WHEN CAN EXPERT CORRECTIONS BE TRUSTED? EVIDENCE, LIMITS, AND RISK-CONTROLLED DECISIONS ACROSS ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

When a strong reference model and an automated expert disagree, the same correction rule can help on one patient's slide and harm on another. We ask when such corrections can be trusted across environments, and propose a unified framework separating the construction of corrective evidence from the decision to act on it. On the evidence side, we introduce CARE-Cell, a cell-classification architecture combining frozen tissue representations, local nuclear features, and target-conditioned context to build both a strong reference predictor and complementary corrective candidates. On the decision side, a risk-controlled layer determines when to revise, retain, or defer. Theory explains why useful evidence alone does not authorize action: we construct observationally indistinguishable environments whose optimal correction decisions are opposite, yielding a worst-case regret lower bound for any router measurable with respect to reference outputs, expert outputs, and confidence. The decision layer accordingly offers two complementary guarantees. With target-environment labels, an independent randomized audit certifies positive aggregate utility on the remaining queries with a finite-sample error guarantee, retaining the reference predictions when certification fails. Without such labels, labeled calibration environments set an accept/defer threshold that, under exchangeability, keeps a new environment's retained-error fraction below a prescribed limit with a marginal probability guarantee. We evaluate on a consecutive dermatopathology cohort collected without case selection based on image appearance or model performance, together with Lizard, PanNuke, Camelyon17-WILDS, and iWildCam-WILDS. CARE-Cell improves inflammation F1 over a UNI2-based cell classifier by 4.59, 8.35, and 1.84 percentage points on the three cell-classification datasets; under identical splits and GT instances, it improves over a CellViT++-based classification baseline by 11.95 points on Lizard. In separate fit/calibration/target evaluations, selective acceptance retains 92.15% of Lizard and 86.26% of PanNuke samples for automatic prediction, giving accepted-subset inflammation F1 of 97.14% and 87.65% without changing predictions, while hospital-shift stress tests illustrate failures outside the calibration assumptions. Together these separate three questions often conflated in expert augmentation: whether corrective evidence exists, whether its utility is identifiable, and which deployment decisions the available information can justify.

1 INTRODUCTION

Pathology foundation models and specialized cell experts expose complementary tissue and nuclear evidence (Chen et al., 2024; Chadoutaud et al., 2026; Hörst et al., 2026). Yet an expert can help on one slide and harm on another despite similar outputs and confidence (Figure 1). Whether an expert can correct a reference and whether its correction rule transfers are therefore distinct questions.

Figure 1c uses blue/red/gray ticks for positive/negative/tie-or-no-action slide utility; Subset A/B are training subsets, Alt. layer an alternative frozen layer, and Strong aug. a strongly augmented expert. Diamonds show pooled utility (3.71; 0.45 without slide 48). We ask: *when can expert corrections be trusted across environments?* Our framework separates evidence construction from action authoriza-

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.

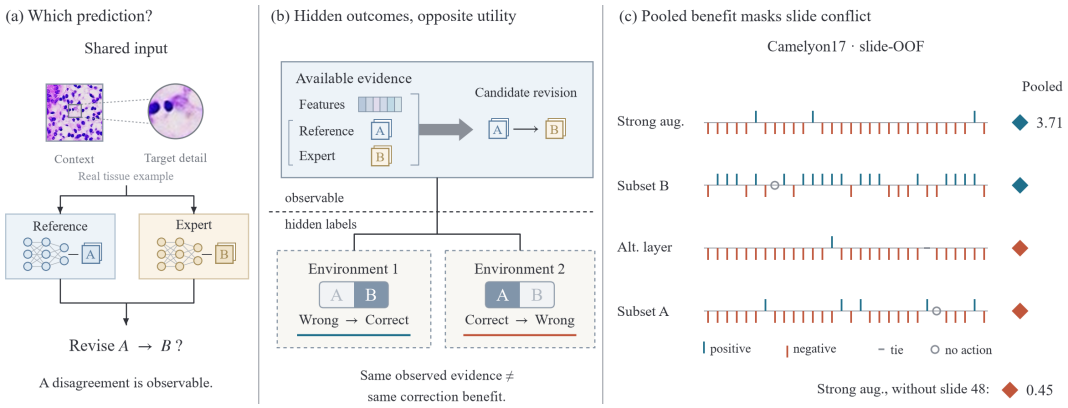


Figure 1: **Observable disagreement can hide opposite utility.** The same observed transcript can support opposite correction outcomes across environments.

tion through a frozen reference and candidate interface. CARE-Cell supplies a concrete pathology architecture; the decision layer uses the available labels to justify revision or selective acceptance. The unifying contribution is this information-dependent decision framework, not a claim that every component adds to the same F1 score.

We propose CARE-Cell to combine frozen tissue representations, local nuclear features, morphology, and target-conditioned context into a strong reference and complementary candidates. Their utility still needs justification. With target labels, a utility learner proposes revisions and an independent randomized audit certifies positive aggregate additive utility on remaining queries, retaining the reference on rejection. Without target labels, labelled calibration environments set an accept/defer threshold with a marginal retained-error guarantee under exchangeability. These are two information-dependent branches of the same decision interface, not interchangeable guarantees.

Theory specifies the information each stage needs. T1 constructs observationally indistinguishable environments with opposite optimal corrections. T2 characterizes the additive-utility oracle for a fixed information partition and separates granularity loss from finite-label sign-estimation loss. T3 provides the exchangeable-environment acceptance guarantee. We test these distinctions on five datasets spanning cell classification, tumor detection, and wildlife recognition. The consecutive clinical cohort exposes naturally occurring patient variation without appearance- or performance-based case selection. CARE-Cell gains 4.59/8.35/1.84 inflammation-F1 points over UNI2-based classification on private validation/Lizard/PanNuke; target-supported revision adds 1.57 points privately but little publicly. Exchangeable-split calibration and hospital-shift stress tests distinguish justified decisions from failures outside the assumptions.

2 Problem: Information Is Not Authorization

Let $f(X)$ be a frozen reference, $h(X)$ a fixed candidate proposal, and E an environment such as a slide or camera. A policy $\pi(W) \in [0, 1]$ revises f to h using observable information W. Include in W every quantity available to the policy: embeddings, probabilities, environment metadata, and the unlabeled target batch. Define multiclass revision utility and its aggregate gain as

$$U = \mathbf{1}\{h(X) = Y\} - \mathbf{1}\{f(X) = Y\}, \qquad G(\pi) = \mathbb{E}[\pi(W)U]. \tag{1}$$

For a finite action set, let B count wrong-to-correct changes and H correct-to-wrong changes. Then $B - H$ is the net correct count. This additive quantity is *not* inflammatory F1, whose denominator also changes under revision.

T1: indistinguishable evidence can conceal opposite utility. Consider two admissible worlds with disagreement mass d, conditional observable-law total variation at most ε, and pointwise conditional utilities at least γ and at most $-\gamma$ on their disagreement region. Any observation-only policy has worst-world regret at least $d\gamma(1 - \varepsilon)/2$ relative to the corresponding informed policy. The conditions concern the whole observable transcript, not merely matching average confidences. This is

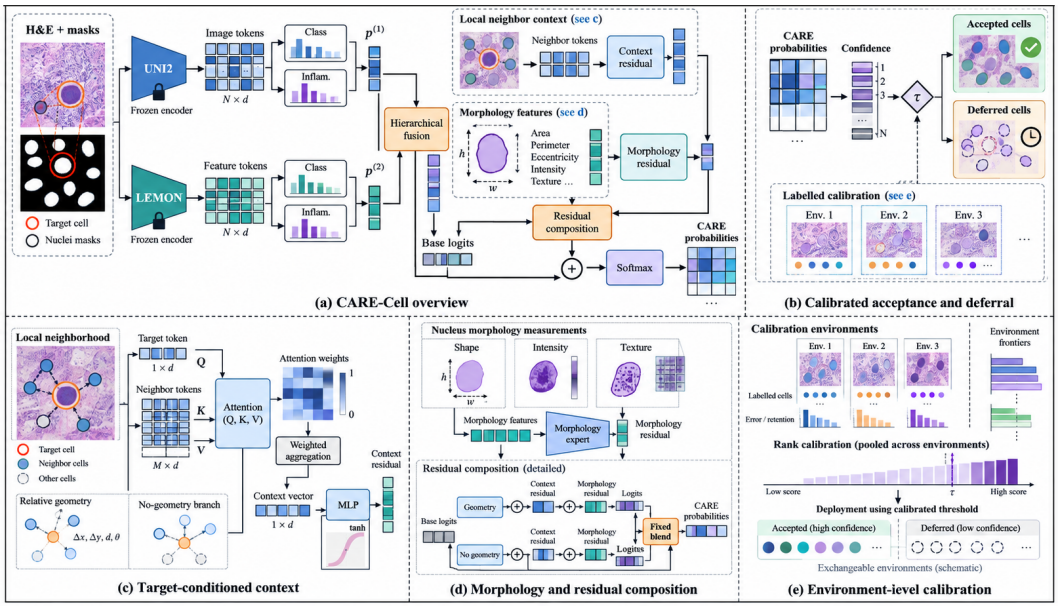


Figure 2: **Evidence construction and calibrated acceptance.** CARE-Cell builds corrective candidates, and the decision layer authorizes, retains, or defers them.

a two-point lower bound specialized to revision, related to domain-adaptation impossibility results (Ben-David et al., 2010); it is not a claim that every real disagreement is unresolvable. Proofs and the exact assumptions are in Appendix A.

T2: the accessible granularity limits recoverable gain. For an information field $\mathcal{G}$, the best $\mathcal{G}$-measurable policy achieves

$$O_{\mathcal{G}} = \mathbb{E}[(\mathbb{E}[U \mid \mathcal{G}])_+], \quad O_{\text{cell}} - O_{\mathcal{G}} = \tfrac{1}{2}\{\mathbb{E}|U| - \mathbb{E}|\mathbb{E}[U \mid \mathcal{G}]|\}. \tag{2}$$

Refining a *nested* information field cannot decrease this oracle. Adding a feature does not necessarily refine useful information about U, however; nor does replacing an expert isolate an information effect. This distinction explains why we separately evaluate evidence construction and revision on a frozen candidate bank.

3 Build Evidence, Then Authorize Revision

3.1 Stage I: Complementary Evidence Construction

CARE-Cell combines frozen tissue features u_i from UNI2 and nuclear features ℓ_i from LEMON (Chen et al., 2024; Mahmood Lab, 2025; Chadoutaud et al., 2026). A dual-view hierarchical head produces $p_i^{(0)}$. Context and explicit morphology evidence supply residual proposals rather than replacing the base prediction indiscriminately (Figure 2). A convenient interface is

$$p_i = \text{softmax}\left[\log(p_i^{(0)} + \epsilon) + \sum_{e \in \{c,m\}} \beta_i^e \widetilde{d}_i^e\right], \qquad \beta_i^e = \lambda_e a_i^e r_i^e. \tag{3}$$

Here $\widetilde{d}_i^e$ is a residual direction, a_i^e measures agreement or applicability, and r_i^e records evidence reliability or availability. In the ideal positive-probability form with $\epsilon = 0$, zero residual recovers $p_i^{(0)}$ exactly; an implementation with clipping must explicitly return the stored base to ensure exact fallback. These reliability weights are model features, *not statistical certificates*.

Target-conditioned context uses the target nucleus to query neighbour evidence. Private Anchor and public set-decoder banks implement this interface differently; coefficients and implementation

details are in Appendix B.8. Explicit relative geometry is an optional context configuration, not a necessary source of improvement.

Stage I improves predictions and exposes complementary candidates, but does not certify their target-environment utility. Stage II freezes this bank and uses sampled target outcomes to justify actions, rather than treating another transformation of the same features as independent evidence of benefit.

3.2 Stage II: target-supported utility learning and independent audit

Freeze a proposal before acquiring outcomes. Let f now be the completed CARE prediction. In the evaluated inflammatory task, the action pool consists of cells for which at least one frozen candidate disagrees with CARE on aggregated inflammatory membership. Among eligible candidates, choose the largest probability margin. This creates one fixed multiclass proposal h_i per eligible cell; neither the proposal nor pool uses target labels. Public and private candidate banks differ, an explicit limitation of cross-dataset mechanism comparisons.

Target-supported utility learning. Draw a uniform pilot from the full target population and additional nonadaptive, environment-stratified support from the fixed action pool. The pilot is part of, not additional to, the construction budget. It estimates the reference inflammatory F1, $\widehat{F}_0$. For each supported action, define

$$z_i = 2\Delta\mathrm{TP}_i - \widehat{F}_0\widehat{\Delta\mathrm{Pos}_i}. \tag{4}$$

If the coefficient were the true reference F1, a positive aggregate sum would characterize an F1 increase, provided its denominator is nonzero. Using the pilot makes this a learnable surrogate, not an F1 certificate. Unobserved labels are masked and cannot enter fitting, including the coefficient estimate.

Fit a pooled regressor $m_0(W)$ and one local regressor $m_e(W)$ per environment. Predict utility using

$$\widehat{z}_i = w_e m_e(W_i) + (1 - w_e)m_0(W_i), \qquad w_e = \frac{n_e}{n_e + 128}. \tag{5}$$

Here n_e counts support actions. We propose $\widehat{z}_i > 0$ actions and compare local-only and pooled-only ablations. Regressor settings are in Appendix B; shrinkage is an engineering choice, not an optimality theorem.

Independent utility audit. Condition on the bank, construction labels, and a resulting proposed set A of size M. Sample n audit actions uniformly without replacement, independently of their labels. If b corrections and h harms are observed, invert the hypergeometric distribution to obtain a lower bound $L_{\eta/2}(b)$ on total corrections and an upper bound $U_{\eta/2}(h)$ on total harms in A. The lower bound on utility in the *unaudited* actions is

$$L_Q = L_{\eta/2}(b) - U_{\eta/2}(h) - (b - h). \tag{6}$$

Deploy the complete remaining proposal only if $L_Q > 0$; otherwise retain CARE. Construction and audit cells also retain CARE and remain in all full-population metrics. Conditional on the fixed construction transcript,

$$\mathbb{P}\left(\text{accept and } \sum_{i \in A \setminus C} U_i \le 0\right) \le \eta. \tag{7}$$

This is a finite-pool unsafe-deployment bound, not a guarantee conditional on acceptance, for every environment, for F1, or for a future dataset. For the three fixed policies, $\eta = 0.05/3$ controls their union within one audit experiment. It does not control an unlimited sequence of experimental searches. The audit separates fitting and testing as in Angelopoulos et al. (2021), with exact finite-population sampling rather than an i.i.d. query assumption.

3.3 Environment-level acceptance and deferral

The final component addresses a different deployment decision: *when may the reference be accepted without further revision?* This is closely related to selective prediction and reject-option decision

rules (Chow, 1970; Geifman & El-Yaniv, 2017). Freeze f and a confidence score s independently of calibration and future environment bags. For bag S_e, define

$$D_e = \sum_{i \in S_e} \mathbf{1}\{f_i \neq Y_i\}, \quad R_e(t) = \frac{\sum_{i \in S_e} \mathbf{1}\{s_i > t\}\mathbf{1}\{f_i \neq Y_i\}}{D_e}, \tag{8}$$

with $R_e = 0$ if $D_e = 0$. This is the fraction of *reference errors* retained, not the error rate among accepted predictions. Let T_e be the $(\lfloor \rho D_e \rfloor + 1)$-st largest error score when this index exists, otherwise $-\infty$. Strict acceptance $s_i > t$ makes $t \geq T_e$ sufficient for $R_e(t) \leq \rho$, including ties.

T3: a scoped future-environment certificate. For n exchangeable calibration bags and one exchangeable future bag, set

$$k = \lceil (n+1)(1-\alpha) \rceil, \qquad \widehat{t} = T_{(k)}, \qquad \mathbb{P}\{R_{\text{new}}(\widehat{t}) \leq \rho\} \geq 1 - \alpha, \tag{9}$$

using $\widehat{t} = +\infty$ if $k > n$. This is an environment-frontier specialization of conformal quantile risk control (Angelopoulos et al., 2024), not a new rank theorem. A conservative upward projection onto a fixed threshold grid preserves the event; unrestricted grid-wise tests do not. We use $\rho = 0.60$ throughout, with $\alpha = 0.15$ for cell-classification acceptance and $\alpha = 0.20$ for the WILDS certificate experiments. A separate PAC tolerance order statistic increases assurance at a coverage cost (Appendix A.4).

Populations and composition. Eligibility ($D_e \geq 20$) must be applied symmetrically and does not certify low-error or unlabeled future bags. T3 is marginal, not hospital-conditional. We use independently fitted references for validation; ordinary OOF alone does not establish exchangeability. Revision and deferral are evaluated separately. Their compatible serial information-flow contract is specified, but not experimentally evaluated, in Appendix B.

4 Experiments: evidence, identifiable utility, and justified action

We first establish the correction problem, then evaluate evidence construction and authorization. Table 1 is the primary classification benchmark. Tables 2 and 3 are separate, within-protocol decision evaluations, not successive improvements to that benchmark. Their exact paired references appear in Appendix B.1. Cell endpoints are inflammation F1, macro-F1, and accuracy on ground-truth nuclei; Lizard pools four inflammatory subtypes (Graham et al., 2021), whereas PanNuke uses its inflammatory class (Gamper et al., 2020).

Evaluation populations. The consecutive dermatopathology cohort was collected over a fixed clinical period without appearance- or performance-based case selection: 927,535 development cells/41 slides and 130,476 validation cells/six slides. The public classification evaluation comprises 112,226 Lizard cells and 189,744 PanNuke cells. Acceptance uses independently refitted CARE models with disjoint fit/calibration/target folds. Camelyon17/iWildCam test the decision principles with task-specific references, not the CARE-Cell architecture (Koh et al., 2021). The private sealed test is not used.

4.1 Correction opportunity and cross-environment utility

Opportunity coexists with harm. Figure 3a is an outcome-selected illustration of the two possibilities, not a prevalence estimate; the complete case-selection rule is in Appendix B.7. Let B and H count wrong-to-correct and correct-to-wrong revisions. Against the matched dual-view reference, CARE yields 4,744/2,353 on Lizard and 4,349/3,045 on PanNuke (Figure 3b). Thus useful corrective evidence is present, but the same disagreement pool also contains harmful revisions. A label-using binary oracle applied after CARE retains upper-bound gains of 5.78/3.90/2.81 points on private/Lizard/PanNuke. This is a diagnostic ceiling, not an F1-optimal or deployable policy.

Environment and input diagnostics. All four Camelyon experts have mixed hospital-OOF slide utility: strong augmentation helps 9 slides and harms 20. Private rich-router fit AUC is 0.753,

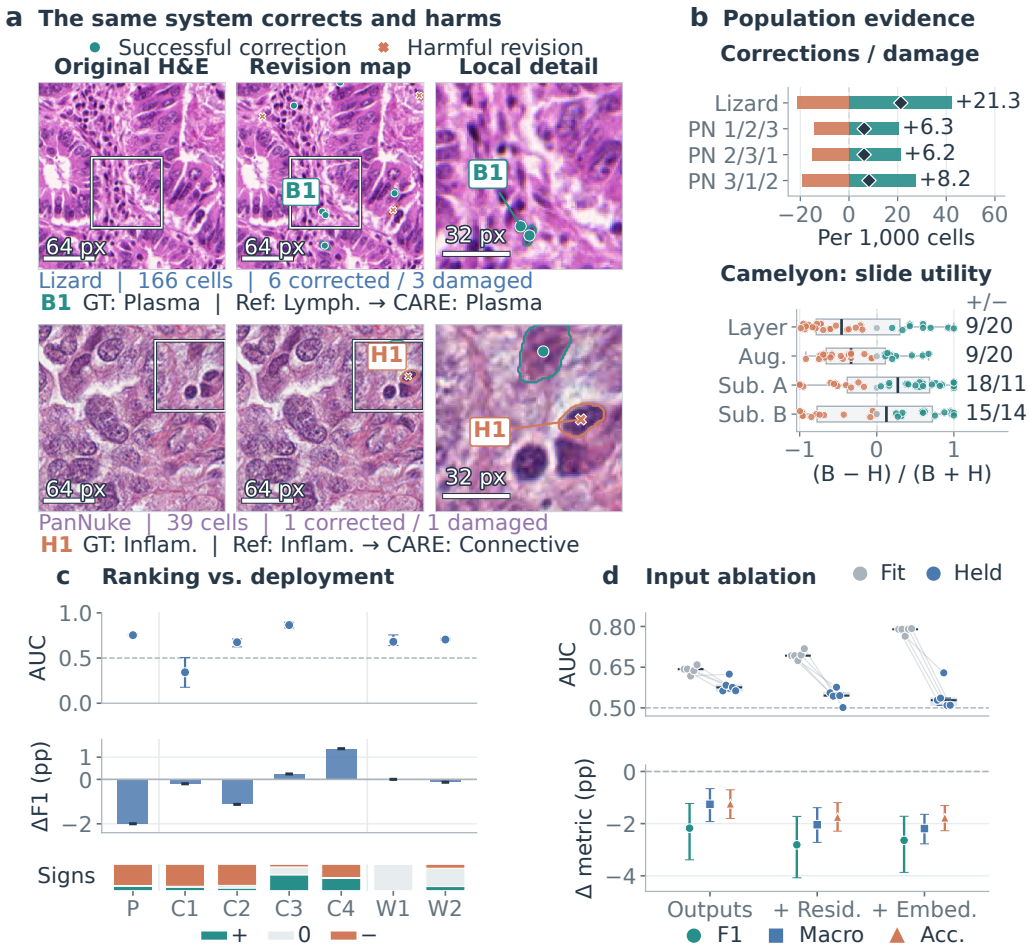


Figure 3: **Correction ability is not transferable utility.** Beneficial and harmful revisions coexist, and disagreement ranking does not determine deployment utility.

yet held-patient F1 falls 2.00 points and 33/41 environments have negative utility. iWildCam rich routing and Camelyon scalar OOD routing also lose F1, whereas the Camelyon hospital-OOF probability router gains 1.381 points. With private candidates and training fixed, all three input ablations lose F1 (Figure 3d; details in Appendix B.8). These diagnostics motivate T1 without proving indistinguishability on real data; T2 concerns additive utility, not cross-bank F1. In Figure 3, B1/H1 denote wrong-to-correct/correct-to-wrong examples; P, C1–C4, and W1–W2 denote private, Camelyon, and iWildCam diagnostics. Panel d reports input ablations; AUC ranges are not CIs and metric changes use patient-bootstrap intervals.

4.2 CARE-Cell constructs complementary evidence

Comparators and overall performance. Table 1 compares tissue-feature and LEMON controls, SSPP, NEP, KongNet, CytoFormer, Cell Graph Transformer, CellViT++, HoVer-Net, and CARE-Cell (Sirinukunwattana et al., 2016; Lv et al., 2026; Zhang et al., 2026; Lou et al., 2024; Chadoutaud et al., 2026; Hörst et al., 2026; Graham et al., 2019). All methods classify the same ground-truth nucleus instances under the dataset splits specified in Appendix B; the PanNuke aggregation convention for each row is identified in the table. CARE-Cell improves inflammation F1 over the UNI2-based classifier by 4.59/8.35/1.84 points on private/Lizard/PanNuke, and exceeds the matched Lizard CellViT++ classifier by 11.95 points. The corresponding detailed records are also provided in Appendix C.

Table 1: **Cell classification on three evaluation banks.** Percent scores for inflammation F1, macro-F1, and accuracy under the stated ground-truth-nucleus protocol.

	Private validation			Lizard			PanNuke		
Method	F1	Macro	Acc.	F1	Macro	Acc.	F1	Macro	Acc.
Single-patch CNN (SSPP)	65.78	55.94	62.14	75.15	44.45	68.52	69.73^{p}	63.33	68.09
CNN + NEP	69.02	59.46	62.93	77.39	46.13	71.02	70.45^{p}	64.30	69.08
LEMON	66.83	56.45	65.78	82.45	54.60	81.66	71.31^{r}	68.11	73.16
UNI2-based classifier	75.59	72.09	78.42	83.25	62.42	87.17	77.55^{r}	81.71	85.36
UNI2 + LEMON dual-view	77.70	72.36	79.78	87.33	65.83	<u>89.14</u>	<u>78.64^{r}</u>	<u>82.17</u>	<u>85.61</u>
KongNet	69.62	69.69	63.22	78.32	47.66	82.32	70.91^{s}	65.21	69.76
CytoFormer	<u>78.75</u>	<u>73.21</u>	<u>80.19</u>	<u>88.12</u>	<u>70.22</u>	88.75	77.92^{s}	79.21	85.43
CellViT++	71.23	65.42	71.83	79.65	49.56	81.08	70.48^{s}	64.12	69.17
Cell Graph Transformer	74.62	71.63	77.84	82.32	<u>68.71</u>	86.44	77.95^{s}	81.91	84.61
HoVer-Net (scratch)	67.48	62.17	71.26	85.22	55.95	84.74	64.58^{p}	59.91	70.17
CARE-Cell	**80.18**	**74.68**	**83.32**	**91.60**	**71.96**	**91.27**	**79.39^{r}**	**82.92**	**86.30**

r: arithmetic mean of rotation-level metrics; p: pooled cell predictions across rotations; s: separately supplied dataset-level score. PanNuke F1 values with different superscripts have different aggregation records; within each metric column, bold and underline denote the largest and second-largest displayed values, respectively. LEMON is a feature-head control.

Components and distribution of benefit. We distinguish configuration comparisons from controlled ablations (Figure 4). Panel a is illustrative, not causal; panels b–e report Lizard controls, component/rotation results, slide heterogeneity, and encoder portability. The private historical sequence changes heads as well as inputs and therefore supports the complete system, not a separate causal gain for every step. Five of six slides improve over dual view; the two lowest-reference-F1 slides average +8.81 points versus +0.79 elsewhere. In the Lizard A5 controls, removing centre–context interaction lowers F1 from 91.226 to 90.857, whereas removing geometry raises it to 91.560. Thus target–context interaction is supported in this comparison; geometry is optional, and no universal benefit is claimed for it.

Scope of the primary comparison. The controlled component analysis uses UNI2 as the primary tissue encoder. Figure 4e summarizes CARE minus encoder-only classifiers for UNI2, Virchow2, Prov-GigaPath, SAM2, and SAM3 across the three banks; detailed records are listed in Appendix C (Zimmermann et al., 2024; Xu et al., 2024; Ravi et al., 2025; Carion et al., 2025).

4.3 Available labels determine justified action

With target labels: construct locally, audit independently. The target-supported branch uses random target labels to construct a fixed proposal and an independent audit of disjoint labels, retaining CARE on rejection. Existing annotations simulate labels; no active learning is used. Across 100 draws per seed, full-population F1 gains are +1.57 points on private validation, +0.06 on Lizard, and 0 on PanNuke (Table 2). The audit certifies the specified aggregate multiclass additive utility on unaudited actions, not F1 and not per-slide benefit. Three private slides improve and three decline; S6 gains 11.49 points. The first-draw slide-bootstrap interval is [0.094,4.789], versus support-repeat range [1.488,1.642]. All PanNuke audits reject amid limited tile disagreement pools; zero executed actions do not establish zero opportunity.

Table 2: **Audited revision relative to the reference.** Full-population changes and audit outcomes; labels support construction and audit.

	ΔInfl. F1	ΔMacro	ΔAcc.	Audit pass	Labels used
Private	+1.57	+0.96	+0.97	100.00	3,072 + 256
Lizard	+0.06	+0.05	+0.05	55.80	2,643 + 221
PanNuke	0.00	0.00	0.00	0.00	2,909–2,935 + 373

Public deltas use multi-seed references, not the classification banks of Table 1; Appendix B.1 gives both endpoints. The near-zero changes delimit practical benefit: the audit limits unsupported revision, rather than promising F1 improvement. Rejection can reflect insufficient power and does not prove harm. Support/audit cells retain the reference in the full-population denominator.

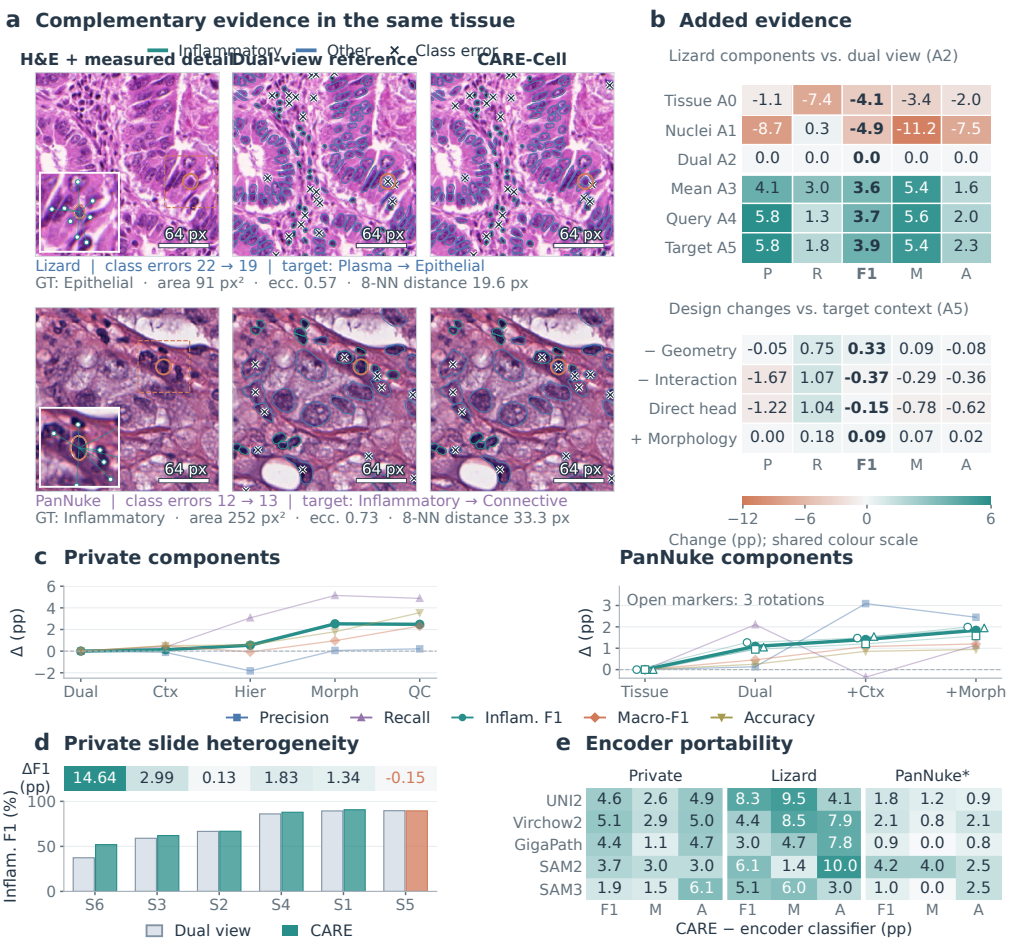


Figure 4: **CARE-Cell adds complementary evidence.** Controlled spatial, component, slide-level, and encoder-transfer analyses show where added evidence changes predictions.

Without target labels: accept or defer the reference. For target fold j, calibration is $(j + 1) \bmod 5$; a CARE model fitted on the other three folds predicts both. Heads, residuals, and preprocessing are refitted with frozen encoders. Five Lizard and fifteen PanNuke fits cover 366,315 and 189,744 cells. The score $|\sum_{c\in\mathcal{I}} p_c - \max_{c\notin\mathcal{I}} p_c|$, $\alpha = .15$, $\rho = .60$, and strict ties are frozen. T3 concerns binary inflammation error retention. Accepted-subset F1 improves in all twenty target folds (Table 3), an empirical quality result rather than the certified event.

Table 3 reports this separate acceptance protocol. Coverage is the accepted cell fraction; environment pass is the fraction satisfying $R_e \leq \rho$. The subset gap compares accepted cells with all cells from the *same refitted model*, not with Table 1. Intervals use 500 paired tile-bootstrap replicates; tiles are not patients. The table includes the private validation summary together with the two public evaluation banks.

Table 3: **Selective acceptance.** Scores are percent; subset gaps are percentage points with paired tile-bootstrap 95% CIs. Predictions are unchanged.

	All F1	Accepted F1	Subset gap (pp), 95% CI	Coverage	Env. pass
Private	81.88	90.05	8.17 [7.39,8.82]	91.24	88.75
Lizard	93.23	97.14	3.91 [3.79,4.04]	92.15	85.85
PanNuke	78.88	87.65	8.77 [8.39,9.17]	86.26	84.77

The practical outcome is selective automation: 91.24% of private, 92.15% of Lizard, and 86.26% of PanNuke cells remain accepted, with higher empirical subset F1. Deferred cells remain part of the cohort and require a separate workflow; no outcome for their subsequent review is assumed.

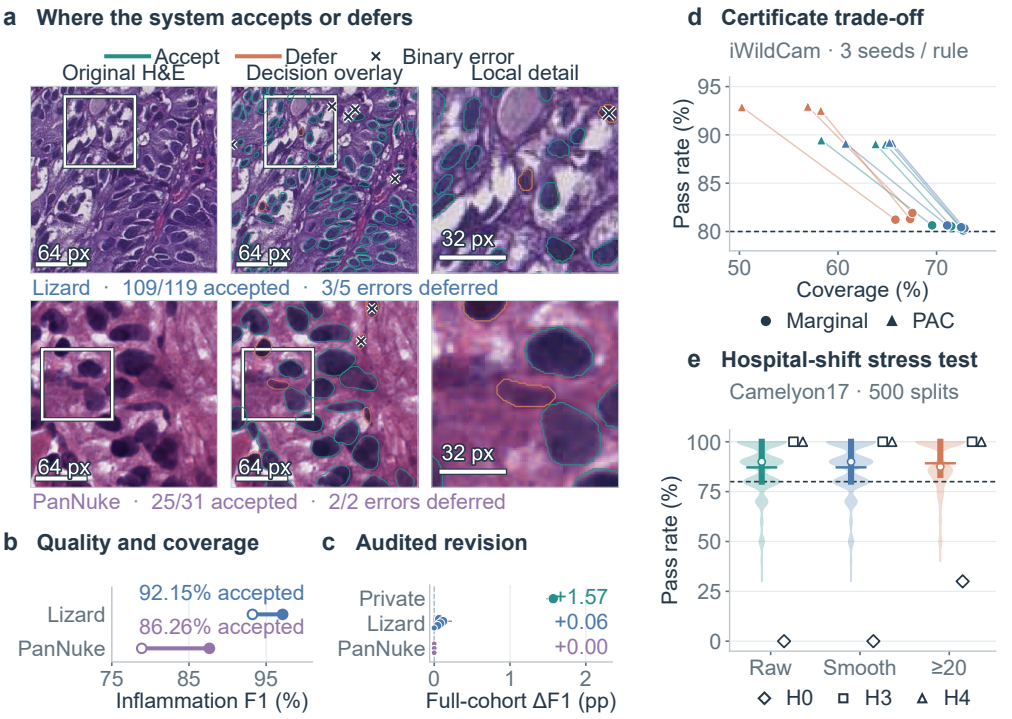


Figure 5: **Risk-controlled decisions have distinct endpoints.** Acceptance, audited revision, calibration, and shift stress tests report different endpoints.

This benefit is a quality–coverage trade-off, not additional full-population correction. Classification, audited revision, and acceptance therefore have distinct denominators and cannot be added as gains of a serial system. Figure 5a–b show accept/defer examples and subset quality, c audited-revision F1, and d–e calibration/shift stress tests; the dashed line is the nominal pass target, not an F1 guarantee. Raw, smoothed, and minimum-support are denominator rules.

Calibration and its boundary. Independent-fit iWildCam uses 150 fit cameras and 93 disjoint calibration/evaluation cameras, with 500 repeated 44/49 partitions per MLP seed. At $\alpha = .20$, the independent-fit theorem-validation pass rates are close to the nominal 0.80: 0.806 on Camelyon17 and 0.799 on iWildCam, with mean coverage reported separately. Figure 5d–e visualizes the coverage trade-off and the hospital-shift failure; the dashed 0.80 line is a certificate target, not an F1 guarantee. Reused OOF partitions are a sensitivity analysis, not an independent-fit certificate. Camelyon hospital-held-out stress testing falls to 0.7667 across hospitals (0.30, 1.00, 1.00), illustrating the boundary when exchangeability fails. Raw and smoothed denominator rules give similar sensitivity values; these are not substituted for the primary conformal statement.

5 Discussion and Conclusion

Our contribution is an information-dependent two-stage framework for deciding what complementary experts justify. CARE-Cell expands evidence and improves inflammation F1 across the three cell-classification cohorts, while target-supported audits authorize only measured aggregate utility and label-free calibration trades coverage for accepted-subset quality without changing predictions. Confidence-based ranking alone is not authorization under shift (Jitkrittum et al., 2023); the two-world construction makes this limitation explicit, while audit and conformal control provide scoped safeguards (Ben-David et al., 2010; Laroche et al., 2019; Angelopoulos et al., 2021; 2024). Sparse support, candidate coverage, and objective mismatch may limit public gains; grouped public tiles and six private validation slides do not establish patient-level generalization. We therefore claim scoped safeguards, not universal correction or clinical readiness.

REPRODUCIBILITY STATEMENT

The appendices in this PDF provide the formal assumptions and proofs, model families, label budgets, sampling seeds, aggregation rules, and panel definitions. Appendix B.1 records the distinct frozen-reference endpoints so that the classification, revision, and acceptance evaluations can be interpreted independently. Reproducing training requires the original feature banks, checkpoints, loaders, and authorized access to private data; these assets are not distributed within this PDF.

AI USE STATEMENT

A generative AI coding and writing assistant was used to organize this manuscript, draft prose and mathematical exposition, inspect experiment artifacts, generate plotting and LaTeX code, and assist with experimental design and implementation. The authors are responsible for the scientific claims, proofs, citations, data permissions, and final manuscript.

ETHICS STATEMENT

The pathology-image analysis project underlying the private cohort was approved by the contributing institution's medical research ethics committee (approval no. [2021]2020-196-2; revised protocol version 2.0). The institution name is withheld for double-blind review. This work is a retrospective computational evaluation, not a clinical intervention or a validated diagnostic system. Target label acquisition is simulated using existing annotations. Patient identifiers are not included in this manuscript; public tile IDs are not treated as patient identities. Private images and clinical metadata are not publicly released. Classification metrics and selective acceptance do not establish clinical benefit or replace professional clinical judgment.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021. URL https://arxiv.org/abs/2110.01052.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pp. 129–136, 2010. URL https://proceedings.mlr.press/v9/david10a.html.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. URL https://arxiv.org/abs/2511.16719.

Loïc Chadoutaud, Alice Blondel, Hana Feki, Jacqueline Fontugne, Emmanuel Barillot, and Thomas Walter. LEMON: a foundation model for nuclear morphology in computational pathology. *arXiv preprint arXiv:2603.25802*, 2026. URL https://arxiv.org/abs/2603.25802.

Richard J. Chen, Tong Ding, Ming Y. Lu, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30:850–862, 2024. doi: 10.1038/s41591-024-02857-3. URL https://www.nature.com/articles/s41591-024-02857-3.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406. URL https://doi.org/10.1109/TIT.1970.1054406.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benes, Simon Graham, Mostafa Jahanifar, Syed Ali Khurram, Ayesha Azam, Katherine Hewitt, and Nasir Rajpoot. PanNuke dataset extension, insights and baselines. *arXiv preprint arXiv:2003.10778*, 2020. URL https://arxiv.org/abs/2003.10778.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017. URL https://papers.nips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf.

Simon Graham, Quoc Dang Vu, Shan E. Ahmed Raza, Ayesha Azam, Yee-Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. HoVer-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019. doi: 10.1016/j.media.2019.101563. URL https://doi.org/10.1016/j.media.2019.101563.

Simon Graham, Mostafa Jahanifar, Ayesha Azam, Mohammed Nimir, Yee-Wah Tsang, Katherine Dodd, Emily Hero, Harvir Sahota, Atisha Tank, Ksenija Benes, et al. Lizard: A large-scale dataset for colonic nuclear instance segmentation and classification. *arXiv preprint arXiv:2108.11195*, 2021. URL https://arxiv.org/abs/2108.11195.

Fabian Hörst, Moritz Rempe, Helmut Becker, Lukas Heine, Julius Keyl, and Jens Kleesiek. CellViT++: Energy-efficient and adaptive cell segmentation and classification using foundation models. *Computer Methods and Programs in Biomedicine*, 277:109206, 2026. doi: 10.1016/j.cmpb.2025.109206. URL https://doi.org/10.1016/j.cmpb.2025.109206.

Wittawat Jitkrittum, Neha Gupta, Aditya K. Menon, Harikrishna Narasimhan, Ankit Rawat, and Sanjiv Kumar. When does confidence-based cascade deferral suffice? In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1f09e1ee5035a4c3fe38a5681cae5815-Abstract-Conference.html.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pp. 3652–3661, 2019. URL https://proceedings.mlr.press/v97/laroche19a.html.

Wei Lou, Guanbin Li, Xiang Wan, and Haofeng Li. Cell graph transformer for nuclei classification. *arXiv preprint arXiv:2402.12946*, 2024. URL https://arxiv.org/abs/2402.12946.

Jiaqi Lv, Esha Sadia Nasir, Kesi Xu, Mostafa Jahanifar, Brinder Singh Chohan, Behnaz Elhaminia, and Shan E. Ahmed Raza. KongNet: a multi-headed deep learning model for detection and classification of nuclei in histopathology images. *Medical Image Analysis*, 114:104233, 2026. doi: 10.1016/j.media.2026.104233. URL https://doi.org/10.1016/j.media.2026.104233.

Mahmood Lab. UNI2: Official model release and model card. Official model repository, 2025. URL https://github.com/mahmoodlab/UNI. UNI2 release, January 2025; accessed September 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha6RTeWMd0.

Korsuk Sirinukunwattana, Shan E. Ahmed Raza, Yee-Wah Tsang, David R. J. Snead, Ian A. Cree, and Nasir M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1196–1206, 2016. doi: 10.1109/TMI.2016.2525803. URL https://doi.org/10.1109/TMI.2016.2525803.

Han Xu, Naoto Usuyama, Jinfeng Bagga, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188, 2024. doi: 10.1038/s41586-024-07441-w. URL https://doi.org/10.1038/s41586-024-07441-w.

Zhi Zhang et al. Cytoformer: A molecularly supervised cell foundation model for histopathology cell classification. *arXiv preprint arXiv:2608.16718*, 2026. URL https://arxiv.org/abs/2608.16718.

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024. URL https://arxiv.org/abs/2408.00738.

A FORMAL STATEMENTS AND PROOFS

A.1 T1: A REVISION-SPECIFIC TWO-WORLD BOUND

Proposition. *Let P_+ and P_- be two admissible joint laws of (W, U), and let D be an observable disagreement event with $P_+(D) = P_-(D) = d$. On D, suppose versions of the conditional mean utility satisfy $\mu_+(w) \geq \gamma$ P_+-almost surely and $\mu_-(w) \leq -\gamma$ P_--almost surely. Assume $\mathrm{TV}(P_+(W \mid D), P_-(W \mid D)) \leq \varepsilon$. For any randomized observation-only policy $\pi : W \to [0, 1]$, its worst-world regret on D, relative to the informed positive-part utility policy, is at least $d\gamma(1 - \varepsilon)/2$.*

Proof. Write $a_\pm = \mathbb{E}_\pm[\pi(W) \mid D]$. In the positive world, declining incurs regret at least $d\gamma(1-a_+)$. In the negative world, revising incurs regret at least $d\gamma a_-$. Their maximum is at least half their sum. Since $\pi \in [0, 1]$, the TV bound gives $a_+ - a_- \leq \varepsilon$, completing the claim. Any additional nonnegative regret outside D cannot weaken the bound. □

The pointwise mean-sign condition is essential: opposite *pooled* signs alone do not suffice, since a policy might select observable beneficial subregions. The observation law must include training information and any unlabeled target adaptation transcript actually available to the learner. In the limiting case $\varepsilon = 0$, no observation-only policy can distinguish the two worlds. Additional target labels change the information set and are not covered by this no-new-outcome restriction.

For a concrete label-budget illustration, suppose a target stratum has utilities in $\{-1, +1\}$ with $\mathbb{P}_+(U = 1) = (1 + \gamma)/2$ and the reverse under P_-. With k independent labels and equal prior weights, the optimal testing error is

$$e_k = \mathbb{P}\{Z < k/2\} + \tfrac{1}{2}\mathbb{P}\{Z = k/2\}, \qquad Z \sim \mathrm{Binomial}(k, (1 + \gamma)/2). \tag{10}$$

The corresponding symmetric minimax regret is $d\gamma e_k$. This is the standard likelihood-ratio test, not a distribution-free budget promise for our heterogeneous data. Sampling from a finite target pool would instead require its without-replacement law.

A.2 T2: GRANULARITY, TRANSPORT, AND FINITE-LABEL ESTIMATION

Proposition. *For integrable $U \in [-1, 1]$ and information field $\mathcal{G}$, the optimum among $\mathcal{G}$-measurable randomized policies is $O_\mathcal{G} = \mathbb{E}[(\mathbb{E}[U \mid \mathcal{G}])_+]$. If $O_{\mathrm{cell}} = \mathbb{E}[U_+]$, the gap is Equation 2. For nested fields $\mathcal{G} \subseteq \mathcal{H}$, $O_\mathcal{G} \leq O_\mathcal{H}$.*

Proof. Conditional expected utility is $\pi\mu$ where $\mu = \mathbb{E}[U \mid \mathcal{G}]$. Choosing $\pi = \mathbf{1}\{\mu > 0\}$ maximizes it pointwise. Applying $x_+ = (x+|x|)/2$ and the tower property yields the gap. Conditional Jensen's inequality for the convex positive-part map yields monotonicity under nested information fields. □

For a source-derived sign policy π_s and target mean utility $\mu_t(g)$, target regret satisfies the exact identity

$$O_{\mathcal{G},t} - G_t(\pi_s) = \mathbb{E}_t\left[|\mu_t(G)|\mathbf{1}\{\pi_s(G) \neq \mathbf{1}\{\mu_t(G) > 0\}\}\right] \tag{11}$$

for deterministic sign decisions, with a corresponding weighted expression for randomized decisions. Good source ranking does not bound this target sign error without an assumption relating source and target utility.

Proposition. *Fix M disjoint strata with population weights p_c. In each stratum, observe $k_c > 0$ independent labels of $U \in [-1, 1]$, independently of a future evaluation draw, and use the sign of its sample mean. With probability at least $1 - \delta$, the regret relative to the stratum oracle is at most*

$$\sum_{c:k_c>0} p_c \min\left\{1, \sqrt{\frac{2\log(2M/\delta)}{k_c}}\right\} + \sum_{c:k_c=0} p_c. \tag{12}$$

Proof. Hoeffding's inequality and a union bound make every sampled stratum mean accurate to its displayed radius simultaneously. A wrong sign then implies $|\mu_c|$ is no larger than that radius. Bound unsampled stratum regret by one and weight by p_c. □

The strata and their meanings are fixed before these labels. Adaptive partitioning, insufficient per-stratum labels, or target/source shifts invalidate this simple use. The result is a standard estimation bound, included to clarify the granularity–sample-size tradeoff rather than as a new rate.

A.3 Independent finite-query utility audit

For clarity, the F1 fitting surrogate and the audited utility are distinct. Let $D_0 = Y_+ + \widehat{Y}_+$, $F_0 = 2\mathrm{TP}_0/D_0$, and let an action collection change true positives by a and predicted positives by b. Then, for positive denominators,

$$F_{\mathrm{new}} - F_0 = \frac{2a - F_0 b}{D_0 + b}. \tag{13}$$

This motivates Equation 4. Replacing F_0 with an estimated pilot coefficient introduces estimation error; moreover, multiclass correction utility is not $2a - F_0 b$. The audit therefore does not certify this F1 objective.

For an observed hypergeometric count b from n draws in a population of M, choose a one-sided lower confidence limit $L_{\eta/2}(b)$ satisfying $\mathbb{P}(L_{\eta/2}(b) > B_A) \le \eta/2$ for every population correction count B_A. Define the analogous upper limit for the harm count H_A. By a union bound, both limits cover with probability at least $1 - \eta$, without assuming independence between correction and harm counts. On their joint coverage event,

$$\sum_{i \in A \setminus C} U_i = (B_A - H_A) - (b - h) \ge L_Q. \tag{14}$$

Thus acceptance when $L_Q > 0$ implies positive query utility on this event. Empty proposals, zero audit evidence, and fully audited proposals have no certified remaining gain and use fallback. The proof conditions on all construction choices; queries must not subsequently be trimmed using audit outcomes. Auditing several policies permits a union bound over the declared family, not post hoc search over unspecified variants.

A.4 T3: environment-frontier rank certificate

Condition on an independently fitted model and score. Let $S_1, \ldots, S_n, S_{n+1}$ be exchangeable complete environment bags. The common bag-to-frontier mapping makes $T_1, \ldots, T_{n+1}$ exchangeable. With randomized tie ranks, the future rank is uniform; nonrandomized conservative thresholding only improves coverage. Hence

$$\mathbb{P}\{T_{n+1} > T_{(k)}\} \le \frac{n + 1 - k}{n + 1} \le \alpha. \tag{15}$$

When $T_{n+1} \le \widehat{t}$, at most $\lfloor \rho D_{n+1} \rfloor$ reference-error scores exceed the threshold. This proves Equation 9. If $k > n$, all-defer makes retained burden zero. This is a direct conformal quantile-risk-control argument (Angelopoulos et al., 2024).

Why no extra grid correction is needed here. The threshold is one order statistic of a single scalar frontier per bag. Rounding it upward to a predeclared grid only shrinks acceptance and preserves $R_e \le \rho$. This is not an argument that arbitrarily testing every coverage level at nominal α is valid. A genuine fixed-sequence testing alternative would need valid tests and a stop-at-first-failure rule; that is not the proof used here.

Eligibility and denominators. Under i.i.d. sampling of environment bags, the same deterministic predicate $D_e \ge 20$ applied to every bag yields i.i.d. accepted bags from the conditional eligible-bag law (conditional on their count); the rank argument applies to a future eligible bag. For more general exchangeable designs the conditioning and symmetric sampling protocol must preserve exchangeability jointly. Filtering does not give an all-environment guarantee. Because future D_e depends on outcomes, eligibility is not an unlabeled deployment detector. The smoothed alternative $(H + 1)/(D + 2)$ is a different loss. Low-denominator absolute risks must be reported separately if an all-environment deployment statement is desired.

Marginal is not conditional. Equation 9 averages over calibration and the next environment. It does not assert $\mathbb{P}(R_{\text{new}} \leq \rho \mid S_1, \ldots, S_n) \geq 1 - \alpha$ for every calibration realization, nor that a finite evaluation batch passes with frequency at least $1 - \alpha$. Those are distinct events. A hospital-held-out shift need not follow the calibrated bag distribution; exchangeability and support assumptions must be justified, not inferred from the word OOF.

PAC tolerance variant. Under i.i.d. frontiers, choose the smallest $k \leq n$ satisfying

$$\mathbb{P}\{Z \leq k - 1\} \geq 1 - \delta, \quad Z \sim \text{Binomial}(n, 1 - \alpha), \tag{16}$$

and use all-defer if there is no such k. The order-statistic tolerance result gives, with calibration probability at least $1 - \delta$, future frontier coverage at least $1 - \alpha$. Ties are conservative. This stronger claim still does not require every finite evaluation fraction to exceed $1 - \alpha$. The iWildCam table uses $\delta = 0.05$.

B Protocols, costs, and evidence accounting

B.1 Bank identity and evaluation units

Table 1 is the sole primary classification comparison. The decision experiments answer different questions and retain their own paired baselines, summarized below. No result is rebased by replacing its measured reference score with a score from another protocol. Scores are percentages.

Protocol	Dataset	Reference/all F1	Revised/accepted F1	Change (pp)
Primary classification	Private	80.18	–	–
	Lizard	91.60	–	–
	PanNuke	79.39	–	–
Target-supported revision	Private	80.18	81.75	+1.57
	Lizard	91.92	91.98	+0.06
	PanNuke	79.50	79.50	0.00
Independent-refit acceptance	Lizard	93.23	97.14	+3.91
	PanNuke	78.88	87.65	+8.77

Revision changes compare predictions on the same full population. Acceptance changes compare an accepted subset with the full population of the same refitted reference; predictions do not change. The Lizard acceptance population has 366,315 cells, rather than the 112,226 cells used for the classification and revision evaluations. PanNuke acceptance uses 189,744 cells with refitted models and separate calibration folds. These are not sequential gains over the primary benchmark.

Private acceptance: descriptive record only. The private acceptance summary reports all-cell F1 81.88, accepted F1 90.05, a subset gap of 8.17 points with interval [7.39,8.82], coverage 91.24%, and environment pass 88.75%. These values are reported alongside the public acceptance results as a supplementary endpoint.

The Stage I private headline is the fixed Anchor 0.40, hierarchical 0.35, morphology 0.35 bank after coordinate QC. The development OOF and reused validation endpoints are separate. Public Stage I results use the historical target-query set decoder plus morphology banks. The later target-supported reference bank includes five Lizard and three PanNuke reference seeds; its PanNuke metrics pool held-out rotation predictions within each seed before averaging seeds, unlike the earlier arithmetic mean of rotation-level scores. F1 is nonlinear, so these values cannot be silently interchanged.

The private target-supported candidates are dual-view, hierarchical, Anchor, and pre-morphology predictions. Public candidates are the archived full and no-geometry-plus-morphology variants (including their seed variants). Their input probabilities are not identical evidence sources to the private residuals. The transferred object is the utility-learning and audit procedure, not a perfectly matched expert mechanism. Public block/tile identities are available; patient identity is not established by those IDs. All claims about these public environments therefore use the term tile.

The pilot is a simple random sample from all cells. Additional support is allocated proportionally to each environment's available action capacity, with a largest-remainder allocation and at least

Table 4: Target-support cost per adaptation draw and single frozen policy. Pilot labels are included in construction. Public acquisition is simulated from existing labels.

	Private	Lizard	PanNuke
Cells	130,476	112,226	189,744
Environment units	6 slides	698 tiles	7,558 block/tiles
Planned construction labels	3,072	2,643	4,468
Actual construction labels	3,072	2,643	2,909–2,935
Included pilot labels	768	661	1,117
Audit labels / policy	256	221	373
Frozen reference seeds	1	5	3
Support repeats / seed	100	100	100

one action retained for query where possible. This is fixed-pool nonadaptive stratified sampling, not active learning. PanNuke's sparse tile pools cannot fill the intended budget under that rule. Support and audit actions retain CARE and are included in the full N denominator. Running all three ablations consumes their respective audit labels in addition to the shared construction labels; the table gives the cost of one deployed policy, not total research labeling.

B.2 Target-supported revision implementation and randomness

The utility regressor is `HistGradientBoostingRegressor`: squared loss, learning rate 0.05, 100 iterations, maximum seven leaves, minimum leaf size 20, and L_2 regularization 10. The pooled model trains on supported action cells; local models use the corresponding environment subset. The feature interface includes reference and expert probabilities, margins, entropies, predicted classes, chosen-candidate identity, and disagreement summaries; the private bank additionally uses its archived reliability/morphology proposal features. Query labels are represented by a sentinel and are asserted absent from fitting. A label-masking self-test changes all unobserved labels and requires identical fitted selections.

The sampling base seed is 20260926, with fixed index offsets for seed, repeat, and audit policy. Reference seeds on public banks are 20260826–20260830 for Lizard and 20260826–20260828 for PanNuke. No best-seed selection is used. The primary policy is shrinkage; local and pooled are fixed ablations. Audit level $0.05/3$ is assigned per policy in a single repetition/reference setting. The earlier 0.81794 result used a full-target-label coefficient and is not a headline result; the reported target-supported protocol estimates the coefficient only from pilot labels. These protocols differ in allocation and audit level, so they are not a one-variable ablation.

B.3 Detailed target-supported controls and uncertainty

Table 5: All predeclared utility policies after their own audits. Deltas in percentage points; B, H are mean full-population multiclass corrections and harms.

Dataset	Policy	ΔF1	ΔAcc.	ΔMacro	B	H
Private	Local	1.56	0.98	1.04	2115.44	833.51
	Pooled	1.24	0.72	0.51	1426.87	491.80
	Shrink	1.57	0.97	0.96	1961.66	700.48
Lizard	Local	0	0	0	0	0
	Pooled	0.06	0.05	0.05	131.38	77.53
	Shrink	0.06	0.05	0.05	129.82	75.22
PanNuke	Local	0	0	0	0	0
	Pooled	0	0	0	0	0
	Shrink	0	0	0	0	0

The private shrinkage gain exceeds local-only by 0.0059 points, too small to present as the central empirical advance. The evidence instead supports local target information relative to purely pooled

utility estimation. Per-slide gain is heterogeneous: in Figure 5, S1–S6 follow the archived slide order, with changes -0.00950, $+0.97295$, -0.02824, -0.06271, $+0.35187$, and $+11.48747$ points. Pseudonyms avoid distributing original identifiers in the manuscript.

Support-repeat percentiles condition on the same target population and frozen model. The repeat-zero cluster bootstrap resamples environments, not support draws, and does not establish generalization from six slides. Public tile bootstrap intervals are not patient-level confidence intervals. A degenerate interval $[0, 0]$ after all-fallback means that realized prediction vector equals CARE; it is not infinitely precise evidence about the effect of a hypothetical accepted revision.

B.4 Diagnostic replay and remaining opportunity

All fixed first-five-repeat replays per seed were checked against original support hashes, query hashes, certificates, and post-audit metrics. They add diagnostics but do not change the model or sampling. The binary cell-label oracle corrects only CARE binary errors fixable by the frozen proposal. Support-excluded oracles retain reference predictions on construction cells but do not yet remove policy-specific audit cells. They are neither all-expert upper bounds nor a multiclass safe policy.

The public failure analysis has four distinct observations. First, sparse local support makes nearly all local models constant. Second, audit samples occupy about 36.4% and 48.1% of proposed actions on Lizard and PanNuke, compared with 7.1% privately. Third, PanNuke's pre-audit mean B/H is approximately $337.60/358.40 < 1$. Fourth, Lizard's aggregate inflammatory endpoint does not resolve internal immune subtype errors. These observations motivate future controlled experiments, but do not identify a single causal explanation for public transfer failure.

B.5 iWildCam conformal refit

The training-only camera inventory contains 243 environments. A predeclared deterministic camera ordering assigns 150 to fitting and 93 to calibration/evaluation, with no fitting-camera overlap. Frozen DINOv2 features (Oquab et al., 2023) are L_2 normalized; an unweighted MLP head is fitted separately for seeds 20262001, 20262002, and 20262003. The acceptance score is the top-two probability margin. The task retains 182 output classes; 142 classes appear in the fitting subset. For each seed, 500 fixed random partitions assign 44 of the remaining cameras to calibration and 49 to evaluation. The summary bundle records no checkpoint selection on those remaining environments and no test access. The displayed sensitivity comparison includes raw-all-bag, smoothed, and eligible-bag rules.

B.6 Information-flow contract for a compatible serial implementation

The following specifies how the components *could* be composed without changing the population after seeing audit labels. It is not an additional measured experiment.

1. Fit and freeze the evidence module, candidate bank, and acceptance score using source training data. Maintain a complete information-access log.
2. Calibrate the environment threshold on independent calibration bags. On a target bag, mark reference acceptance and deferral before any utility audit.
3. Construct a fixed disagreement proposal inside the deferred pool. Draw construction support, estimate the pilot coefficient, and fit local/pooled utility predictors using only that support.
4. Freeze the resulting positive-score proposal. Draw independent uniform audit labels from that exact proposal population and apply the remaining-pool certificate.
5. Revise only unaudited proposed actions when certified. Otherwise retain the reference, marking deferred-but-unrevised cases as unaccepted rather than implying a certificate they did not receive. Retain all cells in whole-population performance metrics.

The target-supported results in this paper skip Step 2 and use the full binary-disagreement pool. Restricting to conformal deferrals changes support, proposals, and audit power, so those results cannot be carried over without rerunning the protocol.

B.7 FIGURE SEMANTICS AND SPATIAL-CASE PROVENANCE

Compact-caption reading key. Table 2 averages 100 draws per seed (one private, five Lizard, three PanNuke); changes are computed before rounding. Table 3 uses separately refitted CARE models; Appendix B.1 records their all-cell F1 values. These references are not interchangeable with the primary classification models. Tiles are not patients, and target labels do not determine acceptance thresholds.

Figure 1c is a Camelyon17 slide-OOF diagnostic, not a hospital-held-out result. In Figure 3a, circles and crosses mark wrong-to-correct and correct-to-wrong multiclass revisions; GT and Ref denote ground truth and the dual-view reference. Panel b reports changes per 1,000 cells (diamonds: net; PN: split rotation), plus 120 expert–slide points with median/IQR boxes and 1.5-IQR whiskers. Panel c combines different scopes: P, private rich router; C1/C2, Camelyon scalar OOF/OOD; C3, non-nested rich OOD; C4, probability hospital-OOF (AUC unavailable); W1/W2, iWildCam scalar/rich. AUC spans are min–max ranges, not confidence intervals. Panel d has five fit/held pairs per input and patient-bootstrap 95% CIs. Figure 4b compares A0–A5 with A2 and A6–A9 with A5; it is not a sequential ladder. Panel c uses private component and PanNuke rotation-mean changes, with open points for individual rotations. Figure 5a uses label-blind illustrative cases. Its audit pass rates concern additive utility; displayed F1 changes are descriptive, not certified F1 improvements.

Figure 4a uses two fixed, label-blind ROI tiles from the stored public evaluation banks. The displayed nucleus is the closest-to-centre eligible instance in each tile; it was not selected by maximizing a metric or by inspecting the plotted CARE output. The small ellipse and nearest-neighbour marks are recomputed descriptive measurements from the stored masks and centroids. They are not network attention maps, the exact logged neighbourhood, or a causal attribution. The Lizard example changes a reference Plasma prediction to the epithelial ground truth; the PanNuke example changes an inflammatory reference prediction to connective and is therefore a harmful multiclass revision. The quantitative ablations in panels b–e, rather than these two illustrative tiles, support the evidence-construction claim. Figure 3a is similarly outcome-stratified and is not a prevalence estimate; its B1/H1 markers denote benefit and harm cases, respectively. Figure 5 uses an independently stored spatial export for the accept/defer visualization, while its quantitative panels report the stated tile and environment summaries.

B.8 EXPANDED IMPLEMENTATION AND EXPERIMENTAL DETAILS

These details accompany the compact main-text presentation; prediction banks and numerical results are unchanged.

Target-conditioned context queries neighbor tokens with attention proportional to $\exp(q_i^\top k_{ij}/\sqrt{d})$. The private bank instead uses frozen Anchor consensus: Anchor coefficient 0.40, hierarchical expert 0.35, morphology 0.35. Public banks use trained set decoders and validation-selected morphology weights. The evidence interface is shared; the networks are not identical.

Here n_e counts support actions. Fixed-capacity tree regressors use a local constant below 20 labels; implementation details are in Appendix B. We propose $\widehat{z}_i > 0$ actions and compare fixed local-only and pooled-only ablations. Shrinkage is an engineering choice, not a new optimality theorem.

Restricting to bags with $D_e \geq 20$ symmetrically preserves an eligible-bag interpretation under the stated sampling assumptions; it does not certify low-error bags, nor identify future eligibility without labels. The guarantee is marginal over calibration and a future environment, not conditional on a particular hospital. Ordinary OOF predictions are not automatically interchangeable with an independently fitted frozen model. Our primary T3 experiment uses disjoint fitting environments. In a serial implementation, freeze the conformal mask before constructing and auditing revisions within the deferred pool, and audit that exact action population independently. We evaluate the ingredients separately, not this unrun serial composition. A deferred cell can retain CARE as an operational fallback, but that prediction is not thereby certified as accepted.

All four Camelyon experts have mixed slide utility under hospital-OOF; strong augmentation helps 9 slides, harms 20, and ties on one. A private rich router reaches fit-set AUC 0.753, yet held-patient F1 falls 2.00 points and 33/41 environments have negative utility. The iWildCam camera-OOF AUC is 0.705 with a -0.128 macro-F1 change. Camelyon's scalar OOD diagnostic has AUC 0.675 and a

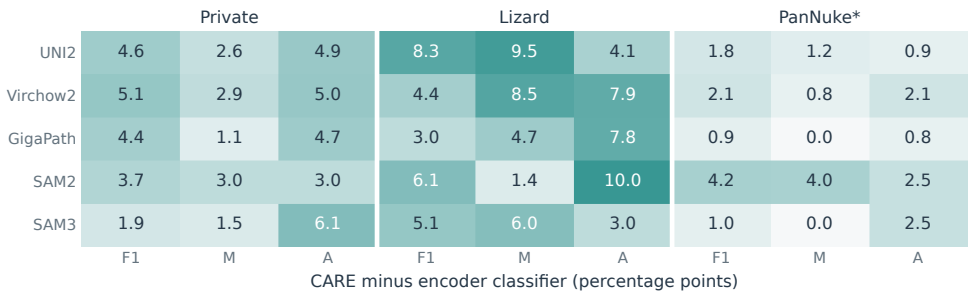


Figure 6: Supplementary encoder portability records. Each cell is CARE minus the corresponding encoder classifier in percentage points; M denotes macro-F1 and A denotes accuracy.

-1.125 positive-F1 change, whereas its hospital-OOF probability router gains 1.381 points. In other words, ranking disagreement is not authorization. Holding private candidates, folds, optimization, and thresholds fixed, the input ablations in Figure 3d lose 2.174–2.812 F1 points, with all nine metric-change intervals excluding zero. These observations motivate T1 without claiming to prove its indistinguishable-world premise on real data; T2 concerns fixed additive utility rather than these cross-bank F1 comparisons.

Private F1 rises from 75.589 to 76.738 with LEMON and reaches 80.177 after the locked dual-view, context, hierarchy, morphology, and coordinate-QC sequence. The component curves are controlled historical fits rather than a claim that every row is a one-variable ladder. The spatial distribution is heterogeneous: five of six private slides improve over the dual view, S6 gains 14.64 points, and the two lowest-reference-F1 slides average +8.81 points versus +0.79 for the others. Public paired tile-bootstrap gains are +4.270 [3.933,4.620] and +0.750 [0.506,0.989]; Table 1 reports rotation means. Lizard controls further show that removing centre–context interaction reduces F1 from 91.226 to 90.857, direct classification scores 91.076, and removing geometry reaches 91.560. These results support complementary evidence, not a universal claim that every residual module helps every tile.

Encoder configurations and comparator records are provided in Appendix C.

B.9 Artifact provenance and scope

The retained analysis archive contains target-supported summaries, diagnostic summaries, the CARE evidence map, independent-fit T3 summaries, and plotting source. The three target-supported metric files have SHA256 prefixes `cd3fbed297ff`, `27152a84713e`, and `dfdcd52852d9` for private, Lizard, and PanNuke. Identifying remote paths are excluded from this review PDF. The approved conceptual and experimental figures are preserved.

These are retrospective research evaluations. The private sealed test is not used. The results support the stated classification and information-dependent decision analyses, not prospective clinical deployment. The private cohort is not publicly distributed, and access to its source data is controlled by the contributing institution.

C Additional classification results

Scores are percentages. The alternative-encoder entries are point estimates from their corresponding evaluation records. Their aggregation rules are listed below.

C.1 PRIVATE VALIDATION: ENCODER CLASSIFIER VERSUS FULL CARE

Encoder	System	Infl. P	Infl. R	Infl. F1	Macro-F1	Accuracy
UNI2	Classifier	82.236	69.937	75.589	72.092	78.416
UNI2	Full CARE	77.912	82.577	80.177	74.677	83.322
Virchow2	Classifier	78.621	66.701	72.172	69.232	76.210
Virchow2	Full CARE	76.221	78.421	77.305	72.135	81.212
Prov-GigaPath	Classifier	77.451	71.326	74.262	71.016	75.416
Prov-GigaPath	Full CARE	77.144	80.163	78.625	72.166	80.133
SAM2	Classifier	76.154	66.152	70.802	66.140	72.182
SAM2	Full CARE	71.165	78.132	74.486	69.140	75.163
SAM3	Classifier	75.152	75.150	75.151	71.534	76.172
SAM3	Full CARE	74.142	80.145	77.027	73.061	82.299

C.2 LIZARD: ENCODER CLASSIFIER VERSUS FULL CARE

Encoder	System	Infl. P	Infl. R	Infl. F1	Macro-F1	Accuracy
UNI2	Classifier	82.711	83.803	83.253	62.425	87.168
UNI2	Full CARE	89.180	94.158	91.602	71.961	91.274
Virchow2	Classifier	84.175	86.144	85.148	60.591	82.237
Virchow2	Full CARE	86.124	93.155	89.502	69.122	90.178
Prov-GigaPath	Classifier	81.236	82.233	81.731	59.421	78.344
Prov-GigaPath	Full CARE	81.134	88.652	84.727	64.124	86.178
SAM2	Classifier	75.141	77.621	76.361	59.415	69.144
SAM2	Full CARE	79.144	86.152	82.499	60.863	79.131
SAM3	Classifier	88.144	82.154	85.044	64.138	88.169
SAM3	Full CARE	88.197	92.159	90.134	70.151	91.171

C.3 PANNUKE: ENCODER CLASSIFIER VERSUS FULL CARE

Encoder	System	Infl. P	Infl. R	Infl. F1	Macro-F1	Accuracy
UNI2	Classifier	76.053	79.171	77.555	81.714	85.362
UNI2	Full CARE	78.502	80.314	79.392	82.915	86.305
Virchow2	Classifier	74.150	79.120	76.554	80.122	81.475
Virchow2	Full CARE	78.145	79.166	78.652	80.899	83.621
Prov-GigaPath	Classifier	76.145	76.145	76.145	79.102	83.315
Prov-GigaPath	Full CARE	79.144	75.144	77.092	79.141	84.133
SAM2	Classifier	69.163	75.144	72.030	72.152	75.632
SAM2	Full CARE	72.942	79.741	76.190	76.196	78.144
SAM3	Classifier	76.144	80.162	78.101	82.166	84.165
SAM3	Full CARE	77.164	81.168	79.115	82.169	86.713

UNI2 values are unweighted means of three rotation-specific metrics. For the other encoders, F1 is computed from the listed precision and recall values. The three dataset blocks retain their recorded aggregation rules.

C.4 ADDITIONAL PROTOCOL RECORDS

Dataset	Method	Infl. F1	Macro-F1	Accuracy
PanNuke	CellViT++ (target-pretrained)	87.25	90.80	92.54

This target-domain-pretrained result is listed with its evaluation protocol.

Separate LEMON record. The archived PanNuke LEMON rotation-F1 values are 71.1885, 70.7721, and 71.9544 (mean 71.3050), which define the Table 1 entry. A second PanNuke LEMON evaluation record gives F1 77.865, macro-F1 74.361, and accuracy 79.125.