CAVE: Counterfactual Auditing of Visual Evidence for Budgeted Local Supervision

Anonymous authors
Paper under double-blind review

Abstract

Medical images often have diagnosis labels but few local masks to supervise predicted evidence. With a limited annotation budget, we must select masks before their target regions are visible. Uncertainty and diversity identify ambiguous or underrepresented images but do not test whether predicted regions contribute to the known-class response. Expert-ROI criteria require the mask that the query will acquire. We introduce Counterfactual Auditing of Visual Evidence (CAVE), which compares known-class responses when a predicted region is retained or removed and measures reflection stability before ranking mask queries. Acquired masks supervise the shared learner. Under the original validation-selected five-seed profiles, CAVE has higher mean Dice budget-AUC than its matched control on BUSI, ISIC2018, and RSNA and the highest RSNA mean among tested selectors, although paired intervals span zero. Under a separate ten-seed protocol counting every queried mask, its mean Dice budget-AUC is lower than the matched control on all three datasets. The audit measures spatial evidence before annotation, but its current score does not reliably identify valuable masks.

1 Introduction

Medical-image classifiers can be trained from image-level diagnoses even when lesion masks or opacity regions are missing. Local masks offer a more direct way to supervise where a model predicts evidence, but acquiring them for every image is costly (Wang et al., 2021; Rädsch et al., 2023). Under a fixed budget, we choose images for local annotation before their target masks are available.

Common acquisition rules prioritize uncertain predictions or images far from those already labeled (Sener & Savarese, 2018; Ash et al., 2020). Selective-supervision and medical-image methods extend these criteria to annotation granularity, predicted segmentation quality, and localized similarity (Vijayanarasimhan & Grauman, 2008; Belharbi et al., 2021; Shi et al., 2024; Zhu et al., 2025). These signals help identify ambiguous or underrepresented images. They do not ask whether the model's current spatial prediction supports the class that is already known. A confident classifier can localize poorly, while an uncertain one can already predict a useful region. Confidence alone does not distinguish the two cases.

Spatial feedback can correct evidence during training (Ross et al., 2017; Schramowski et al., 2020; Teso & Kersting, 2019). Acquisition rules that compare a prediction with an expert region require that region before ranking (Uddin et al., 2026; Wang et al., 2026). Here, the expert mask is precisely what the query will obtain. The available information is the image, its class, the current model, and masks acquired in earlier rounds. Figure 1 shows the information available at ranking time. The predicted region can be tested before a query. Comparison with the target requires the acquired mask.

A discrepancy in predicted evidence may identify an unreliable region without identifying a valuable annotation. Repair after a query also depends on the revealed mask, the other selected images, and the model update. We measure the pre-query response and the post-query improvement separately to test whether the former predicts the latter.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

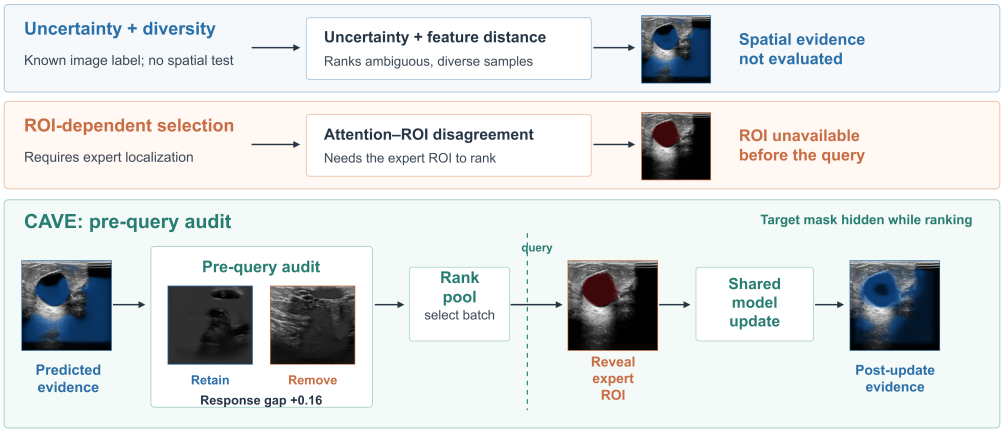

Figure 1: Ranking-time information. Uncertainty and diversity do not inspect spatial evidence. ROI-guided ranking needs the hidden target mask. Our audit uses predicted evidence before the query.

We compare known-class responses under retained and removed views of the predicted region, and use reflection to measure its stability. These signals join uncertainty and feature distance in a budgeted ranking rule. Selected masks then supervise the shared learner (Figure 2). We evaluate selection and post-query repair under the original five-seed dataset profiles and an independent ten-seed, all-counted protocol. We make three contributions:

- We formalize known-label, hidden-mask acquisition with mask revelation after selection.
- We test predicted-region support beyond uncertainty and diversity. The configured five-seed gains have paired intervals spanning zero.
- We test the audit with fully counted queries over ten new seeds. Negative effects on all three datasets show that its current signal does not improve acquisition under this protocol.

2 Related Work

Selection with limited annotation. Uncertainty sampling queries ambiguous predictions, while Core-set, ProbCover, and BADGE add coverage or diverse gradient information (Settles, 2011; Sener & Savarese, 2018; Yehuda et al., 2022; Ash et al., 2020). Medical-image methods estimate segmentation quality, combine pseudo-labeling with annotation selection, or use localized similarity and cold-start features (Shi et al., 2024; Zhou et al., 2024; Li et al., 2022; Zhu et al., 2025; Jiang et al., 2025; Chen et al., 2024). Perturbation consistency provides another image-level ranking cue (Wang et al., 2025). These criteria address difficulty or representation coverage. The known-label, hidden-mask setting also permits a test of the region on which the current class response depends.

Annotation value after updating. An ideal query would improve the learner after its annotation is added. Query value has been estimated from error reduction or selection rewards (Konyushkova et al., 2017; Casanova et al., 2020), and segmentation policies use predicted accuracy or mutual information as tractable proxies (Shi et al., 2024; Föllmer et al., 2024). The mask is hidden at ranking time, so a candidate's update value can be measured only after annotation. We test whether the pre-query score is associated with subsequent repair on queried images.

Spatial evidence and perturbation. Expert regions can constrain a model's explanation during training (Ross et al., 2017; Schramowski et al., 2020; Teso & Kersting, 2019). Retaining or removing image content tests a model's class response to a region (Fong & Vedaldi, 2017; Fong et al., 2019). The result depends on the masking operator (Hooker et al., 2019). ROI-guided acquisition compares a predicted region with expert evidence (Uddin et al., 2026; Wang et al., 2026). We instead perturb the current predicted region before acquisition and compare the ensuing selections with a matched rule that omits the spatial audit.

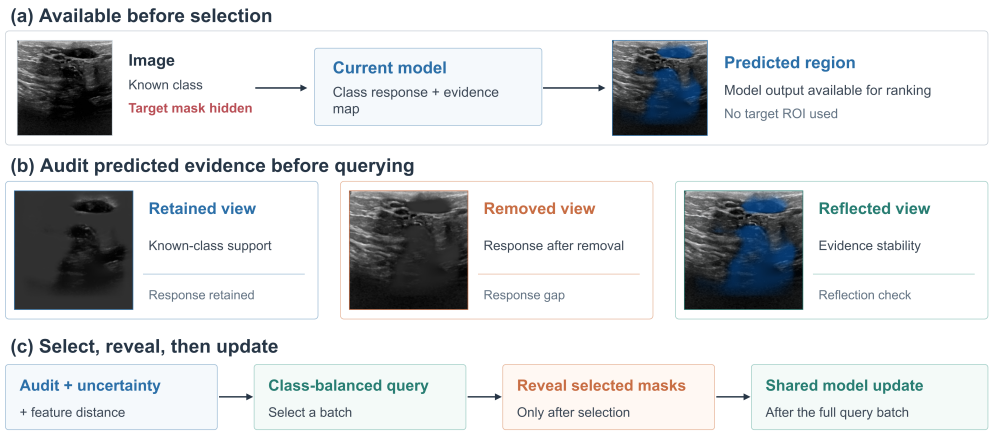

Figure 2: Pre-query audit and selection on BUSI. Known labels and predicted regions produce retained, removed, and reflected views. Their responses, uncertainty, and feature distance rank a class-balanced batch. Queried masks then supervise the shared update.

3 METHOD

At each round, the current model has access to the image-level label of every training sample and the local masks acquired so far. The audit examines the predicted evidence for each remaining image and combines it with standard uncertainty and diversity signals. The selected masks are revealed only after ranking and are used to retrain the model, as shown in Figure 2.

3.1 PROBLEM FORMULATION

Let $\mathcal{D} = \{(x_i, y_i, e_i)\}_{i=1}^{N}$ denote the training set, where $x_i \in \mathbb{R}^{C\times H\times W}$ is an image, $y_i \in \{1, \dots, K\}$ its class, and $e_i \in \{0,1\}^{H\times W}$ its local evidence mask. The learner observes (x_i, y_i) for every image, whereas e_i is hidden until image i is queried. Every method starts with the same small stratified evidence set $\mathcal{Q}_0$. At round t, $\mathcal{Q}_t$ indexes the revealed masks and $\mathcal{P}_t$ the remaining candidates. A batch $B_t \subseteq \mathcal{P}_t$ contains at most b_t indices. Let $\mathcal{F}$ index images whose class labels imply empty masks, and let $\mathcal{E}_t = \mathcal{Q}_t \cup \mathcal{F}$ index the samples contributing evidence supervision. The set $\mathcal{F}$ is nonempty only when the annotation protocol provides these empty masks without query cost. After annotation, $\mathcal{Q}_{t+1} = \mathcal{Q}_t \cup B_t$ and $\mathcal{P}_{t+1} = \mathcal{P}_t \setminus B_t$. Spatial indices h, w range over H, W, and k indexes the K classes.

The model has parameters θ, class logits $f_\theta(x) \in \mathbb{R}^K$, and evidence logits $g_\theta(x) \in \mathbb{R}^{H\times W}$. Training uses every class label and the evidence masks indexed by $\mathcal{E}_t$. Let $\ell_{\mathrm{CE}}^{\mathrm{bal}}$ be class-weighted cross-entropy and ℓ_{ev} the sum of positive-class-weighted binary cross-entropy and Dice loss. The nonnegative coefficient λ weights local supervision. The training objective is

$$\mathcal{L}_t(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell_{\mathrm{CE}}^{\mathrm{bal}}(f_\theta(x_i), y_i) + \frac{\lambda}{|\mathcal{E}_t|}\sum_{i\in\mathcal{E}_t} \ell_{\mathrm{ev}}(g_\theta(x_i), e_i). \tag{1}$$

For the original five-seed protocol, let $\mathcal{T}(\mathcal{E}_t)$ denote training afresh on the current supervision. It returns the model θ_t selected by validation evidence Dice. During acquisition at round t, θ denotes this current model. The separate unified protocol uses a staged update defined below.

Let $\mathcal{V} = \{(x_j, y_j, e_j)\}_{j=1}^{N_{\mathcal{V}}}$ be a validation set disjoint from the query pool and test set. In the original five-seed protocol, its masks select trained models, area references, and profiles. They do not enter $\mathcal{E}_t$. The training-pool budget counts queried masks, and the validation masks form an additional annotation resource. The unified protocol uses validation image labels for decisions and counts ev-

ery queried training mask, including known-empty masks, toward its budget. Both protocols rank candidates without their masks.

3.2 Evidence Audit through Paired Perturbations

For $x_i \in \mathcal{P}_t$, the current model produces

$$p_i(y) = \mathrm{softmax}(f_\theta(x_i))_y, \qquad z_i = \mathrm{sigmoid}(g_\theta(x_i)). \tag{2}$$

Here $z_i \in [0,1]^{H\times W}$ is the soft evidence map at the input resolution. The notation $p_\theta(y \mid x)$ denotes the same class probability on an unindexed input.

Let $\mu_i \in \mathbb{R}^{C\times H\times W}$ contain the per-channel spatial mean of x_i, copied over all locations. The mask z_i is broadcast over the C image channels to form evidence-retained and evidence-removed views:

$$\begin{aligned} \widetilde{x}_i^{\mathrm{ret}} &= x_i \odot z_i + \mu_i \odot (1 - z_i), \\ \widetilde{x}_i^{\mathrm{rem}} &= x_i \odot (1 - z_i) + \mu_i \odot z_i. \end{aligned} \tag{3}$$

Define $p_i^{\mathrm{ret}}(y) = p_\theta(y \mid \widetilde{x}_i^{\mathrm{ret}})$ and $p_i^{\mathrm{rem}}(y) = p_\theta(y \mid \widetilde{x}_i^{\mathrm{rem}})$. The paired inputs measure the known-class response with the predicted region preserved or suppressed. With $[v]_+ = \max(v, 0)$ and $\beta = \frac{1}{2}$, support failure under these counterfactual views is

$$o_i = \left[p_i^{\mathrm{rem}}(y_i) - p_i^{\mathrm{ret}}(y_i)\right]_+ + \beta p_i^{\mathrm{rem}}(y_i). \tag{4}$$

The first term rises when removal gives a stronger known-class response than retention. The second rises when confidence persists after removal.

Let T denote horizontal reflection, with T^{-1} aligning the reflected output. Reflection instability is $s_i = (HW)^{-1}\|z_i - T^{-1}(\mathrm{sigmoid}(g_\theta(Tx_i)))\|_1$. Let $a_i = (HW)^{-1}\sum_{h,w} z_{i,h,w}$ be predicted area and $\bar{a}_{\mathrm{ref}}$ the nonempty-mask area reference from validation, $\epsilon_a > 0$ a fixed known-empty reference, and $\nu > 1$ a fixed area scale. With $\bar{a}(y_i) = \epsilon_a$ for a known-empty class and $\bar{a}(y_i) = \bar{a}_{\mathrm{ref}}$ otherwise, let $\epsilon_r > 0$ stabilize the area-fit ratio. Define area deviation $c_i^{\mathrm{area}} = |a_i - \bar{a}(y_i)|$ and area fit $r_i^{\mathrm{area}} = \exp[-|\log((a_i + \epsilon_r)/\bar{a}_{\mathrm{ref}})|/\log\nu]$. These signals check stability and size without reading the candidate mask.

Let $p_i^{(1)}$ and $p_i^{(2)}$ be the two largest class probabilities. Uncertainty is measured by entropy $u_i = -\sum_k p_i(k)\log p_i(k)/\log K$ and margin $m_i = 1-(p_i^{(1)}-p_i^{(2)})$. For the pooled model feature $\phi_\theta(x_i)$ and a positive stabilizer ϵ_h, define $h_i = \phi_\theta(x_i)/(\|\phi_\theta(x_i)\|_2 + \epsilon_h)$ and $d_i = 1 - \max_{j\in\mathcal{Q}_t} h_i^\top h_j$. Appendix G gives the profile weights.

3.3 Acquisition Score and Batch Selection

Let $\xi_i = (m_i, u_i, o_i, s_i, c_i^{\mathrm{area}}, r_i^{\mathrm{area}}, a_i)$ collect margin, entropy, support failure, reflection instability, area deviation, area fit, and predicted area. For any scalar candidate signal v_i, write $v_{\min} = \min_{j\in\mathcal{P}_t} v_j$ and $v_{\max} = \max_{j\in\mathcal{P}_t} v_j$. Normalize within the current pool by

$$\mathcal{N}_t(v_i) = \begin{cases} (v_i - v_{\min})/(v_{\max} - v_{\min}), & v_{\max} > v_{\min}, \\ 0, & v_{\max} = v_{\min}. \end{cases} \tag{5}$$

For a fixed profile π, let $\alpha_j^{(\pi)} \geq 0$ weight component j of ξ_i, with $\sum_j \alpha_j^{(\pi)} = 1$. Let $q_i^{(\pi)}$ attenuate known-empty candidates, and $w_A^{(\pi)} \in [0,1]$ weight the audit against feature distance. The combined score is

$$\begin{aligned} A_i^{(\pi)} &= q_i^{(\pi)} \sum_j \alpha_j^{(\pi)} \mathcal{N}_t(\xi_{ij}), \\ S_t^{(\pi)}(i) &= w_A^{(\pi)} \mathcal{N}_t(A_i^{(\pi)}) + (1 - w_A^{(\pi)})\mathcal{N}_t(d_i). \end{aligned} \tag{6}$$

MPD, EDD, and HMD are preset profiles whose coefficients appear in Table 16. Nonempty-mask validation Dice selects one profile per dataset, fixed across seeds and budgets.

Candidates are sorted by $S_t^{(\pi)}$. Within each available class, the selector first takes the highest-scoring $\max(1, \lfloor b_t/K \rfloor)$ candidates, subject to the batch budget, and then fills any remaining positions by

the global score. Baselines that assign one score per image use the same class-balanced procedure. Methods that select a batch jointly keep their original batch selection within the same annotation budget. Ranking uses x_i, y_i, the current model, validation-derived area references and profile, and the previously queried set. The hidden mask e_i remains unavailable until the candidate is selected.

3.4 Unified Audit and Matched Control

The ten-seed comparison uses a single rule across datasets. For a quantity v_i, let $P_{y_i}(v_i)$ denote its percentile rank among currently scored images with class label y_i. The control ranks the full unqueried pool using the entropy u_i and feature distance d_i defined above. The audit subsequently reranks a fixed shortlist. The control score is

$$U_i = 0.90P_{y_i}(u_i) + 0.10P_{y_i}(d_i). \tag{7}$$

For a batch of size b_t, the audit first retains up to $2\max(1, \lfloor b_t/K \rfloor)$ highest-U_i candidates per available class. Only this shortlist receives the paired perturbation and reflection forwards. Within the shortlist, the audit combines support failure o_i from Equation 4 with reflection instability s_i, defined above as the mean absolute difference between aligned evidence maps. Its score is

$$A_i^{\mathrm{u}} = 0.50P_{y_i}(o_i) + 0.50P_{y_i}(s_i). \tag{8}$$

The selector ranks shortlisted candidates by A_i^{u}, using U_i to resolve equal audit scores. It reserves an equal query quota per known class and fills any remaining slots from the control ranking. All percentiles use only current model outputs and known labels. Candidate masks are revealed after batch selection.

At 1%, both policies share the same trained model. At each later budget, both replay all acquired masks while continuing from the previous model state. Let $\mathcal{U}_t(\theta_t, \mathcal{E}_{t+1})$ denote this shared staged update. Validation image-label macro-F1 selects the model at each budget, and the local mask threshold remains 0.5. Validation masks may be evaluated after selection but do not choose a score, profile, model, or threshold. The unified protocol has $\mathcal{F} = \varnothing$: every training mask used for evidence supervision is budget-counted.

3.5 Repair Value after Annotation

For the original five-seed repair analysis, the value of an acquired mask is defined by its effect on the scratch-trained model. Let ℓ_{mask} denote unweighted binary cross-entropy plus Dice loss, and define validation evidence risk as

$$\mathcal{R}(\theta) = \frac{1}{N_{\mathcal{V}}} \sum_{j=1}^{N_{\mathcal{V}}} \lambda \ell_{\mathrm{mask}}(g_\theta(x_j), e_j). \tag{9}$$

The repair value after full retraining for candidate i at round t is

$$v_t(i) = \mathcal{R}(\mathcal{T}(\mathcal{E}_t)) - \mathcal{R}(\mathcal{T}(\mathcal{E}_t \cup \{i\}))\,. \tag{10}$$

Evaluating $v_t(i)$ requires its revealed mask and a separate model update. After querying, a one-step response $\widetilde{v}_i^{(1)}$ measures score–repair association, and full retraining measures the value of the selected batch. For a queried batch, replace $\{i\}$ in Equation 10 with B_t. Batch value need not equal the sum of individual values because training updates interact.

For retrospective spatial-error analysis, let $a_i^{\mathrm{gt}} = (HW)^{-1} \sum_{h,w} e_{i,h,w}$ denote the target area after mask revelation. The hidden area gap is $G_i = |a_i - a_i^{\mathrm{gt}}|$. This outcome is not available to the ranking rule.

3.6 Model Update

The initial model is $\theta_0 = \mathcal{T}(\mathcal{E}_0)$. At round t, the selector returns B_t, its masks are revealed, and the queried set becomes $\mathcal{Q}_{t+1} = \mathcal{Q}_t \cup B_t$. The next evidence set is $\mathcal{E}_{t+1} = \mathcal{Q}_{t+1} \cup \mathcal{F}$, and $\theta_{t+1} = \mathcal{T}(\mathcal{E}_{t+1})$ is trained from scratch in the original five-seed protocol. The unified comparison instead uses $\theta_{t+1} = \mathcal{U}_t(\theta_t, \mathcal{E}_{t+1})$. Both policies within either comparison share the split, budget, and update settings. The paired perturbations add acquisition computation, and we report their measured selection-time cost in Section 4.

4 EXPERIMENTS

4.1 EVALUATION SETUP

We evaluate BUSI breast ultrasound images (Al-Dhabyani et al., 2020), ISIC2018 dermoscopy (Codella et al., 2019), and RSNA chest radiographs (Radiological Society of North America, 2018). BUSI uses image-level splits because patient identifiers are unavailable. ISIC2018 keeps each recorded patient within one split, falling back to lesion and image IDs when needed. Melanoma forms the positive diagnosis class. RSNA uses a fixed balanced subset of 2,048 opacity and 2,048 no-opacity images, with overlapping opacity boxes merged into one mask. Table 2 gives split sizes and the first and last queried-mask counts for the reported budgets.

The original five-seed evaluation compares selectors with the same class labels, split, initialization, learner, and 1%, 5%, 10%, and 20% cumulative budgets. BUSI and ISIC2018 count every queried training mask. For RSNA, class-known empty training masks are available without a query. The paid pool contains opacity cases. Nonempty-mask validation Dice selects trained models, area references, and one fixed profile per dataset: MPD for BUSI, EDD for ISIC2018, and HMD for RSNA. The separate ten-seed comparison uses the unified audit and matched control, counts every queried mask, and selects trained models by validation image-label macro-F1. Its policies share the split, initial model, staged update, and all other training settings. Within either protocol, candidate masks are hidden during scoring and test masks are reserved for evaluation. The protocols use different updates and validation decisions, so we compare selectors only within each protocol.

The primary endpoint is hard Dice on test images with nonempty masks at a fixed 0.5 threshold. Normalized trapezoidal Dice budget-AUC spans the four budgets. Within each protocol, differences are paired by seed. The original evaluation reports means and sample standard deviations over five seeds. The unified comparison reports ten-seed paired effects. Each comparison uses the same learner across its acquisition methods. Appendix A gives the paired decision analysis. Medical baselines are compared as ranking rules in the shared learner. EG-AL uses the predicted ROI because the expert ROI is hidden before a query.

4.2 RESULTS

A pre-query signal improves acquisition only if the masks it selects produce better held-out localization under the same annotation budget. The ten-seed comparison tests this against a matched selector with every query counted.

Method	BUSI	ISIC2018	RSNA
Matched control	0.369 ± 0.048	0.702 ± 0.018	0.452 ± 0.019
Audit (ours)	0.358 ± 0.056	0.693 ± 0.017	0.447 ± 0.017
Entropy	0.362 ± 0.051	0.710 ± 0.014	0.446 ± 0.019
Core-set	0.362 ± 0.059	0.695 ± 0.020	0.454 ± 0.012
BADGE	0.376 ± 0.049	0.696 ± 0.030	0.448 ± 0.021
PCDAL	0.378 ± 0.053	0.692 ± 0.026	0.434 ± 0.023

Table 1: Ten-seed Dice budget-AUC on a 0–1 scale, mean ± sample SD. The first two rows are matched; the remaining selectors use the same learner and budget.

Dataset	Split; paid masks
BUSI	546/118/116 6/110
ISIC2018	1816/389/389 19/364
RSNA	2868/614/614 29/574

Table 2: Original five-seed train/val/test sizes (top) and paid masks at 1%/20% (bottom).

Dataset	ΔDice budget-AUC	95% paired CI	Holm p
BUSI	-0.0106	$[-0.0264, 0.0020]$	0.402
ISIC2018	-0.0089	$[-0.0180, -0.0001]$	0.305
RSNA	-0.0052	$[-0.0124, 0.0039]$	0.402

Table 3: Ten-seed audit-minus-control Dice budget-AUC effects and paired tests.

Under the unified, all-counted protocol, the ten-seed audit has lower mean Dice budget-AUC than its matched uncertainty–diversity control on all three datasets (Table 1). The paired differences range from -0.0106 to -0.0052. The ISIC2018 interval lies just below zero, while its Holm-adjusted test and the other two intervals do not meet the registered decision rule (Table 3). The matched pair shares the learner, starting masks, and budget. Its paired differences reflect the masks selected after that common starting state. Among the same-learner contextual selectors, the highest mean belongs to PCDAL on BUSI, Entropy on ISIC2018, and Core-set on RSNA. The audit is below the strongest mean on every dataset. The median within-pair selection-time ratios are 1.12, 1.14, and 1.14 on BUSI, ISIC2018, and RSNA, alongside lower mean localization AUC. Appendix A gives the complete paired decision statistics.

Dataset	5%	10%	20%
BUSI	-0.0204 $[-0.0407, -0.0024]$	-0.0107 $[-0.0339, 0.0096]$	-0.0059 $[-0.0204, 0.0081]$
ISIC2018	-0.0011 $[-0.0129, 0.0127]$	-0.0151 $[-0.0304, 0.00004]$	-0.0101 $[-0.0259, 0.0059]$
RSNA	$+0.0007$ $[-0.0115, 0.0126]$	-0.0106 $[-0.0210, 0.0007]$	-0.0047 $[-0.0150, 0.0084]$

Table 4: Ten-seed audit-minus-control test Dice at each paid-mask budget. Brackets are pointwise seed-paired bootstrap 95% intervals. Both policies share the 1% model, so their 1% difference is zero.

Both policies start from the same 1% model (Table 4). BUSI has its largest audit deficit at 5%, followed by smaller negative means at 10% and 20%. The deficits on ISIC2018 and RSNA are largest at 10%. Later acquisitions narrow some gaps, but the audit mean remains below the control at 20% on all three datasets. Only the BUSI 5% pointwise interval lies wholly below zero. The budget-AUC difference integrates negative mean effects at different stages of acquisition.

Dataset	Overlap 1–5%	Overlap 5–10%	Overlap 10–20%	Audit score–repair ρ
BUSI	0.518	0.352	0.316	-0.010 $[-0.069, 0.050]$
ISIC2018	0.517	0.212	0.282	$+0.037$ $[-0.003, 0.074]$
RSNA	0.488	0.317	0.306	$+0.076$ $[0.031, 0.127]$

Table 5: Ten-seed query overlap and audit-score/post-update evidence-Dice correlation. Brackets give seed-cluster bootstrap 95% intervals.

The policies select about half the same images in the first query batch, with less overlap in later batches (Table 5). Their first query uses the same model, so this divergence begins with the ranking rule. The different mask sets do not produce a higher aggregate localization score. The audit reranks a control-derived shortlist, so this comparison tests the spatial signal among candidates already prioritized for uncertainty and coverage. Within that shortlist, the spatial reranking does not improve the budget-AUC. Among audit-selected cases, the pre-query score has near-zero association with post-update evidence-Dice change on BUSI and ISIC2018. RSNA has a positive but small association, with a confidence interval above zero, while its acquisition budget-AUC remains below the control. A score associated with repair among selected cases thus need not identify a better batch than the matched policy. Appendix A reports the diagnostic statistics.

Dataset	CAVE	Best external
BUSI	0.448 ± 0.023	0.454 ± 0.029
ISIC2018	0.737 ± 0.012	0.744 ± 0.010
RSNA	0.431 ± 0.022	0.429 ± 0.017

Table 6: Original five-seed Dice budget-AUC, mean ± sample SD. The best external mean comes from MedCAL-ALPS, BADGE, and Entropy (tied with Margin), respectively.

Dataset	1-step ρ	Repair	No U
BUSI	$+0.065 \pm 0.225$	$+0.009 \pm 0.086$	-0.039
ISIC2018	-0.044 ± 0.184	$+0.041 \pm 0.059$	-0.006
RSNA	$+0.318 \pm 0.212$	-0.027 ± 0.051	-0.071

Table 7: Exploratory five-seed diagnostics: score–one-step-risk correlation, full-batch validation-risk reduction, and 5% Dice change without uncertainty.

Under the original five-seed dataset profiles, audit-minus-control mean Dice budget-AUC is $+0.0062$, $+0.0049$, and $+0.0010$ on BUSI, ISIC2018, and RSNA (Table 6). Each dataset has three positive seed pairs out of five, with paired intervals spanning zero (Appendix E). Across the complete benchmark, MedCAL-ALPS has the highest BUSI mean, BADGE the highest ISIC2018 mean, and the audit the highest RSNA mean. The paired effects vary across the five runs. External selectors lead on BUSI and ISIC2018.

Removed	BUSI	ISIC2018	RSNA
None	0.435	0.722	0.455
Diversity	0.431	0.731	0.450
Support	0.434	0.723	0.447
Instability	0.435	0.715	0.453
Area	0.442	0.721	0.452
Uncertainty	0.396	0.716	0.384

Table 8: Five-seed 5% Dice means from matched 1% states. Complete mean ± SD values are in Appendix B.

Dataset	Control	Audit	Paired Δ
BUSI	-0.006 (0.020)	0.009 (0.086)	0.016 (0.071)
ISIC2018	0.023 (0.039)	0.041 (0.059)	0.017 (0.028)
RSNA	-0.034 (0.047)	-0.027 (0.051)	0.006 (0.011)

Table 9: Five-seed full-batch validation-risk reduction, mean (SD). External controls are MedCAL-ALPS, BADGE, and Entropy; Δ is audit minus control.

In the matched 5% component study, removing uncertainty lowers mean Dice most on BUSI and RSNA (Table 8). Removing any one spatial term changes the mean by at most 0.008 per dataset. Removing the area term raises the BUSI mean, whereas removing instability lowers the ISIC2018 mean and removing support failure lowers the RSNA mean. The spatial terms contribute differently across the original profiles. On ISIC2018, removing feature distance raises Dice, while removing instability lowers it. Coverage and spatial stability thus shift this profile in opposite directions. Appendix B reports the complete component and query-time results.

For full-batch repair, each selector starts from the same 5% Entropy model, acquires another 1% of masks, and retrains. We compare the audit with the highest-mean external selector for each dataset in the completed benchmark. The audit has higher mean repair on all three datasets (Table 9), although the paired differences are small relative to run variation. Its mean repair is positive on BUSI and ISIC2018. Both RSNA means remain negative. Four of five paired runs favor the audit for repair on each dataset, but the exact sign test gives $p = 0.375$ (Appendix B). On RSNA, the strongest one-step score–repair association (Table 7) does not translate into a mean reduction in validation risk from a complete acquired batch.

Dataset	Control	Audit
BUSI	0.4510	0.4352
ISIC2018	0.7938	0.7844
RSNA	0.4211	0.4193

Table 10: Matched ResNet-18 five-seed Dice budget-AUC means under the original profile.

Dataset	ΔDice [95% CI]	Exact p
BUSI	-0.0158 $[-0.0460, 0.0145]$	0.1875
ISIC2018	-0.0094 $[-0.0146, -0.0043]$	0.0625
RSNA	-0.0018 $[-0.0166, 0.0130]$	0.8125

Table 11: Audit-minus-control paired Dice effects and exact sign-flip p for the same five seeds.

Changing both arms to a matched ResNet-18 learner leaves the five-seed acquisition and validation profile fixed. The audit-minus-control means are negative on all three datasets (Table 10). The ISIC2018 paired interval lies below zero (Table 11). The original five-seed profile's positive means thus do not persist when both arms use this learner. For ISIC2018, the exact five-pair test gives $p = 0.0625$, whereas the paired t interval lies below zero. Appendix D gives the matched-backbone setup and the diagnosis-AUROC budget-AUC differences (-0.0124, -0.0138, and -0.0045). Together with the ten-seed all-counted result, the matched-backbone comparison does not support a consistent acquisition gain.

5 Conclusion

Image-level labels leave the spatial evidence behind a prediction uncertain when local masks are unavailable. We introduced CAVE to audit that evidence through retained and removed views and reflection stability before selecting masks for annotation. Its configured five-seed profiles show small mean gains with uncertain paired effects, while the independent ten-seed all-counted comparison and matched ResNet-18 test show negative effects across all three datasets. Retained and removed views expose how the current model uses a predicted region before mask acquisition. In these experiments, the resulting score did not reliably select masks that improved localization over matched acquisition. Patient-disjoint cohorts with recorded annotation time would provide a direct test of scores calibrated to post-query improvement.

Reproducibility Statement

Released benchmark masks simulate all queries. No new human annotation was collected. Within each paired comparison, the selectors share splits, initial masks, learner, and annotation budgets. The original five-seed protocol uses annotated validation masks for model and profile selection. The independent ten-seed protocol selects models using validation image labels and counts every queried training mask. Appendix E reports the five-seed paired effects, and Appendix A gives the ten-seed decision statistics.

Ethics Statement

This study uses existing medical-imaging benchmarks and their released annotations for retrospective computational evaluation. Lesion masks and opacity boxes are annotation targets, not independent confirmation of a clinical diagnosis. No prospective patient intervention or clinical deployment is evaluated. Reuse and redistribution of images remain subject to the original dataset access conditions. Image-level splits and coarse targets limit claims about patient-level reliability.

AI Use Statement

Generative AI was used for English translation and language polishing, and for adjustments to manuscript formatting.

REFERENCES

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. doi: 10.1016/j.dib.2019.104863.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*, 2020. URL https://openreview.net/forum?id=ryghZJBKPS.

Soufiane Belharbi, Ismail Ben Ayed, Luke McCaffrey, and Eric Granger. Deep active learning for joint classification and segmentation with weak annotator. In *WACV*, pp. 3338–3347, 2021. doi: 10.1109/WACV48630.2021.00338. URL https://openaccess.thecvf.com/content/WACV2021/html/Belharbi_Deep_Active_Learning_for_Joint_Classification__Segmentation_With_Weak_WACV_2021_paper.html.

Arantxa Casanova, Pedro O. Pinheiro, Negar Rostamzadeh, and Christopher J. Pal. Reinforced active learning for image segmentation. In *ICLR*, 2020. URL https://openreview.net/forum?id=SkgC6TNFvr.

Liangyu Chen, Yutong Bai, Siyu Huang, Yongyi Lu, Bihan Wen, Alan Yuille, and Zongwei Zhou. Making your first choice: To address cold start problem in medical active learning. In *MIDL*, volume 227 of *Proceedings of Machine Learning Research*, pp. 496–525, 2024. URL https://proceedings.mlr.press/v227/chen24a.html.

Noel C. F. 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.

Bernhard Föllmer, Kenrick Schulze, Christian Wald, Sebastian Stober, Wojciech Samek, and Marc Dewey. Active learning with the nnUNet and sample selection with uncertainty-aware submodular mutual information measure. In *MIDL*, volume 250 of *Proceedings of Machine Learning Research*, pp. 480–503, 2024. URL https://proceedings.mlr.press/v250/follmer24a.html.

Ruth Fong, Mandela Patrick, and Andrea Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *ICCV*, pp. 2950–2958, 2019. doi: 10.1109/ICCV.2019.00304. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Fong_Understanding_Deep_Networks_via_Extremal_Perturbations_and_Smooth_Masks_ICCV_2019_paper.html.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, pp. 3429–3437, 2017. doi: 10.1109/ICCV.2017.371. URL https://openaccess.thecvf.com/content_iccv_2017/html/Fong_Interpretable_Explanations_of_ICCV_2017_paper.html.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *NeurIPS*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

Zhifan Jiang, Vishwesh Nath, Holger R. Roth, Abhijeet Parida, Nicholas Foreman, Michael J. Fisher, Roger J. Packer, Syed Muhammad Anwar, Robert A. Avery, and Marius George Linguraru. KALM: Knowledge-driven active learning for medical image segmentation using localized similarity. In *IEEE ISBI*, pp. 1–5. IEEE, 2025. doi: 10.1109/ISBI60581.2025.10981164.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. Learning active learning from data. In *NeurIPS*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8ca8da41fe1ebc8d3ca31dc14f5fc56c-Abstract.html.

Wenyuan Li, Jiayun Li, Zichen Wang, Jennifer Polson, Anthony E. Sisk, Dipti P. Sajed, William Speier, and Corey W. Arnold. PathAL: An active learning framework for histopathology image analysis. *IEEE Trans. Med. Imaging*, 41(5):1176–1187, 2022. doi: 10.1109/TMI.2021.3135002.

Dwarikanath Mahapatra, Behzad Bozorgtabar, Sudipta Roy, Imran Razzak, and Mauricio Reyes. VALIANT: Prompt instability for active learning in black-box medical imaging. In *AAAI*, volume 40, pp. 7901–7909, 2026. doi: 10.1609/aaai.v40i10.37734.

Radiological Society of North America. RSNA pneumonia detection challenge. Kaggle competition, 2018. URL https://www.kaggle.com/c/rsna-pneumonia-detection-challenge.

Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D. Tizabi, Nicholas Schreck, A. Emre Kavur, Bünyamin Pekdemir, Tobias Roß, Annette Kopp-Schneider, and Lena Maier-Hein. Labelling instructions matter in biomedical image analysis. *Nat. Mach. Intell.*, 5:273–283, 2023. doi: 10.1038/s42256-023-00625-5.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *IJCAI*, pp. 2662–2670, 2017. doi: 10.24963/ijcai.2017/371.

Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.*, 2:476–486, 2020. doi: 10.1038/s42256-020-0212-3.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A Core-Set approach. In *ICLR*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Burr Settles. From theories to queries: Active learning in practice. In *Active Learning and Experimental Design Workshop in Conjunction with AISTATS 2010*, volume 16 of *Proceedings of Machine Learning Research*, pp. 1–18, 2011. URL https://proceedings.mlr.press/v16/settles11a.html.

Jun Shi, Shulan Ruan, Ziqi Zhu, Minfan Zhao, Hong An, Xudong Xue, and Bing Yan. Predictive accuracy-based active learning for medical image segmentation. In *IJCAI*, pp. 4885–4893, 2024. doi: 10.24963/ijcai.2024/540. URL https://www.ijcai.org/proceedings/2024/540.

Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In *AAAI/ACM AIES*, pp. 239–245, 2019. doi: 10.1145/3306618.3314293.

Ifrat Ikhtear Uddin, Longwei Wang, Xiao Qin, Yang Zhou, and K. C. Santosh. Learning to select like humans: Explainable active learning for medical imaging. In *IEEE CAI*, pp. 458–463. IEEE, 2026. doi: 10.1109/CAI68641.2026.11536249.

Sudheendra Vijayanarasimhan and Kristen Grauman. Multi-level active prediction of useful image annotations for recognition. In *NeurIPS*, volume 21, pp. 1715–1722, 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/hash/4e0928de075538c593fbdabb0c5ef2c3-Abstract.html.

Longwei Wang, Ifrat Ikhtear Uddin, and K. C. Santosh. Expert-guided explainable few-shot learning with active sample selection for medical image analysis. *IEEE J. Biomed. Health Inform.*, pp. 1–14, 2026. doi: 10.1109/JBHI.2025.3650334. Early access.

Shanshan Wang, Cheng Li, Rongpin Wang, Zaiyi Liu, Meiyun Wang, Hongna Tan, Yaping Wu, Xinfeng Liu, Hui Sun, Rui Yang, Xin Liu, Jie Chen, Huihui Zhou, Ismail Ben Ayed, and Hairong Zheng. Annotation-efficient deep learning for automatic medical image segmentation. *Nat. Commun.*, 12:5915, 2021. doi: 10.1038/s41467-021-26216-9.

Tao Wang, Xinlin Zhang, Yuanbo Zhou, Yuanbin Chen, Longxuan Zhao, Tao Tan, and Tong Tong. PCDAL: A perturbation consistency-driven active learning approach for medical image segmentation and classification. *IEEE Trans. Emerg. Top. Comput. Intell.*, 9(4):3162–3177, 2025. doi: 10.1109/TETCI.2025.3547635.

Ofer Yehuda, Avihu Dekel, Guy Hacohen, and Daphna Weinshall. Active learning through a covering lens. In *NeurIPS*, volume 35, pp. 22354–22367, 2022. doi: 10.52202/068431-1624. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/8c64bc3f7796d31caa7c3e6b969bf7da-Abstract-Conference.html.

Taimin Zhou, Jin Yang, Lingguo Cui, Nan Zhang, and Senchun Chai. SBC-AL: Structure and boundary consistency-based active learning for medical image segmentation. In *MICCAI*, volume 15012, pp. 283–293. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72390-2_27.

Ning Zhu, Xiaochuan Ma, Shaoting Zhang, and Guotai Wang. MedCAL-Bench: A comprehensive benchmark on cold-start active learning with foundation models for medical image analysis. arXiv preprint arXiv:2508.03441, 2025. URL https://arxiv.org/abs/2508.03441.

A Unified comparison details

The independent comparison uses ten paired runs. Within each dataset and run, the audit and its matched control share the split, initial queried masks, initial model, cumulative budgets, and training procedure. At later budgets, training continues with all acquired masks. The audit reranks a class-balanced shortlist using support failure and reflection instability. The control uses entropy and feature distance. Every queried mask, including an empty one, counts toward the budget.

For each dataset, the primary endpoint is the normalized trapezoidal area under held-out nonempty-mask Dice across the four cumulative budgets. The paired interval uses a seed-level bootstrap. A two-sided exact sign-flip test with Holm adjustment covers the three dataset-level primary tests. The prespecified dataset gate additionally requires a mean paired gain of at least 0.005, a positive Dice interval lower bound, at least seven positive seed pairs, and a diagnosis-AUROC lower bound above -0.005. No dataset passes. Table 3 gives the complete paired decision statistics.

Dataset	ΔDice [95% CI]	+/10	Holm p	AUROC LCB
BUSI	-0.0106 $[-0.0264, 0.0020]$	3	0.402	-0.0179
ISIC2018	-0.0089 $[-0.0180, -0.0001]$	3	0.305	-0.0046
RSNA	-0.0052 $[-0.0124, 0.0039]$	1	0.402	-0.0030

Table 12: Paired decision statistics for audit minus control. LCB is the one-sided 95% lower bound of the AUROC budget-AUC change.

B Original five-seed repair and component diagnostics

B.1 Repair after annotation

From each model at the 5% Entropy budget, the audit scores the same 24 candidates per run. After their masks are revealed, a one-step evidence-loss update measures the validation-risk response. Spearman correlation measures score alignment with the resulting validation risk change, and top-quartile gain compares the highest-scored candidates with the same candidate pool. For full-batch repair, each selector reveals an additional 1% batch and retrains from a common starting point with all acquired masks. The 5% model supplies the query state and reference risk.

Score–response correlations across five seeds are $+0.065 \pm 0.225$, -0.044 ± 0.184, and $+0.318 \pm 0.212$ on BUSI, ISIC2018, and RSNA. Selecting the top score quartile changes the mean one-step response relative to the candidate pool by $+0.002 \pm 0.004$, -0.002 ± 0.005, and $+0.018 \pm 0.024$, respectively. The largest association and top-quartile difference occur on RSNA.

After retraining with the acquired batch, we compare the audit with the highest-mean external selector for each dataset in Table 14. Entropy is used for the RSNA tie. Mean repair is $+0.009$, $+0.041$, and -0.027 on BUSI, ISIC2018, and RSNA. The paired differences are $+0.016 \pm 0.071$, $+0.017 \pm 0.028$, and $+0.006 \pm 0.011$, with four positive pairs out of five on each dataset. The two-sided exact sign-test is $p = 0.375$ for each comparison.

B.2 Component and query-time results

Table 13: Five-percent Dice from matched 1% states, reported as mean ± sample standard deviation over five seeds. Support and Area denote support failure and area inconsistency.

Removed term	BUSI	ISIC2018	RSNA
None	0.435 ± 0.032	0.722 ± 0.020	0.455 ± 0.019
Diversity	0.431 ± 0.033	0.731 ± 0.018	0.450 ± 0.017
Support	0.434 ± 0.018	0.723 ± 0.021	0.447 ± 0.012
Instability	0.435 ± 0.024	0.715 ± 0.021	0.453 ± 0.012
Area term	0.442 ± 0.027	0.721 ± 0.020	0.452 ± 0.017
Uncertainty	0.396 ± 0.066	0.716 ± 0.018	0.384 ± 0.026

Selected-to-pool mean predicted-area ratios are 1.22, 2.72, and 0.79 on BUSI, ISIC2018, and RSNA. Correlations between the audit score and the revealed area gap are 0.01, 0.37, and -0.14, respectively. HMD assigns 70% of its RSNA weight to uncertainty. Under the shared 1% state in Table 13, removing uncertainty changes 5% Dice by -0.039, -0.006, and -0.071 on BUSI, ISIC2018, and RSNA, respectively. Removing feature distance changes Dice by -0.004, $+0.009$, and -0.005. Other terms stay within 0.008 of the full profile. Support failure has its largest effect on RSNA, instability on ISIC2018, and removing area inconsistency slightly raises the BUSI mean.

Replacing mean filling with blur or zero yields score correlations of 0.821–1.000 and selected-set overlaps of 0.703–0.970. At the 1%–5% transition, query scoring takes 2.27, 4.58, and 4.09 seconds, compared with 0.73, 1.36, and 1.25 seconds for Entropy. The 3.1–3.4$\times$ overhead excludes the shared training update.

C ORIGINAL FIVE-SEED BENCHMARK COMPARISON

Table 14: Original five-seed dataset-profile comparison: nonempty-mask Dice budget-AUC over four annotation budgets, reported as mean ± sample standard deviation across five random seeds. The paired row reports audit minus control, with wins in parentheses. Criterion-only adapters use the shared learner. Bold marks the highest mean within each dataset.

Method	BUSI	ISIC2018	RSNA
General-purpose selectors			
Random	0.435±0.035	0.736±0.009	0.420±0.021
Entropy (2011)	0.435±0.030	0.733±0.015	0.429±0.017
Margin (2011)	0.438±0.025	0.733±0.015	0.429±0.017
Core-set (2018)	0.443±0.028	0.741±0.013	0.408±0.026
BADGE (2020)	0.414±0.030	**0.744±0.010**	0.421±0.020
ProbCover (2022)	0.435±0.035	0.736±0.015	0.415±0.022
Criterion-only medical adapters			
PAAL (2024)	0.414±0.034	0.705±0.011	0.379±0.027
SBC-AL (2024)	0.381±0.028	0.656±0.014	0.421±0.018
MedCAL-ALPS (2025)	**0.454±0.029**	0.739±0.013	0.414±0.018
KALM (2025)	0.424±0.030	0.699±0.018	0.419±0.019
EG-AL (2026)	0.421±0.029	0.700±0.016	0.424±0.022
VALIANT (2026)	0.431±0.023	0.715±0.015	0.425±0.026
PCDAL (2025)	0.436±0.027	0.724±0.019	0.400±0.024
CAVE	0.448±0.023	0.737±0.012	**0.431±0.022**
Audit isolation, ours minus control			
Scratch Δ	+0.006±0.012 (3/5)	+0.005±0.009 (3/5)	+0.001±0.003 (3/5)

D MATCHED-BACKBONE ROBUSTNESS

The ResNet-18 experiment changes the learner in both arms while retaining the original dataset-profile acquisition, annotated-validation model-selection rule, five seeds, and four cumulative budgets. The paired scratch initialization, split, mask budget, and training schedule match within each dataset and seed. Tables 10 and 11 report the means and paired outcomes. The two-sided exact sign-flip p values are 0.1875, 0.0625, and 0.8125 for BUSI, ISIC2018, and RSNA. The ISIC2018 t interval excludes zero. Diagnosis-AUROC budget-AUC differences are -0.0124, -0.0138, and -0.0045.

E PAIRED EFFECTS ACROSS THE FIVE FIXED SEEDS

For within-run differences δ_s, the interval is $\bar{\delta} \pm t_{0.975,4} s_\delta/\sqrt{5}$, with $t_{0.975,4} = 2.7764$ and sample standard deviation s_δ. The interval summarizes paired variation across the five fixed runs. Table 15 gives the five paired effects behind the rounded aggregate in the main text.

F BUDGET-AUC DEFINITION

For budgets $b_1 < \cdots < b_M$, normalized budget-AUC is

$$\frac{1}{b_M - b_1} \sum_{k=1}^{M-1} \frac{D_k + D_{k+1}}{2} (b_{k+1} - b_k), \tag{11}$$

where D_k is test Dice at budget b_k. The same budget levels enter every method, and comparisons are paired within each seed.

Table 15: Audit-minus-control Dice budget-AUC for the fixed-five cohort. Values are differences on the 0–1 Dice scale.

Run	BUSI	ISIC2018	RSNA
1	-0.00294	+0.00268	+0.00182
2	+0.01852	-0.00106	-0.00256
3	+0.00533	-0.00322	-0.00016
4	+0.01755	+0.00557	+0.00458
5	-0.00756	+0.02041	+0.00152
Mean	+0.00618	+0.00488	+0.00104
Sample SD	0.01177	0.00932	0.00263
95% interval	[-0.0084,0.0208]	[-0.0067,0.0164]	[-0.0022,0.0043]

G ACQUISITION PROFILES

Equation 5 normalizes the seven score components within each round. MPD, EDD, and HMD denote margin-predominant diversity, evidence-defect diversity, and hybrid margin diversity. The weights in Table 16 specify $\alpha_j^{(\pi)}$, $w_A^{(\pi)}$, and the known-empty attenuation $q_i^{(\pi)}$ used in Equation 6. Components omitted from a profile have zero weight. All listed component weights sum to one. For other classes, $q_i^{(\pi)} = 1$. Nonempty-mask validation Dice selects one profile per dataset, held fixed across seeds and budgets.

Table 16: Preset acquisition profiles. The first five rows form the spatial audit, and margin and entropy measure class ambiguity. Dashes denote zero weights. The area-fit tolerance ν is 10 for MPD and 8 for HMD. The empty factor applies only when the class label implies an empty mask.

Signal	MPD	EDD	HMD
Support failure o_i	0.05	0.20	0.07
Instability s_i	0.03	0.15	0.05
Area inconsistency c_i^{area}	0.07	0.25	0.08
Area fit r_i^{area}	0.05	–	0.10
Area a_i	–	0.10	–
Margin m_i	0.72	–	0.58
Entropy u_i	0.08	0.30	0.12
Candidate weight $w_A^{(\pi)}$	0.95	0.85	0.90
Empty factor $q_i^{(\pi)}$	0.97	1.00	0.91