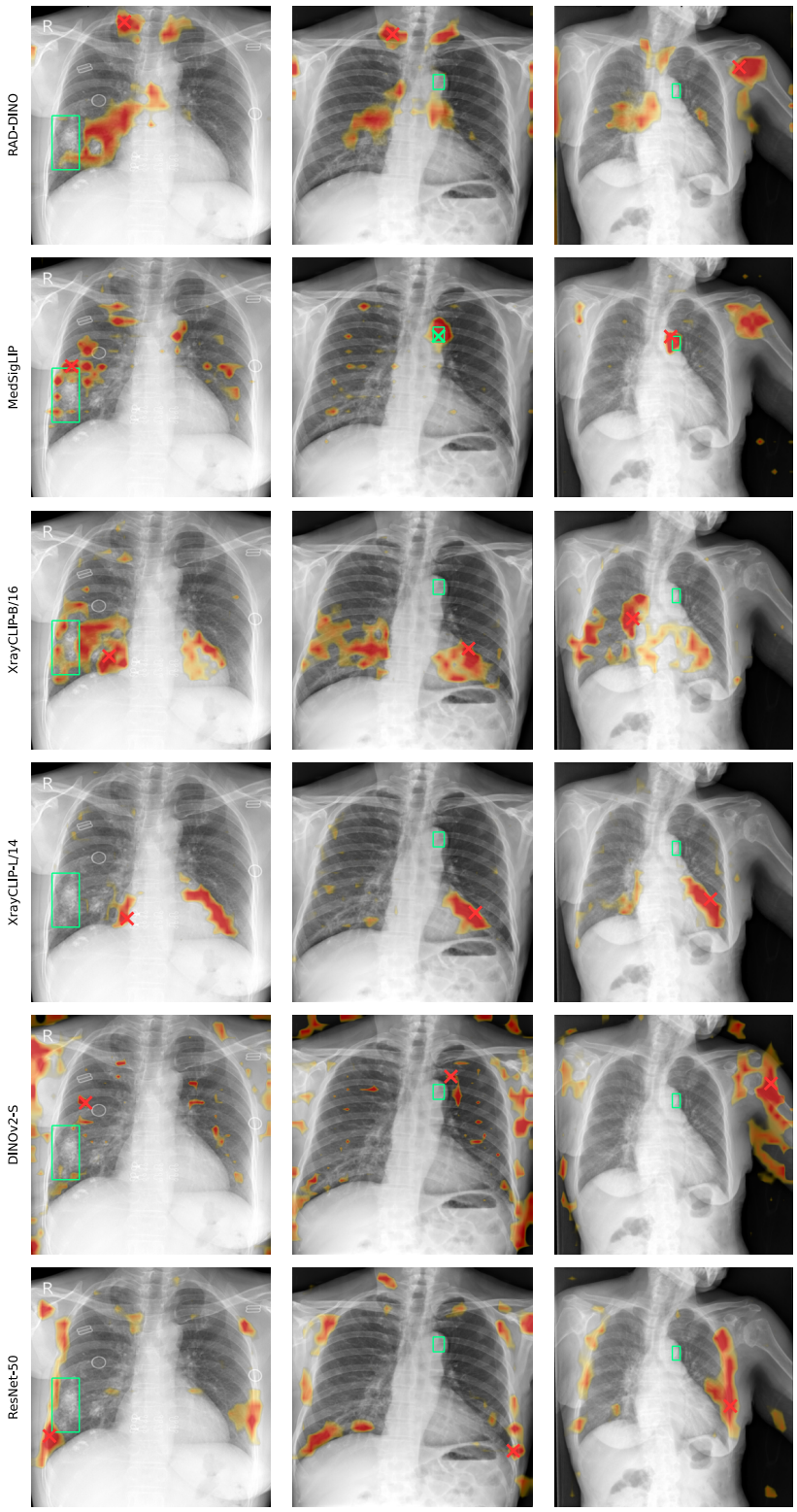


Figure 11: **Probe-CAM across all six encoders** (k=100, seed 0, primary VinDr-CXR calcification bed). The same test positives are displayed in every row. Annotations are marked in green, and the CAM argmax is explicitly indicated.

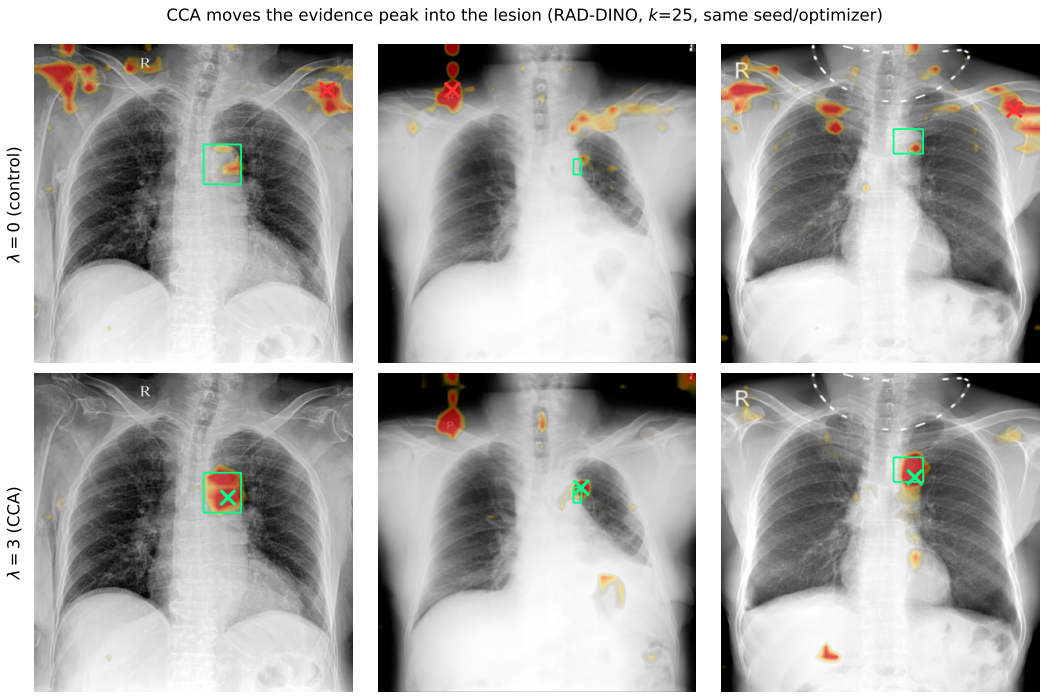


Figure 12: **The identical CCA effect on an encoder starting at chance** (RAD-DINO, VinDr-CXR calcification, k=25). The constraint successfully relocates the peak even where absolute causal grounding is initially low. This explains why the win-rate moves on every encoder in Table 3, while external AUC improves exclusively on the CXR-domain models.

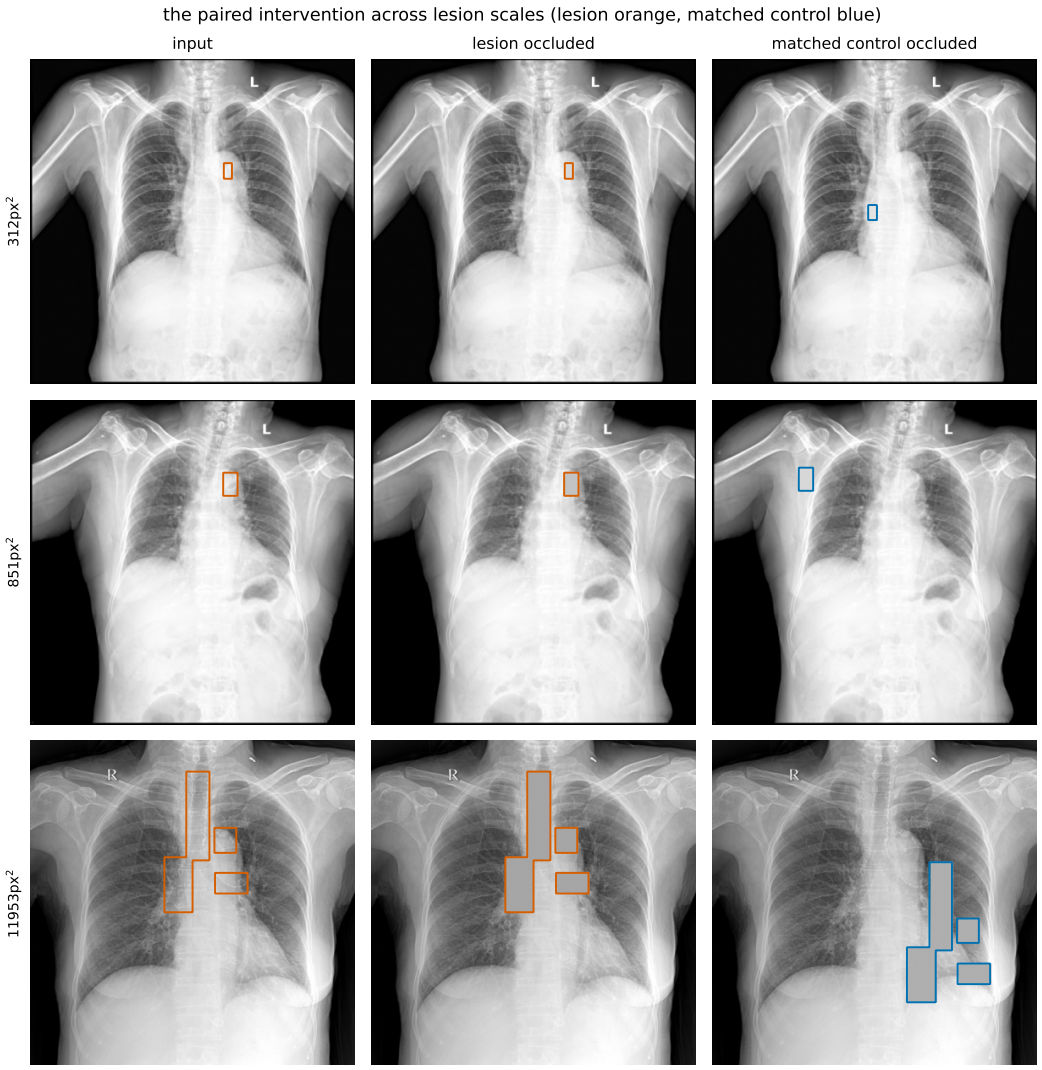


Figure 13: **The occlusion construction** on test positives from VinDr-CXR, SIIM-ACR, and DDR. Each panel displays the lesion, its three shape-matched controls, and the corresponding ring-mean fill.

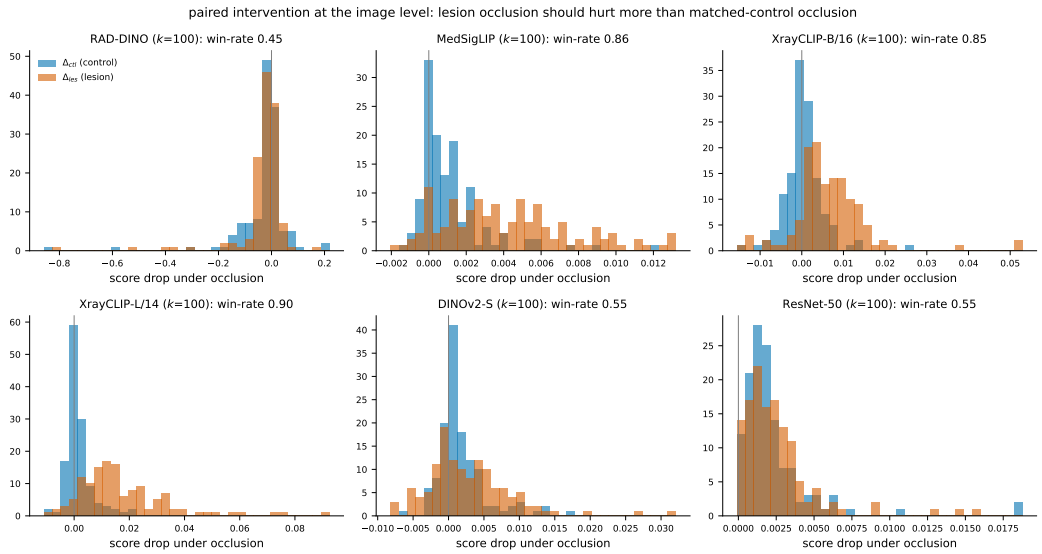


Figure 14: **Per-image intervention deltas** on VinDr-CXR calcification. The scatter visual plots Δ_{les} directly against $\overline{\Delta}_{ctl}$. The aggregate causal win-rate equates precisely to the fraction of data points lying strictly above the diagonal.

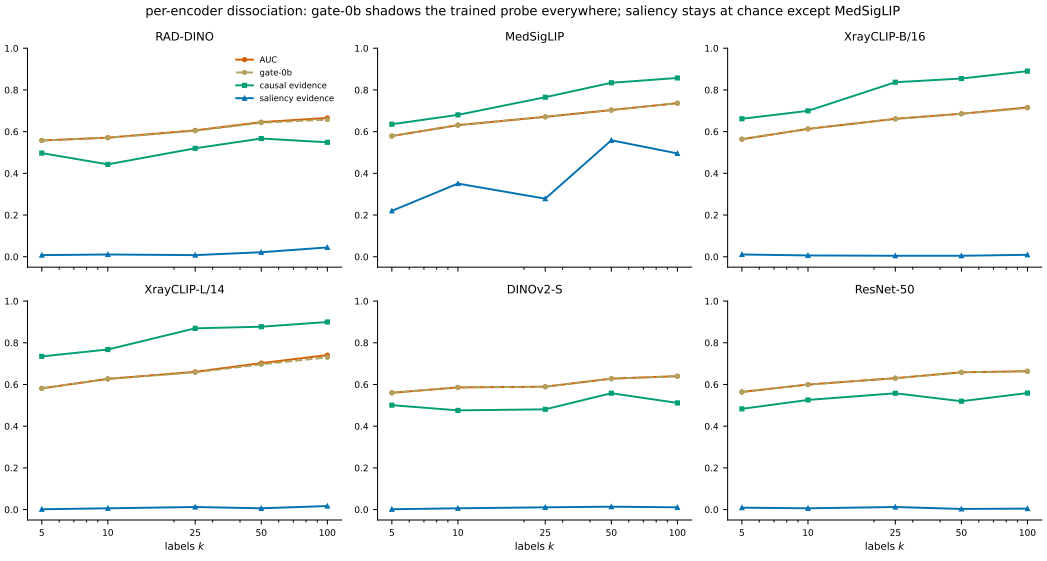


Figure 15: **Per-encoder curves at every budget** evaluated on the primary VinDr-CXR calcification bed.

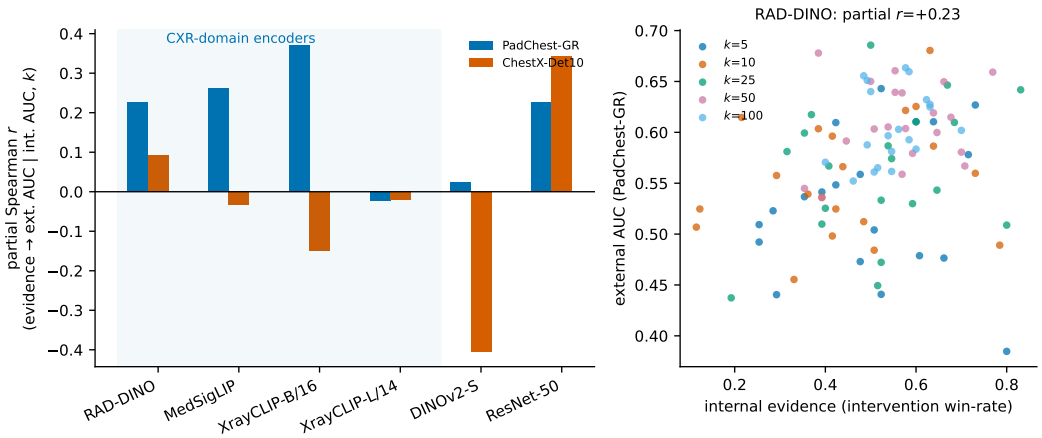


Figure 16: **Evidence–transfer partial correlation per encoder** on VinDr-CXR calcification. The left panel splits results by external site (PadChest-GR and ChestX-Det10). The right panel details the underlying cell for RAD-DINO transferring to PadChest-GR, where each point represents a trained probe colored by label budget. The partial Spearman correlation explicitly controls for internal AUC and $\log k$. While this visualizes the primary bed only, Table 20 reports the 102-cell evaluation across the six replication beds (the primary bed is excluded there).

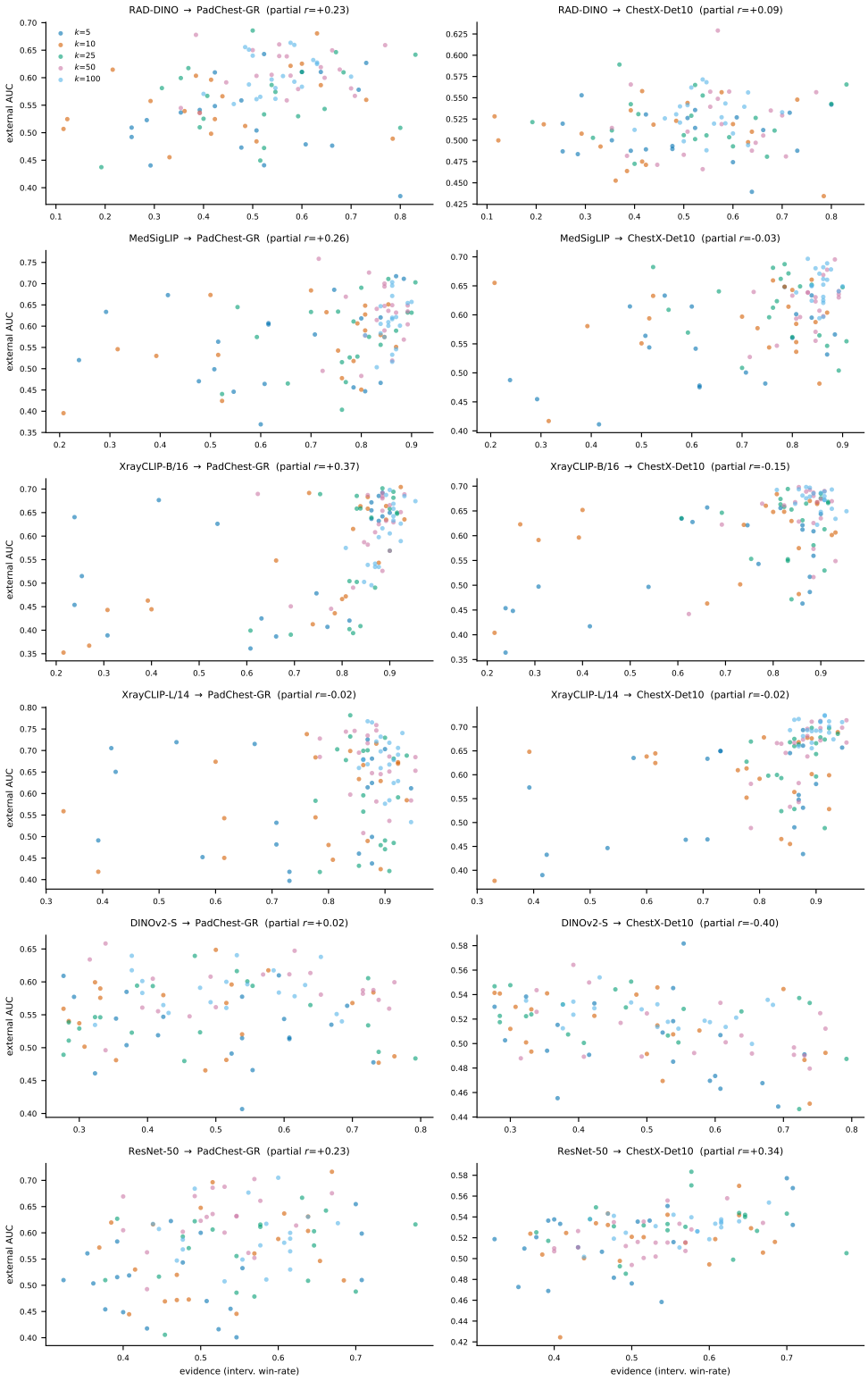


Figure 17: **Evidence versus external AUC across all cells.** The grid spans six public beds, six encoders, and their corresponding public external sites (Table 4). Each point represents an individual trained probe.

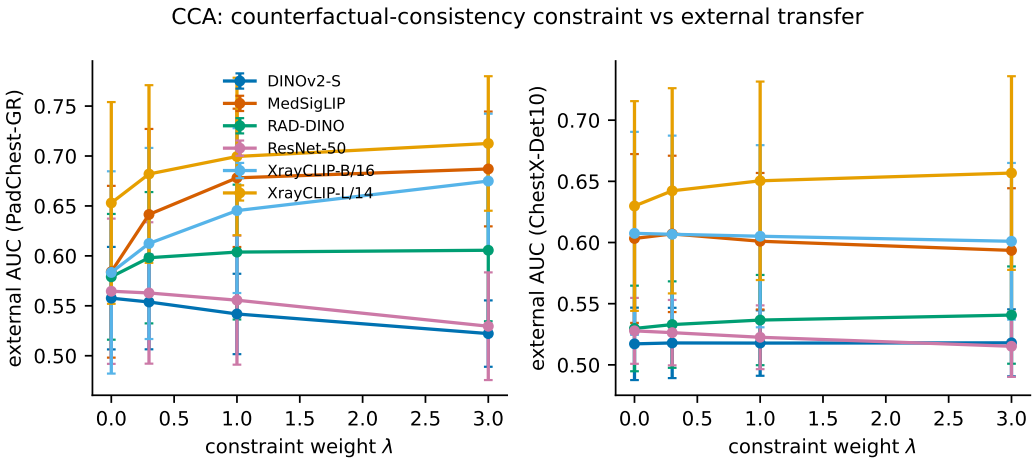


Figure 18: **External AUC against the constraint weight** λ on VinDr-CXR calcification across public external sites (left: PadChest-GR, right: ChestX-Det10). Bands represent ± 1 standard deviation over seeds. The baseline $\lambda = 0$ denotes the optimizer-matched control, excluding framework-level (sklearn vs torch) differences by construction.

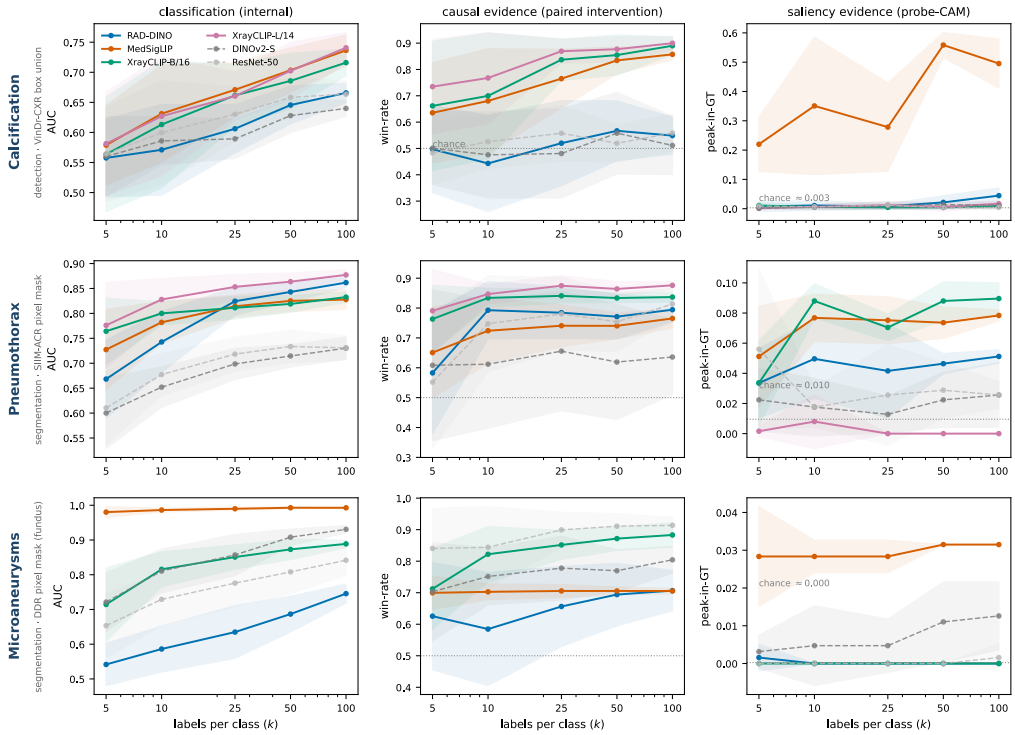


Figure 19: **The dissociation replicates across annotation type and modality.** Rows depict calcification on VinDr-CXR (detection, three-reader box union), pneumothorax on SIIM-ACR (segmentation, pixel mask), and microaneurysms on DDR (segmentation, pixel mask, fundus). Columns represent internal AUC, causal evidence, and saliency evidence. Solid lines indicate CXR-domain encoders. Dashed lines denote natural-image controls. Shading displays ± 1 standard deviation over 20 seeds. Internal AUC climbs consistently in every row, yet neither evidence channel follows. The CAM chance line (lesion area fraction) is computed per bed. Causal evidence and saliency frequently disagree. Encoders exhibiting high causal win-rates can still produce chance-level saliency maps.

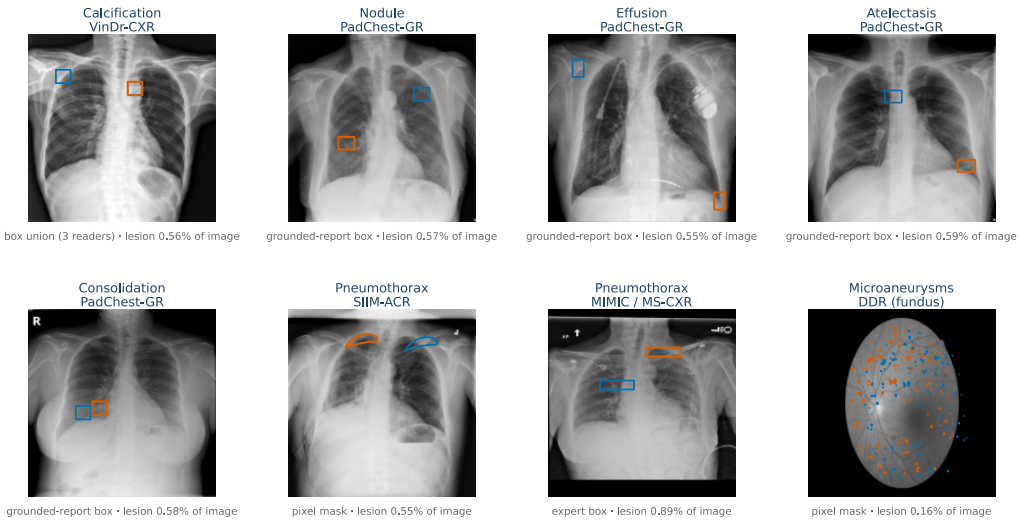


Figure 20: **The eight public task beds.** Each panel displays one test positive with the annotated lesion (orange) and one shape-matched control (blue). Sources and formats include VinDr-CXR (Nguyen et al., 2022) three-reader box union; PadChest-GR (Bustos et al., 2020; De Castro et al., 2025) grounded-report boxes (nodule, effusion, atelectasis, consolidation); SIIM-ACR (Zawacki et al., 2019) pixel masks; MIMIC / MS-CXR (Johnson et al., 2019; Boecking et al., 2022) expert boxes; and DDR (Li et al., 2019) fundus pixel masks. The protocol construction remains identical throughout. Only the annotation format and lesion scale vary.

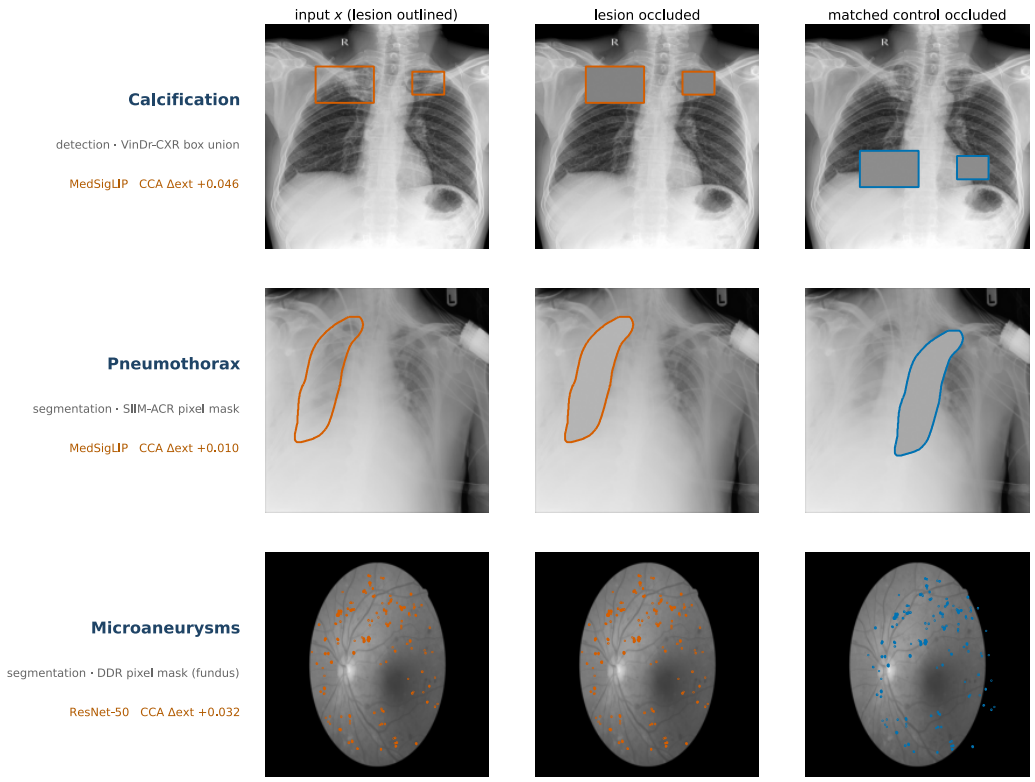


Figure 21: **The constraint's actual inputs across annotation formats.** Each panel highlights the largest annotated test positive for the encoder achieving the highest fixed-λ CCA gain on that bed. It displays the paired lesion region and shape-matched control differenced by the CCA constraint during training. Box annotations (VinDr-CXR), coarse pixel masks (SIIM-ACR), and fine pixel masks in a distinct modality (DDR fundus) supply the identical paired causal signal.