CRS-Bench: A Reference-Relative Reliability Benchmark for Medical Image Encoders

Xingtao Lin[1] Hangqi Ren[1] Caiwan Sun[1] You Chen[1,2]

[1]Vanderbilt University [2]Vanderbilt University Medical Center (VUMC)

Abstract

*Pretrained image encoders are central to medical image classification, where expert annotation is costly and task-specific cohorts are often limited. As the model space expands from general-purpose to broad-medical and specialty-specific encoders, selecting the representation becomes a substantive modeling decision. Clean-test discrimination alone is insufficient for this purpose: encoders with similar AUROC can differ in calibration, label efficiency, and stability under acquisition perturbations or distribution shift. We introduce **CRS-Bench**, a controlled benchmark for multi-objective medical encoder selection. CRS-Bench evaluates 15 pretrained encoder families across dermatology, ophthalmology, and radiology using ISIC 2019, APTOS 2019, and CheXpert, with CheXpert→MIMIC-CXR as an observed institutional shift, yielding 17,575 controlled run records and 3,515 seed-aggregated metric rows. Each encoder is characterized along four operational reliability dimensions (discrimination, calibration, label efficiency, and robustness) and summarized by the **Clinical Reliability Score (CRS)**, a Pareto-aware, reference-relative score combining dominance, profile balance, and worst-axis performance. AUROC and CRS are positively associated but not decision-equivalent: **21 of 105 pairwise orderings reverse**, with a mean absolute rank displacement of 1.87. Paired-seed bootstrap analysis identifies PanDerm, MedSigLIP, and MedGemma as a stable leading reliability tier rather than a statistically resolved single leader. Reference-panel, scalarization, estimator, and cross-axis analyses further characterize criterion stability and information content. A matched 135-run adaptation arm keeps the leading three encoders intact ($\tau = 0.562$, Top-3 overlap 1.000) while reordering the middle of the ranking, so frozen comparisons bound rather than determine post-adaptation behavior. CRS-Bench provides a controlled framework for selecting medical image encoders from their **multi-axis reliability profiles**, rather than from clean AUROC alone.*

1 Introduction

Pretrained image encoders are a standard component of medical image classification, especially where expert annotation is expensive, task-specific cohorts are modest, and acquisition protocols vary across institutions. Large-scale pretraining transfers reusable visual structure into these low-data regimes and can reduce downstream supervision. The available model space has consequently become heterogeneous, spanning general-purpose encoders such as ResNet [8], ViT [5], CLIP [24], DINOv2 [21], and MAE [9]; broad-medical models such as BiomedCLIP [32], MedSigLIP, and MedGemma [25]; and specialty encoders for dermatology, ophthalmology, and radiology [2, 23, 26, 29, 33]. The resulting problem is how to select among substantially different encoders for a clinical task.

Current selection practice is dominated by clean-test discrimination, typically AUROC or AUPRC. Although necessary, discrimination is not sufficient to characterize an encoder for transfer. Representations with similar AUROC can differ materially in probability calibration, supervision demand, and stability under acquisition perturbations or distribution shift. An encoder may therefore rank highly on a clean split while remaining miscalibrated, supervision-intensive, or unstable under conditions that differ from the evaluation distribution [3, 7, 10, 13, 18, 20, 22]. Existing transfer and medical benchmarks measure subsets of these properties, but rarely all four under one matched clinical protocol, limiting guidance when selection targets a multidimensional reliability profile rather than a single predictive metric.

We introduce **CRS-Bench**, a controlled benchmark for multi-objective medical encoder selection. Fifteen pretrained encoder families are evaluated across dermatology, ophthalmology, and radiology using ISIC 2019, APTOS 2019, and CheXpert, with CheXpert→MIMIC-CXR as an observed institutional shift. Matched downstream capacity, data splits, label fractions, perturbations, random seeds, and tuning budgets yield 17,575 run records and 3,515 seed-aggregated metric rows. Each encoder is characterized by four operational dimensions: **discrimination, calibration, label**

efficiency, and robustness.

To summarize these trade-offs, we propose the **Clinical Reliability Score (CRS)**. CRS combines Pareto dominance, proximity to a balanced high-performing profile, and worst-axis performance. Because the score is comparative, its reference panel and normalization anchors are explicit and fixed, allowing later encoders to be evaluated without changing previously reported values. We then examine whether this criterion remains stable under reference-panel and scalarization perturbations, whether its constituent dimensions contribute information beyond AUROC, how sampling uncertainty affects the leading set, and how far the ordering persists under downstream adaptation and excluded data.

The results establish that clean discrimination and multi-axis reliability are related but not decision-equivalent: **21 of 105 pairwise encoder orderings reverse** between AUROC and CRS, with mean absolute displacement 1.87. Bootstrap analysis supports a stable leading tier of PanDerm, MedSigLIP, and MedGemma rather than a statistically resolved single leader. The benchmark further reveals systematic differences in encoder behavior: specialty pretraining can yield large matched-domain gains with heterogeneous transfer; calibration is often substantially improved post hoc without changing AUROC; intermediate representations can outperform final-layer features; corruption can preserve discrimination while degrading confidence; and medical pretraining provides its largest relative advantage when labels are scarce.

Contributions. **(i) CRS-Bench.** A controlled multi-domain evaluation of 15 pretrained encoders across four complementary reliability dimensions. **(ii) CRS.** A Pareto-aware, reference-relative summary whose published values remain comparable as new encoders are evaluated. **(iii) Criterion validation.** Systematic analysis of reference stability, construct informativeness, scalarization sensitivity, sampling uncertainty, estimator dependence, and decision consequences beyond AUROC. **(iv) Encoder analysis.** Empirical characterization of specialization, calibration, feature depth, label scarcity, perturbation robustness, prompting, and downstream adaptation.

2 Related Work

Pretrained encoders for medical imaging. Generalist models use supervised natural-image pretraining (ResNet [8], ViT [5]), image–text contrastive learning (CLIP [24], SigLIP [31]), or self-supervision (DINOv2 [21], MAE [9]). Broad-medical encoders adapt these recipes to biomedical data (BiomedCLIP [32], LLaVA-Med [17], MedSigLIP, MedGemma [25]), while RETFound [33], PanDerm [29], RAD-DINO [23], BioViL [2], and CheXzero [26] target particular clinical domains. Prior comparisons generally cover narrower model sets and emphasize clean discrimination or task-specific adaptation.

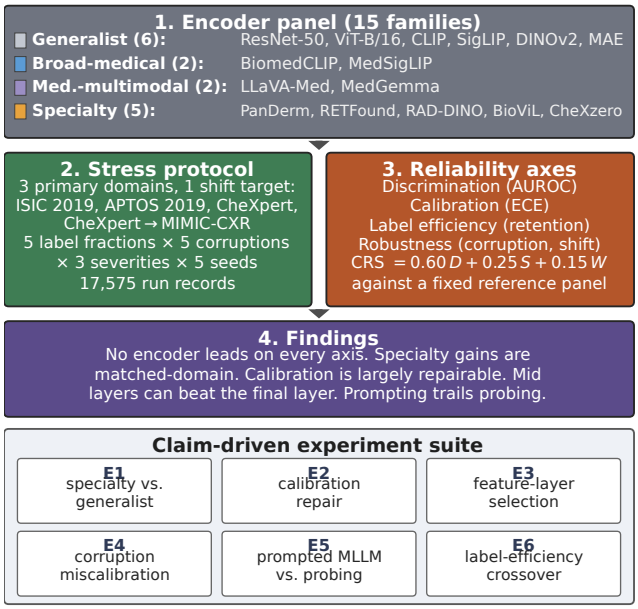


Figure 1. **CRS-Bench evaluation pipeline.** Fifteen encoder families are evaluated under one controlled protocol, and four reliability dimensions define the profile summarized by CRS.

Benchmarking reliability. VTAB [30] and ELEVATER [16] measure representation transfer, ImageNet-C standardizes corruption robustness [10], and medical benchmarks emphasize few-shot adaptation, modality breadth, radiology, or fairness [14, 19, 27, 28]. Calibration and uncertainty studies further show that high discrimination does not guarantee reliable probabilities and that confidence degrades under shift [3, 7, 13, 18, 22]. CRS-Bench instead measures four reliability dimensions under one controlled clinical protocol and defines its aggregate against an explicit reference panel.

3 The CRS-Bench Benchmark

CRS-Bench separates three methodological tasks: *measurement*, *score validation*, and *behavioral interpretation* (Fig. 1). Secs. 3.1–3.4 measure four reliability dimensions under matched frozen probing, label scarcity, acquisition perturbations, and institutional shift. Secs. 3.5–3.6 define the reference-relative CRS and test reference stability (panel perturbation and insertion), construct informativeness/scalarization (correlations and weight perturbations), sampling uncertainty (paired-seed bootstrap), estimator dependence, and ranking transfer (LoRA and held-out data). Sec. 3.7 defines E1–E6 for specialization, calibration, feature depth, corruption, prompting, and label scarcity. This structure separates what is measured, how the aggregate is validated, and what the benchmark reveals about encoders.

3.1 Selection Problem and Encoder Panel

CRS-Bench studies *controlled encoder selection*: given a declared clinical evaluation suite, how do pretrained repre-

Table 1. The 15 encoder families by pretraining scope. "Input" is the released input size used by each encoder (Sec. 3.2). Domain marks the matched specialist domain: Derm. = ISIC, Eye = APTOS, Rad. = CheXpert/MIMIC-CXR. † indicates no APTOS frozen cell.

Model	Pretraining data	Paradigm	Input	Domain
Generalist natural-image encoders				
ResNet-50 [8]	ImageNet-1K	Supervised	224	n/a
ViT-B/16 [5]	ImageNet-21K	Supervised	224	n/a
CLIP ViT-B [24]	WIT-400M	Contrastive	224	n/a
SigLIP ViT-B [31]	WebLI	Contrastive	224	n/a
DINOv2 ViT-B [21]	LVD-142M	Self-sup.	224	n/a
MAE ViT-B [9]	ImageNet-1K	Self-sup.	224	n/a
Broad-medical and medical-multimodal encoders				
BiomedCLIP [32]	PMC-15M	Contrastive	224	Med.
LLaVA-Med [17]	PMC, instruct.	MLLM	336	Med.
MedSigLIP [25]	Med. image/text	Contrastive	448	Med.
MedGemma [25]	Med. image/text, IT	MLLM	own	Med.
Specialty-specific encoders				
PanDerm [29]	Skin images	Self-sup.	224	Derm.
RETFound [33]	Fundus / OCT	MAE-SSL	224	Eye
RAD-DINO [23]	Chest X-rays	DINO-SSL	224	Rad.
BioViL† [2]	MIMIC-CXR	Contrastive	480	Rad.
CheXzero† [26]	MIMIC-CXR	Contrastive	320	Rad.

sentations compare when selection is based jointly on several observable reliability properties under matched downstream conditions? For encoder panel $\mathcal{M} = \{m_1, \ldots, m_K\}$, representation $z = m_k(x) \in \mathbb{R}^{d_k}$, and lightweight probe $\hat{y} = h_\theta(z)$, we measure

$$\mathbf{s}^{(k)} = (s^{(k)}_{\text{disc}}, s^{(k)}_{\text{cal}}, s^{(k)}_{\text{LE}}, s^{(k)}_{\text{rob}}), \tag{1}$$

for discrimination, calibration, label efficiency, and robustness. CRS (Sec. 3.5) is a reference-relative summary of this measured profile rather than an estimator of universal clinical utility. Candidate models may evolve while the comparison reference remains explicit, so later evaluations do not implicitly redefine earlier scores.

The 15-family panel spans general-purpose, broad-medical or medical-multimodal, and specialty-specific encoders (Tab. 1). MedSigLIP and MedGemma are kept distinct because we probe different released visual representations: the standalone MedSigLIP tower at native 448 resolution and the MedGemma-4B vision stack after projection into its generative interface. The frozen panel contains 43/45 encoder–dataset cells; BioViL and CheXzero have no APTOS cell because their released interfaces are chest-radiograph specific. Their aggregate values use the two observed domains and are never imputed.

3.2 Controlled Representation Evaluation

Frozen probing defines the primary estimand: representation quality under a common downstream decision rule. Fixing probe family and optimization budget reduces confounding from task-specific optimization and isolates differences in the pretrained feature space. LoRA is evaluated as a separate regime (Sec. 5.5) that measures how much of the frozen ordering persists after task-specific adaptation.

Each encoder uses the input transform required by its released interface (Tab. 1). After feature extraction, data splits, label subsamples, perturbation seeds, probe capacity, hyperparameter budget, wall-clock cap, and five seeds $s \in \{0, \ldots, 4\}$ are matched. Features are standardized with training-split statistics. The primary probe is ℓ_2-regularized logistic regression ($\lambda = 1/N_{\text{train}}$, L-BFGS), one-vs-rest for CheXpert; a two-layer MLP and k=5 nearest neighbor provide probe-dependence checks. For specialists that accept the common 224-pixel ImageNet transform, a standardized-input sensitivity changes AUROC by < 0.006 (supplement).

Two-tier evaluation. A factorial core varies five label fractions, five perturbation types, three severity levels, and five shared seeds on representative encoders for ISIC and APTOS. A broad sweep evaluates all 15 families on ISIC, APTOS, and CheXpert, with CheXpert probes additionally tested on MIMIC-CXR without retraining. The benchmark contains 17,575 runs and 3,515 seed-aggregated metric rows.

Adaptation regime. LoRA uses rank 8, $\alpha = 16$, dropout 0.05, learning rate 2×10^{-4}, eight epochs, attention/MLP projection adapters, and three seeds. Each encoder–dataset–seed triple is evaluated over the same label-fraction and perturbation grid (135 encoder–dataset–seed runs), with all four dimensions recomputed after adaptation. The analysis therefore tests transfer of the frozen selection ordering rather than redefining CRS around fine-tuning.

3.3 Reliability Dimensions

Each dimension is computed per dataset before aggregation in Sec. 3.5. They are operationally distinct but not assumed statistically independent; empirical dependence is tested in Sec. 5.3.

Discrimination. We use macro-AUROC,

$$\text{AUROC}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} \text{AUROC}_c, \tag{2}$$

computed one-vs-rest for multiclass tasks and per-label then macro-averaged for CheXpert. Macro AUPRC and sensitivity at fixed specificity are secondary metrics. Because APTOS grades are ordinal, QWK and ordinal MAE are additionally substituted for the discrimination axis in Sec. 5.7.

Calibration. ECE with B=15 fixed-width bins is

$$\text{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{N} \big|\text{acc}(S_b) - \text{conf}(S_b)\big|. \tag{3}$$

Confidence is maximum class probability for multiclass tasks; CheXpert ECE is macro-averaged across labels. Fixed-width ECE provides a deterministic common estimator across the model panel. Adaptive ECE tests sensitivity to binning, while Brier score evaluates whether the conclusions persist under a proper scoring rule (Sec. 5.7).

Label efficiency. Absolute low-label AUROC mixes sample efficiency with attainable discrimination. We instead measure retention of each encoder's own full-label performance.

For $\mathcal{P} = \{0.01, 0.05, 0.10, 0.25\}$,

$$\mathrm{LE} = \frac{1}{|\mathcal{P}|} \sum_{\rho \in \mathcal{P}} \frac{A(\rho)}{A(1)}, \tag{4}$$

where $A(\rho)$ and $A(1)$ are macro-AUROC at fraction ρ and full labels for the same encoder–dataset pair. The full-label point appears only in the denominator, so LE measures preservation under label scarcity rather than absolute attainment. This distinction matters because conventional area under the label-fraction curve (AULC) rewards a high full-label endpoint even when a model loses a similar fraction of performance as labels are removed. We compare AULC and relative retention directly in Sec. 5.3 rather than assuming the two encode distinct information. Ratios slightly above one are kept (maximum 1.003) and bounded only by the fixed CRS normalization; a log-spaced trapezoidal variant is reported in the supplement.

Robustness. With clean AUROC A_0 and perturbed AUROC A_j, let $\delta_j = \max(0, (A_0 - A_j)/A_0)$ and let $\bar{\delta}, \delta_{\max}$ be the mean and worst perturbation drops. CheXpert additionally uses δ_{shift} from MIMIC-CXR. We define

$$R = \max\left(0, 1 - \frac{\bar{\delta} + \delta_{\max} + \mathbf{1}_{\text{shift}}\delta_{\text{shift}}}{2 + \mathbf{1}_{\text{shift}}}\right), \tag{5}$$

with $\mathbf{1}_{\text{shift}} = 1$ for CheXpert and 0 otherwise. The denominator counts the perturbation statistics that were actually observed, so a cell without a natural shift is scored as the mean of its two corruption terms rather than being charged a third term that was never measured. Robustness is consequently comparable across cells in construction while remaining heterogeneous in evidence: only CheXpert contributes an observed institutional shift. We therefore report a matched corruption-only variant, in which every cell is scored from corruption alone, as the direct test of whether that heterogeneity carries the ranking (Sec. 5.7); E4 separately quantifies calibration degradation under corruption.

3.4 Stress Conditions

Acquisition-quality perturbations. Five perturbations are applied at three severities: Gaussian blur, JPEG compression, contrast reduction, Rician noise [6], and field-of-view crop. Rician noise uses $\tilde{x} = \sqrt{(x + n_r)^2 + n_i^2}$, $n_r, n_i \sim \mathcal{N}(0, \sigma^2)$. Perturbation seeds are derived from a SHA-256 hash of the image identifier, ensuring that the same transformed image reaches every encoder. These stresses model acquisition and image-quality degradation, not changes in anatomy, pathology, or population.

Observed institutional shift. CheXpert-trained probes are applied to MIMIC-CXR without retraining. The datasets share 14 labels but differ in institution, population, protocol, and labeling, testing preservation of discriminative structure across institutions.

3.5 Reference-Relative Clinical Reliability Score

After orienting all four dimensions so larger is better, CRS combines Pareto dominance over a declared reference panel, proximity to an ideal profile, and worst-axis performance,

$$D^{(k)} = \frac{1}{n_k} \sum_{j \in \mathcal{R} \setminus \{k\}} \mathbf{1}[\tilde{\mathbf{s}}^{(k)} \succ \tilde{\mathbf{s}}^{(j)}], \tag{6}$$

$$S^{(k)} = 1 - \sum_a w_a (1 - \tilde{s}_a^{(k)})^2, \quad W^{(k)} = \min_a \tilde{s}_a^{(k)}, \tag{7}$$

$$\mathrm{CRS}^{(k)}_{\mathcal{R},\mathcal{A}} = \alpha_D D^{(k)} + \alpha_S S^{(k)} + \alpha_W W^{(k)}. \tag{8}$$

Here D measures Pareto-consistent improvement over the reference panel, S measures proximity to the ideal profile, and W retains sensitivity to the weakest axis. The three terms provide complementary summaries without treating the four raw metrics as directly commensurate. $n_k = |\mathcal{R}| - \mathbf{1}[k \in \mathcal{R}]$.

Reference standard. Candidate-dependent min–max scaling and dominance make an unchanged encoder's score move when new candidates are inserted. CRS therefore fixes a reference $(\mathcal{R}, \mathcal{A})$, where $\mathcal{R}$ supplies dominance peers and $\mathcal{A} = \{(l_a, u_a)\}_a$ the normalization anchors,

$$\tilde{s}_a = \mathrm{clip}\left(\frac{s_a - l_a}{u_a - l_a}, 0, 1\right). \tag{9}$$

The 15-family panel under Sec. 3.2 defines the aggregate reference. A future encoder is normalized by the same anchors and compared with the same peers, so previously reported scores do not change. Per-domain CRS uses an analogous domain-specific reference and is interpreted only within that domain. Repeated anchor saturation signals when a successor reference is needed.

Scalarization preferences. Axis weights w_a affect only the ideal-point term and are uniform $(1/4)$ by default. Component weights $(\alpha_D, \alpha_S, \alpha_W) = (0.60, 0.25, 0.15)$ prioritize Pareto-consistent comparative evidence while retaining ideal-profile and worst-axis penalties. Both are explicit design parameters rather than learned clinical utilities. The two weight families are perturbed separately in Sec. 5.3; our analysis concerns the stability of the induced ordering, not optimality of a particular coefficient vector.

3.6 Validation Design

We evaluate five properties of the selection criterion. **Reference stability/extensibility** uses 15 leave-one-encoder-out panels, 200 random subsets at sizes 12 and 10, and 30 random 12+3 insertion trials comparing cohort recomputation with the fixed reference (Sec. 5.2). **Construct informativeness** uses Spearman correlations at encoder and encoder–dataset levels and directly contrasts relative-retention LE with AULC; **weight sensitivity** separately perturbs 500

axis-weight and 500 component-weight draws (Sec. 5.3). **Decision consequence** compares AUROC and CRS using Kendall τ, Spearman ρ, pairwise reversals, rank displacement, and Top-k overlap (Sec. 5.4).

Sampling uncertainty uses 10,000 paired-seed bootstrap replicates over the five shared seeds and 43 observed cells, recomputing anchors and CRS within each replicate (Sec. 5.1). Estimator substitutions use adaptive ECE/Brier, APTOS QWK/ordinal MAE, and matched corruption-only robustness (Sec. 5.7). LoRA and leave-one-dataset-out/MIMIC analyses test ranking transfer under adaptation and excluded data without redefining future-task prediction as the CRS objective (Secs. 5.5 and 5.6).

3.7 Encoder-Level Analysis Suite

Six analyses explain the profiles themselves. **E1** separates specialist matched-domain gain from transfer using $\Delta_{\text{spec}} = \text{AUROC}_{\text{in}} - \text{AUROC}_{\text{out}}$ and $\Delta_{\text{ood}} = \text{CRS}_{\text{in}} - \text{CRS}_{\text{out}}$. **E2** fits temperature, vector, isotonic, and Dirichlet calibration on a held-out calibration split and recomputes ECE/Brier while checking unchanged AUROC. **E3** probes transformer features at 25%, 50%, 75%, and 100% depth. **E4** pairs AUROC drop with ECE, confidence shift, and overconfidence-under-shift. **E5** compares frozen probing with simple, multiple-choice, chain-of-thought, refined, and few-shot prompts for LLaVA-Med/MedGemma, reporting parse coverage. **E6** reports AUROC at 1%/10% labels and the first label fraction matching ViT-B/16. Together, these analyses characterize the empirical mechanisms underlying the observed profiles rather than treating scalar rank as the sole endpoint.

4 Experimental Setup

Datasets. Three primary datasets and one external shift target: CheXpert [12] (191,229 chest radiographs, 14 labels, multilabel); ISIC 2019 [4, 11] (25,331 dermoscopy images, 8 classes); APTOS 2019 [1] (3,662 fundus photographs, 5 ordinal grades); and MIMIC-CXR [15] as the institutional-shift target for CheXpert-trained probes. Together they span distinct modalities, anatomical scales, and acquisition pipelines.

Implementation. Following Sec. 3.2, all encoders are frozen and evaluated in a single feature pass through their required input transform, with shared splits, seeds, probe capacity, and tuning budget. For generative VLMs in E5 we use simple, multiple-choice, chain-of-thought, refined, and few-shot prompts, parsed deterministically with coverage recorded per condition. Frozen experiments run on a single GB10 DGX Spark; the 135 adaptation runs are distributed over a compute cluster.

5 Results

Our evaluation addresses two goals. For criterion validity, Sec. 5.1 combines four-axis profiles with 10,000 bootstrap replicates; Sec. 5.2 tests reference stability using 415 panel perturbations and 30 insertion trials; Sec. 5.3 tests informativeness and scalarization using cross-axis correlations and 1,000 weight perturbations; Sec. 5.4 quantifies AUROC–CRS decision differences; and Secs. 5.5–5.7 test LoRA transfer, held-out/MIMIC-CXR transfer, and estimator dependence. For encoder behavior, Secs. 5.8–5.11 execute E1 plus institutional shift, E2/E4, E3/E6, and E5 to study specialization/transfer, calibration/corruption, feature depth/label scarcity, and prompting/probing.

5.1 Multi-Axis Profiles and Sampling Uncertainty

Clean AUROC and CRS yield related but distinct orderings (Tab. 2 and Fig. 2). MedGemma/MedSigLIP define the AUROC frontier (0.909/0.908), whereas PanDerm and MedSigLIP are separated by 0.002 in CRS (0.703/0.701). Figure 2 explains the reordering: PanDerm pairs competitive discrimination with the strongest robustness; MedSigLIP/MedGemma concentrate more advantage in discrimination and low-label retention; BiomedCLIP is comparatively strong in calibration; and CheXzero combines lower discrimination with favorable calibration and robustness. Per-domain leaders also change, with PanDerm leading on ISIC CRS, MedGemma on APTOS, and MedSigLIP on CheXpert, so the aggregate ranking summarizes distinct reliability profiles rather than a rescaled accuracy statistic.

Across 10,000 paired-seed bootstrap replicates, PanDerm/MedSigLIP attain rank 1 with probability 0.664/0.336 and their paired CRS difference has 95% interval $[-0.097, 0.017]$. PanDerm, MedSigLIP, and MedGemma each have $P(\text{top-3}) = 1.000$, whereas every other encoder has probability zero. The bootstrap therefore supports a three-model leading tier rather than a statistically resolved single leader. Because anchors are recomputed within each replicate, the reported intervals describe the sampling distribution of the full scoring procedure rather than intervals centered on the fixed-anchor point estimate.

5.2 Reference Stability and Incremental Extensibility

Panel perturbation tests dependence on current peers; incremental insertion tests whether future candidates move

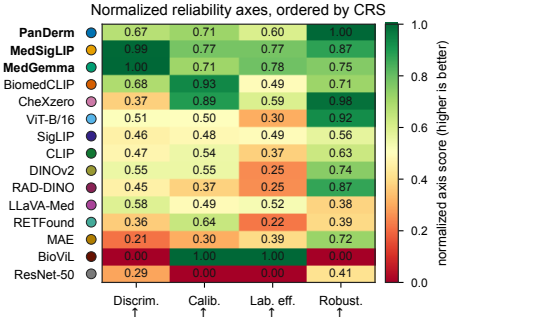


Figure 2. **Normalized reliability profiles.** Encoders are ordered by CRS; higher is better on every axis. Similar discrimination can coexist with different calibration, label retention, and robustness.

Table 2. Encoder reliability by dataset. CRS uses domain-specific references per dataset and the global reference in aggregate. BioViL/CheXzero lack APTOS. CI is the paired-seed sampling interval; P_1 is rank-1 probability.

		ISIC		APTOS		CheXpert		Aggregate			
Model	Grp	AUROC	CRS	AUROC	CRS	AUROC	CRS	AUROC	CRS	CI	P_1
PanDerm	S	0.957	**0.583**	0.922	0.685	0.733	0.346	0.871	**0.703**	[0.600,0.720]	**0.66**
MedSigLIP	BM	0.957	0.485	0.951	0.755	**0.816**	**0.524**	0.908	0.701	[0.682,0.703]	0.34
MedGemma	MM	**0.960**	0.333	**0.958**	**0.763**	0.808	0.359	**0.909**	0.645	[0.598,0.647]	0.00
BiomedCLIP	BM	0.899	0.449	0.908	0.454	0.812	0.246	0.873	0.424	[0.406,0.522]	0.00
CheXzero	S	0.886	0.253	n/a	n/a	0.788	0.418	0.837	0.398	[0.391,0.403]	0.00
ViT-B/16	G	0.926	0.323	0.906	0.297	0.727	0.332	0.853	0.320	[0.228,0.443]	0.00
SigLIP	G	0.935	0.409	0.920	0.220	0.686	0.142	0.847	0.298	[0.278,0.305]	0.00
CLIP	G	0.927	0.396	0.898	0.303	0.720	0.154	0.848	0.284	[0.248,0.352]	0.00
DINOv2	G	0.935	0.480	0.912	0.363	0.725	0.202	0.857	0.266	[0.168,0.352]	0.00
RAD-DINO	S	0.891	0.165	0.846	0.138	0.803	0.244	0.847	0.251	[0.142,0.331]	0.00
LLaVA-Med	MM	0.938	0.364	0.924	0.250	0.723	0.078	0.861	0.241	[0.231,0.247]	0.00
RETFound	S	0.900	0.257	0.907	0.098	0.702	0.186	0.836	0.187	[0.131,0.225]	0.00
MAE	G	0.896	0.279	0.875	0.137	0.687	0.145	0.819	0.184	[0.172,0.194]	0.00
BioViL	S	0.809	0.125	n/a	n/a	0.782	0.323	0.795	0.125	[0.125,0.125]	0.00
ResNet-50	G	0.881	0.144	0.880	0.160	0.724	0.097	0.828	0.072	[0.071,0.085]	0.00

published scores. Across 415 perturbations, Kendall τ is 0.975/0.926/0.891 for leave-one-out, size-12, and size-10 panels, with Top-3 overlap 1.000/0.985/0.958. Agreement weakens gradually as the panel is thinned, but the leading tier remains stable. In 30 random 12+3 trials (90 insertions), cohort renormalization moves existing scores by mean/median/maximum 0.053/0.042/0.167, whereas the fixed-reference rule yields exactly zero displacement. The zero displacement under fixed-reference insertion is a definitional invariance rather than an empirical coincidence: a future encoder requires only its own evaluation and cannot alter previously published scores.

5.3 Construct Informativeness and Scalarization Stability

Axis informativeness. AULC label efficiency is nearly a re-expression of full-label AUROC ($\rho = 0.986$ across encoders; 0.977 across 43 cells). Relative retention (Eq. (4)) reduces these to 0.339 and -0.342, consistent with the intended interpretation of LE as relative performance retention under reduced supervision. The remaining dimensions are dependent but non-interchangeable: the largest $|\rho|$ is 0.625/0.460 at encoder/cell level, while discrimination versus oriented calibration changes from 0.764 on APTOS to -0.279 on CheXpert and -0.029 on ISIC.

Scalarization stability. Across 500 axis-weight and 500 component-weight draws, mean Kendall agreement is 0.986/0.985 (minimum 0.943/0.924) and Top-3 is unchanged in all 1,000 draws; alternative axis weights $(0.30, 0.30, 0.15, 0.25)$ give $\tau = 1.000$. These perturbations support stability of the leading set, but do not imply uniqueness or optimality of the default coefficient vector.

5.4 Decision Consequences beyond AUROC

Because discrimination is one CRS dimension, positive association is expected; the question is whether the remaining axes change selections. Across 15 encoders, $\tau = 0.600$, $\rho = 0.811$, 21/105 pairwise choices reverse, mean absolute displacement is 1.87 ranks, and Top-3/Top-5 overlap is 0.667/0.800 (Fig. 3). CheXzero moves 11→5, LLaVA-Med 5→11, PanDerm 4→1, and DINOv2 6→9. CRS is therefore correlated with, but not decision-equivalent to, clean discrimination; the disagreement is distributed across the ordering rather than being confined to a single change at the top.

5.5 Transfer of the Ranking under Adaptation

Frozen probing fixes the downstream decision rule and therefore isolates the pretrained representation; LoRA changes the representation itself. Agreement between the two regimes consequently tests persistence of the frozen ordering after task-specific optimization rather than serving as another estimate of the same quantity. The matched adaptation arm recomputes discrimination, calibration, label efficiency, robustness, and CRS on the same evaluation grid. Adaptation raises clean discrimination in every domain, by $+0.026/+0.030/+0.010$ mean macro-AUROC on ISIC/APTOS/CheXpert, and preserves the discrimination ordering unevenly: Kendall $\tau = 0.829/0.667/0.448$ with Top-3 overlap 0.667 in each domain, and the matched-domain leader changes in all three (MedGemma to MedSigLIP on ISIC and APTOS, MedSigLIP to RAD-DINO on CheXpert). The four-axis comparison is more demanding: against the frozen CRS the adapted ranking gives $\tau = 0.562$ and $\rho = 0.707$ with a mean absolute rank shift of 2.0. PanDerm is retained at rank 1 and the leading three are unchanged (Top-3 overlap 1.000), whereas the middle reorganizes, with BiomedCLIP falling from 4 to 13 and LLaVA-Med rising from 11 to 4. Frozen rankings therefore characterize the pretrained representation and bound, rather than determine, the ordering after parameter-efficient adaptation.

5.6 External Validity of a Reference-Relative Score

Leave-one-dataset-out CRS correlates with the held-out ranking at $\rho = 0.550/0.709/0.271$ for ISIC/APTOS/CheXpert, with Top-3 overlap 1.000/0.333/0.667 and the held-out winner recovered in no fold. Removing MIMIC-CXR from CRS and predicting MIMIC AUROC gives $\rho = 0.529$, versus 0.864 for CheXpert AUROC. The latter is more directly aligned because the target is discrimination on a closely related radiographic dataset. These tests therefore bound CRS as a

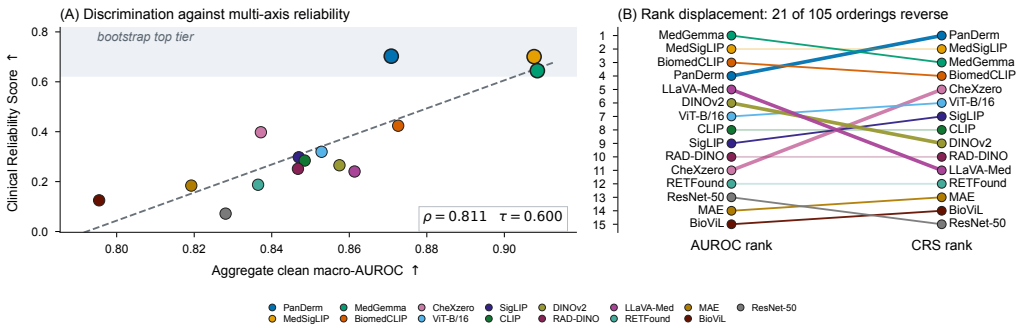


Figure 3. **AUROC versus multi-axis reliability.** (A) Aggregate AUROC versus CRS ($\rho = 0.811$, $\tau = 0.600$). (B) Rank displacement: 21/105 pairwise orderings reverse (mean 1.87). The complete key identifies every model; panel (b) also labels each model directly.

multi-objective selection summary under a declared suite rather than a universal predictor of unseen-task AUROC.

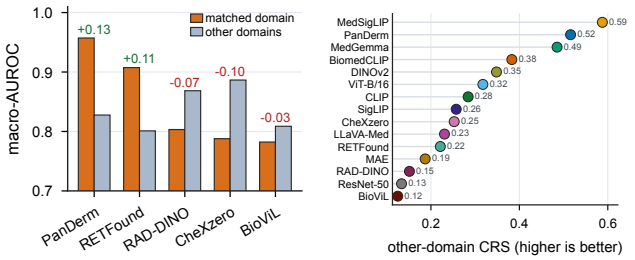
5.7 Sensitivity to Metric Representation

Adaptive ECE leaves the ranking unchanged ($\tau = 1.000$) and Brier gives $\tau = 0.981$, preserving the leading tier; the primary conclusion is therefore not determined by the fixed-width binning scheme. On APTOS, QWK/ordinal-MAE give $\tau = 0.923/0.949$ with Top-3 preserved, indicating that the leading set is not an artifact of representing the ordinal task with macro-AUROC. Matched corruption-only robustness gives $\tau = 0.962$ with Top-3/Top-5 preserved, although absolute CRS moves more (mean 0.031; maximum 0.171). Thus the two robustness variants are different measurements that support the same leading set.

5.8 Domain Specialization and Cross-Domain Transfer

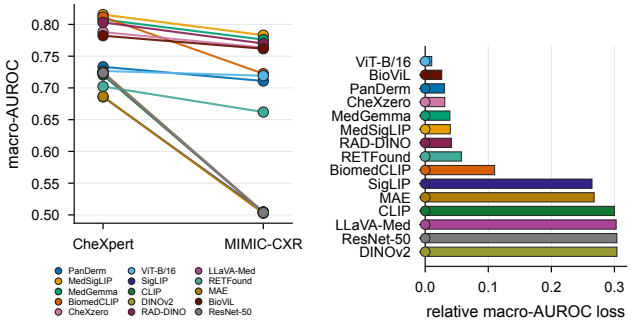
Specialty pretraining yields substantial but heterogeneous matched-domain gains (Fig. 4). PanDerm is the clearest case: ISIC AUROC is 0.957 versus 0.828 out of domain (+0.130), with only a 0.068 out-of-domain CRS penalty. RETFound gains ophthalmic discrimination without a commensurate balanced-reliability advantage. For radiology specialists, raw in-minus-out AUROC is less comparable because CheXpert is multilabel; the institutional-shift analysis provides a more directly interpretable transfer setting. Within this panel, broad-medical encoders are more consistent across domains, while PanDerm shows that specialization need not imply narrow transfer.

Under CheXpert→MIMIC-CXR (Fig. 5), MedSigLIP/MedGemma show small relative AUROC gaps (0.040/0.039), and medically pretrained encoders retain 0.717 macro-AUROC versus 0.540 for generalists (Welch $p = 0.0029$). Preservation is not uniform within the medical group: LLaVA-Med falls to 0.504, so provenance alone does not secure it. The site shift therefore exposes preservation differences that clean source-domain AUROC does not reveal.



(a) Matched vs. other-domain AUROC. (b) Other-domain CRS by encoder.

Figure 4. **Specialization versus transfer (E1).** Matched-domain discrimination and cross-domain reliability are complementary. Panel (b) ranks and labels every encoder; values are printed beside the corresponding model markers.



(a) Source vs. shifted AUROC. (b) Relative AUROC loss.

Figure 5. **Institutional shift.** CheXpert→MIMIC-CXR preservation. Model colours in (a) match the complete labels in (b).

5.9 Calibration Repair and Corruption-Induced Miscalibration

Across six primary tests, temperature, vector, isotonic, or Dirichlet scaling reduces ECE by 54–77% (mean 0.067, 95% CI $[0.024, 0.110]$) while preserving AUROC (mean ΔAUROC approximately $+0.001$). DINOv2/APTOS improves 0.150→0.034 and MedSigLIP/APTOS 0.121→0.028 (Fig. 6). Because calibration changes the probability mapping, it can repair confidence but cannot recover lost discrimination or transfer.

Corruption exposes a complementary failure mode: class ranking can remain useful while probability estimates degrade. APTOS LLaVA-Med shows mean ECE increase

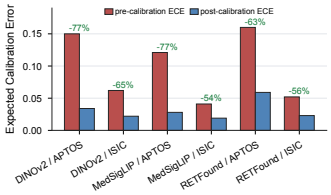
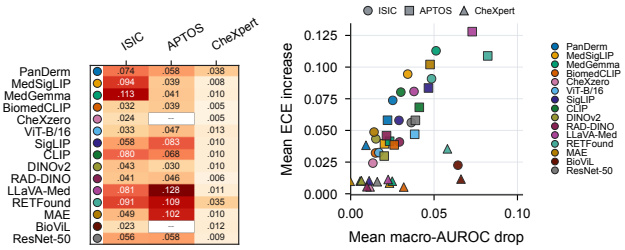


Figure 6. **Post-hoc calibration (E2).** ECE falls by 54–77% at preserved AUROC.



(a) Mean ECE increase. (b) AUROC loss vs. ECE increase.

Figure 7. **Corruption-induced miscalibration (E4).** The losses are moderately associated ($r = 0.63$). Colour = encoder, with the complete model key at right; shape = dataset.

0.128 (worst 0.418); on ISIC, MedGemma/MedSigLIP remain strong discriminators while ECE rises 0.113/0.094, whereas CheXpert is comparatively stable throughout (maximum 0.038). Discrimination loss and calibration loss are only moderately associated across the 43 cells ($r = 0.63$), so the failure mode is invisible to an AUROC-only notion of robustness.

5.10 Representation Depth and Supervision Efficiency

Intermediate features win 5/8 depth sweeps: RAD-DINO gains 1.3–1.6 AUROC points around 50% depth, DINOv2 prefers 75%, and shallower MedSigLIP/MedGemma features can lower ECE. Final blocks are therefore not uniformly optimal clinical representations, making layer choice a low-cost but consequential evaluation decision.

Medical pretraining shows its largest relative advantage in the low-label regime. With 1% of labels, MedGemma reaches 0.884 AUROC on APTOS and MedSigLIP/MedGemma/PanDerm reach 0.819/0.815/0.806 on ISIC. On CheXpert, six medically pretrained encoders meet ViT-B/16 at 1%, whereas MAE and ResNet-50 never cross it. Crossover analysis complements relative-retention LE by expressing supervision demand against a common reference.

5.11 Prompting versus Feature Probing

MedGemma exceeds prompted LLaVA-Med on APTOS/ISIC/CheXpert (74.7/51.0/82.0% vs. 49.3/17.8/79.5%) but remains below frozen-feature probing; one APTOS chain-of-thought setting parses only 64.6% of cases. Representation quality and the generative prediction interface are therefore distinct sources of reliability variation.

6 Discussion and Conclusion

CRS-Bench frames encoder selection as a controlled multi-objective comparison. Because discrimination is one profile component, positive AUROC–CRS association is expected; the key result is that adding calibration, label efficiency, and robustness changes 21 of 105 pairwise choices. Together with paired-seed uncertainty, this supports a stable leading tier and profile-specific trade-offs rather than a statistically resolved universal winner.

The reference-relative construction separates score stability from ranking transfer. Panel perturbations quantify sensitivity to reference composition, whereas fixed-anchor insertion keeps previously reported scores invariant to new candidates. LoRA and held-out data instead test whether the frozen ordering persists after representation or distribution change, delimiting transfer scope without altering the within-suite selection objective.

Encoder analyses further show why the profile should accompany CRS. Specialty pretraining improves matched-domain discrimination without uniform cross-domain gains, while broad medical pretraining is comparatively consistent in this panel under label scarcity and the observed radiology shift. Calibration is substantially repairable post hoc, but corruption can preserve AUROC while degrading probability quality; representation depth and prompting add variation beyond final-layer clean AUROC.

The scope is classification across three primary domains and one observed radiology shift; controlled perturbations model acquisition/image quality, and BioViL/CheXzero lack APTOS frozen cells. CRS is a benchmark-level comparative summary rather than a clinical-utility endpoint; prospective validation, subgroup analysis, operating-point selection, and human–AI interaction remain outside this protocol. CRS-Bench enables controlled, reference-relative multi-objective selection of medical image encoders beyond clean-test discrimination. Across 15 encoders, 21 of 105 AUROC–CRS pairwise choices reverse, and paired-seed resampling supports a stable leading tier rather than a single resolved leader. The resulting ranking reflects encoder reliability profiles rather than arbitrary comparison effects: it remains stable under reference-panel perturbations, while fixed-reference anchoring prevents previously reported scores from changing as new encoders are added.

CRS-Bench further shows that reliability dimensions capture complementary properties: specialization, calibration, label efficiency, and robustness reveal differences that clean AUROC alone cannot characterize. Its scope is a benchmark-level comparative summary rather than a clinical-utility endpoint, and future work should extend evaluation to additional modalities and prospective settings. Within these limits, reliability profiles should accompany, rather than replace, scalar performance metrics when selecting medical image encoders.

References

[1] Asia Pacific Tele-Ophthalmology Society. APTOS 2019 blindness detection, 2019.

[2] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision–language processing. In *ECCV*, pages 1–21. Springer, 2022.

[3] Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11: 3673, 2020.

[4] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1902.03368*, 2019.

[5] 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 *ICLR*, 2021.

[6] Hákon Gudbjartsson and Samuel Patz. The rician distribution of noisy MRI data. *Magnetic Resonance in Medicine*, 34(6): 910–914, 1995.

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, pages 1321–1330. PMLR, 2017.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.

[10] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.

[11] Carlos Hernández-Pérez, Marc Combalia, Sebastian Podlipnik, Noel C. F. Codella, Veronica Rotemberg, Allan C. Halpern, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Brian Helba, Susana Puig, Veronica Vilaplana, and Josep Malvehy. BCN20000: Dermoscopic lesions in the wild. *Scientific Data*, 11(1):641, 2024.

[12] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, pages 590–597, 2019.

[13] Xiaoqian Jiang, Melanie Osl, Jihoon Kim, and Lucila Ohno-Machado. Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, 19(2):263–274, 2012.

[14] Ruinan Jin, Zikang Xu, Yuan Zhong, Qiongsong Yao, Qi Dou, S. Kevin Zhou, and Xiaoxiao Li. FairMedFM: Fairness benchmarking for medical imaging foundation models. In *NeurIPS Datasets and Benchmarks Track*, 2024.

[15] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

[16] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. ELEVATER: A benchmark and toolkit for evaluating language-augmented visual models. In *NeurIPS Datasets and Benchmarks Track*, 2022.

[17] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *NeurIPS*, 2023.

[18] Alireza Mehrtash, William M. Wells, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 39 (12):3868–3878, 2020.

[19] Shentong Mo, Xufang Luo, Yansen Wang, and Dongsheng Li. A large-scale medical visual task adaptation benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2024.

[20] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, pages 2901–2907, 2015.

[21] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[22] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, pages 13991–14002, 2019.

[23] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 2025.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

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

[26] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12): 1399–1406, 2022.

[27] Dequan Wang, Xiaosong Wang, Lilong Wang, Mengzhang Li, Qian Da, Xiaoqiang Liu, Xiangyu Gao, Jun Shen, Junjun He, Tian Shen, Qi Duan, Jie Zhao, Kang Li, Yu Qiao, and Shaoting Zhang. MedFMC: A real-world dataset and benchmark for foundation model adaptation in medical image classification. *Scientific Data*, 10:574, 2023.

[28] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16:7866, 2025.

[29] Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Lie Ju, H. Peter Soyer, and Zongyuan Ge. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, 31(8):2691–2702, 2025.

[30] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark, 2019.

[31] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023.

[32] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2024.

[33] Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robbert R. Struyven, Timing Liu, Moucheng Xu, Mateo G. Lozano, Peter Woodward-Court, Yuka Kihara, Andre Altmann, Aaron Y. Lee, Eric J. Topol, Alastair K. Denniston, Daniel C. Alexander, and Pearse A. Keane. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981): 156–163, 2023.