Post-Training Clinical-Signal Auditing of Medical Diagnostic Models

Anonymous authors
Paper under double-blind review

Abstract

In diagnostic imaging, specific findings vary in appearance with disease severity or clinical state, and these relationships, accumulated as clinical knowledge, provide an important basis for diagnostic interpretation. Diagnostic performance is a central axis for evaluating medical AI, but from diagnostic performance alone it can be difficult to determine whether models with similar performance respond differently to these changes. Using clinical ordering information withheld from training and final model selection as a reference, we assess how closely frozen-model responses follow the order of these changes within the same diagnostic label, and propose this assessment as a retraining-free post-training audit method. Across 32 backbone–task combinations on four medical image diagnosis tasks (Liver US, Breast US, Knee X-ray, and Fundus), this degree of agreement ranged from -0.24 to $+0.63$, and its relationship with diagnostic performance varied by task. In the three tasks with a coarse spatial scale prespecified for the relevant clinical findings, higher agreement was associated with better preservation of this ordering after controlled perturbations, and the association remained positive after adjustment for relevant covariates ($\rho = +0.475$). Because the audit method requires only frozen-model forward outputs on the model side, it can be applied to publicly released diagnostic checkpoints without access to their original training pipelines or internal representations. These results provide a complementary post-training criterion for assessing how closely model responses agree with the direction of change indicated by clinical knowledge, beyond what conventional diagnostic performance alone readily captures.

1 Introduction

Medical image diagnosis involves not only assigning findings to diagnostic labels but also interpreting task-specific findings and how they vary with disease severity or clinical state. These relationships are specified as clinical knowledge accumulated in clinical grading criteria and the quantitative imaging literature (Leung et al., 2025; Moon et al., 2013; Ferraioli et al., 2022), and give clinicians a basis for interpreting findings in relation to clinical state.

Multiple evaluation axes, including diagnostic performance, calibration, and spatial attribution, have been used to assess the trustworthiness of medical imaging diagnostic models (Guo et al., 2017; Selvaraju et al., 2017). These axes evaluate label discrimination, the appropriateness of confidence, and the spatial evidence associated with predictions, respectively, but do not directly indicate whether model responses vary consistently with the ordered relation formulated in the established clinical knowledge for the task. A more complete assessment of trustworthiness therefore requires evaluating this relationship, not captured by existing evaluation axes, from a separate perspective.

Prior work has taken two perspectives: incorporating clinical knowledge into model learning, and analyzing model behavior after training. At training time, specified evidence or clinical processes have been included in the learning objective to encourage model behavior consistent with clinical knowledge (Ross et al., 2017; Rieger et al., 2020; Liu et al., 2025a). Even with such interventions, whether the resulting frozen model's responses follow the ordered variation of task-specific findings indicated by clinical knowledge must be checked separately after training. Post-training work has diagnosed the behavior of trained models by analyzing external-site failure, shortcut and feature dependence, and concept sensitivity (Zech et al., 2018; Oakden-Rayner et al., 2020; DeGrave et al.,

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.

2021; Kim et al., 2018; Graziani et al., 2018). These analyses provide important information on failure, dependence, and sensitivity, but do not directly target how closely model responses follow the ordered variation of a prespecified clinical signal.

We therefore propose a retraining-free post-training clinical-signal audit for frozen diagnostic models. We refer to a task-specific imaging finding that varies in an ordered manner with disease severity or clinical state as a *clinical signal*, and to the order of this variation given by established clinical knowledge as the ordered relation. Using an ordered clinical factor withheld from training as a reference for this ordering, the audit quantifies clinical-signal response alignment: how closely frozen-model responses follow the ordered relation within the same diagnostic label. This enables evaluation of differences in within-label clinically ordered response that conventional diagnostic evaluation alone does not reveal, even among models with similar diagnostic performance.

In a controlled model panel of 32 backbone–task cells and 96 checkpoints across four medical imaging tasks, clinical-signal alignment varied widely across models, and its relationship with diagnostic performance differed by task. In the three coarse-anchor tasks, higher alignment was associated with better preservation of clinical ordering after controlled perturbations, and lower alignment with greater controlled-corruption vulnerability even after controlling for clean diagnostic performance. The same forward-only audit was also applied to public diagnostic checkpoints whose original training pipelines are inaccessible.

Contributions. Our main contributions are as follows:

1. **(Post-training clinical-signal audit)** We define clinical-signal response alignment as the degree to which frozen-model responses follow the ordered variation of a prespecified clinical signal within the same diagnostic label, and propose a retraining-free audit that measures it using only frozen-model forward outputs on the model side.
2. **(Behavioral relevance and stress-testing utility)** We show that clinical-signal alignment is associated with opposite-scale dependence and, in the three coarse-anchor tasks, with clinical-order retention after controlled perturbations, and present its potential use as follow-up stress-testing information complementary to clean diagnostic performance.
3. **(Practical post-training applicability)** The audit can be applied to existing frozen checkpoints without retraining and without access to gradients, intermediate representations, or the original training pipeline, as we confirm on 11 public Breast US checkpoints.

2 Related Work

Medical Diagnostic Models and Trustworthiness Evaluation. Pretrained representations and model-development tools have been used to build diagnostic models efficiently even with limited labeled data (Thirunavukarasu et al., 2023; Zhang & Metaxas, 2024; Schäfer et al., 2024). Model evaluation, meanwhile, has covered diagnostic discrimination, calibration (Guo et al., 2017), generalization (Wu et al., 2021), and spatial attribution (Selvaraju et al., 2017), among other axes. Models with similar diagnostic performance can still differ in their decision rules, as discussed from underspecification and Rashomon perspectives (D'Amour et al., 2022; Semenova et al., 2022). Post-hoc attribution such as Grad-CAM shows prediction-related spatial evidence (Selvaraju et al., 2017), but the localization validity of such explanations is not guaranteed (Arun et al., 2021), and their ability to detect spurious correlations not known in advance is limited (Adebayo et al., 2022). These evaluations assess complementary aspects of model behavior, but whether model responses are consistent with clinically known relationships remains a separate evaluation problem.

Clinical Knowledge in Diagnostic Model Learning. Clinical knowledge has been used in model learning through multi-task objectives that jointly predict descriptors or grades (Zhang et al., 2021; Tiulpin & Saarakkala, 2020; Li et al., 2020) and through clinical-concept bottlenecks (Koh et al., 2020; Bunnell et al., 2024). More generally, differentiable penalties that align model explanations with specified evidence regions have been proposed (Ross et al., 2017; Rieger et al., 2020), and for medical diagnosis, a causal alignment loss in the learning objective has been proposed to align the model's decision process with clinicians' causal process (Liu et al., 2025a). These approaches encourage model behavior consistent with clinical knowledge during training, but how consistent the frozen model's responses are with such clinical relations is left to separate post-training evaluation.

Post-Training Auditing of Medical Diagnostic Models. Post hoc auditing of trained models has extended from pointwise prediction reliability (Schulam & Saria, 2019) to medical algorithmic audits of task-specific clinical errors and risks (Liu et al., 2022a). Related studies have analyzed external-site generalization failure, hidden stratification, and subgroup-specific underdiagnosis (Zech et al., 2018; Oakden-Rayner et al., 2020; Seyyed-Kalantari et al., 2021). For shortcut learning (Geirhos et al., 2020), shortcut utilization and feature dependence have been diagnosed through image interventions and internal model analyses (DeGrave et al., 2021; Boland et al., 2024; Vigneshwaran et al., 2026; Wong et al., 2026), and verification and mitigation approaches using controlled training, modified data, and model revision have also been proposed (Brown et al., 2023; Ong Ly et al., 2024; Pahde et al., 2025; Cooke et al., 2026; Unnikrishnan et al., 2026). Computational stress testing with systematic perturbations reveals clinic-readiness gaps and perturbation-specific vulnerabilities that conventional performance metrics miss (Young et al., 2021; Islam et al., 2023). Concept-based post-hoc methods have measured prediction sensitivity to human-defined concepts (Kim et al., 2018; Graziani et al., 2018; Bai et al., 2023), and a concept model grounded in the medical literature has supported concept-level data and model auditing (Kim et al., 2024). An audit framework using clinician-interpreted counterfactual images and an approach that builds auditability into model design through clinically grounded concept representations have also been proposed (DeGrave et al., 2025; Han et al., 2026).

These studies provide important information on model failure, feature dependence, perturbation robustness, and concept representation and sensitivity, but from this information alone it can be difficult to determine how closely model responses follow the task-specific ordering of a particular clinical signal within the same diagnostic label. We therefore quantify this relationship as clinical-signal response alignment and, unlike representation-level concept sensitivity, directly evaluate label-controlled clinical ordering at the level of model outputs.

3 Problem Setup & Background

3.1 Problem Setup

A trained diagnostic model f_θ predicts a diagnostic label y from an input image x. We evaluate at the post-training stage whether the frozen-model response $f_\theta(x)$ follows the ordered variation of a clinical signal. The clinical ordering reference is an ordered clinical factor c, withheld from model training and checkpoint selection. This setting does not assume statistical independence between y and c; for example, y may be defined by thresholding c, or, as in Breast US, y and c may be defined from pathology and BI-RADS, respectively.

For each task, a representative clinical finding linked to c and its spatial scale are prespecified before audit outputs are examined. This spatial scale is the *clinical anchor scale*, and the corresponding input in the fixed scale-band representation is the *anchor band*. The audit evaluates how closely the anchor-band response follows the ordering of c within the same diagnostic label y. Table 1 summarizes the resulting audit specification for each task.

3.2 Background

Scale-band representation. We evaluate the clinical signals of different tasks in a common multiscale image representation. The Laplacian pyramid and scale-space representation are classical multiscale representations long used to express an image as components at different spatial scales (Burt & Adelson, 1983; Lindeberg, 1994), and multiscale representations have also been used in medical imaging to analyze image structures and information at different scales (Vu et al., 2026; Liu et al., 2025b).

We decompose each input with a fixed Laplacian pyramid and represent its components as three *scale bands*, *fine*, *mid*, and *coarse*, by spatial extent. The decomposition is applied only at audit time, independently of model training, and the relative scale ordering of the three implemented bands was confirmed by a synthetic calibration without clinical data (§A).

Table 1: **Task-specific clinical audit specifications.** Cohort sizes are given in §5.1.

Task	y / factor c	Representative clinical finding	Spatial-scale rationale	Anchor scale
Liver US	$c \geq 5$ / biopsy steatosis % 0–85	Liver–kidney echogenicity contrast (Ballestri et al., 2012) and posterior/deep beam attenuation (Ballestri et al., 2012; Ferraioli et al., 2022)	Liver–renal cortex brightness ratio and depth-dependent attenuation (Kwon et al., 2016)	coarse
Breast US	pathology / BI-RADS 2–5	Lesion shape, orientation, and margin (Leung et al., 2025; Hong et al., 2005; Moon et al., 2013)	Whole-lesion shape/orientation and overall contour geometry (Moon et al., 2013; Lee et al., 2009)	coarse
Knee X-ray	$c > 1$ / KL grade 0–4	Joint-space narrowing and associated structural change (Kohn et al., 2016; Altman & Gold, 2007; Irfan et al., 2026)	Tibiofemoral joint-space geometry measured across the joint and compartment (Marijnissen et al., 2008; Cheung et al., 2021)	coarse
Fundus	$c > 1$ / DR grade 0–4	Microaneurysms and small dot haemorrhages (Lim et al., 2025; Wilkinson et al., 2003; EyePACS, n.d.)	Small, localized punctate retinal lesions (Quellec et al., 2006; EyePACS, n.d.)	fine

Clinical anchor specification. The clinical anchor scale of each task is prespecified independently of audit outputs. We determine the representative finding from clinical grading criteria and its spatial scale from the quantitative imaging literature, and map this scale to the anchor band (Table 1).

4 Methodology

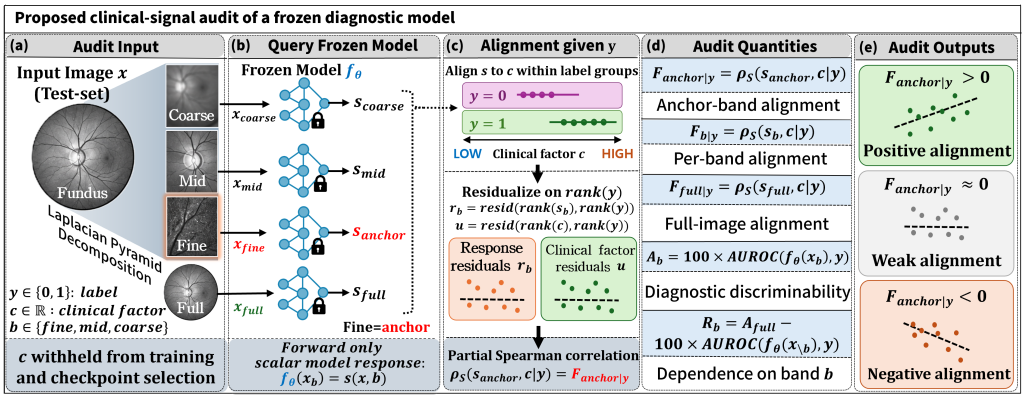

Figure 1: **Overview of the proposed post-training clinical-signal audit. (a–e)** Audit inputs, frozen-model queries, label-controlled alignment, audit quantities, and interpretation of $F_{\text{anchor}|y}$. The modality illustrations are synthetic schematics, are not used as audit data, and convey no anatomical or diagnostic evidence.

Figure 1 summarizes the five audit steps defined below: (a) constructing audit inputs, (b) querying the frozen model, (c) computing label-controlled alignment, (d) computing companion quantities, and (e) interpreting audit outputs.

4.1 Audit Inputs and Frozen-Model Queries

We first construct the audit inputs (Figure 1(a)) and query the frozen model for the responses used in the audit (Figure 1(b)). Each audit sample consists of a test image x, its diagnostic label y, and the ordered clinical factor c withheld from model training and checkpoint selection. Only the image

is input to the frozen model f_θ; y and c are used to compute audit quantities. The scalar response $f_\theta(x)$ is the softmax probability of the positive class.

A fixed Laplacian pyramid represents each image by raw scale-band components $\tilde{x}_b$, $b \in \{\text{fine}, \text{mid}, \text{coarse}\}$. The model input x_b is $\tilde{x}_b$ after per-image $[0, 1]$ min–max normalization and the training preprocessing (§C.2), and x_{anchor} is the input corresponding to the clinical anchor scale. Responses are $s_b = f_\theta(x_b)$, $s_{\text{anchor}} = f_\theta(x_{\text{anchor}})$, and $s_{\text{full}} = f_\theta(x)$.

Given an audit cohort with x, y, and c, the audit requires only frozen-model forward outputs on the model side; it requires neither additional training nor spatial annotations such as lesion masks. The main model ordering was largely preserved under two alternative normalizations (§B.3).

4.2 Label-Controlled Clinical-Signal Alignment

Next, we quantify how closely the anchor-band response s_{anchor} follows the ordering of the clinical factor c within the same diagnostic label (Figure 1(c)). Because y and c can be associated, a simple correlation between s_{anchor} and c may reflect both clinical ordering and diagnostic class separation. To separate the two, we define the primary audit quantity, controlling for y-related differences, as

$$F_{\text{anchor}|y} = \rho_S(s_{\text{anchor}},\, c \mid y) = \rho_S(f_\theta(x_{\text{anchor}}),\, c \mid y)\,. \tag{1}$$

We mean-rank transform s_{anchor}, c, and y, residualize $\text{rank}(s_{\text{anchor}})$ and $\text{rank}(c)$ on $\text{rank}(y)$, and compute the Pearson correlation of the two residuals. Here $\rho_S(\cdot, \cdot \mid \cdot)$ denotes this label-controlled rank statistic, the traditional partial Spearman correlation. Estimator implementation, tie handling, and permutation controls are given in §B and the released code.

$F_{\text{anchor}|y} > 0$, $F_{\text{anchor}|y} \simeq 0$, and $F_{\text{anchor}|y} < 0$ respectively mean that the anchor-band response follows the ordering of the clinical factor, shows no clear ordered relation, or shows the opposite ordering.

4.3 Companion Quantities and Audit Scope

Finally, to interpret the primary alignment alongside other model-response properties, we compute companion quantities (Figure 1(d)). Band-wise response alignment is $F_{b|y} = \rho_S(f_\theta(x_b), c \mid y)$. Its value at the prespecified anchor band is $F_{\text{anchor}|y}$. Full-image alignment is $F_{\text{full}|y} = \rho_S(f_\theta(x), c \mid y)$.

We measure the diagnostic discriminability of band b as $A_b = 100 \times \text{AUROC}(f_\theta(x_b), y)$, denoted A_{anchor} and A_{full} for the anchor band and the full image; all AUROC-based values are reported on a 0–100 scale. Reporting A_b and $F_{b|y}$ together separates band-level diagnostic discriminability from within-label clinical ordering.

Band dependence is defined as

$$R_b = A_{\text{full}} - 100 \times \text{AUROC}\big(f_\theta(x_{\setminus b}),\, y\big)\,, \tag{2}$$

where $x_{\setminus b}$ is the input reconstructed from the pyramid components excluding band b (§B.3), so that R_b is the AUROC percentage-point loss from removing band b. R_{fine} denotes fine-band dependence, and R_{opp} denotes dependence on the scale opposite to the anchor: R_{fine} for the coarse-anchor tasks and R_{coarse} for Fundus, whose anchor is fine. A_b, R_b, and R_{opp} are computed without sample-level values of c.

Band-restricted inputs serve as controlled probes of the prespecified spatial scale and may differ from the training distribution. The final audit output is therefore reported as the frozen model's *functional response alignment* measured on the prespecified anchor-band probe, centered on $F_{\text{anchor}|y}$ and interpreted with the companion quantities (Figure 1(e)).

5 Experiments

5.1 Experimental Setup

We used four clinical imaging tasks across three modalities, differing in clinical factor and anchor role: Liver US (Byra et al., 2018b), Breast US (Gómez-Flores et al., 2024), Knee X-ray (Chen, 2018),

and Fundus (EyePACS & California Healthcare Foundation, 2015). The audit cohorts comprise 170, 561, 1289, and 1375 images from 17, 320, 919, and 1100 patients or cases, respectively.

For each task, we fine-tuned the same 8 ImageNet-pretrained CNN and Transformer backbones with 3 seeds, forming a *controlled model panel* of 32 backbone–task cells and 96 checkpoints. All models were trained only on full images and diagnostic labels y; neither the clinical factor c nor any audit quantity was used for training or checkpoint selection.

Backbone–task analyses use three-seed means. Confidence intervals are bootstrapped over patient or case clusters or over backbone–task cells, depending on the analysis; each table states its analysis and resampling units. Preprocessing, splits, architectures, and training details are in §C, detailed protocols in §C.3–§C.6, and additional analyses in §D.1–§D.4.

5.2 Evaluation Protocol

The evaluation has four parts. **(a) Diagnostic performance.** We computed $F_{b|y}$ for all scale bands and the prespecified $F_{\text{anchor}|y}$, and related them to A_{full}, A_{anchor}, and $F_{\text{full}|y}$.

(b) Behavioral relevance and stress-testing utility. We evaluated the relation between $F_{\text{anchor}|y}$ and opposite-scale dependence R_{opp} over the full panel, applied the same Gaussian blur, JPEG compression, and downsampling to all four tasks, and analyzed diagnostic AUROC loss and clinical-order retention in the three coarse-anchor tasks. Clinical-order retention D is the mean change in full-image alignment $F_{\text{full}|y}$ from its clean value under the three corruptions; larger values indicate better preservation of clinical ordering (§D.2.1). We applied a role-matched test to Fundus and assessed operational utility with an exploratory task-stratified budgeted-prioritization analysis (§D.2.4).

(c) Complementarity to existing evaluation. On the 24 checkpoints of Breast US, the only task in the panel with lesion masks, we compared diagnostic performance, ground-truth-class Grad-CAM lesion attribution mass, and $F_{\text{anchor}|y}$. On the 24 checkpoints of the Fundus replication panel, we compared $F_{\text{anchor}|y}$ with regression concept vector (RCV; Graziani et al., 2018) relevance computed on the same x_{anchor}.

(d) Practical applicability. We applied the same forward-only audit to 11 public Breast US checkpoints outside our training protocol.

5.3 Audit Outputs

Table 2 reports the controlled-panel audit outputs for the 32 backbone–task cells as three-seed means: the primary audit quantity $F_{\text{anchor}|y}$, band-wise alignment $F_{b|y}$, full-image and anchor-band AUROC, and full-image alignment $F_{\text{full}|y}$. $F_{\text{anchor}|y}$ ranged from -0.24 to $+0.63$ in Liver US and from $+0.19$ to $+0.30$ in Fundus. The prespecified anchor band gave the largest band-wise alignment $F_{b|y}$ in 7 out of 8 cells for Knee X-ray, 5 out of 8 for Breast US, 4 out of 8 for Liver US, and 7 out of 8 for Fundus. The prespecified anchor scale thus often coincided with the strongest within-label clinical ordering of model responses, especially in Knee X-ray and Fundus, but not consistently across tasks and models. Liver US has 17 independent patients, wider confidence intervals for single-checkpoint alignment estimates than the other tasks, and the lowest seed-swap backbone rank correlation; its $4/8$ should therefore be read with this lower precision (§D.1). To test whether the observed alignment depends on the sample-wise correspondence between the clinical factor and the model response, we used a permutation control that removes this correspondence while preserving both distributions. Under this correspondence-breaking permutation, $F_{\text{anchor}|y}$ moved to near 0. The observed $|F_{\text{anchor}|y}|$ exceeded the 95th percentile of the image-level permutation null distribution in 27 of 32 cells, and in 26 under a group-preserving null reference that retains the repeated-measures structure (§B.2).

Public checkpoints. On the 11 public Breast US checkpoints outside our training protocol, $F_{\text{anchor}|y}$ ranged from $+0.07$ to $+0.31$ and exceeded the 95th percentile of its model-specific within-y permutation distribution in all cases. The prespecified coarse anchor band gave the largest $F_{b|y}$ in $10/11$ (Table 17; §D.4).

Table 2: **Controlled-panel audit outputs.** Three-seed means for 32 backbone–task cells; bold marks the largest $F_{b|y}$. Grad-CAM attribution is shown only for Breast US (uniform baseline 0.087).

	AUROC		Alignment $F_{b\|y}$			Audit quantities		Lesion
Model	full	anchor	fine	mid	coarse	$F_{\text{anchor}\|y}$	$F_{\text{full}\|y}$	attribution
Breast US (clinical anchor: *coarse*)								
ConvNeXt-T	90.8	85.7	+0.17	+0.35	**+0.41**	+0.41	+0.48	0.192
ConvNeXt-S	90.7	85.8	+0.23	+0.28	**+0.35**	+0.35	+0.48	0.208
Swin-T	90.4	83.4	+0.15	+0.23	**+0.30**	+0.30	+0.44	0.181
ViT-S	86.0	79.8	+0.15	**+0.26**	+0.24	+0.24	+0.40	0.111
DeiT-S	87.6	78.5	+0.06	+0.24	**+0.27**	+0.27	+0.43	0.131
DenseNet121	86.8	64.4	+0.06	+0.07	**+0.17**	+0.17	+0.42	0.227
ResNet18	81.4	52.9	+0.07	**+0.16**	+0.02	+0.02	+0.34	0.208
EfficientNet-B0	81.5	54.7	+0.05	**+0.15**	+0.06	+0.06	+0.32	0.150
Liver US (clinical anchor: *coarse*)								
ConvNeXt-T	92.5	84.2	+0.04	+0.07	**+0.48**	+0.48	+0.20	–
ConvNeXt-S	95.5	83.3	+0.05	+0.29	**+0.52**	+0.52	+0.26	–
Swin-T	85.9	92.2	+0.02	+0.02	**+0.63**	+0.63	+0.29	–
ViT-S	94.3	76.9	−0.03	**+0.35**	−0.09	−0.09	+0.25	–
DeiT-S	94.8	87.9	−0.06	+0.27	**+0.58**	+0.58	+0.42	–
DenseNet121	88.7	61.7	+0.14	**+0.33**	+0.12	+0.12	+0.46	–
ResNet18	96.2	32.3	**+0.11**	−0.26	−0.24	−0.24	+0.40	–
EfficientNet-B0	94.0	54.9	**+0.26**	+0.01	−0.03	−0.03	+0.48	–
Knee X-ray (clinical anchor: *coarse*)								
ConvNeXt-T	92.4	86.0	+0.38	+0.37	**+0.39**	+0.39	+0.48	–
ConvNeXt-S	92.4	86.5	+0.39	+0.41	**+0.43**	+0.43	+0.49	–
Swin-T	91.3	82.5	+0.35	+0.35	**+0.36**	+0.36	+0.47	–
ViT-S	89.4	83.0	+0.31	+0.31	**+0.36**	+0.36	+0.48	–
DeiT-S	89.5	79.1	+0.11	+0.12	**+0.34**	+0.34	+0.47	–
DenseNet121	90.4	77.8	+0.24	+0.26	**+0.28**	+0.28	+0.46	–
ResNet18	88.7	69.4	+0.21	**+0.24**	+0.21	+0.21	+0.44	–
EfficientNet-B0	87.7	60.9	−0.01	+0.12	**+0.12**	+0.12	+0.41	–
Fundus (clinical anchor: *fine*)								
ConvNeXt-T	89.8	86.4	**+0.27**	+0.25	+0.17	+0.27	+0.33	–
ConvNeXt-S	89.2	86.4	**+0.26**	+0.25	+0.18	+0.26	+0.30	–
Swin-T	88.3	85.7	**+0.30**	+0.24	+0.17	+0.30	+0.33	–
ViT-S	85.7	80.7	**+0.25**	+0.19	+0.17	+0.25	+0.29	–
DeiT-S	86.5	80.1	**+0.22**	+0.20	+0.16	+0.22	+0.27	–
DenseNet121	85.9	78.5	+0.22	**+0.26**	+0.12	+0.22	+0.32	–
ResNet18	85.2	79.5	**+0.27**	+0.16	+0.09	+0.27	+0.30	–
EfficientNet-B0	84.4	71.9	**+0.19**	+0.14	+0.10	+0.19	+0.33	–

6 Discussion

6.1 Clinical-Signal Alignment Beyond Diagnostic Performance

Within each task, clinical-signal alignment was related to diagnostic performance, but the relationship differed across tasks. Diagnostic performance alone therefore could not determine the within-label clinical ordering of a frozen model (Figure 2; §D.1). This shows that clinically ordered responses can differ even among models with similar label discrimination.

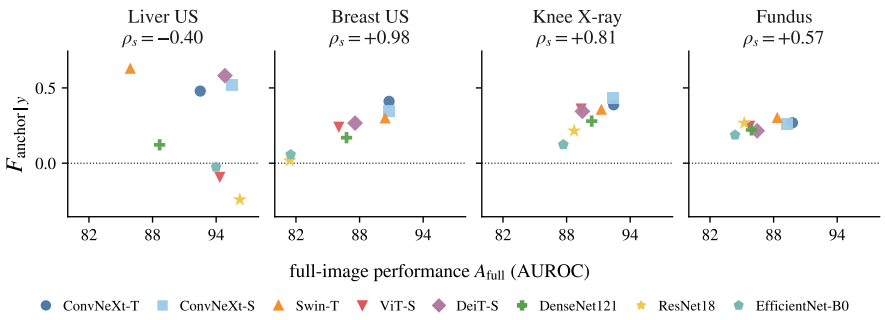

Figure 2: **Diagnostic performance and clinical-signal alignment.** Each point represents one backbone–task cell; ρ_s denotes the within-task Spearman correlation.

The distinction also held for anchor-band discriminability. A_{anchor} covaried strongly with $F_{\text{anchor}|y}$, but A_{anchor} evaluates the diagnostic discrimination of the anchor-band input, whereas $F_{\text{anchor}|y}$ eval-

uates how closely the response follows clinical ordering within the same label. Indeed, shuffle-c moved $F_{\text{anchor}|y}$ to near 0 while leaving A_{anchor} unchanged (§B.2). In the three coarse-anchor tasks, when the adjusted associations of the two quantities with corruption vulnerability and clinical-order retention were compared, the paired crossover contrast was positive ($\Delta = +0.776\,[+0.11, +1.06]$; §D.1). Thus, high anchor-band discriminability does not by itself determine the within-label clinical ordering of the response.

These results show that diagnostic performance and clinically ordered response reflect different aspects of model behavior.

6.2 Behavioral Relevance and Stress-Testing Utility

Clinical-signal alignment measured on the anchor-band probe was associated with full-image clinical-order retention after controlled perturbations. With the three coarse-anchor tasks pooled, models with higher baseline $F_{\text{anchor}|y}$ better preserved clinical ordering, even after adjusting for anchor-band discriminability and clean full-image alignment ($\rho = +0.475\,[+0.17, +0.71]$; Table 9). This relation, between quantities measured on different inputs, shows that clinical-signal alignment is a model characteristic also related to preserving clinically ordered behavior after perturbation.

Across the full panel, higher $F_{\text{anchor}|y}$ was associated with lower dependence on the scale opposite to the prespecified anchor. In the three coarse-anchor tasks, it was also associated with smaller performance loss under controlled corruptions that preferentially degrade the opposite scale (Figure 3; §D.2), even after adjusting for clean diagnostic performance ($\rho = -0.693\,[-0.89, -0.51]$; Table 13). In Fundus, whose anchor is fine, $F_{\text{anchor}|y}$ was negatively associated with the role-defined R_{opp} ($\rho = -0.619\,[-0.81, -0.10]$), whereas its correlation estimate with the physical R_{fine} was near 0 ($+0.048\,[-0.52, +0.31]$; Table 12). The pattern is thus difficult to attribute to a single fixed scale band and may instead depend on the task-specific anchor role, a possibility further examined in the role-matched analysis (§D.2.3).

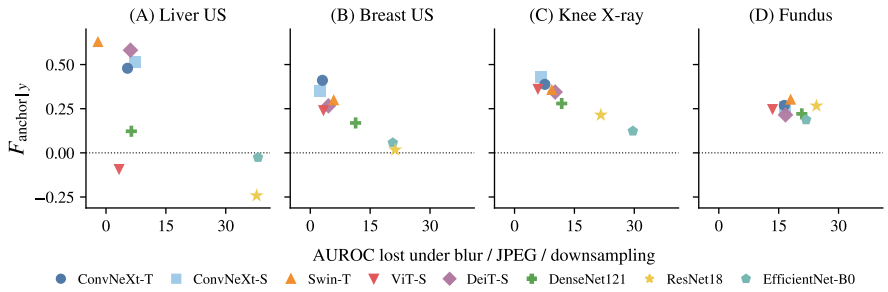

Figure 3: **Clinical-signal alignment and controlled-corruption vulnerability.** Each point represents one backbone–task cell, with the color key of Figure 2. Fundus is shown separately because the same corruptions directly affect its fine anchor.

In the exploratory prioritization over the three coarse-anchor tasks, low-$F_{\text{anchor}|y}$ ranking captured more task-average vulnerability than low-A_{full} at every budget in Table 15, whereas A_{anchor} ranking captured as much as or more than $F_{\text{anchor}|y}$ (§D.2.4). The value of this analysis thus lies less in establishing $F_{\text{anchor}|y}$ as an independent robustness ranking than in allowing clinically ordered response, a separate model characteristic, to be considered alongside clean diagnostic performance when setting follow-up stress-testing priorities.

These results show that clinical-signal alignment links the clinically ordered response measured on the clean model to model behavior after perturbation, and can be used in prioritizing follow-up stress testing under a limited evaluation budget.

6.3 Complementarity to Attribution and Concept Sensitivity

On the 24 Breast US checkpoints, diagnostic performance, Grad-CAM lesion attribution, and $F_{\text{anchor}|y}$ produced different model orderings (Figure 4), and clinical-signal alignment differed even among models with similar lesion attribution mass.

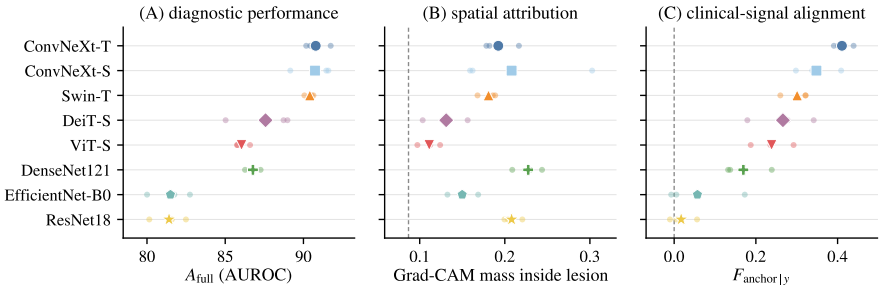

Figure 4: **Diagnostic performance, lesion attribution, and clinical-signal alignment in Breast US.** Small markers are individual checkpoints and large markers are three-seed means; rows are ordered by $F_{\text{anchor}|y}$.

This difference shows that spatial attribution and clinical-signal alignment reflect different evaluation targets: the concentration of attribution on the lesion and within-label clinical ordering need to be evaluated separately (§D.3.1).

A concept-sensitivity measure using the same clinical factor also showed a different pattern. On the Fundus replication panel, both quantities used the same anchor-band inputs and within-y group-preserving permutation draws, and each was calibrated against its own reference distribution; $24/24$ checkpoints exceeded their model-specific 95th percentile for $F_{\text{anchor}|y}$, and $8/24$ for RCV relevance (§D.3.2). Rather than ranking the two methods, this supports the view that, even with the same clinical factor, representation-level concept sensitivity and label-controlled output ordering reflect different evaluation targets.

The two comparisons show that spatial evidence, representation-level sensitivity, and output-level within-label ordering are different evaluation targets, providing grounds for interpreting clinical-signal alignment separately from existing evaluations.

6.4 Forward-Only Applicability to Existing Checkpoints

Applying the audit to public Breast US checkpoints outside our training protocol shows that the same clinical-signal audit specification can be used for existing frozen models. Given an eligible audit cohort and forward access, clinical-signal alignment can be evaluated under the same audit specification without reconstructing the original training pipeline or accessing internal representations. Selection and provenance are described in §C.6.

7 Conclusion

We proposed a retraining-free post-training clinical-signal audit that evaluates how closely a trained diagnostic model follows the ordered variation of a prespecified clinical signal within the same diagnostic label. Across four tasks, clinical-signal alignment revealed differences in clinically ordered response not evident from diagnostic performance alone and, in the three coarse-anchor tasks, was associated with clinical-order retention after controlled perturbations. The results also suggest that the scale-dependent pattern may depend on the task-specific anchor role, and the same audit was applied to existing checkpoints whose original training pipelines are inaccessible. This study is scoped to the post-training characterization of frozen-model responses and their relation to existing diagnostic evaluation and perturbation behavior. These results present clinical-signal alignment as an explicit post-training evaluation target for clinically grounded ordering within the same label, to be considered alongside diagnostic label discrimination. Future work should further verify finding-level validity with fine-grained clinical annotation and the external-cohort utility of audit-guided model selection or adaptation.

AI Use Statement

Generative AI tools were used to create the synthetic modality illustrations in Figure 1, to assist with language polishing, code review and debugging. All AI-assisted suggestions were reviewed and verified by the authors, who take responsibility for the manuscript and code.

Ethics Statement

This study collected no new human-subject data and used four previously released, de-identified retrospective datasets under their respective access and redistribution terms (Byra et al., 2018a; Gómez-Flores et al., 2023; Chen, 2018; EyePACS & California Healthcare Foundation, 2015). No medical images, participant identifiers, or potentially identifying metadata are redistributed in the paper, supplementary material, or released code. The proposed audit is a research evaluation tool, not a clinical decision rule or safety certification, and its outputs do not establish deployment readiness.

Reproducibility Statement

Controlled-panel audit quantities are recomputed from released forward-output caches without retraining. Public-checkpoint tables and figures are regenerated from released aggregate artifacts; the corresponding checkpoint files are pinned by revision and SHA-256, while model-specific loading and inference adapters are not included. The audit and experimental protocols are specified in §4 and Appendices A–C; all train–test splits are fixed at the patient or case level, all split and training seeds are reported, and the released code reproduces the analyses supported by the released artifacts.

References

Julius Adebayo, Michael Muelly, Hal Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2212.04629. 2

R. D. Altman and G. E. Gold. Atlas of individual radiographic features in osteoarthritis, revised. *Osteoarthritis and Cartilage*, 15:A1–A56, 2007. doi: 10.1016/j.joca.2006.11.009. 4

Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, Julius Adebayo, Matthew D. Li, and Jayashree Kalpathy-Cramer. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021. doi: 10.1148/ryai.2021200267. 2

Andrew Bai, Chih-Kuan Yeh, Neil Y. C. Lin, Pradeep Ravikumar, and Cho-Jui Hsieh. Concept Gradient: Concept-based interpretation without linear assumption. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=_01dDd3f78. 3

Stefano Ballestri, Amedeo Lonardo, Dante Romagnoli, Lucia Carulli, Luisa Losi, Christopher P. Day, and Paola Loria. Ultrasonographic fatty liver indicator, a novel score which rules out NASH and is correlated with metabolic parameters in NAFLD. *Liver International*, 32(8):1242–1252, 2012. doi: 10.1111/j.1478-3231.2012.02804.x. 4

biradarakash0909. breast-cancer-ultrasound-mobilenet. Hugging Face model repository, n.d. https://huggingface.co/biradarakash0909/breast-cancer-ultrasound-mobilenet/tree/8dd3b3b283e878e6a2b289969dd3988b03cd1541. Accessed: 2026-08-11. 30

Christopher Boland, Keith A. Goatman, Sotirios A. Tsaftaris, and Sonia Dahdouh. There are no shortcuts to anywhere worth going: Identifying shortcuts in deep learning models for medical image analysis. In *Proceedings of the 7th International Conference on Medical Imaging with Deep Learning (MIDL)*, volume 250 of *Proceedings of Machine Learning Research*, pp. 131–150, 2024. URL https://proceedings.mlr.press/v250/boland24a.html. 3

Alexander Brown, Nenad Tomasev, Jan Freyberg, Yuan Liu, Alan Karthikesalingam, and Jessica Schrouff. Detecting shortcut learning for fair medical AI using shortcut testing. *Nature Communications*, 14:4314, 2023. doi: 10.1038/s41467-023-39902-7. 3

BTX24. hiera-finetuned-busi. Hugging Face model repository, n.d. https://huggingface.co/BTX24/hiera-finetuned-busi/tree/fbeaa620fea3fbe992619f15279ff271db1ac429. Accessed: 2026-08-11. 30

Arianna Bunnell, Yannik Glaser, Dustin Valdez, Thomas Wolfgruber, Aleen Altamirano, Carol Zamora González, Brenda Y. Hernandez, Peter Sadowski, and John A. Shepherd. Learning a clinically-relevant concept bottleneck for lesion detection in breast ultrasound. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 650–659, 2024. doi: 10.1007/978-3-031-72384-1_61. 2

Peter J. Burt and Edward H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983. doi: 10.1109/TCOM.1983.1095851. 3, 18

Michał Byra, Grzegorz Styczynski, Cezary Szmigielski, Piotr Kalinowski, Łukasz Michałowski, Rafał Paluszkiewicz, Bogna Ziarkiewicz-Wróblewska, Krzysztof Zieniewicz, Piotr Sobieraj, and Andrzej Nowicki. Dataset of B-mode fatty liver ultrasound images. https://doi.org/10.5281/zenodo.1009146, 2018a. CC BY 4.0; accessed: 2026-08-25. 10, 20

Michał Byra, Grzegorz Styczynski, Cezary Szmigielski, Piotr Kalinowski, Łukasz Michałowski, Rafał Paluszkiewicz, Bogna Ziarkiewicz-Wróblewska, Krzysztof Zieniewicz, Piotr Sobieraj, and Andrzej Nowicki. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. *International Journal of Computer Assisted Radiology and Surgery*, 13(12):1895–1903, 2018b. doi: 10.1007/s11548-018-1843-2. 5

Pingjun Chen. Knee osteoarthritis severity grading dataset. https://doi.org/10.17632/56rmx5bjcr.1, 2018. CC BY 4.0; organised from the Osteoarthritis Initiative. 5, 10, 20

James Chung-Wai Cheung, Andy Yiu-Chau Tam, Lok-Chun Chan, Ping-Keung Chan, and Chunyi Wen. Superiority of multiple-joint space width over minimum-joint space width approach in the machine learning for radiographic severity and knee osteoarthritis progression. *Biology*, 10(11): 1107, 2021. doi: 10.3390/biology10111107. 4

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*, 9:324, 2026. doi: 10.1038/s41746-026-02497-6. 3

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://www.jmlr.org/papers/v23/20-1335.html. 2

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3:610–619, 2021. doi: 10.1038/s42256-021-00338-7. 1, 3

Alex J. DeGrave, Zhuo Ran Cai, Joseph D. Janizek, Roxana Daneshjou, and Su-In Lee. Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians. *Nature Biomedical Engineering*, 9(3):294–306, 2025. doi: 10.1038/s41551-023-01160-9. 3

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. 21

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy. 21

EyePACS. EyePACS digital retinal image grading: Protocol narrative, grading procedures and rules. https://www.eyepacs.org/consultant/Clinical/grading/EyePACS-DIGITAL-RETINAL-IMAGE-GRADING.pdf, n.d. Grading protocol for the retinal images used in this study. Undated document; accessed: 2026-08-20. 4

EyePACS and California Healthcare Foundation. Diabetic Retinopathy Detection. https://www.kaggle.com/c/diabetic-retinopathy-detection, 2015. Competition data; the competition rules grant no licence to redistribute or publish the images; accessed: 2026-08-25. 6, 10, 20

Giovanna Ferraioli, Viksit Kumar, Arinc Ozturk, Kibo Nam, Chris L. de Korte, and Richard G. Barr. US attenuation for liver fat quantification: An AIUM-RSNA QIBA Pulse-Echo Quantitative Ultrasound Initiative. *Radiology*, 302(3):495–506, 2022. doi: 10.1148/radiol.210736. 1, 4

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. 3

Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems. https://doi.org/10.5281/zenodo.8231412, 2023. Version 1.0; CC BY 4.0; accessed: 2026-08-25. 10, 20

Wilfrido Gómez-Flores, Maria Julia Gregorio-Calas, and Wagner Coelho de Albuquerque Pereira. BUS-BRA: A breast ultrasound dataset for assessing computer-aided diagnosis systems. *Medical Physics*, 51(4):3110–3123, 2024. doi: 10.1002/mp.16812. 5

Mara Graziani, Vincent Andrearczyk, and Henning Müller. Regression concept vectors for bidirectional explanations in histopathology. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications (iMIMIC), MICCAI Workshop*, volume 11038 of *Lecture Notes in Computer Science*, pp. 124–132. Springer, 2018. doi: 10.1007/978-3-030-02628-8_14. 2, 3, 6, 22

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html. 1, 2

Tianyu Han, Riga Wu, Yu Tian, Firas Khader, Lisa C. Adams, Keno K. Bressem, Christos Davatzikos, Jakob Nikolas Kather, Li Shen, David A. Mankoff, Eduardo Mortani Barbosa Jr., and Daniel Truhn. CLEAR: An auditable foundation model for radiology grounded in clinical concepts. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-026-01741-4. Published online 22 July 2026. 3

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90. 21

Andrea S. Hong, Eric L. Rosen, Mary S. Soo, and Jay A. Baker. BI-RADS for sonography: Positive and negative predictive values of sonographic features. *American Journal of Roentgenology*, 184 (4):1260–1265, 2005. doi: 10.2214/ajr.184.4.01841260. 4

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2261–2269, 2017. doi: 10.1109/CVPR.2017.243. 21

Azmul A. Irfan, Nur Ahmad Khatim, Alfan Alfian Irfan, Achmad Zaki, Erike A. Suwarsono, and Mansur M. Arief. Knee-xRAI: An explainable AI framework for automatic Kellgren–Lawrence grading of knee osteoarthritis. https://doi.org/10.48550/arXiv.2604.23435, 2026. URL https://arxiv.org/abs/2604.23435. 4

Mobarakol Islam, Zeju Li, and Ben Glocker. Robustness stress testing in medical image classification. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging (UNSURE), MICCAI Workshop*, volume 14291 of *Lecture Notes in Computer Science*, pp. 167–176, 2023. doi: 10.1007/978-3-031-44336-7_17. 3

ITSheep. breastcancer-ultrasound-ViT. Hugging Face model repository, n.d. https://huggingface.co/ITSheep/breastcancer-ultrasound-ViT/tree/a7bdc0ef5d6851930087ed1b6d5209c6f4ccc5dc. Accessed: 2026-08-11. 30

Ken2707. ultrasound-breast-cancer-V27. Hugging Face model repository, n.d. https://huggingface.co/Ken2707/ultrasound-breast-cancer-V27/tree/5e802fd5ef0d8124071173359076cb99fa2a9504. Accessed: 2026-08-11. 30

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677, 2018. URL https://proceedings.mlr.press/v80/kim18d.html. 2, 3

Chanwoo Kim, Soham U. Gadgil, Alex J. DeGrave, Jesutofunmi A. Omiye, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. Transparent medical image AI via an image–text foundation model grounded in medical literature. *Nature Medicine*, 30(4):1154–1165, 2024. doi: 10.1038/s41591-024-02887-x. 3

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, 2020. URL https://proceedings.mlr.press/v119/koh20a.html. 2

Mark D. Kohn, Adam A. Sassoon, and Navin D. Fernando. Classifications in Brief: Kellgren–Lawrence classification of osteoarthritis. *Clinical Orthopaedics and Related Research*, 474(8):1886–1893, 2016. doi: 10.1007/s11999-016-4732-4. 4

Heon-Ju Kwon, Kyoung Won Kim, Jin-Hee Jung, Sang Hyun Choi, Woo Kyoung Jeong, Bohyun Kim, Gi-Won Song, and Sung-Gyu Lee. Noninvasive quantitative estimation of hepatic steatosis by ultrasound: a comparison of the hepato-renal index and ultrasound attenuation index. *Medical Ultrasonography*, 18(4):431–437, 2016. doi: 10.11152/mu-868. 4

Hsieh-Wei Lee, Bin-Da Liu, King-Chu Hung, Sheau-Fang Lei, Po-Chin Wang, and Tsung-Lung Yang. Breast tumor classification of ultrasound images using wavelet-based channel energy and ImageJ. *IEEE Journal of Selected Topics in Signal Processing*, 3(1):81–93, 2009. doi: 10.1109/JSTSP.2008.2011160. 4

J. W. T. Leung, J. A. Baker, R. Hooley, V. A. Loving, and J. A. Rapelyea. Ultrasound. In *ACR BI-RADS v2025 Manual*. American College of Radiology, Reston, VA, 2025. URL https://www.acr.org/Clinical-Resources/Clinical-Tools-and-Reference/Reporting-and-Data-Systems/BI-RADS. Accessed: 2026-08-25. 1, 4

Xiaomeng Li, Xiaowei Hu, Lequan Yu, Lei Zhu, Chi-Wing Fu, and Pheng-Ann Heng. CANet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. *IEEE Transactions on Medical Imaging*, 39(5):1483–1493, 2020. doi: 10.1109/TMI.2019.2951844. 2

Jennifer I. Lim, Stephen J. Kim, Steven T. Bailey, Jaclyn L. Kovach, G. Atma Vemulakonda, Guishuang Ying, Christina J. Flaxel, and American Academy of Ophthalmology Preferred Practice Pattern Retina/Vitreous Committee. Diabetic Retinopathy Preferred Practice Pattern. *Ophthalmology*, 132(4):P75–P162, 2025. doi: 10.1016/j.ophtha.2024.12.020. 4

Tony Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21(1–2):225–270, 1994. doi: 10.1080/757582976. 3

Mingzhou Liu, Ching-Wen Lee, Xinwei Sun, Xueqing Yu, Yu Qiao, and Yizhou Wang. Learning causal alignment for reliable disease diagnosis. In *International Conference on Learning Representations (ICLR)*, 2025a. URL https://openreview.net/forum?id=ozZG5FXuTV. 1, 2

Qi Liu, Chun Li, Valentine Wanga, and Bryan E. Shepherd. Covariate-adjusted Spearman's rank correlation with probability-scale residuals. *Biometrics*, 74(2):595–605, 2018. doi: 10.1111/biom.12812. 19

Qiong Liu, Chaofan Li, Jinnan Teng, Liping Chen, and Jianxiang Song. Medical image segmentation based on frequency domain decomposition SVD linear attention. *Scientific Reports*, 15:2833, 2025b. doi: 10.1038/s41598-025-86315-1. 3

Xiaoxuan Liu, Ben Glocker, Melissa M. McCradden, Marzyeh Ghassemi, Alastair K. Denniston, and Lauren Oakden-Rayner. The medical algorithmic audit. *The Lancet Digital Health*, 4(5): e384–e397, 2022a. doi: 10.1016/S2589-7500(22)00003-6. 3

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021. doi: 10.1109/ICCV48922.2021.00986. 21

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966–11976, 2022b. doi: 10.1109/CVPR52688.2022.01167. 21

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7. 21

A. C. A. Marijnissen, K. L. Vincken, P. A. J. M. Vos, D. B. F. Saris, M. A. Viergever, J. W. J. Bijlsma, L. W. Bartels, and F. P. J. G. Lafeber. Knee Images Digital Analysis (KIDA): a novel method to quantify individual radiographic features of knee osteoarthritis in detail. *Osteoarthritis and Cartilage*, 16(2):234–243, 2008. doi: 10.1016/j.joca.2007.06.009. 4

MarvelousXHJ. vit-finetuning-breastcancer-ultrasound-images. Hugging Face model repository, n.d. https://huggingface.co/MarvelousXHJ/vit-finetuning-breastcancer-ultrasound-images/tree/a177268d3f6c11af4c26259ec6ecf9986041177b. Accessed: 2026-08-11. 30

Woo Kyung Moon, Chung-Ming Lo, Jung Min Chang, Chiun-Sheng Huang, Jeon-Hor Chen, and Ruey-Feng Chang. Quantitative ultrasound analysis for classification of BI-RADS category 3 breast masses. *Journal of Digital Imaging*, 26(6):1091–1098, 2013. doi: 10.1007/s10278-013-9593-8. 1, 4, 29

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning (CHIL)*, pp. 151–159, 2020. doi: 10.1145/3368555.3384468. 1, 3

Cathy Ong Ly, Balagopal Unnikrishnan, Tony Tadic, Tirth Patel, Joe Duhamel, Sonja Kandel, Yasbanoo Moayedi, Michael Brudno, Andrew Hope, Heather Ross, and Chris McIntosh. Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data. *npj Digital Medicine*, 7(1):124, 2024. doi: 10.1038/s41746-024-01118-4. 3

Osteoarthritis Initiative. The Osteoarthritis Initiative. https://nda.nih.gov/oai/, n.d. NIH data archive; a public–private partnership funded by the National Institutes of Health; accessed: 2026-08-25. 20

Frederik Pahde, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Ensuring medical AI safety: Interpretability-driven detection and mitigation of spurious model behavior and associated data. *Machine Learning*, 114(9):206, 2025. doi: 10.1007/s10994-025-06834-w. 3

Gwénolé Quellec, Mathieu Lamard, Pierre Marie Josselin, Guy Cazuguel, Béatrice Cochener, and Christian Roux. Detection of lesions in retina photographs based on the wavelet transform. In *Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2618–2621, 2006. doi: 10.1109/IEMBS.2006.260220. 4

Laura Rieger, Chandan Singh, W. James Murdoch, and Bin Yu. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8116–8126, 2020. URL https://proceedings.mlr.press/v119/rieger20a.html. 1, 2

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. 1, 2

Raphael Schäfer, Till Nicke, Henning Höfener, Annkristin Lange, Dorit Merhof, Friedrich Feuerhake, Volkmar Schulz, Johannes Lotz, and Fabian Kiessling. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nature Computational Science*, 4(7):495–509, 2024. doi: 10.1038/s43588-024-00662-z. 2

Peter Schulam and Suchi Saria. Can you trust this prediction? Auditing pointwise reliability after learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1022–1031, 2019. URL https://proceedings.mlr.press/v89/schulam19a.html. 3

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74. 1, 2

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 1827–1858, 2022. doi: 10.1145/3531146.3533232. 2

sergiopaniego. fine_tuning_vit_custom_dataset_breastcancer-ultrasound-images. Hugging Face model repository, n.d. https://huggingface.co/sergiopaniego/fine_tuning_vit_custom_dataset_breastcancer-ultrasound-images/tree/d2474044753880d5463402737f9edbc1c09b8ea3. Accessed: 2026-08-11. 30

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021. doi: 10.1038/s41591-021-01595-0. 3

sidushdid. ViT-base-patch16-BUSI-Mendeley-AdamW. Hugging Face model repository, n.d.a. https://huggingface.co/sidushdid/ViT-base-patch16-BUSI-Mendeley-AdamW/tree/e15fc9fddbbba15b8aa200d18147f738a34b389d. Accessed: 2026-08-11. 30

sidushdid. ViT-base-patch16-BUSI-Mendeley-Lion. Hugging Face model repository, n.d.b. https://huggingface.co/sidushdid/ViT-base-patch16-BUSI-Mendeley-Lion/tree/edcb3ce2703fe829d36b64b0d7456cde59836205. Accessed: 2026-08-11. 30

sidushdid. ViT-base-patch16-BUSI-Mendeley-Adafactor. Hugging Face model repository, n.d.c. https://huggingface.co/sidushdid/ViT-base-patch16-BUSI-Mendeley-Adafactor/tree/92fdd4ac33a2bb963fba492ff1eee351a7265079. Accessed: 2026-08-11. 30

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114, 2019. URL https://proceedings.mlr.press/v97/tan19a.html. 21

Arun James Thirunavukarasu, Kabilan Elangovan, Laura Gutierrez, Yong Li, Iris Tan, Pearse A. Keane, Edward Korot, and Daniel Shu Wei Ting. Democratizing artificial intelligence imaging analysis with automated machine learning: Tutorial. *Journal of Medical Internet Research*, 25: e49949, 2023. doi: 10.2196/49949. 2

Aleksei Tiulpin and Simo Saarakkala. Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks. *Diagnostics*, 10(11):932, 2020. doi: 10.3390/diagnostics10110932. 2

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357, 2021. URL https://proceedings.mlr.press/v139/touvron21a.html. 21

Balagopal Unnikrishnan, Michael Brudno, and Chris McIntosh. SilverLining: Data-first mitigation of spatial and spectral shortcuts without introducing new confounders. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1294–1303, 2026. doi: 10.1109/WACV61042.2026.00132. 3

Vibujithan Vigneshwaran, Emma A. M. Stanley, Raissa Souza, Erik Ohara, Matthias Wilms, and Nils Forkert. Evaluating shortcut utilization in deep learning disease classification through counterfactual analysis. In *Proceedings of the 8th International Conference on Medical Imaging with Deep Learning (MIDL)*, volume 301 of *Proceedings of Machine Learning Research*, pp. 1653–1668, 2026. URL https://proceedings.mlr.press/v301/vigneshwaran26a.html. 3

Thi Lan Nhi Vu, Vimal Chandran, Christina Haberl, Otmar Scherzer, and Julia Binder. Multiresolution tone mapping for high dynamic range medical ultrasound images. *PLOS One*, 21(1): e0340777, 2026. doi: 10.1371/journal.pone.0340777. 3

Ross Wightman. PyTorch Image Models. https://github.com/huggingface/pytorch-image-models, 2019. Software; DOI 10.5281/zenodo.4414861. Accessed: 2026-08-28. 21

C. P. Wilkinson, Frederick L. Ferris, III, Ronald E. Klein, Paul P. Lee, Carl David Agardh, Matthew Davis, Diana Dills, Anselm Kampik, R. Pararajasegaram, Juan T. Verdaguer, and Global Diabetic Retinopathy Project Group. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9):1677–1682, 2003. doi: 10.1016/S0161-6420(03)00475-5. 4

Chun Kit Wong, Paraskevas Pegios, Nina Weng, Emilie Pi Fogtmann Sejer, Martin Grønnebæk Tolsgaard, Anders Nymark Christensen, and Aasa Feragen. Weight space correlation analysis: Quantifying feature utilization in deep learning models. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning (MIDL)*, volume 315 of *Proceedings of Machine Learning Research*, pp. 2711–2737, 2026. URL https://proceedings.mlr.press/v315/wong26a.html. 3

Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E. Ho, and James Zou. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nature Medicine*, 27(4):582–584, 2021. doi: 10.1038/s41591-021-01312-x. 2

Albert T. Young, Kristen Fernandez, Jacob Pfau, Rasika Reddy, Nhat Anh Cao, Max Y. von Franque, Arjun Johal, Benjamin V. Wu, Rachel R. Wu, Jennifer Y. Chen, Raj P. Fadadu, Juan A. Vasquez, Andrew Tam, Michael J. Keiser, and Maria L. Wei. Stress testing reveals gaps in clinic readiness of image-based diagnostic artificial intelligence models. *npj Digital Medicine*, 4:10, 2021. doi: 10.1038/s41746-020-00380-6. 3

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. doi: 10.1371/journal.pmed.1002683. 1, 3

Boyu Zhang, Aleksandar Vakanski, and Min Xian. BI-RADS-Net: An explainable multitask learning approach for cancer diagnosis in breast ultrasound images. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2021. doi: 10.1109/MLSP52302.2021.9596314. 2

Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *Medical Image Analysis*, 91:102996, 2024. doi: 10.1016/j.media.2023.102996. 2

The appendix has four parts.
§A presents the scale-band operator that produces the audit inputs and its calibration.
§B presents the estimator and permutation controls for the audit quantities.
§C presents the experimental protocols for data, training, attribution, corruption, and public checkpoints.
§D presents additional results and sensitivity analyses corresponding to the four subsections of §6.

A Scale-Band Operator and Calibration

The task-specific clinical anchor scales were prespecified from clinical grading criteria and the quantitative imaging literature (Table 1). The relative fine–mid–coarse ordering of the fixed scale-band operator was calibrated with synthetic inputs independent of the clinical factor. No independent reader study of the anchor mapping was performed.

Fixed Laplacian-pyramid operator. We build a Gaussian pyramid $G_0 \to G_1 \to G_2$ and define

$$L_0 = G_0 - \mathrm{expand}(G_1), \qquad L_1 = G_1 - \mathrm{expand}(G_2). \tag{3}$$

Represented at the original image resolution, L_0, L_1, and G_2 give $\tilde{x}_{\text{fine}}$, $\tilde{x}_{\text{mid}}$, and $\tilde{x}_{\text{coarse}}$, the raw scale-band components of §4.1; the model query input x_b is obtained from them by per-image normalization and the common preprocessing. Each *reduce* halves each side by bilinear resampling after σ=1 Gaussian smoothing, and *expand* restores the original resolution by bilinear resampling. This follows the Laplacian-pyramid construction of Burt & Adelson (1983), with the smoothing and resampling operations specified above.

Sinusoidal-grating calibration. We passed sinusoidal gratings with periods of 2–224px, varying orientation and phase, through the fixed operator and measured the share of energy in each band (Table 3). The fine band responded most strongly at a 2.5px period and the mid band peaked at 11px, while the coarse response increased toward longer periods. The band with the largest energy switched from fine to mid near 8px and was coarse at all orientations and phases from 16px upward.

Table 3: **Responses of the fixed scale bands to sinusoidal gratings.** The share is the band's mean-removed squared-L_2 energy divided by the summed energy of the three bands, averaged over orientations and phases.

Band	Max-response period in sweep (share)	Half-maximum range	Modal-argmax window
fine	2.5px (0.999)	2–6px	$\leq$ 6px
mid	11px (0.670)	8–16px	8–11px
coarse	224px (1.000)	$\geq$ 16px	$\geq$ 16px

Reconstruction residual. Because the components use different upsampling paths, their sum $\hat{x} = \tilde{x}_{\text{fine}} + \tilde{x}_{\text{mid}} + \tilde{x}_{\text{coarse}}$ does not exactly equal the input. We define the relative residual as $\|x - \hat{x}\|_2 / \|x\|_2$. For each task, 300 images were sampled without replacement with seed 0 from the source image pool used for this calibration analysis; the median was 0.012 for Liver US, 0.009 for Breast US, 0.002 for Knee X-ray, and 0.007 for Fundus, and the maximum did not exceed 0.018.

B Audit Estimation and Controls

B.1 Estimator and Tie Handling

Because all audit quantities depend on the model response through its ranks or AUROC, strictly increasing transformations of the response do not change them.

Estimator. We use $F_{\text{anchor}|y}$ as an operational rank statistic for comparing frozen-model responses within the same audit cohort. Following §4.2, it is the Pearson correlation between the mean-rank-transformed response and clinical factor after removing the rank effect of the diagnostic label from each, which corresponds to the traditional partial Spearman statistic. The population-parameter interpretation of the traditional partial Spearman has limitations, and an alternative based on probability-scale residuals has been proposed (Liu et al., 2018). Conventions for subsequent model-level associations are given in §B.4.

Ties and mean ranks. All rank transforms assign mean ranks to tied values. The clinical factor c takes 13, 4, 5, and 5 distinct values in the Liver US, Breast US, Knee X-ray, and Fundus audit cohorts, shared by 170, 561, 1289, and 1375 images, respectively. Ties occur even in Liver US, whose clinical factor is a continuous biopsy percentage: images from the same patient share one value, so the 17 test patients yield 13 distinct values.

An ordinal-rank implementation that breaks ties by array order assigns different ranks to equal values; under the within-y permutation control, it produced an artificial alignment of $+0.18$ on average. With mean ranks, the order of tied observations does not enter the statistic.

Estimator sensitivity. To check sensitivity to the estimator choice, we compared against a variant that re-ranks the residuals and computes their Spearman correlation. In this variant, residuals within 10^{-9} were treated as ties so that numerically equal residuals are not separated by floating-point error. Across the 32 backbone–task cells, the two estimators differed by at most 0.061, the Spearman correlation of model ordering across cells was 0.996, and signs agreed in $32/32$. The cross-cell ordering and signs of alignment were therefore stable to this estimator choice.

B.2 Permutation Controls

For each checkpoint, we ran two permutation controls with $B{=}200$ each. *Shuffle-c* permutes the clinical factor c within each diagnostic class, and *shuffle-s* permutes the anchor-band response s_{anchor} across all samples. Both preserve the marginal distribution of each variable while removing the sample correspondence between the clinical factor and the model response. Shuffle-c leaves s_{anchor} intact and thus exactly preserves A_{anchor}, whereas shuffle-s also removes the relation between s_{anchor} and y.

The reference distribution of each backbone–task cell pools the permutation values from its three seeds, giving 600 values per cell and 19,200 over all 96 checkpoints for each control. The shuffle-c distribution had mean $+0.001$ and standard deviation 0.050, and shuffle-s had 0.000 and 0.048. The pooled 95th percentiles of $|F_{\text{anchor}|y}|$ were 0.107 and 0.101, respectively. The observed values in the main text (mean $+0.258$, range -0.242 to $+0.631$) exceeded the 95th percentile of the corresponding image-level permutation distribution in 27 of 32 cells.

Cell-level reference. We also built a reference matched in unit to the observed values by applying the same permutation to the three seeds and averaging. Its 95th percentile of $|F_{\text{anchor}|y}|$ was 0.889 times that of the seed-level reference (median over cells), and 29 of 32 cells exceeded it. The main text reports the seed-level judgments.

Group-preserving permutation reference. Some audit cohorts contain multiple images from the same patient or case, inducing within-group dependence. To preserve it, we separately built a group-preserving permutation reference that reassigns whole clinical-factor vectors among groups with compatible size and diagnostic-label composition ($B{=}5000$). Under this reference, the 95th percentile of $|F_{\text{anchor}|y}|$ rose by a median factor over cells of 3.51 for Fundus, 3.02 for Liver US, 1.28 for Breast US, and 1.26 for Knee X-ray. Of the 32 backbone–task cells, 26 exceeded its 95th

percentile. Relative to the 29 under the unit-matched cell-level reference, the decrease occurred in Liver US ($6 \to 4$) and Breast US ($7 \to 6$), whereas Knee X-ray and Fundus remained at $8/8$. Two cells (Liver US ResNet18 and Breast US EfficientNet-B0) lay very close to this 95th percentile. We use the group-preserving permutation as a sensitivity analysis for the repeated-measures structure and apply the same permutation scheme to the concept-sensitivity comparison in §C.4.

Shuffle-c removes only the sample-wise correspondence between c and s_{anchor} while exactly preserving A_{anchor}; the return of $F_{\text{anchor}|y}$ to near 0 under shuffle-c confirms that the observed alignment depends on this correspondence.

B.3 Input Normalization and Band Removal

We refer to the image-wise $[0, 1]$ min–max normalization of the main text as *per-image* normalization. The band-removed input $x_{\setminus b}$ in Eq. (2) is built from the pre-normalization pyramid components $\tilde{x}_b$: the components other than band b are summed, and the reconstructed image is clipped to the valid intensity range. No separate per-image normalization is applied to $x_{\setminus b}$; after clipping, it receives the same common preprocessing as x_b.

We compared per-image normalization with two alternatives. *Global* normalization applies a fixed linear map using dataset-level 1st/99th percentiles and clips to $[0, 1]$, and *signed* normalization is

$$x_b^{\text{signed}} = \text{clip}_{[0,1]}\left(\frac{\tilde{x}_b}{2\tau_b} + 0.5\right), \qquad \tau_b = \text{percentile}_{99}(|\tilde{x}_b|)\,. \tag{4}$$

Because the alternative normalizations require new forward queries, the sensitivity analysis used replication sets with weight access: 4 backbones for Liver US and, for Breast US, 24 checkpoints corresponding to the same 8 backbones as the controlled panel.

On the Breast US panel, the Spearman correlation of backbone ordering was $+0.98$ for both per-image versus global and per-image versus signed, with mean $|\Delta F_{\text{anchor}|y}|$ of 0.051 and 0.057, respectively. Cells whose sign changed all had per-image $|F_{\text{anchor}|y}| \leq 0.17$, and cells with $|F_{\text{anchor}|y}| > 0.19$ kept their sign under all three normalizations. In the Liver US replication, ResNet18 was negative under all three (per-image -0.38, global -0.25, signed -0.14), and ConvNeXt-T, ConvNeXt-S, and Swin-T all remained positive. In these replication sets, the main model ordering and sign pattern were largely preserved across the three normalization choices.

B.4 Model-Level Correlation Conventions

These conventions apply to all model-level associations in §D. The performance–alignment, anchor-discriminability, and opposite-scale-dependence analyses use Pearson correlations of rank quantities, computed between rank residuals when covariates are adjusted. The corruption-vulnerability adjustment, clinical-order retention, Fundus role-matched analysis, and corruption breakdown use Spearman correlations between rank residuals after covariate adjustment. Within-task ranking and resampling units are stated for each analysis.

Residual correlations are computed only when both residual vectors are non-degenerate and are recorded as undefined otherwise. The degeneracy criterion and the number of valid bootstrap replicates are given in the released code and artifacts.

C Experimental Protocols

C.1 Data, Preprocessing, and Splits

Datasets and preprocessing. The four datasets are Liver US (Byra et al., 2018a), Breast US (Gómez-Flores et al., 2023), Knee X-ray (Osteoarthritis Initiative, n.d.; Chen, 2018), and Fundus (EyePACS & California Healthcare Foundation, 2015); access and redistribution terms are given in the Ethics Statement. Because only Fundus is in color, all images were converted to grayscale and resized to 224×224 by bilinear interpolation so that the same scale-band decomposition and band-restricted input protocol apply to all four tasks. For Fundus, the circular field of view was cropped before resizing to the bounding box of pixel values above 12 (no crop if that region spans fewer than

32 rows or columns), and no further local-contrast enhancement was applied. Because the Gaussian background subtraction of Ben Graham preprocessing alters the spatial-frequency distribution of the image, it was omitted for the Fundus task, which audits a fine-scale signal.

Audit cohorts and splits. The audit cohorts of §5.1 are fixed test splits defined once at the patient or case level. Liver US comprises 10 images from each of 55 patients, and its test split has 170 images from 17 patients. Breast US uses 561 images from 320 of its 1,064 cases (1,875 images) for testing. For OAI and EyePACS, after `random.seed(0)`, images were shuffled within each grade and subsampled to at most 1,000 per grade before the same group-wise split, yielding final audit cohorts of 1,289 images from 919 cases and 1,375 images from 1,100 cases, respectively. The group identifier is the participant number at the start of the file name for OAI (shared by left and right knees) and the patient part of the `<patient>_<side>` file name for EyePACS.

All splits were generated once with `GroupShuffleSplit(test_size=0.30, random_state=100)` and fixed across all experiments. The training seeds of each backbone–task cell are 8000, 8001, and 8002.

C.2 Model Training

The controlled panel uses ResNet (He et al., 2016), DenseNet (Huang et al., 2017), EfficientNet (Tan & Le, 2019), ViT (Dosovitskiy et al., 2021), DeiT (Touvron et al., 2021), Swin Transformer (Liu et al., 2021), and ConvNeXt (Liu et al., 2022b) backbones, all initialized from timm (Wightman, 2019) checkpoints pretrained on ImageNet (Deng et al., 2009). Common settings are AdamW (Loshchilov & Hutter, 2019), a classification-head learning rate of 5×10^{-4}, weight decay 10^{-4}, batch size 16, and cross-entropy loss weighted by inverse class frequency in the training split. Inputs are $[0, 1]$ grayscale images replicated to three channels and normalized with ImageNet statistics; band-restricted inputs receive the same preprocessing.

No data augmentation, validation-based model selection, or early stopping was used, and cuDNN was set to deterministic mode. Each backbone was trained for a prespecified number of epochs, and the final checkpoint was audited. Per-backbone encoder learning rates and epochs are listed in Table 4.

Table 4: **Per-backbone training settings.**

Backbone	timm checkpoint	Encoder lr	Epochs
ConvNeXt-T	`convnext_tiny.fb_in1k`	10^{-4}	16
ConvNeXt-S	`convnext_small.fb_in1k`	10^{-4}	16
Swin-T	`swin_tiny_patch4_window7_224.ms_in1k`	3×10^{-5}	20
ViT-S	`vit_small_patch16_224.augreg_in21k_ft_in1k`	10^{-5}	12
DeiT-S	`deit_small_patch16_224.fb_in1k`	10^{-5}	12
DenseNet121	`densenet121.ra_in1k`	10^{-4}	20
ResNet18	`resnet18.a1_in1k`	10^{-4}	20
EfficientNet-B0	`efficientnet_b0.ra_in1k`	10^{-4}	20

C.3 Attribution and Localization Protocol

Attribution is used only in the Breast US case study of §6.3. For the 24 checkpoints, Grad-CAM was computed on the same fixed image subset with each image's ground-truth class as the target, and localization was evaluated against lesion masks. Sensitivity to the attribution-target convention is reported in §D.3.1.

Scored images. From the 561 Breast US test images, 160 were selected at evenly spaced indices from the first to the last image. The subset was not stratified by diagnostic label and was shared by all checkpoints; it contains 111 benign and 49 malignant images.

Grad-CAM implementation. For each backbone, the last spatial representation served as the target layer.

Backbone	Target layer
ResNet18	`layer4.1.act2`
DenseNet121	`features.norm5`
EfficientNet-B0	`bn2.act`
ConvNeXt-T/S	`norm_pre`
Swin-T	`layers.3.blocks.1`
ViT-S / DeiT-S	`blocks.11.norm1`

Swin-T activations were transposed to a channel-first layout, and for ViT-S and DeiT-S the class token was removed and the remaining tokens were reshaped into a square spatial grid. Activations were weighted by channel-wise mean gradients, and a ReLU was applied to the weighted sum. The resulting map was bilinearly upsampled to 224×224 and normalized by its per-image maximum.

Localization measures. For an attribution map H and lesion mask $M \subseteq \Omega$, we define

$$\text{mass} = \frac{\sum_{p \in M} H(p)}{\sum_{p \in \Omega} H(p)}, \qquad \text{pointing} = \mathbf{1}\left[\arg\max_p H(p) \in M\right], \qquad \text{IoU} = \frac{|\hat{M} \cap M|}{|\hat{M} \cup M|}, \tag{5}$$

where $\hat{M}$ is the top-20% pixel set of H. Each measure is computed per image and then averaged.

C.4 Concept-Sensitivity Comparison Protocol

RCV (Graziani et al., 2018) finds the direction of the clinical factor in an intermediate activation by regression and uses the directional derivative of the prediction along that direction. The comparison was run on the 24 checkpoints of the Fundus replication panel (§D.2.3), using the same x_{anchor} inputs as $F_{\text{anchor}|y}$.

The layer was fixed to the output of the penultimate stage for staged architectures and to the 9th of 12 blocks for ViT and DeiT; per-model layers are specified in the released code. We do not use the pre-classifier representation: because the path from it to the logit is linear, the sign of the sensitivity is fixed regardless of the sample, which makes the sign-aggregation variant degenerate.

For each checkpoint, activations are spatially averaged per channel to give $\Phi \in \mathbb{R}^{n \times D}$, and the clinical factor is regressed on the standardized Φ by least squares to obtain a unit concept vector v and the coefficient of determination R^2 of that regression. The per-sample sensitivity is $S_i = \langle \nabla_\Phi f_\theta(x_i), v \rangle$, where f_θ is the positive-class softmax probability defined in §4.1. Following the original definition, the bidirectional relevance is

$$\text{Br} = R^2 \frac{\hat{\mu}_S}{\hat{\sigma}_S}, \tag{6}$$

and we use $|\text{Br}|$ as the statistic to assess the magnitude of relevance regardless of sign. The original also rescales the Br of multiple concepts to $[-1, 1]$ by dividing by their maximum absolute value, but each audit task has a single clinical factor, for which this normalization degenerates to ± 1; we therefore use unnormalized Br. The concept vector and R^2 are fit on one fold and $\hat{\mu}_S$ and $\hat{\sigma}_S$ on the other, patient-disjoint fold, and the result is averaged over the two fold assignments.

The reference distribution applies the group-preserving permutation of §B.2 B=2000 times, refitting the concept vector and R^2 at every draw, with empirical $p = (1 + \#\{|\text{Br}^\pi| \geq |\text{Br}|\})/(B + 1)$. The same permutation sequence is applied to $F_{\text{anchor}|y}$, but because the two quantities are different statistics, each has its own reference distribution and threshold. As aggregation sensitivity analyses, we also compute a label-controlled variant that residualizes the clinical factor on y and a TCAV-style variant that aggregates only the sign of the sensitivity, $\frac{1}{n}\sum_i \mathbf{1}[S_i > 0]$.

C.5 Corruption Protocols

Corruption operators. The three corruptions were applied to all tasks with the same parameters; we denote them $\mathcal{T} = \{\text{blur, JPEG, downsampling}\}$. Gaussian blur applies `scipy.ndimage.gaussian_filter` with σ=2.2 to the $[0, 1]$ image, with the function's default reflection boundary. JPEG compression saves the single-channel image with PIL as a grayscale JPEG at quality 12 and reloads it. Downsampling uses PIL `resize` to reduce each side

to 56 pixels and back to 224×224, with PIL's default (bicubic) interpolation in both steps, unlike the bilinear resize used in data preparation. Corruptions were applied directly to the full image, without prior scale-band decomposition. The prespecified role-matched operators for Fundus, whose anchor is fine, are used in the separate analysis of §D.2.3.

Scale-band effect. Each clean and corrupted image was decomposed with the same fixed scale-band operator, and the corrupted-to-clean energy ratio was computed per band. Band energy is the sum of squared raw-component coefficients, $\|\tilde{x}_b\|_2^2$, and task-level values are medians of per-image ratios over 300 images sampled with a fixed seed. No model output or clinical factor was used in this computation.

In all 12 task–corruption combinations, the fine band retained the least energy (Table 5). Blur and downsampling left 0.030–0.120 of the fine-band energy while preserving at least 0.984 of the coarse-band energy. JPEG compression changed the bands less, leaving 0.631–0.940 of the fine-band energy.

Table 5: **Scale-band energy retained under the three corruptions.** Bold marks the band with the least retained energy in each row. Per image, the ratio of band energy after to before the corruption; entries are medians over 300 images sampled without replacement with seed 0 from each task's source image pool.

Task	Corruption	fine	mid	coarse
Liver US	blur	**0.030**	0.293	0.984
	JPEG	**0.631**	0.942	1.006
	downsampling	**0.061**	0.451	1.002
Breast US	blur	**0.030**	0.262	0.994
	JPEG	**0.940**	1.018	1.000
	downsampling	**0.067**	0.442	0.999
Knee X-ray	blur	**0.036**	0.369	0.999
	JPEG	**0.713**	1.118	1.000
	downsampling	**0.091**	0.577	1.000
Fundus	blur	**0.062**	0.379	0.996
	JPEG	**0.932**	1.003	1.001
	downsampling	**0.120**	0.554	1.000

Fundus role-matched operator. Because the Fundus anchor is fine, the three corruptions above degrade the anchor band itself. For a role-matched contrast, we used three operators that perturb only the coarse band and leave the other two components unchanged. *Attenuation* multiplies the coarse component by 0.5; *block shuffle* shuffles the coarse component in 32-pixel blocks and projects it back onto the same band; and *band-limited bias* adds the coarse component of standard normal noise, divided by its standard deviation and scaled by that of the original coarse component. In all three, the input is reconstructed by summing the three components and clipping to the valid intensity range. The noise seed was fixed at 20260811, and the composite score is the mean over the three operators. We also computed an anchor-targeted control that applies the same operators to the fine band.

C.6 Public Checkpoint Selection and Audit

Selection and audit protocol. We searched for public Breast US classifiers via the Hugging Face API and GitHub search, and included checkpoints for which no BI-RADS supervision was indicated in public documentation and that were accessible and loadable at submission time.

Of the 11 checkpoints, 7 used the preprocessing configuration provided by their repositories. For the remaining 4, whose preprocessing and class ordering were undocumented, we evaluated both $[0, 1]$ and ImageNet normalization and used the interface with the higher full-image AUROC for the pathology label y, which selected ImageNet normalization for 2 and $[0, 1]$ normalization for 2. Interface selection used full-image AUROC for the diagnostic label y on the same audit cohort and did not use the clinical factor c or $F_{\text{anchor}|y}$. The full-image AUROC of these 4 checkpoints is therefore an estimate after interface selection within the same cohort.

Each checkpoint's permutation reference is the distribution of $|F_{\text{anchor}|y}|$ over $B{=}2000$ permutations of c within diagnostic classes, and the main text compares each observed value with the 95th percentile of its model-specific distribution. The judgments for all 11 checkpoints were unchanged when the number of permutations was increased from 200 to 2000 or the permutation seed was changed.

Eligibility and provenance. Public repository documentation alone could not fully verify the training data, any undisclosed clinical supervision, or image-level overlap with BUS-BRA. Four repository names include BUSI, but we did not treat a repository name as documented training provenance. No repository stated use of BUS-BRA.

The public-checkpoint audit was applied to Breast US, which met the audit-eligibility and documentation requirements. Of the 50 Fundus candidates examined, 10 of the 11 with readable configurations used the DR grade itself as the output label, so the withheld factor c was part of their training supervision (§3). For the 6 Knee X-ray candidates and the 1 Liver US candidate, no configuration sufficient to judge audit eligibility could be obtained.

D Additional Results

This section follows the four subsections of §6 in order and presents the detailed results and sensitivity analyses supporting each.

D.1 Performance, Discriminability, and Model Stability

Performance–alignment control. With backbones ranked within each task and the 32 backbone–task cells pooled, $F_{\text{anchor}|y}$ showed a positive relation with full-image diagnostic performance A_{full} ($\rho = +0.488\,[+0.16, +0.75]$). Task-wise relations ranged from -0.405 in Liver US to $+0.976$ in Breast US. After controlling for A_{anchor}, the partial correlation between A_{full} and $F_{\text{anchor}|y}$ was $+0.016\,[-0.46, +0.52]$.

Relation to anchor-band discriminability. Table 6 reports the partial associations among $F_{\text{anchor}|y}$, A_{anchor}, and corruption vulnerability. $F_{\text{anchor}|y}$ and A_{anchor} covaried at $+0.808$ at the checkpoint level and $+0.893\,[+0.68, +0.98]$ at the cell level.

Table 6: **Adjusted associations with opposite-scale dependence and corruption vulnerability.** Checkpoint-level rank correlations, pooled after ranking within task; intervals resample backbone cells. Each row is shown with and without the other quantity partialled out.

Outcome	Predictor	unadjusted	adjusted for the other
R_{opp} (4 tasks, $N{=}96$)	$F_{\text{anchor}\|y}$	$-0.607\,[-0.72, -0.33]$	$-0.077\,[-0.32, +0.08]$
	A_{anchor}	$-0.711\,[-0.82, -0.45]$	$-0.472\,[-0.66, -0.24]$
vulnerability (coarse-anchor, $N{=}72$)	$F_{\text{anchor}\|y}$	$-0.683\,[-0.81, -0.41]$	$-0.035\,[-0.39, +0.13]$
	A_{anchor}	$-0.781\,[-0.88, -0.52]$	$-0.518\,[-0.70, -0.26]$

Because $F_{\text{anchor}|y}$ and A_{anchor} covary strongly, we do not interpret the unique associations of the two quantities with corruption vulnerability separately.

Crossover analysis. Table 7 relates the two audit quantities to two criteria of a different kind, corruption vulnerability (discrimination) and clinical-order retention (ordering). With both criteria oriented so that higher values are favorable, the paired crossover contrast was $\Delta = +0.776\,[+0.11, +1.06]$. The row difference in each column and its confidence interval are given in the table.

Table 7: **Association pattern of the two audit quantities across two criteria.** Partial rank correlations over the 72 coarse-anchor checkpoints, computed on the Spearman path of §B.4. *Vulnerability* is the mean diagnostic AUROC lost under the three standard corruptions and is adjusted for the other audit quantity; *retention* is D of Eq. (7) and is adjusted for the other audit quantity and for clean $F_{\text{full}|y}$. Writing $Q = -\text{vulnerability}$ so that a higher value is the favourable direction in both columns, the row differences are $d_{\text{vul}} = r(A_{\text{anchor}}, Q) - r(F_{\text{anchor}|y}, Q)$ and $d_{\text{ret}} = r(F_{\text{anchor}|y}, D) - r(A_{\text{anchor}}, D)$, and the crossover contrast is $\Delta = d_{\text{vul}} + d_{\text{ret}}$. All seven quantities are computed on the same bootstrap draws, which resample backbone cells within each task and keep the three seeds of a cell together.

Audit quantity	vulnerability (discrimination)	retention (ordering)
A_{anchor}	-0.539 $[-0.67, -0.25]$	$+0.241$ $[-0.06, +0.54]$
$F_{\text{anchor}\|y}$	$+0.004$ $[-0.39, +0.14]$	$+0.475$ $[+0.17, +0.71]$
Row difference	d_{vul} $+0.543$ $[-0.03, +0.72]$	d_{ret} $+0.233$ $[-0.31, +0.71]$
Crossover contrast Δ		$+0.776$ $[+0.11, +1.06]$

Seed stability. Each backbone–task cell in Table 2 is a three-seed mean. We assessed checkpoint-level stability by the variance of backbone means, the mean within-backbone seed variance, and backbone rank correlations across seed pairs.

Table 8: **Seed stability of $F_{\text{anchor}|y}$.** For each task the table reports the variance across the eight three-seed backbone means, the mean within-backbone variance across the three seeds, and the range of backbone rank correlations over the three seed pairs.

Task	Checkpoints	Variance of backbone means	Mean within-backbone variance	Seed-swap rank ρ
Liver US	24	0.1187	0.0294	$+0.62$ to $+0.81$
Breast US	24	0.0188	0.0037	$+0.76$ to $+0.81$
Knee X-ray	24	0.0102	0.0013	$+0.86$ to $+0.98$
Fundus	24	0.0014	0.0005	$+0.76$ to $+0.95$

In each task, the variance of backbone mean alignments exceeded the mean within-backbone seed variance, and the backbone ordering was largely preserved across seeds (Table 8). Liver US had the lowest seed-swap backbone rank correlation of the four tasks.

Precision of a single checkpoint. The median width of patient- or case-level cluster-bootstrap intervals for single-checkpoint $F_{\text{anchor}|y}$ was 0.878 for Liver US, 0.184 for Breast US, 0.105 for Knee X-ray, and 0.103 for Fundus. The Liver US test split has 170 images but 17 independent patients, so single-checkpoint alignment estimates were less precise than in the other three tasks.

D.2 Post-Training Model-Characterization Analyses

This subsection presents the results reported in §6.2 in four parts: clinical-order retention (§D.2.1), scale dependence and corruption vulnerability (§D.2.2), the role-matched test for the fine-anchor task (§D.2.3), and exploratory stress-testing prioritization (§D.2.4). All model-level correlations follow the conventions of §B.4.

D.2.1 Clinical-Order Retention

Clinical-order retention is the mean change in full-image alignment over the three standard corruptions,

$$D = \frac{1}{3} \sum_{T \in \mathcal{T}} \left(F^{T}_{\text{full}|y} - F^{\text{clean}}_{\text{full}|y} \right). \tag{7}$$

Larger D means that clinical ordering is better preserved after corruption, and $D = 0$ means no average change in full-image alignment across the three corruptions. Table 9 reports task-wise and pooled values of this endpoint.

Table 9: **Clinical-order retention under the standard corruptions.** D is defined in Eq. (7). $F_{\text{anchor}|y}$, $F^{\text{clean}}_{\text{full}|y}$ and D are means over the 24 checkpoints of each task. The correlation adjusts for A_{anchor} and the clean full-image alignment, and intervals resample backbone cells. The pooled row covers the three coarse-anchor tasks.

Task	ckpt	$F_{\text{anchor}\|y}$	$F^{\text{clean}}_{\text{full}\|y}$	D	$\rho(F_{\text{anchor}\|y}, D \mid A_{\text{anchor}}, F^{\text{clean}})$
Liver US	24	$+0.246$	$+0.344$	-0.100	$+0.429$ $[-0.25, +0.76]$
Breast US	24	$+0.226$	$+0.413$	-0.110	$+0.715$ $[+0.17, +0.86]$
Knee X-ray	24	$+0.312$	$+0.463$	-0.154	$+0.458$ $[-0.11, +0.80]$
Fundus	24	$+0.246$	$+0.309$	-0.156	-0.250 $[-0.71, +0.50]$
Pooled, coarse-anchor tasks (N=72)					$+0.475$ $[+0.17, +0.71]$
patient/case-disjoint cross-fit, 50/50 splits positive					$+0.375$

When baseline $F_{\text{anchor}|y}$ and the retention score were computed on patient- or case-disjoint folds, all 50 fold-direction combinations showed a positive association, with a mean of $+0.375$.

D.2.2 Scale Dependence and Corruption Vulnerability

Cell-level quantities. Table 10 reports the backbone–task-level corruption vulnerability and R_{opp} used in §6.2. Pooled confidence intervals were computed by a hierarchical bootstrap that resamples patients or cases while keeping the task composition.

Table 10: **Corruption vulnerability and opposite-scale dependence, all 32 cells.** Vulnerability is the mean AUROC drop under blur, JPEG and downsampling. R_{opp} is the AUROC drop when the band *opposite* the prespecified clinical anchor is removed, so the removed band differs by task: fine for Liver US, Breast US and Knee X-ray, and coarse for Fundus, whose anchor is fine. Both quantities are positive when the perturbation hurts.

	Liver US		Breast US		Knee X-ray		Fundus	
Backbone	vuln.	R_{opp}	vuln.	R_{opp}	vuln.	R_{opp}	vuln.	R_{opp}
ConvNeXt-T	$+5.4$	-4.4	$+3.0$	$+1.4$	$+7.6$	$+3.9$	$+16.4$	$+9.4$
ConvNeXt-S	$+7.2$	$+4.8$	$+2.4$	$+1.4$	$+6.7$	$+2.8$	$+16.5$	$+10.9$
Swin-T	-2.0	-1.8	$+5.9$	$+2.8$	$+9.3$	$+3.8$	$+18.0$	$+12.2$
ViT-S	$+3.3$	$+2.6$	$+3.3$	$+1.9$	$+5.9$	$+2.3$	$+13.5$	$+19.0$
DeiT-S	$+6.1$	$+1.7$	$+4.6$	$+2.0$	$+10.2$	$+3.0$	$+16.7$	$+13.6$
DenseNet121	$+6.3$	-7.2	$+11.4$	$+6.2$	$+11.8$	$+3.6$	$+20.8$	$+16.3$
ResNet18	$+37.7$	$+41.2$	$+21.3$	$+21.3$	$+21.6$	$+10.0$	$+24.5$	$+17.0$
EfficientNet-B0	$+38.1$	$+35.2$	$+20.7$	$+18.2$	$+29.6$	$+19.9$	$+21.9$	$+19.6$

Scale dependence. Table 11 reports task-wise anchor- and opposite-scale dependence and their associations with $F_{\text{anchor}|y}$. R_b is computed without the clinical factor c. The Fundus boundary-case values reported in §6.2 follow the same cell-level conventions as the main analysis (three-seed-mean cells, mean loss under blur, JPEG, and downsampling). Table 12 compares, over the eight backbone–task cells of each task, how physical fine-scale dependence R_{fine} and role-based opposite-scale dependence R_{opp} each relate to $F_{\text{anchor}|y}$.

Table 11: **Anchor- and opposite-scale dependence by task.** Entries are panel means; pooled rows are rank correlations with $F_{\text{anchor}|y}$ across all 96 checkpoints, with intervals bootstrapped over backbone cells. Bold marks an interval excluding zero.

Task	Anchor	R_{anchor}	R_{opp}	$R_{\text{anchor}} - R_{\text{opp}}$
Liver US	coarse	28.45	9.01	$+19.44$
Breast US	coarse	19.33	6.93	$+12.40$
Knee X-ray	coarse	22.45	6.18	$+16.27$
Fundus	fine	8.48	14.75	-6.28
Pooled $\rho(F_{\text{anchor}\|y},\ R_{\text{opp}})$				$\mathbf{-0.607}$ $[-0.72, -0.33]$
$\rho(F_{\text{anchor}\|y},\ R_{\text{anchor}})$				$\mathbf{-0.347}$ $[-0.55, -0.02]$
$\rho(F_{\text{anchor}\|y},\ R_{\text{anchor}} - R_{\text{opp}})$, A_{full} controlled				$+0.249$ $[-0.06, +0.48]$

Table 12: **Physical fine-scale and role-based opposite-scale dependence.** Spearman correlations are computed across the eight backbone-level cells in each task. Bold marks an interval excluding zero.

Task	Anchor scale	$\rho(R_{\text{fine}}, F_{\text{anchor}\|y})$ *physical*	$\rho(R_{\text{opp}}, F_{\text{anchor}\|y})$ *by role*
Liver US	coarse	$\mathbf{-0.500}$ $[-0.79, -0.05]$	$\mathbf{-0.500}$ $[-0.79, -0.05]$
Breast US	coarse	$\mathbf{-0.881}$ $[-0.98, -0.57]$	$\mathbf{-0.881}$ $[-0.98, -0.57]$
Knee X-ray	coarse	$\mathbf{-0.643}$ $[-0.88, -0.41]$	$\mathbf{-0.643}$ $[-0.88, -0.41]$
Fundus	**fine**	$+0.048$ $[-0.52, +0.31]$	$\mathbf{-0.619}$ $[-0.81, -0.10]$

Corruption vulnerability. Table 13 reports the association between $F_{\text{anchor}|y}$ and corruption vulnerability before and after adjustment. Vulnerability is the mean diagnostic AUROC lost under blur, JPEG, and downsampling, analyzed over the 24 backbone–task cells of the three coarse-anchor tasks. Confidence intervals resample patient or case clusters while keeping the number of evaluation units in each task, with ranks taken within task before pooling.

Table 13: **Robustness relevance of clinical-signal alignment.** ρ between $F_{\text{anchor}|y}$ and the mean diagnostic AUROC lost under blur, JPEG and downsampling, over the 24 backbone–task cells of the three coarse-anchor tasks. Ranks are taken within task before pooling; each adjustment residualises both rank vectors on the stated quantity; intervals resample evaluation units.

Analysis	ρ [95% CI]
Unadjusted	-0.762 $[-0.88, -0.56]$
Adjusted for clean A_{full}	-0.693 $[-0.89, -0.51]$
Adjusted for R_{fine}	-0.450 $[-0.74, -0.13]$
Adjusted for both	-0.212 $[-0.71, -0.03]$

Individual-checkpoint sensitivity. Without three-seed averaging, we analyzed the 72 checkpoints of the three coarse-anchor tasks. The correlation between $F_{\text{anchor}|y}$ and corruption vulnerability was -0.683 $[-0.81, -0.40]$; after controlling for A_{full} and for R_{fine}, it was -0.665 $[-0.79, -0.42]$ and -0.344 $[-0.57, -0.04]$, respectively. These values are similar in direction and magnitude to the cell-level associations in the main text. The 72 checkpoints are, however, not independent experimental units: seeds are nested within the 24 backbone–task cells and share the same test cohort. Confidence intervals were computed by a cluster bootstrap that resamples backbone–task cells.

D.2.3 Fundus Role-Matched Analysis

The plan for the test that reverses the corruption role in the fine-anchor task was fixed before analysis (Table 14). Because the weights of the original Fundus campaign were not retained, we trained a replication panel with the same protocol (§C.2). The Fundus role-matched analysis and the RCV comparison use its 24 checkpoints, corresponding to the same 8 backbones as the controlled panel.

The replication panel closely reproduced the A_{full}, A_{anchor}, $F_{\text{anchor}|y}$, and $F_{\text{full}|y}$ of the original Fundus panel, and the backbone ordering matched for all four quantities (Spearman $+1.000$).

Table 14: **Fundus role-reversal analysis.** Operator and severity are held fixed; only the band each operator acts on changes. The analysis plan was fixed before any value below was computed.

Perturbation	Band it acts on	$\rho(F_{\text{anchor}\|y}, D \mid A_{\text{anchor}}, F^{\text{clean}})$
coarse-targeted	opposite to the anchor	$+0.068$ $[-0.56, +0.65]$
fine-targeted	the anchor itself	-0.183 $[-0.65, +0.49]$
blur / JPEG / downsampling	the anchor itself	-0.279 $[-0.72, +0.51]$
$\Delta\rho$ = coarse-targeted − fine-targeted, paired		$+0.250$ $[-0.66, +0.78]$
share of bootstrap draws above zero		69%

Per operator, coarse-targeted: atten -0.170, shuffle $+0.235$, bias $+0.067$; fine-targeted: atten -0.123, shuffle -0.077, bias -0.251. The coarse-targeted correlations differ in sign across operators.

Because standard blur, JPEG compression, and downsampling degrade the Fundus anchor band itself, the vulnerability comparison used for the coarse-anchor tasks involves a different scale role here. Under this condition, the alignment–vulnerability association was $\rho = -0.190$ [−0.90, +0.62] (8 cells). The paired role-matched contrast was $\Delta\rho = +0.250$ [−0.66, +0.78]. The range of $F_{\text{anchor}|y}$ in Fundus, $+0.187$ to $+0.304$, was the narrowest of the four tasks.

D.2.4 Exploratory Stress-Testing Prioritization

Evaluation setup and capture metric. This subsection presents the exploratory prioritization analysis of §6.2 over the 24 backbone–task cells of the three coarse-anchor tasks. The vulnerability of cell i is

$$V_i = \tfrac{1}{3} \sum_{T \in \mathcal{T}} \left(A^{\text{clean}}_{\text{full},\, i} - A^{T}_{\text{full},\, i} \right),$$

averaged over the three seeds, with negative losses kept as is rather than clipped. With $S_r(k)$ the k cells selected by ranking r and $S_V(k)$ the top k cells by actual vulnerability, we compute

$$\text{Captured}_r(k) = \frac{\sum_{i \in S_r(k)} V_i}{\sum_{i \in S_V(k)} V_i}$$

in each task and macro-average over the three tasks. The random reference consists of 20,000 stratified randomizations that select k cells at random in each task, and the table reports its 2.5–97.5th percentiles.

Table 15: **Task-stratified stress-testing prioritization.** Captured vulnerability is macro-averaged over the three coarse-anchor tasks. The random reference reports the mean and the 2.5–97.5th percentiles of the stratified randomization.

k/task	low $F_{\text{anchor}\|y}$	low A_{full}	low A_{anchor}	random [2.5, 97.5]%
1	0.997	0.649	0.997	0.400 [0.133, 0.787]
2	0.847	0.686	1.000	0.425 [0.195, 0.701]
3	0.984	0.624	0.996	0.527 [0.287, 0.780]
4	0.971	0.803	0.978	0.630 [0.399, 0.860]
5	0.970	0.817	0.977	0.719 [0.508, 0.919]
6	0.984	0.849	0.984	0.810 [0.627, 0.960]
7	0.994	0.864	0.997	0.897 [0.764, 0.985]

Budget-wise results. Table 15 reports the macro-averaged captured vulnerability at each per-task budget k. Budget k=8 selects all cells of each task, giving 1.000 for all three rankings and random, and is omitted. At every budget shown (k=1–7), the low-$F_{\text{anchor}|y}$ macro-average exceeded the 97.5th percentile of the random reference, whereas low-A_{full} did not. A_{anchor} ranking captured as much as or more than $F_{\text{anchor}|y}$.

Task-wise and pooled sensitivity. At the 25% budget (k=2, 6 cells in total), low-$F_{\text{anchor}|y}$ captured 5 of the 6 most vulnerable cells, low-A_{full} 4, and low-A_{anchor} 6; stratified random selection captured 1.49 on average. At this budget, per-task capture of low-$F_{\text{anchor}|y}$ versus low-A_{full} was 0.541 versus 0.057 in Liver US and 1.000 versus 1.000 in Breast US and Knee X-ray. The aggregate gain was thus largest in Liver US, and both rankings saturated in the other two tasks at this budget. Even in the macro-average excluding Liver US, low-$F_{\text{anchor}|y}$ was not lower than low-A_{full} at any evaluated budget.

In a pooled analysis that ranks all 24 cells without regard to task, the contrast at the 25% budget was larger, 0.754 versus 0.350. Because this does not control for differences in vulnerability level across tasks, we use the task-stratified analysis as the primary prioritization result.

The scope of this analysis is retrospective prioritization within the controlled panel.

D.3 Complementary Evaluation Comparisons

This subsection details the two comparisons of §6.3: spatial attribution (§D.3.1) and concept sensitivity (§D.3.2).

D.3.1 Attribution Measures and Target Sensitivity

This part reports the reference levels of the localization measures and the attribution-target sensitivity for the Breast US case study of §6.3. The scored images, Grad-CAM protocol, and the three measures of Eq. (5) are defined in §C.3.

Reference levels. The random-map references of mass and pointing are determined by the lesion area fraction. Over the 160 scored images, the lesion area fraction had mean 0.087, median 0.068, and 10th–90th percentiles 0.019–0.185. A random-map simulation returned mass 0.087, pointing 0.076, and IoU 0.057. Because the top-20% region has a fixed size, IoU reached only 0.390, the mean upper bound for this image set, even in a unit test with $H = M$.

Lesion masks mark lesion extent but do not include all evidence that may be clinically meaningful in ultrasound. Attribution to perilesional tissue and to posterior acoustic shadowing or enhancement can be clinically relevant (Moon et al., 2013) but can be penalized by mask-overlap measures. We therefore do not read a low localization score directly as an incorrect explanation.

Model-level contrasts. In three-seed means, ConvNeXt-S and ResNet18 had the same lesion attribution mass (0.208) but $F_{\text{anchor}|y}$ of $+0.348$ and $+0.017$, respectively, and DenseNet121, with the highest attribution mass (0.227), had $F_{\text{anchor}|y} = +0.169$. All models had attribution mass above the uniform-map baseline (0.087), so the ordering differences reported in §6.3 are not explained solely by an absence of lesion-focused attribution.

Attribution target class. The main text uses each image's ground-truth class as the Grad-CAM target. With the malignant class as a fixed target for all images, the maps for the 111 benign images would explain a scalar other than their ground-truth class. Switching from the fixed malignant target to the ground-truth-class target increased mean attribution mass from 0.141 to 0.176 and pointing from 0.256 to 0.360, with a Spearman correlation of $+0.71$ between the backbone orderings. The association between Grad-CAM localization and $F_{\text{anchor}|y}$ was also sensitive to the target convention: for attribution mass, it changed from $+0.571$ under the fixed target to $+0.024$ under the ground-truth-class target, and for pointing, its sign reversed from $+0.548$ to -0.095. Reflecting this sensitivity, the main-text comparison focuses on the model orderings of the three evaluations.

D.3.2 Concept-Sensitivity Comparison

Table 16 reports the results of the comparison described in §C.4.

Table 16: **Correspondence specificity under a matched permutation scheme.** All quantities are computed on the same x_{anchor} inputs and the same 24 Fundus replication checkpoints, and share the same within-y group-preserving clinical-factor permutation draws (B=2000); each is calibrated against its own reference distribution.

Quantity	Exceeded its own 95th percentile	Median empirical p
$F_{\text{anchor}\|y}$	$24/24$	< 0.001
Br, Eq. (6)	$8/24$	0.185
label-controlled variant	$7/24$	0.135
sign aggregation	$1/24$	0.335

The raw RCV relevance averaged Br $= +0.295$ (range -0.235 to $+0.764$), and 8 of the 24 checkpoints exceeded their own 95th percentile. Exceedances occurred across four backbones (ConvNeXt-S, ConvNeXt-T, DeiT-S, and ResNet18), with no backbone exceeding in all three seeds. Under TCAV-style aggregation of the sensitivity sign alone, the count was $1/24$. This comparison is limited to the Fundus replication panel and the layer rule above and does not assess the relative merit of the two methods.

D.4 Public-Checkpoint Audit Outputs

Figure 5 and Table 17 report the full audit outputs for the 11 public Breast US checkpoints analyzed in §6.4. Selection, preprocessing interface, permutation reference, eligibility, and provenance are described in §C.6.

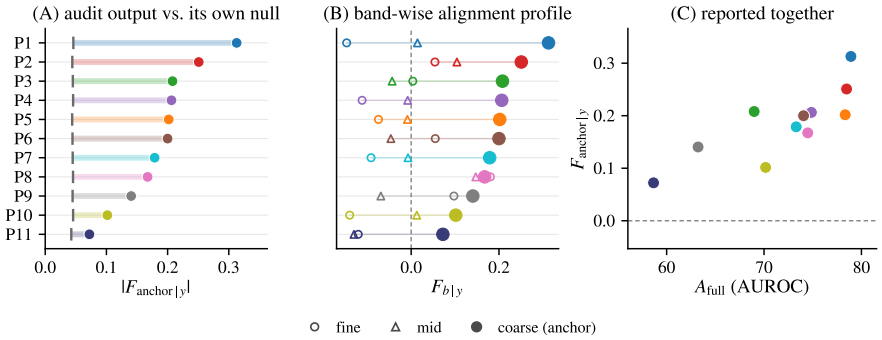

Figure 5: **Forward-only clinical-signal audit of public Breast US checkpoints. (A)** Observed $|F_{\text{anchor}|y}|$ and each model's within-y permutation 95th percentile; **(B)** band-wise alignment profiles; **(C)** a descriptive plot of $F_{\text{anchor}|y}$ against full-image AUROC. Colors identify the same checkpoint of Table 17 across panels.

The high-performance/low-alignment dissociation seen in the controlled panel did not appear in this set of public checkpoints.

Table 17: **Audit results for the** 11 **public Breast US checkpoints.** $F_{\text{anchor}|y}$ is read on the prespecified coarse anchor; *null* q_{95} is the 95th percentile of $|F_{\text{anchor}|y}|$ over permutations of c within each class of y; *top band* is the band with the highest $F_{b|y}$. **These are community uploads audited on data outside their intended use and must not be read as a quality verdict.**

ID	Backbone	A_{full}	$F_{\text{full}\|y}$	$F_{\text{anchor}\|y}$	null q_{95}	top band	Repository
P1	ViT-B	78.9	+0.27	+0.31	0.05	coarse	sergiopaniego (n.d.)
P2	ViT-B	78.5	+0.39	+0.25	0.04	coarse	sidushdid (n.d.a)
P3	Hiera	69.0	+0.08	+0.21	0.05	coarse	BTX24 (n.d.)
P4	ViT-B	74.9	+0.15	+0.21	0.05	coarse	sidushdid (n.d.b)
P5	EfficientNet	78.3	+0.29	+0.20	0.04	coarse	Ken2707 (n.d.)
P6	ViT-B	74.1	+0.27	+0.20	0.04	coarse	MarvelousXHJ (n.d.)
P7	ViT-B	73.3	+0.26	+0.18	0.04	coarse	sidushdid (n.d.c)
P8	DenseNet121	74.5	+0.36	+0.17	0.05	fine	Ken2707 (n.d.)
P9	ViT-L	63.2	+0.06	+0.14	0.04	coarse	ITSheep (n.d.)
P10	ResNet50	70.2	+0.17	+0.10	0.05	coarse	Ken2707 (n.d.)
P11	MobileNet	58.7	+0.02	+0.07	0.04	coarse	biradarakash0909 (n.d.)