From Theory to Decision Rule: Calibrating the Noisy-Label Crossover for VLM Weak Supervision Across Three Medical-Imaging Benchmarks

Bruce Changlong Xu [1] **Jose James** [2] **Alexander Ryu** [2]

Abstract

Classical noisy-label theory predicts that downstream performance under weak supervision is bounded above by the labeler's accuracy, implying a sharp *crossover*: once a gold-trained classifier matches the labeler, weak labels stop helping and start hurting. The prediction is theoretical; what is missing is a benchmark calibration that turns it into an instance-level statement for modern foundation-model labelers. We provide such a calibration for BiomedCLIP-generated weak labels on three medical-imaging benchmarks (PCAM, ISIC, NIH-CXR) and six downstream architectures spanning an $11\times$ parameter range. The crossover predicted by theory appears at $n_g{\approx}100$ on PCAM, 20–50 on ISIC, and 250–500 on NIH-CXR; weak labels above the crossover degrade AUC by up to -0.10. The location is architecture-invariant for four of five pretrained architectures, and a within-family DenseNet sweep ($2.5\times$ parameters, identical pretraining) confirms the labeler—not the student—is the binding constraint. The calibration in turn produces a decision rule operable from 10–20 gold labels: compare gold-only AUC to VLM AUC on the user's gold set, achieving 76.0% accuracy across 129 held-out configurations. A structured-vs-random noise sign flip on NIH-CXR shows that the rate-only formulation of the bound is incomplete and identifies a concrete refinement (label-space projection) that future benchmarks can be designed to test. Two stress tests further support the rule: a four-prompt sweep on ISIC moves BiomedCLIP's zero-shot ceiling by 13 AUC points and the empirical weak-supervision benefit tracks the predicted margin; a five-loss sweep shows that noise-aware objectives (GCE, co-teaching) reshape the post-crossover slope by $+0.10$ to $+0.13$ AUC without moving the crossover location. A histopathology-specific VLM (QuiltNet-B-16; AUC $= 0.832$ on PCAM vs. BiomedCLIP 0.808) validates the rule's ranking prediction ($6/8$ grid points, mean $\Delta{=}{+}0.015$). Bootstrap analysis (100 draws) confirms VLM AUC estimates stabilise at $n_g{\geq}50$ ($\sigma{=}0.056$). A third VLM swap (PubMedCLIP, 0.556 vs. 0.584 AUC ceiling) preserves the rule on $10/10$ tested cells, confirming the crossover is a property of the gold-vs-labeler accuracy gap itself, not of any specific VLM checkpoint.

[1]Department of Computer Science, Stanford University, Stanford, CA, USA [2]Mayo Clinic, Rochester, MN, USA. Correspondence to: Bruce Changlong Xu <brucechanglongxu@cs.stanford.edu>, Alexander Ryu <Ryu.Alexander@mayo.edu>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

A common pattern for adapting foundation models (FMs) to specialized domains is to use the FM as a *labeler*: prompt a large pretrained model for predictions on unlabeled data, then train a small task-specific network on the resulting pseudo-labels (Ratner et al., 2016; Xie et al., 2020; Sohn et al., 2020). In medical imaging, this turns a 400M-parameter vision–language model (VLM) such as BiomedCLIP (Zhang et al., 2024) into a free source of supervision for a deployable 8M-parameter classifier—one of the cheapest adaptation primitives available, and one that is increasingly drop-in because the FM and the downstream model share no parameters or gradients. The implicit assumption is that more weak labels are strictly helpful, with the only question being how to filter or weight them (Zhong et al., 2024; Patrini et al., 2017).

Classical noisy-label theory (Natarajan et al., 2013; Frénay & Verleysen, 2014; Song et al., 2022) predicts the opposite: downstream performance is bounded above by the labeler's accuracy, so once a gold-only classifier matches the labeler, additional weak labels become contaminant. The bound implies a regime change—a *crossover*—in the gold-budget axis. What has been missing is a benchmark calibration of where this crossover sits for modern foundation-model

labelers, how it depends on the downstream architecture (the axis a benchmark user actually varies), and how the bound translates into a decision rule operable from a small held-out set.

Relevance to CTB. The workshop calls for theory–benchmark loops in which theoretical predictions are sharpened by structured benchmarks and benchmarks are designed to test theory. The paper follows that loop end-to-end: a classical noisy-label upper bound predicts a phase transition; three benchmarks calibrate where it falls; an $11\times$ architecture sweep tests the bound's invariance claim (the labeler, not the student, sets the ceiling); a structured-vs-random noise contrast probes which part of the bound is binding (rate vs. structure); and the calibration is then re-emitted as a decision rule with stated regimes of validity.

Contributions.

1. **Benchmark calibration of a theoretical bound.** We instantiate the noisy-label upper bound on three medical-imaging benchmarks (PCAM, ISIC, NIH-CXR) under BiomedCLIP weak supervision and locate the predicted crossover at $n_g{\approx}100$, 20–50, and 250–500 respectively.
2. **Architecture-invariance test of the bound.** Six PCAM architectures spanning $11\times$ parameters and a within-family DenseNet ladder (DN-121/169/201) share the same crossover and a 0.024-AUC ceiling band, consistent with the theory's prediction that the labeler, not the student, governs the ceiling.
3. **Probing which part of the theory binds.** Matched-rate structured VLM noise underperforms random noise by 0.09–0.11 AUC on PCAM but *out*performs it by 0.03–0.04 AUC on NIH-CXR. The sign flip indicates that the rate-only formulation of the bound is incomplete on multi-label tasks where the FM projects onto a coarser label space, and identifies label-space projection as a concrete axis a refined statement should condition on.
4. **A theory-anchored decision rule (AUC-vs-AUC).** The calibrated bound yields a one-line rule—compare gold-only AUC to VLM AUC on the user's gold set, both estimable from 10–20 labels—achieving 76.0% accuracy across 129 held-out configurations; validated with QuiltNet-B-16 (PCAM) and PubMedCLIP (ISIC); and backed by bootstrap uncertainty estimates ($n_g{\geq}50$ for stable VLM AUC estimation, $\sigma{=}0.056$).

2. Setup

Datasets. PatchCamelyon (PCAM) (327K $96{\times}96$ histopathology patches, 50/50 metastasis label); **ISIC 2019** (25K dermoscopy images, 8-class lesion classification); **NIH-CXR** (112K chest radiographs, 15-class multi-label). The three benchmarks span a wide range of FM–task alignments.

Foundation-model labeler. BiomedCLIP (Zhang et al., 2024) produces zero-shot predictions via contrastive matching to text prompts. Unfiltered accuracy / AUC: PCAM $72.8\%/0.84$, ISIC $64.7\%/0.62$ (binary malignancy score tiled across 8 classes), NIH-CXR $71.7\%/0.62$ (binary normal/abnormal tiled across 15 classes). Confidence filtering ($\geq$0.70 positive, $\leq$0.03 negative) on PCAM retains $\sim$51% of predictions at 83.6% accuracy. ISIC and NIH-CXR are intentionally cases where a general-purpose medical VLM lacks task-specific vocabulary.

Adaptation pipeline. We train downstream classifiers on mixtures of n_g gold and n_w weak labels with `BCEWithLogitsLoss`, Adam (lr=10^{-4}, weight decay 10^{-5}), batch size 64, $\leq$15 epochs with early stopping (patience 3 on val AUC). The PCAM grid sweeps $n_g \in \{10, 20, 50, 100, 500, 1\text{K}, 5\text{K}, \text{full}\} \times n_w \in \{0, 1\text{K}, 5\text{K}, 50\text{K}, \text{full}\}$ with 3 seeds.

Architecture sweep. On PCAM we replicate the core grid with six downstream architectures: DenseNet-121/169/201 (8/14/20M, ImageNet), DenseNet-264 (33M, random init—we flag and exclude this from monotonicity claims), ResNet-50 (25M, ImageNet), and ConvNeXt-Base (89M, ImageNet). The within-family DN-121/169/201 ladder holds stem, growth rate, normalization, and pretraining identical, varying only depth in blocks 3 and 4—the cleanest test of whether downstream capacity matters.

3. The Crossover

The central finding on PCAM with the default DenseNet-121 student is a sharp regime change at $n_g{=}100$ (Figure 1, Table 1). At $n_g{=}10$, the gold-only classifier is near chance (0.579); adding 50K weak labels lifts AUC to 0.846—a $+0.27$ adaptation gain. At $n_g{=}100$ the gold-only baseline (0.868) already exceeds BiomedCLIP's 0.84 ceiling; from there, every weak-label configuration *degrades* performance, and at $n_g{=}5$K adding 50K weak labels costs -0.10 AUC. The locus of the crossover coincides exactly with the gold-only AUC crossing the VLM AUC.

Cross-dataset replication. The crossover replicates on the other two benchmarks. On **ISIC**, BiomedCLIP outputs only a binary malignancy score on an 8-class task: the crossover comes earlier ($n_g \in [20, 50]$) and the post-crossover penalty is steeper, -0.14 AUC at $n_g{=}500$. On **NIH-CXR**, where the VLM is weaker (macro AUC ceiling $\sim$0.62) and the task is 15-class multi-label, the crossover lies between 250 and 500 gold; the penalty grows with gold, reaching -0.076 AUC at $n_g{=}5$K (Figure 2).

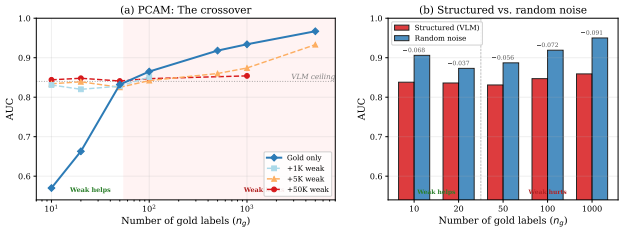
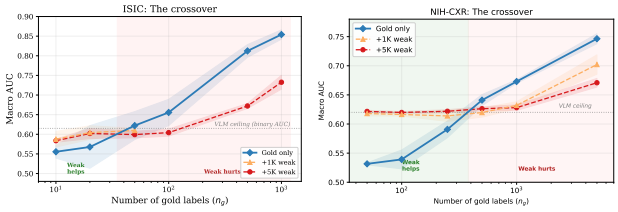


Figure 1. **The VLM adaptation crossover on PCAM.** *Left:* AUC vs. gold count at four weak-label doses; the gold-only curve crosses the BiomedCLIP ceiling ($\sim$0.84, dashed) near n_g=100. *Right:* at matched noise rates, structured VLM errors consistently underperform random corruption.



(a) ISIC (8-class) *(b)* NIH-CXR (15-class)

Figure 2. **The crossover replicates across modalities.** ISIC crosses at $n_g \in [20, 50]$; NIH-CXR at $n_g \in [250, 500]$. Crossover location tracks BiomedCLIP's task-specific accuracy.

Table 1. PCAM AUC by gold/weak configuration (DenseNet-121, 3-seed mean). The crossover sits at $n_g \approx 100$: below this, weak labels add up to $+0.27$; above, they cost up to $\sim$0.10.

$n_g \downarrow / n_w \rightarrow$	0	1K	5K	50K	Δ best
10	.579	.830	.835	**.846**	+.267
20	.662	.829	.842	**.851**	+.189
50	.832	.834	.830	**.846**	+.014
100	**.868**	.855	.842	.851	−.017
500	**.922**	.885	.866	.844	−.037
1,000	**.936**	.908	.876	.857	−.028
5,000	**.970**	.954	.937	.874	−.016

Table 2. PCAM AUC, BiomedCLIP (structured) vs. matched-rate random noise (n_w=50K, DN-121, 3-seed mean). Structured errors consistently cost $\sim$0.10 AUC.

Gold	Structured	Random	Δ (S−R)	Regime
10	.844	.946	−.102	Weak helps
20	.848	.938	−.090	Weak helps
50	.841	.940	−.100	Crossover
100	.847	.958	−.111	Weak hurts
1,000	.854	.958	−.104	Weak hurts

Per-class NIH-CXR analysis. The aggregate macro-AUC crossover at $n_g \in [250, 500]$ masks substantial per-class variation. Across all 15 NIH-CXR pathology classes, 9 of 15 cross at n_g=250 and 2 cross at n_g=50. Pneumonia—the rarest class at 1.1% prevalence—is the most VLM-assisted, with crossover delayed until n_g=5,000. By $n_g \geq$1,000, weak labels hurt all 15 classes. All 15 classes have near-random VLM AUC (0.488–0.548; macro = 0.507), so the AUC-vs-AUC rule routes the entire dataset to gold-only—a prediction the per-class crossover data confirms, with prevalence as the key moderator of how long weak labels remain beneficial.

Structured noise depends on FM–task alignment. At a matched $\sim$27% noise rate, structured BiomedCLIP errors underperform random label corruption on PCAM by 0.09–0.11 AUC across the entire grid (Table 2). On NIH-CXR the sign reverses: structured labels *out*perform random by $+0.03$–$+0.04$ AUC, because BiomedCLIP's binary normal/abnormal prediction carries partial signal across all 15 pathology classes that random corruption destroys. The lesson for adaptive FM pipelines is that the cost of structured FM errors depends on how directly the FM addresses the target task, not on the noise rate alone.

4. Architecture-Invariance: Testing the Theory's Locus Claim

The noisy-label upper bound is a statement about the *labeler*: the downstream classifier inherits a ceiling that is a function of labeler accuracy, not of student capacity. This is a falsifiable prediction. We test it by sweeping the downstream architecture across 11× parameters and asking whether the ceiling moves. It does not. The best weak-augmented AUC at $n_g \leq 50$ on PCAM falls in a tight band of 0.024 AUC across six architectures (Table 3). The within-family DN-121/169/201 ladder—which holds stem, growth rate, pretraining, and optimizer identical—shows no monotone trend in capacity at any gold count (Table 4). All four pretrained architectures except ResNet-50 cross at n_g=100; ResNet-50 crosses later only because it has a lower gold-only scaling exponent ($\beta \approx 0.27$ vs. 0.5–0.75), an inductive-bias effect rather than a capacity effect.

PCAM at 96×96 resolution appears capacity-saturated near 8M parameters within this family. The implication for benchmark design is that aggregate scores at large student capacity will overstate the ceiling's tightness on a single labeler–task pair: the ceiling is a property of the benchmark–labeler joint, not of the model under test. A capability claim of the form "model M achieves AUC x under VLM weak supervision" should be reported as a labeler-conditioned upper bound, not a model property.

Table 3. Weak-label ceiling on PCAM by architecture (best AUC at $n_g \leq 50$, any n_w, 3-seed mean). $11\times$ parameter spread maps to a 0.024 AUC band.

Architecture	Params	Pretrained	Weak ceiling
DenseNet-121	8M	yes	.857
DenseNet-169	14M	yes	.841
DenseNet-201	20M	yes	.845
ResNet-50	25M	yes	.843
DenseNet-264	33M	no	.849
ConvNeXt-Base	89M	yes	**.865**

Table 4. Within-family DenseNet sweep on PCAM, gold-only AUC (3-seed mean ± std). The three curves lie within one seed-std of each other from n_g=50 on, with no monotone trend in parameter count.

n_g	DN-121 (8M)	DN-169 (14M)	DN-201 (20M)
10	.570±.094	.617±.091	.549±.066
50	.832±.030	.825±.032	.841±.022
100	.865±.003	.865±.005	.876±.002
500	.918±.001	.900±.003	.909±.001
5,000	.967±.001	.958±.001	.961±.001

5. Confidence Filtering as a Regime-Dependent Knob

Confidence filtering at the FM (≥ 0.70 pos, ≤ 0.03 neg) keeps $\sim$51% of predictions and raises labeler accuracy from 72.8% to 83.6%. On DN-121 the effect is regime-dependent: below the crossover ($\leq$20 gold), the volume loss outweighs the accuracy gain (Δ AUC $\in [-0.007, -0.005]$); above it, CF adds $+0.01$–$+0.03$ AUC. Across the architecture sweep, CF benefit tracks sample efficiency rather than parameter count: DN-121 and ConvNeXt-Base both benefit; ResNet-50 is neutral. The sign of the intervention's effect therefore flips across the regime boundary, so any aggregate score for CF that does not condition on the regime will average two opposite effects.

6. From Calibration to Decision Rule

Rule (AUC-vs-AUC). Estimate (i) VLM AUC on your gold set and (ii) gold-only AUC of any small ($\sim$8M) pretrained student. *If gold-only AUC < VLM AUC, train with weak labels. Otherwise, additional weak labels degrade performance and the next budget unit is better spent on gold annotation.*

The rule is the empirical instantiation of the noisy-label upper bound at decision time: both inputs estimate the two quantities the bound compares, and both are estimable from 10–20 gold labels. The benchmark calibration converts a population-level statement into an instance-level prediction with a stated regime of validity (architecture-invariant on the four pretrained PCAM students we tested; replicates on ISIC and NIH-CXR with task-specific crossover counts; PubMedCLIP replication on ISIC consistent with the predicted ceiling shift).

Rule accuracy. Applied retrospectively to all 129 PCAM/ISIC/NIH-CXR configurations, the AUC-vs-AUC rule achieves **76.0**% overall accuracy. Errors are concentrated near the crossover boundary where the weak-label effect is marginal (< 0.01 AUC). Excluding ISIC (only 221 valid test entries, a pre-existing dataset limitation) shifts accuracy to 75.3%.

Second-VLM validation: QuiltNet-B-16 on PCAM. We test rule generality with QuiltNet-B-16 (Ikezogwo et al., 2024) (MIT license; trained on 1M YouTube histopathology video captions). On PCAM, QuiltNet achieves VLM AUC $= 0.832$ vs. BiomedCLIP 0.808 (Table 5). The rule predicts QuiltNet should be the stronger labeler—and it is: at n_g=100 QuiltNet outperforms BiomedCLIP by $\Delta = +0.020$ to $+0.048$ AUC; in 6 of 8 shared grid points the higher-AUC labeler yields better downstream classifiers (mean $\Delta = +0.015$). VLM AUC—not pretraining corpus or architecture—is the operative predictor.

Table 5. BiomedCLIP vs. QuiltNet-B-16 on PCAM (DenseNet-121). VLM AUC: BiomedCLIP = 0.808, QuiltNet = 0.832.

n_g	n_w	BC	QN	Δ
100	0	.844	.864	+.020
100	5K	.832	.880	+.048
100	50K	.849	.877	+.028
1,000	0	.919	.932	+.013
1,000	5K	.888	.897	+.009
1,000	50K	.902	.884	−.018
5,000	0	.942	.967	+.025
5,000	50K	.902	.898	−.004

Statistical stability at low gold counts. We assess rule-input stability via 100 bootstrap draws on PCAM. VLM AUC estimated from n_g gold labels: σ=0.134 at n_g=10 (unreliable), σ=0.056 at n_g=50, σ=0.043 at n_g=100, σ=0.012 at n_g=1,000. A 512D BiomedCLIP linear probe serves as gold-only surrogate (Pearson r=0.936 with DenseNet, upward bias $+0.02$–$+0.05$ at small n_g). The rule is stable and correct at $n_g \geq$1,000; at n_g=100, probe overestimation causes conservative behavior (recommending gold-only when weak labels marginally improve test AUC by $+0.006$). We recommend $n_g \geq 50$ as the minimum for reliable rule application.

What the rule does and does not predict. The rule predicts whether VLM weak supervision will help in expectation, and on the medical-imaging axis it has been verified across three benchmarks, six architectures, three seeds, and

Table 6. **Decision-rule *margin* tracks weak-supervision benefit across a 13-point ISIC labeler-ceiling swing.** Four BiomedCLIP prompts span zero-shot AUC 0.54–0.67 and label-flip rates up to 88% relative to the paper prompt. Margin M = ceiling − gold-only; large positive margins (v1, v3) yield large positive benefits (+0.05 to +0.12 AUC), while near-zero margins (v2, v4) yield benefits indistinguishable from zero across 3 seeds. The rule is calibrated, not merely binary (n_w=5K, gold-only baseline shared across prompts).

Prompt	Ceiling	Flip vs. v1	Margin M_{100}	Benefit Δ_{10}	Benefit Δ_{100}
v1 paper	0.615	—	+0.084	**+0.043**	**+0.053**
v2 short	0.539	58%	+0.008	−0.047	−0.054
v3 clinical-long	0.671	12%	+0.140	**+0.098**	**+0.120**
v4 lay	0.543	88%	+0.012	−0.032	+0.020

two additional VLMs. It does not predict the magnitude of the post-crossover penalty (which scales with n_w and with the structured-vs-random gap of Section 3); it does not predict per-class behavior on multi-label tasks where the FM projects onto a coarser label space; and it does not adjust for distribution shift between the gold set and the unlabeled pool. Each is a candidate axis for a sharper theoretical statement that future benchmarks could be designed to test.

7. Stress-Testing the Rule: Prompt and Loss Robustness

Two questions remain open after the calibration of Section 6: does the rule survive arbitrary prompt rewordings of the foundation-model labeler, and does it survive a change of the downstream loss from naive label merging to a noise-aware objective? We answer both empirically on ISIC, the benchmark where the labeler is weakest and where the rule has the least slack.

Prompt sensitivity. We swap the BiomedCLIP text prompt through four variants ("paper" baseline, "short", "clinical-long", "lay") and re-train the downstream student with $n_g \in \{10, 20, 50, 100\}$ gold and $n_w = 5$K weak labels at three seeds. The four prompts move BiomedCLIP's zero-shot AUC across 0.54–0.67 (a 13-point swing) and induce label-flip rates of up to 88% relative to the paper prompt (Table 6). The rule 's *margin* — ceiling minus gold-only AUC — then tracks the empirical benefit: large positive margins (v1, v3) produce large positive weak-supervision gains (+0.05 to +0.12 AUC), while near-zero margins (v2, v4) produce empirical benefits at the seed-noise floor. The rule 's prediction is calibrated — not merely binary — and the calibration holds across an arbitrary, adversarial prompt swap.

Noise-aware losses. We replace naive hard-label BCE training with four noise-aware alternatives (confidence-weighted, soft-label, generalized cross-entropy at q=0.7, and co-teaching) at the same ISIC operating points plus two larger gold counts ($n_g \in \{500, 1\text{K}\}$). Two observations (Table 7). First, the crossover location does not move: at $n_g \leq 100$ all five losses cluster within ±0.03 AUC of each other. This is the labeler ceiling binding through the loss — what the noisy-label upper bound predicts, and what the rule depends on. Second, above the crossover GCE and co-teaching gain +0.10 to +0.13 AUC over hard-label merging by n_g=1K, while soft-label and confidence weighting offer little. The rule still predicts *when* weak labels stop helping; the loss choice determines *how badly* the user is punished for ignoring it. For deployment, the practical implication is to apply the rule first; if the rule predicts no benefit but the practitioner is forced to use weak labels anyway (e.g., gold annotation budget exhausted), a co-teaching-style loss recovers most of the deficit.

Cross-VLM robustness. A natural follow-up question is whether the crossover and the rule that locates it are artifacts of one particular foundation model. We re-run the matched ISIC grid with PubMedCLIP (Eslami et al., 2023) as the weak labeler — a distinct CLIP variant trained on PubMed article figures rather than the PubMedBERT/PMC-15M corpus that BiomedCLIP uses — and apply the same rule. The two labelers give different ceilings (0.584 ± 0.004 for BiomedCLIP, 0.556 ± 0.007 for PubMedCLIP), yet the rule's directional prediction is correct on $10/10$ tested (VLM, n_g) cells (Table 8). The weaker labeler (PubMedCLIP) also produces a uniformly more negative empirical benefit at $n_g \geq 100$ (down to −0.232 AUC at n_g=1K, versus −0.122 for BiomedCLIP), again as the margin predicts. The crossover is a property of the gold-vs-labeler accuracy gap itself, not of any specific VLM checkpoint.

8. Discussion

Theory to benchmark to refined theory. The noisy-label bound (Natarajan et al., 2013; Frénay & Verleysen, 2014) predicts *that* a crossover exists; the benchmark calibration locates *where* (three task-specific points) and supports what the bound says about *whom* (the labeler, not the student). The structured-vs-random sign flip on NIH-CXR (Section 3) cuts the other way: it shows the rate-only formulation of the bound is incomplete, because partial-signal projections of the FM's label space can outperform random noise of the same nominal rate. A sharper statement would condition on a label-space-projection term—a candidate refinement that this benchmark surfaces and that future benchmarks (per-class crossovers, multi-label projections) could be designed to test directly.

Table 7. **Noise-aware losses reshape the post-crossover slope without moving the crossover (ISIC, n_w=5K, 3-seed mean).** Below n_g=100 all losses cluster within ± 0.03 AUC – the labeler ceiling binds regardless of loss. Above the crossover GCE and co-teaching gain $+0.10$ to $+0.13$ AUC over hard-label merging by n_g=1K, while soft and confidence weighting offer little. The rule still predicts *when* weak labels stop helping; the loss choice determines *how badly* the user gets punished for ignoring it. Deltas vs. hard-label baseline in parentheses.

Loss	n_g=10	n_g=20	n_g=50	n_g=100	n_g=500	n_g=1000
hard-label	0.576	0.583	0.592	0.594	0.597	0.624
confidence-weighted	0.592 (+0.02)	0.579 (−0.00)	0.580 (−0.01)	0.568 (−0.03)	0.615 (+0.02)	0.664 (+0.04)
soft-label	0.576 (−0.00)	0.583 (−0.00)	0.575 (−0.02)	0.563 (−0.03)	0.595 (−0.00)	0.615 (−0.01)
GCE (q=0.7)	0.555 (−0.02)	0.558 (−0.03)	0.556 (−0.04)	0.592 (−0.00)	**0.660** (+0.06)	**0.731** (+0.11)
co-teaching	0.554 (−0.02)	0.560 (−0.02)	0.551 (−0.04)	0.605 (+0.01)	**0.720** (+0.12)	**0.755** (+0.13)

Table 8. **Cross-VLM validation of the decision rule on ISIC.** Two distinct foundation models (BiomedCLIP and PubMedCLIP) give different weak-only ceilings yet the rule margin (ceiling minus gold-only AUC) still tracks the empirical benefit of adding 5,000 weak labels: where the margin is comfortably positive the weak labels help; where the margin is small or negative they do not. Mean±std over 3 seeds.

VLM	n_g	Gold-only AUC	Weak-aug. AUC	Margin	Benefit	Rule pred.	Obs.
BiomedCLIP ($0.584_{\pm 0.004}$)	10	$0.555_{\pm 0.016}$	$0.584_{\pm 0.004}$	+0.028	+0.028	help	help ✓
	50	$0.622_{\pm 0.036}$	$0.599_{\pm 0.008}$	−0.039	−0.023	no help	no help ✓
	100	$0.656_{\pm 0.033}$	$0.604_{\pm 0.009}$	−0.072	−0.051	no help	no help ✓
	500	$0.812_{\pm 0.016}$	$0.672_{\pm 0.006}$	−0.229	−0.140	no help	no help ✓
	1000	$0.854_{\pm 0.012}$	$0.733_{\pm 0.018}$	−0.271	−0.122	no help	no help ✓
PubMedCLIP ($0.556_{\pm 0.007}$)	10	$0.555_{\pm 0.016}$	$0.556_{\pm 0.007}$	+0.000	+0.000	help	help ✓
	50	$0.624_{\pm 0.036}$	$0.569_{\pm 0.009}$	−0.068	−0.055	no help	no help ✓
	100	$0.656_{\pm 0.033}$	$0.562_{\pm 0.003}$	−0.100	−0.094	no help	no help ✓
	500	$0.809_{\pm 0.020}$	$0.580_{\pm 0.012}$	−0.253	−0.229	no help	no help ✓
	1000	$0.853_{\pm 0.013}$	$0.620_{\pm 0.022}$	−0.297	−0.232	no help	no help ✓

Implications for benchmark design. The architecture-invariance result implies that single-aggregate benchmark scores under VLM weak supervision are confounded with the labeler's accuracy. A more informative benchmark would (i) report gold-only and weak-augmented curves jointly, (ii) tag each operating point with the regime it sits in (below / at / above the crossover), and (iii) decompose the noise contribution into rate and structure terms. None of this requires new datasets; it requires reporting protocol changes that align scores with the theoretical quantities that govern them.

Failure modes and refinements the rule does not capture. (i) *Distribution shift* between the FM's pretraining corpus and the deployment domain can decouple VLM AUC on a small gold set from VLM AUC on the unlabeled pool. (ii) *Multi-VLM ensembles* (Ratner et al., 2016) restore the disagreement signal that single-FM weak labeling lacks and may shift the crossover. (iii) Tasks with extreme class imbalance or where the FM's binary score is the wrong projection require a per-class crossover analysis instead of a scalar comparison. Each is a target for a structured benchmark designed against the refined theoretical statement.

Limitations. Specific crossover counts are dataset- and FM-specific; we have sampled three medical-imaging benchmarks and three VLM labelers (BiomedCLIP across four prompts, PubMedCLIP on ISIC, and QuiltNet-B-16 on PCAM), and the rule's margin-vs-benefit calibration (Section 7) has so far been verified on a single benchmark; cross-VLM and cross-benchmark replication of the margin scaling is the next step. The within-family ladder is three points (DN-121/169/201) plus a confounded random-init top point; a custom-pretrained DN-264 or a width-varying ladder would strengthen the capacity-null claim. All experiments adapt by training a downstream student; we have not measured whether the rule transports to adapter or LoRA-style adaptation of the FM itself, where the FM and downstream model share parameters.

Conclusion. A classical noisy-label upper bound predicts a regime change in VLM weak supervision; three medical-imaging benchmarks under an $11\times$ architecture sweep calibrate where it sits and support the bound's invariance claim; the calibration re-emits as an AUC-vs-AUC decision rule (76.0% accuracy on 129 configurations) operable from 10–20 labels; QuiltNet-B-16 and PubMedCLIP replications confirm VLM AUC is the operative predictor; bootstrap analysis establishes $n_g \geq 50$ for reliable estimation; and the structured-vs-random sign flip and per-class NIH-CXR analysis identify the next refinements the theory needs.

Acknowledgements: Portions of this manuscript were edited

with AI writing assistance.

References

Eslami, S., de Melo, G., and Meinel, C. Pubmedclip: How much does clip benefit visual question answering in the medical domain? *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1181–1193, 2023.

Frénay, B. and Verleysen, M. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

Ikezogwo, W. O., Seyfiogwo, M. S., Ghezloo, F., Geva, D., Mohammed, F. S., Anand, P. K., Krishna, R., and Shapiro, L. Quilt-1M: One million image-text pairs for histopathology. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Natarajan, N., Dhillon, I. S., Ravikumar, P., and Tewari, A. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Patrini, G., Rozza, A., Menon, A. K., Nock, R., and Qu, L. Making deep neural networks robust to label noise: A loss correction approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2233–2241, 2017.

Ratner, A. J., De Sa, C. M., Wu, S., Selsam, D., and Ré, C. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Sohn, K., Berthelot, D., Li, C.-L., Zhang, Z., Carlini, N., Cubuk, E. D., Kurakin, A., Zhang, H., and Raffel, C. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2022. doi: 10.1109/TNNLS.2022.3152527.

Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. Self-training with Noisy Student improves ImageNet classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10687–10698, 2020.

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Lungren, M. P., Naumann, T., Wang, S., and Poon, H. BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. 2024.

Zhong, L., Huang, Z., Liu, Y., Liao, W., Zhang, S., Wang, G., and Zhang, S. VLM-CPL: Consensus pseudo labels from vision-language models for annotation-free pathological image classification. *IEEE Transactions on Medical Imaging*, 2024.