Supplementary Material

Well-Calibrated but Unusable: Calibration Error Does Not Certify Zero-Shot Chest X-Ray VLMs. Tables self-number S1–S9 and are referenced by those labels from the main text.

Table S1. Experimental design: per-finding valid N, positive/negative counts, and native prevalence π on the patient/image-disjoint CheXpert calibration and official-test splits (confident labels in $\{0, 1\}$ only; the official-test split has no uncertain labels).

	Calibration				Test			
Finding	Valid N	Pos	Neg	π	Valid N	Pos	Neg	π
Atelectasis	870	195	675	0.224	500	153	347	0.306
Cardiomegaly	980	183	797	0.187	500	151	349	0.302
Consolidation	898	90	808	0.100	500	29	471	0.058
Edema	949	244	705	0.257	500	78	422	0.156
Pleural Effusion	959	307	652	0.320	500	104	396	0.208

Table S2. Per-finding test metrics at each backbone's ECE-floor method (the macro-ECE-minimizing post-hoc head), refit on calibration and scored on the disjoint official-test split. ECE is 15-bin; bracketed intervals are patient-level bootstrap 95% CIs (2000 resamples). AUROC gives the discrimination retained at the floor.

Finding	AUROC	ECE [95% CI]
medsiglip — Pairwise log-linear		
Atelectasis	0.837	0.129 [0.099, 0.165]
Cardiomegaly	0.917	0.188 [0.158, 0.222]
Consolidation	0.894	0.047 [0.034, 0.066]
Edema	0.896	0.059 [0.056, 0.094]
Pleural Effusion	0.919	0.086 [0.074, 0.117]
chexzero — Platt (per-finding bias)		
Atelectasis	0.771	0.092 [0.071, 0.135]
Cardiomegaly	0.879	0.162 [0.132, 0.197]
Consolidation	0.916	0.054 [0.041, 0.074]
Edema	0.891	0.034 [0.035, 0.070]
Pleural Effusion	0.926	0.084 [0.074, 0.116]
biomedclip — Pairwise log-linear		
Atelectasis	0.692	0.069 [0.032, 0.110]
Cardiomegaly	0.851	0.164 [0.130, 0.200]
Consolidation	0.883	0.049 [0.041, 0.070]
Edema	0.807	0.038 [0.030, 0.075]
Pleural Effusion	0.878	0.094 [0.076, 0.125]

Finding	AUROC	ECE [95% CI]
medclip — Pairwise log-linear		
Atelectasis	0.626	0.071 [0.032, 0.112]
Cardiomegaly	0.537	0.085 [0.048, 0.126]
Consolidation	0.604	0.061 [0.042, 0.081]
Edema	0.802	0.116 [0.097, 0.149]
Pleural Effusion	0.709	0.125 [0.098, 0.161]
openclip — Matrix scaling (ODIR)		
Atelectasis	0.471	0.082 [0.046, 0.122]
Cardiomegaly	0.602	0.139 [0.101, 0.179]
Consolidation	0.507	0.051 [0.031, 0.072]
Edema	0.580	0.065 [0.041, 0.100]
Pleural Effusion	0.633	0.074 [0.046, 0.112]

Table S3. SCA-T sign-flip robustness. Each SCA-T hyperparameter is varied around the published setting (lr 0.01, entropy-weight 0.1, 100 iterations), plus a no-adaptation control, over all 625 Protocol-A cells. 'Below-chance' = test AUROC < 0.5; 'low-ECE below-chance' also requires ECE < 0.05 (the hazardous tail). The basin persists across all 7 settings (60–89/625) and vanishes without adaptation (0/625); the baseline reproduces the main-text 89/625.

SCA-T setting	Below-chance (/625)	Low-ECE below-chance	Lowest-ECE cell (ECE / AUROC)
baseline (published)	89	5	0.029 / 0.443
lr = 0.001	66	10	0.021 / 0.434
lr = 0.1	60	0	0.071 / 0.283
entropy-wt = 0.01	86	0	0.057 / 0.300
entropy-wt = 1	79	5	0.036 / 0.363
iterations = 50	83	0	0.056 / 0.338
iterations = 200	86	2	0.019 / 0.487
no adaptation (zero-shot)	0	0	—

Table S4. Floor mechanism. A supervised linear probe on the frozen image features (L2-logistic, C by 5-fold CV on calibration, scored on the disjoint official-test split; same 15-bin and debiased smooth ECE as the main text). It does not breach the debiased smooth floor on any usable-discriminator backbone (probe smECE $\geq$ floor smECE), so the floor is not a low-capacity zero-shot-head artifact; its in-distribution CV-ECE drops below the floor on only two of five backbones. On the near-random backbones it instead rescues *discrimination* (MedCLIP AUROC 0.62→0.87), so discrimination is head-limited while no tested head removes the floor.

Backbone	Probe ECE	Probe smECE	Probe CV-ECE	Probe AUROC	Floor smECE	ZS AUROC
MedSigLIP	0.192	0.194	0.113	0.796	0.088	0.889
CheXzero	0.097	0.090	0.072	0.840	0.075	0.877
BiomedCLIP	0.089	0.082	0.087	0.780	0.073	0.798
MedCLIP	0.091	0.088	0.062	0.873	0.089	0.617
OpenCLIP	0.162	0.142	0.144	0.631	0.076	0.539

Table S5. SCA-T below-chance failure-mode taxonomy. The 89 cells with test AUROC < 0.5 split by prediction spread (std over the 500 test images, cut at 0.02): base-rate collapses (std < 0.02) flatline near prevalence and earn low ECE without ranking; active mis-rankers (std ≥ 0.02) vary yet rank below chance (most inverted AUROC 0.165) at ECE ≥ 0.125 overlapping well-ranking cells, so only AUROC separates them. "Healthy" (AUROC ≥ 0.5) shown for reference. The lowest-ECE exemplar in every row is MedCLIP × consolidation, at the stated prevalence π.

Mode	Cells	Median std	AUROC range	Lowest-ECE exemplar (π; ECE / AUROC / std)
Healthy (AUROC ≥ 0.5)	536	0.001	0.50–0.91	π=0.05: 0.020 / 0.548 / 0.005
Base-rate collapse	49	0.004	0.24–0.50	π=0.10: 0.029 / 0.443 / 0.012
Active mis-ranker	40	0.132	0.16–0.37	π=0.20: 0.125 / 0.293 / 0.074

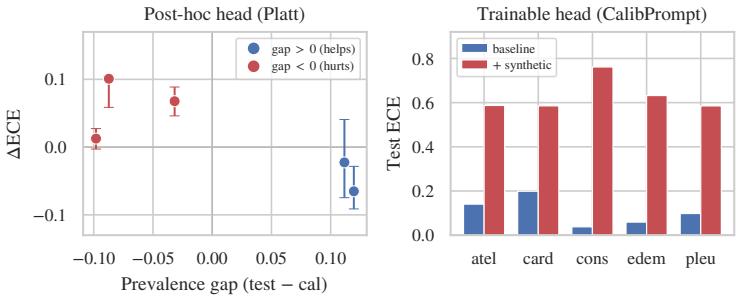

Fig. S1. Synthetic data-budget intervention. *Left:* post-hoc ΔECE vs. the test-minus-calibration prevalence gap (helps/hurts by its sign; finding-mean correlation 95% CI $[-0.92, -0.54]$). *Right:* folding positives into a soft prompt (CalibPrompt) overfits, lifting test ECE far above the floor (ΔECE +0.52/+0.50).

Table S6. The floor is a calibration-to-test gap. Per-finding-bias debiased smECE by fit regime: the cal→test deployment fit (c→t) exceeds a matched perfectly-calibrated Bernoulli null by $\Delta = 0.045$–0.067 on every backbone ($p < 1/400$), while the in-distribution 5-fold-CV refit (id) matches the null ($|\Delta| \leq 0.002$, $p \geq 0.19$). Up to sampling noise the floor is therefore a calibration-to-test transfer gap rather than a barrier; Table S8 separates how much of it is the prevalence mismatch of Table S1 specifically (§5). The id column here is a single 5-fold split (seed 0); Table S8 averages five, so the two differ by $\leq$0.003.

Backbone	$\text{smECE}_{\text{c}\to\text{t}}$	$\Delta_{\text{c}\to\text{t}}$	smECE_{id}	Δ_{id}
MedSigLIP	0.089	+0.058	0.032	+0.002
CheXzero	0.075	+0.045	0.030	−0.001
BiomedCLIP	0.076	+0.050	0.028	−0.002
MedCLIP	0.092	+0.067	0.021	−0.002
OpenCLIP	0.086	+0.065	0.020	−0.001

Table S7. Protocol A resampling, realized. Sampling is without replacement and one-directional, so a target above a finding's native π is reached by dropping negatives (marked $\downarrow n$) and one below it by dropping positives ($\downarrow p$); positives are never duplicated. Each cell gives realized π and calibration N. Values are seed- and backbone-invariant. No cell was rejected by the $\geq$ 5-positive / $\geq$ 50-row guards.

Finding	native π	π=0.01	π=0.02	π=0.05	π=0.1	π=0.2
Atelectasis	0.224	0.010/682 $\downarrow p$	0.020/689 $\downarrow p$	0.051/711 $\downarrow p$	0.100/750 $\downarrow p$	0.200/844 $\downarrow p$
Cardiomegaly	0.187	0.010/805 $\downarrow p$	0.020/813 $\downarrow p$	0.050/839 $\downarrow p$	0.100/886 $\downarrow p$	0.200/915 $\downarrow n$
Consolidation	0.100	0.010/816 $\downarrow p$	0.019/824 $\downarrow p$	0.051/851 $\downarrow p$	0.100/898 $\downarrow p$	0.200/450 $\downarrow n$
Edema	0.257	0.010/712 $\downarrow p$	0.019/719 $\downarrow p$	0.050/742 $\downarrow p$	0.100/783 $\downarrow p$	0.200/881 $\downarrow p$
Pleural Effusion	0.320	0.011/659 $\downarrow p$	0.020/665 $\downarrow p$	0.050/686 $\downarrow p$	0.099/724 $\downarrow p$	0.200/815 $\downarrow p$

Table S8. How much of the per-finding Platt calibration-to-test gap is prevalence mismatch. Per-finding Platt, debiased smooth ECE. *native*: fit on the calibration split at its own prevalence (within 0.003 of each backbone's reported floor except OpenCLIP, whose floor method is matrix scaling). *matched*: calibration subsampled so π_{cal} equals each finding's test prevalence, removing the base-rate gap only. *in-dist.*: 5-fold CV on test, scored out-of-fold. *closed* is the fraction of the native-to-in-distribution gap that prevalence matching alone removes. Matching accounts for about half this gap on the backbones that discriminate and nearly all of it on the two near-random ones (§5). The in-dist. column averages five 5-fold splits; Table S6 reports a single split (seed 0), so the two agree to $\leq$ 0.003.

Backbone	$\text{smECE}_{\text{native}}$	$\text{smECE}_{\text{matched}}$	$\text{smECE}_{\text{in-dist.}}$	closed
MedSigLIP	0.089	0.063	0.035	48%
CheXzero	0.075	0.052	0.031	51%
BiomedCLIP	0.076	0.048	0.028	58%
MedCLIP	0.092	0.025	0.021	95%
OpenCLIP	0.086	0.019	0.018	98%

Table S9. Triage operating points under institutional prevalence shift. The per-finding threshold is fit on the CheXpert calibration split at 90% sensitivity and then frozen. *missed* is 1−sensitivity and *flag* is the per-finding flag rate, both macro-averaged over the five findings; *study* is the fraction of studies referred, i.e. flagged for at least one finding, which is what a worklist actually absorbs and is strictly larger. *NIH*† re-tunes the threshold in-distribution on NIH and is an oracle, not a deployable rule. Sensitivity transfers; burden and PPV do not, and re-tuning recovers almost none of it (§4).

	CheXpert test				NIH test				$\text{NIH}^{\dagger}$ re-tuned		
Backbone	missed	flag	study	PPV	missed	flag	study	PPV	missed	flag	PPV
MedSigLIP	0.046	0.523	0.744	0.369	0.078	0.716	0.831	0.157	0.083	0.675	0.161
CheXzero	0.062	0.488	0.830	0.383	0.126	0.553	0.810	0.184	0.083	0.595	0.177
BiomedCLIP	0.089	0.596	0.900	0.312	0.059	0.762	0.940	0.147	0.083	0.719	0.152
MedCLIP	0.106	0.843	1.000	0.217	0.065	0.849	1.000	0.131	0.083	0.829	0.131
OpenCLIP	0.183	0.828	1.000	0.206	0.133	0.831	1.000	0.128	0.082	0.881	0.132

Table S10. Itemized version of the benchmark overview in Sec. 2 of the main paper: every dataset, protocol and intervention that carries a result, with the question each answers. Unless a row says otherwise, the regime is the deployment one: post-hoc maps are fit on a calibration split and scored on a disjoint test split. The cross-finding mean-bias analysis and the CalibPrompt trainable arm read off the same CheXpert fits.

Evaluation	Data	Varies	Question
Protocol A	CheXpert	cal. π, 5 levels	does the calibration prior move test ECE?
Protocol B	CheXpert	cal. N, native π	is any effect just calibration-set size?
Floor sweep	CheXpert	post-hoc map family	does any map we test breach the floor?
Data budget	CheXpert+synth.	# injected positives	does more data lower the floor?
Prevalence match	CheXpert	π_{cal}=π_{test}	how much of the floor is base-rate gap?
In-dist. refit	CheXpert	fit split (5-fold CV)	is the floor a barrier or a transfer gap?
Linear probe	CheXpert	recalibration head	is the floor a head-capacity artifact?
Acquisition shift	CheXpert cal.	AP vs. PA view	does the channel track a non-prevalence shift?
SCA-T stress test	CheXpert	transductive adapt.	can low ECE co-occur with bad ranking?
Triage thresholds	CheXpert, NIH	institutional shift	what does the transfer cost a clinic?
CheXphoto	CheXphoto	acquisition domain	do floor and decoupling reproduce?
NIH CXR14	NIH	institution + labeler	do floor and decoupling reproduce?