EVIDENCE ACQUISITION AND STOPPING VIA LATTICE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequential evidence acquisition requires deciding what to observe and when to stop from partial information. We introduce a complete-record view of policy supervision: each fully observed training instance defines a lattice of evidence subsets, with a stop/query decision at every node. For a modest number of evidence groups, this structure can be enumerated even when individual groups contain rich representations. Our method, *Subset Lattice Distillation* (SLD), scores subsets with a shared masked predictor, constructs multi-step decision targets by cost-sensitive backward planning, and distills them into an observed-history controller. Uniformly resampled masks train the student across the subset space, while combinatorial planning remains offline. Separating subset scores from acquisition prices allows the predictor to be reused when the policy is replanned for new costs. Our analysis identifies action-margin conditions under which privileged supervision preserves optimal online decisions. On retrospective influenza monitoring, SLD achieves higher phase F1 at lower acquisition cost than both the closest-cost adaptive baselines and the fixed full-evidence model. We further examine geographic–seasonal transfer, price-responsive replanning, and the roles of the model components.

1 INTRODUCTION

Useful evidence is not always available at the start of a prediction task. Additional inputs may improve an assessment, but obtaining them incurs a cost. When inputs are acquired as groups, such as a clinical stream or a contextual report, the relevant choice is which group to request for the current instance and when to stop collecting evidence. These decisions are coupled: acquiring a group changes both the prediction and the value of subsequent acquisitions. A cost-sensitive policy compares stopping with the terminal loss and cumulative cost of continuing.

Retrospective influenza monitoring provides a concrete setting. Wastewater measurements (WW) are available initially; emergency-department visits (ED), hospital admissions, laboratory positivity, and source-context summaries are optional evidence groups. Each state–week record is labelled for epidemic phase and alert status. A fixed WW+ED workflow uses wastewater and emergency-department evidence for every episode, whereas an adaptive policy can acquire different groups or stop with a smaller subset. Each query reveals aligned evidence from the stored record and incurs its assigned acquisition cost.

Active feature acquisition learns such policies through reinforcement learning (Shim et al., 2018), models of partially observed data (Ma et al., 2019; Li & Oliva, 2021), and greedy dynamic feature selection (GDFS) (Covert et al., 2023). Nongreedy oracle supervision, including the acquisition conditioned oracle (ACO), accounts for jointly informative subsets (Valancius et al., 2024). We study the construction of supervision from complete grouped records. An outcome label evaluates a prediction, but does not directly specify how to act under each partial observation. The learning problem is to turn those records into cost-sensitive targets for alternative acquisition paths, while keeping the deployed controller restricted to the evidence it has acquired.

Our key observation is that *rich evidence representations can coexist with a small acquisition space*. With a modest number of groups, observed subsets form an enumerable lattice: each node is an evidence mask, and each edge adds one group. A complete record supplies the values needed to evaluate every node. Backward planning reuses successor values to construct a table of decision

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.

targets. We treat each complete record as a structured collection of acquisition and stopping problems, accessed through uniformly resampled masks during policy learning.

We implement this perspective with *Subset Lattice Distillation* (SLD). A shared masked predictor first scores the evidence subsets of each record. A cost-sensitive teacher then compares stopping at each subset with acquiring a group and continuing through its successors. Finally, a student learns from uniformly resampled masks using the corresponding evidence representation, cost parameters, and stop/query targets. This construction places multi-step planning in the training stage; deployment requires only encoding the observed evidence and applying the controller. Subset scoring is separate from acquisition prices, so the predictor can be reused when new prices call for replanning and redistillation.

The lattice teacher evaluates realized continuations, while the student learns from partial observations. We characterize the conditional preferences induced by soft-target distillation and derive decision-margin conditions for their transfer to online control. The analysis separates changes in action ordering from changes in confidence, and quantifies robustness to student error and approximate planning.

We evaluate SLD alongside fixed evidence workflows and adaptive acquisition methods. It improves phase prediction over ACO and GDFS at lower acquisition cost on the retrospective influenza benchmark. We assess geographic and seasonal transfer using a held-out test set. Price-profile experiments examine replanning with a fixed predictor, while component studies assess sensitivity to representation and policy learning.

Contributions.

- **A complete-record supervision perspective.** We organize each training record into a reusable table of subset scores and multi-step acquisition decisions, separating evidence evaluation from cost-sensitive planning.
- **A lattice-distillation algorithm.** We introduce SLD, combining shared subset evaluation, cost-sensitive backward planning, and observed-history distillation, with a soft-target analysis and sufficient conditions for preserving online decisions.
- **Empirical evaluation.** SLD improves phase prediction over the closest-cost adaptive baselines while incurring lower acquisition cost; holdout, price-profile, and component studies examine transfer and acquisition behavior.

2 Related Work

Cost-aware acquisition methods balance predictive benefit against the expense of obtaining additional evidence, learning when to query and when to stop (Trapeznikov & Saligrama, 2013; Shim et al., 2018). Their central design choice is how to assess unobserved information. EDDI estimates information gain with a partial generative model, while generative surrogate methods learn feature dependencies to support acquisition (Ma et al., 2019; Li & Oliva, 2021). These approaches model missing evidence; SLD instead scores observed realizations in complete training records. GDFS avoids explicit density estimation through a greedy conditional-information objective (Covert et al., 2023), but its one-step criterion can miss benefits that emerge only after further acquisition. Stochastic encodings reason over possible realizations in a learned latent space (Norcliffe et al., 2025); in SLD, a shared representation supplies subset predictions and a separate teacher evaluates multi-step continuations. Our complete-record perspective exploits a modest number of evidence groups to turn an enumerable table of subset scores into stop/query supervision, without fitting a generative model of missing groups or restricting the teacher to immediate predictive gain.

A complementary question is how training-time information becomes supervision for a deployable policy. Asymmetric imitation shows that a fully informed teacher can induce unsuitable actions under partial observation (Warrington et al., 2021); cost-sensitive imitation relates action errors to decision cost (Ross & Bagnell, 2014). Direct value learning instead regresses continuation costs, as in BRiG-AFA (Feng et al., 2026). Within oracle-based acquisition, ACO evaluates nongreedy candidate subsets conditional on the current observations and learns its policy by behavioral cloning of oracle rollouts(Valancius et al., 2024). SLD instead constructs a realized target lattice for each complete

record and trains on uniformly resampled masks. The lattice organizes multi-step preferences independently of a particular rollout, while sampling provides stochastic updates from the full subset space.

3 Problem Formulation

We study *cost-sensitive grouped evidence acquisition*: predicting a target while deciding which additional evidence to obtain and when to stop. Each episode provides mandatory evidence $X^{(0)}$, optional groups $\{X^{(m)} : m \in \mathcal{M}\}$, and a target Y, where $\mathcal{M}$ indexes the $K = |\mathcal{M}|$ optional groups. After acquiring $S \subseteq \mathcal{M}$, the observed history H_S contains the mandatory input, acquired group values, their binary mask, and available metadata. A policy π uses H_S to query an unobserved group $m \in \mathcal{M} \setminus S$ at cost $c_m \geq 0$, or to stop and invoke a terminal prediction rule. Each group is acquired at most once. The objective is to minimize expected terminal loss plus accumulated acquisition cost. Training uses complete labeled records; deployment reveals only queried values and never the target. We model each query as revealing stored evidence within an episode, not as waiting for new measurements over calendar time. Optional deferral is a terminal-output choice after acquisition stops.

4 Subset Lattice Distillation

Overview. A complete record can supervise decisions at multiple evidence subsets, so SLD moves the combinatorial comparison of acquisition paths into training. The method has three components: *shared subset evaluation* turns partial observations into prediction scores; a *cost-sensitive lattice teacher* converts those scores into multi-step stop/query targets; and *observed-history distillation* learns a deployable controller from the targets. Figure 1 shows this flow and the separation between complete training records and partial observations at deployment. Subset evaluation is performed before prices enter the planner, allowing the predictor to be reused under new acquisition costs.

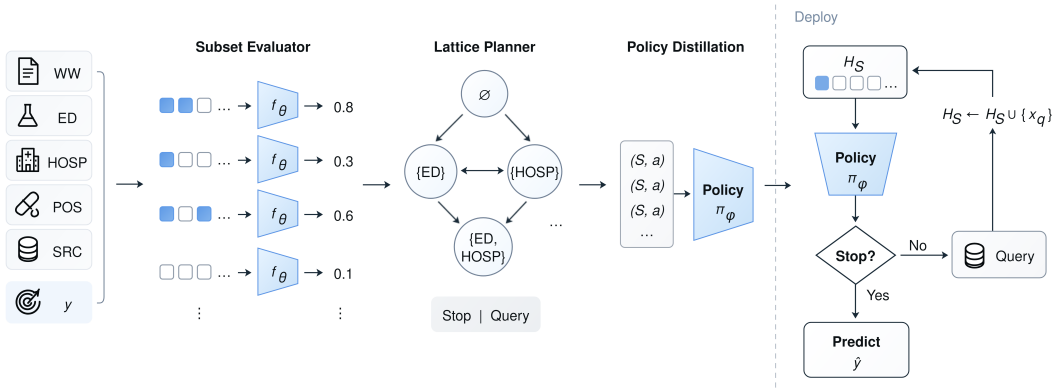


Figure 1: **From complete records to acquisition policies.** Shared subset evaluation produces one score table per training record. A cost-sensitive lattice teacher uses backward planning to assign stop/query targets, and uniformly resampled masks train an observed-history controller. Online, a query reveals one group and updates the history; a stop invokes the terminal rule. Labels and unqueried values enter offline target construction, not the deployed controller's inputs.

4.1 Shared Subset Evaluation

Comparing acquisition paths requires predictions from different evidence subsets. A single masked predictor shares statistical strength across these subsets and avoids fitting a separate model for each mask. The input retains observed blocks, zero-fills unobserved numeric and text blocks, and includes a binary observation mask so that missing blocks remain distinguishable from measured zeros. The encoder and decoder are

$$q_\theta(z \mid H_S) = \mathcal{N}\big(\mu_S, \mathrm{diag}(\sigma_S^2)\big), \quad f_\psi(H_S) = \mathrm{decode}_\psi\big([\mu_S; \log \sigma_S^2]\big). \tag{1}$$

Here $z \in \mathbb{R}^d$ is a latent vector of dimension d; $\mu_S \in \mathbb{R}^d$ and $\sigma_S^2 \in \mathbb{R}^d_{>0}$ are its mean and variance. The operator diag forms a diagonal covariance matrix, brackets with a semicolon denote concatenation, and the logarithm is elementwise. The parameters θ belong to the encoder and ψ to the decoder. We write $f(H)$ for the prediction produced by this model at history H. The Gaussian state summarizes the observed evidence. The decoder adopts natural-parameter networks (NPN) (Wang et al., 2016), propagating moments deterministically to the task outputs.

To train one predictor across masks, let t index a complete training record and $H_{t,S}$ its history under subset S. In the influenza task, $Y_t = (y_t, b_t)$ consists of a four-class phase label y_t and a binary alert label b_t. The decoder defines their predictive distribution $p_\psi(y_t, b_t \mid H_{t,S})$ and is trained jointly with the encoder using

$$\mathcal{L}_{\mathrm{state}} = \mathbb{E}_{t,S}\big[-\log p_\psi(y_t, b_t \mid H_{t,S}) + \beta D_{\mathrm{KL}}(q_\theta(\cdot \mid H_{t,S}) \| q_\theta(\cdot \mid H_{t,\varnothing}))\big]. \tag{2}$$

The expectation averages training records and sampled subsets; $\beta \geq 0$ weights the Kullback–Leibler divergence D_{KL}. The empty subset $\varnothing$ retains only mandatory evidence, providing an instance-specific reference for the regularizer. Task likelihood learns predictive content, while the regularizer discourages unconstrained changes between subset representations. A frozen text encoder supplies the representations of observed summaries.

The planner uses a scalar terminal score for each subset. After freezing θ and ψ, a fixed scoring procedure produces a realized terminal score $\rho_t(S)$ from each complete record. With compulsory prediction, this can be a prediction-loss score. With an optional reject decision (Geifman & El-Yaniv, 2019), a generic construction uses a calibrated realized prediction score $\widetilde{r}_t(S)$ and a deferral cost $\tau > 0$:

$$\rho_t(S) = \min\{\tau, \widetilde{r}_t(S)\}, \qquad d_t(S) = \mathbf{1}\{\widetilde{r}_t(S) < \tau\}. \tag{3}$$

The terminal target $d_t(S) = 1$ indicates prediction after stopping; $d_t(S) = 0$ indicates deferral. These are realized supervision targets, not conditional expected losses. In our experiments, we use the task-specific terminal scoring and output procedure described in Appendix D; Equation (3) specifies the generic optional-deferral interface. The resulting table $\{\rho_t(S) : S \subseteq \mathcal{M}\}$ is the input to the lattice teacher.

4.2 Cost-sensitive lattice teacher

A subset score alone does not indicate whether another group is worth acquiring: a useful group may enable further acquisitions before prediction improves. The teacher therefore compares stopping with complete continuations under one loss–cost objective. For a policy π including its terminal rule, let T be its stopping time, S_T its final acquired subset, and L_T its terminal loss. The objective is

$$J_\lambda(\pi) = \mathbb{E}_\pi\left[L_T + \lambda \sum_{m \in S_T} c_m\right], \tag{4}$$

where $\lambda \geq 0$ converts acquisition costs to the terminal-loss scale. The expectation includes the record distribution and any policy randomization. A prediction incurs $L_T = \ell(f(H_{S_T}), Y)$ for a nonnegative integrable task loss ℓ; optional deferral incurs $L_T = \tau$. Thus stopping is compared with the loss and total cost of continuing, not with the next query's predictive gain alone.

For each record, the 2^K subsets form an acyclic lattice whose edges add one group. A complete record supplies a realized successor for every edge. Starting from the full subset, where no query remains, the teacher computes

$$\widetilde{V}_{\lambda,t}(S) = \min\left\{\rho_t(S), \min_{m \in \mathcal{M} \setminus S}\left[\lambda c_m + \widetilde{V}_{\lambda,t}(S \cup \{m\})\right]\right\}, \tag{5}$$

with boundary $\widetilde{V}_{\lambda,t}(\mathcal{M}) = \rho_t(\mathcal{M})$. Here $\widetilde{V}_{\lambda,t}(S)$ is the minimum remaining cost for record t at mask S. The first term stops immediately; each query term pays for m and uses the already-computed successor value. Processing subsets in decreasing size reuses these successor values across paths. The recurrence is exact for the realized record and its terminal scores. Its outputs provide cost-sensitive preferences; Section 5 analyzes how learning them from partial observations relates to online decision cost.

To obtain supervision for the controller, write $\mathcal{A}_S^{\mathrm{acq}} = \{\mathsf{stop}\} \cup (\mathcal{M} \setminus S)$ for the valid actions. Define $\widetilde{Q}_{\lambda,t}(\mathsf{stop}, S) = \rho_t(S)$ and $\widetilde{Q}_{\lambda,t}(m, S) = \lambda c_m + \widetilde{V}_{\lambda,t}(S \cup \{m\})$ for a query. Near-tied action costs can make hard targets unstable, so we use the smoothed distribution

$$p_T(a \mid t, S, \lambda) = \frac{\exp[-\widetilde{Q}_{\lambda,t}(a, S)/T_Q]}{\sum_{a' \in \mathcal{A}_S^{\mathrm{acq}}} \exp[-\widetilde{Q}_{\lambda,t}(a', S)/T_Q]}. \tag{6}$$

Here a is an available action, a' ranges over alternatives, and $T_Q > 0$ is the target temperature, distinct from the stopping time T. The subscript T on p_T denotes the teacher. Soft targets assign similar probabilities to actions with similar teacher costs-to-go; the planning recurrence retains the hard minimum.

The target table spans every mask of each complete record. Student training accesses this table through repeated uniform mask sampling. Because subset scores $\rho_t(S)$ are separate from prices $c = (c_m)_{m \in \mathcal{M}}$, changing prices requires replanning and redistillation while reusing the predictor and score table when the terminal procedure is fixed.

4.3 Observed-history distillation and deployment

The student learns the teacher's action targets from the corresponding partial history. Let $s(H_S)$ concatenate the encoder moments, evidence mask, available metadata, and cost features. A policy $\pi_\eta(a \mid s(H_S), \lambda)$, with parameters η, predicts a distribution over $\mathcal{A}_S^{\mathrm{acq}}$. When deferral is enabled, a terminal head $h_\omega(s(H_S)) \in (0, 1)$, with parameters ω, predicts the answer indicator $d_t(S)$. We train them with

$$\begin{aligned}\mathcal{L}_{\mathrm{distill}} = \mathbb{E}_{t,S,\lambda}\Big[&- \sum_{a \in \mathcal{A}_S^{\mathrm{acq}}} p_T(a \mid t, S, \lambda) \log \pi_\eta(a \mid s(H_{t,S}), \lambda) \\ &+ \gamma \,\mathrm{BCE}(d_t(S), h_\omega(s(H_{t,S})))\Big].\end{aligned} \tag{7}$$

The first term is policy cross-entropy; the second is binary cross-entropy (BCE) for the terminal output, weighted by $\gamma \geq 0$. Masks are repeatedly sampled from $w = \mathrm{Unif}(2^{\mathcal{M}})$, where $2^{\mathcal{M}}$ is the set of all subsets and $w_S = 2^{-K}$. Each minibatch combines records, fresh masks, and training multipliers. Uniform resampling is an unbiased stochastic estimate of the full-mask average; individual updates need not traverse the lattice. Compulsory prediction omits the terminal head. Labels and complete evidence supply targets only.

At deployment, $S = \varnothing$ and only mandatory evidence is visible. Each query reveals one group and updates H_S; the next controller evaluation can select another query or stop. Repeated or unavailable queries are masked, and the full subset terminates. A stop invokes the terminal procedure: in the generic interface, prediction is issued when $h_\omega(s(H_S))$ exceeds a development-selected threshold, and the system otherwise defers, while the experiments use the procedure in Appendix D. Algorithm 1 summarizes training and deployment.

Computational scope and scaling. Let $N = |\mathcal{D}_{\mathrm{tr}}|$ and $|\Lambda|$ be the number of training multipliers. Exact target construction uses $N2^K$ subset evaluations and $O(N|\Lambda|K2^K)$ scalar planning operations. Here K counts acquisition *groups*, not the dimensionality within a group; the present benchmark has four groups and therefore 16 subset states. Student updates evaluate resampled masks, and deployment requires at most K queries and $K + 1$ controller evaluations with no lattice search.

Extending the offline planner to a sparse subset DAG is straightforward: the same backward recurrence is applied only to retained nodes and edges. For a sampled supervision mask, retain every feasible first query and use a budgeted search for its continuations. Once the graph is constructed, evaluation and planning scale with its nodes and edges, rather than 2^K. The student and deployment interface are unchanged. Appendix A.5 gives the construction cost and a theorem relating pruning error to decision stability.

Algorithm 1 SLD: complete-record supervision and observed-history control

Require: Complete training records $\mathcal{D}_{\mathrm{tr}}$, disjoint development records $\mathcal{D}_{\mathrm{dev}}$, prices c, multiplier grid Λ, terminal scoring procedure.
1: Fit the masked predictor using Equation (2); freeze its parameters.
2: Score all training record–mask pairs: $\rho_t(S)$ and, when used, $d_t(S)$.
3: **for** $\lambda \in \Lambda$ and each training record t **do**
4: Solve Equation (5) in decreasing subset size.
5: Store the stop/query targets in Equation (6) at every mask.
6: **end for**
7: **for** each student update **do**
8: Sample records and multipliers; redraw $S \sim \mathrm{Unif}(2^{\mathcal{M}})$.
9: Update the student and optional terminal head with Equation (7).
10: **end for**
11: Fix operating parameters on $\mathcal{D}_{\mathrm{dev}}$; report realized student cost.
Online: initialize $S = \varnothing$ with mandatory evidence observed.
12: **while** the controller selects an available unobserved group m **do**
13: Pay c_m, reveal the group, and update H_S and S.
14: **end while**
15: Stop and invoke the terminal rule on H_S.

5 From Lattice Supervision to Online Decisions

Fix the fitted functions and costs. Let $W \sim P$ be a fresh complete record, $S \sim \mathrm{Unif}(2^{\mathcal{M}})$ an independent mask, $\nu = P \times w$, and $U = s(H_S)$ the student input, including the mask. We analyze the preferences learned by distillation and when their ordering preserves online decisions.

5.1 What the student learns

Proposition 1 (Conditional preference projection)**.** *The policy cross-entropy over unrestricted observed-input distributions is minimized by*

$$\bar{p}(a \mid U) = \mathbb{E}_\nu[p_T(a \mid W, S) \mid U]. \tag{8}$$

The projection follows asymmetric imitation (Warrington et al., 2021). Its KL decomposition appears in Appendix A.2; we focus here on action ordering.

5.2 Decision stability under privileged supervision

Fix the deployed predictor and terminal rule g. Let $Q_g^*(H, a)$ be the expected cost of action a followed by optimal observed-history acquisition using the same g. Write $\bar{q}(U) = \mathbb{E}_\nu[\widetilde{Q}_W(\cdot, S) \mid U]$ and $\mathrm{osc}(x) = \max_a x_a - \min_a x_a$. Define

$$\begin{aligned} b(H) &= \mathrm{osc}\big(\bar{q}(U) - Q_g^*(H, \cdot)\big), \quad e(U) = \|\pi_\eta(\cdot \mid U) - \bar{p}(\cdot \mid U)\|_\infty, \\ v^2(U) &= \mathbb{E}_\nu\left[\mathrm{osc}\Big(\widetilde{Q}_W - \bar{q}(U)\Big)^2 \mid U\right]. \end{aligned} \tag{9}$$

Here b measures action-relative mean distortion, including terminal-score mismatch and representation compression; v^2 measures fluctuations of cost differences, and e is student error. At a history with $n > 1$ feasible actions, let $\Delta(H)$ be the gap between the best and second-best online action costs.

Theorem 2 (Decision-margin transfer)**.** *Assume the static evidence setting, integrable action costs and finite v^2. If $\Delta(H) > b(H)$ and*

$$M_T(H) := \frac{1 - e^{-(\Delta(H) - b(H))/T_Q}}{n} - \frac{v^2(U)}{2T_Q^2} > 2e(U), \tag{10}$$

argmax student decoding selects the unique online-optimal action at H. If these conditions hold at every feasible multi-action history, the student is optimal for the fixed-output loss–cost objective.

From action margins to policy cost. Let $J^{\star}_{\lambda,g}$ be the minimum expected loss–cost among observed-history policies using the same predictor and terminal rule g. Let $\mathcal{B}_\pi$ contain multi-action histories where $\Delta \le b$ or $M_T \le 2e$, including optimal-action ties, and set $B_g(H) = \mathrm{osc}(Q^*_g(H,\cdot))$. The argmax student's excess cost satisfies

$$J_{\lambda,g}(\widehat{\pi}) - J^{\star}_{\lambda,g} \le \mathbb{E}_{\widehat{\pi}} \sum_{t=0}^{T} B_g(H_t)\mathbf{1}\{H_t \in \mathcal{B}_\pi\}. \tag{11}$$

Here T indexes the final decision, $A_T = \mathsf{stop}$, and the expectation is over records and student paths. Single-action histories have $B_g(H) = 0$. The bound accumulates cost only where the margin condition is not satisfied. If $B_g(H) \le B_*$, it is at most B_* times the expected number of such decisions. Appendix A.3 gives the proof, and Appendix A.5 treats pruned teachers.

6 Experiments

We evaluate the prediction–cost trade-off, the roles of the model components, replanning under new evidence prices, and geographic–seasonal transfer.

6.1 Experimental setup

Task and evidence. The retrospective influenza benchmark contains 5,933 state–week prediction episodes. Each record has an epidemic-phase label and a binary alert label. Wastewater measurements (WW) provide the initial evidence; queries reveal emergency-department visits (ED), hospital admissions (HOSP), laboratory positivity (POS), or source-context records (SRC). Their default costs are $(1.0, 1.5, 1.2, 0.4)$, respectively, so one ED query defines one cost unit and the full optional bundle costs 4.1. These are benchmark acquisition costs, not monetary prices. All queried evidence is restricted to information available by the corresponding state–week cutoff, and train/development/test partitions are chronological. We fit the predictor and controller on the training split and select calibration parameters and operating points on the development split.

Baselines. *XGBoost (WW)* uses boosted trees (Chen & Guestrin, 2016), while *SelectiveNet (WW)* jointly models prediction and rejection (Geifman & El-Yaniv, 2019); both use only the initial wastewater evidence. The multilayer perceptron (MLP) references *MLP (WW+SRC)* and *MLP (WW+ED)* add source context and emergency-department evidence, respectively. *MLP (All)* uses WW and all optional groups.

For adaptive acquisition, *EDDI* estimates expected information gain with a partial variational autoencoder (Ma et al., 2019); *GDFS* learns a greedy conditional-information selector (Covert et al., 2023); and *ACO* uses a nongreedy acquisition-conditioned oracle (Valancius et al., 2024). SLD, ACO, and GDFS share the same trained backbone, prediction head, evidence representations, data splits, and acquisition costs. The remaining controls are *ED-First*, a fixed acquisition workflow beginning with ED; *ReAct Router*, the benchmark's ReAct-style evidence-routing baseline (Yao et al., 2023); and *Supervised Router*, which learns routing actions through supervised training.

Implementation. The shared predictor has a 16-dimensional Gaussian state, a width-128 residual encoder (two blocks, layer normalization, dropout 0.10), and an NPN decoder. The student uses a 60-dimensional input and a width-128 MLP with one residual block and dropout 0.10. Both train for 1,000 Adam steps, with batch size 64 and learning rate 10^{-3}. Predictor training uses $\beta = 0.02$ and empty/partial/full mask probabilities $0.10/0.30/0.60$; student masks are uniform over subsets. Further settings appear in Appendix D.

Evaluation. Phase macro-F1 and realized mean acquisition cost are the primary endpoints; expected calibration error (ECE) is complementary. Alert AUROC and Brier scores appear in Appendix C; Appendix D gives the task-specific scoring and training protocol.

6.2 PREDICTION QUALITY AND ACQUISITION COST

The first question is whether the complete pipeline improves prediction without simply collecting more evidence. Table 1 compares prediction quality and realized acquisition cost across no-query, fixed-workflow, and adaptive methods.

Table 1: **Prediction and evidence cost under the default price profile.** We report the mean and standard deviation over three seeds for realized acquisition cost, phase macro-F1, and ECE. MLP (WW+ED) and MLP (All) provide fixed-input references for the adaptive methods.

System	Mean cost ↓	Phase F1 ↑	ECE ↓
XGBoost (WW)	$0.000_{\pm 0.000}$	$0.725_{\pm 0.001}$	$0.031_{\pm 0.004}$
SelectiveNet (WW)	$0.000_{\pm 0.000}$	$0.623_{\pm 0.008}$	$0.197_{\pm 0.042}$
MLP (WW+SRC)	$0.400_{\pm 0.000}$	$0.671_{\pm 0.004}$	$0.025_{\pm 0.001}$
ReAct Router	$0.472_{\pm 0.031}$	$0.622_{\pm 0.004}$	$0.116_{\pm 0.002}$
MLP (WW+ED)	$1.000_{\pm 0.000}$	$0.964_{\pm 0.003}$	$0.073_{\pm 0.010}$
EDDI	$1.761_{\pm 0.071}$	$0.814_{\pm 0.031}$	$0.138_{\pm 0.026}$
GDFS	$2.187_{\pm 0.046}$	$0.873_{\pm 0.011}$	$0.130_{\pm 0.034}$
ACO	$2.357_{\pm 0.099}$	$0.879_{\pm 0.006}$	$0.131_{\pm 0.036}$
ED-First	$2.500_{\pm 0.000}$	$0.831_{\pm 0.031}$	$0.122_{\pm 0.018}$
Supervised Router	$3.017_{\pm 0.262}$	$0.866_{\pm 0.014}$	$0.126_{\pm 0.030}$
MLP (All)	$4.100_{\pm 0.000}$	$0.936_{\pm 0.004}$	$0.046_{\pm 0.006}$
SLD (ours)	$\mathbf{2.179}_{\pm 0.027}$	$\mathbf{0.983}_{\pm 0.004}$	$\mathbf{0.014}_{\pm 0.001}$

Comparison with adaptive acquisition. SLD achieves phase F1 0.983 at mean cost 2.179. Compared with ACO, F1 increases by 0.104 while cost decreases by 7.6%. GDFS likewise has lower F1 (0.873) at higher cost (2.187). SLD also has lower ECE than ACO and GDFS. Relative to the fixed full-evidence MLP (All), SLD improves phase F1 by 0.047 while reducing mean acquisition cost by 46.9%. Thus, its phase-prediction gain is not explained by greater evidence consumption: it achieves higher F1 while acquiring substantially less evidence on average.

Comparison with a strong fixed workflow. MLP (WW+ED) reaches F1 0.964 at cost 1.000, providing a strong low-cost alternative. SLD adds 0.019 F1 at 1.179 additional cost; MLP (WW+ED) also has better alert AUROC/Brier ($0.975/0.059$ versus $0.966/0.068$). SLD therefore offers higher phase F1, while MLP (WW+ED) remains a strong low-cost alert predictor.

6.3 ABLATIONS AND SUPERVISION COMPARISON

The design separates subset evaluation, cost-sensitive planning, and online control. Table 2 compares component removals, cost scaling, and oracle-trajectory supervision. We examine both prediction and acquisition cost, since each intervention can change the evidence used.

Table 2: **Ablations and supervision comparison.**

Variant / method	Mean cost ↓	Phase F1 ↑	ECE ↓
SLD (ours)	**2.179**	**0.983**	**0.014**
W/O probabilistic state	4.100	0.930	0.040
W/O frozen text adapter	2.263	0.963	0.023
W/O policy network	0.000	0.670	0.108
Raw unaligned units	0.000	0.150	0.032
Oracle trajectories (ACO)	2.357	0.879	0.131

Subset representation. The probabilistic state supports both subset prediction and policy inputs. Removing it increases mean cost to 4.100, the full-bundle cost, while F1 falls to 0.930. The full model predicts more accurately with less evidence, supporting the state representation's role in the acquisition pipeline.

The frozen text adapter encodes contextual evidence. Removing it lowers F1 from 0.983 to 0.963 and raises cost from 2.179 to 2.263. Replacing E5-Mistral with gte-Qwen2 yields F1 0.972 at cost 2.134 (Table 6). Both text encoders achieve higher phase F1 and lower mean acquisition cost than the variant without the text adapter.

Cost calibration and online control. The teacher compares stopping loss with continuation cost. Raw, unaligned units produce a no-query policy with F1 0.150. Removing the policy network also leaves a no-query system, with F1 0.670 at zero cost.

Oracle trajectories and uniform mask supervision. ACO's behavioral-cloning formulation learns from oracle rollouts (Valancius et al., 2024). SLD instead repeatedly samples masks uniformly from each record's target lattice. Compared with ACO, SLD has higher F1 (0.983 versus 0.879), lower cost (2.179 versus 2.357), and lower ECE (0.014 versus 0.131).

6.4 Replanning under new costs with a fixed predictor

For each price profile, we keep the predictor and terminal procedure fixed, recalibrate the acquisition-cost multiplier on the development set, regenerate the teacher targets, and redistill the student. We do not use test labels for any of these steps. Six profiles vary ED prices; three alter HOSP or SRC costs or swap the full cost ordering.

Across the nine nondefault profiles, 25.7%–63.6% of actions change while phase F1 remains in $[0.981, 0.987]$ (Appendix C.1). The full cost-rank swap changes 36.2% of actions with F1 0.985. These results support price-responsive replanning while preserving prediction quality across the tested costs, without refitting the predictor.

6.5 Transfer across geography and season

The price study changes the acquisition objective while retaining the prediction model. We next ask whether the complete system remains useful when the evaluation population changes. HHS Region 3 is excluded from model fitting, threshold selection, cost calibration, and operating-point selection. We evaluate the frozen systems on 305 episodes from the held-out 2024/25 season.

Table 3: **Frozen geographic and seasonal transfer.** Results on HHS Region 3 in 2024/25, excluded from fitting and model selection.

System	Mean cost ↓	Phase F1 ↑
ACO	2.975	0.928
MLP (All)	4.100	0.914
SLD (ours)	**2.700**	**0.958**

SLD reaches F1 0.958 at cost 2.700, improving F1 by 0.030 over ACO while using 0.275 less cost (Table 3). Against MLP (All), the corresponding differences are 0.044 higher F1 and 1.400 lower cost. The higher-F1/lower-cost pattern relative to the full-evidence reference therefore also persists under the tested geographic and seasonal shift.

7 Conclusion

SLD converts complete records into acquisition policies through shared subset evaluation, cost-sensitive lattice planning, and observed-history distillation. Uniform mask sampling trains the controller across partial observations, with combinatorial planning confined to training. Reusable subset scores support replanning under new prices. Our analysis characterizes online action-order preservation and sensitivity to student approximation.

On retrospective influenza monitoring, SLD achieves higher phase F1 at lower acquisition cost than adaptive baselines. The geographic and seasonal holdout also shows higher F1 and lower cost than ACO and MLP (All). Sparse subset planning extends the offline construction to larger acquisition spaces, with pruning error linked to decision stability.

Reproducibility Statement

The appendix supplies proofs, implementation details, and additional experimental results. Code will accompany the submission.

AI Use Statement

Generative AI tools were used for assistance with manuscript editing and code-related tasks. All scientific contributions, experimental design, analysis, and conclusions were developed and verified by the authors.

References

Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. Association for Computing Machinery, 2016. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Ian Covert, Wei Qiu, Mingyu Lu, Nayoon Kim, Nathan White, and Su-In Lee. Learning to Maximize Mutual Information for Dynamic Feature Selection. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6424–6447. PMLR, 2023. URL https://proceedings.mlr.press/v202/covert23a.html.

Jiaorong Feng, Qian Li, and Ying Li. BRiG-AFA: Bellman Risk-to-Go Learning for Non-Myopic Active Feature Acquisition. arXiv preprint arXiv:2608.02305, 2026. URL https://arxiv.org/abs/2608.02305.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Yang Li and Junier B. Oliva. Active Feature Acquisition with Generative Surrogate Models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6450–6459. PMLR, 2021. URL https://proceedings.mlr.press/v139/li21p.html.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient Dynamic Discovery of High-Value Information with Partial VAE. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4234–4243. PMLR, 2019. URL https://proceedings.mlr.press/v97/ma19c.html.

Alexander Norcliffe, Changhee Lee, Fergus Imrie, Mihaela van der Schaar, and Pietro Liò. 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. PMLR, 2025. URL https://proceedings.mlr.press/v267/norcliffe25a.html.

Stéphane Ross and J. Andrew Bagnell. Reinforcement and Imitation Learning via Interactive No-Regret Learning. arXiv preprint arXiv:1406.5979, 2014. URL https://arxiv.org/abs/1406.5979.

Hajin Shim, Sung Ju Hwang, and Eunho Yang. Joint Active Feature Acquisition and Classification with Variable-Size Set Encoding. In *Advances in Neural Information Processing Systems*, volume 31, pp. 1368–1378, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/e5841df2166dd424a57127423d276bbe-Abstract.html.

Kirill Trapeznikov and Venkatesh Saligrama. Supervised Sequential Classification Under Budget Constraints. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pp. 581–589. PMLR, 2013. URL https://proceedings.mlr.press/v31/trapeznikov13a.html.

Michael Valancius, Maxwell Lennon, and Junier Oliva. Acquisition Conditioned Oracle for Non-greedy Active Feature Acquisition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48957–48975. PMLR, 2024. URL https://proceedings.mlr.press/v235/valancius24a.html.

Hao Wang, Xingjian Shi, and Dit-Yan Yeung. Natural-Parameter Networks: A Class of Probabilistic Neural Networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/fe9fc289c3ff0af142b6d3bead98a923-Abstract.html.

Andrew Warrington, J. Wilder Lavington, Adam Ścibior, Mark Schmidt, and Frank Wood. Robust Asymmetric Learning in POMDPs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11013–11023. PMLR, 2021. URL https://proceedings.mlr.press/v139/warrington21a.html.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629v3.

A SOFT-TARGET DISTILLATION: ANALYSIS AND PROOFS

This appendix analyzes the soft action targets and policy cross-entropy of SLD. Conditional averaging is a standard property of asymmetric imitation (Warrington et al., 2021); the pathwise cost argument is related to cost-sensitive imitation analysis (Ross & Bagnell, 2014). We specialize these tools to an exhaustive, static evidence lattice with outcome-dependent terminal scores.

A.1 POPULATION SETTING AND COMPARISON CLASS

Condition on all fitted functions: the predictor, score map, representation, student, and deployed terminal rule g. Fix prices $c = (c_m)_{m\in\mathcal{M}}$, the multiplier λ, and target temperature $T_Q > 0$. A fresh complete record W has distribution P and contains the mandatory evidence, optional group values, metadata, and outcome. It is independent of the randomness used to fit these functions. All expectations and conditional distributions below refer to this same P. The analysis is conditional on the fitted functions and concerns their performance under P.

For each mask $S \subseteq \mathcal{M}$, let $H_S(W)$ be the corresponding observed history and let $U_S = s(H_S)$ be the actual student input. It includes S and any availability information needed to determine the action set. Each group is queried at most once, queries only reveal stored values, and acquisition does not change the underlying record. The analyzed action set is $\mathcal{A}_S^{\mathrm{acq}} = \{\mathsf{stop}\} \cup (\mathcal{M} \setminus S)$. A fixed observed restriction may reduce it, provided the teacher, student, and comparison policies use the same feasible actions. The full mask always stops.

Independently of W, sample $S \sim w = \mathrm{Unif}(2^{\mathcal{M}})$, matching the repeated uniform mask sampling used for student optimization. Thus $w_S = 2^{-K}$, $\nu = P \times w$, and $w_{\min} = 2^{-K}$. Uniformity is over individual subsets, not over subset sizes. The proof below is written for general positive weights w_S, with the uniform scheme as the implementation used here.

The fixed rule g is measurable with respect to the observed history. It specifies prediction and, when used, whether to defer. Write $L_g(W, S)$ for its realized loss; independent terminal randomization can be integrated into this loss. Define

$$J_{\lambda,g}(\pi) = \mathbb{E}\left[L_g(W, S_T) + \lambda \sum_{m\in S_T} c_m\right], \qquad J^\star_{\lambda,g} = \inf_{\pi\in\Pi_{\mathrm{obs}}} J_{\lambda,g}(\pi), \tag{12}$$

where Π_{obs} consists of nonanticipative acquisition policies using the same predictor, terminal rule, costs, and feasible actions. These policies may use the full observed history, not only its compressed representation. They do not observe the outcome or unqueried values. Thus the comparator can choose a better acquisition policy but cannot replace the terminal predictor or rejection procedure.

Let $\rho_W(S)$ be the realized score supplied to the teacher and define

$$J_\rho(\pi) = \mathbb{E}\left[\rho_W(S_T) + \lambda \sum_{m\in S_T} c_m\right]. \tag{13}$$

Losses and scores are integrable. For notational brevity, suppress the fixed λ in $\widetilde{V}_W(S)$ and $\widetilde{Q}_W(a, S)$. These are precisely the realized teacher values from Section 4.2.

A.2 PROOF OF THE CONDITIONAL PROJECTION

For any student with finite expected divergence, define

$$\mathcal{E}_{\mathrm{priv}} = \mathbb{E}_\nu D_{\mathrm{KL}}(p_T \| \bar{p}), \qquad \mathcal{E}_{\mathrm{fit}} = \mathbb{E}_\nu D_{\mathrm{KL}}(\bar{p} \| \pi_\eta). \tag{14}$$

Then $\mathbb{E}_\nu D_{\mathrm{KL}}(p_T \| \pi_\eta) = \mathcal{E}_{\mathrm{priv}} + \mathcal{E}_{\mathrm{fit}}$.

For each feasible action, define $p_W(a, S) = p_T(a \mid W, S)$ and $\bar{p}(a \mid U) = \mathbb{E}_\nu[p_W(a, S) \mid U]$. Since the mask is included in U, all distributions in a conditional average have the same action set. For any observed-input distribution π,

$$\begin{aligned}\mathbb{E}_\nu\left[-\sum_a p_W(a, S) \log \pi(a \mid U)\right] &= \mathbb{E}_\nu\left[-\sum_a \bar{p}(a \mid U) \log \pi(a \mid U)\right] \\ &= \mathbb{E}_\nu\left[\mathsf{H}(\bar{p}(\cdot \mid U)) + D_{\mathrm{KL}}(\bar{p} \| \pi)\right],\end{aligned} \tag{15}$$

where H is Shannon entropy with natural logarithms. Nonnegativity of KL proves Equation (8). The minimizer is unique up to null sets over unrestricted distributions. A finite network or shared training with the terminal head need not attain it.

For finite expected divergences, add and subtract $\log \bar{p}$ inside the KL and condition its second term on U:

$$\begin{aligned}\mathbb{E}_\nu D_{\mathrm{KL}}(p_W \| \pi) &= \mathbb{E}_\nu D_{\mathrm{KL}}(p_W \| \bar{p}) + \mathbb{E}_\nu \sum_a p_W(a, S) \log \frac{\bar{p}(a \mid U)}{\pi(a \mid U)} \\ &= \mathbb{E}_\nu D_{\mathrm{KL}}(p_W \| \bar{p}) + \mathbb{E}_\nu D_{\mathrm{KL}}(\bar{p} \| \pi). \end{aligned} \tag{16}$$

This proves Proposition 1. With finite teacher costs and $T_Q > 0$, teacher probabilities are positive on feasible actions. A student assigning zero probability where the teacher is positive has infinite KL divergence.

$\mathcal{E}_{\mathrm{priv}}$ is the irreducible discrepancy for a *fixed* teacher and student input: it measures variation of preferences across records indistinguishable through that input. A richer representation can change it, but enumerating additional masks does not reveal hidden values at deployment. It can also be positive when action rankings agree, so small $\mathcal{E}_{\mathrm{priv}}$ is a sufficient, not necessary, condition for good decisions. The projection averages probabilities, not costs:

$$\mathbb{E}_\nu[\mathrm{softmax}(-\widetilde{Q}_W/T_Q) \mid U] \neq \mathrm{softmax}(-\mathbb{E}_\nu[\widetilde{Q}_W \mid U]/T_Q) \quad \text{in general.} \tag{17}$$

A.3 Decision margins and conditional target averaging

The projection in Proposition 1 averages probabilities rather than action costs. This subsection bounds the resulting change in action ordering.

For the fixed predictor and deployed terminal rule g, define

$$\begin{aligned} R_g(H) &= \mathbb{E}[L_g(W, S) \mid H_S = H], \\ Q_g^*(H, \mathsf{stop}) &= R_g(H), \\ Q_g^*(H, m) &= \lambda c_m + \mathbb{E}[V_g^*(H_m^+) \mid H], \\ V_g^*(H) &= \min_{a \in \mathcal{A}_S^{\mathrm{acq}}} Q_g^*(H, a). \end{aligned} \tag{18}$$

The finite, acyclic acquisition process determines these values by backward induction, with only stop available at a full mask. The terminal rule is fixed, including any prediction/deferral decision; the comparison does not give the reference policy a different predictor or terminal head. All conditional expectations refer to the fresh-record distribution in Appendix A.1.

At input $U = u$, write $q(u) = \bar{q}(u)$ and $Z = \widetilde{Q}_W - q(u)$, so that $\mathbb{E}[Z \mid U = u] = 0$. Teacher and student action sets agree because U contains the mask and observed availability information. Define $p^0(u) = \mathrm{softmax}(-q(u)/T_Q)$. The actual projected target $\bar{p}$ need not equal p^0.

Lemma 3 (Averaging error for pairwise preferences)**.** *For any two feasible actions i, j and finite $v^2(u)$,*

$$\left|\bar{p}_i(u) - \bar{p}_j(u) - \left(p_i^0(u) - p_j^0(u)\right)\right| \leq \frac{v^2(u)}{2T_Q^2}. \tag{19}$$

Proof. Fix a cost perturbation z and put $p_i(s) = \mathrm{softmax}_i(-(q + sz)/T_Q)$ for $s \in [0, 1]$. Let $m(s) = \sum_a p_a(s) z_a$. Direct differentiation gives

$$p_i''(s) = \frac{p_i(s)}{T_Q^2}\left[(z_i - m(s))^2 - \mathrm{Var}_{p(s)}(z)\right]. \tag{20}$$

Both terms in brackets are nonnegative before subtraction and at most $\mathrm{osc}(z)^2$. Hence $|p_i''(s) - p_j''(s)| \leq (p_i(s) + p_j(s))\,\mathrm{osc}(z)^2/T_Q^2 \leq \mathrm{osc}(z)^2/T_Q^2$. Taylor's formula with integral remainder bounds the pairwise remainder by $\mathrm{osc}(z)^2/(2T_Q^2)$. The first derivative at zero is linear in z, so its conditional expectation vanishes. Taking expectations proves the claim. Finite conditional first moments and $v^2(u)$ justify these operations. □

Proof of Theorem 2. Let a^* be the unique minimizer of $Q_g^*(H, \cdot)$. By the definition of $b(H)$, every competitor $a \neq a^*$ satisfies

$$\bar{q}_a(U) - \bar{q}_{a^*}(U) \geq Q_g^*(H, a) - Q_g^*(H, a^*) - b(H) \geq \Delta(H) - b(H) > 0. \tag{21}$$

Thus a^* also minimizes the conditional mean teacher costs. With n feasible actions, $p_{a^*}^0 \geq 1/n$, and

$$p_{a^*}^0 - p_a^0 = p_{a^*}^0 \left[1 - e^{-(\bar{q}_a - \bar{q}_{a^*})/T_Q}\right] \geq \frac{1 - e^{-(\Delta - b)/T_Q}}{n}. \tag{22}$$

Lemma 3 implies $\bar{p}_{a^*} - \bar{p}_a \geq M_T(H)$. Student approximation can decrease a pairwise probability gap by at most $2e(U)$, so the strict criterion in Equation (10) makes a^* the unique student maximizer. When the criterion holds at every multi-action history, backward induction identifies the resulting policy with an optimal observed-history acquisition policy. Single-action histories are automatic. This proves the theorem.

Proof of Equation (11)**.** Substituting Equation (18) into the sum of action disadvantages telescopes the conditional successor values in expectation. The terminal step contributes $R_g(H_T)$, whose expectation is the actual terminal loss. This gives the performance-difference identity

$$J_{\lambda,g}(\widehat{\pi}) - J_{\lambda,g}^\star = \mathbb{E}_{\widehat{\pi}} \sum_{t=0}^{T} [Q_g^*(H_t, A_t) - V_g^*(H_t)].$$

Outside $\mathcal{B}_\pi$, Theorem 2 gives zero action disadvantage. Inside $\mathcal{B}_\pi$, the disadvantage is bounded by $B_g(H_t)$. At a single-action history, $B_g(H_t) = 0$. If $B_g(H) \leq B_*$, at most $K + 1$ decisions contribute, giving the bound $(K + 1)B_*$.

Effect of teacher temperature. Suppose uniformly $\Delta - b \geq d_0 > 0$, $v^2 \leq V_0 < \infty$, and $n \leq n_{\max}$. For $T_Q \geq d_0$, the elementary inequality $1 - e^{-x} \geq x/2$ for $0 \leq x \leq 1$ gives

$$M_T(H) \geq \frac{d_0}{2n_{\max}T_Q} - \frac{V_0}{2T_Q^2}. \tag{23}$$

The right side is positive when $T_Q > n_{\max}V_0/d_0$. Any student with error below half this bound selects optimal actions. Thus finite fluctuations need not prevent consistency when mean action ordering is compatible. Increasing temperature also narrows the probability margin and requires more accurate fitting.

A.4 Why absolute teacher variability need not create regret

Corollary 4 (Common-score-shift invariance)**.** *Replace every terminal score of a record by $\rho'_W(S) = \rho_W(S) + \zeta(W)$, where the same integrable $\zeta(W)$ is used at every subset. Teacher soft targets, their projection, $b(H)$, $v^2(U)$, and the decision criterion are unchanged.*

Proof. Backward induction in Equation (5) gives $\widetilde{V}'_W(S) = \widetilde{V}_W(S) + \zeta(W)$ and $\widetilde{Q}'_W(a, S) = \widetilde{Q}_W(a, S) + \zeta(W)$ for every action. Softmax removes this common additive term. Conditional mean costs shift by the action-independent $\mathbb{E}[\zeta \mid U]$, leaving b unchanged. The centered cost vector shifts by $(\zeta - \mathbb{E}[\zeta \mid U])\mathbf{1}$, leaving its oscillation and hence v^2 unchanged. □

For a large positive constant ζ, the absolute terminal mismatch used by the global bound can be large while every acquisition decision remains identical. Action-dependent score changes, including arbitrary nonlinear clipping, need not have this invariance; their effects enter b and v^2.

Proposition 5 (Symmetric binary preferences)**.** *With two feasible actions a, b, put $D = \widetilde{Q}_W(b, S) - \widetilde{Q}_W(a, S)$ and $\mu(U) = \mathbb{E}[D \mid U]$. If the conditional distribution of $D - \mu(U)$ is symmetric about zero, then, for every finite $T_Q > 0$,*

$$\operatorname{sign}\big(\bar{p}(a \mid U) - \tfrac{1}{2}\big) = \operatorname{sign}(\mu(U)). \tag{24}$$

Thus an exact projection student ranks actions by their conditional mean teacher costs. This is online-optimal whenever that ranking agrees with $Q_g^*(H, \cdot)$.

Proof. The teacher probability of a is $\sigma(D/T_Q)$, where $\sigma(x) = (1 + e^{-x})^{-1}$. If $Z = D - \mu$ is symmetric, $\mathbb{E}[\sigma(Z/T_Q) \mid U] = 1/2$ by $\sigma(x) + \sigma(-x) = 1$. The expectation of $\sigma((x + Z)/T_Q)$ is strictly increasing in x for finite Z and $T_Q > 0$. Its sign relative to $1/2$ is therefore the sign of μ. □

A positive information gap with an optimal projected policy. Consider a balanced binary target Y, and an independent success indicator Z with probability $1/2$. Initially the fixed terminal rule defers at loss 0.3. A query costs 0.1 and reveals (Z, Y) on success or a failure symbol otherwise. After a successful query the terminal rule predicts Y with loss zero; on failure it defers at loss 0.3. Use these exact realized losses as $\rho_W(S)$.

The initial teacher cost vector $(\widetilde{Q}_{\mathsf{stop}}, \widetilde{Q}_{\mathrm{query}})$ is $(0.3, 0.1)$ on success and $(0.3, 0.4)$ on failure. The online-optimal query costs $0.1 + 0.5(0.3) = 0.25$, whereas immediate stopping costs 0.3. The hindsight teacher attains $0.5(0.1) + 0.5(0.3) = 0.20$, giving information advantage 0.05. Nevertheless,

$$\bar{p}(\mathrm{query}) = \tfrac{1}{2}\sigma(0.2/T_Q) + \tfrac{1}{2}\sigma(-0.1/T_Q) > \tfrac{1}{2} \qquad (T_Q > 0). \tag{25}$$

The exact argmax projection therefore queries and has zero online regret for every finite positive temperature. Teacher preferences still vary across successful and failed records, so their conditional KL variability is positive. This analytical construction shows that a positive information-relaxation gap is compatible with optimal deployed decisions. The asymmetric counterexample in Appendix B.4 shows why compatibility cannot be presumed for all distributions.

A.5 Sparse subset planning and decision stability

Sparse planning changes the offline search budget, not the observed-history deployment interface. For a supervised root S, let $\mathcal{G}_{W,S} = (\mathcal{V}, \mathcal{E})$ be a finite DAG of subsets containing S, with edges $R \to R \cup \{m\}$. Every retained node permits stopping at its original score $\rho_W(R)$. At the supervised root, include every feasible first query and its successor; deeper continuations can be pruned. The recurrence on this graph is

$$\widetilde{V}^{\mathcal{G}}_W(R) = \min\left\{\rho_W(R), \min_{(R, R\cup\{m\})\in\mathcal{E}}[\lambda c_m + \widetilde{V}^{\mathcal{G}}_W(R \cup \{m\})]\right\}. \tag{26}$$

A leaf stops; the inner minimum of an empty set is $+\infty$. This specifies an exact backward pass on the retained DAG. At the root it yields finite costs for all original actions: $\widetilde{Q}^{\mathcal{G}}_W(\mathsf{stop}, S) = \rho_W(S)$ and $\widetilde{Q}^{\mathcal{G}}_W(m, S) = \lambda c_m + \widetilde{V}^{\mathcal{G}}_W(S \cup \{m\})$. Every retained continuation is feasible in the full lattice, so $\widetilde{Q}^{\mathcal{G}}_W(a, S) \geq \widetilde{Q}_W(a, S)$. Equality is recovered when the graph contains a full-lattice optimal continuation for that first action. Merely retaining the optimal path from the initial empty mask need not give equality for other roots or actions.

Computation. For constructed graphs indexed by j (record–root pairs), scoring takes at most $\sum_j |\mathcal{V}_j|$ predictor evaluations, and each price profile requires $O(\sum_j |\mathcal{V}_j| + \sum_j |\mathcal{E}_j|)$ scalar operations. Shared nodes can be cached. These are post-construction costs; proposal generation and scoring discarded candidates also consume work. A width-b, depth-D beam that expands at most K children per retained node generates at most $1 + K + (D - 1)bK$ candidate states per supervised root, before duplicates are removed. Thus its candidate-scoring count is $O(bKD)$, with separate ranking/bookkeeping costs. Stopping at every leaf defines the truncation. Search budgets bound computation, while discarded continuations determine approximation error.

Lemma 6 (Soft-target sensitivity to pruning)**.** *For cost vectors q, q' on the same action set and $p = \mathrm{softmax}(-q/T_Q)$, $p' = \mathrm{softmax}(-q'/T_Q)$,*

$$\|p' - p\|_1 \leq \frac{\mathrm{osc}(q' - q)}{2T_Q}. \tag{27}$$

Proof. Write $d = q' - q$ and $p(s) = \mathrm{softmax}(-(q + sd)/T_Q)$. Differentiation gives $\sum_a |p'_a(s)| = T_Q^{-1}\mathbb{E}_{p(s)}|d - \mathbb{E}_{p(s)}d| \leq \mathrm{osc}(d)/(2T_Q)$. The last bound follows from Cauchy–Schwarz and the variance bound for a variable in an interval of length $\mathrm{osc}(d)$. Integrate from zero to one. □

Let $\bar{p}_{\mathcal{G}}(U) = \mathbb{E}[p_T^{\mathcal{G}}(\cdot \mid W, S) \mid U]$, where $p_T^{\mathcal{G}}$ uses the same temperature on the sparse root costs. Let $e_{\mathcal{G}}(U) = \|\pi_{\mathcal{G}}(\cdot \mid U) - \bar{p}_{\mathcal{G}}(U)\|_\infty$, and define the conditional pruning distortion

$$s_{\mathcal{G}}(U) = \mathbb{E}\Big[\operatorname{osc}(\widetilde{Q}_W^{\mathcal{G}}(\cdot, S) - \widetilde{Q}_W(\cdot, S)) \mid U\Big]. \tag{28}$$

Random search seeds can be included in this expectation. They are used only in offline target construction.

Corollary 7 (Sparse-target decision preservation)**.** *Under Theorem 2's assumptions, a sparse-target student selects the same unique online-optimal action at H whenever $\Delta(H) > b(H)$ and*

$$M_T(H) > \frac{s_{\mathcal{G}}(U)}{2T_Q} + 2e_{\mathcal{G}}(U). \tag{29}$$

The regret bound in Equation (11) *applies with failure of this augmented condition in place of $\mathcal{B}_\pi$.*

Proof. Lemma 6 bounds the change of every pairwise projected probability difference by $s_{\mathcal{G}}(U)/(2T_Q)$. Theorem 2's proof supplies the full-target margin $M_T(H)$. Student approximation reduces the remaining margin by at most $2e_{\mathcal{G}}(U)$. The strict inequality preserves the optimum. Equation (11) uses the same actual output rule and online costs and therefore applies unchanged. □

Retaining all first queries at each supervised root is essential to the stated comparison: it keeps the target action space fixed. Deeper graph membership can depend on the complete record, but it is not an input to the deployed controller. The experiments use the exact lattice teacher; the sparse formulation extends its offline planning procedure. No general pruning-error bound follows from beam width alone.

A.6 Global soft-target transfer bound

The parameter δ bounds the conditional-mean discrepancy between the teacher score $\rho_W(S)$ and the loss $L_g(W, S)$ of the fixed terminal rule, including effects of calibration, subset regularization, and clipping. The decision-margin result instead uses the action-relative distortion in Equation (9). Appendix B.6 analyzes terminal-score perturbations and nonlinear clipping.

For the global KL bound, assume

$$\max_{a \in \mathcal{A}_S^{\mathrm{acq}}} \widetilde{Q}_W(a, S) - \min_{a \in \mathcal{A}_S^{\mathrm{acq}}} \widetilde{Q}_W(a, S) \le B_Q \tag{30}$$

for every mask and almost every record, with $B_Q > 0$. For example, scores in $[0, B_\rho]$ and finite prices imply $0 \le \widetilde{V}_W(S) \le B_\rho$ and permit $B_Q = B_\rho + \lambda \max_m c_m$. The bound assumes a finite action-cost span; an unbounded log loss requires a separate span control.

Fix the deployed terminal rule g, including any deferral. Let $L_g(W, S)$ be its loss at S, $J_{\lambda,g}$ the objective in Equation (4), and $J^\star_{\lambda,g}$ its infimum over observed-history acquisition policies using the *same* predictor and rule. For $\delta \ge 0$, assume

$$|\mathbb{E}[\rho_W(S) - L_g(W, S) \mid H_S]| \le \delta \tag{31}$$

for every mask and almost every history; realized scores need not equal conditional risks.

Theorem 8 (Soft-target transfer bound)**.** *Assume the static evidence setting of Section 3, integrable losses, Equation* (31)*, and teacher action-cost span at most $B_Q > 0$. Set $w_{\min} = 2^{-K}$. For $\widehat{\pi}(U) \in \arg\max_a \pi_\eta(a \mid U)$ and $T_Q > 0$,*

$$0 \le J_{\lambda,g}(\widehat{\pi}) - J^\star_{\lambda,g} \le C\sqrt{\mathcal{E}_{\mathrm{priv}} + \mathcal{E}_{\mathrm{fit}}} + 2\delta. \tag{32}$$

The constant C depends only on $K, T_Q, B_Q, w_{\min}$ and is given below.

A.7 From soft preference error to path cost

Define the record-level disadvantage

$$\Gamma_W(S, a) = \widetilde{Q}_W(a, S) - \widetilde{V}_W(S) \ge 0. \tag{33}$$

For a realized acquisition path, $\mathcal{P}$ denotes its sequence of (S, a) pairs, including the final stop. If S_T is its terminal subset, set $\mathcal{C}_\rho(W, \mathcal{P}) = \rho_W(S_T) + \lambda \sum_{m \in S_T} c_m$.

Lemma 9 (Pathwise telescoping)**.** *Every feasible path satisfies*

$$\mathcal{C}_\rho(W,\mathcal{P}) - \widetilde{V}_W(\varnothing) = \sum_{(S,a)\in\mathcal{P}} \Gamma_W(S,a). \tag{34}$$

Proof. For a query m, the summand is $\lambda c_m + \widetilde{V}_W(S\cup\{m\}) - \widetilde{V}_W(S)$. For the final stop it is $\rho_W(S_T) - \widetilde{V}_W(S_T)$. Consecutive values cancel. The remaining terms are the query costs, terminal score, and negative initial teacher value. Each summand is nonnegative because $\widetilde{V}_W(S)$ is the minimum action cost. □

The next lemma translates the actual soft-target loss into a cost disadvantage for deterministic decoding.

Lemma 10 (Softmax ranking control)**.** *Let $q\in\mathbb{R}^n$ have span at most B_Q, $p_a \propto e^{-q_a/T_Q}$, and let $\widehat{a}$ maximize any probability vector v. Then*

$$q_{\widehat{a}} - \min_a q_a \le \frac{nB_Q}{1-e^{-B_Q/T_Q}}\|p-v\|_1 \le \frac{nB_Q}{1-e^{-B_Q/T_Q}}\sqrt{2D_{\mathrm{KL}}(p\|v)}. \tag{35}$$

Proof. Choose a minimum-cost action a^* and put $d = q_{\widehat{a}} - q_{a^*} \in [0, B_Q]$. Then $p_{a^*} \ge 1/n$ and

$$p_{a^*} - p_{\widehat{a}} = p_{a^*}(1-e^{-d/T_Q}) \ge \frac{d}{nB_Q}(1-e^{-B_Q/T_Q}).$$

The last inequality uses concavity of $1-e^{-d/T_Q}$ on $[0,B_Q]$. Since $v_{\widehat{a}} \ge v_{a^*}$, the same probability difference is at most $|p_{a^*}-v_{a^*}| + |p_{\widehat{a}} - v_{\widehat{a}}| \le \|p-v\|_1$. This proves the first inequality; Pinsker's inequality proves the second. The case $\widehat{a}=a^*$ is immediate. Arbitrary observed-input tie-breaking is allowed. □

A.8 Proof of the soft-target transfer bound

Write $h = K+1$ and $A_Q = hB_Q/(1-e^{-B_Q/T_Q})$. Let $I_S(W)$ indicate that the argmax student's path visits S, and let $k_W(S) = D_{\mathrm{KL}}(p_W(\cdot,S)\|\pi_\eta(\cdot\mid U_S))$. Each mask is visited at most once and $\sum_S I_S \le h$. Lemmas 9–10 imply

$$\begin{aligned} J_\rho(\widehat{\pi}) - \mathbb{E}\widetilde{V}_W(\varnothing) &\le A_Q\sqrt{2}\,\mathbb{E}\sum_S I_S\sqrt{k_W(S)} \\ &\le A_Q\sqrt{2h\,\mathbb{E}\sum_S I_S k_W(S)} \\ &\le A_Q\sqrt{\frac{2h}{w_{\min}}\mathbb{E}_\nu k_W(S)}. \end{aligned} \tag{36}$$

The second line applies Cauchy–Schwarz along each path and then Jensen's inequality. The last line follows pointwise from $0\le I_S\le 1$ and $w_S \ge w_{\min}$. Importantly, the visited masks may depend on observed values: no independence between visitation and teacher errors is assumed. Static revelation makes all counterfactual masks of the *same* fresh record available for this domination argument.

The complete-record teacher minimizes realized score-plus-cost over all feasible paths. Consequently,

$$\mathbb{E}\widetilde{V}_W(\varnothing) \le \inf_{\pi\in\Pi_{\mathrm{obs}}} J_\rho(\pi). \tag{37}$$

It remains to connect score cost to the actual fixed terminal procedure. For a deterministic nonanticipative acquisition policy, the event $\{S_T = S\}$ is measurable from H_S: deciding whether a path ends at S only requires values in that final subset. For randomized policies the same statement holds conditional on independent policy randomness. Therefore,

$$J_\rho(\pi) - J_{\lambda,g}(\pi) = \sum_S \mathbb{E}\left[\mathbf{1}\{S_T=S\}\,\mathbb{E}[\rho_W(S) - L_g(W,S)\mid H_S]\right], \tag{38}$$

with independent randomization integrated out. Equation (31) and $\sum_S \mathbf{1}\{S_T = S\} = 1$ give $|J_\rho(\pi) - J_{\lambda,g}(\pi)| \le \delta$ for every comparison policy. Only one terminal loss is paid, so this error does not multiply by the number of queries.

Combining Equations (36)–(38) and Proposition 1,

$$J_{\lambda,g}(\widehat{\pi}) - J^\star_{\lambda,g} \le J_\rho(\widehat{\pi}) - \inf_{\pi \in \Pi_{\mathrm{obs}}} J_\rho(\pi) + 2\delta$$
$$\le \frac{(K+1)B_Q}{1 - e^{-B_Q/T_Q}} \sqrt{\frac{2(K+1)}{w_{\min}}(\mathcal{E}_{\mathrm{priv}} + \mathcal{E}_{\mathrm{fit}})} + 2\delta. \tag{39}$$

The lower bound is the definition of the infimum, since $\widehat{\pi} \in \Pi_{\mathrm{obs}}$. This proves Theorem 8. For the uniform scheme, substitute $w_{\min} = 2^{-K}$; the factor inside the square root is $2(K+1)2^K$.

The global bound controls path error through the uniform-mask population loss and therefore carries a coverage factor. Theorem 2 instead uses conditional action margins and approximation errors at observed inputs. These are different sufficient conditions: increasing temperature reduces nonlinear aggregation error in the latter, but also narrows probability margins. Both analyses refer to the specified population P.

A.9 Action compatibility and improvement over a reference

A useful exact case does not require reconstructing unqueried measurements. Suppose there is an observed-input action $a^\dagger(U)$ that minimizes $\widetilde{Q}_W(\cdot, S)$ for almost every record consistent with U, at every mask. Then it maximizes every corresponding soft teacher distribution. It therefore maximizes $\bar{p}$; any other maximizer of $\bar{p}$ is also a teacher maximizer almost surely, since the nonnegative teacher-probability gap has mean zero. An exact projection student with argmax decoding incurs zero disadvantage on each visited mask. Lemma 9 gives $J_\rho(\widehat{\pi}) = \mathbb{E}\widetilde{V}_W(\varnothing)$, so it is online-optimal for the score objective. With $\delta = 0$ it is optimal for $J_{\lambda,g}$ as well. This sufficient compatibility condition concerns action preferences, not equality of all hidden evidence values.

More generally, let π_b be any reference acquisition policy using the same fixed output rule and define $\Delta_b = J_{\lambda,g}(\pi_b) - J^\star_{\lambda,g}$. If the right side of Equation (32) is strictly smaller than Δ_b, the distilled controller improves on that reference in the loss–cost objective. This implication compares acquisition policies for the same fixed output rule and population.

A.10 Sampled-action policies

For a stochastic controller, actions are sampled from the learned policy at each visited history. The following result complements the argmax guarantees in Theorems 2 and 8.

Corollary 11 (Sampled-policy bound)**.** *Under the assumptions of Theorem 8, sampling $a \sim \pi_\eta(\cdot \mid U)$ independently at each visited state yields*

$$J_{\lambda,g}(\pi_\eta) - J^\star_{\lambda,g} \le h \min\{B_Q, T_Q \log h\} + B_Q \sqrt{\frac{h}{2w_{\min}}(\mathcal{E}_{\mathrm{priv}} + \mathcal{E}_{\mathrm{fit}})} + 2\delta, \quad h = K+1. \tag{40}$$

Proof. At a fixed record and mask, shift the teacher costs by their minimum to obtain disadvantages $d_a \in [0, B_Q]$. For $p_a = e^{-d_a/T_Q}/Z$ with $Z \ge 1$,

$$\mathbb{E}_{a\sim p} d_a = T_Q\big(\mathsf{H}(p) - \log Z\big) \le \min\{B_Q, T_Q \log |\mathcal{A}_S^{\mathrm{acq}}|\}.$$

Changing the action distribution from p to π_η changes this expectation by at most $B_Q\,\mathrm{TV}(p, \pi_\eta) \le B_Q\sqrt{D_{\mathrm{KL}}(p\|\pi_\eta)/2}$. Apply the pathwise identity conditional on independent action randomness, sum at most h entropy terms, and use the same coverage and terminal-bias arguments as above. This proves Equation (40). □

The additional temperature term accounts for sampling suboptimal actions, even when the student matches the soft teacher. Argmax and sampled decoding therefore have distinct guarantees.

B Decision-Theoretic Benchmarks and Examples

B.1 Information and prediction conventions

Let $(\Omega, \mathcal{F}, P)$ carry one coherent joint distribution of mandatory evidence, optional groups, and labels. The sigma-field $\mathcal{F}_S$ is generated by the observed history H_S. A query refines it to $\mathcal{F}_{S\cup\{m\}}$. Regular conditional distributions are assumed to exist. All conditional expectations in the online recursion are taken with respect to this joint distribution.

The predictor $f(H_S)$ is fixed during the decision analysis. Unless stated otherwise, the proofs below spell out the optional-deferral instantiation: its terminal actions are prediction and deferral, with fixed $\tau > 0$. Under compulsory prediction, stopping invokes the predictor and the Bellman argument uses $R(H) = r(H)$ instead. Query costs are finite and nonnegative. A policy is measurable with respect to the current history (and, for randomized policies, independent randomization). No query changes the underlying target or measurements. These assumptions describe retrospective static evidence revelation. Calendar-time waiting, query failures, and unavailable future measurements would require an enlarged state and are not modeled by the present experiment.

A compressed input can merge histories with different useful actions, contributing to the irreducible discrepancy $\mathcal{E}_{\mathrm{priv}}$ defined in Appendix A.2. Observed metadata means metadata genuinely available at decision time, not merely present in the complete record. The ideal terminal rule used below is a decision benchmark; Theorem 8 instead holds the deployed terminal rule g fixed.

Online decision benchmark. The learned controller acts without knowing future evidence values. For a fixed predictor f, define its conditional prediction risk $r(H)$ and the optimal cost $R(H)$ of the allowed terminal rule by

$$r(H) = \mathbb{E}[\ell(f(H), Y) \mid H], \quad R(H) = \begin{cases} r(H), & \text{compulsory prediction,} \\ \min\{r(H), \tau\}, & \text{optional deferral.} \end{cases} \tag{41}$$

Here H denotes any observed history. With deferral, ties favor deferring. These conditional risks describe an online decision benchmark, whereas the teacher's $\rho_t(S)$ is a realized score. For the objective in Equation (4), the minimum conditional cost-to-go is

$$V_\lambda(H_S) = \min\left\{ R(H_S), \min_{m\in\mathcal{M}\setminus S} \left[\lambda c_m + \mathbb{E}[V_\lambda(H_{S\cup\{m\}}) \mid H_S]\right] \right\}. \tag{42}$$

The expectation averages possible values of the next group, conditional on H_S; write $H_m^+ = H_{S\cup\{m\}}$ for the enlarged history including that value. At the full subset, $V_\lambda(H_{\mathcal{M}}) = R(H_{\mathcal{M}})$. Backward induction gives the optimum over policies using only observed evidence. Fixing f separates its prediction error from error in the acquisition controller.

B.2 Terminal loss and acquisition stopping in the Bellman recurrence

Proof. At a fixed history, predicting incurs conditional loss $r(H)$ and deferring incurs τ. Their minimum is $R(H) = \min\{r(H), \tau\}$. Prediction is strictly preferred exactly when $r(H) < \tau$; at equality either action is optimal, and the stated tie rule selects deferral.

At the full subset, no query remains, so the optimal value is $R(H_{\mathcal{M}})$. Suppose the optimal values at all strict successor subsets are known. Any policy at H_S either stops or first queries an unacquired group m. The minimum cost in the first case is $R(H_S)$. In the second case it is $\lambda c_m + \mathbb{E}[V_\lambda(H_{S\cup\{m\}}) \mid H_S]$. Minimizing over this exhaustive finite first-action set gives Equation (42). Induction in the number of unacquired groups completes the proof. The recurrence includes an expectation precisely because the query's value is not known at the time of choosing it. □

For a fixed imperfect predictor, the one-query value need not be nonnegative: its predictions under different masks can be inconsistent. With a Bayes-optimal predictor and permission to ignore new evidence, information cannot increase minimum expected decision loss before cost. The nonnegativity claim requires those extra conditions; it is not an automatic property of the fitted network.

B.3 Terminal loss and acquisition value

For the optional-deferral terminal rule $R(H) = \min\{r(H), \tau\}$, write $H_m^+ = H_{S\cup\{m\}}$ and $r^+ = r(H_m^+)$. The one-query value before paying acquisition cost is $\Delta_m(H) = R(H) - \mathbb{E}[R(H_m^+) \mid H]$.

Proposition 12 (Threshold-sensitive acquisition value)**.** *For integrable conditional prediction risks,*

$$\Delta_m(H) = r(H) - \mathbb{E}[r^+ \mid H] + \mathbb{E}[[r^+ - \tau]_+ \mid H] - [r(H) - \tau]_+. \tag{43}$$

Equal expected prediction-risk reductions can therefore have different acquisition values.

The correction accounts for prediction loss replaced by deferral.

Proof. Use $\min\{u, \tau\} = u - [u - \tau]_+$ at both the current and successor risks. Then

$$\begin{aligned} R(H) - \mathbb{E}[R(H_m^+) \mid H] &= r(H) - [r(H) - \tau]_+ - \mathbb{E}[r^+ - [r^+ - \tau]_+ \mid H] \\ &= r(H) - \mathbb{E}[r^+ \mid H] + \mathbb{E}[[r^+ - \tau]_+ \mid H] - [r(H) - \tau]_+. \end{aligned}$$

Integrability justifies the rearrangement. The two observation channels constructed below have equal expected risk reductions and different values of the correction term, proving the final statement. □

B.4 Observation channels realizing the numerical examples

Equal average risk, unequal selective value. Let $Y \sim \mathrm{Bernoulli}(1/2)$. For query A, reveal a binary symmetric channel output with independent flip probability 0.2. The posterior Bayes error is 0.2 for either output. For query B, draw an independent $Z \sim \mathrm{Bernoulli}(1/2)$ and reveal Z as part of the observation. If $Z = 1$, reveal Y exactly; if $Z = 0$, reveal a binary symmetric channel with flip probability 0.4. Thus posterior error is 0 with probability $1/2$ and 0.4 otherwise. Both queries have expected prediction error 0.2.

With $\tau = 1/4$, the current selective risk is $1/4$. Query A's expected stopping risk is $1/5$, giving value $1/20 = 0.05$. Query B's expected stopping risk is $\frac{1}{2}(0) + \frac{1}{2}(1/4) = 1/8$, giving value $1/8 = 0.125$. The current prediction error is 0.5. At acquisition price $\lambda c = 2/25 = 0.08$, only query B is worth buying, despite the equal expected prediction errors. Both observation channels share the same balanced prior.

Complementary evidence. Let $X^{(1)}, X^{(2)}$ be independent fair bits and $Y = X^{(1)} \oplus X^{(2)}$. A Bayes predictor has error $1/2$ before any query and after either single query, and zero after both. With $\tau = 0.3$ and cost 0.05 per group, every one-step gain is zero, so a myopic stop/query rule stops. The two-query plan has total value 0.1, less than stopping at 0.3. With compulsory prediction, stopping costs 0.5 and a single query followed by immediate prediction costs $0.05 + 0.5 = 0.55$, whereas the two-query continuation costs 0.1. The preference for multi-step continuation therefore does not rely on deferral. A subset-lattice planner can represent either continuation.

Teacher-majority counterexample. Take a balanced binary label, an independent success event Z of probability 0.6, and one observation returning (Z, Y) on success and a failure symbol otherwise. On success the terminal risk is zero; on failure it is 0.3, because deferral beats error 0.5. The online query cost is $0.2 + 0.4(0.3) = 0.32$, so the optimum is to defer at 0.3. The clairvoyant reference teacher pays 0.2 on success and otherwise stops at 0.3, with relaxed mean value $0.6(0.2)+0.4(0.3) = 0.24$. The information gap is 0.06.

Under hard-label cross-entropy distillation from a constant initial history, the population-optimal student's query probability is 0.6. Argmax decoding therefore queries and incurs regret 0.02. Sampling from the predicted probabilities also does not recover the optimal defer-only rule: its expected objective is $0.6(0.32) + 0.4(0.3) = 0.312$. This is a hard-label inconsistency example. The same observation model also permits a soft-target example. For temperature $T_Q = 0.02$, the teacher's query probability is $\sigma(0.1/T_Q)$ on success and $\sigma(-0.2/T_Q)$ on failure, where $\sigma(v) = (1 + e^{-v})^{-1}$. The exact projected probability is $0.6\sigma(5) + 0.4\sigma(-10) \approx 0.5960 > 1/2$, so its argmax still queries. Here $\mathcal{E}_{\mathrm{fit}} = 0$ but $\mathcal{E}_{\mathrm{priv}} > 0$; the transfer bound does not imply zero regret. The conclusion is specific to the stated observation model and temperature.

B.5 COMPLETE-RECORD INFORMATION ADVANTAGE

To isolate anticipation from score-estimation error, first use the true stopping cost $R(H_S)$ inside the realized teacher. For a complete record W, denote the clairvoyant reference value by $V^{\mathrm{cl}}(W, S)$ and define

$$L(H_S) = \mathbb{E}[V^{\mathrm{cl}}(W, S) \mid H_S], \quad j(H) = \min_a \mathbb{E}[Q_a^{\mathrm{cl}} \mid H] - \mathbb{E}[\min_a Q_a^{\mathrm{cl}} \mid H], \tag{44}$$

where $Q_{\mathsf{stop}}^{\mathrm{cl}} = R(H)$ and $Q_m^{\mathrm{cl}} = \lambda c_m + V^{\mathrm{cl}}(W, S \cup \{m\})$.

Theorem 13 (Local information-relaxation gap)**.** *For finite K, nonnegative finite costs, and integrable terminal risks, $L(H) \le V_\lambda(H)$ and $j(H) \ge 0$. With $d(H) = V_\lambda(H) - L(H)$,*

$$j(H) \le d(H) \le j(H) + \max\left\{0, \max_{m \notin S} \mathbb{E}[d(H_m^+) \mid H]\right\}. \tag{45}$$

At the full subset, $d = 0$. If $\bar{j}_i$ bounds j at histories with i unacquired groups, then $d(H) \le \sum_{i=1}^{K-|S|} \bar{j}_i$.

The gap quantifies the benefit of selecting actions with access to unqueried evidence values. Realized outcome-dependent terminal scores add a separate discrepancy; they are not identified with conditional risks by this theorem.

Proof. A clairvoyant reference planner minimizes the realized path cost among all subset paths, while every nonanticipative policy chooses one of those paths. Its clairvoyant minimum is therefore no larger than the realized conditional-risk path cost of any online policy. Taking conditional expectations and then minimizing over online policies gives $L(H) \le V_\lambda(H)$. The same fact follows by backward induction using the tower property.

At history H, put $u_a = \mathbb{E}[Q_a^{\mathrm{cl}} \mid H]$. Since minimization is concave as a function of the action-cost vector,

$$j(H) = \min_a u_a - \mathbb{E}[\min_a Q_a^{\mathrm{cl}} \mid H] \ge 0.$$

Let $b_{\mathsf{stop}} = 0$ and $b_m = \mathbb{E}[d(H_m^+) \mid H] \ge 0$. The tower property yields $\mathbb{E}[V^{\mathrm{cl}}(W, S \cup \{m\}) \mid H] = \mathbb{E}[L(H_m^+) \mid H]$, hence

$$V_\lambda(H) = \min_a (u_a + b_a), \qquad d(H) = j(H) + \min_a (u_a + b_a) - \min_a u_a.$$

The last difference lies between zero and $\max_a b_a$. This is Equation (45). At the full subset, both planners must stop, so $d = 0$. Applying the upper bound recursively with depth-wise bounds $\bar{j}_i$ gives $d(H) \le \sum_{i=1}^{K-|S|} \bar{j}_i$. □

B.6 TERMINAL-SCORE PERTURBATIONS

Theorem 13 uses $R(H)$ in the clairvoyant reference teacher to isolate the information advantage. The implementation instead uses realized terminal supervision $\rho_t(S)$. Suppose, as a deterministic additional assumption, that every subset in a complete episode obeys $|\rho_t(S) - R(H_{t,S})| \le \eta$. Then

$$|\widetilde{V}_{\lambda,t}(S) - V^{\mathrm{cl}}(W_t, S)| \le \eta \tag{46}$$

for every subset. To prove this, use the boundary inequality at the full set and the sup-norm nonexpansiveness of a minimum at each earlier subset. Each path pays only one terminal loss, so this particular perturbation does not accumulate once per query.

Clipping also changes the conditional mean of a realized loss. Even if an individual loss L_t is an unbiased sample of conditional prediction risk, concavity gives

$$\mathbb{E}[\min\{\tau, L_t\} \mid H] \le \min\{\tau, \mathbb{E}[L_t \mid H]\}. \tag{47}$$

For example, if L_t is zero or one with equal probability and $\tau = 0.3$, the left side is 0.15 while the right side is 0.3. Predicting whether an individual realized loss falls below a threshold is not the same as estimating conditional expected loss. Thus unbiased prediction losses do not justify setting the terminal mismatch δ in Equation (31) to zero after clipping or a different terminal-output rule.

B.7 Cost multipliers and optimal expenditure

The multiplier λ penalizes evidence cost; it is not a hard budget. A per-episode cap would additionally require remaining budget in the state and removal of unaffordable queries. For exact optimizers π_1, π_2 at $0 \leq \lambda_1 < \lambda_2$, let L_i and C_i denote their expected terminal loss and evidence cost. Optimality gives

$$L_1 + \lambda_1 C_1 \leq L_2 + \lambda_1 C_2, \qquad L_2 + \lambda_2 C_2 \leq L_1 + \lambda_2 C_1.$$

Adding yields $(\lambda_2 - \lambda_1)(C_2 - C_1) \leq 0$, hence $C_2 \leq C_1$. The argument fixes relative prices and varies only λ; expenditures under distinct price vectors are reported directly in Table 4.

C Additional Experimental Results

The following tables report the complete acquisition-price profiles and additional predictive measurements for the same implementation used in the main experiments.

C.1 Acquisition-price profiles

Table 4 gives all price profiles evaluated in Section 6.4. The predictor and terminal procedure remain fixed; each profile has its own development-set multiplier calibration and student redistillation.

Table 4: **Policy replanning with a fixed predictor.** Each price profile has its own multiplier calibration and student redistillation. "Changed" is the reported action-change rate relative to the default policy. Costs are evaluated under the price vector shown in each row.

Price profile	Mean cost	Phase F1 ↑	Changed
Default	**2.179**	**0.983**	**0.0%**
ED 0.8×	2.280	0.987	37.0%
ED 0.9×	2.103	0.984	25.7%
ED 1.1×	2.192	0.983	51.0%
ED 1.2×	2.450	0.987	48.3%
ED 1.5×	2.715	0.984	44.9%
ED 2.0×	3.166	0.981	51.2%
HOSP 0.5×	2.053	0.983	47.6%
SRC 3.0×	2.150	0.982	63.6%
Full cost-rank swap	1.768	0.985	36.2%

C.2 Additional predictive measurements

Table 5: Alert-prediction metrics under the default acquisition costs.

System	Mean cost	Alert AUROC	Alert Brier
XGBoost (WW)	0.000	0.933	0.094
SelectiveNet (WW)	0.000	0.906	0.117
MLP (WW+SRC)	0.400	0.907	0.102
ReAct Router	0.472	0.919	0.121
MLP (WW+ED)	1.000	0.975	0.059
SLD (ours)	**2.179**	**0.966**	**0.068**

Table 5 reports alert AUROC and Brier scores. MLP (WW+ED) achieves higher alert AUROC and lower Brier score than SLD, at lower acquisition cost. Rankings can differ across phase prediction, alert prediction, and calibration; these endpoints are therefore reported separately.

SLD achieves a teacher-sequence alignment score of 0.599, compared with 0.469 for EDDI, 0.365 for GDFS, 0.355 for ACO, 0.323 for Supervised Router, and 0.056 for ReAct Router. This is agreement with realized teacher behavior.

Table 6: Evidence-representation comparison at default prices. The full model uses E5-Mistral.

Frozen representation	Mean cost	Phase F1
SLD (ours)	**2.179**	**0.983**
gte-Qwen2 replacement	2.134	0.972
W/O frozen text adapter	2.263	0.963

D Implementation Details

The implementation follows the shared subset evaluator, cost-sensitive lattice teacher, and observed-history student in Sections 4.1–4.3.

D.1 Evidence preprocessing and frozen text adapter

The fixed block order is WW, ED, HOSP, POS, and SRC. Structured features use training-split standardization, categorical fields use one-hot encoding, and week-of-year uses sine/cosine features. Development and test records use the same preprocessing. An observation mask identifies acquired evidence groups.

Frozen text adapter. Short textual or categorical summaries are represented by the frozen E5-Mistral-7B-Instruct encoder (`intfloat/e5-mistral-7b-instruct`). The task instruction is projected to 16 dimensions and each observed evidence summary to 64 dimensions, denoted *i16/e64*. Unqueried evidence embeddings are zero-filled. Cached, prompt-conditioned encoder outputs are reused across runs; its frozen parameters are excluded from trainable counts.

D.2 Shared subset evaluation

For the shared masked predictor in Section 4.1, $r_{t,S} \in \mathbb{R}^{d_r}$ denotes the input assembled from history $H_{t,S}$. The Gaussian encoder uses a residual MLP with width 128, two residual blocks, layer normalization, and dropout 0.10. It outputs the mean and log-variance used in Equation (1):

$$\mathrm{enc}_\theta(r_{t,S}) = (\mu_{t,S}, \log \sigma^2_{t,S}), \qquad \mu_{t,S}, \log \sigma^2_{t,S} \in \mathbb{R}^{16}. \tag{48}$$

The log-variance output is unconstrained. The decoder decode_ψ adopts the natural-parameter networks (NPN) of Wang et al. (2016). It takes the Gaussian state represented by $[\mu_{t,S}; \log \sigma^2_{t,S}] \in \mathbb{R}^{32}$ and propagates moments deterministically to the four-class phase distribution and binary alert probability. Table 7 summarizes the learned components.

Table 7: **SLD architecture.** Dimensions refer to i16/e64.

Component	Input	Architecture	Output
Frozen text adapter	Task instruction; observed summaries	Frozen E5-Mistral-7B-Instruct; 16-/64-dimensional projections	Task and evidence representations
Masked Gaussian encoder	Masked vector $r_{t,S} \in \mathbb{R}^{d_r}$	Width 128; two residual blocks; layer norm; dropout 0.10	16 means and 16 log-variances
Moment decoder	Gaussian moments $[\mu_S; \log \sigma^2_S] \in \mathbb{R}^{32}$	NPN moment propagation	Phase distribution; alert probability
Student controller	60-dimensional input $s(H_S)$	Width 128; one residual block; dropout 0.10	Five query/stop logits; terminal answer probability

Predictor optimization. The encoder and decoder are trained for 1,000 gradient steps with batch size 64, Adam learning rate 10^{-3}, and KL coefficient $\beta = 0.02$ in Equation (2). The probabilities of sampling the mandatory-evidence-only mask, a proper nonempty partial mask, and the full-evidence mask are

$$(p_\varnothing, p_{\mathrm{partial}}, p_{\mathrm{full}}) = (0.10, 0.30, 0.60). \tag{49}$$

This mixture retains wastewater-only examples while emphasizing full-evidence prediction. The checkpoint is selected on development data. Observed-history distillation uses the uniform mask distribution specified in Section 4.3.

D.3 COST-SENSITIVE LATTICE TEACHER

The selected predictor parameters (θ, ψ) are frozen before scoring all $2^4 = 16$ subsets of each labeled training and development record. The task-specific terminal procedure supplies the scores $\rho_t(S)$ and answer indicators $d_t(S)$ used in Section 4.1. A log-risk calibration map fitted on the training split transforms the negative log-likelihood plus subset KL into terminal scores; the deferral cost is $\tau = 3.0$. Each target budget has a separate development-selected acquisition-cost multiplier.

Multiplier search. For budget B, the multiplier λ_B is the value of λ selected on the development split from

$$\begin{aligned}\Lambda = \{&0.02, 0.05, 0.075, 0.1, 0.15, 0.2, 0.25, 0.35, 0.5, 0.75,\\ &1, 1.5, 2, 3, 5, 8, 13, 21, 34, 55, 89\}.\end{aligned} \tag{50}$$

The benchmark budgets are $B_0 = 0.0$, $B_1 = 1.2$, $B_2 = 2.5$, and $B_3 = 4.1$, under the default costs $c_{\mathrm{ED}} = 1.0$, $c_{\mathrm{HOSP}} = 1.5$, $c_{\mathrm{POS}} = 1.2$, and $c_{\mathrm{SRC}} = 0.4$. The lattice teacher applies Equation (5) to each complete record; hard stop/query actions and soft preferences p_T from Equation (6) are cached for distillation. Deployment uses the learned policy and terminal heads.

D.4 OBSERVED-HISTORY DISTILLATION AND DEPLOYMENT

The student input $s(H_{t,S})$ combines encoder moments, the optional-group mask, the task embedding, budget-cost features, and a reliability summary. For i16/e64 its dimension is

$$d_\pi = \underbrace{2 \times 16}_{\text{moments}} + \underbrace{4}_{\text{mask}} + \underbrace{16}_{\text{task}} + \underbrace{4}_{\text{budget/cost}} + \underbrace{4}_{\text{reliability}} = 60. \tag{51}$$

The residual policy MLP has width 128, one residual block, and dropout 0.10. Its heads produce five query/stop logits for π_η and a terminal answer probability h_ω, as in Section 4.3. Repeated or unavailable queries are masked, stop is always valid, and the full subset terminates.

Student optimization. The policy and terminal head use 1,000 gradient steps, batch size 64, and Adam learning rate 10^{-3}. They minimize Equation (7), combining cross-entropy against soft preferences p_T with terminal answer-head BCE. Equation (6) smooths near-tied teacher costs. Each update draws a fresh mask from $w = \mathrm{Unif}(2^{\mathcal{M}})$ independently of the record and multiplier and uses the corresponding cached target. Validation and test metrics retain the original episode distribution. Table 8 collects the training settings.

Table 8: **SLD training settings.** Network settings describe the masked encoder and student policy, respectively.

Setting	Shared subset evaluation	Observed-history distillation
Hidden width	128	128
Residual blocks	2	1
Dropout probability	0.10	0.10
Gradient steps	1,000	1,000
Batch size	64	64
Optimizer	Adam	Adam
Learning rate	10^{-3}	10^{-3}
KL coefficient	$\beta = 0.02$	—
Mask selection	Empty/partial/full: $0.10/0.30/0.60$	Repeated uniform draws from $2^{\mathcal{M}}$
Model selection	Development checkpoint	Development calibration

Deployment calibration. The development search for the action-logit correction uses the grid

$$\mathcal{C} = \{0, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.2\}. \tag{52}$$

The selected adjustment is fixed before evaluation; the calibrated policy uses only acquired evidence.

D.5 HARDWARE AND EXECUTION

Trainable neural components were optimized on NVIDIA RTX PRO 6000 Blackwell PCIe GPUs and NVIDIA A100 80 GB PCIe GPUs. Outputs of the frozen E5-Mistral text encoder are cached and reused. Neither this text encoder nor the local ReAct text model is fine-tuned; the masked Gaussian encoder, moment decoder, and student controller are trainable.

D.6 TEMPORAL PROTOCOL AND REPORTED EXPENDITURE

The benchmark uses fixed state–week cutoffs and chronological partitions. Queries expose only evidence available by the episode cutoff; record construction is fixed before fitting. Development data select operating points, and evaluation reports realized student expenditure.

The multiplier λ weights evidence cost in Equation (4). Appendix B.7 gives the corresponding cost-monotonicity result for exact optimizers.