StatOrch: Evidence-Conditioned Stagewise Inference under Unknown Distribution Shift

Anonymous authors
Paper under double-blind review

Abstract

Inference on an unlabeled target population requires deciding how to respond to distribution shift: correct class proportions, adapt a predictor, acquire measurements, or retain the current procedure. We present StatOrch, a framework connecting observable statistical evidence to these choices through stage-specific effect prediction and fixed fallbacks. It organizes established tools around population targets, dependent contexts, and resource constraints. Controlled evaluations show routing gains over validation-selected fixed actions. An executed adaptation–quantification–classification chain improves the joint population-and-prediction objective on seen mechanisms across four fitting seeds on a fixed cohort, with no corresponding advantage established on unseen mechanisms. A matched one-shot complete-pipeline selector attains similar losses; these results do not establish a stagewise-selection advantage. Separately, an archival measurement/resource instance with fixed downstream heads records 21.1% lower known-class conditional prevalence error, 5.26 percentage points higher Macro-F1, and 10.5% lower recorded measurement cost than a near-cost fixed policy on 1,232 held-out units, under its preserved ledger. Acquisition, re-entry, and real-sensor stress tests characterize operational behavior under their respective protocols. The controlled results support conditional selection and dependent execution within the evaluated settings.

1 Introduction

An unlabeled target population creates two coupled problems: infer what is present, and decide which statistical procedure the evidence justifies. Changed class proportions can favor prior correction; changed class-conditional features can favor adaptation; damaged measurements can favor a different sensor bundle. These conditions can coexist. Acquiring observations, updating a model, and correcting its probabilities address different errors, making intervention choice part of inference.

Established tools address each case: quantification estimates population proportions, test-time adaptation changes a predictor, and active acquisition purchases information (Saerens et al., 2002; Wang et al., 2021; Li & Oliva, 2021). Yet a fixed choice need not suit every target task. Prior correction need not repair a distorted representation; adaptation changes the references needed by a quantifier; and additional observations have value relative to the estimator, sampling policy, and measurement cost.

We ask: *how can observable evidence guide choices across these interventions without target labels?* A controller must connect diagnostics to task effects, respect observation permissions, preserve a fallback, and update downstream quantities after an upstream change. Our focus is this decision structure and its statistical targets.

We introduce StatOrch, an evidence-conditioned stagewise framework. Confidence, estimator disagreement, fit, stability, and source–target discrepancy form its inputs. Nonexclusive diagnostics enrich them without forcing a causal diagnosis. Offline task losses supervise candidate-effect predictors; validation selects when to replace a fixed fallback. Dedicated resource and open-world rules enforce cost and metric constraints. The forward chain produces prevalence estimates, predictions, risk summaries, screened residuals, and optional clusters, with their decision and measurement records (Figure 1).

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.

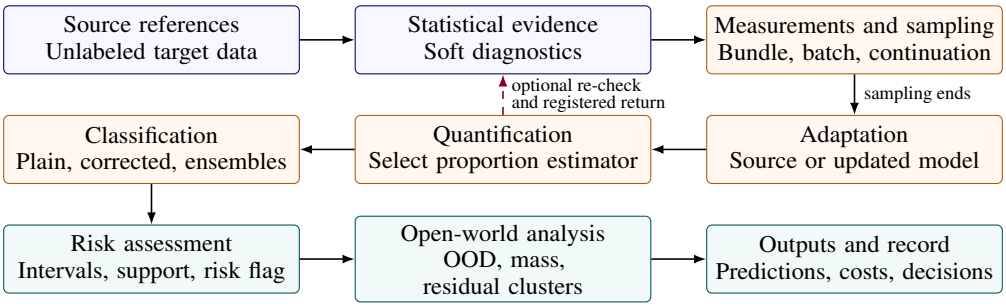


Figure 1: **From observations to statistical decisions.** Solid arrows show forward dependencies. Learned stages compare predicted effects against fixed fallbacks; measurement and open-world stages use dedicated constrained rules. The dashed edge illustrates optional registered re-entry. Stopping acquisition still permits downstream inference.

The core evidence comes from controlled stage-routing comparisons against validation-selected fixed actions and a directly executed adaptation–quantification–classification chain. The chain evaluation includes a matched one-shot complete-pipeline selector and four fitting seeds on a fixed cohort. Separately, archival resource-control records on 1,232 held-out tasks provide complementary endpoint evidence under their preserved measurement ledger. Measurement and re-entry studies, together with real-sensor stress tests, characterize operational behavior.

Our contributions are: (i) a population-inference framework connecting observable evidence to statistical interventions; (ii) a complete stagewise specification of effect prediction, fallbacks, resources, risk, and dependency updates; and (iii) an empirical account of prediction–cost gains, action specialization, and intervention control. The methodological contribution organizes established effect-prediction tools around population targets and dependent statistical contexts; the experiments identify the settings in which the evaluated decisions are useful.

2 Related Work

Quantification estimates population proportions through posterior re-estimation or distribution matching (Saerens et al., 2002; González-Castro et al., 2013). Adaptation changes the predictor (Niu et al., 2023), potentially changing the quantifier's reference compatibility. StatOrch treats these as separate, dependent decisions with source-model and EMQ fallbacks.

Performance-based algorithm selection learns candidate outcomes from instance features, as implemented by ASML (Gómez-Casares et al., 2025). Automatic configuration (Feurer et al., 2015) and MetaML's offline recommendation of dependent pipelines (Maia et al., 2025) provide related foundations. Our contribution is a statistical-task interface: population targets, observation permissions, stage-specific effects, and downstream updates. It does not introduce performance regression as a new selection principle.

Unsupervised adaptation validation includes DEV (You et al., 2019), importance-weighted group accuracy estimation (IW-GAE) (Joo & Klabjan, 2024), and ensemble-based model selection (EnsV) (Hu et al., 2024). DEV estimates target risk; IW-GAE and EnsV assess target predictions. StatOrch instead learns stage effects from labeled offline tasks and observable online evidence; it therefore depends on transfer across tasks. These methods address complementary unsupervised model-selection objectives; direct same-bank comparison is outside the present evaluation. The structured supplement compares information and supervision requirements.

Active feature acquisition uses generative surrogates (Li & Oliva, 2021), stochastic encodings (Norcliffe et al., 2025), or explanation-driven rankings (Guney et al., 2025) to choose measurements. StatOrch's acquisition target is population-proportion error, with task-level bundles and sample continuation; label-querying active test-time adaptation (Gui et al., 2024) has different online permissions.

Individual OOD scores (Liu et al., 2020) and category discovery (Vaze et al., 2022) answer different questions from unknown population mass. StatOrch preserves these distinctions, including a no-discovery output.

3 Problem Setup

We study inference for a fixed, unlabeled target population P_T when the useful intervention is not known in advance. Let $\mathcal{Y} = \{1, \dots, K\}$ be the source classes and aggregate unknown classes as $\perp$. The estimands are the class proportions conditional on membership in the known classes and the unknown mass:

$$\beta_k = P_T(Y = k \mid Y \in \mathcal{Y}), \qquad \rho = P_T(Y = \perp), \qquad \beta \in \Delta^{K-1}, \quad 0 \le \rho < 1, \tag{1}$$

where $\Delta^{K-1} = \{v \ge 0 : \mathbf{1}^\top v = 1\}$. The joint population masses are $\pi_T = ((1-\rho)\beta^\top, \rho)^\top$; a closed-set task has $\rho = 0$. Outputs include proportion estimates, known-class predictions, uncertainty, unknown-sample screening, optional residual clusters, and the decisions and resources used to obtain them.

Labeled source data are split into training and reference calibration sets. Separate offline task sets $\mathcal{T}$ and $\mathcal{V}$ train and validate the controller. Within a target task, fitting indices I_t, acquisition candidates P_t, and prediction indices J are disjoint. Adaptation and proportion fitting use only I_t. Classification and open-world analysis may read prediction covariates X_J when the observation protocol permits them, but never Y_J. True proportions, unknown mass, mechanism identities, and evaluation labels are available only for offline supervision or evaluation.

For measurement channels $\mathcal{S}$, a candidate bundle $b \in \mathcal{B}$ is a subset of channels with additive cost $c(b) = \sum_{s \in b} c_s$. An execution protocol $\mathcal{P}$ fixes initial observations, available tools, information permissions, resource checkpoints and limits, and whether sequential acquisition or action verification is enabled. These choices determine the admissible actions; they do not change the estimands. Appendix A.1 formalizes the observation permissions.

4 StatOrch

StatOrch turns observable evidence into a sequence of statistical decisions. At each learned routing stage, it predicts candidate-specific loss or utility and replaces a fixed fallback only when the predicted advantage exceeds a validation-selected threshold. Measurement and open-world analysis additionally enforce their own resource or metric rules. Figure 1 shows the dependencies; Algorithm 1 gives the online order.

4.1 Statistical Evidence and Stage-wise Control

Frozen source references. For each bundle, select a source model using Macro-F1, calibration error, and Brier score, and its calibrator by reference negative log-likelihood. Freeze the calibrated probabilities, prior, hard and soft confusion matrices, feature references, prototypes, and neural OOD references; adaptation uses model copies. Train offline modules in execution order, with task-level out-of-fold upstream actions generating downstream training contexts. Validation follows the trained upstream pipeline (Appendices A.2 and A.4).

Evidence, rather than a hard shift diagnosis. Initial observations under bundle b_0 yield source predictions and a preliminary EMQ estimate. Evidence includes current proportions, confidence and entropy, quantifier disagreement, bootstrap widths, subsample stability, confusion-matrix reconstruction residuals, and source–target distribution differences. Channel health statistics use only channels already observed. A fixed dictionary forms x_j^{raw} for stage j; a reference transform ψ_j combines empirical quantiles and robust standardized coordinates. Independent classifiers produce soft scores for applicability, conditional mismatch, representation failure, insufficient information, out-of-reference structure, and conflicting evidence:

$$e_j = \left[\psi_j(x_j^{\mathrm{raw}}), \left(d_{j,h}(\psi_j(x_j^{\mathrm{raw}}))\right)_{h \in \mathcal{H}}\right]. \tag{2}$$

Scores need not sum to one; their maximizer does not mandate an action. Recompute affected evidence after new observations or context changes (Appendix A.3).

Candidate-specific predictions and fallback. Offline candidate losses train predictors $\widehat{L}_{j,a}(e)$. Given fallback a_0 and $a^- \in \arg\min_{a \in \mathcal{A}_j(e)} \widehat{L}_{j,a}(e)$, choose

$$a_j^\star = \begin{cases} a^-, & \widehat{L}_{j,a_0}(e) - \widehat{L}_{j,a^-}(e) > \delta_j, \\ a_0, & \text{otherwise.} \end{cases} \tag{3}$$

Utility stages require a predicted utility gain above their threshold. Validation retains controls within a prescribed multiple of the best mean loss's standard error, then prefers lower cost or switching rate. Unmet dependencies exclude tools; numerical failures invoke the fallback, with fallback failures recorded and handled by the protocol.

4.2 Measurement and Resource Control

Measurement prediction targets each bundle's EMQ proportion error at the reference initial sample size. The near-optimal, low-cost protocol retains bundles close to the predicted optimum and chooses the cheapest:

$$\mathcal{B}_{\text{near}}(e) = \left\{ b : \widehat{E}_b(e) \le \min_{b'} \widehat{E}_{b'}(e) + \delta_B \right\}, \qquad b^\star \in \arg\min_{b \in \mathcal{B}_{\text{near}}(e)} c(b). \tag{4}$$

An alternative full-channel health protocol filters bundles by source-reference consistency and obtains a default bundle from validation EMQ error. A learned bundle replaces that default only above its advantage threshold; if all bundles fail the health filter, the registered fallback set is used. The protocol selects exactly one of these measurement rules, with its own diagnostic permissions and costs (Appendix A.5).

After selecting $b^\star$, existing features are reused and fitting observations are acquired at checkpoints $n_0 < n_1 < \cdots < n_T$. A feature-independent protocol follows a fixed random order. A feature-driven protocol selects among available Random, Margin, Entropy, BALD, BADGE, CoreSet, and EPIG actions using visible candidate information; its utility target is the negative proportion error after the next batch. Scoring measurements enter the ledger. Feature-driven acquisition is treated as a separate protocol and does not presume representativeness of the selected fitting set (Appendix A.5).

At a nonterminal checkpoint, gain models predict $G_t = L_{\text{keep},t} - L_{\text{more},t}$ for the selected next batch. Their mean μ_t and disagreement σ_t determine

$$\text{CONTINUE} \iff \mu_t - \kappa\sigma_t > \tau_G, \quad |P_t| > 0, \quad C_t^{\text{obs}} + \Delta C_t \le C_{\max}. \tag{5}$$

Otherwise acquisition stops and downstream inference continues. The bundle remains fixed within this sequential protocol. Gain supervision matches the batch schedule and out-of-fold acquisition policy; multiple-checkpoint controls are calibrated on complete validation trajectories. Stopping reasons, including resource exhaustion and numerical failure, are retained.

4.3 Adaptation and Quantification

Choose whether and how to adapt. Candidates comprise the unchanged classic source model, CORAL, a neural source model, and dependency-compatible TENT-, EATA-, SAR-, SHOT-, CoTTA-, and COME-style updates. Each tool fits only on the final fitting set I, then returns probabilities and representations paired with its registered fixed source-reference context. Appendix A.6 specifies the associated compatibility assumptions. Offline candidate a is supervised jointly by a common reference quantifier (EMQ) and known-class Macro-F1 $F_{u,a}$ on an independent prediction set:

$$L_{u,a}^{\text{adapt}} = \|\widehat{\beta}_{u,a}^{\text{ref}} - \beta_u\|_1 + \lambda_{\text{cls}}(1 - F_{u,a}). \tag{6}$$

This is a stage-local proxy for adaptation selection, not the terminal chain objective; Section 5.4 evaluates the composed endpoint. Equation (3) uses the classic source model as fallback. The selected probabilities, features, and references establish a new quantification context. The tools have their own update objectives; the routing loss does not replace them (Appendices A.6 and A.14).

Select the proportion estimator. Within that context, candidates are CC, PCC, ACC, PACC, BBSE, RLLS, EMQ, a CPMCN-style solver, HDy-OVR, DyS-OVR, and a reference-aligned GMM. They

cover counting, confusion correction, probability matching, distribution matching, and mixture fitting. Candidate predictors learn

$$L_{u,q}^{\text{quant}} = \|\widehat{\beta}_{u,q} - \beta_u\|_1. \tag{7}$$

The loss rule selects $q^\star$ with EMQ as fallback and returns $\widehat{\beta}$, separately from preliminary diagnostic EMQ. Candidates use registered source references and produce simplex-valued estimates (Appendix A.7).

4.4 Classification Policy Selection

Classification preserves ordinary predictions as an actionable choice. Let p_S denote calibrated source probabilities and p_A the selected adaptation probabilities, with fixed reference priors π_S and π_A. Define

$$\mathcal{C}(p; \widehat{\beta}, \pi)_k = \frac{p_k \widehat{\beta}_k / \pi_k}{\sum_\ell p_\ell \widehat{\beta}_\ell / \pi_\ell}. \tag{8}$$

The six policies use p_S, $\mathcal{C}(p_S; \widehat{\beta}, \pi_S)$, p_A, $\mathcal{C}(p_A; \widehat{\beta}, \pi_A)$, a source ensemble, and an adapted ensemble. The source ensemble equally averages the classic and available neural source probabilities. The adapted ensemble equally averages p_A, its prior-corrected version, and p_S, retaining duplicate terms with their prescribed weights. Weights are fixed without target labels. Thus retaining an ordinary classifier, correcting proportions, adapting, and combining adaptation with correction are separate choices.

Utility predictors $\widehat{V}_c(e)$ target known-class Macro-F1 on offline independent prediction sets. For a fixed baseline c_0 selected on offline training tasks and $c^+ \in \arg\max_c \widehat{V}_c(e)$,

$$c^\star = \begin{cases} c^+, & \widehat{V}_{c^+}(e) - \widehat{V}_{c_0}(e) > \delta_{\text{class}}, \\ c_0, & \text{otherwise.} \end{cases} \tag{9}$$

Return $p_i^\star = p_{c^\star}(\cdot \mid x_i)$ and $\widehat{y}_i = \arg\max_k p_{i,k}^\star$. Unknown samples retain known-class conditional predictions and a separate OOD flag (Appendix A.8).

4.5 Risk and Open-world Analysis

Uncertainty and selective status. Holding the selected representation and quantifier fixed, bootstrap refits produce componentwise proportion intervals. If all refits fail, a soft-count Dirichlet fallback is explicitly marked as a different interval type. A separate model of final proportion error combines predicted mean, model scale, and validation residuals into an empirical risk radius r_U. Evidence distance to training references determines support status. States are reliable (relative low predicted risk within reference support), cautious, high risk, and out of support. The last two produce a terminal abstention flag; the others produce a stop flag. These flags do not imply that samplewise rejection has been performed. Intervals, risk radius, support, and terminal status are recorded separately. The radius is an empirical selective-risk summary rather than a distribution-free coverage guarantee (Appendix A.9).

Three distinct open-world outputs. Available MSP, ODIN, Energy, Mahalanobis, ReAct, and ViM scores use the frozen source neural model. An offline supervised fusion model combines scores standardized against source references. Score-distribution summaries and the current classification state form x_O. Direct unknown-mass regression and a separate task-existence classifier output

$$\widehat{\rho} = \text{clip}(f_\rho(x_O), 0, \rho_{\max}), \qquad p_{\text{open}} = d_O(e_O), \tag{10}$$

where e_O transforms $[x_O, \widehat{\rho}]$. Individual OOD scores, task-level existence, and estimated mass have distinct meanings (Appendix A.10).

Screening and optional discovery. Screening uses source-reference false-positive quantiles or the top $\text{round}(|J|\widehat{\rho})$ scores. It first creates OOD flags U, then removes confident, source-prototype-compatible known samples to obtain the clustering pool U^{clean}. Both sets are retained. A closed-set-leaning task selects a no-discovery action. An open-set-leaning task first selects a detector

and screening rule by predicted metrics, then considers K-means, GMM-BIC, HDBSCAN, a semi-supervised K-means core, or NONE. Nonempty discovery actions must satisfy

$$\widehat{\mathrm{FPR}} \le \alpha_{\max}, \quad \widehat{\mathrm{Contam}} \le \gamma_{\max}, \quad \widehat{\mathrm{Coverage}} \ge \zeta_{\min}, \quad \widehat{\mathrm{PairF1}} \ge \xi_{\min}. \tag{11}$$

Predicted cluster quality ranks feasible actions. No feasible action, an undersized cleaned pool, or clustering failure returns no discovery. These are constraints on predicted metrics, and residual clusters do not acquire semantic class labels automatically (Appendix A.11).

4.6 Execution and Outputs

Dependencies determine what must be recomputed: new fitting observations update evidence; a new representation rebuilds quantification context; a new proportion estimate regenerates prior-corrected candidates; and a new detector or screen rebuilds the residual pool. Only compatible intermediate objects are reused. Optional verification compares visible before/after risk, uncertainty, and fit statistics, returns supported, unsupported, or inconclusive, and updates registered re-entry memory. It never reads true target loss. Verification and sequential-acquisition return edges are independently specified by the protocol; a single forward pass does not require per-action verification.

Let O_t contain the acquired target sample–channel pairs. Each pair is charged once:

$$C_t^{\mathrm{obs}} = \sum_{(i,s)\in O_t} c_s, \qquad \Delta C_t = \sum_{(i,s)\in O_{t+1}\setminus O_t} c_s. \tag{12}$$

Candidate screening, proportion fitting, prediction measurements, and computation retain their respective accounting scopes. The terminal object contains proportions, intervals and risk, classification probabilities and labels, unknown mass, OOD flags, residual clusters, stop and failure states, costs, and a decision record of actual actions, predicted advantages, and fallbacks (Appendix A.12). Appendix A.13 specifies training and execution; each experiment must identify the tools, checkpoints, and verification protocol actually enabled.

Algorithm 1 StatOrch inference for one target task

1: Acquire initial observations; compute source predictions, preliminary proportions, and permitted diagnostics.
2: Select the measurement bundle. At enabled resource checkpoints, choose sampling and continuation actions; update evidence after acquisition and retain the final fitting set.
3: Compare unchanged and available adapted models; fit the selected action on the fitting set and fix its probability, representation, and reference context.
4: Select a quantifier and compute $\widehat{\beta}$; regenerate ordinary, corrected, adapted, and ensemble classification candidates and select the final policy.
5: Compute proportion intervals, risk radius, reference support, and selective-output status.
6: Compute OOD and fusion scores, $\widehat{\rho}$, and task existence; choose a detector and screen, then construct and clean the residual pool.
7: Discover structure only when predicted constraints and actual pool size permit; otherwise return no discovery. If verification is enabled, follow only registered return edges.
8: Save proportions, classification, open-world results, uncertainty, costs, and decisions; return terminal status.

5 Experiments

5.1 Protocols and Metrics

Controlled stage routing and the executed short chain provide the core evidence for conditional selection and dependent inference. Complementary archival resource records contain 1,232 held-out units: Seen (384), Severity (336), Unseen mechanism (192), Unseen combination (192), and Discovery pairs (128). Target composition, injected mechanisms, and labels are evaluator-only.

The archival instance fixes Raw Classic, EMQ, prior-corrected source classification, and a legacy hurdle unknown-mass head. Controlled stage and chain evaluations use separate cohorts with

Table 1: **Prediction and cost for the archival measurement/resource instance.** Results use 1,232 held-out units; OOD metrics use 360 eligible units. Bold marks the best predictive value. Costs follow the recorded measurement ledger.

Method	β $L_1\downarrow$	Macro-F1↑	ρ err.↓	AUROC↑	AUPR↑	FPR95↓	Cost↓
ValidationMatched	0.294	0.720	0.128	0.669	0.543	0.692	672†
EMQ	0.292	0.710	0.121	0.671	0.543	0.688	672†
RLLS	0.348	0.712	0.123	0.666	0.540	0.691	672†
PACC	0.289	0.711	0.122	0.669	0.542	0.691	672†
CPMCN	0.294	0.711	0.120	0.671	0.543	0.688	672†
NearCost	0.284	0.721	0.130	0.671	0.544	0.689	896†
HighBudget	0.268	0.721	0.130	0.670	0.543	0.688	1344†
STATORCH	**0.224**	**0.773**	**0.091**	**0.716**	**0.594**	**0.638**	802

† Preset acquisition budget; STATORCH reports the realized mean. Its adaptation, quantification, and classification heads are fixed in this evaluation.

Table 2: **Stage routing against validation-selected fixed actions.** All columns share a held-out cohort and controlled reduced candidate bank. BestFixed-Val freezes one validation-selected action per stage. Bold: best deployable mean; Oracle is evaluator-only.

Stage	Split	Fallback	BestFixed-Val	HealthRule	Greedy	Conservative	Oracle
Adapt. loss↓	seen	0.0703	0.0640	0.0561	**0.0540**	0.0566	0.0525
	unseen	0.6419	**0.5589**	0.5708	**0.5589**	**0.5589**	0.5589
Quant. $L_1\downarrow$	seen	0.2215	0.2089	0.1794	**0.1682**	0.1723	0.1410
	unseen	0.1981	0.1904	**0.1719**	0.1793	0.1766	0.1331
Class. F1↑	seen	0.6231	0.6231	0.6342	**0.6621**	0.6596	0.6801
	unseen	0.6697	0.6697	0.7008	**0.7207**	0.6697	0.7383
Open loss↓	seen	0.3216	0.2191	0.1726	**0.1427**	0.1465	0.1282
	unseen	0.4727	0.3991	0.3991	**0.3715**	**0.3715**	0.3092

Seen units = 96/120/96/120 by stage; unseen = 48 per stage. Paired intervals and historical broader-bank results: Appendix D. Protocol and fit scope: Appendix H.

preserved fitted objects and paired outputs. The chain uses fixed observations; random-order sampling and optional feature-driven acquisition are separate protocols. No experiment jointly evaluates every stage in Section 4.

We report known-class conditional β-L_1, known-class Macro-F1, and recorded measurement cost; archival open-world metrics use their stated eligible subsets. Appendix B specifies instances and metrics, and Appendix H distinguishes controlled evidence from historical reconstruction limits.

5.2 HISTORICAL RESOURCE-CONTROL ENDPOINT

Archival endpoints provide complementary resource evidence. Under the preserved measurement ledger, the historical instance records 21.1% lower prevalence error, 5.26 percentage points higher Macro-F1, and 10.5% lower recorded measurement cost than NearCost (Table 1). We interpret these archival records separately from the controlled evaluations below, which support the main selection and composition claims.

The split results clarify where the gain occurs (Appendix C). On Unseen combination, classification and recorded cost favor the archival instance, while prevalence error is 0.522 versus 0.512/0.501 for NearCost/HighBudget. Population-proportion recovery and individual prediction therefore remain distinct evaluation targets.

5.3 CONDITIONAL TOOL SELECTION

We conduct a controlled stage-routing evaluation comparing every deployable selector with a validation-selected fixed action on the same held-out cohort and candidate bank (Table 2). Val-

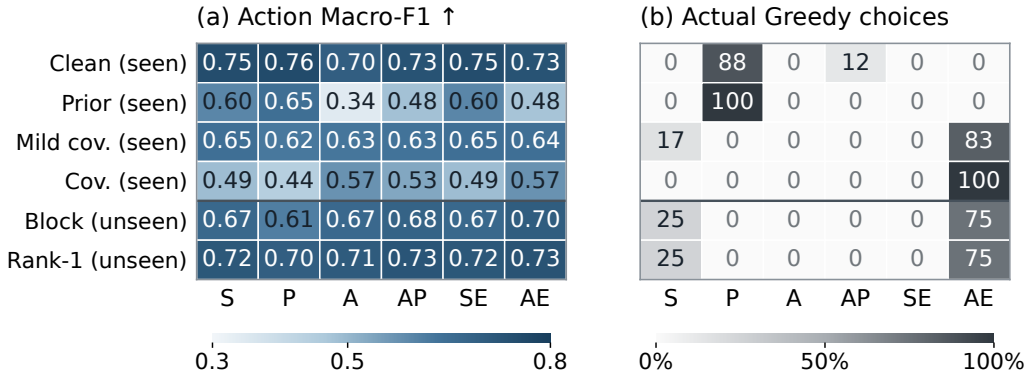


Figure 2: **Different regimes favor different classification actions.** The historical suite has six scenarios with 24 tasks each: action-level F1 (left) and Greedy's actual choice percentages (right). S/P: source/plain and prior-corrected; A/AP: adapted/plain and prior-corrected; SE/AE: source/adapted ensembles. Appendix D reports all policies and choices.

Table 3: **Short-chain routing and one-shot selection.** Same observations, training/validation tasks and $2 \times 6 \times 6$ named bank. One-shot selects a complete tuple from initial evidence; stagewise refreshes evidence. Original fit (seed 20260926); bold marks the best mean.

		BestFixed	Stagewise		One-shot	
Split	Metric	Val	Cons.	Greedy	Cons.	Greedy
Seen	β-L_1 ↓	0.2206	0.2058	**0.2005**	0.2054	0.2008
	Macro-F1↑	0.7735	0.7915	**0.7953**	0.7925	**0.7953**
	Joint loss↓	0.3565	0.3309	**0.3234**	0.3300	0.3237
Unseen	β-L_1 ↓	0.1587	0.1587	0.1586	0.1587	**0.1578**
	Macro-F1↑	**0.7248**	0.7178	0.7225	**0.7248**	0.7230
	Joint loss↓	**0.3238**	0.3280	0.3251	**0.3238**	0.3240

Joint loss: β-$L_1 + 0.60(1 - \text{F1})$; 288 seen/96 unseen tasks. Cons.: validation-calibrated conservative selector. All four fits and paired contrasts: Appendix H.

idation alone determines fixed actions and switching margins. Greedy improves seven of eight stage/split means and ties on unseen adaptation. Its paired 95% intervals favor routing in six comparisons: all four seen stages, unseen classification, and unseen open-world (Appendix D). Conservative improves all four seen stages.

Two evidence layers separate conditional value from portfolio breadth. The controlled reduced bank isolates selection against BestFixed-Val; historical neural-inclusive suites characterize specialization in a broader portfolio. Figure 2 connects historical classification performance to actual choices: prior correction is strongest under prior shift, adaptation under strong covariance shift, and an adapted ensemble under the unseen regimes.

5.4 Dependency-Aware Composition

We execute adaptation → quantification → classification on fixed ALL observations, with 96 unlabeled fit and 96 disjoint audit samples per task. BestFixed-Val selects a complete tuple on 192 validation tasks by minimizing β-$L_1 + 0.60(1 - \text{F1})$. One-shot predicts this loss for all 72 named tuples from initial evidence and selects once; stagewise refreshes downstream contexts. Both use the same data, bank and forest settings per decision model. Routed training uses task-level out-of-fold upstream choices.

Table 4: **External sensor acquisition at representative operating budgets.** Bold marks the best displayed predictive value in each regime.

Dataset	Regime	Budget	Policy	β-L_1 ↓	F1↑	Cost
REALDISP	moderate	3	StatOrch	**0.691**	**0.411**	1,716.00
		3	GSM-AFA	0.965	0.289	1,524.00
		3	BestFixed	1.054	0.214	1,500.00
REALDISP	severe	3	StatOrch	1.406	0.130	1,716.00
		3	GSM-AFA	1.461	0.097	1,524.00
		3	BestFixed	**1.233**	**0.142**	1,500.00
Gas Drift	temporal	4	GSM-AFA	**0.898**	**0.256**	2,031.25
		4	StatOrch	0.995	0.232	2,500.00
		4	BestFixed	1.017	0.216	2,000.00

Budget is sensor count; costs are mean sensor–sample pairs, including pilots, for 500 evaluation observations. Equal sensor counts do not imply equal total costs. GSM-AFA denotes the archived policy. Full curves: Appendix G.

Conditional pipeline selection across fitting seeds. Conservative lowers β-L_1 from 0.2206 to 0.2058 and raises F1 from 0.7735 to 0.7915 (Table 3). Across seeds 20260926–20260929, its seen joint-loss improvement over fixed remains 0.0257. One-shot Conservative attains 0.3300 versus stagewise's 0.3309; their original-fit paired difference is $+0.0009$ with a 95% interval spanning zero. Greedy variants are also close. The similar one-shot losses support conditional pipeline selection without identifying an advantage from staged re-evaluation. The experiment evaluates the dependency-aware chain as a composed policy rather than attributing its gain to individual stages. Appendix H reports every fit, paired interval, and original action-path summary.

5.5 Measurement and Resource Control

At matched recorded cost 432, both evidence-conditioned policies improve over BestFixed. StatOrch has the lowest prevalence error on Standard; HealthRule is strongest on Shift and in Macro-F1 (Appendix E). Within the separate A0/A1/A2 study, most endpoint change occurs when measurement routing is introduced, whereas checkpoint stopping primarily changes measurement cost; the routing effect concentrates in severe Sensor-4 failure (Appendix E). Separate re-entry experiments reduce repeated unsupported reuse (Appendix F).

5.6 External Sensor Stress Tests

Archival real-sensor records evaluate evidence-conditioned acquisition under separate protocols. On moderate REALDISP displacement, StatOrch improves both prevalence estimation and classification over fixed sensing at the representative budget (Table 4). Severe displacement favors BestFixed, while archived GSM-AFA performs best on temporal Gas Drift among the displayed policies. These results show regime-dependent acquisition performance rather than uniform dominance. All methods share frozen downstream inference; full cost–accuracy curves appear in Appendix G, with environment-level results in the structured supplement.

6 Discussion and Conclusion

StatOrch connects observable evidence to statistical intervention choice through stage-specific effect prediction and dependency-aware context updates. Controlled comparisons show gains over validation-selected fixed actions, and the executed adaptation–quantification–classification chain improves the joint endpoint on seen mechanisms across four fitting seeds on a fixed cohort. Matched one-shot selection attains similar losses, supporting conditional pipeline selection within the evaluated family. Generalization beyond the trained mechanism families remains an open direction. Complementary resource, re-entry, and external studies characterize operational behavior under their respective protocols.

AI USE STATEMENT

AI tools were used for retrieval and discovery (e.g., finding related work) and for generating synthetic datasets.

ETHICS STATEMENT

The empirical study uses controlled synthetic tasks and existing sensor datasets. Unknown-mass and out-of-distribution scores indicate statistical mismatch; they do not identify people or certify semantic categories. Decisions in consequential applications require domain-specific validation beyond the protocols studied here. The supplement describes the source datasets and does not redistribute external raw archives.

REPRODUCIBILITY STATEMENT

Appendix A specifies the method; Appendices B–H identify evaluated instances and statistical scope. Controlled stage and chain evaluations release source/router checkpoints, train/validation matrices, frozen selections, observations, and paired outputs. The one-shot follow-up includes all four fitting seeds, complete-tuple supervision and inference-before-evaluation records; it reuses the original raw cohort. Historical records lack the original primary runner, fitted controller, complete trajectories, and some resampling inputs. Historical resource results use the preserved measurement ledger; the original execution does not support reconstruction of a complete modern end-to-end sensing ledger. The supplement retains all detailed result tables and records moved out of the PDF, with a file index and display-rebuilding scripts. Formatting these records is distinct from rerunning their experiments.

REFERENCES

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1906.03671.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1903.09734.

Oresti Banos, Mate Toth, and Oliver Amft. Realdisp activity recognition dataset, 2014. URL https://archive.ics.uci.edu/dataset/305/realdisp+activity+recognition+dataset.

Antonio Bella, Cesar Ferri, Jose Hernandez-Orallo, and Maria Jose Ramirez-Quintana. Quantification via probability estimators. In *Proceedings of the 10th IEEE International Conference on Data Mining*, 2010. doi: 10.1109/ICDM.2010.75. URL https://ieeexplore.ieee.org/document/5694031/.

Freddie Bickford Smith, Andreas Kirsch, Sebastian Farquhar, Yarin Gal, Adam Foster, and Tom Rainforth. Prediction-oriented bayesian active learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 7331–7348, 2023. URL https://proceedings.mlr.press/v206/bickfordsmith23a.html.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 160–172. Springer, 2013. doi: 10.1007/978-3-642-37456-2_14. URL https://link.springer.com/chapter/10.1007/978-3-642-37456-2_14.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x. URL https://academic.oup.com/jrsssb/article/39/1/1/7027539.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/11d0e6287202fced83f79975ec59a3a6-Abstract.html.

George Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17:164–206, 2008. doi: 10.1007/s10618-008-0097-y. URL https://link.springer.com/article/10.1007/s10618-008-0097-y.

Ignacio Gómez-Casares, Beatriz Pateiro-López, Brais González-Rodríguez, and Julio González-Díaz. ASML: An R package for algorithm selection with machine learning. *The R Journal*, 17(4): 217–237, 2025. doi: 10.32614/RJ-2025-045. URL https://journal.r-project.org/articles/RJ-2025-045/.

Víctor González-Castro, Rocío Alaiz-Rodríguez, and Enrique Alegre. Class distribution estimation based on the Hellinger distance. *Information Sciences*, 218:146–164, 2013. doi: 10.1016/j.ins.2012.05.028. URL https://doi.org/10.1016/j.ins.2012.05.028.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Shurui Gui, Xiner Li, and Shuiwang Ji. Active test-time adaptation: Theoretical analyses and an algorithm. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/c150ebe1b9d1ca0eb61502bf979fa87d-Abstract-Conference.html.

Osman Berke Guney, Ketan Suhaas Saichandran, Karim Elzokm, Ziming Zhang, and Vijaya B Kolachalama. Active feature acquisition via explainability-driven ranking. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20748–20765, 2025. URL https://proceedings.mlr.press/v267/guney25a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1610.02136.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011. URL https://arxiv.org/abs/1112.5745.

Dapeng Hu, Mi Luo, Jian Liang, and Chuan-Sheng Foo. Towards reliable model selection for unsupervised domain adaptation: An empirical study and a certified baseline. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4316. URL https://proceedings.nips.cc/paper_files/paper/2024/hash/f50cebc22663df45ce619645bfabb3b3-Abstract-Datasets_and_Benchmarks_Track.html.

Taejong Joo and Diego Klabjan. IW-GAE: Importance weighted group accuracy estimation for improved calibration and model selection in unsupervised domain adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22509–22529, 2024. URL https://proceedings.mlr.press/v235/joo24a.html.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://arxiv.org/abs/1807.03888.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6450–6459. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/li21p.html.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6028–6039, 2020. URL https://proceedings.mlr.press/v119/liang20a.html.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1706.02690.

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3122–3130, 2018. URL https://proceedings.mlr.press/v80/lipton18a.html.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html.

Cynthia Moreira Maia, Lucas B. V. de Amorim, George D. C. Cavalcanti, and Rafael M. O. Cruz. MetaML: a multi-label meta-learning approach for pipeline recommendation. *Machine Learning*, 114:278, 2025. doi: 10.1007/s10994-025-06909-8. URL https://link.springer.com/article/10.1007/s10994-025-06909-8.

André Maletzke, Denis dos Reis, Everton Cherman, and Gustavo Batista. DyS: A framework for mixture models in quantification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4552–4560, 2019. doi: 10.1609/aaai.v33i01.33014552. URL https://ojs.aaai.org/index.php/AAAI/article/view/4376.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16888–16905, 2022. URL https://proceedings.mlr.press/v162/niu22a.html.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2302.12400.

Alexander Luke Ian Norcliffe, Changhee Lee, Fergus Imrie, Mihaela Van Der Schaar, and Pietro Lio. Stochastic encodings for active feature acquisition. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 46784–46814, 2025. URL https://proceedings.mlr.press/v267/norcliffe25a.html.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002. doi: 10.1162/089976602753284446. URL https://doi.org/10.1162/089976602753284446.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. doi: 10.1214/aos/1176344136. URL https://projecteuclid.org/journals/annals-of-statistics/volume-6/issue-2/Estimating-the-Dimension-of-a-Model/10.1214/aos/1176344136.full.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1708.00489.

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. doi: 10.1609/aaai.v30i1.10306. URL https://ojs.aaai.org/index.php/AAAI/article/view/10306.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/01894d6f048493d2cacde3c579c315a3-Abstract.html.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7492–7501, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Vaze_Generalized_Category_Discovery_CVPR_2022_paper.html.

Alexander Vergara. Gas sensor array drift dataset, 2012. URL https://archive.ics.uci.edu/dataset/224/gas+sensor+array+drift+dataset.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.10726.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4921–4930, 2022a. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_ViM_Out-of-Distribution_With_Virtual-Logit_Matching_CVPR_2022_paper.html.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022b. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Continual_Test-Time_Domain_Adaptation_CVPR_2022_paper.html.

Hongwei Wen, Annika Betken, and Hanyuan Hang. Class probability matching with calibrated networks for label shift adaption. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7e3767db483c942b883eb4f8cfb74e31-Abstract-Conference.html.

Kaichao You, Ximei Wang, Mingsheng Long, and Michael Jordan. Towards accurate model selection in deep unsupervised domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7124–7133, 2019. URL https://proceedings.mlr.press/v97/you19a.html.

Qingyang Zhang, Yatao Bian, Xinke Kong, Peilin Zhao, and Changqing Zhang. COME: Test-time adaption by conservatively minimizing entropy. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2ea18fdc667e0ef2ad82b2b4d65147ad-Abstract-Conference.html.

A Complete Method Specification

This appendix specifies the statistical quantities, candidate operators, training and validation rules, observation permissions, and execution protocols of StatOrch. Numerical settings belong to the registered configuration of each experiment.

A.1 Problem and Observation Protocol

Target quantities. The known classes are $\mathcal{Y} = \{1, \dots, K\}$; all unknown classes are represented by $\perp$. A task corresponds to a fixed population P_T, with

$$\beta_k = P_T(Y = k \mid Y \in \mathcal{Y}), \quad \rho = P_T(Y = \perp), \quad \beta \in \Delta^{K-1}, \quad 0 \le \rho < 1, \tag{13}$$

where $\Delta^{K-1} = \{v \in \mathbb{R}^K : v \ge 0, \mathbf{1}^\top v = 1\}$. The joint class masses are

$$\pi_T = ((1-\rho)\beta^\top, \rho)^\top. \tag{14}$$

Closed-set tasks have $\rho = 0$. Known-class conditional proportions, known-class predictions, and unknown-sample screening are separate outputs. Residual clusters are not directly added to the source class set.

Data partitions and information permissions. Let $\mathcal{D}_S = \{(x_i^S, y_i^S)\}_{i=1}^{N_S}$, partitioned into source training data $\mathcal{D}_S^{\mathrm{tr}}$ and reference calibration data $\mathcal{D}_S^{\mathrm{ref}}$. Offline tasks are partitioned into $\mathcal{T}$ for learning controllers and $\mathcal{V}$ for selecting validation parameters. Each target task has fitting indices I_t, candidate acquisition indices P_t, and prediction indices J, satisfying

$$I_t \cap J = \varnothing, \qquad P_t \cap J = \varnothing. \tag{15}$$

The current fitting set and candidate pool also satisfy $I_t \cap P_t = \varnothing$. Adaptation and proportion fitting use only I_t. Classification, open-world analysis, and their evidence may use prediction covariates X_J permitted by the protocol. Prediction labels Y_J, true β, unknown mass ρ, and mechanism identities are restricted to offline supervision or final evaluation.

Let O_t be the target sample–channel pairs acquired through stage t, and $\mathcal{K}_{\mathrm{off}}$ all frozen source models, reference objects, and offline controllers. The online information is

$$\mathscr{F}_t = \sigma\Big(\mathcal{K}_{\mathrm{off}}, \{x_i^{(s)} : (i,s) \in O_t\}, \{a_j, o_j^{\mathrm{obs}} : j < t\}, \xi\Big), \tag{16}$$

where o_j^{obs} is an observable tool output and ξ contains prespecified randomization variables. Every online action a_t is $\mathscr{F}_t$-measurable.

Measurements, tools, and protocol. Channels form $\mathcal{S}$ and candidate bundles $\mathcal{B} = \{b_1, \dots, b_M\}$ satisfy $b \subseteq \mathcal{S}$. Each channel has cost $c_s > 0$ and $c(b) = \sum_{s \in b} c_s$. A candidate tool is defined by its input type, output quantities, fitting and prediction operators, applicability conditions, and failure rules. $\mathcal{A}_j(e)$ contains the candidates meeting the input and dependency requirements of stage j.

The execution protocol $\mathcal{P}$ fixes b_0, n_0, checkpoints $n_0 < n_1 < \cdots < n_T$, measurement and sampling rules, candidate tools, measurement budget $C_{\max}$, and any verification operators. Full-channel diagnostics use $b_0 = \mathcal{S}$; reference-bundle diagnostics use a prespecified $b_0 = b_{\mathrm{anc}}$. Their information permissions and diagnostic costs are recorded separately.

A.2 Source Models and Statistical References

For every $b \in \mathcal{B}$, train candidate classifiers $\mathcal{F}_b$ on $\mathcal{D}_S^{\mathrm{tr}}$ and select by the source validation criterion

$$f_b^\star \in \arg\max_{f \in \mathcal{F}_b} \{\mathrm{MacroF1}(f) - \lambda_E \,\mathrm{ECE}(f) - \lambda_B \,\mathrm{Brier}(f)\}. \tag{17}$$

Classic candidates are Logistic, LDA, LinearSVM, RBFSVM, KNN, RandomForest, ExtraTrees, MLP, and XGBoost when its dependencies are available. Calibrators comprise Identity, Temperature, PlattOVR, IsotonicOVR, VectorScaling, and Dirichlet; source probability calibration follows standard held-out calibration practice (Guo et al., 2017). With reference indices I_S^{ref} and $N_{\mathrm{ref}} = |I_S^{\mathrm{ref}}|$, choose

$$g_b^\star \in \arg\min_{g \in \mathcal{C}_b} -\frac{1}{N_{\mathrm{ref}}} \sum_{i \in I_S^{\mathrm{ref}}} \log[g \circ f_b^\star]_{y_i^S}(x_i^{S,b}), \qquad p_S^b = g_b^\star \circ f_b^\star. \tag{18}$$

The source prior is $\pi_{S,k} = |\mathcal{D}_S^{\mathrm{tr}}|^{-1} \sum_{i \in I_S^{\mathrm{tr}}} \mathbf{1}\{y_i^S = k\} > 0$. Set $N_k^{\mathrm{ref}} = |\{i \in I_S^{\mathrm{ref}} : y_i^S = k\}|$. Under the same reference prediction mapping, the hard and soft confusion matrices and joint matrices are

$$A_{jk}^b = \frac{1}{N_k^{\mathrm{ref}}} \sum_{i \in I_S^{\mathrm{ref}} : y_i^S = k} \mathbf{1}\left\{ \arg\max_{\ell} p_S^b(\ell \mid x_i^{S,b}) = j \right\}, \tag{19}$$

$$B_{jk}^b = \frac{1}{N_k^{\mathrm{ref}}} \sum_{i \in I_S^{\mathrm{ref}} : y_i^S = k} p_S^b(j \mid x_i^{S,b}), \quad C_A^b = A^b \operatorname{diag}(\pi_S), \quad C_B^b = B^b \operatorname{diag}(\pi_S). \tag{20}$$

For the source representation $z = \phi_b(x^b)$, retain its mean, covariance, class prototypes μ_k^b, pooled within-class covariance, and reference prediction samples. Neural tools additionally retain the source neural model, prototypes, feature subspaces, and OOD reference thresholds. All source models and reference objects remain fixed throughout one target task.

A.3 Statistical Evidence, Reference Transforms, and Soft Diagnostics

Preliminary estimation and predictive uncertainty. For an observed set I, write $n = |I|$ and $p_{ik} = p_S^b(k \mid x_i^b)$. Compute

$$c_i = \max_k p_{ik}, \qquad H_i = -\sum_k p_{ik} \log p_{ik}, \qquad (\bar{c}, \mathrm{sd}(c), \bar{H}, \mathrm{sd}(H)). \tag{21}$$

Preliminary quantification uses EMQ (Saerens et al., 2002), initialized at $\beta^{(0)} = \pi_S$ and iterated as

$$r_{ik}^{(m)} = \frac{p_{ik} \beta_k^{(m)} / \pi_{S,k}}{\sum_\ell p_{i\ell} \beta_\ell^{(m)} / \pi_{S,\ell}}, \qquad \beta_k^{(m+1)} = \frac{1}{n} \sum_{i \in I} r_{ik}^{(m)}. \tag{22}$$

Stop when $\|\beta^{(m+1)} - \beta^{(m)}\|_1 < \varepsilon_{\mathrm{EM}}$ or after R_{EM} iterations. For diagnostic quantifiers $\mathcal{Q}_{\mathrm{diag}}$, define

$$D_Q = \max_{q, q' \in \mathcal{Q}_{\mathrm{diag}}^{\mathrm{ok}}} \|\widehat{\beta}^{(q)} - \widehat{\beta}^{(q')}\|_1, \tag{23}$$

where $\mathcal{Q}_{\mathrm{diag}}^{\mathrm{ok}}$ contains tools returning successful, finite simplex vectors. The initial diagnostic set contains PCC, PACC, EMQ, BBSE, RLLS, and CPMCN.

Draw B_{boot} bootstrap samples with replacement from the observed samples, producing EMQ estimates $\widehat{\beta}^{[r]}$. Component widths and summaries are

$$\begin{aligned} W_k &= Q_{1-\alpha_{\mathrm{boot}}/2}(\widehat{\beta}_k^{[r]}) - Q_{\alpha_{\mathrm{boot}}/2}(\widehat{\beta}_k^{[r]}), \\ \bar{W} &= K^{-1} \sum_k W_k, \qquad W_{\max} = \max_k W_k. \end{aligned} \tag{24}$$

For R_{sub} subsamples drawn without replacement at fraction q_{sub}, use

$$D_{\mathrm{sub}} = \max_{r, r'} \|\widehat{\beta}^{\langle r \rangle} - \widehat{\beta}^{\langle r' \rangle}\|_1. \tag{25}$$

Fit residuals and distribution differences. Let $\widehat{q}_j = n^{-1} \sum_{i \in I} \mathbf{1}\{\arg\max_k p_{ik} = j\}$. Compute

$$D_{\mathrm{res}} = \|A^b \widehat{\beta}^{\mathrm{EMQ}} - \widehat{q}\|_2, \qquad T_{\mathrm{P}} = \sum_j \frac{(n\widehat{q}_j - n[A^b \widehat{\beta}^{\mathrm{EMQ}}]_j)^2}{\max\{n[A^b \widehat{\beta}^{\mathrm{EMQ}}]_j, \varepsilon_{\mathrm{P}}\}}. \tag{26}$$

The Pearson reference-tail score $p_{\mathrm{P}} = 1 - F_{\chi^2_{K-1}}(T_{\mathrm{P}})$ is used only as a fit-diagnostic coordinate. Let $Z_S = \{z_i^S\}_{i=1}^m$ and $Z_T = \{z_i^T\}_{i=1}^n$. With $k_\gamma(x, y) = \exp(-\gamma\|x - y\|_2^2)$, compute

$$D_{\mathrm{MMD}} = \frac{1}{m^2}\sum_{i,j} k_\gamma(z_i^S, z_j^S) + \frac{1}{n^2}\sum_{i,j} k_\gamma(z_i^T, z_j^T) - \frac{2}{mn}\sum_{i,j} k_\gamma(z_i^S, z_j^T), \tag{27}$$

$$D_{\mathrm{en}} = \frac{2}{mn}\sum_{i,j}\|z_i^S - z_j^T\|_2 - \frac{1}{m^2}\sum_{i,j}\|z_i^S - z_j^S\|_2 - \frac{1}{n^2}\sum_{i,j}\|z_i^T - z_j^T\|_2, \tag{28}$$

$$D_{\mathrm{W}} = \frac{1}{d}\sum_{j=1}^{d} W_1(\widehat{P}_{S,j}, \widehat{P}_{T,j}), \quad D_\mu = \|\widehat{\mu}_S - \widehat{\mu}_T\|_2, \quad D_\Sigma = \frac{\|\widehat{\Sigma}_S - \widehat{\Sigma}_T\|_F}{d}. \tag{29}$$

The kernel discrepancy is the empirical MMD statistic (Gretton et al., 2012). The kernel scale is determined by the median of pooled reference distances; sampling caps and random seeds are fixed by the protocol. The domain-discrimination score is the training AUC of a Logistic discriminator fitted to pooled source and target representations. Predictive-distribution discrepancy is

$$\begin{aligned} D_{\mathrm{JS}} &= D_{\mathrm{JS}}(\bar{p}_S, \bar{p}_T), \\ D_{\mathrm{JS}}(p, q) &= \tfrac{1}{2}D_{\mathrm{KL}}(p\|m) + \tfrac{1}{2}D_{\mathrm{KL}}(q\|m), \qquad m = \tfrac{1}{2}(p + q). \end{aligned} \tag{30}$$

Channel reference consistency. Health scores are computed only when the diagnostic protocol has acquired the relevant channels. For d_s features in channel s, define

$$D_{\mu,s} = \frac{1}{d_s}\sum_j \left|\frac{\mu_{T,sj} - \mu_{S,sj}}{\max(\sigma_{S,sj}, \varepsilon_s)}\right|, \tag{31}$$

$$D_{\sigma,s} = \frac{1}{d_s}\sum_j \left|\log\frac{\max(\sigma_{T,sj}, \varepsilon_s)}{\max(\sigma_{S,sj}, \varepsilon_s)}\right|, \quad h_s = \exp(-D_{\mu,s} - \lambda_h D_{\sigma,s}). \tag{32}$$

h_s measures marginal consistency with the source reference; it is not a probability that a physical fault has been identified.

Stage evidence mapping. The fixed feature dictionary combines current proportions, predictive statistics, disagreement, resampling summaries, and distribution differences into x_j^{raw}. Measurement uses initial statistics from the reference bundle. Resource and adaptation stages add sample size, bundle cost, and bundle indicators. Quantification uses the selected representation. Classification and risk add the selected proportions and predictive statistics; the open-world stage adds OOD score distributions.

For a coordinate x requiring transformation, use offline training reference values $r_1, \ldots, r_N$:

$$F(x) = \mathrm{clip}\left(\frac{\frac{1}{2}(N_{<x} + N_{\le x}) + \frac{1}{2}}{N + 1}, \varepsilon_q, 1 - \varepsilon_q\right), \tag{33}$$

$$\begin{aligned} Z(x) &= L_z^{-1}\,\mathrm{clip}\left(\frac{x - m}{s}, -L_z, L_z\right), \\ m &= \mathrm{median}(r_i), \qquad s = \max\{c_{\mathrm{MAD}}\,\mathrm{median}\,|r_i - m|, \varepsilon_z\}. \end{aligned} \tag{34}$$

Proportion coordinates, sample size, bundle identifiers, and other context coordinates are retained directly as specified by the stage dictionary. Denote the complete transform by ψ_j. For the diagnostic labels

$$\begin{aligned} \mathcal{H} = \{&\text{applicable}, \text{conditional mismatch}, \text{representation failure}, \\ &\text{insufficient information}, \text{out-of-reference structure}, \text{evidence conflict}\}, \end{aligned} \tag{35}$$

train separate binary ExtraTrees classifiers $d_{j,h}$ with class-balanced weights. The deployed evidence is

$$e_j = \left[\psi_j(x_j^{\mathrm{raw}}), \left(d_{j,h}(\psi_j(x_j^{\mathrm{raw}}))\right)_{h\in\mathcal{H}}\right]. \tag{36}$$

Diagnostic supervision follows a fixed mapping from offline mechanisms to labels; applicability may include tasks with only prior change. Scores need not sum to one and their maximizer is not a mandatory routing condition. Mechanism identities never enter online features.

A.4 Effect Prediction, Validation, and Stage-wise Training

Effect prediction and fallback selection. Stage j has actions $\mathcal{A}_j$ with offline task loss $L_{u,a}^{(j)}$ or utility $V_{u,a}^{(j)}$. Multi-output ExtraTrees regressors use the respective objectives

$$\begin{aligned}&\sum_{u\in\mathcal{T}}\sum_{a\in\mathcal{A}_j}[f_{j,a}(e_{u,j}) - L_{u,a}^{(j)}]^2,\\ &\sum_{u\in\mathcal{T}}\sum_{a\in\mathcal{A}_j}[v_{j,a}(e_{u,j}) - V_{u,a}^{(j)}]^2.\end{aligned} \tag{37}$$

Tree count, leaf restrictions, feature-subsampling fraction, and random seeds are configuration parameters. For a loss-based stage, let a_0 be the fixed fallback and $a^- \in \arg\min_{a\in\mathcal{A}_j(e)} \widehat{L}_{j,a}(e)$. Execute

$$\mathcal{R}_j^L(e) = \begin{cases} a^-, & \widehat{L}_{j,a_0}(e) - \widehat{L}_{j,a^-}(e) > \delta_j,\\ a_0, & \text{otherwise.}\end{cases} \tag{38}$$

For a utility-based stage, let $a^+ \in \arg\max_{a\in\mathcal{A}_j(e)} \widehat{V}_{j,a}(e)$ and execute

$$\mathcal{R}_j^V(e) = \begin{cases} a^+, & \widehat{V}_{j,a^+}(e) - \widehat{V}_{j,a_0}(e) > \delta_j,\\ a_0, & \text{otherwise.}\end{cases} \tag{39}$$

All ties follow a fixed order. A selected tool's numerical failure invokes the stage fallback. If the fallback also fails, record stage failure and terminate or return an unresolved state according to the protocol.

Validation selection. Execute every candidate control parameter $\theta \in \Theta_j$ on validation tasks. With $N_V = |\mathcal{V}| \geq 2$ and terminal losses $L_u(\theta)$, define

$$\bar{L}(\theta) = \frac{1}{N_V}\sum_{u\in\mathcal{V}} L_u(\theta), \qquad \mathrm{SE}(\theta) = \left[\frac{\sum_{u\in\mathcal{V}}(L_u(\theta) - \bar{L}(\theta))^2}{N_V(N_V-1)}\right]^{1/2}. \tag{40}$$

For $\theta_L \in \arg\min_\theta \bar{L}(\theta)$, retain

$$\Theta_j^{\mathrm{adm}} = \{\theta : \bar{L}(\theta) \leq \bar{L}(\theta_L) + \lambda_V\,\mathrm{SE}(\theta_L)\}. \tag{41}$$

Within this set, measurement and stopping controls prioritize minimum resource cost. Fallback-routing thresholds prioritize minimum switching rate, then loss. Utility rules use $-V$ in place of L.

Stage-wise training order. Train measurement, sampling and stopping, adaptation, quantification, classification, risk, and open-world modules in that order. Downstream training contexts use task-level out-of-fold upstream actions: for a partition $\mathcal{T} = \bigcup_{r=1}^F \mathcal{T}_r$, the upstream action for $u \in \mathcal{T}_r$ comes from a router fitted on $\mathcal{T} \setminus \mathcal{T}_r$. Freeze upstream outputs before generating downstream candidate results. Validation contexts follow the already trained upstream pipeline. Fit score fusion, unknown-mass regression, and open-world metric prediction in their respective supervised order. Predictions fitted and evaluated on the same training tasks are not labeled out of fold.

A.5 Measurement Selection and Sequential Resource Control

Low cost within a near-optimal error set. For indices I, define $E_{u,b}(I) = \|\widehat{\beta}^{\mathrm{EMQ}}_{u,b}(I) - \beta_u\|_1$. Let I_n be the first n indices of a prespecified nested sample sequence and abbreviate $E_{u,b}(n) = E_{u,b}(I_n)$. Measurement predictors take initial reference-bundle evidence and are supervised by $(E_{u,b}(n_0))_{b\in\mathcal{B}}$. Select

$$\begin{aligned} \mathcal{B}_{\mathrm{near}}(e) &= \{b \in \mathcal{B} : \widehat{E}_b(e) \le \min_{b'\in\mathcal{B}} \widehat{E}_{b'}(e) + \delta_B\}, \\ b^\star &\in \arg\min_{b\in\mathcal{B}_{\mathrm{near}}(e)} c(b). \end{aligned} \tag{42}$$

Calibrate δ_B using Eq. (41), prioritizing lower validation measurement cost.

Health default and residual replacement. Under full-channel diagnostics, set

$$\mathcal{B}_{\mathrm{pass}} = \{b \in \mathcal{B} : \min_{s\in b} h_s \ge \eta\}, \quad \mathcal{B}_H = \begin{cases} \mathcal{B}_{\mathrm{pass}}, & \mathcal{B}_{\mathrm{pass}} \neq \varnothing, \\ \mathcal{B}, & \mathcal{B}_{\mathrm{pass}} = \varnothing. \end{cases} \tag{43}$$

Rank bundles by mean validation EMQ error at a fixed reference size n_{ref}, defining

$$b_H = \arg\min_{b\in\mathcal{B}_H} \frac{1}{N_V} \sum_{u\in\mathcal{V}} E_{u,b}(n_{\mathrm{ref}}), \qquad b_L = \arg\min_{b\in\mathcal{B}_H} \widehat{E}_b(e). \tag{44}$$

The action is

$$b^\star = \begin{cases} b_L, & b_L \neq b_H \text{ and } \widehat{E}_{b_H}(e) - \widehat{E}_{b_L}(e) > \delta_M, \\ b_H, & \text{otherwise.} \end{cases} \tag{45}$$

Choose η and δ_M on validation tasks. Among near-optimal reference-error candidates, δ_M prioritizes a lower replacement rate. Equations (42) and (45) are mutually exclusive measurement protocols; they are not applied sequentially as a combined default rule.

Acquisition policy. The sampling tools are

$$\mathcal{A}_{\mathrm{acq}} = \{\mathrm{Random}, \mathrm{Margin}, \mathrm{Entropy}, \mathrm{BALD}, \mathrm{BADGE}, \mathrm{CoreSet}, \mathrm{EPIG}\}. \tag{46}$$

Given I_t and P_t, tool q returns

$$J_{t,q} = \mathcal{S}_q(P_t, I_t, \Delta n_t), \qquad I^+_{t,q} = I_t \cup J_{t,q}, \qquad \Delta n_t = n_{t+1} - n_t. \tag{47}$$

Offline utility and online selection are

$$V^{\mathrm{acq}}_{u,q,t} = -E_{u,b^\star}(I^+_{t,q}), \qquad q^\star_t \in \arg\max_{q\in\mathcal{A}_{\mathrm{acq}}(e_t)} \widehat{V}^{\mathrm{acq}}_{q,t}(e_t). \tag{48}$$

Appendix A.15 specifies the operators. A feature-independent sequential protocol uses a fixed random permutation to produce nested samples. Feature-driven sampling reads only declared visible candidate information, and inspected target channels are charged to the observation ledger. An uncorrected average from a selectively acquired subset is not asserted to be a consistent estimate of the original population proportions.

The estimand remains the original population's conditional proportion β in Eq. (13), even when acquisition depends on features. This specification does not define inclusion probabilities, weighted quantifiers, or a separate representative fitting subsample for that case. The unweighted counts and EMQ updates then operate on the selected fitting data, whose distribution may differ from the target population. Label-free access satisfies the observation permissions but does not establish representativeness. Supervision against offline β_u evaluates candidate performance in the offline task distribution; it is not a selection-bias correction. Experimental conclusions must therefore retain the distinction between fixed random-order and feature-driven acquisition.

Incremental gain and stopping. Stopping supervision follows the batch selected by the learned acquisition policy. For a training task, use its out-of-fold sampling action $q^{\mathrm{oof}}_{u,t}$ and define

$$G_{u,t} = E_{u,b^\star}(I_t) - E_{u,b^\star}(I^+_{t,q^{\mathrm{oof}}_{u,t}}). \tag{49}$$

Gain regressors $g_r, r = 1, \ldots, R_G$, predict this quantity from current visible evidence e_t. Define

$$\mu_t = R_G^{-1} \sum_r g_r(e_t), \quad \sigma_t = \left[R_G^{-1} \sum_r (g_r(e_t) - \mu_t)^2 \right]^{1/2}, \quad S_t = \mu_t - \kappa\sigma_t. \tag{50}$$

With incremental observation cost ΔC_t for the selected next batch, a nonterminal checkpoint executes

$$a_t^{\text{resource}} = \begin{cases} \text{CONTINUE}, & S_t > \tau_G,\ |P_t| > 0,\ C_t^{\text{obs}} + \Delta C_t \le C_{\max}, \\ \text{STOP_ACQUISITION}, & \text{otherwise.} \end{cases} \tag{51}$$

After further acquisition, recompute EMQ and stage evidence. Bundle $b^\star$ remains fixed within this sequential protocol. Stopping acquisition retains the fitting set for downstream inference.

Multiple-checkpoint protocols train gain models on (u, t) records and execute complete validation trajectories for each (κ, τ_G). If $t_u^\star$ is the final checkpoint, then

$$L_u(\kappa, \tau_G) = E_{u,b_u^\star}(I_{t_u^\star}), \qquad C_u(\kappa, \tau_G) = \sum_{(i,s) \in O_{t_u^\star}} c_s. \tag{52}$$

Retain near-optimal-error policies using Eq. (41), then minimize mean cost. A single-continuation protocol contains only one nonterminal decision. S_t is a disagreement-penalized gain score and is not defined as a statistical lower confidence bound.

A.6 Adaptation Context and Joint Task Supervision

Candidates and output context. Candidates comprise the classic source model, CORAL, a neural source model, and dependency-compatible TENT-, EATA-, SAR-, SHOT-, CoTTA-, and COME-style updates. For final fitting set I, adaptation tool a returns

$$\mathcal{Z}_a = (p_{a,I}, z_{a,I}, p_{a,J}, z_{a,J}, \mathcal{R}_a), \tag{53}$$

where $\mathcal{R}_a$ contains the corresponding fixed source references. Fit adaptation parameters only from X_I, then produce outputs on X_J according to the tool's prediction specification without using Y_J.

The CORAL-style transform (Sun et al., 2016) uses means and covariances of the source and target fitting representations:

$$T_a(z) = (z - \mu_T)(\Sigma_T + \lambda_{\text{coral}} I)^{-1/2} (\Sigma_S + \lambda_{\text{coral}} I)^{1/2} + \mu_S. \tag{54}$$

Fix this transform, apply it to fitting and prediction sets, then invoke the matching source classifier and calibrator. Neural updates are specified in Appendix A.14; the methods do not share a single adaptation objective.

Here, the corresponding reference $\mathcal{R}_a$ identifies the registered source objects accompanying candidate a; this association does not establish statistical compatibility after adaptation. Source-only candidates retain their respective source prediction mappings and references. The CORAL path applies Eq. (54) before the fixed source classifier and calibrator. Neural updates can change probabilities and feature coordinates while source reference objects remain fixed. No step above recomputes source confusion matrices using the adapted model or fits a new calibrator. The probability mapping, calibrator, reference matrices, and feature coordinates paired by a neural tool must be identified in its execution configuration. When adapted outputs are paired with original references, confusion correction, EMQ, and prior correction are working models with possible reference mismatch. Source calibration does not automatically transfer to the adapted probabilities, and rebuilding the context does not establish preservation of class-conditional moments.

Joint supervision and selection. For candidate a on offline task u, the common reference quantifier EMQ yields $\widehat{\beta}_{u,a}^{\text{ref}}$. Compute known-class Macro-F1 $F_{u,a}$ on the independent prediction set and use

$$L_{u,a}^{\text{adapt}} = \|\widehat{\beta}_{u,a}^{\text{ref}} - \beta_u\|_1 + \lambda_{\text{cls}}(1 - F_{u,a}). \tag{55}$$

This stage-local proxy uses a reference quantifier; it is distinct from the downstream composed objective evaluated in Section 5.4. Learn $\widehat{L}_a^{\text{adapt}}(e)$ and select $a^\star$ by Eq. (38), using the classic source model as fallback. Rebuild quantification evidence from $\mathcal{Z}_{a^\star}$ after adaptation.

A.7 Complete Quantifier Definitions

Context and direct estimates. In the selected context, let π_a denote the source prior, $P = (p_{ik})$ the target probability matrix, A_a, B_a the hard and soft source confusion matrices, and $C_{A,a}, C_{B,a}$ their joint forms. Project every quantification output onto Δ^{K-1} using

$$\Pi_{\Delta^{K-1}}(v) = \arg\min_{q \in \Delta^{K-1}} \|q - v\|_2^2. \tag{56}$$

The hard and soft counts are

$$\widehat{q}_k = n^{-1} \sum_i \mathbf{1}\{\arg\max_j p_{ij} = k\}, \qquad \widehat{m}_k = n^{-1} \sum_i p_{ik}. \tag{57}$$

CC and PCC return $\Pi_{\Delta^{K-1}}(\widehat{q})$ and $\Pi_{\Delta^{K-1}}(\widehat{m})$, respectively.

Confusion-matrix correction. The adjusted-counting candidates ACC and PACC (Forman, 2008; Bella et al., 2010) solve nonnegative least-squares problems and then project:

$$\begin{aligned} \widehat{\beta}^{\mathrm{ACC}} &= \Pi_{\Delta^{K-1}}\left(\arg\min_{v \geq 0} \|A_a v - \widehat{q}\|_2^2\right), \\ \widehat{\beta}^{\mathrm{PACC}} &= \Pi_{\Delta^{K-1}}\left(\arg\min_{v \geq 0} \|B_a v - \widehat{m}\|_2^2\right). \end{aligned} \tag{58}$$

BBSE (Lipton et al., 2018) estimates weights and restores proportions:

$$\widehat{w} = \arg\min_{w \geq 0} \|C_{A,a} w - \widehat{q}\|_2^2, \qquad \widehat{\beta}^{\mathrm{BBSE}} = \Pi_{\Delta^{K-1}}(\pi_a \odot \widehat{w}). \tag{59}$$

Under these unregularized formulations and strictly positive source priors, $v = \mathrm{diag}(\pi_a) w$ gives a bijection between the nonnegative feasible sets, so ACC and BBSE are mathematically equivalent up to reparameterization. Both names are retained for historical portfolio identity; the pair is not counted as independent methodological diversity. Numerical tie-breaking can still produce different action labels. The regularized label-shift candidate (Azizzadenesheli et al., 2019) solves

$$\begin{aligned} \widehat{w} &= \arg\min_{w \geq 0,\, \pi_a^\top w = 1} \{\|C_{B,a} w - \widehat{m}\|_2^2 + \lambda_R \|w - \mathbf{1}\|_2^2\}, \\ \widehat{\beta}^{\mathrm{RLLS}} &= \Pi_{\Delta^{K-1}}(\pi_a \odot \widehat{w}). \end{aligned} \tag{60}$$

EMQ and probability matching. EMQ uses Eq. (22) in the selected context, replacing the source prior by π_a. Its working objective is

$$\min_{\beta \in \Delta^{K-1}} -\frac{1}{n} \sum_i \log\left(\sum_k p_{ik} \frac{\beta_k}{\pi_{a,k}}\right). \tag{61}$$

Under closed-set label shift and accurate source probabilities, this is the marginal-likelihood objective for the target prior. Under other conditions, it remains a working model.

The CPMCN-style candidate solver (Wen et al., 2024) parameterizes

$$w_k(\theta) = \frac{\exp[\mathrm{clip}(\theta_k, -L_\theta, L_\theta)]}{\sum_\ell \pi_{a,\ell} \exp[\mathrm{clip}(\theta_\ell, -L_\theta, L_\theta)]} \tag{62}$$

and numerically solves

$$\begin{aligned} \widehat{\theta} &\in \arg\min_\theta \sum_k \left[\pi_{a,k} - \frac{1}{n} \sum_i \frac{p_{ik}}{\sum_\ell p_{i\ell} w_\ell(\theta)}\right]^2, \\ \widehat{\beta}^{\mathrm{CPM}} &= \Pi_{\Delta^{K-1}}(\pi_a \odot w(\widehat{\theta})). \end{aligned} \tag{63}$$

Distribution matching and a reference-aligned mixture. For each class k and histogram bin count $h \in \mathcal{G}_h$, construct positive-class, remaining-class, and target score histograms $H^+_{k,h}, H^-_{k,h}, H^T_{k,h}$ from reference and target probabilities. Search a prespecified grid $\mathcal{G}_\alpha \subset [0,1]$:

$$\begin{aligned} \widehat{\alpha}_{k,h} &= \arg\min_{\alpha \in \mathcal{G}_\alpha} D\Big(H^T_{k,h}, \alpha H^+_{k,h} + (1-\alpha) H^-_{k,h}\Big), \\ \widehat{\beta} &= \Pi_{\Delta^{K-1}}((\text{median}_h\, \widehat{\alpha}_{k,h})_k). \end{aligned} \tag{64}$$

HDy-OVR uses Hellinger distance (González-Castro et al., 2013), while DyS-OVR uses Topsøe distance (Maletzke et al., 2019):

$$\begin{aligned} D_H(p,q) &= \left[\tfrac{1}{2}\sum_j (\sqrt{p_j} - \sqrt{q_j})^2\right]^{1/2}, \\ D_{\text{Top}}(p,q) &= \sum_j \left[p_j \log \frac{2p_j}{p_j+q_j} + q_j \log \frac{2q_j}{p_j+q_j}\right]. \end{aligned} \tag{65}$$

Both are defined here as one-versus-rest extensions, rather than native multiclass estimators.

The reference-aligned GMM fits K diagonal-covariance components on target representations by

$$\max_{w \in \Delta^{K-1},\, m_j,\, \Sigma_j} \sum_{i \in I} \log \left[\sum_{j=1}^{K} w_j \varphi(z_i; m_j, \Sigma_j)\right], \tag{66}$$

where φ is a Gaussian density and covariance regularization is prespecified. Align fitted centers m_j and weights w_j to source classes by linear assignment:

$$\sigma^\star = \arg\min_{\sigma \in \mathfrak{S}_K} \sum_j \|m_j - \mu^{\text{ref}}_{\sigma(j)}\|_2^2, \qquad \widehat{\beta}_k = w_{(\sigma^\star)^{-1}(k)}. \tag{67}$$

The component count, covariance regularization, and initialization are fixed by the protocol.

Final quantifier selection. For every quantifier q, offline supervision is

$$L^{\text{quant}}_{u,q} = \|\widehat{\beta}_{u,q} - \beta_u\|_1. \tag{68}$$

Learn $\widehat{L}^{\text{quant}}_q(e)$ and use Eq. (38) with EMQ as fallback to select $q^\star$ and return $\widehat{\beta} = \widehat{\beta}^{(q^\star)}$. Outputs use the specified population estimand; consistency is not presumed under unknown contamination or conditional mismatch.

A.8 Classification Candidates, Scoring, and Utility Routing

Candidate probabilities. Let p_S be calibrated classic source probabilities, p_A the selected adaptation probabilities, and p_N available neural source probabilities. Their corresponding fixed reference priors are π_S and π_A. Define

$$\mathcal{C}(p; \widehat{\beta}, \pi)_k = \frac{p_k \widehat{\beta}_k / \pi_k}{\sum_\ell p_\ell \widehat{\beta}_\ell / \pi_\ell}. \tag{69}$$

The six classification policies are

$$p_{\text{source}} = p_S, \qquad p_{\text{source+prior}} = \mathcal{C}(p_S; \widehat{\beta}, \pi_S), \tag{70}$$

$$p_{\text{adapted}} = p_A, \qquad p_{\text{adapted+prior}} = \mathcal{C}(p_A; \widehat{\beta}, \pi_A), \tag{71}$$

$$p_{\text{source-ens}} = |\mathcal{P}_S|^{-1} \sum_{p \in \mathcal{P}_S} p, \qquad p_{\text{adapted-ens}} = |\mathcal{P}_A|^{-1} \sum_{p \in \mathcal{P}_A} p. \tag{72}$$

$\mathcal{P}_S$ is the ordered sequence containing p_S and available p_N, and $\mathcal{P}_A = (p_A, \mathcal{C}(p_A; \widehat{\beta}, \pi_A), p_S)$. Identical probability terms retain their individual weights. Without adaptation, p_A may equal a source candidate. Ensembles use prespecified equal weights, never weights fitted to target labels.

Utility supervision and final prediction. For offline known-class prediction indices $J_u^K = \{i \in J_u : y_i \in \mathcal{Y}\}$, candidate utility is

$$V_{u,c}^{\text{class}} = \frac{1}{K} \sum_{k=1}^{K} \frac{2\text{TP}_{u,k,c}}{2\text{TP}_{u,k,c} + \text{FP}_{u,k,c} + \text{FN}_{u,k,c}}. \tag{73}$$

A term with zero denominator follows the prespecified zero-value rule. Select the fixed baseline on offline training tasks:

$$c_0 \in \arg\max_{c \in \mathcal{A}_{\text{class}}} \frac{1}{|\mathcal{T}|} \sum_{u \in \mathcal{T}} V_{u,c}^{\text{class}}. \tag{74}$$

Learn utility predictors $\widehat{V}_c(e)$, choose δ_{class} on validation tasks, and select $c^\star$ by Eq. (39). Return

$$p_i^\star = p_{c^\star}(\cdot \mid x_i), \qquad \widehat{y}_i = \arg\max_{k \in \mathcal{Y}} p_{i,k}^\star, \qquad i \in J. \tag{75}$$

The action follows predicted candidate utility and is not mandated by one shift test. For unknown samples, retain known-class conditional predictions and independent OOD flags; known-class probabilities are not interpreted as joint posteriors that include an unknown class.

A.9 Proportion Intervals, Risk Calibration, and Selective Outputs

Component intervals. Hold the selected representation and quantifier fixed, bootstrap the fitting set, and rerun $q^\star$. For successful estimates $\{\widehat{\beta}^{[r]}\}_{r \in \mathcal{I}_B}$, report

$$[\ell_k, u_k] = \left[Q_{\alpha_B/2}(\widehat{\beta}_k^{[r]}), Q_{1-\alpha_B/2}(\widehat{\beta}_k^{[r]}) \right]. \tag{76}$$

If every resample fails, use a soft-count Dirichlet fallback with parameters $a_k = \sum_{i \in I} p_{ik} + \lambda_D$, recording the interval type. This fallback is not reported as a bootstrap interval for the selected quantifier.

These intervals describe resampling variation in the observed fitting data with the representation and quantifier held fixed. They do not rerun acquisition or tool selection. In particular, resampling a feature-selected fitting set does not recover unobserved regions of the population or correct its selection bias. Interval width and the separately learned risk radius below are different outputs; one does not validate the other.

Risk model, reference support, and calibration. Train R_U error regressors on final proportion error $L_u = \|\widehat{\beta}_u - \beta_u\|_1$. Their mean and scale are

$$\begin{aligned} m(e) &= R_U^{-1} \sum_r f_r(e), \\ s(e) &= \max\left\{ \left[R_U^{-1} \sum_r (f_r(e) - m(e))^2 \right]^{1/2}, s_{\min} \right\}. \end{aligned} \tag{77}$$

Fit a standardization map ζ on training evidence. Define the average distance to its k_U nearest training neighbors and the validation support threshold by

$$\begin{aligned} d_U(e) &= \frac{1}{k_U} \sum_{v \in \text{kNN}_{\mathcal{T}}(e)} \|\zeta(e) - \zeta(e_v)\|_2, \\ d_{\max} &= Q_{q_U}(d_U(e_u) : u \in \mathcal{V}). \end{aligned} \tag{78}$$

The risk calibration set is $\mathcal{V}_U = \{u \in \mathcal{V} : d_U(e_u) \le d_{\max}\}$. If it is smaller than a prespecified minimum, use all validation tasks. Define

$$\begin{aligned} s_u^+ &= \max\left\{0, \frac{L_u - m(e_u)}{s(e_u)}\right\}, \qquad q_{\text{cal}} = Q_{\ell_U}^{\text{higher}}(s_u^+ : u \in \mathcal{V}_U), \\ \ell_U &= \min\left\{1, \frac{\lceil (|\mathcal{V}_U| + 1)(1 - \alpha_U) \rceil}{|\mathcal{V}_U|}\right\}. \end{aligned} \tag{79}$$

The empirical risk radius is

$$r_U(e) = \max\{0, m(e) + q_{\text{cal}} s(e)\}. \tag{80}$$

Let $r_{\text{rel}} \le r_{\text{cau}}$ be prespecified quantiles of validation risk radii. The state is

$$\nu_U(e) = \begin{cases} \text{OUT_OF_SUPPORT}, & d_U(e) > d_{\max}, \\ \text{RELIABLE}, & d_U(e) \le d_{\max},\ r_U(e) \le r_{\text{rel}}, \\ \text{CAUTIOUS}, & d_U(e) \le d_{\max},\ r_{\text{rel}} < r_U(e) \le r_{\text{cau}}, \\ \text{HIGH_RISK}, & \text{otherwise.} \end{cases} \tag{81}$$

Q^{higher} denotes the empirical quantile with upward rounding. This is selective empirical calibration; it does not establish unconditional coverage under arbitrary shifts or random stopping.

The label `RELIABLE` denotes a relative predicted-risk category: evidence lies within the reference-support threshold and its radius is below the validation quantile r_{rel}. Equation (81) contains no absolute acceptable β-error tolerance, so this label does not certify that such a tolerance is met. The risk thresholds rank outputs relative to the validation reference, including when that reference has large errors.

The same task set $\mathcal{V}$ is used for upstream control selection in Appendix A.4, the support threshold, residual quantile, and risk-state thresholds here. This specification does not provide a separate calibration split after those choices. The reported radius is therefore an empirical quantity after validation-based selection; the upward-quantile formula alone does not supply independent-calibration coverage. Task-level out-of-fold construction on $\mathcal{T}$ does not make the reused $\mathcal{V}$ independent of parameter selection. Neither this reuse nor fixed-set bootstrapping supplies a guarantee at a data-dependent stopping time.

A.10 OOD Scores, Fusion, and Unknown-Mass Models

Individual OOD scores. The OOD module uses logits $\ell(x)$ and representation $z(x)$ of the fixed source neural model. Source prototypes, covariance, and subspaces are fixed together with that model. The maximum-softmax, energy, and Mahalanobis scores follow their respective reference constructions (Hendrycks & Gimpel, 2017; Liu et al., 2020; Lee et al., 2018). Larger scores indicate greater departure from the reference:

$$s_{\text{MSP}}(x) = 1 - \max_k p_N(k \mid x), \tag{82}$$

$$s_{\text{Energy}}(x) = -T_E \log \sum_k \exp(\ell_k(x)/T_E), \tag{83}$$

$$s_{\text{Maha}}(x) = \min_k (z(x) - \mu_k)^\top \Sigma^{-1} (z(x) - \mu_k). \tag{84}$$

ODIN (Liang et al., 2018) sets $\widehat{y} = \arg\max_k \ell_k(x)$ and works in standardized input space:

$$\begin{aligned} x' &= x - \varepsilon_O \operatorname{sign} \nabla_x [-\log \operatorname{softmax}(\ell(x)/T_O)_{\widehat{y}}], \\ s_{\text{ODIN}}(x) &= 1 - \max_k \operatorname{softmax}(\ell(x')/T_O)_k. \end{aligned} \tag{85}$$

ReAct (Sun et al., 2021) clips the representation at a source quantile threshold τ_R and computes energy:

$$z_R = \min\{z, \tau_R\}, \qquad s_{\text{ReAct}}(x) = -\log \sum_k \exp([W z_R + b]_k). \tag{86}$$

ViM (Wang et al., 2022a) uses source origin o, residual-subspace basis $U_\perp$, and source-fitted scale α_V:

$$s_{\text{ViM}}(x) = \alpha_V \|U_\perp^\top (z(x) - o)\|_2 - \log \sum_k \exp \ell_k(x). \tag{87}$$

Do not register a detector with unmet dependencies. A failed execution returns its failure flag and tool-specified replacement score; the replacement is not treated as a newly obtained detector result.

Score fusion. For detector d and bundle b, standardize against the corresponding source-reference mean and standard deviation:

$$\widetilde{s}_d^b(x) = \frac{s_d^b(x) - \mu_{S,d}^b}{\sigma_{S,d}^b + \varepsilon_O}. \tag{88}$$

Concatenate standardized scores and bundle indicators, then fit linear standardization and a class-balanced Logistic fusion model on training samples. Absorbing the linear standardization gives the equivalent form

$$s_F(x) = \sigma\big(w^\top \widetilde{s}^b(x) + v^\top e_b + a\big). \tag{89}$$

Offline sample labels are $o_i = \mathbf{1}\{y_i \notin \mathcal{Y}\}$. With class-balanced weights ω_{o_i}, fit

$$\begin{aligned} \min_{w,v,a} \; & -\sum_i \omega_{o_i}[o_i \log s_F(x_i) + (1-o_i)\log(1-s_F(x_i))] \\ & + \lambda_{\text{fus}}(\|w\|_2^2 + \|v\|_2^2). \end{aligned} \tag{90}$$

Test labels are not used to fit fusion.

Open-world evidence and unknown mass. For available individual detectors $\mathcal{M}_O$ and the fusion score, extract the mean, standard deviation, prespecified quantiles $\{Q_\gamma : \gamma \in \Gamma_O\}$, and exceedance rates at each source threshold. Concatenate these with final routing evidence and the selected classification probabilities' average confidence and entropy to obtain x_O.

The unknown-mass regressor is supervised directly by $(x_{O,u}, \rho_u)$:

$$\mathcal{L}_\rho(f) = \sum_{u \in \mathcal{T}} [f(x_{O,u}) - \rho_u]^2, \tag{91}$$

and returns

$$\widehat{\rho} = \text{clip}(f_\rho(x_O), 0, \rho_{\max}). \tag{92}$$

Apply the open-world evidence mapping to $[x_O, \widehat{\rho}]$ to obtain e_O. A separate task-existence classifier produces

$$p_{\text{open}} = d_O(e_O), \qquad \text{with offline target } \mathbf{1}\{\rho_u > 0\}. \tag{93}$$

Record p_{open}, $\widehat{\rho}$, and individual sample scores $s_d(x)$ separately.

A.11 Residual Screening, Clustering, and Metric Constraints

Screening and removal of known-class samples. For detector d and reference false-positive level $\alpha \in \mathcal{G}_O$, define

$$\tau_{d,\alpha} = Q_{1-\alpha}(s_d(X_S^{\text{ref}})), \qquad U_{d,\alpha} = \{i \in J : s_d(x_i) > \tau_{d,\alpha}\}. \tag{94}$$

Unknown-mass-matched screening selects

$$U_{d,\rho} = \text{Top}_{\text{round}(|J|\widehat{\rho})}\{s_d(x_i) : i \in J\}. \tag{95}$$

Resolve score ties by fixed sample order. For a candidate screen U, compute distance to fixed neural source prototypes:

$$D_i^{\text{proto}} = \min_k (z_i^N - \mu_k)^\top \Sigma^{-1} (z_i^N - \mu_k), \tag{96}$$

and construct the actual clustering pool

$$U^{\text{clean}} = U \setminus \{i : \max_k p_{i,k}^\star \geq \tau_{\text{conf}},\ D_i^{\text{proto}} \leq \tau_{\text{proto}}\}. \tag{97}$$

The OOD flags U and clustering input U^{clean} are saved separately.

Clustering tools. Discovery actions are NONE, KMeans, GMM-BIC, HDBSCAN, and the SS-KMeans core. Mixture fitting uses EM (Dempster et al., 1977), with BIC model selection (Schwarz, 1978). K-means solves

$$\min_{\mu_1,\dots,\mu_m,\, c_i\in[m]} \sum_{i\in U^{\text{clean}}} \|z_i - \mu_{c_i}\|_2^2 \tag{98}$$

and chooses the number of clusters from a prespecified set by maximum average silhouette score $\bar{s} = |U^{\text{clean}}|^{-1}\sum_i (b_i - a_i)/\max(a_i, b_i)$. Here a_i is the mean within-cluster distance and b_i the smallest mean distance to another cluster. GMM fits every candidate component count by maximum likelihood and minimizes

$$\text{BIC}(m) = -2\ell_m(\widehat{\vartheta}_m) + d_m \log |U^{\text{clean}}|. \tag{99}$$

HDBSCAN (Campello et al., 2013) uses fixed minimum-cluster-size and neighborhood parameters to return cluster or noise labels. The semi-supervised K-means core (Vaze et al., 2022) fixes known-class prototypes and iterates assignments and new-cluster centers:

$$c_i^{(r)} = \arg\min_j \|z_i - \mu_j^{(r)}\|_2^2, \qquad \mu_j^{(r+1)} = \frac{\sum_i \mathbf{1}\{c_i^{(r)} = j\} z_i}{\sum_i \mathbf{1}\{c_i^{(r)} = j\}}, \quad j > K. \tag{100}$$

Known prototypes are not updated. Empty new clusters retain their previous centers. Residual samples assigned to a known prototype receive noise labels. This core operator is not equivalent to the complete GCD, SimGCD, or CMS training pipelines.

Supervision of action metrics. An open-world action $a = (d, g, m)$ specifies the detector, screen, and clustering tool. In offline tasks, let o_i indicate an unknown sample, $k_i = \mathbf{1}\{c_i \geq 0\}$ indicate a non-noise clustering assignment, and $u_i = \mathbf{1}\{i \in U\}$ indicate screening. Define

$$\text{FPR}(a) = \frac{\sum_i (1 - o_i) u_i}{\sum_i (1 - o_i)}, \qquad \text{Recall}(a) = \frac{\sum_i o_i u_i}{\sum_i o_i}, \tag{101}$$

$$\text{Residual}(a) = |J|^{-1} \sum_i u_i, \qquad E_g(a) = |\text{Residual}(a) - \rho|, \tag{102}$$

$$\text{Coverage}(a) = \frac{\sum_i o_i u_i k_i}{\sum_i o_i}, \qquad \text{Contam}(a) = \frac{\sum_i (1 - o_i) u_i k_i}{\max\{\sum_i u_i k_i, 1\}}. \tag{103}$$

E_g is screening-mass error, distinct from unknown-mass regression error $|\widehat{\rho} - \rho|$. Also compute OOD AUROC, clustering ARI on unknown samples, and pairwise F1:

$$\text{PairF1} = \frac{2\text{TP}_{\text{pair}}}{2\text{TP}_{\text{pair}} + \text{FP}_{\text{pair}} + \text{FN}_{\text{pair}}}. \tag{104}$$

A pair is predicted to share a cluster only if both samples have the same non-noise cluster label. Missing evaluation samples and zero denominators use fixed degeneracy rules. Assemble these metrics as $M_{u,a}$ and fit a multi-output regressor with objective

$$\sum_{u\in\mathcal{T}} \sum_{a\in\mathcal{A}_O} \|f_{O,a}(e_{O,u}) - M_{u,a}\|_2^2. \tag{105}$$

At deployment, use the predicted metrics, clipped to their prespecified ranges.

Existence, detector, and discovery selection. Select the existence threshold τ_O on validation tasks: first constrain false triggering on closed-set tasks, then maximize open-task recall and binary F1 in that order. If no threshold is feasible, use the prespecified F1 fallback ordering.

If $p_{\text{open}} < \tau_O$, select among actions with $m = \text{NONE}$ by lexicographically minimizing

$$(\widehat{\text{FPR}}, \widehat{\text{Residual}}, \widehat{E}_g). \tag{106}$$

Mark the task as closed-set-leaning and do not cluster. This state does not require all OOD scores to be zero.

If $p_{\text{open}} \geq \tau_O$, first retain no-discovery actions satisfying $\widehat{\text{FPR}} \leq \alpha_{\max}$. If none is feasible, restore the corresponding candidate set and record a fallback. Within the fixed validation-best detector and within all detectors, respectively, lexicographically maximize

$$(\widehat{\text{AUROC}}, \widehat{\text{Recall}}, -\widehat{E}_g, -\widehat{\text{FPR}}) \tag{107}$$

to obtain default action a_B and learned candidate a_L. Select a_L only when

$$\widehat{\text{AUROC}}(a_L) \geq \widehat{\text{AUROC}}(a_B) + \delta_O; \tag{108}$$

otherwise retain a_B.

With detector and screen fixed, nonempty discovery actions must satisfy

$$\widehat{\text{FPR}} \leq \alpha_{\max}, \quad \widehat{\text{Contam}} \leq \gamma_{\max}, \quad \widehat{\text{Coverage}} \geq \zeta_{\min}, \quad \widehat{\text{PairF1}} \geq \xi_{\min}. \tag{109}$$

Among feasible actions, lexicographically maximize $(\widehat{\text{PairF1}}, \widehat{\text{Coverage}}, \widehat{\text{ARI}}, -\widehat{\text{Contam}})$. If none is feasible, choose $m = \text{NONE}$. If the actual cleaned pool has fewer than n_{disc} samples or clustering fails, return no valid clusters. Equation (109) constrains predicted metrics, not target ground truth.

A.12 Dependencies, Verification Protocols, Costs, and Outputs

Dependency updates. New fitting observations require recomputing affected preliminary quantification and evidence. A changed representation or model requires rebuilding the quantification context from the new probabilities, features, and corresponding fixed references. A changed quantification output requires recomputing all prior-corrected candidates. A changed detector or screen requires regenerating and cleaning the residual pool. Reuse only intermediate results consistent with the current analysis context.

Rebuilding associates current outputs with their registered fixed source-reference objects and invalidates stale intermediate calculations. Appendix A.6 specifies the compatibility conditions for these objects.

Registered action verification. When verification is enabled, let R, U, D_{res}, and p_{fit} denote that protocol's observable risk score, uncertainty score, fit residual, and reference-fit score. Its operator $\mathcal{V}_{\mathcal{P},a}$ receives the before/after changes

$$\Delta_a = (R^- - R^+,\ U^- - U^+,\ D^-_{\text{res}} - D^+_{\text{res}},\ p^+_{\text{fit}} - p^-_{\text{fit}}) \tag{110}$$

and returns $v_a \in \{\text{supported}, \text{unsupported}, \text{inconclusive}\}$. The protocol fixes the metrics and decision thresholds. True task loss is never an online verification input. A universal threshold system common to every stage is not specified, so there is no global verification decision equation.

Let ν_t be the evidence version and $M_t(a)$ blocked-action memory. Version-based re-entry uses

$$M_{t+1}(a) = \begin{cases} \nu_t, & v_a = \text{unsupported}, \\ M_t(a), & \text{otherwise}, \end{cases} \qquad \text{Eligible}(a,t) = \mathbf{1}\{a \notin \text{dom}\, M_t \text{ or } \nu_t > M_t(a)\}. \tag{111}$$

Update ν_t only after receiving protocol-recognized new observations or changing the effective analysis context. Health- and gain-based re-entry protocols, when present, use their own registered rules. Return edges from different verification protocols are not combined and presented as one executed controller.

Cost and terminal output. The complete observation ledger is

$$\Delta C_t = \sum_{(i,s) \in O_{t+1} \setminus O_t} c_s, \qquad C_t^{\text{obs}} = \sum_{(i,s) \in O_t} c_s. \tag{112}$$

Only when the first n_0 samples measure all channels, subsequent samples measure a fixed $b^\star$, and no additional candidate-screening measurements occur does fitting cost simplify to

$$C^{\text{fit}} = n_0 c(\mathcal{S}) + (n^\star - n_0) c(b^\star). \tag{113}$$

Candidate-scoring measurements, prediction-set measurements, and computation are accounted for separately. A report of C^{fit} alone is explicitly identified as fitting cost and is not called complete run cost.

The terminal state is

$$\nu_{\mathrm{term}} = \begin{cases} \mathrm{ABSTAIN}, & \nu_U \in \{\mathrm{OUT_OF_SUPPORT}, \mathrm{HIGH_RISK}\}, \\ \mathrm{STOP}, & \text{otherwise.} \end{cases} \tag{114}$$

Also retain the acquisition stopping reason, open-world state, and tool-failure status. The output object is

$$\mathcal{O} = \left(\widehat{\beta}, [\ell, u], r_U, \{p_i^\star, \widehat{y}_i\}_{i\in J}, \widehat{\rho}, \{s_i, u_i, c_i\}_{i\in J}, \nu_U, \nu_{\mathrm{term}}, C, \mathcal{L}\right), \tag{115}$$

where the decision record $\mathcal{L}$ contains the stage, actual action, diagnostic scores, predicted advantage, fallback status, and resource information.

A.13 Detailed Offline and Online Algorithms

Offline fitting and calibration.

1. Train source classifiers and calibrators for each measurement bundle. Construct priors, confusion matrices, representation references, and OOD references.
2. Compute measurement evidence and each bundle's reference errors from initial observations of offline tasks. Fit measurement-effect predictors and calibrate either the near-optimal tolerance or the health and replacement thresholds required by the protocol.
3. Generate training bundles using out-of-fold measurement actions. Execute sampling tools, fit their utility model, and form gain supervision from out-of-fold outputs of the selected sampling policy.
4. Fit gain regressors and calibrate stopping parameters. For multiple-checkpoint protocols, use complete validation trajectories. Generate adaptation fitting sets and their stage evidence.
5. Fit each adaptation candidate on the fitting set and evaluate its joint loss using an independent prediction set. Train adaptation routing and generate out-of-fold selected representations.
6. Evaluate each quantifier's proportion error in the selected representation. Fit and calibrate quantification routing. Construct classification candidates using the selected proportions, then fit and calibrate classification utility routing.
7. Fit risk models to final proportion error. Construct the reference-support region, risk calibration scores, and terminal thresholds.
8. Fit OOD score fusion and source screening thresholds. Fit unknown-mass, task-existence, and open-world action-metric models; calibrate existence and detector-selection parameters.
9. Freeze all offline objects and the execution protocol. Save randomization settings, tool versions, feature dictionaries, and fallback rules.

Staged inference for one target task.

1. Acquire n_0 samples under b_0 and update the observation ledger. Compute preliminary EMQ, diagnostic statistics, and measurement evidence.
2. Select $b^\star$ under the registered measurement protocol. Reuse existing features and acquire remaining measurements required for the first resource checkpoint.
3. At each nonterminal checkpoint, compute current resource evidence, a sampling action, and gain scores. If the continuation condition holds, acquire the selected batch and recompute evidence; otherwise stop acquisition. Record numerical failure, insufficient resources, reaching a limit, or other registered stopping reasons.
4. Select adaptation using the final fitting-set evidence. Fit only on that set, freeze the output context, and compute independent prediction-set probabilities and representations.
5. Rebuild quantification evidence, select a quantifier, and return $\widehat{\beta}$. Compute all available classification candidates and select ordinary, corrected, adapted, or ensemble probabilities through utility routing.

6. Compute component intervals, the risk radius, and reference-support state for the selected quantification result. Retain all uncertainty types and fallback flags.
7. Compute individual and fused OOD scores, unknown mass, and task-existence score. Select the detector and screen, then construct U and U^{clean}.
8. Execute the selected clustering tool only if predicted constraints and the actual sample size permit it; otherwise return no discovery. Save screening and cluster labels.
9. If verification is enabled, run its registered checks and update re-entry memory. Re-execute affected stages only along return edges specified by that protocol.
10. Return final proportions, classification and open-world outputs, uncertainty, costs, the decision record, and terminal state.

Execution scope. The complete forward dependency chain is

$$\begin{aligned}&\text{initial evidence} \to \text{measurement and sampling control} \to \text{adaptation}\\ &\quad\to \text{quantification} \to \text{classification} \to \text{risk assessment}\\ &\to \text{open-world analysis and residual discovery} \to \text{terminal output.}\end{aligned} \tag{116}$$

A single forward pass does not automatically include multiple acquisition checkpoints or per-action verification. Sequential protocols must declare trained checkpoints and actual loop edges; verification protocols must declare their decision operator and re-entry rules. Module, measurement, and verification experiments each correspond to the action sets actually enabled.

A.14 Neural Adaptation Update Operators

Fix target fitting set I, source parameters ϑ_S, update count R_a, learning-rate schedule, and permitted parameter subset Θ_a. Every tool operates on a copy of the source model and never reads prediction labels. The following equations specify candidate update implementations, not the complete training and evaluation protocols of the original methods.

TENT- and EATA-style updates. Write $H_i(\vartheta) = -\sum_k p_\vartheta(k \mid x_i) \log p_\vartheta(k \mid x_i)$. The TENT-style update (Wang et al., 2021) changes the designated normalization affine parameters with

$$\mathcal{L}_{\text{TENT}}(\vartheta) = |I|^{-1} \sum_{i \in I} H_i(\vartheta). \tag{117}$$

The EATA-style update (Niu et al., 2022) selects samples by their current entropy and cosine similarity to the running probability average $\bar{p}_r$:

$$I_r^E = \{i : H_i(\vartheta_r) < \tau_E,\ \cos(p_{\vartheta_r}(x_i), \bar{p}_r) < 1 - \delta_E\}. \tag{118}$$

Omit the similarity condition in the first round and terminate on an empty set. Optimize

$$\mathcal{L}_E(\vartheta) = |I_r^E|^{-1} \sum_{i \in I_r^E} H_i(\vartheta) + \lambda_F \sum_j F_j(\vartheta_j - \vartheta_{S,j})^2, \tag{119}$$

where F_j are source-reference Fisher weights. Update the running average with prespecified momentum and optimize only the designated parameters.

SAR-style updates. The operator follows the sharpness-aware entropy-minimization principle (Niu et al., 2023). Fix the round's selected set $I_r^S = \{i : H_i(\vartheta_r) < \tau_S\}$ and define

$$L_r(\vartheta) = |I_r^S|^{-1} \sum_{i \in I_r^S} H_i(\vartheta), \qquad \epsilon_r = \rho_S \frac{\nabla L_r(\vartheta_r)}{\|\nabla L_r(\vartheta_r)\|_2 + \varepsilon_S}. \tag{120}$$

Evaluate the gradient at $\vartheta_r + \epsilon_r$, restore the original parameters, and perform a descent update with the configured momentum. Terminate when the selected set is empty.

SHOT-style updates. The operator uses a source-hypothesis transfer objective (Liang et al., 2020). Freeze the classification head and update representation parameters. Form centers from current hard-prediction groups, substituting prescribed source centers for empty groups. Nearest-center assignments give pseudolabels $\widetilde{y}_i$. Use

$$\begin{aligned}\mathcal{L}_{\mathrm{SHOT}}(\vartheta) &= \frac{1}{|I|}\sum_i H_i(\vartheta) + \sum_k \bar{p}_k \log \bar{p}_k - \frac{\lambda_P}{|I|}\sum_i \log p_\vartheta(\widetilde{y}_i \mid x_i),\\ \bar{p} &= |I|^{-1}\sum_i p_\vartheta(x_i).\end{aligned} \tag{121}$$

Recompute pseudolabels each round and update representation parameters with the prescribed optimizer.

CoTTA-style updates. The operator uses teacher consistency and stochastic restoration (Wang et al., 2022b). Let student and teacher parameters be ϑ and $\bar{\vartheta}$. For the specified augmentation set or random augmentation distribution, construct

$$\begin{aligned}q_i &= R_{\mathrm{aug}}^{-1}\sum_{r=1}^{R_{\mathrm{aug}}} p_{\bar{\vartheta}}(\cdot \mid \mathcal{T}_r(x_i)),\\ \mathcal{L}_{\mathrm{CoTTA}} &= -|I|^{-1}\sum_i\sum_k q_{ik}\log p_\vartheta(k \mid x_i).\end{aligned} \tag{122}$$

After the student update, set $\bar{\vartheta} \leftarrow \lambda_T\bar{\vartheta} + (1-\lambda_T)\vartheta$. Restore student parameter coordinates to their source values with the prescribed restoration probability. Return the teacher model.

COME-style evidential objective. The scoped operator uses an evidential uncertainty objective (Zhang et al., 2025). For logits ℓ_i, use the gradient-normalized expression

$$\widetilde{\ell}_i = \frac{\ell_i}{\|\ell_i\|_2 + \varepsilon_C}\,\mathrm{stopgrad}(\|\ell_i\|_2), \quad E_{ik} = e^{\widetilde{\ell}_{ik}}, \quad D_i = \sum_k E_{ik} + \lambda_{\mathrm{ev}}. \tag{123}$$

Define $b_{ik} = E_{ik}/D_i$ and $u_i = K/D_i$, with protected quantities $\bar{b}_{ik} = b_{ik} + \varepsilon_C$ and $\bar{u}_i = u_i + \varepsilon_C$. The implemented objective is

$$\mathcal{L}_{\mathrm{COME}} = -\frac{1}{|I|}\sum_i\left[\sum_k \bar{b}_{ik}\log\bar{b}_{ik} + \bar{u}_i\log\bar{u}_i\right]. \tag{124}$$

This is the specified evidential objective; when $\lambda_{\mathrm{ev}} \neq K$, (b_i, u_i) is not defined as a normalized probability vector.

A.15 Sample Acquisition Operators

Fix candidate indices P, already measured representations Z_I, and batch size m. Candidates have probabilities p_i and representations z_i. Random samples uniformly from P without replacement. Margin selects the largest $-(p_{i,(1)} - p_{i,(2)})$, and Entropy the largest $H(p_i)$, where $p_{i,(1)} \geq p_{i,(2)}$ are the two largest probabilities.

For R_{MC} stochastic model passes, write $p_i^{[r]}$ and $\bar{p}_i = R_{\mathrm{MC}}^{-1}\sum_r p_i^{[r]}$. BALD (Houlsby et al., 2011) uses

$$s_i^{\mathrm{BALD}} = H(\bar{p}_i) - R_{\mathrm{MC}}^{-1}\sum_r H(p_i^{[r]}). \tag{125}$$

For reference candidates P_{ref}, EPIG (Bickford Smith et al., 2023) defines

$$\begin{aligned}P_{ij}(k,\ell) &= R_{\mathrm{MC}}^{-1}\sum_r p_{ik}^{[r]}p_{j\ell}^{[r]},\\ s_i^{\mathrm{EPIG}} &= |P_{\mathrm{ref}}|^{-1}\sum_{j\in P_{\mathrm{ref}}}\sum_{k,\ell} P_{ij}(k,\ell)\log\frac{P_{ij}(k,\ell)}{\bar{p}_{ik}\bar{p}_{j\ell}}.\end{aligned} \tag{126}$$

Both methods take the highest-scoring batch. The protocol fixes stochastic passes and reference-candidate selection.

BADGE (Ash et al., 2020) constructs gradient representations

$$g_i = (p_i - e_{\widehat{y}_i}) \otimes z_i, \qquad \widehat{y}_i = \arg\max_k p_{ik}, \tag{127}$$

then performs k-means++ center sampling on $\{g_i\}$. CoreSet (Sener & Savarese, 2018) initializes centers with measured representations and repeatedly chooses

$$i^\star = \arg\max_{i \in P \setminus J_{\text{batch}}} \min_{z \in Z_I \cup Z_{J_{\text{batch}}}} \|z_i - z\|_2^2, \qquad J_{\text{batch}} \leftarrow J_{\text{batch}} \cup \{i^\star\}, \tag{128}$$

until the batch is full. Here J_{batch} is the within-batch selected set, distinct from prediction indices J. Without initial centers, use the deterministic starting point in the tool specification. Measured samples used as centers do not require target labels.

A.16 Symbols and Execution Configuration

Symbols	Definition or selection rule
$K, \mathcal{S}, \mathcal{B}, c_s$	Number of known classes, channels, candidate bundles, and unit measurement costs; specified by the task.
$b_0, b_{\text{anc}}, n_0, \{n_t\}$	Initial and reference bundles, initial sample size, and checkpoints; fixed by the execution protocol.
$\mathcal{A}_j, \mathcal{H}$	Stage candidates and nonexclusive diagnostic labels; availability follows input types and dependencies.
$\psi_j, d_{j,h}, f_{j,a}, v_{j,a}$	Stage reference transforms, diagnostic classifiers, and effect predictors; learned from offline tasks.
$\eta, \delta_B, \delta_M, \delta_j$	Health threshold, bundle tolerance, and switching thresholds; selected by their validation rules.
$\kappa, \tau_G, \Delta n_t$	Continuation disagreement penalty, gain threshold, and next-batch size; gain training matches the batch schedule.
λ_V, Θ_j	Validation tolerance factor and candidate controls; fixed before training and testing.
$R_G, R_U, B_{\text{boot}}, R_{\text{sub}}$	Ensemble and resampling settings; configuration parameters rather than task estimands.
$\widehat{\beta}, \widehat{\rho}, p_{\text{open}}$	Known-class conditional proportions, unknown mass, and task-level openness score; estimated separately.
$\mathcal{V}_{\mathcal{P},a}, M_t, \nu_t$	Optional verification operator, re-entry memory, and evidence version; used only under a registered protocol.
$O_t, C^{\text{obs}}, C^{\text{fit}}$	Observed target sample–channel pairs, complete observation cost, and the specified fitting cost.

Numerical protection constants, optimization tolerances, update counts, tree-model capacity, quantile levels, and random seeds are fixed in the configuration. Rare known classes have no separate mandatory routing branch. Reference-decomposition tools such as RCTD and information-design tools such as FWSP may be registered separately when their input and model assumptions match; they are not enabled members of the general candidate sets above.

B EXPERIMENTAL PROTOCOLS

Controlled stage routing and the executed short chain provide the main conditional-selection evidence. The archival 1,232-unit endpoint fixes Raw Classic, EMQ, prior-corrected source classification, and a hurdle unknown-mass head. Separate resource, re-entry, and external suites evaluate the protocol instances summarized below.

Table 5: Evaluated instances. Controlled stage and chain records support the core selection claims; other instances provide complementary evidence under separate protocols.

Instance	Selected and fixed components	Held-out units
Controlled stages	One stage varies; validation-selected fixed control on the same reduced bank.	624 total
Executed chain	Adaptation → quantification → classification; ALL sensors and fit/audit 96/96.	288 seen; 96 unseen
Archival endpoint	Measurement/resource decisions; Raw Classic, EMQ, prior-corrected source, and hurdle heads fixed.	1,232; five families
Historical stages	Broader portfolios; one stage varies in each fixed context.	236 seen; 108 unseen
Measurement studies	A0/A1/A2 isolates routing/stopping; independent cap comparison uses shared nominal caps.	384 per split in each suite
Re-entry simulator	Full and NoVerifier compute the same diagnostic; only Full changes eligibility.	1,784 reused at two horizons
External acquisition	Exact sensor budgets with common frozen downstream fusion.	REALDISP: 3 folds; Gas: 4 batches

B.1 PARTITIONS AND DATA VISIBILITY

Source labels fit source models and references. Training-task truth supervises candidate losses offline; validation selects fixed actions, margins, and thresholds. Held-out target truth, injected mechanisms, true composition, and audit labels are evaluator-only. Route-visible fitting observations and evaluator audit observations are disjoint.

Table 6: Data roles and visibility. Audit observations are disjoint from route-visible fitting observations.

Role	Permitted use
Source training	Labeled source observations fit classifiers and reference objects.
Offline training	Training-task truth scores candidate actions and supervises routers.
Validation	Select fixed actions, margins, and decision thresholds before held-out scoring.
Target fitting	Unlabeled observations and public costs form routing evidence.
Held-out audit	Labels, mechanisms, composition, and action outcomes are evaluator-only.

The historical primary partition is 32 pilot, 160 route-pool, and 64 audit observations; the separately released consolidated configuration uses 48/144/96. Neither identifies the other runner. The controlled chain instead fixes ALL sensors with 96 fit and 96 audit observations. Random-order sampling and optional feature-driven acquisition are distinct protocols, specified in Appendix A.5.

The archival families contain 384 Seen, 336 Severity, 192 Unseen-mechanism, 192 Unseen-combination, and 128 Discovery-pair units. Severity has seven levels of 48; Discovery contains 64 matched backgrounds. Separate targeted and injection records, released feature fields, detailed partitions, and historical configuration constants remain in the structured supplement. Appendix H specifies the controlled fitting seeds, bank sizes, and released artifacts.

B.2 METRICS AND COST SCOPE

Prevalence error is known-class conditional $\|\widehat{\beta} - \beta\|_1$; Macro-F1 scores known-class prediction. Unknown-mass error is $|\widehat{\rho} - \rho|$. Archival OOD AUROC, AUPR, and FPR95 use 360 eligible units;

prevalence error, Macro-F1, unknown-mass error, and cost use all 1,232. The archival hurdle estimate is $\widehat{P}(\rho > 0 \mid S)\widehat{\mathbb{E}}[\rho \mid \rho > 0, S]$, not the method's direct mass head or a certified calibrated probability.

The adaptation stage suite uses $0.55(1 - \mathrm{F1}) + 0.45\|\widehat{\beta} - \beta\|_1/2$, with EMQ and independent audit predictions; a failed view or EMQ has loss 2. This stage-local scale is distinct from the framework's parameterized proxy and the chain's terminal objective. For open-world loss, let f, r, e_ρ, c, p, v denote known false-positive rate, unknown recall, mass error, residual contamination, cluster pair-F1, and coverage. With discovery action d, the saved suite uses

$$L_{\mathrm{open}} = \begin{cases} 0.70f + 0.25\,\text{residual fraction} + 0.05\mathbb{I}\{d \neq \mathrm{NONE}\}, & \rho = 0, \\ 0.28f + 0.30(1 - r) + 0.16e_\rho + 0.10c + J, & \rho > 0, \end{cases}$$

where $J = 0.12(1 - p) + 0.04(1 - v)$ for at least two unknown subtypes; otherwise $J = 0.03(1 - p)$ if discovery is enabled and zero if not. Truth-dependent terms are used only offline. These losses have different units and must not be compared numerically across stages.

Costs retain each experiment's recorded scope. Historical all-channel fixed policies use 96/128/192 observations at costs 672/896/1,344; NearCost is a nearby-cost reference, not a matched-cost control. External cost is $n_{\mathrm{eval}}k + n_{\mathrm{pilot}}s_{\mathrm{pilot}}$, with pilot observations excluded from evaluation. Equal nominal caps need not yield equal realized cost. The historical ledger limitation is stated in Appendix H.

C HISTORICAL ENDPOINT EVIDENCE

Table 1 and the following family/severity tables preserve complementary archival endpoints. All 1,232 StatOrch units retain Raw Classic, EMQ, and prior-corrected source classification; its open-world endpoint is the historical hurdle output.

C.1 FIVE HELD-OUT FAMILIES

The aggregate result is heterogeneous. Unseen-combination prevalence error is 0.522 for StatOrch versus 0.512/0.501 for NearCost/HighBudget, despite favorable F1 and recorded cost. Seen and unseen-mechanism OOD metrics likewise do not improve uniformly. The complete family table preserves these boundaries.

Table 7: Historical endpoint results in all five held-out families. The row count is the total number of units; AUROC/AUPR use only units with defined OOD truth. Lower known-conditional prevalence error, unknown-mass error, and measurement cost are better; higher Macro-F1 and OOD metrics are better. Every StatOrch row uses Raw Classic, EMQ, prior-corrected-source classification, and the historical hurdle head.

Split	Policy	n	β L1	F1	ρ err.	AUROC	AUPR	Cost
Seen	StatOrch	384	0.172	0.827	0.074	0.802	0.654	849
	NearCost	384	0.218	0.795	0.110	0.813	0.658	896
	HighBudget	384	0.201	0.795	0.109	0.816	0.662	1344
Severity	StatOrch	336	0.168	0.807	0.076	0.648	0.475	770
	NearCost	336	0.317	0.700	0.152	0.620	0.439	896
	HighBudget	336	0.307	0.701	0.152	0.620	0.439	1344
Unseen mechanism	StatOrch	192	0.199	0.826	0.106	0.893	0.854	878
	NearCost	192	0.226	0.821	0.130	0.909	0.874	896
	HighBudget	192	0.206	0.822	0.131	0.909	0.874	1344
Unseen combination	StatOrch	192	0.522	0.536	0.122	0.623	0.561	677
	NearCost	192	0.512	0.458	0.136	0.477	0.405	896
	HighBudget	192	0.501	0.458	0.136	0.469	0.402	1344
Discovery pairs	StatOrch	128	0.117	0.799	0.113	0.656	0.465	817
	NearCost	128	0.136	0.793	0.124	0.617	0.424	896
	HighBudget	128	0.111	0.793	0.124	0.616	0.421	1344

C.2 SEVERITY

Table 8: Historical endpoint results across all seven severity levels, with 48 held-out units per level and policy. Lower known-conditional prevalence error, unknown-mass error, and measurement cost are better; higher Macro-F1 and OOD metrics are better. OOD metrics use the eligible subset within each level. Bold marks the best displayed value at that severity.

Severity	Policy	n	β L1↓	F1↑	ρ err.↓	AUROC↑	AUPR↑	Cost↓
0.05	StatOrch	48	**0.180**	**0.773**	**0.075**	**0.682**	**0.520**	**783**
0.05	NearCost	48	0.332	0.656	0.149	0.635	0.458	896
0.05	HighBudget	48	0.312	0.657	0.150	0.635	0.458	1344
0.20	StatOrch	48	**0.152**	**0.805**	**0.064**	**0.621**	**0.355**	**781**
0.20	NearCost	48	0.325	0.691	0.130	0.610	0.346	896
0.20	HighBudget	48	0.318	0.692	0.131	0.610	0.346	1344
0.40	StatOrch	48	**0.145**	**0.838**	**0.064**	0.595	0.339	**769**
0.40	NearCost	48	0.288	0.719	0.138	**0.611**	**0.341**	896
0.40	HighBudget	48	0.290	0.720	0.137	**0.611**	**0.341**	1344
0.60	StatOrch	48	**0.150**	**0.770**	**0.077**	**0.687**	**0.540**	**765**
0.60	NearCost	48	0.354	0.678	0.152	0.652	0.496	896
0.60	HighBudget	48	0.341	0.678	0.153	0.652	0.496	1344
0.80	StatOrch	48	**0.159**	**0.840**	**0.086**	0.636	**0.542**	**791**
0.80	NearCost	48	0.299	0.730	0.166	**0.647**	0.537	896
0.80	HighBudget	48	0.292	0.731	0.165	**0.647**	0.537	1344
1.00	StatOrch	48	**0.176**	**0.836**	**0.075**	**0.674**	**0.529**	**726**
1.00	NearCost	48	0.290	0.719	0.159	0.637	0.478	896
1.00	HighBudget	48	0.283	0.717	0.157	0.637	0.478	1344
1.20	StatOrch	48	**0.213**	**0.789**	**0.093**	**0.640**	**0.497**	**773**
1.20	NearCost	48	0.333	0.709	0.174	0.551	0.421	896
1.20	HighBudget	48	0.314	0.710	0.173	0.551	0.421	1344

D SINGLE-STAGE ROUTING AND ACTION SPECIALIZATION

D.1 CONTROLLED FIXED-ACTION COMPARISONS

Table 9 retains every controlled stage/split contrast against BestFixed-Val. Positive intervals favor Greedy in six of eight comparisons and Conservative in five. Unseen adaptation ties; the learned-router intervals on unseen quantification include zero. These pointwise task-bootstrap intervals condition on one fitted seed and have no multiplicity correction. The quantification bank has nine named candidates (ACC/BBSE aliases; Appendix A.7).

Table 9: **Controlled stage improvements over BestFixed-Val.** Each cell gives the mean and saved 95% paired interval; positive favors the column policy.

Stage / split	Fallback	HealthRule	Greedy	Conservative	Oracle
Adaptation (S)	-.0063 [-.0155,.0022]	.0079 [.0035,.0124]	.0100 [.0060,.0144]	.0074 [.0028,.0124]	.0115 [.0075,.0159]
Adaptation (U)	-.0830 [-.0953,-.0710]	-.0118 [-.0218,-.0037]	.0000 [.0000,.0000]	.0000 [.0000,.0000]	.0001 [.0000,.0002]
Quantification (S)	-.0126 [-.0437,.0153]	.0295 [.0186,.0408]	.0406 [.0219,.0572]	.0366 [.0177,.0535]	.0678 [.0569,.0794]
Quantification (U)	-.0077 [-.0429,.0263]	.0185 [-.0011,.0369]	.0111 [-.0166,.0368]	.0138 [-.0143,.0399]	.0573 [.0403,.0752]
Classification (S)	.0000 [.0000,.0000]	.0111 [-.0033,.0259]	.0389 [.0243,.0545]	.0364 [.0235,.0507]	.0569 [.0444,.0706]
Classification (U)	.0000 [.0000,.0000]	.0311 [.0185,.0449]	.0511 [.0358,.0666]	.0000 [.0000,.0000]	.0687 [.0549,.0831]
Open-world (S)	-.1025 [-.1352,-.0681]	.0466 [.0271,.0673]	.0764 [.0562,.0965]	.0727 [.0513,.0933]	.0909 [.0730,.1090]
Open-world (U)	-.0736 [-.1049,-.0423]	.0000 [.0000,.0000]	.0276 [.0080,.0466]	.0276 [.0080,.0466]	.0899 [.0789,.1006]

S/U: seen/unseen. Improvement: fixed minus policy loss, or policy minus fixed F1. BestFixed-Val self-comparisons are 0 $[0, 0]$. Saved 10,000-resample pointwise intervals condition on one fitted seed, without multiplicity correction. Oracle is evaluator-only; scenario intervals are supplied as CSV.

D.2 HISTORICAL PORTFOLIO AND ROUTING RULES

Table 10: Saved one-feature HealthRules and conservative switch margins for the separate stage suites. Alternative actions are used only when the rule condition is satisfied; otherwise the listed fallback remains. The scalar thresholds are validation-selected diagnostics, not calibrated causal failure detectors. Conservative margins concern predicted stage advantage and are not significance thresholds.

Stage	Fallback	HealthRule condition and alternative	Margin
Adaptation	Raw Classic	Mean predictive entropy ≥ 0.0169471: EATA-style update	0.03
Quantification	EMQ	Class-4 predictive marginal ≥ 0.0826586: BBSE	0.05
Classification	Source calibrated	Relative covariance discrepancy ≥ 1.59758: adapted ensemble	0.05
Open world	NONE	Median ClassicMSP score $\geq 8.90217 \times 10^{-9}$: ClassicMahalanobis at nominal FPR 0.10	0.08

Each experiment in this appendix selects one stage action in a fixed task context. The following tables preserve the historical candidate banks, scenarios, action effects, and choices; the controlled reduced-bank comparison with validation-selected BestFixed is reported separately in Table 2. Greedy uses the learned ranking directly. Conservative retains the stage fallback unless the validation-calibrated margin is exceeded. HealthRule uses the explicit rule in Table 10. Oracle is an evaluator-only upper reference for the available candidate set.

The deployed policies show regime-dependent effects. On unseen classification, HealthRule's F1 is 0.7177 versus Greedy's 0.7100; Conservative remains at the source fallback, 0.6936. Unseen adaptation is nearly tied between the three deployable alternatives. Open-world Greedy capture drops from 86.2% on seen tasks to 44.4% on unseen tasks, and the near-warp scenario has little available headroom.

D.3 SCENARIO HEADROOM AND SELECTED OUTCOMES

Table 11 lists fallback and Oracle outcomes together with scenario denominators. A scenario's test-best fixed action is a descriptive, evaluator-only summary, distinct from the per-unit Oracle, which may choose a different action for each unit. Neither is a fixed baseline selected using training/validation tasks. The historical suites lack that stronger fixed-action comparator; fallback-to-Oracle headroom therefore does not isolate the benefit of conditioning the route from the benefit of having stronger candidates available. Table 12 then gives each deployed router's selected mean and captured fraction. The closed open-world case has zero Oracle headroom; its capture ratios are marked n/a while its losses remain visible.

Table 11: Scenario-level fallback and evaluator Oracle references for every stage suite. Adaptation, quantification, and open-world values are losses (lower is better); classification is Macro-F1 (higher is better). Headroom is the mean oriented fallback-to-Oracle improvement; Pos. is the fraction with strictly positive unit-level headroom. Oracle action is the best fixed action by scenario mean; the Oracle value permits unit-specific actions and is not that fixed-action mean. Abbreviations are defined with the action matrices.

Split	Scenario	n	Fallback	Oracle	Oracle action	Headroom	Pos.
Adaptation							
seen	BN affine	10	0.1176	0.0440	COME	0.0735	0.900
seen	Clean	10	0.0389	0.0302	SHOT	0.0087	0.900
seen	Covariance shear	10	0.0664	0.0352	COME	0.0312	1.000
seen	Nonlinear	10	0.0432	0.0328	COME	0.0104	0.900
unseen	Mixed	10	0.6440	0.5765	COME	0.0675	0.900
unseen	Rotation	10	0.6253	0.5342	CORAL	0.0911	1.000
Quantification							
seen	Clean label shift	10	0.1531	0.1246	CPMCN	0.0285	0.700
seen	Noisy long tail	10	0.1869	0.1142	PACC	0.0727	1.000
seen	Overconfident	10	0.2036	0.1200	PACC	0.0836	0.900
seen	Posterior corruption	10	0.3442	0.1470	ACC	0.1972	0.900
seen	Underconfident	10	0.0924	0.0664	EMQ	0.0260	0.700
unseen	Asymmetric mix	10	0.2991	0.1428	ACC	0.1563	0.800
unseen	Temperature mix	10	0.1958	0.1586	BBSE	0.0372	0.800
Classification							
seen	Clean	24	0.7517	0.7727	Prior	0.0210	0.708
seen	Covariance	24	0.4914	0.6039	Adapt	0.1126	0.917
seen	Mild covariance	24	0.6456	0.6745	Src	0.0290	0.708
seen	Prior only	24	0.5952	0.6617	Prior	0.0665	0.833
unseen	Block mixture	24	0.6682	0.7184	AEns	0.0502	0.833
unseen	Rank-one covariance	24	0.7190	0.7558	AEns	0.0368	0.708
Open world							
seen	Closed	10	0.0000	0.0000	NONE	0.0000	0.000
seen	Far	10	0.4227	0.1342	ViM/G	0.2885	1.000
seen	Multimodal	10	0.5158	0.1790	CMah/K	0.3368	1.000
seen	Near covariance	10	0.3452	0.2290	CMah/N	0.1162	1.000
seen	Tiny	10	0.3065	0.0676	CMah/N	0.2389	1.000
unseen	Mixed density	10	0.5194	0.2014	ViM/K	0.3180	1.000
unseen	Near warp	10	0.4599	0.4185	Mah/N	0.0414	0.800

Table 12: Complete scenario-level outcomes of the three deployable stage routers. Each cell is selected mean / captured fraction of fallback-to-Oracle headroom. The original capture convention clips negative gains to zero; the selected mean preserves any degradation. Zero headroom is marked n/a. Lower values are better except for classification Macro-F1. Fallback and Oracle means appear once in Table 11.

Split	Scenario	n	HealthRule	Greedy	Conservative
Adaptation					
seen	BN affine	10	0.0451 / 0.985	0.0451 / 0.986	0.0464 / 0.968
seen	Clean	10	0.0396 / 0.000	0.0350 / 0.442	0.0389 / 0.000
seen	Covariance shear	10	0.0392 / 0.873	0.0356 / 0.988	0.0448 / 0.693
seen	Nonlinear	10	0.0392 / 0.385	0.0369 / 0.604	0.0424 / 0.079
unseen	Mixed	10	0.5775 / 0.985	0.5775 / 0.985	0.5775 / 0.985
unseen	Rotation	10	0.5495 / 0.832	0.5495 / 0.832	0.5495 / 0.832
Quantification					
seen	Clean label shift	10	0.1616 / 0.000	0.1556 / 0.000	0.1531 / 0.000
seen	Noisy long tail	10	0.2096 / 0.000	0.1161 / 0.974	0.1404 / 0.640
seen	Overconfident	10	0.1757 / 0.333	0.1332 / 0.842	0.2036 / 0.000
seen	Posterior corruption	10	0.2402 / 0.527	0.1812 / 0.827	0.1688 / 0.890

Continued on next page

Table 12 (continued)

Split	Scenario	n	HealthRule	Greedy	Conservative
seen	Underconfident	10	0.1365 / 0.000	0.1798 / 0.000	0.0924 / 0.000
unseen	Asymmetric mix	10	0.1998 / 0.635	0.1766 / 0.784	0.1839 / 0.737
unseen	Temperature mix	10	0.1881 / 0.206	0.1785 / 0.466	0.2017 / 0.000
Classification					
seen	Clean	24	0.7517 / 0.000	0.7580 / 0.299	0.7517 / 0.000
seen	Covariance	24	0.5717 / 0.714	0.5717 / 0.714	0.5717 / 0.714
seen	Mild covariance	24	0.6438 / 0.000	0.6467 / 0.039	0.6456 / 0.000
seen	Prior only	24	0.5952 / 0.000	0.6464 / 0.770	0.6340 / 0.583
unseen	Block mixture	24	0.7036 / 0.706	0.6908 / 0.451	0.6682 / 0.000
unseen	Rank-one covariance	24	0.7318 / 0.348	0.7291 / 0.274	0.7190 / 0.000
Open world					
seen	Closed	10	0.1257 / n/a	0.0238 / n/a	0.0000 / n/a
seen	Far	10	0.2179 / 0.710	0.1774 / 0.850	0.1774 / 0.850
seen	Multimodal	10	0.2778 / 0.707	0.1980 / 0.943	0.1980 / 0.943
seen	Near covariance	10	0.2449 / 0.863	0.2303 / 0.989	0.2937 / 0.444
seen	Tiny	10	0.2655 / 0.172	0.1151 / 0.801	0.1825 / 0.519
unseen	Mixed density	10	0.3485 / 0.537	0.3127 / 0.650	0.3127 / 0.650
unseen	Near warp	10	0.5010 / 0.000	0.5071 / 0.000	0.5071 / 0.000

Table 13: Aggregate captured fractions of fallback-to-Oracle headroom for all deployable routing strategies. Historical scenario counts and policy means appear in Table 12. These fractions concern separate stage suites; they are not additive components of an end-to-end effect. Conservative denotes the direct thresholded stage router.

Stage	Split	n	HealthRule	Greedy	Conservative
Adaptation	seen	40	0.831	0.916	0.756
Adaptation	unseen	20	0.897	0.897	0.897
Quantification	seen	50	0.139	0.525	0.544
Quantification	unseen	20	0.553	0.723	0.565
Classification	seen	96	0.343	0.607	0.520
Classification	unseen	48	0.555	0.376	0.000
Open world	seen	50	0.468	0.862	0.753
Open world	unseen	20	0.361	0.444	0.444

D.4 Every candidate action by scenario

The action matrices retain both mean effect and evaluator-best frequency. Mean effect describes committing to that action within the scenario; best frequency describes how often it is optimal for an individual evaluation unit. Neither is an actual learned-router choice frequency. For adaptation, the candidate set includes the no-adaptation Raw Classic and neural Source options, CORAL, and the evaluated TENT-, EATA-, SAR-, SHOT-, and COME-style updates. These implementation labels do not claim full reproduction of each original method's training and evaluation pipeline.

Table 14: Adaptation action effects on seen scenarios. Each cell is mean loss / evaluator-best fraction; lower mean is better. Best fraction is the rate at which a candidate is evaluator-optimal, not a deployed router choice rate. Each scenario has 10 units. Raw = Raw Classic; Source is the neural source candidate. Named neural tools use the evaluated method-style update implementations.

	BN		Covariance	
Action	affine	Clean	shear	Nonlinear
Raw	0.1176 / 0.100	0.0389 / 0.100	0.0664 / 0.000	0.0432 / 0.100
CORAL	0.0581 / 0.100	0.0409 / 0.100	0.1019 / 0.000	0.0440 / 0.400
Source	0.1175 / 0.000	0.0383 / 0.200	0.0581 / 0.000	0.0414 / 0.100
TENT	0.0451 / 0.000	0.0371 / 0.000	0.0356 / 0.000	0.0407 / 0.000
EATA	0.0451 / 0.100	0.0371 / 0.000	0.0356 / 0.200	0.0407 / 0.000
SAR	0.0451 / 0.000	0.0371 / 0.000	0.0356 / 0.100	0.0406 / 0.000
SHOT	0.0985 / 0.100	0.0327 / 0.300	0.0506 / 0.100	0.0430 / 0.100
COME	0.0450 / 0.600	0.0370 / 0.300	0.0355 / 0.600	0.0406 / 0.300

Table 15: Adaptation action effects on unseen scenarios. Each cell is mean loss / evaluator-best fraction; lower mean is better. Best fraction is the rate at which a candidate is evaluator-optimal, not a deployed router choice rate. Each scenario has 10 units. Raw = Raw Classic; Source is the neural source candidate. Named neural tools use the evaluated method-style update implementations.

Action	Mixed	Rotation
Raw	0.6440 / 0.100	0.6253 / 0.000
CORAL	0.5971 / 0.000	0.5378 / 0.700
Source	0.6529 / 0.000	0.6236 / 0.000
TENT	0.5775 / 0.200	0.5495 / 0.000
EATA	0.5775 / 0.100	0.5495 / 0.000
SAR	0.5774 / 0.000	0.5494 / 0.100
SHOT	0.6324 / 0.100	0.6182 / 0.000
COME	0.5772 / 0.500	0.5493 / 0.200

Table 16: Quantification action effects on seen scenarios. Each cell is mean β-L1 / evaluator-best fraction; lower mean is better. Best fraction is the rate at which a candidate is evaluator-optimal, not a deployed router choice rate. Each scenario has 10 units. GMM-ref is the source-reference Gaussian-mixture quantifier; CPMCN denotes the evaluated solver implementation.

Action	Clean label shift	Noisy long tail	Overconfident	Posterior corruption	Underconfident
CC	0.3546 / 0.000	0.4085 / 0.000	0.2933 / 0.000	0.3585 / 0.000	0.2961 / 0.000
PCC	0.4175 / 0.000	0.4546 / 0.000	0.3000 / 0.000	0.6874 / 0.000	0.4487 / 0.000
ACC	0.1680 / 0.000	0.2105 / 0.100	0.1484 / 0.100	0.1812 / 0.500	0.1764 / 0.000
PACC	0.1545 / 0.100	0.1161 / 0.900	0.1333 / 0.500	0.5406 / 0.100	0.2470 / 0.100
BBSE	0.1680 / 0.300	0.2105 / 0.000	0.1484 / 0.300	0.1812 / 0.100	0.1764 / 0.000
RLLS	0.4489 / 0.000	0.5319 / 0.000	0.3431 / 0.000	0.7073 / 0.000	0.5082 / 0.000
EMQ	0.1531 / 0.300	0.1869 / 0.000	0.2036 / 0.100	0.3442 / 0.100	0.0924 / 0.300
CPMCN	0.1459 / 0.300	0.1843 / 0.000	0.2188 / 0.000	0.4391 / 0.200	0.1164 / 0.600
GMM-ref	0.8898 / 0.000	1.3612 / 0.000	0.8278 / 0.000	1.0702 / 0.000	1.0823 / 0.000

Table 17: Quantification action effects on unseen scenarios. Each cell is mean β-L1 / evaluator-best fraction; lower mean is better. Best fraction is the rate at which a candidate is evaluator-optimal, not a deployed router choice rate. Each scenario has 10 units. GMM-ref is the source-reference Gaussian-mixture quantifier; CPMCN denotes the evaluated solver implementation.

Action	Asymmetric mix	Temperature mix
CC	0.3004 / 0.000	0.3298 / 0.000
PCC	0.5804 / 0.000	0.4549 / 0.000
ACC	0.1856 / 0.300	0.1748 / 0.000
PACC	0.4312 / 0.000	0.2912 / 0.000
BBSE	0.1856 / 0.200	0.1748 / 0.300
RLLS	0.6068 / 0.000	0.4842 / 0.000
EMQ	0.2991 / 0.200	0.1958 / 0.200
CPMCN	0.2884 / 0.300	0.1956 / 0.500
GMM-ref	0.9066 / 0.000	0.8739 / 0.000

The classification candidates include ordinary source prediction, prior correction, adaptation, adaptation followed by prior correction, and the source and adapted ensembles. Ordinary classification can remain preferable when an intervention does not help. The final seen/unseen cohorts are used throughout the classification effect and choice tables.

Table 18: Classification action effects on seen scenarios. Each cell is mean Macro-F1 / evaluator-best fraction; higher mean is better. Best fraction is the rate at which a candidate is evaluator-optimal, not a deployed router choice rate. Each scenario has 24 units. Src = source calibrated; Prior = prior-corrected source; Adapt = adapted; APrior = adapted and prior corrected; SEns/AEns = source/adapted ensemble.

Action	Clean	Covariance	Mild covariance	Prior only
Src	0.7517 / 0.292	0.4914 / 0.083	0.6456 / 0.292	0.5952 / 0.167
Prior	0.7584 / 0.417	0.4447 / 0.000	0.6232 / 0.167	0.6464 / 0.792
Adapt	0.6973 / 0.000	0.5729 / 0.417	0.6263 / 0.167	0.3436 / 0.000
APrior	0.7342 / 0.208	0.5335 / 0.125	0.6332 / 0.125	0.4832 / 0.042
SEns	0.7517 / 0.000	0.4914 / 0.000	0.6456 / 0.000	0.5952 / 0.000

Continued on next page

Table 18 (continued)

Action	Clean	Covariance	Mild covariance	Prior only
AEns	0.7343 / 0.083	0.5717 / 0.375	0.6438 / 0.250	0.4767 / 0.000

Table 19: Classification action effects on unseen scenarios. Each cell is mean Macro-F1 / evaluator-best fraction; higher mean is better. Best fraction is the rate at which a candidate is evaluator-optimal, not a deployed router choice rate. Each scenario has 24 units. Src = source calibrated; Prior = prior-corrected source; Adapt = adapted; APrior = adapted and prior corrected; SEns/AEns = source/adapted ensemble.

Action	Block mixture	Rank-one covariance
Src	0.6682 / 0.167	0.7190 / 0.292
Prior	0.6111 / 0.000	0.6984 / 0.167
Adapt	0.6654 / 0.125	0.7097 / 0.083
APrior	0.6846 / 0.208	0.7261 / 0.167
SEns	0.6682 / 0.000	0.7190 / 0.000
AEns	0.7036 / 0.500	0.7318 / 0.292

The open-world candidate is a detector/threshold/cluster combination, with a NONE fallback. The nominal detector FPR setting is 0.10; it is a calibration configuration, not an asserted target FPR guarantee. Routing losses and matched-injection responses describe decision quality and directional sensitivity, respectively; neither establishes calibrated unknown mass or semantic class discovery. Detector and clustering names in the matrices identify the actual evaluated combinations.

Table 20: Open world action effects on seen scenarios. Each cell is mean loss / evaluator-best fraction; lower mean is better. Best fraction is the rate at which a candidate is evaluator-optimal, not a deployed router choice rate. Each scenario has 10 units. Action detector/cluster uses CMSP = ClassicMSP, CMah = ClassicMahalanobis, Mah = Mahalanobis, and N/K/G/H = NONE/KMeans/GMM-BIC/HDBSCAN. Every detector threshold is calibrated at nominal FPR 0.10; NONE is the no-analysis fallback.

Action	Closed	Far	Multimodal	Near covariance	Tiny
NONE	0.0000 / 1.000	0.4227 / 0.000	0.5158 / 0.000	0.3452 / 0.000	0.3065 / 0.000
CMSP/N	0.1346 / 0.000	0.2156 / 0.000	0.3323 / 0.000	0.2784 / 0.200	0.1243 / 0.300
CMSP/K	0.1846 / 0.000	0.2177 / 0.100	0.2838 / 0.000	0.3679 / 0.000	0.2185 / 0.000
CMSP/G	0.1846 / 0.000	0.2134 / 0.000	0.2858 / 0.000	0.3674 / 0.000	0.2169 / 0.000
CMSP/H	0.1846 / 0.000	0.2255 / 0.000	0.3131 / 0.000	0.3401 / 0.000	0.1943 / 0.000
CMah/N	0.2138 / 0.000	0.1917 / 0.200	0.2778 / 0.000	0.2303 / 0.800	0.0949 / 0.400
CMah/K	0.2637 / 0.000	0.1829 / 0.000	0.1974 / 0.400	0.3091 / 0.000	0.1916 / 0.000
CMah/G	0.2637 / 0.000	0.1787 / 0.000	0.2043 / 0.100	0.3075 / 0.000	0.1876 / 0.000
CMah/H	0.2637 / 0.000	0.1798 / 0.200	0.2165 / 0.200	0.2877 / 0.000	0.1786 / 0.000
Energy/N	0.1187 / 0.000	0.2249 / 0.000	0.3250 / 0.000	0.3123 / 0.000	0.1429 / 0.000
Energy/K	0.1688 / 0.000	0.2296 / 0.000	0.2793 / 0.100	0.4120 / 0.000	0.2397 / 0.000
Energy/G	0.1688 / 0.000	0.2229 / 0.000	0.2796 / 0.000	0.4115 / 0.000	0.2388 / 0.000
Energy/H	0.1688 / 0.000	0.2378 / 0.000	0.3151 / 0.000	0.3887 / 0.000	0.2256 / 0.000
Mah/N	0.1811 / 0.000	0.3407 / 0.000	0.4152 / 0.000	0.2808 / 0.000	0.2503 / 0.000
Mah/K	0.2311 / 0.000	0.3998 / 0.000	0.4196 / 0.000	0.3714 / 0.000	0.3658 / 0.000
Mah/G	0.2311 / 0.000	0.3983 / 0.000	0.4182 / 0.000	0.3710 / 0.000	0.3654 / 0.000
Mah/H	0.2311 / 0.000	0.3921 / 0.000	0.4187 / 0.000	0.3469 / 0.000	0.3594 / 0.000
ReAct/N	0.1247 / 0.000	0.2178 / 0.000	0.3147 / 0.000	0.3011 / 0.000	0.1460 / 0.000
ReAct/K	0.1747 / 0.000	0.2163 / 0.000	0.2572 / 0.000	0.3978 / 0.000	0.2438 / 0.000
ReAct/G	0.1747 / 0.000	0.2133 / 0.000	0.2635 / 0.100	0.3971 / 0.000	0.2428 / 0.000
ReAct/H	0.1747 / 0.000	0.2294 / 0.000	0.2936 / 0.000	0.3693 / 0.000	0.2237 / 0.000
ViM/N	0.1207 / 0.000	0.1807 / 0.300	0.2986 / 0.000	0.3061 / 0.000	0.1106 / 0.300
ViM/K	0.1707 / 0.000	0.1719 / 0.000	0.2304 / 0.100	0.4035 / 0.000	0.2032 / 0.000
ViM/G	0.1707 / 0.000	0.1616 / 0.200	0.2385 / 0.000	0.4029 / 0.000	0.2005 / 0.000
ViM/H	0.1707 / 0.000	0.1905 / 0.000	0.2766 / 0.000	0.3821 / 0.000	0.1904 / 0.000

Table 21: Open world action effects on unseen scenarios. Each cell is mean loss / evaluator-best fraction; lower mean is better. Best fraction is the rate at which a candidate is evaluator-optimal, not a deployed router choice rate. Each scenario has 10 units. Action detector/cluster uses CMSP = ClassicMSP, CMah = ClassicMahalanobis, Mah = Mahalanobis, and N/K/G/H = NONE/KMeans/GMM-BIC/HDBSCAN. Every detector threshold is calibrated at nominal FPR 0.10; NONE is the no-analysis fallback.

Action	Mixed density	Near warp
NONE	0.5194 / 0.000	0.4599 / 0.200
CMSP/N	0.3311 / 0.000	0.4889 / 0.000
CMSP/K	0.2922 / 0.000	0.5035 / 0.000
CMSP/G	0.2903 / 0.000	0.4936 / 0.000
CMSP/H	0.3103 / 0.000	0.5211 / 0.000
CMah/N	0.3485 / 0.000	0.5010 / 0.000
CMah/K	0.3129 / 0.000	0.5071 / 0.000
CMah/G	0.3104 / 0.100	0.4812 / 0.300
CMah/H	0.3127 / 0.100	0.5190 / 0.000
Energy/N	0.3093 / 0.000	0.4976 / 0.000
Energy/K	0.2575 / 0.000	0.5172 / 0.000
Energy/G	0.2575 / 0.000	0.4988 / 0.000
Energy/H	0.2878 / 0.000	0.5330 / 0.000
Mah/N	0.4863 / 0.000	0.4429 / 0.200
Mah/K	0.5324 / 0.000	0.5067 / 0.000
Mah/G	0.5320 / 0.000	0.4984 / 0.100
Mah/H	0.5214 / 0.000	0.4766 / 0.000
ReAct/N	0.3035 / 0.000	0.5009 / 0.000
ReAct/K	0.2462 / 0.100	0.5138 / 0.000
ReAct/G	0.2477 / 0.000	0.4958 / 0.200
ReAct/H	0.2814 / 0.000	0.5362 / 0.000
ViM/N	0.2904 / 0.000	0.5116 / 0.000
ViM/K	0.2213 / 0.500	0.5236 / 0.000
ViM/G	0.2287 / 0.100	0.5047 / 0.000
ViM/H	0.2637 / 0.100	0.5453 / 0.000

D.5 Actual routing distributions

The next four tables retain every scenario's HealthRule, Greedy, and Conservative choice count. Fractions are count divided by the displayed row denominator. The full per-unit choice records preserve the unit identifiers, fallback, and Oracle choice. Fallback choices are constant within each stage; Oracle choice frequencies are already given by the action matrices' best-frequency entries. This separation prevents evaluator-optimal action shares from being presented as the controller's deployed behavior.

Table 22: Actual adaptation routing choices on each held-out scenario. Cells give action: count; the choice fraction is count divided by the row denominator. Omitted actions have zero choices. Fallback always selects Raw Classic. The corresponding evaluator Oracle action frequencies are the best fractions in the action matrices; they are not deployed choices. Action abbreviations match those matrices.

Split	Scenario	n	HealthRule	Greedy	Conservative
seen	Clean	10	EATA: 7, Raw: 3	COME: 7, SHOT: 2, Raw: 1	Raw: 10
seen	Covariance shear	10	EATA: 9, Raw: 1	EATA: 10	Raw: 7, EATA: 3
seen	BN affine	10	EATA: 10	EATA: 8, COME: 2	EATA: 8, COME: 1, Raw: 1
seen	Nonlinear	10	EATA: 6, Raw: 4	COME: 6, Raw: 2, EATA: 1, SHOT: 1	Raw: 9, EATA: 1
unseen	Rotation	10	EATA: 10	EATA: 8, COME: 2	EATA: 8, COME: 2
unseen	Mixed	10	EATA: 10	EATA: 10	EATA: 10

Table 23: Actual quantification routing choices on each held-out scenario. Cells give action: count; the choice fraction is count divided by the row denominator. Omitted actions have zero choices. Fallback always selects EMQ. The corresponding evaluator Oracle action frequencies are the best fractions in the action matrices; they are not deployed choices. Action abbreviations match those matrices.

Split	Scenario	n	HealthRule	Greedy	Conservative
seen	Clean label shift	10	BBSE: 7, EMQ: 3	EMQ: 5, PACC: 5	EMQ: 10
seen	Overconfident	10	BBSE: 5, EMQ: 5	PACC: 8, BBSE: 2	EMQ: 10
seen	Underconfident	10	BBSE: 6, EMQ: 4	EMQ: 4, BBSE: 3, PACC: 3	EMQ: 10
seen	Posterior corruption	10	BBSE: 8, EMQ: 2	BBSE: 10	BBSE: 7, EMQ: 3

Continued on next page

Table 23 (continued)

Split	Scenario	n	HealthRule	Greedy	Conservative
seen	Noisy long tail	10	EMQ: 6, BBSE: 4	PACC: 10	PACC: 6, EMQ: 4
unseen	Asymmetric mix	10	BBSE: 7, EMQ: 3	BBSE: 9, EMQ: 1	BBSE: 8, EMQ: 2
unseen	Temperature mix	10	BBSE: 6, EMQ: 4	BBSE: 6, EMQ: 4	EMQ: 9, BBSE: 1

Table 24: Actual classification routing choices on each held-out scenario. Cells give action: count; the choice fraction is count divided by the row denominator. Omitted actions have zero choices. Fallback always selects source calibrated. The corresponding evaluator Oracle action frequencies are the best fractions in the action matrices; they are not deployed choices. Action abbreviations match those matrices.

Split	Scenario	n	HealthRule	Greedy	Conservative
seen	Clean	24	Src: 24	Prior: 21, APrior: 3	Src: 24
seen	Prior only	24	Src: 24	Prior: 24	Prior: 12, Src: 12
seen	Mild covariance	24	AEns: 24	AEns: 20, Src: 4	Src: 24
seen	Covariance	24	AEns: 24	AEns: 24	AEns: 24
unseen	Block mixture	24	AEns: 24	AEns: 18, Src: 6	Src: 24
unseen	Rank-one covariance	24	AEns: 24	AEns: 18, Src: 6	Src: 24

Table 25: Actual open world routing choices on each held-out scenario. Cells give action: count; the choice fraction is count divided by the row denominator. Omitted actions have zero choices. Fallback always selects NONE. The corresponding evaluator Oracle action frequencies are the best fractions in the action matrices; they are not deployed choices. Action abbreviations match those matrices.

Split	Scenario	n	HealthRule	Greedy	Conservative
seen	Closed	10	CMah/N: 6, NONE: 4	NONE: 9, CMah/N: 1	NONE: 10
seen	Far	10	CMah/N: 9, NONE: 1	CMah/G: 7, CMah/K: 3	CMah/G: 7, CMah/K: 3
seen	Near covariance	10	CMah/N: 9, NONE: 1	CMah/N: 10	CMah/N: 5, NONE: 5
seen	Multimodal	10	CMah/N: 10	CMah/K: 8, CMah/G: 2	CMah/K: 8, CMah/G: 2
seen	Tiny	10	NONE: 8, CMah/N: 2	CMah/N: 9, NONE: 1	CMah/N: 6, NONE: 4
unseen	Near warp	10	CMah/N: 10	CMah/K: 10	CMah/K: 10
unseen	Mixed density	10	CMah/N: 10	CMah/K: 8, CMah/G: 2	CMah/K: 8, CMah/G: 2

D.6 Paired effects and classification uncertainty

The archived non-classification paired intervals are retained as reported uncertainty summaries in Table 26; they were not newly estimated from retrained models. Classification has two inconsistent historical bootstrap summaries and lacks the original paired action-value matrices needed to resolve them with one resampling procedure. Consequently, Table 27 retains all saved mean policy contrasts without asserting confidence intervals. Direct policy means support the seen/unseen comparisons above. The historical conflicting interval records remain in the provenance data, rather than selecting the more favorable set. Controlled-bank intervals are reported separately in Table 9.

Table 26: Source-reported paired-bootstrap gains of Greedy routing in the adaptation, quantification, and open-world suites. Gain is reference loss minus Greedy loss, so positive values favor Greedy. These saved intervals concern held-out units within each fixed suite and were not recomputed from original resampling inputs. Classification uses direct mean contrasts in Table 27; no classification interval is asserted.

Stage	Split	Reference	Mean gain	95% low	95% high	n
adaptation	seen	Fallback	0.0284	0.0155	0.0441	40
adaptation	seen	HealthRule	0.0026	0.0003	0.0054	40
adaptation	unseen	Fallback	0.0712	0.0512	0.0916	20
adaptation	unseen	HealthRule	0.0000	0.0000	0.0001	20
quantification	seen	Fallback	0.0429	0.0054	0.0856	50
quantification	seen	HealthRule	0.0316	0.0030	0.0658	50
quantification	unseen	Fallback	0.0699	0.0079	0.1391	20
quantification	unseen	HealthRule	0.0165	0.0007	0.0363	20
open world	seen	Fallback	0.1691	0.1321	0.2048	50
open world	seen	HealthRule	0.0774	0.0527	0.1029	50
open world	unseen	Fallback	0.0798	0.0143	0.1487	20
open world	unseen	HealthRule	0.0149	-0.0086	0.0376	20

Table 27: Historical mean Macro-F1 contrasts for all saved classification policy/reference comparisons. Positive values favor the named policy. The seen and unseen sets contain 96 and 48 units. Original paired action-value inputs are unavailable and two saved interval summaries conflict; no confidence interval is asserted. Scenario policy means appear in Table 12.

Split	Policy	Reference	Mean F1 gain	n
seen	HealthRule	Fallback	0.0197	96
seen	Greedy	Fallback	0.0347	96
seen	Greedy	HealthRule	0.0151	96
seen	Conservative	Fallback	0.0298	96
seen	Conservative	HealthRule	0.0101	96
unseen	HealthRule	Fallback	0.0241	48
unseen	Greedy	Fallback	0.0164	48
unseen	Greedy	HealthRule	-0.0078	48
unseen	Conservative	Fallback	0.0000	48
unseen	Conservative	HealthRule	-0.0241	48

D.7 Matched Open-World Response

The 64 matched backgrounds reuse the 128 Discovery-pair units in the archival endpoint evaluation. Each pair shares a background mechanism, with unknown classes injected in the positive target. Figure 3 compares the saved open-presence output and historical hurdle mass estimate before and after injection; it does not evaluate the direct mass regressor in Section 4.

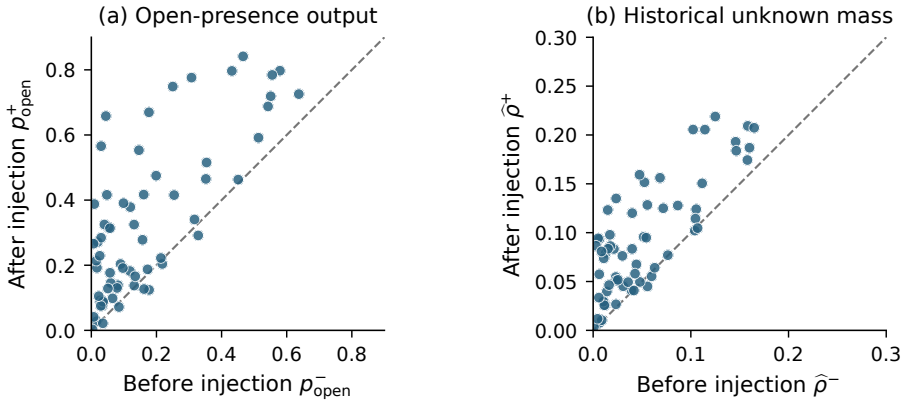


Figure 3: **Matched open-world response.** Unknown injection increases p_{open} in 57/64 pairs (89.1%) and $\widehat{\rho}$ in 59/64 pairs (92.2%). Each point is a matched background; points above $y = x$ indicate an increase. The mass output is the historical hurdle estimate. This evaluates directional sensitivity rather than probability calibration or semantic discovery.

E Measurement and Resource Analysis

E.1 Measurement and stopping endpoints

Table 28: **Aggregate measurement and stopping endpoints.** Each split contains 384 units. A0 fixes measurement; A1 adds routing; A2 adds stopping. Bold marks the best split/metric value.

Split	Policy	β L_1 ↓	Macro-F1↑	Mean cost↓
Standard	A0 fixed	0.2314	0.7922	**432.0**
	A1 + routing	**0.1840**	**0.8286**	623.5
	A2 + stopping	0.1843	0.8280	617.9
Shift	A0 fixed	0.2822	0.7644	**432.0**
	A1 + routing	0.2508	**0.8064**	620.2
	A2 + stopping	**0.2502**	0.8050	602.5

Recorded measurement costs; accounting scope: Appendix B.

This independent five-class, four-sensor suite has 384 units per split. A0 fixes measurement, A1 adds routing, and A2 adds checkpoint stopping. A0-to-A1 gains accompany higher mean cost: 432.0 to 623.5 on Standard and 620.2 on Shift. Stopping reduces those costs to 617.9 and 602.5 with small metric changes. This is a prediction/resource trade-off, distinct from the matched-cost experiment in Table 30.

E.2 Sensitivity to severe Sensor-4 failure

Table 29: **A0-to-A1 sensitivity excluding Sensor-4 failure.** Both predictive metrics worsen after excluding this scenario. Means are recomputed from the saved unrounded scenario results, weighted by unit counts.

Split	Metric	A0 fixed	A1 routing	A1 minus A0
Standard	β-L_1 ↓	0.1734	0.1806	+0.0073
Standard	Macro-F1↑	0.8448	0.8363	-0.0084
Shift	β-L_1 ↓	0.2439	0.2594	+0.0155
Shift	Macro-F1↑	0.8148	0.8040	-0.0109

Each remaining split has 15 scenarios and 360 units. This is a sensitivity summary of the unequal-cost A0/A1 experiment, not the separate matched-cost comparison. All 16 leave-one-scenario-out summaries, including A2, remain in CSV.

The aggregate A0-to-A1 improvement concentrates in the severe Sensor-4 failure scenario. Its prevalence error decreases from 1.1014 to 0.2355 on Standard and from 0.8558 to 0.1216 on Shift. Excluding that scenario, routing worsens both average prevalence error and F1 in each split (Table 29). We interpret this experiment primarily as evidence of recovery from catastrophic sensor failure, rather than uniform gains across heterogeneous shifts. This deterministic sensitivity summary uses the saved scenario means and counts; it neither retrains a model nor changes the held-out cohort retrospectively. All leave-one-scenario-out summaries and the full A0/A1/A2 scenario records are supplied as CSV.

E.3 MATCHED COST AND BROADER RESOURCE RECORDS

Table 30: **Comparison at identical recorded measurement cost 432.** Each split contains 384 units; all policies are feasible. Bold marks the best split/metric mean.

Split	Metric	BestFixed	HealthRule	StatOrch
Standard	β-$L_1 \downarrow$	0.2274	0.2033	**0.2009**
Standard	Macro-F1↑	0.8007	**0.8402**	0.8353
Shift	β-$L_1 \downarrow$	0.2801	**0.2695**	0.2725
Shift	Macro-F1↑	0.7755	**0.8186**	0.8131

This independent common-cost experiment is separate from A0/A1/A2. The full cost-cap frontier remains in the structured supplement.

Table 30 retains all three policies at identical recorded cost 432. HealthRule has higher F1 on both splits and lower prevalence error on Shift. The full frontier at caps 336, 432, 576, 768, and 1,008 is supplied as structured data; on Shift, HealthRule has lower error and higher F1 than StatOrch at every recorded cap. This matched-cost point represents the trade-off, not uniform superiority. High caps can remain partly unused.

The separate broader acquisition comparison, historical failure/bundle audits, and checkpoint reasons remain in the structured supplement. In the archival primary run, 98.5% of units stop at 144 and the permitted 192 maximum is not reached; stopping effects are modest and stopping is not a declaration of statistical correctness. These records retain their original costs and scenario denominators.

F Re-entry Control

F.1 Constraints and eligibility

Eligibility conditions reduce unsupported-action exposure at different realized costs (Table 31); they do not establish equal-cost predictive superiority or statistical validity.

Table 31: **Eligibility constraints and unsupported-action exposure.** Each split has 384 held-out units; realized acquisition costs differ.

Split	Full unsupported	No-contract unsupported	Full cost	No-contract cost
Standard	0.00%	2.08%	564.2	344.6
Shift	1.30%	8.33%	573.0	364.9

Unsupported rates are evaluator-identified action exposures. This comparison does not certify statistical validity or equal-cost predictive superiority.

F.2 Repeated-use control

The Full/NoVerifier comparison uses a separate potential-outcome simulator on the same 1,784 held-out IDs. Both arms compute the same post-action diagnostic. Full can block immediate reuse until the configured new-evidence condition is met; NoVerifier logs the status without changing eligibility. Fixed horizon uses two 48-observation rounds after 48 initial observations; adaptive horizon uses 96/144/192 checkpoints on the same IDs. These horizons are overlapping evaluations, not independent replications.

Table 32: **Re-entry control reduces repeated unsupported reuse.** Both arms use the same 1,784 held-out IDs at each horizon. Endpoint metrics change only slightly.

Horizon	Arm	β-$L_1 \downarrow$	F1$\uparrow$	ρ err.$\downarrow$	Cost	Repeated unsupported
Fixed 144	Full	0.1477	0.8552	0.0265	218.3	0.0006
Fixed 144	NoVerifier	0.1481	0.8550	0.0265	218.2	0.0135
Adaptive	Full	0.1488	0.8554	0.0269	220.0	0.0028
Adaptive	NoVerifier	0.1494	0.8551	0.0269	220.0	0.0174

Full uses diagnostic status to gate reuse; NoVerifier is the archived ungated arm. Repeated unsupported is the mean count per unit. Costs use the simulator ledger; the two horizons reuse IDs and are not independent replications.

Repeated unsupported reuse falls from 0.0135 to 0.0006 at fixed horizon and from 0.0174 to 0.0028 under adaptive stopping, with negligible endpoint changes. Diagnostic status is an operational re-entry signal rather than an estimate of true action benefit or a correctness certificate. The simulator permits an all-blocked fallback, so it does not prohibit every unsupported reuse. Detailed status/family audits and matched-injection records remain in the structured supplement.

G External Stress Tests

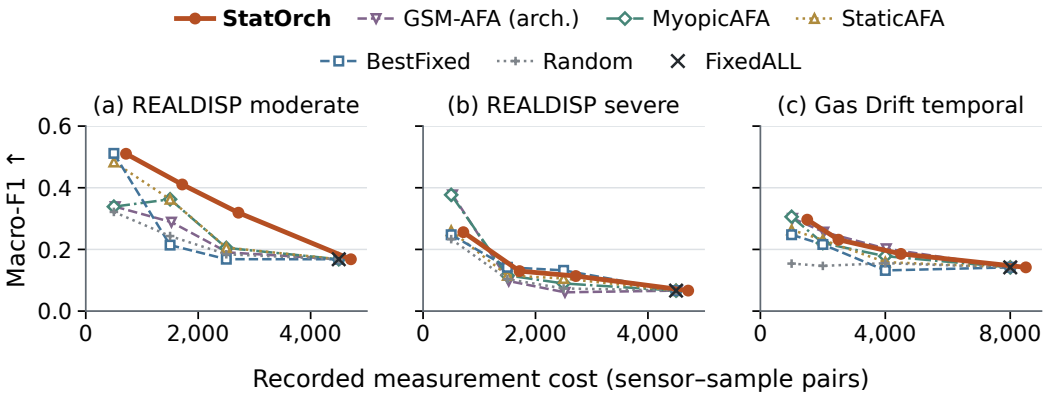


Figure 4: **Real-sensor acquisition has regime-dependent cost–accuracy trade-offs.** All recorded policies share downstream inference. Costs include charged pilots; lines connect observed exact-budget outcomes, not a post-hoc best-so-far envelope. Complete budget, mode, fold, and batch records are in the structured supplement.

Table 33: External stress-test protocol and charged pilots. Evaluation uses 500 observations per episode, separate from target pilots.

Protocol	REALDISP	Gas Drift
Source / calibration	Training-subject ideal placement / self-placement.	B1–B3 / B4–B6.
Held-out environments	Three effective folds; self-placement (moderate), induced displacement (severe).	Four temporal batches B7–B10.
Episodes / mode	20 per fold and severity.	30 per batch.
Prevalence modes	Balanced; Dirichlet $\alpha = 1, 0.2$.	Balanced; Dirichlet $\alpha = 1, 0.2$.
Exact sensor budgets	$\{1, 3, 5, 9\}$.	$\{2, 4, 8, 16\}$.
StatOrch pilot	24 observations on 9 sensors; 216 sensor–sample measurements.	29–32 observations on 16 sensors; mean charge 500.
Archived GSM-AFA pilot	24 observations on 1 sensor; charge 24.	29–32 observations on 1 sensor; mean charge 31.25.
Other policies	No charged target pilot.	No charged target pilot.

REALDISP (Banos et al., 2014) and Gas Drift (Vergara, 2012) are acquisition stress tests with frozen source models, downstream quantification/classification, and calibration-reliability-weighted fusion. The recorded policies are BestFixed, Random, StaticAFA, MyopicAFA, archived GSM-AFA, FixedALL, and StatOrch. Their differences concern acquisition under this shared context. Adding a shifted sensor can worsen fusion; the nonmonotone curves preserve the observed outcomes.

External results are mixed and provide boundary evidence. Moderate REALDISP shows favorable behavior for StatOrch; severe REALDISP favors simpler fixed sensing in some comparisons. On Gas, archived GSM-AFA has stronger prevalence/F1 outcomes in the paired comparison and lower pilot overhead. Thus these records do not establish universal acquisition dominance. The full exact-budget tables, all folds/batches and prevalence modes, hardest-quartile summaries, paired contrasts, and sensor-harm audits remain in the structured supplement without selecting policies or rows by outcome.

The REALDISP subject-to-fold mapping and historical external runner are not recovered. The archived GSM-AFA label identifies a saved policy, not an independently verified reproduction of every original algorithm detail. Released-code BestFixed/MyopicAFA aliases cannot recreate their distinct historical rows. Episode/budget resampling does not create additional independent subjects or temporal environments; this evidence does not validate a real-data full routed chain.

H Reproducibility and Statistical Reporting

H.1 Controlled evidence and uncertainty

Controlled stage and chain evaluations preserve fitted source/router objects, training and validation matrices, frozen selections, observations/truth, and paired task outputs. Both use 300/100 source train/validation samples per class, 32/16/24 target train/validation/test tasks per mechanism, and 260-tree routers with minimum leaf size 4. The stage suite retains one fitted seed (20260926). The short chain is evaluated at seeds 20260926–20260929, using fixed raw data, five-fold task-level out-of-fold upstream choices, ALL observations, and 96/96 fit/audit samples.

The reduced CPU-compatible stage banks have 2/9/6/9 named adaptation/quantification/classification/open-world candidates on 624 held-out tasks, including ACC/BBSE aliases (Appendix A.7). The chain uses a $2 \times 6 \times 6$ named bank on 288 seen and 96 unseen tasks, with 192 validation tasks for complete-tuple selection. Candidate counts refer to named implementations, with aliases retained. The broader historical neural-inclusive suites assess specialization separately.

Per-fit 10,000-resample mechanism-stratified paired intervals quantify held-out task variation conditional on that fit. They are pointwise, without multiplicity correction. Across-fit sample SD instead describes fitting variability on the same raw data; repeated predictions are not additional independent tasks. Table 9 remains the single-fit stage comparison. Four chain fits do not establish robustness across independently sampled cohorts or for the full neural-inclusive bank.

H.2 Matched one-shot selection and fitting-seed sensitivity

The one-shot selector receives the same pre-adaptation state and fitted evidence/diagnosis augmentation as the first stage. Offline, all 72 complete tuples are executed in their own upstream contexts on 384 training tasks. A multioutput ExtraTrees regressor predicts each tuple's terminal joint loss from initial evidence (260 trees, minimum leaf 4, maximum-feature fraction 0.75). One-shot Greedy takes the minimum; One-shot Conservative retains the validation-selected fixed tuple unless the predicted improvement exceeds a margin selected on the same 192 validation tasks. It uses the existing error router's one-standard-error rule and margin grid $\{0, .0025, .005, .01, .02, .03, .05, .08\}$. Both variants are reported without test-based selection.

All policies share source models, observations, candidate bank and offline labels within each fit. One-shot receives complete-tuple terminal-loss supervision; stagewise retains its original local objectives and refreshed contexts. Forest settings match per decision model, but stagewise has three forests and one-shot one. This comparison therefore does not isolate context refresh as a single causal factor or match total computation.

The follow-up protocol and four seeds were fixed before examining the new results, on an already reported test cohort. Original stage predictions are preserved. Each additional seed refits all eight source candidates, source-selected calibration, evidence/diagnosis models, stage forests, OOF upstream choices and validation margins. Every fit selects LDA with Dirichlet calibration; all BestFixed choices remain Raw Classic $\rightarrow$ EMQ $\rightarrow$ adapted ensemble. The recorded runs saved fitted objects and selections before predictions, and all policy predictions before evaluator truth was loaded. Four later serialization repairs are documented in the checkpoint-recovery record, which reports routing-score discrepancies below 10^{-15} and unchanged recorded predictions and metrics.

Table 34: **Joint loss across four fitting seeds.** Raw source and task data are fixed; the three additional seeds refit source candidates, evidence/diagnosis models, stage routers, OOF contexts and validation margins. Every planned fit is reported.

			Stagewise		One-shot	
Split	Fit seed	BestFixed Val	Cons.	Greedy	Cons.	Greedy
Seen	20260926	0.3565	0.3309	**0.3234**	0.3300	0.3237
	20260927	0.3565	0.3309	0.3237	0.3300	**0.3232**
	20260928	0.3565	0.3309	**0.3231**	0.3299	**0.3231**
	20260929	0.3565	0.3309	**0.3232**	0.3300	0.3238
	Mean (SD)	0.3565 (0.00000)	0.3309 (0.00000)	0.3234 (0.00028)	0.3300 (0.00001)	0.3234 (0.00035)
Unseen	20260926	**0.3238**	0.3280	0.3251	**0.3238**	0.3240
	20260927	**0.3238**	0.3280	0.3239	**0.3238**	0.3243
	20260928	**0.3238**	0.3280	0.3247	**0.3238**	0.3243
	20260929	**0.3238**	0.3280	0.3247	**0.3238**	0.3250
	Mean (SD)	0.3238 (0.00000)	0.3280 (0.00000)	0.3246 (0.00050)	0.3238 (0.00000)	0.3244 (0.00041)

SD is the sample standard deviation across four fits, not across tasks. The same 288 seen and 96 unseen tasks are reused; four fits do not multiply the independent test sample size. The source selection is LDA/Dirichlet for every fit.

Conservative's decisions and endpoints are unchanged across the four stagewise fits. The corresponding one-shot means are slightly lower on both splits. Greedy differences between architectures change sign across fits. The paired intervals in Table 35 do not establish stagewise superiority. These observed similarities are not an equivalence-test guarantee.

Table 35: **Paired joint-loss differences by fitted seed.** Negative favors stagewise. C/G denote conservative/greedy variants; no variant is selected using test outcomes.

Split	Fit seed	Stagewise-C − fixed	Stagewise-C − one-shot-C	Stagewise-G − one-shot-G
Seen	20260926	-0.0257 [-0.0311, -0.0200]	+0.0009 [-0.0013, +0.0033]	-0.0003 [-0.0026, +0.0014]
	20260927	-0.0257 [-0.0311, -0.0200]	+0.0009 [-0.0013, +0.0033]	+0.0006 [-0.0005, +0.0017]
	20260928	-0.0257 [-0.0311, -0.0200]	+0.0009 [-0.0013, +0.0033]	+0.0000 [-0.0010, +0.0010]
	20260929	-0.0257 [-0.0311, -0.0200]	+0.0009 [-0.0013, +0.0033]	-0.0006 [-0.0029, +0.0011]
Unseen	20260926	+0.0042 [-0.0005, +0.0098]	+0.0042 [-0.0005, +0.0098]	+0.0011 [-0.0006, +0.0030]
	20260927	+0.0042 [-0.0005, +0.0098]	+0.0042 [-0.0005, +0.0098]	-0.0004 [-0.0038, +0.0022]
	20260928	+0.0042 [-0.0005, +0.0098]	+0.0042 [-0.0005, +0.0098]	+0.0003 [-0.0012, +0.0020]
	20260929	+0.0042 [-0.0005, +0.0098]	+0.0042 [-0.0005, +0.0098]	-0.0003 [-0.0014, +0.0007]

Each entry gives a paired mean difference and its 95% interval from 10,000 mechanism-stratified task bootstrap draws. Intervals condition on the individual fit; they are pointwise, without multiplicity correction. Complete metric contrasts and per-task records are supplied as CSV.

H.3 Original-fit action paths

Table 36 retains the original seed's complete path summary. Conservative selects CORAL on 23/288 seen tasks and retains EMQ and prior-corrected adapted classification with refreshed inputs. Greedy differs from the fixed tuple on 71.9% of seen tasks, with at least two named stages different on 36.8% and all three on 2.8%. Candidate aliases prevent interpreting these counts as independent functional or causal contributions.

Table 36: **Action-path differences from the frozen fixed tuple.** Counts and percentages of tasks with at least one, at least two, or all three named stage actions different from Raw Classic → EMQ → adapted ensemble.

Split	Policy	≥ 1 stage	≥ 2 stages	All 3 stages
Seen	BestFixed-Val	0/288 (0.0%)	0/288 (0.0%)	0/288 (0.0%)
Seen	Conservative	288/288 (100.0%)	23/288 (8.0%)	0/288 (0.0%)
Seen	Greedy	207/288 (71.9%)	106/288 (36.8%)	8/288 (2.8%)
Unseen	BestFixed-Val	0/96 (0.0%)	0/96 (0.0%)	0/96 (0.0%)
Unseen	Conservative	96/96 (100.0%)	0/96 (0.0%)	0/96 (0.0%)
Unseen	Greedy	68/96 (70.8%)	28/96 (29.2%)	0/96 (0.0%)

These are label differences, not counts of distinct functional changes: candidate aliases can produce identical outputs. Conservative always uses EMQ and prior-corrected adapted classification, whose label differs from the fixed classifier; its 23 seen CORAL choices account for the two-stage differences. Paths alone do not isolate each stage's causal contribution.

H.4 Historical evidence and preserved ledgers

Historical resource results use the preserved measurement ledger; the original execution does not support reconstruction of a complete modern end-to-end sensing ledger. The historical runner, fitted controller, full 1,232-unit trajectories, paired-bootstrap inputs, and complete action-to-audit linkage are not recovered. Consequently, the archival 10.5% reduction is a recorded-cost comparison, not a complete sensing or computation-cost claim. A separately released controller cannot retroactively establish the historical execution.

Historical intervals retain their source-reported status. Conflicting historical classification interval summaries lack the paired inputs needed for reconciliation; that suite therefore reports means without a unified interval claim. Historical OOD intervals use 360 eligible units, whereas other primary endpoints use 1,232. Matched backgrounds, reused horizons, and repeated external budgets remain linked observations. These summaries are not confidence sequences under optional stopping.

H.5 What is released and how to locate it

Table 37: Released records and remaining reconstruction limits. Structured supplements preserve the detailed displays removed from the PDF.

Evidence	Released information and remaining limits
Controlled stages / chain	Fitted objects, inputs/truth, candidate matrices, validation selections, paths, and paired outputs. Stage suite: one fit. Short chain and one-shot comparison: four fits on the same raw cohort and reduced bank.
Archival endpoints / stages	Aggregates, scenario effects, choices, and source-reported intervals; the original primary runner, fitted objects, full trajectories, and some historical paired inputs are unavailable.
Re-entry simulator	Potential outcomes, same-ID endpoints, diagnostic records, and implementation; conclusions concern operational reuse in this simulator.
External acquisition	Episode and full frontier records, pilot costs, environment summaries, and audits; original runner/raw caches and REALDISP subject mapping are not recovered.

The supplement preserves the complete records behind the compact displays: complete cost-cap and A0/A1/A2 records, diagnostic audits, complete matched-injection records and original figure, complete exact-budget external rows, and full historical protocol tables. A table-to-file index maps the former displays, including every policy, split, and outcome, to CSV and retained source tables. Display scripts format or aggregate saved records without training or new evaluation. Missing historical observations, model objects, and subject mappings are not inferred or synthesized.