CAPABILITY IS NOT UTILITY: CONTEXT-DEPENDENT ROUTING OF FROZEN BLACK-BOX MIL EXPERTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Routing among frozen heterogeneous experts is challenging because expert capability does not directly determine intervention utility. In computational pathology, independently trained multiple-instance learning (MIL) models often exhibit complementary strengths across patients, tasks, and expert sets, yet their relative usefulness can change with the evaluation context. We identify two key mismatches in capability-based routing: capability estimated from historical observations may not translate into anchor-relative decision utility, and expert preferences learned in one environment may not remain stable after transfer. We introduce EpiAtlas-Safe, a post-hoc routing framework for frozen black-box MIL experts. EpiAtlas-Safe combines episodic expert-set training, a two-region Capability Atlas for modeling heterogeneous expert behavior, and selective class-balanced safe coupling (SCSC) to align routing decisions with anchor-relative utility while reducing unreliable routing decisions. The framework operates only on class-probability predictions, enabling plug-compatible coordination of independently trained experts without accessing internal representations or joint retraining. On patient-disjoint CAMELYON+ and BRACS evaluations, EpiAtlas-Safe improves patient-weighted slide-level Macro-F1 over a matched capability-transfer router by 1.30 and 2.28 percentage points, respectively, with positive dataset-level gains across all fixed router seeds. In a separately held-out BRCA evaluation, it achieves a 1.08-point improvement on ER while remaining near parity on the prespecified ER/PAM50 primary-task average. Ablation studies support the contributions of SCSC and the Capability Atlas. Additional analyses of expert-ranking changes, utility variation, and residual oracle headroom support modeling expert capability as a context-dependent decision quantity rather than a static global score. Together, these results support viewing frozen-expert routing as context-dependent Anchor-relative intervention selection rather than direct transfer of a static capability ranking. Code is available at `https://anonymous.4open.science/r/epiatlas-safe-review-6D07/README.md`.

1 INTRODUCTION

Computational pathology increasingly produces independently trained whole-slide classifiers with different aggregation mechanisms and inductive biases (Ilse et al., 2018; Lu et al., 2021; Li et al., 2021; Shao et al., 2021; Zhang et al., 2022). Reusing these complete multiple-instance learning (MIL) systems as frozen experts turns this diversity into a coordination problem across heterogeneous model families. Relative to a designated default model, the practical question is input dependent: when and how should an available expert influence the default prediction?

Capability-based routing often treats expert competence as a proxy for the value of using that expert. We distinguish the *Competence-to-Utility Gap*, which separates likely correctness from improvement over an Anchor, from the *Capability-to-Utility Transfer Gap*, which captures how the relationship between source-side capability and target utility changes across patient environments, downstream tasks, and expert-set compositions. This framing separates observable prediction behavior, context-dependent capability, and realized intervention utility.

We study model-level routing over independently trained complete MIL systems. At inference, frozen experts expose only class-probability predictions; their parameters, architectures, and inter-

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.

nal representations are not accessed. The router combines these current predictions with frozen capability characterization estimated on permitted router-training data. The available expert composition may change between training and deployment. This setting connects to sparse MoE routing, model routing, and learning to defer: sparse MoE coordinates internal expert modules in a shared representation space (Fedus et al., 2022), model-level routing selects among complete pretrained systems (Ong et al., 2025), and learning to defer optimizes whether a prediction should be retained or delegated (Mozannar and Sontag, 2020; Verma et al., 2023; Mao et al., 2025).

We introduce **EpiAtlas-Safe**, a post-hoc routing framework with a unified inference-time black-box probability interface and frozen training-side capability characterization. Episodic expert-set training exposes the router to changing expert roles and compositions. A two-region Capability Atlas combines current prediction behavior with local training-side capability information. We introduce selective class-balanced safe coupling (SCSC), an Anchor-relative ranking objective that aligns routing preference with class-balanced utility while penalizing interventions that overturn a correct Anchor prediction. The resulting runtime interface does not require explicit architecture identity and supports heterogeneous frozen experts through their probability outputs and frozen capability context.

Our contributions are threefold. First, we formulate frozen black-box expert routing as a context-dependent intervention problem and separate competence, capability, and realized utility. Second, we instantiate this formulation in EpiAtlas-Safe through episodic expert-set training, a compact Capability Atlas, and the SCSC objective. Third, we characterize capability-to-utility transfer across patient environments, tasks, and expert sets, and quantify residual expert complementarity with label-dependent Oracle headroom. These analyses identify patient-specific intervention-benefit estimation as a key bottleneck in converting expert complementarity into routing utility.

2 Related Work

2.1 MIL diversity and expert combination in computational pathology

MIL represents a slide as a bag of patch instances with slide-level supervision. Attention aggregation (Ilse et al., 2018) and later clustering, dual-stream, Transformer, feature-re-embedding, hierarchical, and pseudo-bag systems (Lu et al., 2021; Li et al., 2021; Shao et al., 2021; Tang et al., 2024; Jin et al., 2025; Zhang et al., 2022) encode distinct inductive biases and can produce complementary errors when trained independently.

Recent work exploits this diversity at different levels. Mammadov et al. use model fusion to reduce run-to-run MIL variability (Mammadov et al., 2025); MoA routes among jointly trained aggregators (Ozlugedik et al., 2026); MAMMOTH introduces mini-experts within an MIL model (Shao et al., 2026); and SIB-MIL uses sparse Bayesian modeling for robust MIL and uncertainty estimation (Chen et al., 2026). EpiAtlas-Safe focuses on post-hoc coordination of independently trained complete MIL models through an inference-time probability interface together with frozen capability characterization from router-training data.

2.2 MoE and routing among complete pretrained models

MoE gates typically operate over internal experts in a jointly trained representation space (Fedus et al., 2022). RoMA studies router generalization with fixed MoE experts while retaining shared task representations (Li et al., 2026). RouteLLM learns cost–quality routing among complete language models from preference data (Ong et al., 2025). Our setting considers complete frozen MIL classifiers accessed through class probabilities at inference, augmented by frozen training-side capability descriptors, with Anchor-relative class-balanced utility and changing expert compositions.

2.3 Learning to defer, selective prediction, and expert uncertainty

Learning to defer optimizes whether a predictor should retain or delegate a decision (Mozannar and Sontag, 2020), with extensions to multiple experts, calibrated decisions, and consistent losses (Verma et al., 2023; Mao et al., 2025). Population deferral and Identity-Free Deferral consider

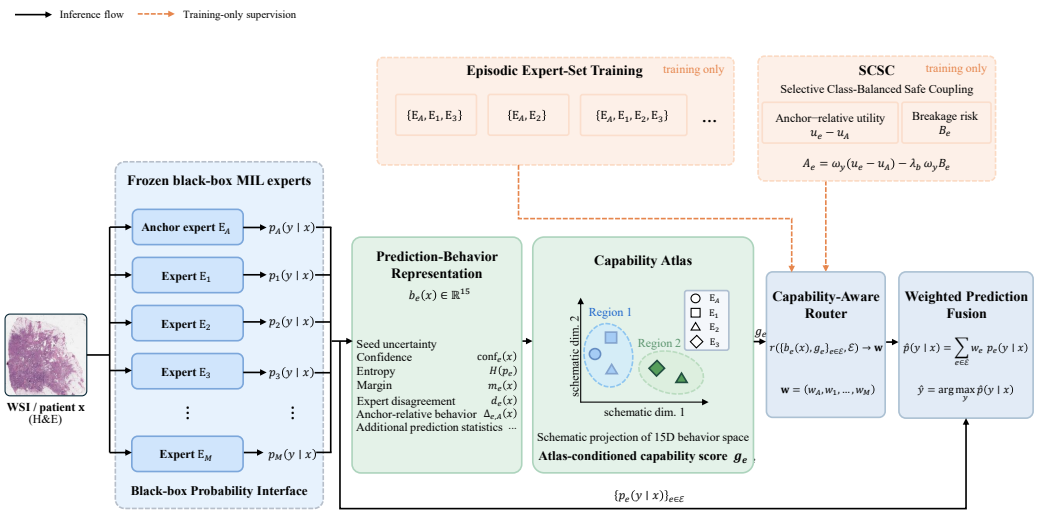

Figure 1: Overview of EpiAtlas-Safe. Frozen MIL experts communicate through a common black-box probability interface at inference, and frozen training-side capability characterization provides expert context for routing. Current prediction behavior $b_e(x)$ is matched against frozen Atlas regions to produce the input-dependent capability score $g_e(x)$, while original expert probabilities enter weighted fusion directly. The Atlas regions are constructed separately for each expert in the 15-dimensional behavior space; the two-dimensional drawing is schematic. For compactness, g_e in the figure denotes the input-dependent score $g_e(x)$. Episodic expert-set training and SCSC provide training-time supervision. Solid arrows denote inference flow and orange dashed arrows denote training-only supervision.

changing or unseen experts using context examples or permutation-symmetric competence representations (Tailor et al., 2024; Strong et al., 2026).

Selective prediction evaluates risk against coverage (Geifman and El-Yaniv, 2019), while deep ensembles and calibration provide reliability estimates whose behavior can vary across evaluation settings (Lakshminarayanan et al., 2017; Park et al., 2020; Kumar et al., 2022). Predictive reliability estimates likely correctness, whereas intervention utility measures Anchor-relative gain. We therefore model competence, capability, routing propensity, and realized utility as separate quantities.

3 Problem Formulation and Method

3.1 Problem Setting and Notation

Let x have label $y \in \{1, \dots, C\}$ and available expert set

$$\mathcal{E}(x) = \{A, e_1, \dots, e_M\},$$

where A denotes the Anchor. Each independently trained frozen expert exposes class probabilities through a common black-box inference interface,

$$p_e(y \mid x) \in \Delta^{C-1}.$$

These probabilities define the runtime expert interface. Frozen capability descriptors characterized from labeled router-training data provide training-side expert context. Together, probability outputs and frozen descriptors form a black-box routing interface that does not require explicit architecture identity.

The routed prediction is

$$\hat{p}(y \mid x) = \sum_{e \in \mathcal{E}(x)} w_e(x) p_e(y \mid x), \qquad \hat{y} = \arg\max_y \hat{p}(y \mid x), \tag{1}$$

where $w_e(x) \geq 0$ and $\sum_{e \in \mathcal{E}(x)} w_e(x) = 1$.

A *correction* changes a wrong Anchor prediction to a correct one. A *breakage* changes a correct Anchor prediction to an incorrect one.

3.2 PREDICTION-BEHAVIOR REPRESENTATION

For each expert, the router constructs a standardized prediction-behavior token and a separate frozen global capability descriptor,

$$b_e(x) \in \mathbb{R}^{15}, \qquad \mathbf{h}_e \in \mathbb{R}^{94}. \tag{2}$$

The token summarizes seed uncertainty, confidence, entropy, prediction margin, expert disagreement, Anchor-relative behavior, and additional prediction-derived statistics. The 15-dimensional token is computed for the current sample x and is therefore input dependent. Source-side capability is represented separately by $\mathbf{h}_e$, which is estimated from the router-training partition and remains fixed during target evaluation.

3.3 EPISODIC EXPERT-SET TRAINING

To reduce dependence on a fixed ordered expert list, EpiAtlas-Safe samples

$$m^{(t)} \in \{0,1\}^M$$

and routes over

$$\mathcal{E}^{(t)} = \{A\} \cup \{e_j : m_j^{(t)} = 1\},$$

varying candidate availability and slot roles while keeping the Anchor active. Episodic sampling is used during training only; inference operates on the currently available set.

3.4 CAPABILITY ATLAS

To capture local variation in expert behavior, the Capability Atlas partitions each expert's router-training tokens by weighted deterministic K-means into $K = 2$ regions with centers

$$c_{e,k} \in \mathbb{R}^{15}.$$

For each expert–region pair, the Atlas estimates local Anchor-relative utility $\mu_{e,k}$ and shrinks sparse estimates toward expert-level utility μ_e:

$$\widetilde{\mu}_{e,k} = \frac{n_{e,k}}{n_{e,k} + \kappa} \mu_{e,k} + \frac{\kappa}{n_{e,k} + \kappa} \mu_e, \tag{3}$$

where $n_{e,k}$ is effective patient mass and $\kappa = 8$.

The frozen capability descriptor is

$$a_e = \left[\mathbf{h}_e; \mathrm{vec}\{(c_{e,k}, \widetilde{\mu}_{e,k})\}_{k=1}^{K}\right] \in \mathbb{R}^{126}.$$

The descriptor a_e stores the frozen training-side Atlas structure for expert e and is independent of the current sample x.

At inference, the current behavior token $b_e(x)$ is matched against these frozen local regions. The 94-dimensional global descriptor and Atlas slots therefore play different computational roles: the base router reads $\mathbf{h}_e$, while the Atlas produces the input-dependent geometric capability score

$$g_e(x) = \log\left[\frac{1}{K}\sum_{k=1}^{K} \exp\left(\frac{\cos(b_e(x), c_{e,k}) + \widetilde{\mu}_{e,k}}{\tau}\right)\right], \tag{4}$$

with $\tau = 1$. The final routing logit is

$$s_e(x) = s_e^{\mathrm{base}}(b_e(x), \mathbf{h}_e) + 0.1\, g_e(x). \tag{5}$$

Because $g_e(x)$ depends on the current behavior token $b_e(x)$, it is the Atlas-conditioned, context-dependent capability signal for the current sample, whereas a_e remains a frozen expert-level descriptor. Here, $g_e(x)$ is not the realized utility of expert e on the current sample; it is a training-side capability signal obtained by transferring historical local utility through similarity in prediction-behavior space. The Atlas therefore combines similarity to training-side behavior regions with their local utility. Figure 1 shows a schematic projection of this matching process in the 15-dimensional behavior space; for compactness, the figure writes g_e for $g_e(x)$.

3.5 SELECTIVE CLASS-BALANCED SAFE COUPLING

We introduce SCSC to map capability information to Anchor-relative intervention preferences using class-balanced correction and breakage signals. For a labeled router-training example (x, y), define

$$u_e(x, y) = \mathbf{1}\left[\arg\max_c p_e(c \mid x) = y\right]$$

and

$$B_e(x, y) = \mathbf{1}\left[u_A(x, y) = 1 \wedge u_e(x, y) = 0\right].$$

SCSC uses the class-balanced safe advantage

$$A_e(x, y) = \omega_y\big(u_e(x, y) - u_A(x, y)\big) - \lambda_b \omega_y B_e(x, y), \tag{6}$$

where $\omega_y \propto n_y^{-1/2}$ and $\lambda_b = 1$. The first term captures the change in correctness relative to the Anchor, while the second imposes an additional asymmetric penalty when an intervention breaks a correct Anchor prediction.

Anchor–expert and expert–expert pairs are used when their safe advantages differ by at least $\delta = 0.05$:

$$\mathcal{P}(x, y) = \{(i, j) : |A_i(x, y) - A_j(x, y)| \geq \delta\}, \qquad v_{ij} = \omega_y |A_i - A_j|. \tag{7}$$

With routing logits s_i and s_j, the weighted SCSC loss is

$$\mathcal{L}_{\text{SCSC}} = \frac{\sum_{(i,j)\in\mathcal{P}} v_{ij}\,\text{ReLU}\left[m - \text{sign}(A_i - A_j)(s_i - s_j)\right]}{\sum_{(i,j)\in\mathcal{P}} v_{ij}}, \tag{8}$$

with $m = 0.1$. Pairs with smaller safe-advantage separation are treated as neutral.

3.6 CAPABILITY-AWARE ROUTING, RISK CONTROL, AND INFERENCE

Routing weights are obtained by

$$w_e(x) = \frac{\exp s_e(x)}{\sum_{j\in\mathcal{E}(x)} \exp s_j(x)}$$

and fuse probabilities according to Equation 1.

Episodic training evaluates each legal candidate subset $\mathcal{S}$. Writing $q_{ij} = w_j(x_i; \mathcal{S})$, the routing-weighted conditional breakage risk is

$$R_{\mathcal{S}} = \frac{\sum_i \sum_{j\in\{A\}\cup\mathcal{S}} q_{ij}\mathbf{1}[\hat{y}_{A,i} = y_i]\mathbf{1}[\hat{y}_{j,i} \neq y_i]}{\max\left(1, \sum_i \mathbf{1}[\hat{y}_{A,i} = y_i]\right)}. \tag{9}$$

Each subset uses the frozen reference-router budget

$$\epsilon_{\mathcal{S}} = \min\left(1, 1.5R_{\mathcal{S}}^{\text{ref}}\right), \tag{10}$$

and routing loss

$$\mathcal{L}_{\text{route},\mathcal{S}} = 1 - \text{softF1}_{\mathcal{S}} + 2.0[R_{\mathcal{S}} - \epsilon_{\mathcal{S}}]_+. \tag{11}$$

The complete subset objective is

$$\mathcal{L}_{\mathcal{S}} = \mathcal{L}_{\text{route},\mathcal{S}} + 0.1\mathcal{L}_{\text{SCSC},\mathcal{S}}. \tag{12}$$

Optimization uses an unweighted full-batch slide-level class-macro differentiable soft-F1 surrogate, while primary CAM/BRACS evaluation uses patient-weighted hard Macro-F1. The exact training surrogate is given in Appendix B.

Experiments use three fixed router seeds (R1–R3), whose slide probabilities are averaged before evaluation; BRCA probabilities are subsequently averaged and renormalized within patient. At inference, the current prediction behavior $b_e(x)$ is combined with the frozen global descriptor $\mathbf{h}_e$ in the base router and matched against the frozen Atlas to produce $g_e(x)$; the resulting routing logits determine the fusion weights. Architecture, optimization, clustering, seed, and schedule details are reported in Appendix B.

Table 1: Datasets, evaluation roles, and primary metrics.

Dataset	Sample and role	Isolation	Primary metric
CAM	813 slides / 350 patients; development	Five patient-disjoint outer splits	Patient-weighted slide-level Macro-F1
BRACS	154 slides / 51 patients; expert-transfer evaluation	Patient + held-out expert	Patient-weighted slide-level Macro-F1
TCGA-BRCA	184 held-out patients; cross-task evaluation	Global patient hold-out across tasks	Patient-level Macro-F1

4 Experimental Protocol

4.1 Datasets and evaluation units

We evaluate CAMELYON+ (CAM) (Ling et al., 2025), BRACS (Brancati et al., 2022), and a TCGA-BRCA-derived cross-task held-out evaluation (BRCA; Table 1). CAM uses 813 slides from 350 patients in five predefined patient-disjoint outer splits. BRACS contains seven breast-pathology classes; held-out experts are excluded from router optimization, while their capability descriptors are computed only from the designated training-side characterization partition and remain fixed during target evaluation. Target-evaluation labels are not used to construct $\mathbf{h}_e$, Atlas regions, or $\mu_{e,k}$. The BRCA evaluation is derived from publicly released TCGA-BRCA pathology slides and associated clinical and molecular annotations available through TCGA/GDC resources (The Cancer Genome Atlas Network, 2012; Heath et al., 2021). It uses globally patient-locked roles, with ER and PAM50 prespecified as primary tasks and PR and HER2 as secondary tasks.

For CAM and BRACS, we use patient-weighted slide-level Macro-F1: all slides from a patient receive total weight one while slide-level labels are retained, preventing patients with more slides from contributing proportionally more weight. BRCA slide probabilities are averaged and renormalized within patient before patient-level evaluation. AUROC and AP are one-vs-rest macro quantities; the exact patient-weighted F1 definition is given in Appendix A.

4.2 Evaluation contexts

We evaluate three complementary axes: patient-environment variation instantiated by the prespecified patient-disjoint stratified CAM/BRACS outer folds, downstream tasks (ER, PAM50, PR, HER2), and predefined expert-set compositions.

4.3 Expert pool and comparison methods

The Anchor is independently trained ABMIL (Ilse et al., 2018). The fixed candidate pool includes CLAM-SB (Lu et al., 2021), HMIL (Jin et al., 2025), RRT-MIL (Tang et al., 2024), DTFD-MIL (Zhang et al., 2022), and the predefined KAN-attention MIL variant where included by the evaluation protocol. All experts are trained independently and frozen before router training.

We compare **Anchor**, **Equal** fusion, fixed non-contextual **Reliability** weighting, the matched learned **Capability-Transfer** router, and **EpiAtlas-Safe**. Capability-Transfer uses the same frozen predictions, prediction-behavior interface, and training-side capability characterization, while EpiAtlas-Safe adds episodic expert-set training, SCSC, and the $K = 2$ Capability Atlas. Additional sensitivity analyses appear in Appendix D.

4.4 Statistical analysis

Primary paired differences use patient-cluster bootstrap with 5,000 resamples and a fixed bootstrap seed. All method components, hyperparameters, expert choices, checkpoints, and the router seed set R1–R3 were frozen before the held-out TCGA-BRCA-derived evaluation.

Table 2: Primary development performance. Macro-F1 is patient weighted at the slide level; each patient contributes total slide weight one.

Method	CAM	BRACS
Anchor	0.5640	0.4251
Equal	0.6047	0.3857
Reliability	0.6021	**0.4457**
Capability-Transfer	0.5943	0.3915
EpiAtlas-Safe	**0.6073**	0.4143
Δ vs. Capability-Transfer	**+0.0130**	**+0.0228**

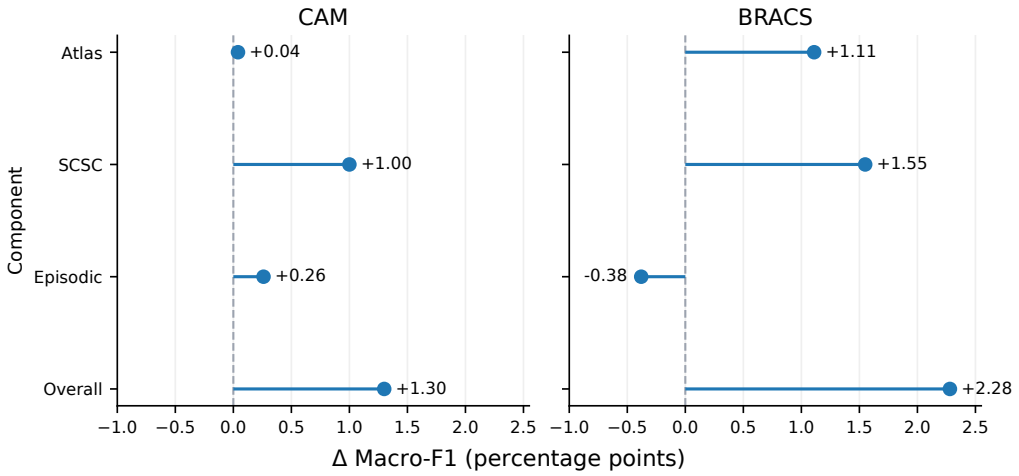

Figure 2: Matched development gains and component contributions. Left: CAM. Right: BRACS. In each panel, the bottom row reports the overall EpiAtlas-Safe–Capability-Transfer difference, and the upper rows report matched component changes. Component effects vary across evaluation contexts, consistent with the context-dependent nature of expert utility.

5 Results

5.1 EpiAtlas-Safe improves matched capability transfer in development

EpiAtlas-Safe produced positive patient-weighted slide-level Macro-F1 differences relative to Capability-Transfer on both development datasets (Table 2). CAM increased from 0.5943 to 0.6073 (ΔF1 $= +0.0130$, 95% CI $[-0.0027, 0.0289]$), and BRACS increased from 0.3915 to 0.4143 (ΔF1 $= +0.0228$, 95% CI $[-0.0134, 0.0645]$). For R1, R2, and R3, the dataset-level differences were $+0.0143, +0.0091, +0.0086$ on CAM and $+0.0283, +0.0216, +0.0247$ on BRACS, respectively; all three fixed router seeds were positive on both development datasets.

On BRACS, Reliability reaches 0.4457 and provides the strongest static reference; the EpiAtlas-Safe–Reliability difference was -0.0314 (95% CI $[-0.0617, 0.0075]$). Relative to the matched contextual Capability-Transfer router, EpiAtlas-Safe improves on both development datasets. Composition-specific minima and correction/breakage analyses are reported in Appendix C.

5.2 SCSC and local capability modeling account for the development gain

SCSC produced the largest consistent matched component change, improving Macro-F1 by $+0.0100$ on CAM and $+0.0155$ on BRACS. The Capability Atlas added $+0.0004$ on CAM and $+0.0111$ on BRACS. Figure 2 summarizes the matched overall gains and incremental component contributions on a common scale. Full correction/breakage counts, RoMA-inspired regularization, and $K = 4$ sensitivity are reported in Appendix D.

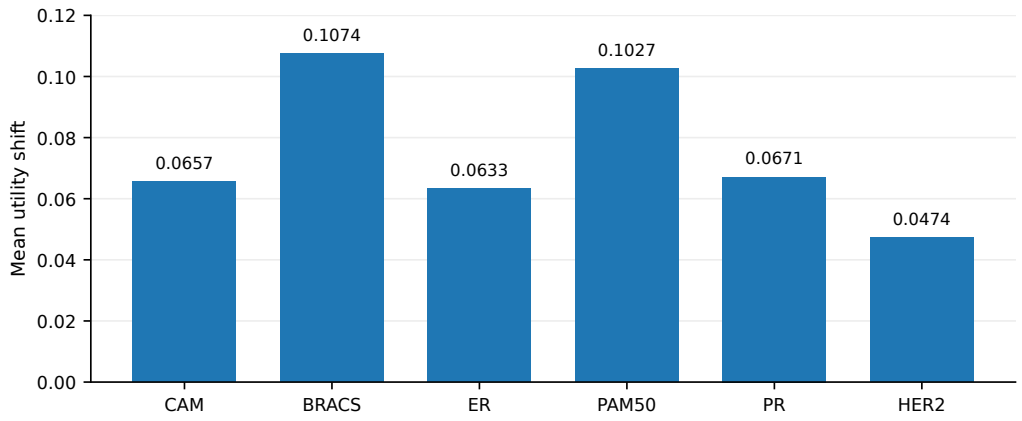

Figure 3: Capability-to-utility transfer across evaluation contexts. Bars show the mean absolute source-to-target utility change for each evaluation setting. Additional analyses of expert ranking changes are provided in the supplementary material. The variation across contexts indicates that intervention value is context dependent after transfer.

5.3 Capability-to-utility correspondence changes across evaluation contexts

Expert capability ordering changed substantially after transfer. Source-to-target ranking inversions occurred in 4/5 CAM and 1/2 BRACS environments, while utility signs changed in 6/10 CAM and 1/4 BRACS comparisons. Mean absolute source-to-target utility shifts were 0.0657 and 0.1074, respectively.

Across BRCA tasks, ranking inversions occurred in 6/10 ER, 2/10 PAM50, 5/10 PR, and 6/10 HER2 comparisons, with mean absolute utility shifts of 0.0633, 0.1027, 0.0671, and 0.0474. Figure 3 visualizes the mean absolute utility shifts; the corresponding ranking inversions and sign changes are reported above. Across BRCA diagnostics, median AUROC was 0.613 for competence discrimination and 0.561 for Anchor-relative utility discrimination, indicating that likely correctness is easier to identify than Anchor-relative improvement.

Expert ordering and intervention value therefore vary across deployment contexts, producing a capability-to-utility transfer gap in which target utility can reorder relative to source-side capability.

5.4 Residual expert complementarity reveals unrealized intervention headroom

To separate the availability of complementary predictions from the router's ability to exploit them, we evaluate a label-dependent intervention Oracle. For each case, the Oracle preserves a correct Anchor prediction when routing would break it and accepts a routed prediction when it corrects an Anchor error. It serves as a non-deployable upper-bound diagnostic for intervention selection.

Positive Oracle headroom appears in five of six evaluation settings. Relative to EpiAtlas-Safe, the Oracle improves Macro-F1 by $+0.0156$ on CAM, $+0.0613$ on BRACS, $+0.0093$ on ER, $+0.0315$ on PR, and $+0.0501$ on HER2; PAM50 contributes no additional Oracle intervention gain. The BRACS gap is especially informative: a strong static Reliability baseline coexists with $+6.13$ points of residual Oracle headroom. Learned and static routing therefore capture only part of the complementary decision opportunity available in the frozen expert set.

The residual gap localizes the next modeling opportunity to patient-specific Anchor-relative intervention-benefit estimation from frozen expert outputs. The expanded expert-pool evaluation in Appendix G reinforces this interpretation: Oracle opportunity increases even as EpiAtlas-Safe performance decreases.

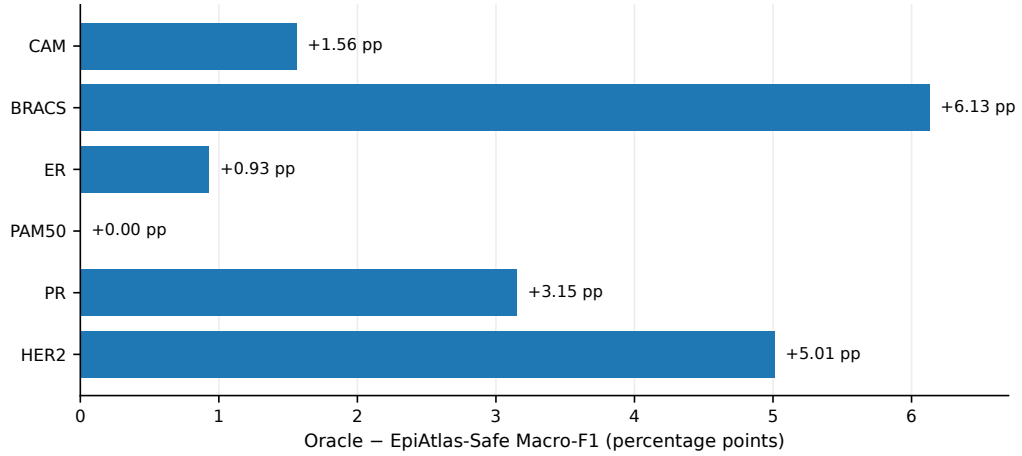

Figure 4: Residual intervention headroom from the label-dependent Oracle, reported as Oracle minus EpiAtlas-Safe Macro-F1 (percentage points). The Oracle is a non-deployable upper bound.

5.5 TCGA-BRCA-derived held-out evaluation

On ER, EpiAtlas-Safe improved Macro-F1 over Capability-Transfer by $+0.0108$ (95% CI $[-0.0191, 0.0401]$). Across the prespecified ER/PAM50 primary pair, the average difference was -0.0011, indicating near parity overall. On PAM50, AUROC and AP increased from 0.7878 and 0.5701 to 0.8021 and 0.6157, respectively, while NLL, Brier score, and ECE were lower; Macro-F1 decreased by 0.0131. This pattern is consistent with the task- and metric-dependent transfer observed in the BRCA evaluation. Additional PAM50 probability-level metrics are reported in Appendix E.

6 Discussion

EpiAtlas-Safe yields positive matched development point estimates over Capability-Transfer across all three fixed router seeds; SCSC contributes the largest matched gains and the Atlas adds a BRACS-specific benefit (Figure 2). Across patient environments and tasks, expert ranking and utility change, while competence is easier to discriminate than Anchor-relative intervention utility (Figure 3). The TCGA-BRCA-derived held-out evaluation extends this pattern to task- and metric-dependent transfer.

BRACS combines a strong Reliability baseline with $+0.0613$ Macro-F1 of label-dependent Oracle headroom above EpiAtlas-Safe, while the expanded expert-pool evaluation increases complementarity as realized routing utility decreases (Figure 4; Appendix G). Together, these analyses show that the frozen expert pool contains additional decision opportunity beyond current routing performance and identify patient-specific intervention-benefit estimation as the next target for converting expert complementarity into realized utility.

7 Conclusion

Capability is not utility. EpiAtlas-Safe provides context-dependent routing for frozen black-box MIL experts and shows that expert capability and complementarity translate into realized routing utility in a context-dependent manner. Across patient environments, tasks, and expert sets, the results support modeling capability as an Anchor-relative decision quantity rather than a static global score.

AI USE STATEMENT

Generative AI tools supported literature discovery and synthesis; feedback on research framing, experimental design, and methodological presentation; result interpretation; plotting and workflow code; scientific-figure drafting; and manuscript drafting, editing, organization, and consistency checking. The authors reviewed and verified all AI-assisted methodological suggestions, code, analyses, citations, numerical results, and interpretations against the underlying experiments, source code, artifacts, and cited literature. All reported observations, labels, and performance results derive from the underlying experimental artifacts. The authors take responsibility for the final scientific content of the submission.

ETHICS STATEMENT

This study uses only publicly released research datasets and features derived from those resources. CAMELYON+, BRACS, and TCGA-BRCA were used in accordance with the access and data-use terms of their originating repositories. The TCGA-BRCA evaluation uses pathology slides and associated public clinical and molecular annotations from TCGA/GDC resources (The Cancer Genome Atlas Network, 2012; Heath et al., 2021); UNI2-h features are derived from the released whole-slide images. No new participants were recruited, no intervention was performed, and no new identifiable participant information was collected for this study. The work consists of secondary computational analysis of previously released research data.

REPRODUCIBILITY STATEMENT

The study records patient split registries, frozen method configurations, base-expert and router seed mappings, bootstrap settings, prediction outputs, and content hashes. The main text and Appendix specify the routing formulation, principal hyperparameters, evaluation units, comparison methods, and statistical protocol. The anonymous repository provides preprocessing, split, training, inference, capability-descriptor, risk-budget, and prediction-lineage artifacts.

CAMELYON+ and BRACS are publicly available through their cited dataset releases. TCGA-BRCA pathology slides and associated annotations are available through the NCI Genomic Data Commons and TCGA resources (Heath et al., 2021; The Cancer Genome Atlas Network, 2012). The anonymous repository provides split registries, configurations, capability-descriptor schemas, evaluation definitions, and code required to reproduce the reported routing and evaluation analyses from prepared frozen-expert predictions and released metadata.

REFERENCES

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, et al. BRACS: A Dataset for BReAst Carcinoma Subtyping in H&E Histology Images. *Database*, 2022:baac093, 2022. https://doi.org/10.1093/database/baac093.

Yihang Chen, Tsai Hor Chan, Jianning Chen, Li Liang, Guosheng Yin, and Lequan Yu. SIB-MIL: Sparsity-Induced Bayesian Neural Network for Robust Multiple Instance Learning on Whole Slide Image Analysis. *IEEE Transactions on Medical Imaging*, 45(4):1638–1650, 2026. https://doi.org/10.1109/TMI.2025.3638243.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *ICML*, volume 97 of PMLR, pages 2151–2159, 2019.

Allison P. Heath, Vincent Ferretti, Stuti Agrawal, et al. The NCI Genomic Data Commons. *Nature Genetics*, 53:257–262, 2021. https://doi.org/10.1038/s41588-021-00791-5.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, volume 80 of PMLR, pages 2127–2136, 2018.

Cheng Jin, Luyang Luo, Huangjing Lin, Jun Hou, and Hao Chen. HMIL: Hierarchical Multi-Instance Learning for Fine-Grained Whole Slide Image Classification. *IEEE Transactions on Medical Imaging*, 44(4):1796–1808, 2025.

Ananya Kumar, Tengyu Ma, Percy Liang, and Aditi Raghunathan. Calibrated ensembles can mitigate accuracy tradeoffs under distribution shift. In *UAI*, volume 180 of PMLR, pages 1041–1051, 2022.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, volume 30, 2017.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *CVPR*, 2021.

Zhongyang Li, Ziyue Li, and Tianyi Zhou. Routing Manifold Alignment Improves Generalization of Mixture-of-Experts LLMs. In *ICLR*, 2026.

Xitong Ling, Yuanyuan Lei, Jiawen Li, et al. Comprehensive Benchmark Dataset for Pathological Lymph Node Metastasis in Breast Cancer Sections. *Scientific Data*, 12:1381, 2025. https://doi.org/10.1038/s41597-025-05586-5.

Ming Y. Lu et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:555–570, 2021. https://doi.org/10.1038/s41551-020-00682-w.

Ali Mammadov, Loic Le Folgoc, Guillaume Hocquet, and Pietro Gori. Reducing Variability of Multiple Instance Learning Methods for Digital Pathology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 506–516, 2025.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Mastering multiple-expert routing: Realizable H-consistency and strong guarantees for learning to defer. In *ICML*, volume 267 of PMLR, pages 43035–43066, 2025.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *ICML*, volume 119 of PMLR, 2020.

Isaac Ong et al. RouteLLM: Learning to route LLMs from preference data. In *ICLR*, 2025.

Fatih Ozlugedik et al. MoA: Mixture of Aggregators Improves Slide-Level Diagnosis in Computational Pathology. In *Medical Imaging with Deep Learning*, 2026.

Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *AISTATS*, volume 108 of PMLR, pages 3219–3229, 2020.

Daniel Shao, Joel Runevic, Richard J. Chen, Drew F. K. Williamson, Ahrong Kim, Andrew H. Song, and Faisal Mahmood. Mixture of Mini Experts: Overcoming the Linear Layer Bottleneck in Multiple Instance Learning. In *ICLR*, 2026.

Zhuchen Shao et al. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *NeurIPS*, volume 34, 2021.

Joshua Strong, Pramit Saha, Yasin Ibrahim, Cheng Ouyang, and J. Alison Noble. Identity-Free Deferral for Unseen Experts. In *ICLR*, 2026.

Dharmesh Tailor, Aditya Patra, Rajeev Verma, Putra Manggala, and Eric Nalisnick. Learning to defer to a population: A meta-learning approach. In *AISTATS*, volume 238 of PMLR, pages 3475–3483, 2024.

Wenhao Tang et al. Feature Re-Embedding: Towards Foundation Model-Level Performance in Computational Pathology. In *CVPR*, pages 11343–11352, 2024.

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490:61–70, 2012. https://doi.org/10.1038/nature11412.

Rajeev Verma, Daniel Barrejon, and Eric Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *AISTATS*, volume 206 of PMLR, pages 11415–11434, 2023.

Hongrun Zhang et al. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *CVPR*, pages 18802–18812, 2022. https://doi.org/10.1109/CVPR52688.2022.01824.

A Evaluation Metric Definition

For CAM and BRACS, slide-level labels are retained while patient contribution is balanced. If slide i belongs to patient $p(i)$ with $n_{p(i)}$ slides, its evaluation weight is $\alpha_i = 1/n_{p(i)}$. For class c, weighted TP_c, FP_c, and FN_c are formed by summing these slide weights over the corresponding outcomes, and

$$F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c}, \qquad \text{MacroF1} = \frac{1}{C}\sum_{c=1}^{C} F1_c.$$

Thus every patient contributes total slide weight one. BRCA instead averages and renormalizes slide probabilities within patient before computing patient-level Macro-F1.

B Reproducibility and Implementation Details

Randomness layers and seed labels. To distinguish randomness at different stages, we use semantic seed labels in the manuscript. Base-expert training seeds E1–E3 correspond to integer seeds 42, 43, and 44, whereas the frozen EpiAtlas-Safe router seeds R1–R3 correspond to 20260911, 20260917, and 20260923. These two seed sets control different training layers and are not interchangeable. Where independent-seed reliability checks are reported, validation seeds V1–V3 correspond to 45, 46, and 47 and are used only for that dedicated validation.

Router and optimization. The router uses latent dimension 16, one 32-unit ReLU hidden layer, and no dropout. AdamW uses learning rate 3×10^{-3}, weight decay 10^{-4}, and full-batch updates. Training allows at most 4,000 epochs, at least 200 epochs, patience 250, and minimum validation improvement 10^{-5}. The epoch count is selected on the validation partition; the router and optimizer are then reinitialized and refitted for exactly that count. The frozen router seed set is denoted R1–R3 throughout the manuscript.

Atlas construction. Weighted deterministic K-means uses Lloyd's algorithm, 10 initializations, and random state 20260911. The frozen configuration uses $K = 2$, shrinkage $\kappa = 8$, temperature $\tau = 1$, and geometry coefficient 0.1.

SCSC weighting. Class weights use normalized inverse-square-root frequency,

$$\omega_c = \frac{n_c^{-1/2}}{C^{-1}\sum_{d=1}^{C} n_d^{-1/2}}.$$

SCSC forms all unordered Anchor–expert and expert–expert pairs whose safe-advantage separation is at least $\delta = 0.05$, with pair weight $v_{ij} = \omega_y |A_i - A_j|$, ranking margin $m = 0.1$, and outer coefficient $\lambda_{sc} = 0.1$.

Capability descriptor specification. The frozen 94-dimensional vector $\mathbf{h}_e$ is generated by the training-side capability-characterization pipeline and is held fixed during target evaluation. It comprises 75 distributional summaries of the 15-dimensional behavior token, four predicted-class prevalence statistics, and 15 train-side few-shot performance, calibration, and Anchor-relative utility statistics. Supplementary material provides the coordinate-level descriptor schema and extraction implementation for exact feature ordering and reproduction.

Soft-F1 training surrogate. Optimization uses an unweighted full-batch slide-level class-macro soft-F1 surrogate, whereas reported CAM/BRACS performance uses patient-weighted hard Macro-F1. For a legal subset $\mathcal{S}$, let r_{ic} denote the fused class probability for slide i in the full router-training batch and $y_{ic} \in \{0, 1\}$ its one-hot label. The surrogate is computed as

$$\widetilde{TP}_c = \sum_i y_{ic} r_{ic}, \tag{13}$$

$$\widetilde{FP}_c = \sum_i (1 - y_{ic}) r_{ic}, \tag{14}$$

$$\widetilde{FN}_c = \sum_i y_{ic} (1 - r_{ic}), \tag{15}$$

$$\mathrm{softF1}_{\mathcal{S}} = \frac{1}{C} \sum_{c=1}^{C} \frac{2\widetilde{TP}_c}{2\widetilde{TP}_c + \widetilde{FP}_c + \widetilde{FN}_c + 10^{-8}}. \tag{16}$$

The training surrogate uses unweighted slide-level soft counts; reported CAM/BRACS Macro-F1 uses hard predictions with patient-normalized slide weights as defined in Appendix A.

Reference-risk budget. The subset-specific risk budget uses the hard conditional breakage risk of a patient-grouped out-of-fold logistic correction–breakage reference policy over the Anchor, Equal, and Reliability actions. For the non-Anchor actions, logistic models estimate correction and breakage probabilities, and the policy selects the action with the largest strictly positive predicted correction-minus-breakage utility; otherwise, it retains the Anchor. Its subset-specific risk defines $R^{\mathrm{ref}}_{\mathcal{S}}$ in Equation 10.

Statistical details. Paired bootstrap uses 5,000 resamples and seed 20260915. For CAM and BRACS, patients are resampled within each outer fold and the same draw multiplicities are shared across compared methods. For BRCA, paired patient resampling is performed separately within each task, and the ER/PAM50 primary-task mean is averaged draw-by-draw. Confidence intervals use the 2.5th and 97.5th percentiles of the paired bootstrap distribution.

TCGA-BRCA-derived cohort. The TCGA-BRCA task cohort is derived from publicly released pathology slides and associated clinical and molecular annotations available through TCGA/GDC resources (The Cancer Genome Atlas Network, 2012; Heath et al., 2021). Patient identifiers follow the first 12 characters of TCGA barcodes, and the task labels are taken from the corresponding public TCGA-BRCA annotations. UNI2-h 1536-dimensional patch features are derived representations extracted from the released whole-slide images. The registry assigns 552 patients to development and 184 to the globally patient-locked held-out cohort; task-specific held-out cohorts contain 117–176 patients.

C Development Composition and Intervention Details

Table 3: Full development performance including composition minima and intervention counts. "Worst" denotes the minimum over predefined expert compositions. C/B/N denotes corrections, breakages, and their net difference.

Method	CAM F1	CAM worst	BRACS F1	BRACS worst	CAM C/B/N	BRACS C/B/N
Anchor	0.5640	0.4975	0.4251	0.3773	0/0/0	0/0/0
Equal	0.6047	0.5466	0.3857	0.3290	74/17/57	12/9/3
Reliability	0.6021	0.5469	0.4457	0.3788	63/13/50	12/6/6
Capability-Transfer	0.5943	0.4948	0.3915	0.3219	54/9/45	10/7/3
EpiAtlas-Safe	0.6073	0.5394	0.4143	0.3215	72/14/58	9/5/4

EpiAtlas-Safe exceeded Capability-Transfer in 8/10 CAM fold-by-held-out-expert evaluations. On BRACS, the full-composition difference was positive in one of two outer splits, and singleton-composition differences were positive in one of four cases.

D Full Ablations and Sensitivity Analyses

Table 4: Full matched development ablations.

Data	Contrast	ΔF1	Δcorr	Δbreak
CAM	Episodic − Capability-Transfer	+0.0026	+7	0
BRACS	Episodic − Capability-Transfer	-0.0038	-4	0
CAM	SCSC − Episodic	+0.0100	+12	+6
BRACS	SCSC − Episodic	+0.0155	+2	-1
CAM	$K = 2$ Atlas on SCSC	+0.0004	-1	-1
BRACS	$K = 2$ Atlas on SCSC	+0.0111	+1	-1
CAM	RoMA-inspired regularization	-0.0004	+1	0
BRACS	RoMA-inspired regularization	+0.0015	0	0

SCSC increased both corrections and breakages on CAM, but increased corrections while reducing breakage on BRACS. Across the complete Capability-Transfer–EpiAtlas-Safe comparison, breakage changed from 9 to 14 on CAM and from 7 to 5 on BRACS. Increasing the Atlas from $K = 2$ to $K = 4$ changed Macro-F1 by approximately $+0.0004$ on CAM and -0.0056 on BRACS. The RoMA-inspired regularizer produced little or no Macro-F1 gain and did not contract routing-neighbor distance as intended.

E Complete TCGA-BRCA-Derived Task Results

Table 5: Patient-level Macro-F1 on the prespecified TCGA-BRCA-derived tasks. ER and PAM50 are primary; PR and HER2 are secondary.

Task	Anchor	Reliability	Capability-Transfer	EpiAtlas-Safe	Δ
ER	0.8244	**0.8544**	0.8436	**0.8544**	**+0.0108**
PAM50	**0.5886**	0.5617	0.5792	0.5661	-0.0131
Primary mean	0.7065	0.7081	**0.7114**	0.7103	-0.0011
PR	**0.7233**	0.7075	0.7180	0.7002	-0.0178
HER2	**0.6026**	0.5602	0.5888	0.5738	-0.0150

For PAM50, the full expert composition yielded NLL values of 1.1208 and 0.9590, multiclass Brier scores of 0.2801 and 0.2592, and 10-bin ECE values of 0.1855 and 0.1049 for Capability-Transfer and EpiAtlas-Safe, respectively.

F Intervention Propensity and Risk–Coverage

For

$$S(x) = 1 - \frac{1}{3}\sum_{s=1}^{3} w_A^{(s)}(x),$$

benefit AUROC was 0.448 on CAM, 0.357 on BRACS, 0.133 on ER, 0.161 on PR, and 0.200 on HER2; PAM50 contained no beneficial EpiAtlas-Safe correction.

At 50% coverage, CAM corrections decreased from 72 to 15 and breakages from 14 to 5; BRACS corrections decreased from 9 to 2 and breakages from 5 to 1. No BRCA task retained a correction.

The label-dependent Oracle summarized in Section 5.4 is computed only as an upper-bound diagnostic for intervention selection. It uses outcome labels to identify routed corrections and Anchor-preserving decisions and is excluded from model training, selection, and deployable inference.

G Exploratory Expanded Expert-Pool Evaluation

We performed an exploratory expert-pool transfer evaluation on the prespecified outer split by adding SIB-MIL and MAMMOTH-ABMIL to the base expert pool while keeping the EpiAtlas-Safe method configuration, training protocol, and router seed set unchanged. The added experts increased diversity and Oracle opportunity, while their individual Macro-F1 remained below the strongest expert under the matched protocol.

Table 6: Exploratory expert-pool transfer results on the prespecified outer split. The two middle columns report patient-weighted slide-level Macro-F1 of EpiAtlas-Safe separately fitted and evaluated under the full base and expanded expert pools. Predictions are probability ensembles over the three fixed router seeds. Oracle-headroom gain denotes the increase in Oracle–Anchor headroom after pool expansion.

Dataset	EpiAtlas-Safe F1, base pool	EpiAtlas-Safe F1, expanded pool	Oracle-headroom gain
CAM	0.7182	0.6927	**+0.0047**
BRACS	0.3215	0.2895	**+0.0257**

The expanded pool increased complementary opportunity while EpiAtlas-Safe Macro-F1 decreased. On CAM, Equal fusion increased from 0.6720 to 0.6963 after full pool expansion, indicating that simple averaging captured part of the added diversity. On BRACS, the Anchor remained stronger than the expanded static and learned fusion policies. The prespecified progressive expert-set evaluation protocol ended at this stage, so we treat the result as exploratory evidence.

H Frozen Constituent Expert Characterization

Table 7 summarizes the constituent experts used in the EpiAtlas-Safe evaluation protocol and their available outer-evaluation Macro-F1 scores.

Table 7: Frozen constituent expert characterization. Rows marked "Membership only" document expert participation without reporting a directly comparable outer-evaluation Macro-F1.

Dataset	Expert	Role	Evaluation	Macro-F1
CAM	ABMIL/Anchor	anchor	Reported	0.5640
CAM	HMIL	query/source	Reported	0.6333
CAM	RRT-MIL	query/source	Reported	0.6190
BRACS	ABMIL/Anchor	anchor	Reported	0.4251
BRACS	HMIL	source expert	Membership only	–
BRACS	CLAM-SB	source expert	Membership only	–
BRACS	KanMIL	source expert	Membership only	–
BRACS	RRT-MIL	held-out query	Reported	0.3468
BRACS	DTFD-MIL	held-out query	Reported	0.4573