Q-SAFE: COMPILING BOUNDED-LOSS CONTRACTS FOR RISK-CONTROLLED POLICY REPLACEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive imaging can fail before learning begins: a nonuniform schedule may have no equal-cost oracle advantage, its objective may not match the scientific query, or its proxy may not align with a downstream human endpoint. We organize these failures into a diagnose–then–replace protocol and present Q-Safe as a model-agnostic compiler for the remaining decision: whether a proposed action may replace an executable, budget-matched fallback. It compiles declared bounded losses into an auditable replacement contract and emits certificates for schedule headroom, calibration, and deployment risk; the proposal itself may come from a neural model or a fixed rule. An initial frozen binary-harm contract improves ConnectomeBench2 utility over ExtraTrees by 1.16 points (paired 95% CI: 0.81–1.51), although the gain is mouse-driven. We then freeze a severity-aware fallback-regret contract before opening four additional shards. All pass the 2% endpoint; pooled regret is 0.191%, and relative to the frozen incidence contract, coverage and utility rise by 13.99 and 3.30 points. Conversely, on 36 registered EMDiffuse regions from three tissues, an equal-cost tile oracle fails to beat uniform acquisition in all 18 tested conditions for one fixed preview-plus-complete-rescan schedule. The contribution is therefore an auditable replacement interface, not a claim of architectural superiority or universal benefit from learned scanning.

1 INTRODUCTION

Scientific microscopes do not acquire information for free. Photon exposure, electron dose, detector readout, and pixel dwell time trade image quality against sample damage and throughput. Deep restoration can recover useful images from substantially fewer measurements (Weigert et al., 2018; Lu et al., 2024), while adaptive microscopes can use a low-cost survey to trigger localized high-resolution acquisition (Mahecic et al., 2022; Alvelid et al., 2022). In scanning electron microscopy even the choice among dense low-dwell, low-resolution, and sparse acquisition has been studied under a fixed time budget (Trampert et al., 2018). These advances make a new operational question unavoidable. Most directly, SmartEM performs a rapid full-field scan followed by learned slow rescans, optimizes preview dwell, rescan dwell, and rescan fraction under an effective-dwell constraint, and deploys the resulting pipeline on a commercial SEM (Meirovitch et al., 2026). We therefore do not claim preview–rescan acquisition itself as a contribution. We ask the next operational question: given the same total budget, *when should a learned, spatially nonuniform policy replace a known fallback?*

This question is not answered by showing that a score ranks difficult regions or that a restoration model improves a global image metric. We distinguish three failures that require different remedies. *Schedule infeasibility* means that, after paying for a preview, even an oracle nonuniform allocation is worse than equal-cost uniform acquisition; no better ranker can repair it. *Objective mismatch* means that the allocation score optimizes the wrong scientific quantity: pixel error, membrane continuity, mitochondrial visibility, and proofreading utility need not induce the same ordering. *Proxy mismatch* means that a nominally target-specific score still fails to preserve the downstream endpoint, for example human interpretability. Learned reconstruction is an ill-posed inverse map whose failures

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.

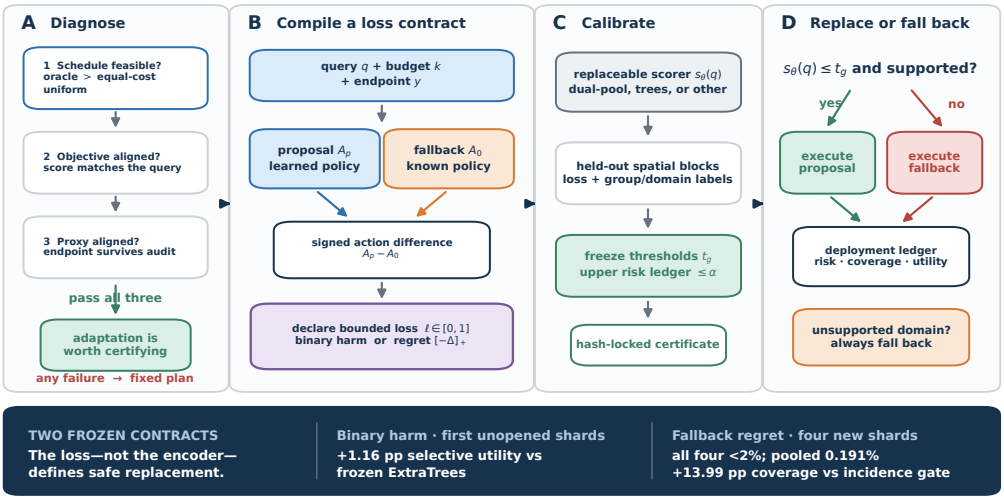

Figure 1: Q-Safe is a loss-contract compiler for policy replacement. It first diagnoses schedule, objective, and proxy failures; then compiles a proposal, equal-cost fallback, and declared bounded loss into a block-calibrated certificate. The frozen gate executes the proposal only when supported and otherwise returns the fallback. Two prospectively frozen losses illustrate that the contract, rather than the encoder brand, changes deployment.

can be structured or unstable (Belthangady and Royer, 2019; Antun et al., 2020); high PSNR or uncertainty alone therefore resolves none of these three questions.

Existing literatures each expose only part of this decision. Active and smart microscopy decide where or when to measure, but generally deploy the learned trigger directly (Mahecic et al., 2022; Alvelid et al., 2022). Selective prediction trades error against coverage by rejecting uncertain predictions (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2019; Gangrade et al., 2021), and risk-control methods calibrate prediction sets or ranked retrieval on held-out data (Bates et al., 2021; Angelopoulos et al., 2024; Xu et al., 2025). Our setting is different: rejection must still produce an image or proofreading queue, so it selects a conservative, *budget-matched action* rather than returning no answer. Safety consequently depends on the counterfactual difference between two executable policies, not on the confidence of either policy in isolation.

We formulate the final decision as *risk-controlled policy replacement* and introduce Q-Safe. Its central object is not a particular encoder but a certificate-producing compiler around two executable policies. Given a scientific query, budget, proposal, and fallback, it (i) rejects schedule classes with no equal-cost oracle headroom, (ii) records the signed action difference and the endpoint used to define harm, (iii) calibrates any scalar proposal score on spatial blocks, and (iv) emits both block- and query-marginal deployment ledgers. A dual-pooling network is one scorer instantiation. Q-Safe routes unsupported species or domains to the fallback. The resulting workflow is *diagnose feasibility, validate the objective and proxy, then calibrate replacement*, rather than train an adaptive controller and audit it afterward.

The empirical conclusion is deliberately narrower than "learned scanning is better." On four frozen connectomic proofreading shards (Brown et al., 2026), a binary-harm contract improves utility over a handcrafted gate but is mouse-driven. After those data became open, we declared fallback regret, froze a second protocol, and opened four new shards; all pass, and the additional utility is fly-driven. Conversely, on registered multi-dwell micrographs (Lu et al., 2024), an equal-cost oracle fails across three tissues. The interface therefore supports both prospective policy replacement and a principled decision not to adapt.

Our contributions are:

- a model-agnostic policy-replacement interface that compiles declared bounded losses into an executable proposal-or-fallback decision and explicit certificate, rather than returning a confidence score or abstention;
- a staged compiler that separates schedule infeasibility, objective and proxy mismatch, and calibrated replacement risk, with callable checks for oracle headroom, signed action differences, and two deployment estimands; and
- two hash-frozen multi-species confirmations of distinct contracts— binary harm and severity-aware fallback regret—plus negative external tests delimiting when the interface should refuse adaptation. The elementary action identities are implementation checks, not a standalone theory claim.

2 Related work

Permutation-invariant evidence. Deep Sets and Set Transformers provide general encoders for unordered inputs (Zaheer et al., 2017; Lee et al., 2019). Q-Safe uses the same invariance but imposes decision structure: it pools the executed and fallback top-k sets separately and explicitly presents their difference to the harm predictor. The comparison is between two actions, not between a query and an unstructured neighborhood.

Selective prediction, robust learning, and risk control. Selective classification formalizes the risk–coverage tradeoff through a reject option (El-Yaniv and Wiener, 2010); SelectiveNet learns that option jointly with a predictor (Geifman and El-Yaniv, 2019), while one-sided and post-hoc variants attach abstention to an existing score (Gangrade et al., 2021). Q-Safe instead selects between two budget-matched *actions*: abstention invokes a query-relevant fallback rather than returning no prediction. GroupDRO targets worst-group training loss (Sagawa et al., 2020), while domain-generalization studies show that model selection and regularization can matter as much as a new objective (Gulrajani and Lopez-Paz, 2021). Risk- controlling prediction sets, conformal risk control, and two-stage ranked- retrieval control calibrate bounded losses on held-out data (Bates et al., 2021; Angelopoulos et al., 2024; Xu et al., 2025). Our claim is more modest: a finite-sample-corrected empirical harmful-action rate over pre-declared spatial blocks, with thresholds intersected across observed domains. We do not claim unknown-domain or conditional coverage. Concurrent policy post-processing calibrates switching from an existing policy to a fallback (Joshi et al., 2026), while OPAL authorizes adaptation at experiment or compound level (Bi et al., 2026). Q-Safe instead compares two query-compiled, equal-budget imaging actions; these are adjacent guarantees, not direct imaging baselines.

Adaptive scientific imaging. Active acquisition methods learn where to measure next (Mahecic et al., 2022; Alvelid et al., 2022), and biological image restoration often optimizes perceptual or pixel metrics (Weigert et al., 2018; Lu et al., 2024). Sequential dose-constrained acquisition has also been studied with deep RL for electron ptychography (Schloz et al., 2023). Realistic simulators such as pySTED expose photobleaching, point-scanning dynamics, and sequential control (Bilodeau et al., 2024), while STEMGym separates perception, navigation, and planning under a unified electron-dose budget (Polat et al., 2026). These are prospective policy environments rather than direct baselines for our fixed compiled-action benchmark. Extending Q-Safe to them would require a state-conditional fallback and trajectory-level replacement loss, not merely a one-shot top-k comparison. SmartEM is the closest acquisition system: it predicts membrane-segmentation errors from a fast scan, locally rescans at longer dwell, and processes the resulting mixed-dwell image (Meirovitch et al., 2026). Q-Safe can treat such a rescan map as the *proposed action*; it neither replaces SmartEM's error detector nor claims its acquisition pattern. Our focus is the supervisory interface between such predictors and a scientific decision: the proposed action must be relevant to the current query, compared with an explicit equal-cost fallback, and rejected when calibrated evidence is insufficient or when an oracle feasibility test shows no headroom. Recent counterfactual prediction sets couple uncertainty coverage to the policy induced by that uncertainty (Zheng and Jin, 2026), whereas active sensing optimizes where to acquire information. Q-Safe occupies the interface between them: it compiles a proposed acquisition and a conservative acquisition into directly comparable evidence, then calibrates whether the proposal may replace the fallback. EMDiffuse is particularly relevant on the imaging side because it couples dwell-time EM restoration with stochas-

tic self-assessment (Lu et al., 2024); unlike restoration uncertainty, our endpoint is the downstream loss of choosing one acquisition action over another.

3 PROBLEM FORMULATION

Let q encode a scientific target, spatial support, and action budget. Each candidate $i \in \mathcal{N}(q)$ has an observable visual score p_i, query relevance $r_i(q) \geq 0$, metadata m_i, and measurement cost c_i. The proposed action A_p selects the largest $p_i r_i(q)$ under budget k; the conservative fallback A_0 selects the largest $r_i(q)$. Their realized utility difference is

$$\Delta(q,k) = U(A_p(q,k);q) - U(A_0(q,k);q), \quad A_p = \mathrm{TopK}(p \odot r, k), \quad A_0 = \mathrm{TopK}(r,k). \tag{1}$$

Q-Safe produces an acceptance decision $a(q,k) \in \{0,1\}$; the deployed policy is A_p if accepted and A_0 otherwise. We evaluate the harmful-action loss

$$\ell(q,k) = \mathbb{1}\{a(q,k)=1\}\mathbb{1}\{\Delta(q,k) < 0\} \tag{2}$$

The primary estimand samples a spatial block uniformly and then averages its queries:

$$R_{\mathrm{block}} = |\mathcal{B}|^{-1} \sum_{b \in \mathcal{B}} |b|^{-1} \sum_{q \in b} \ell(q,k). \tag{3}$$

Block-marginal coverage and utility use the same weighting. We additionally report query-marginal metrics, which weight every query equally, and the conditional harm $\Pr(\Delta < 0 \mid a = 1)$ and $\mathbb{E}[\Delta \mid a = 1]$ are secondary. The fallback defines zero *relative replacement loss*; this is not an absolute biological-safety claim. Queries are grouped into spatial blocks $b \in \mathcal{B}$ before calibration, preventing dense regions from dominating the empirical guarantee.

For the acquisition experiments, we distinguish the unavailable downstream loss y_q from an observable proxy $\widetilde{y}_q$. A score can fail in two ways. *Objective mismatch* uses a proxy for the wrong query, for example a membrane score for a mitochondrial request. *Proxy mismatch* occurs when the named proxy targets the right structure but its magnitude does not preserve the ordering of y_q. Q-Safe controls action risk only for the outcome on which it is calibrated; it cannot convert an invalid $\widetilde{y}_q$ into biological ground truth. We therefore treat independent human scoring as a falsification test rather than as another in-domain metric.

3.1 DIAGNOSTIC HIERARCHY

Figure 1 fixes the order. If an equal-cost oracle cannot beat uniform acquisition, the schedule is rejected before training. If the objective does not represent the query, it is replaced; if its proxy fails downstream audit, no biological deployment claim is made. Only a schedule and endpoint that survive these checks reach Q-Safe, where failed or unsupported replacement executes the budget-matched fallback. A later stage cannot repair an earlier failure.

4 METHOD

4.1 A CERTIFICATE-PRODUCING POLICY COMPILER

The public interface accepts a proposed action, an equal-budget reference, a scalar replacement score from any source, spatial groups, and a declared loss endpoint. It returns four machine-readable records: (1) a fixed-schedule oracle-headroom certificate, (2) a signed proposed-minus-reference action difference, (3) a calibration protocol containing thresholds, sampling unit, and unsupported groups, and (4) a deployment audit containing query-marginal, block-marginal, and conditional outcomes. This separation is intentional: the same calibration and audit code can wrap a neural scorer, a tree, a heuristic, or a human-authored policy. The compiler certifies only the declared schedule, endpoint, and observed calibration domains; it does not certify biological safety or distribution shift in general.

4.2 QUERY-COMPILED EVIDENCE TOKENS

The compiler maps $(q, \mathcal{N}(q), k)$ to a set $X_q = \{x_i\}_{i=1}^{n_q}$, where

$$x_i = [p_i,\ p_i(1-p_i),\ m_i,\ r_i,\ p_i r_i,\ d_i^p,\ d_i^0,\ \mathbb{1}(i \in A_p),\ \mathbb{1}(i \in A_0)]. \tag{4}$$

d_i^p and d_i^0 are signed margins to the respective budget boundaries. No ground-truth candidate label is an input. A global vector g_q contains only query scale, normalized budget, support size, and coarse observable summaries. Thus the network can inspect all local evidence without depending on candidate order or a fixed neighborhood size.

4.3 COUNTERFACTUAL SET ENCODER

An MLP embeds each candidate, $e_i = \phi(x_i)$. A query seed $z_q = \psi(g_q)$ attends separately to the proposed and fallback sets:

$$h_p = \mathrm{Attn}(z_q, \{e_i : i \in A_p\}), \qquad h_0 = \mathrm{Attn}(z_q, \{e_i : i \in A_0\}). \tag{5}$$

We concatenate $h_p, h_0, h_p - h_0$, a mean–max DeepSets summary of all e_i, and g_q. Two heads predict harm probability $s_\theta(q,k)$ and utility $\widehat{\Delta}_\theta(q,k)$. The explicit action difference exposes the comparison directly; we test rather than assume whether this inductive bias is useful. Let $y_i^h = \mathbb{1}\{\Delta_i < 0\}$. With species groups c, training uses

$$\begin{aligned} L_c &= \mathbb{E}_{i:c_i=c}\Big[\mathrm{BCE}_{c_+}(s_i, y_i^h) + 0.15\,\mathrm{SmoothL1}(\widehat{\Delta}_i, \Delta_i)\Big], \\ \min_\theta \mathcal{L}_B(\theta) &= \sum_{c \in C_B} \widetilde{w}_c L_c, \\ w_c &\leftarrow \frac{w_c \exp(\eta L_c)}{\sum_{c'} w_{c'} \exp(\eta L_{c'})}. \end{aligned} \tag{6}$$

where $c_+ = \min\{10, \sqrt{(1-\pi_h)/\pi_h}\}$, $\eta = 0.02$, and raw fractional Δ is not rescaled. Here C_B contains species present in minibatch B and $\widetilde{w}_c = w_c / \sum_{c' \in C_B} w_{c'}$. The fixed global training harm prevalence defines π_h. We initialize $w_c = 1/4$, update weights from the pre-parameter-update minibatch losses, and then take the gradient step. Appendix D gives the optimizer and selection schedule. GroupDRO is an optimization heuristic, not the source of the deployed risk statement.

4.4 AUDITABLE ACTION COMPARISON

The explicit difference has a simple implementation check. If $U(A;q) = C(q) + \sum_{i \in A} v_i(q)$, shared candidates cancel exactly:

$$\Delta(q,k) = \sum_{i \in A_p \setminus A_0} v_i(q) - \sum_{i \in A_0 \setminus A_p} v_i(q). \tag{7}$$

Thus the signed symmetric difference is sufficient in the additive regime and motivates presenting $h_p - h_0$ explicitly; the full-support pool permits departures from that regime. Appendix C gives the elementary proofs, support and cost assumptions, and a secondary cycle-consistent multi-policy extension. We use these results as auditable constraints rather than as theoretical novelty.

4.5 HIERARCHICAL BOUNDED-LOSS CONTRACTS

The neural score orders queries but does not itself provide a guarantee. A user declares nonnegative losses $L_j(q) \in [0, M_j]$ and targets α_j; rejection executes the fallback and incurs zero *relative* loss. Examples are harm incidence $L_h = \mathbb{1}\{\Delta < 0\}$ and fallback regret $L_r = \max\{-\Delta, 0\}$. Let $S_t = \{q : s_\theta(q,k) \le t\}$ and $\bar{\ell}_{b,j}(t)$ be the block mean of $\mathbb{1}\{q \in S_t\} L_j(q)$. For species c and calibration domain d, we choose the largest observed-score threshold satisfying every declared contract,

$$\widehat{R}^+_{c,d,j}(t) = \frac{\sum_{b \in \mathcal{B}_{c,d}} \bar{\ell}_{b,j}(t) + M_j}{|\mathcal{B}_{c,d}| + 1} \le \alpha_j \quad \text{for every } j. \tag{8}$$

For exchangeable calibration and future blocks and this nested threshold family, the add-one bounded-loss correction controls each expected block- marginal loss (Angelopoulos et al., 2024); selecting a threshold no larger than each single-loss threshold preserves every monotone constraint. This is not a high-probability or conditional-on-acceptance bound. The frozen confirmation declares only L_h at $\alpha_h = 0.02$; severity contracts are post-confirmation diagnostics. We deploy $t_c = \min_d t_{c,d}$ and unsupported groups always fall back. Blind novel-domain performance is empirical, not guaranteed.

4.6 Evidence and schedule checks

Let $S(q) = \{i : r_i(q) > 0\}$. With deterministic pre-acquisition tie breaking and $k \leq |S(q)|$, candidates outside $S(q)$ cannot enter either ranking; acquiring visual evidence only on $S(q)$ therefore preserves both actions and the frozen gate. This is a view-count equivalence, not a hardware timing claim.

Before training, we also test oracle headroom for a narrowly defined schedule: n independent equal-cost regions, an additive loss, a mandatory full preview at dwell t_0, and complete t_1 rescans of exactly $m = \rho n$ regions. Selected-region loss $L_i(t_1; q)$ is measured on the complete rescan image, which replaces rather than fuses with the preview; preview cost is nevertheless paid. The equal-cost uniform dwell is $B = t_0 + \rho t_1$. With marginal value $g_i = L_i(t_0; q) - L_i(t_1; q)$ and uniform benefit $u_i = L_i(t_0; q) - L_i(B; q)$, adaptation is feasible exactly when

$$n^{-1} \sum_{i \in \mathrm{TopM}(g)} g_i > n^{-1} \sum_{i=1}^{n} u_i. \tag{9}$$

If this top-m oracle fails, no learned ranking can repair this schedule. Arbitrary top-m selection upper-bounds contiguous selection, and adding nonnegative switching overhead can only make this result less favorable. Heterogeneous dwell choices, variable ROI costs, cross-tile interactions, and adaptive stopping define different policy classes and are not ruled out.

5 Experiments

5.1 Primary task and protocol

We use the official ConnectomeBench2 v2 false-split task (Brown et al., 2026), which asks whether a candidate segment association is false across fly, human, mouse, and zebrafish connectomes. A frozen shared-view ResNet supplies p_i. A local proofreading query defines relevance by geometric and graph proximity; utility is the fraction of query-relevant false associations recovered at a fixed action budget. The fallback spends the same budget using relevance alone.

Train shards 0–3 fit the visual and risk models. Validation 0–1 select the set architecture; validation 2–3 calibrate thresholds; validation 4 is an external check. Test 13–24 are explicitly marked as development reuse. Test 25–28 remained absent from disk until the model, thresholds, comparator, evaluation script, and expected data hashes were frozen. The blind result is produced by a single command that first verifies those hashes. Spatial blocks, not individual queries, are the unit for calibration and bootstrap resampling.

Baselines and selection metric. We compare summary MLP, DeepSets, generic attention, counterfactual ERM/GroupDRO, and a frozen ExtraTrees gate. Three-seed architecture selection uses worst-domain AUROC; no blind shard selects a model or seed. The 66,914-parameter counterfactual GroupDRO model attains harm AP 0.249 and is the only counterfactual variant to pass every reused-development check (worst risk 1.58%, coverage 58.39%, deployed utility 11.69 points); ERM narrowly fails one domain at 2.04%. We therefore freeze GroupDRO even though ERM has slightly higher AUROC. Full selection results and hyperparameters are in Appendix D.

Capacity-matched component ablation. Post-confirmation ablations zero each representation while preserving 66,914 parameters and never reopen blind shards. Explicit difference and fallback evidence improve harm AP and worst-development risk, but deployed utility moves by at most

Table 1: Post-confirmation capacity-matched ablation. Every row is separately calibrated on validation 2–3 at 2% before reused test 13–24; blind shards stay closed. Harm is worst-shard unconditional incidence; Cond. is pooled conditional harm among accepted queries.

Representation	AP	Harm (%)	Cond. (%)	Coverage (%)	Utility (pp)
Full	**0.249**	**1.579**	**0.659**	58.39	**11.692**
No explicit difference	0.224	1.835	0.722	59.51	11.677
No fallback pool	0.232	1.912	0.889	**60.25**	11.671
No full-support pool	0.240	1.707	0.791	57.82	11.664

Table 2: Frozen blind confirmation. Harm is the spatial-block mean incidence of accepted-and-harmful queries; conditional harm is query-level among accepted actions. A/H gives accepted/harmful-accepted query counts.

Shard	A/H	Harm (%)	Cond. (%)	Coverage (%)	Utility (pp)
Test 25	641/3	0.250	0.468	66.14	11.55
Test 26	616/2	0.270	0.325	58.84	12.58
Test 27	550/4	0.229	0.727	54.74	10.78
Test 28	556/0	0.000	0.000	59.72	16.53
Pooled Q-Safe	2363/9	**0.190**	0.381	**59.79**	**12.82**
Frozen ExtraTrees	2289/11	0.248	0.481	58.71	11.66

0.028 points (Table 1). The representation is thus a harm-ranking inductive bias, not an explanation for a large utility gain.

The secondary three-policy potential removes pairwise cycles but behaves almost identically to product-only routing; we therefore report it only in Appendix G.

5.2 Initial frozen binary-harm confirmation

Table 2 reports the first frozen evaluation. All four shards pass the primary criterion. Pooled risk is 0.190% (95% block-bootstrap CI: 0.061–0.345), with 59.79% coverage and 12.82-point deployed utility gain. Blind block sizes have min/median/95th-percentile/max $3/3/12/45$ queries. Under query weighting, risk, coverage, and utility are 0.189%, 49.63%, and 10.68 points; the similar risk but different coverage shows why the sampling unit must be explicit.

Across the same 951 blocks, paired Q-Safe-minus-ExtraTrees differences are $+1.08$ coverage points (95% CI: -0.30 to $+2.46$), -0.058 harmful-action points (95% CI: -0.213 to $+0.085$), and $+1.157$ utility points (95% CI: $+0.812$ to $+1.512$). Thus only the utility improvement is statistically resolved. It is a 9.9% relative increase over the comparator's 11.66-point utility, but its operational meaning remains benchmark-specific rather than a measured reduction in proofreading time. Harm-ranking AUROC is 0.902 and AP is 0.355 on the blind pool.

5.3 Frozen fallback-regret confirmation

After test25–28 became open, we selected $L_r = \max\{-\Delta, 0\}$, retained the same scorer, and froze the complete protocol before downloading test29–32. The primary endpoint requires block-marginal regret below 2% separately on every shard. Source data, model, code, and prediction-export hashes all match the frozen record.

All four shards pass. Across 963 blocks, regret-minus-incidence differences are $+13.99$ coverage points (95% CI: $+12.38$ to $+15.67$), $+3.30$ utility points ($+2.74$ to $+3.89$), and $+0.104$ regret points ($+0.040$ to $+0.187$). This is a declared trade-off, not dominance: accepted harm increases by 0.479 points. Unlike the first confirmation, the utility change is fly-driven: fly coverage rises from 31.70% to 81.53% and utility from 1.64 to 13.20 points; mouse is nearly saturated, human is unchanged, and unsupported zebrafish always falls back.

Table 3: Prospective fallback-regret confirmation. Harm is reported but not the declared endpoint. The final row applies the original incidence thresholds to the same new data. All metrics except conditional counts are block-weighted.

Shard/contract	A/H	Harm (%)	Regret (%)	Coverage (%)	Utility (pp)
Test 29	923/15	1.373	0.320	75.55	12.42
Test 30	818/8	0.575	0.060	73.86	17.06
Test 31	889/19	0.927	0.180	75.35	17.06
Test 32	882/19	1.134	0.186	74.05	13.02
Pooled regret contract	3512/61	1.018	**0.191**	**74.72**	**14.75**
Pooled incidence contract	2419/19	0.539	0.088	60.73	11.46

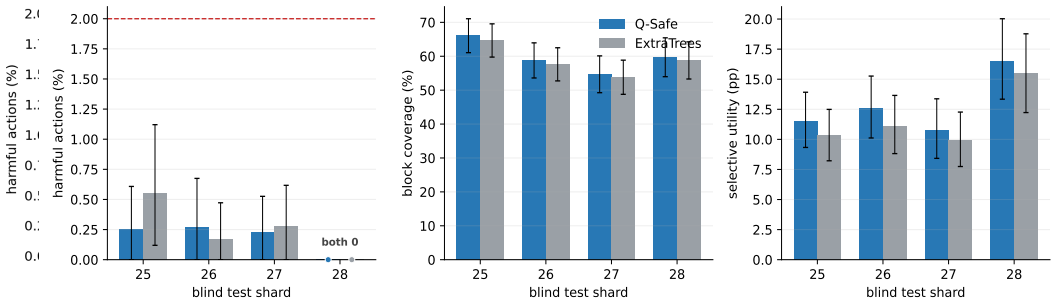

Figure 2: Frozen blind comparison. The dashed line is the pre-registered 2% harmful-action target; lower is better for the left panel and higher is better for coverage and utility. Error bars are 95% spatial-block bootstrap intervals within each frozen shard; they visualize sampling variation and were not used for model or threshold selection.

5.4 Measurement composition and budget sensitivity

Support-only acquisition uses 1.210 views per candidate on average (1.199/1.206/1.210/1.224 across shards), a 59.68% reduction from three views, while exactly preserving all tokens, actions, and gate decisions. At 1%, 2%, and 5% action budgets, Q-Safe obtains block coverage 54.62/57.66/67.10%, harm risk 0.342/0.228/0.000%, and utility 16.54/15.07/6.85 points. The corresponding frozen-gate utility is 15.75/12.65/6.58 points. The benefit is therefore largest at the central 2% operating budget rather than monotonically increasing with the number of actions.

5.5 External stress tests

The main experiment tests the complete learned Q-Safe system. We additionally ask whether its two design premises—compare actions directly and align scores with the scientific query—survive in unrelated imaging settings. These are mechanism tests with pre-existing frozen pipelines, not claims that the connectomics network transfers across modalities.

Live-cell restoration: when learning should abstain. On held-out Fluo-N2DL-HeLa (Ulman et al., 2017), only 3.1% of training examples favor the event-safe restoration, so a learned selector collapses. A deterministic counterfactual verifier reduces event shift by 0.0658 (95% CI: 0.0197–0.1184) for a 0.071-dB PSNR cost. Explicit comparison is useful, but the rare decision does not justify a learned gate.

Serial block-face EM: query-aligned acquisition. At a 25% rescan budget on a held-out SBEM volume, a membrane-aligned Ridge score reaches Spearman 0.853 (95% CI: 0.822–0.882) and captures 40.9% of SegNeuron disagreement, versus 28.4% for pixel risk and 39.9% for matched entropy (Zhang et al., 2024). This supports objective alignment, not learned superiority or biological ground truth.

2D multi-dwell EM: a multi-tissue schedule falsification. Registered EMDiffuse fields (Lu et al., 2024) comprise 10 bone-marrow, 10 heart, and 16 liver regions at 1/2/3/24 μs dwell. We compare a full 1-μs preview plus complete 3-μs rescans of one third of native tiles against the equal-cost observed uniform 2-μs image, using 24 μs as reference. Despite monotone loss improvement with dwell, the top-tile oracle is worse than uniform acquisition for every tissue, pixel/edge/dark-structure query, and normalization: all 18 region-bootstrap intervals are below zero. The adaptive schedule is therefore infeasible regardless of ranker accuracy; full values and preprocessing checks appear in Appendix I.

5.6 Blinded human proxy audit

Three independent annotators completed frozen membrane/mitochondria packs. On the physical-scale-corrected PFIB cases, raw MitoNet probability disagreement has negative rank correlation with all three raters and captures only 20.9/21.7/24.4% of rated loss at 25% budget. However, ordinal reliability is low ($\kappa = -0.037$ to 0.167) and a pre-frozen mass normalization remains unresolved on held-out raters. We conclude only that the raw score fails this noisy audit; protocols and full results are in Appendix I.

6 Discussion

What is Q-Safe useful for? Q-Safe decides whether an adaptive policy may replace a concrete equal-budget reference. Oracle headroom is checked before training; failure rejects the schedule regardless of ranker accuracy. Surviving proposals are compiled with a declared loss, spatial sampling unit, calibration domains, and executable fallback. A higher-level agent may propose the query and policy, but the replacement decision remains deterministic and auditable.

What did the evidence isolate? The frozen +1.16-point gain is predominantly mouse-driven, while representation ablations change utility by at most 0.028 points. Post-confirmation matched-gate tests also find that simpler encoders can match or exceed the selected network (Appendix E); architecture is therefore not the central claim. In contrast, the independently frozen regret contract reproduces the exploratory effect on new shards: relative to incidence control, coverage and utility rise by 13.99 and 3.30 points. The scientific loss contract—not the scorer brand—is therefore operationally consequential. The price is more low-severity harmful actions, so users must declare rather than hide this trade-off. Support-only acquisition and hardware caveats appear in Appendix C.

7 Limitations and falsification criteria

The end-to-end frozen evaluations are both ConnectomeBench2, utility is a benchmark proxy rather than measured proofreading time, and zebrafish always falls back. The regret loss was chosen after the first blind set became open; test29–32 provide an independent second-generation confirmation, not a repair of the original endpoint. The add-one rule is conservative, not a new conformal theorem or protection against arbitrary shift. Action identities are implementation checks, and matched-architecture comparisons remain exploratory.

External imaging studies test mechanisms, not encoder transfer: EMDiffuse rejects one schedule geometry; PFIB/SBEM have limited specimens; and low rater reliability precludes definitive human ground truth. The claim should be falsified if a new frozen cohort violates the declared block risks, consensus adjudication rescues the rejected raw proxy, or full-view evidence changes a support-only decision under its assumptions. Detailed scope and data limitations are retained in Appendix I.

8 Conclusion

Q-Safe evaluates adaptive imaging as replacement of a budget-matched fallback. It tests oracle headroom, objective alignment, and proxy validity before calibrating a declared bounded loss. Two prospective confirmations and negative external tests show that the contract—and earlier diagnostic failures—change deployment. The contribution is a falsifiable loss-contract compiler, not an architectural claim.

REFERENCES

Jonatan Alvelid, Martina Damenti, Chiara Sgattoni, and Ilaria Testa. Event-triggered STED imaging. *Nature Methods*, 19:1268–1275, 2022. doi: 10.1038/s41592-022-01588-y.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020. doi: 10.1073/pnas.1907377117.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535.

Chinmay Belthangady and Loic A. Royer. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nature Methods*, 16:1215–1225, 2019. doi: 10.1038/s41592-019-0458-z.

Jia Bi, Samuel Pinilla, and Chenyang Zhu. Testing when adaptive data acquisition can replace fixed measurement plans, 2026.

Anthony Bilodeau, Albert Michaud-Gagnon, Julia Chabbert, Benoit Turcotte, J"orn Heine, Audrey Durand, and Flavie Lavoie-Cardinal. Development of AI-assisted microscopy frameworks through realistic simulation with pySTED. *Nature Machine Intelligence*, 6:1197–1215, 2024. doi: 10.1038/s42256-024-00903-w.

Jeff Brown, Tim Farkas, Gleb Razgar, and Edward S. Boyden. ConnectomeBench2: A unified benchmark for automated connectomic proofreading, 2026.

Ryan Conrad and Kedar Narayan. Instance segmentation of mitochondria in electron microscopy images with a generalist deep learning model trained on a diverse dataset. *Cell Systems*, 14(1):58–71.e5, 2023. doi: 10.1016/j.cels.2022.12.006.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://www.jmlr.org/papers/v11/el-yaniv10a.html.

Aditya Gangrade, Anil Kag, and Venkatesh Saligrama. Selective classification via one-sided prediction. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2179–2187. PMLR, 2021. URL https://proceedings.mlr.press/v130/gangrade21a.html.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Sunay Joshi, Tao Wang, Hamed Hassani, and Edgar Dobriban. Risk-controlled post-processing of decision policies, 2026.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Chixiang Lu, Kai Chen, Heng Qiu, Xiaojun Chen, Gu Chen, Xiaojuan Qi, and Haibo Jiang. Diffusion-based deep learning method for augmenting ultrastructural imaging and volume electron microscopy. *Nature Communications*, 15:4677, 2024. doi: 10.1038/s41467-024-49125-z.

Dora Mahecic, Willi L. Stepp, Chen Zhang, Juliette Griffié, Martin Weigert, and Suliana Manley. Event-driven acquisition for content-enriched microscopy. *Nature Methods*, 19:1262–1267, 2022. doi: 10.1038/s41592-022-01589-x.

Yaron Meirovitch, Ishaan Singh Chandok, Core Francisco Park, Pavel Potocek, Lu Mi, Shashata Sawmya, Yicong Li, Thomas L. Athey, Vladislav Susoy, Neha Karlupia, Yuelong Wu, Daniel R. Berger, Richard Schalek, Caitlyn A. Bishop, Daniel Xenes, Hannah Martinez, Jordan Matelsky, Brock A. Wester, Hanspeter Pfister, Remco Schoenmakers, Maurice Peemen, Jeff W. Lichtman, Aravinthan D. T. Samuel, and Nir Shavit. SmartEM: Machine-learning guided electron microscopy. *Nature Methods*, 23(1):193–204, 2026. doi: 10.1038/s41592-025-02929-3. Published online 29 December 2025; issue date January 2026.

Can Polat, Erchin Serpedin, Mustafa Kurban, and Hasan Kurban. STEMGym: Benchmarking sequential decision-making under dose budgets in autonomous electron microscopy, 2026.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ryxGuJrFvS`.

Marcel Schloz, Johannes M"uller, Thomas C. Pekin, Wouter Van den Broek, Jacob Madsen, Toma Susi, and Christoph T. Koch. Deep reinforcement learning for data-driven adaptive scanning in ptychography. *Scientific Reports*, 13:8732, 2023. doi: 10.1038/s41598-023-35740-1.

Patrick Trampert, Faysal Bourghorbel, Pavel Potocek, Maurice Peemen, Christian Schlinkmann, Tim Dahmen, and Philipp Slusallek. How should a fixed budget of dwell time be spent in scanning electron microscopy to optimize image quality? *Ultramicroscopy*, 191:11–17, 2018. doi: 10.1016/j.ultramic.2018.03.007.

Vladimír Ulman, Martin Maška, Klas E. G. Magnusson, et al. An objective comparison of cell-tracking algorithms. *Nature Methods*, 14:1141–1152, 2017. doi: 10.1038/nmeth.4473.

Martin Weigert, Uwe Schmidt, Tobias Boothe, Andreas Müller, Alexandr Dibrov, Akanksha Jain, Benjamin Wilhelm, Deborah Schmidt, Coleman Broaddus, Siân Culley, Mauricio Rocha-Martins, Fabián Segovia-Miranda, Caren Norden, Ricardo Henriques, Marino Zerial, Michele Solimena, Jochen Rink, Pavel Tomancak, Loic Royer, Florian Jug, and Eugene W. Myers. Content-aware image restoration: Pushing the limits of fluorescence microscopy. *Nature Methods*, 15:1090–1097, 2018. doi: 10.1038/s41592-018-0216-7.

Yunpeng Xu, Mufang Ying, Wenge Guo, and Zhi Wei. Two-stage risk control with application to ranked retrieval. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 9104–9111, 2025. doi: 10.24963/ijcai.2025/1012.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html`.

Yanchao Zhang, Jinyue Guo, Hao Zhai, Jing Liu, and Hua Han. SegNeuron: 3d neuron instance segmentation in any EM volume with a generalist model. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15008 of *Lecture Notes in Computer Science*, pages 589–600. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72111-3_55.

Yurui Zheng and Ying Jin. Prediction sets for counterfactual decisions: Coverage, optimality, and conformal prediction, 2026.

A Reproducibility and freeze record

All architecture choices, hashes, thresholds, and negative development results were stored before downloading test shards 25–28. The protocol was frozen at 2026-09-10 20:38:51 UTC. The selected checkpoint SHA-256 begins `76fe864c`; the frozen shared-view backbone begins `c0e984`; and the old comparator begins `2237b8`. The blind script rejects any mismatch between these values, the protocol, and expected held-out-shard hashes. Full digests are distributed in `artifacts/frozen_protocol.json` rather than abbreviated in the paper.

The fallback-regret extension has a separate prospective record because its loss was chosen after test25–28 became open. Before downloading test29–32, we froze the 2% target, $[0, 1]$ loss bound, species thresholds, primary per-shard success rule, official source hashes, and hashes of the unchanged checkpoint, backbone, compiler, toolkit, exporter, and confirmation script. All four raw and derived prediction files were absent at freeze. The complete records are `artifacts/frozen_regret_contract_protocol.json` and `artifacts/regret_contract_confirmation.json`.

A direct implementation test on 128 validation queries permutes every valid candidate and appends seven arbitrary masked tokens. Maximum absolute changes in either output head are 9.54×10^{-7} and 4.77×10^{-7}, respectively, within a 10^{-6} floating-point tolerance. This verifies that the implementation preserves the claimed permutation and padding invariances.

The distributed command-line/Python interface independently reproduces the frozen blind ledger from immutable per-query records: 2363 accepted queries, 9 harmful accepted queries, 0.190% block risk, 59.79% block coverage, and 12.82-point block utility. Unit tests cover signed cancellation, feasible and infeasible oracle schedules, block-versus-query weighting, tied scores, domain-intersected thresholds, and unsupported groups. Thus the claimed decision layer can be invoked and checked without loading the neural scorer.

The frozen species thresholds on harm logits are 0.02403 for fly, 0.01347 for human, and 0.62307 for mouse. Zebrafish is assigned $-\infty$ because it lacks the required calibration support. Thresholds are not comparable as calibrated probabilities; they are only cut points on the frozen model score.

Calibration assumes that, within each species–domain pair, the pre-declared spatial blocks and a future block are exchangeable under the deployment data- generating process. It makes no claim for an unobserved domain. Table 4 exposes the sample support and the add-one barrier; zebrafish cannot attain 2% even with zero accepted harm because $1/(B + 1) > 0.02$.

Table 4: Frozen calibration ledger. Harm labels count all harmful proposals, not only accepted ones; Corr. is add-one corrected block risk at the selected threshold, and Cov. is block-weighted calibration coverage.

Species	Domain	Blocks	Harm labels	Threshold	Corr. (%)	Cov. (%)
Fly	val-2	62	26	0.04616	1.867	64.06
Fly	val-3	66	31	0.02403	1.990	33.76
Human	val-2	52	17	0.02326	1.887	41.35
Human	val-3	58	31	0.01347	1.695	24.71
Mouse	val-2	104	8	0.62307	1.905	97.41
Mouse	val-3	115	3	0.90897	1.437	100.00
Zebrafish	val-2	6	0	–	14.286	0.00
Zebrafish	val-3	14	5	–	6.667	0.00

B Compiled token specification

C Design properties and proofs

Signed difference. For additive utility, expand $U(A_p; q) - U(A_0; q)$. The constant and every candidate in $A_p \cap A_0$ cancel, leaving Equation 7. For linear candidate value $v_i(q) = \beta(q)^\top z_i$, this becomes $\beta(q)^\top (\sum_{i \in A_p} z_i - \sum_{i \in A_0} z_i)$. The identity applies to any two feasible action sets,

Table 5: Observable information supplied to Q-Safe. Labels used to compute realized utility never enter the model.

Group	Contents
Visual	frozen probability, Bernoulli uncertainty
Geometry	normalized relative offsets and query distance
Query	relevance and probability–relevance product
Decision	signed margins to both top-k boundaries; membership in proposed and fallback sets
Global	query scale, budget fraction, support size, observable summaries
Excluded	candidate ground truth, realized action utility, shard outcome

including sets selected under heterogeneous cost constraint $\sum_{i \in A} c_i \leq B$; Top-$k$ is the equal-cost special case used in the primary benchmark.

Decision-preserving support. If both policies break ties using pre-acquisition metadata, $k \leq |S(q)|$, and the gate reads only compiled tokens and query metadata, candidates outside $S(q)$ have zero relevance and product score and cannot displace a supported candidate. Every encoder input, action, and threshold decision is unchanged by acquiring complete visual views only on $S(q)$. This proves information equivalence, not wall-clock savings when raster traversal or switching dominates.

Oracle feasibility. Every size-m complete-rescan policy starts from preview loss and subtracts the sum of selected marginal gains g_i. An exchange argument maximizes this sum with $\mathrm{TopM}(g)$; comparing that loss with uniform dwell gives Equation 9. Failure of the inequality rules out every ranking policy in this schedule class.

Cycle-consistent policy library. For fixed q, path independence $D_q(A, B) + D_q(B, C) = D_q(A, C)$ is equivalent to a scalar potential difference. Fixing reference A_0 and defining $V_q(A) = D_q(A, A_0)$ gives $D_q(A, B) = V_q(A) - V_q(B)$; the converse follows by cancellation. The learned potential may depend on query, domain, and budget but not pair-dependent statistics, otherwise the value assigned to A changes with its comparison partner.

D Architecture selection and ablation details

Table 6: Architecture selection on disjoint validation shards (three seeds).

Method	Worst-domain AUROC	Harm AP	Parameters
Summary MLP	0.781 ± 0.007	0.186	0.5K
DeepSets	0.810 ± 0.023	0.171	27.2K
Q-Safe (ERM)	0.795 ± 0.012	0.183	117.8K
Q-Safe (GroupDRO)	0.796 ± 0.011	0.207	117.8K
Counterfactual Q-Safe	0.825 ± 0.012	0.219	66.9K
Counterfactual Q-Safe + GroupDRO	0.822 ± 0.012	0.249	66.9K

Generic attention does not improve over DeepSets. Counterfactual ERM has the highest worst-domain AUROC but narrowly violates one reused-development risk check; GroupDRO is frozen because it passes the complete protocol. Removed ablation vectors are replaced by zeros, preserving parameter count. Exact optimizer, early-stopping, seed, and checkpoint records accompany the artifact. All training queries appear once per shuffled epoch in minibatches of 256; no harm or species resampling is used. AdamW uses learning rate 4×10^{-4} and weight decay 2×10^{-4}, with gradient norm clipped at 5. The positive-class weight and 0.15 regression coefficient in Equation 6, the GroupDRO step $\eta = 0.02$, and the maximum 100 epochs were fixed before the reused-development audit. Early stopping has patience 18 and maximizes the minimum AUROC across validation shards 0–1; validation shards 2–3 are used only for independent per-model calibration.

E POST-CONFIRMATION MATCHED-GATE AUDIT

After the blind set had been opened, we applied the identical public spatial- block calibrator to checkpoints that all predate blind opening. This analysis cannot change the frozen primary result and is explicitly exploratory. It asks whether the selected architecture is necessary for the interface.

Table 7: Post-confirmation common-calibration audit on the now-open blind set. ΔU is relative to the frozen selected counterfactual GroupDRO gate; all risk, coverage, and utility values are block-marginal.

Score source	Harm (%)	Coverage (%)	Utility (pp)	ΔU (pp)
Summary MLP harm	0.112	57.54	12.12	-0.70
DeepSets harm	0.178	59.12	13.17	$+0.35$
Generic attention harm	0.331	64.45	13.98	$+1.16$
Selected counterfactual harm	0.190	59.79	12.82	0.00
Selected predicted gain	0.191	32.28	13.23	$+0.42$

The DeepSets and generic-attention utility differences are respectively $+0.35$ points (95% block-bootstrap CI: $+0.08$ to $+0.68$) and $+1.16$ points ($+0.79$ to $+1.57$). The predicted-gain score attains higher mean utility only by reducing coverage by 27.51 points. These results falsify an architectural-superiority interpretation; they instead show that the same replacement compiler can expose materially different risk–coverage–utility operating points for interchangeable scorers. Only the selected row remains a protocol-frozen confirmation.

F FALLBACK-REGRET CONTRACT SELECTION RECORD

We next keep the frozen harm scorer but change the declared bounded loss from binary incidence to $L_r = \max\{-\Delta, 0\} \in [0, 1]$. The target remains 2%, and validation shards 2–3 are recalibrated with exactly the generalized Equation 8. Because the blind set is already open, this is a hypothesis-generating test of the compiler rather than a revised primary result.

Table 8: Post-confirmation loss-contract audit. Regret is the block mean of accepted fallback loss; all entries use the same frozen harm score.

Calibration contract	Harm (%)	Regret (%)	Coverage (%)	Utility (pp)
2% harm incidence (frozen)	0.190	0.019	59.79	12.82
2% fallback regret (post hoc)	0.659	0.104	74.30	16.29
Both constraints	0.190	0.019	59.79	12.82

Relative to the frozen incidence contract, the regret contract changes block coverage by $+14.51$ points (95% block-bootstrap CI: $+12.90$ to $+16.14$), utility by $+3.47$ points ($+2.89$ to $+4.08$), harm incidence by $+0.469$ points ($+0.260$ to $+0.707$), and realized regret by $+0.085$ points ($+0.038$ to $+0.142$). The utility increase is mainly fly rather than mouse: fly coverage rises from 35.18% to 82.31% and utility from 1.33 to 12.80 points, whereas mouse is already near saturation. Human is unchanged and zebrafish remains unsupported. Thus the scientific meaning of the loss contract materially changes deployment; it cannot be treated as a cosmetic choice of calibration metric.

This exploratory result defined, rather than confirmed, the second-generation hypothesis. At 2026-09-18 04:02:45 UTC we froze its endpoint, thresholds, test29–32 identifiers, source hashes, code hashes, and success rule while all four raw shards and prediction files were absent. Table 3 reports the subsequent one-shot confirmation; no threshold or endpoint was changed between freeze and evaluation.

G SECONDARY MULTI-POLICY AUDIT

Three equal-budget actions—relevance, probability-times-relevance, and uncertainty-times-relevance—supervise all unordered action pairs using additive benchmark utility. The exact poten-

tial eliminates cross-pair cycles and reduces pairwise error but does not materially improve oracle recovery.

Table 9: Post-confirmation three-policy audit (three seeds); blind shards are not reopened. Cycle is the maximum gain-cycle residual.

Comparator	Params	AUROC	Gain MAE	Oracle (%)	Cycle
Signed pairwise	40.7K	0.928 ± 0.003	0.122	**95.96**	0.526
Exact potential	24.0K	$\mathbf{0.932 \pm 0.001}$	**0.101**	95.89	$\mathbf{1.2 \times 10^{-7}}$

After identical 2% calibration, both the potential and product-only routers pass all 12 development shards. Their worst risk is 1.77%; coverage is 27.89% versus 27.84%, and deployed gain is 11.36 versus 11.34 points. Accepted actions are almost entirely the product policy, so this experiment supports structural consistency rather than practical library-level gain.

H Development history and negative results

The project retains rather than overwrites unsuccessful variants. The summary MLP is weak; DeepSets improves worst-domain ranking; generic query attention does not; and the counterfactual ERM model attains the best AUROC but violates one reused development risk check. The GroupDRO counterfactual model is selected because it passes the complete development protocol, not because it is best on every scalar metric. Test 13–24 are labeled development reuse after earlier inspection and are never described as blind.

At the blind stage, Q-Safe does not significantly improve coverage or harmful action rate over Extra-Trees. Post-hoc species inspection finds concentration of benefit in mouse. These results preclude three tempting claims: universal cross-species improvement, a new distribution-free conditional guarantee, and learned-gate superiority in low-sample rare-event regimes.

I External diagnostic details

Table 10: Held-out SBEM objective-alignment audit at 25% rescan budget.

Acquisition score	Spearman	Top-k overlap	Risk captured
Random expectation	0.000	25.0%	25.0%
Pixel-risk calibrator	0.361	34.2%	28.4%
SegNeuron entropy	0.786	61.7%	39.9%
Query-aligned Ridge	**0.853**	**69.2%**	**40.9%**
Oracle	1.000	100.0%	44.4%

Table 11: Equal-cost EMDiffuse oracle loss reduction versus uniform 2 μs, multiplied by 100. Positive would favor adaptive rescanning; brackets are 95% region-bootstrap intervals.

Tissue	Query loss	Per-image norm.	Fixed-ref. norm.
Bone marrow	pixel	$-2.75\ [-3.09, -2.49]$	$-6.92\ [-8.57, -5.37]$
Bone marrow	edge	$-0.98\ [-1.06, -0.90]$	$-0.38\ [-0.64, -0.14]$
Bone marrow	dark structure	$-3.75\ [-4.25, -3.23]$	$-14.00\ [-15.48, -12.36]$
Heart	pixel	$-3.01\ [-3.22, -2.87]$	$-2.97\ [-3.87, -2.18]$
Heart	edge	$-1.13\ [-1.16, -1.09]$	$-0.67\ [-0.78, -0.58]$
Heart	dark structure	$-1.84\ [-2.70, -1.04]$	$-4.96\ [-7.01, -3.26]$
Liver	pixel	$-2.96\ [-3.13, -2.80]$	$-2.14\ [-2.98, -1.34]$
Liver	edge	$-0.91\ [-0.98, -0.83]$	$-0.66\ [-0.75, -0.57]$
Liver	dark structure	$-3.30\ [-4.11, -2.58]$	$-6.59\ [-8.03, -5.31]$

Table 12: Independent-rater agreement. Eval. is agreement on endpoint evaluability; QWK and exact/within-one agreement use cases evaluable to both.

Pack	Endpoint	Eval.	QWK	Exact	Within 1	R2−R1
V2	membrane	100.0%	0.064	22.5%	75.0%	−0.725
V2	mitochondria	65.0%	−0.016	5.3%	73.7%	−1.211
V3 PFIB	membrane	100.0%	0.006	31.3%	81.3%	−0.875
V3 PFIB	mitochondria	93.8%	0.081	44.4%	92.6%	−0.630

Table 13: Corrected V3 PFIB mitochondrial audit. Annotator 1 is development; annotator 2 is held out for the pre-frozen mass-normalized score. Capture is at 25% budget; random expectation is 25%.

Rater	n	Raw ρ	Raw capture	Norm. ρ	Norm. capture
Annotator 1 (development)	29	−0.149	20.9%	0.289	27.9%
Annotator 2 (held out)	27	−0.091	21.7%	0.078	26.1%
Annotator 3 (additional)	32	−0.093	24.4%	0.065	26.7%

For held-out annotator 2, the mass-normalized-minus-raw capture difference is $+3.85$ points with 95% paired spatial-group bootstrap interval $[-11.76, +17.86]$. The frozen rescue hypothesis is therefore unresolved. The additional annotator gives $[-4.65, +8.70]$ points and does not resolve it. The three-rater median is descriptive only: raw MitoNet has $\rho = -0.101$ and 22.6% capture, but low ordinal agreement makes this an unreliable substitute for adjudicated consensus.

HeLa. The two candidates differ only in a fixed restoration setting. Mitosis/event locations are detected by a fixed spatiotemporal rule. Bootstrap intervals resample test samples. The deterministic selector compares candidate outputs and routes disagreement cases to the event-safe action; it is not trained on sequence 02.

SBEM. Each volume contributes 15 cases with 64 tiles per case. The Ridge model is fit on all 960 training-volume tiles; case-held-out validation is used only inside the training volume. Evaluation uses the second volume and case bootstrap. A slow scan is assumed perfectly registered and directly replaces its fast-scan tile, so physical motion, boundary overhead, and online adaptation are omitted.

Real dwell-time stress test. After the initial draft, we received registered mouse-brain acquisitions with PFIB dwell times 0.25–2 μs and an SBEM 0.5/2 μs pair. On 100 fixed spatial crops per modality, frozen SegNeuron disagreement increases as dwell falls (PFIB rank correlation 0.90). A PFIB-trained structural-risk Ridge model captures 32.83% of SBEM structural proxy risk at 25% budget, versus 16.24% for an otherwise matched pixel-risk model; their difference is 16.58 points (spatial-group bootstrap CI: 9.98–24.41). However, SegNeuron entropy captures 33.20%, statistically indistinguishable from the learned structural model. Thus real dwell data strengthen the objective-alignment claim but not learned selector superiority.

The six-level PFIB ladder also permits an exploratory marginal-value replay. An incremental-exposure counterfactual reaches 92.0/93.9% of oracle gain at mean 0.75/1 μs budgets, versus 72.9/82.2% for uniform dwell and 85.5/85.9% for entropy. The acquisition protocol was subsequently confirmed to be a complete same-site scan at every dwell, rather than an in-place exposure extension. We therefore retain the incremental result only as a mechanism analysis and use an additive preview-plus-rescan cost for all deployable claims.

This conditional signal is not explained solely by the dwell-action curve. At 0.75 μs, action-only models recover 51.4–65.3% of oracle gain across three structural targets and four spatial split schemes, whereas all observable image features recover 87.8–92.2%. As a post-hoc sensitivity analysis, we project each tile's structural-risk curve to be non-increasing with dwell and repeat the spatial OOF experiment. The learned policy then recovers 89.3–91.4% and 94.1–94.9% of monotone oracle gain at 0.75 and 1 μs, respectively. For all three targets and both budgets, 500-sample spatial-group bootstrap intervals for the risk improvement over both uniform and entropy exclude zero. These au-

dits remove the 1.5 μs proxy anomaly as an explanation, but they do not remove the single-specimen or retrospective-replay limitations.

Under the confirmed protocol, every tile pays a 0.25 μs preview and a selected tile then pays the full dwell of a second scan; the comparator is a uniform single pass with no preview overhead. This conservative replay reveals a narrow budget regime rather than a general advantage. At 0.75 and 1 μs, learned policies are statistically unresolved on monotone SegNeuron risk and are often significantly worse on monotone MitoNet risk, although an oracle shows allocation headroom. At 1.5 μs, both proxies become positive: the SegNeuron risk-times-curve policy improves remaining risk by 0.02309 (spatial-group 95% CI: 0.02002–0.02613), and its MitoNet counterpart improves it by 0.00624 (0.00413–0.00843). HGB and entropy are also positive for both targets at this budget. A stronger content-free oracle, which knows the mean risk–dwell curve and optimally mixes dwell levels, already explains 0.01494 and 0.00307 of these improvements. Nevertheless, risk-times-curve remains better than this global mixture by 0.00815 (0.00630–0.00993) on SegNeuron and 0.00317 (0.00099–0.00561) on MitoNet. Thus complete-rescan adaptation has a reproducible within-volume sweet spot with a separable content-allocation component, but neither universal budget robustness nor learned superiority over uncertainty is supported.

To test whether this is merely SegNeuron-specific, we freeze the official MitoNet-v1-mini TorchScript model (Conrad and Narayan, 2023) and extract semantic mitochondrial probabilities on the same 100 PFIB crops. MitoNet disagreement with the 2 μs reference rises to 0.0837 probability MAE and 9.17% mask XOR at 0.25 μs; foreground inflation at low dwell also reveals noise-induced false positives. Target specificity is pronounced. At 0.75/1 μs, the SegNeuron-objective policy attains 91.9/94.3% of the SegNeuron oracle but only 67.8/68.7% of the MitoNet oracle; the MitoNet-objective policy attains 90.4/93.1% on MitoNet but 79.6/82.8% on SegNeuron. In both transfer directions and all three budgets, 500-sample spatial-group intervals favor the target-specific policy. A normalized joint objective is a useful but weaker compromise. The learned MitoNet policy is statistically indistinguishable from MitoNet entropy. Thus the new model supports query-conditioned acquisition, not a universal learned agent or universal superiority over uncertainty. MitoNet outputs remain model proxies without expert mitochondrial labels, and no 3D instance accuracy is claimed.

We then freeze a stricter PFIB-to-SBEM transfer at the common 0.5 μs endpoint. A MitoNet-objective Ridge model is trained only on PFIB spatial groups and evaluated against MitoNet disagreement with the registered SBEM 2 μs scan. It retains 0.798 Spearman correlation and captures 40.73% of SBEM proxy risk at 25% budget, compared with 22.59% for a SegNeuron-objective Ridge and 17.51% for a pixel-objective Ridge. Spatial-group bootstrap differences are 18.48 points (95% CI: 12.57–25.36) and 22.45 points (17.21–28.39), respectively. However, MitoNet entropy captures 40.17%, and its difference from MitoNet-objective Ridge is unresolved (-2.15 to $+1.66$ points). This external replication survives a mean feature shift of 2.67 PFIB standard deviations and isolates the supported conclusion: the acquisition score must match the scientific objective, while learning is not automatically better than target-matched uncertainty. This conclusion is unchanged when the proxy target is replaced post hoc by MitoNet mask XOR at the frozen 0.3 threshold: Mito-objective Ridge captures 40.17% of SBEM XOR risk and exceeds SegNeuron- and pixel-objective Ridge by 16.36 points (9.29–24.75) and 20.89 points (14.55–28.15). Policies trained for MitoNet probability MAE and mask XOR are statistically indistinguishable, suggesting robustness within the mitochondrial objective family rather than to a single continuous error definition. Nor is the tie with entropy explained by an impoverished feature vector. We add 13 observable probability-geometry statistics, including quantiles, spatial and axial variation, uncertain-area fractions, boundary density, and component density. Rich Ridge improves PFIB group-OOF captured risk from 41.65% to 43.58%, apparently exceeding entropy at 42.07%. Frozen SBEM transfer reverses this order: rich Ridge captures 39.47%, versus 40.73% for basic Ridge and 40.17% for entropy; both paired intervals cross zero. Rich mito-only Ridge and HGB fall further. This is a concrete domain-shortcut failure and argues against selecting a more complex agent from within-volume performance.

Manual annotation protocols. The v2 annotation pack was frozen before manual labels and contains 64 outcome-independent confirmatory cases and 16 disagreement-enriched diagnostic cases. Candidate and reference are registered 0.5- and 2-μs complete scans. The interface hides modality, cohort, coordinates, and all proxy scores; ratings are stored locally and exported as CSV. The two primary ordinal endpoints ask whether membrane or mitochondrial structure visible in the reference

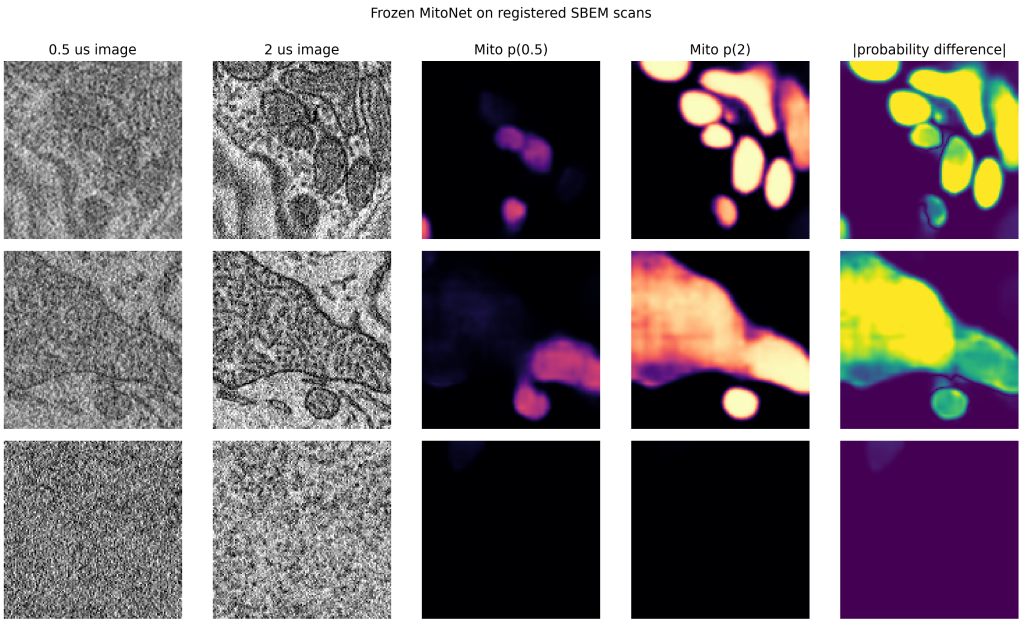

Figure 3: Post-hoc qualitative audit on registered SBEM scans. Rows show the highest MitoNet proxy-risk tile, a tile whose MitoNet-entropy rank exceeds its SegNeuron-entropy rank by 96 percentiles, and the lowest-risk tile. Columns show matched 0.5/2 μs images, frozen semantic probabilities, and their absolute difference. This panel explains objective mismatch; it was not used to select models or thresholds.

remains interpretable in the candidate. An endpoint-specific NA is permitted when the reference lacks the structure, and exclusion defaults to zero.

The v3 PFIB supplement was frozen after inspecting the v2 pilot but before its own ratings. Eligibility uses only reference MitoNet foreground at least 0.01 and boundary fraction at least 0.0005. Within each of 16 pre-declared spatial groups, two cases are selected by a fixed SHA-256 order. Raw TIFFs remain on the native 5-nm grid; previews use nearest-neighbor rendering with no linear or area resampling. A yellow box identifies the original scored ROI inside the larger context. Confirmatory and post-pilot results are reported separately. Sixteen cases appear in both packs; expansion reduces NA from six to one among these repeats, and the ten cases evaluable twice have 80% exact mitochondrial-score agreement (weighted $\kappa = 0.524$). This is a within-rater context audit, not inter-rater reliability.