EARLY WARNING OF PERSONALIZATION DEGRADATION FROM USER HISTORIES AND HIDDEN STATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalization encoders continually update hidden representations of evolving user interactions to guide recommendations or user-specific generation. Future quality can degrade even while recent output quality appears normal, limiting timely intervention. We ask whether user interaction history and the corresponding encoder states form a joint record that exposes warning evidence of such degradation earlier. We call this relation *Temporal State-History Matching* (TSHM). We establish conditions under which a future-quality-relevant change remains visible in the observed state component of this record while recent utility responds only weakly. We then propose `CoTRAC`, a monitoring framework in which a pluggable sentry learns a broad History-State warning and refines it through Regularity, which measures local organization of projected state motion, and Transition, which compares the latest update with a causal prediction. We evaluate dataset-specific subsets of eight sentry designs on MovieLens-1M, PENS, and Amazon Product Reviews, monitoring 13 frozen personalization host instances. Against matched recent-utility controls, History-State warning improves mean normalized area under the precision-recall curve (nAUPRC) by 0.11 on MovieLens and is higher in all 160 encoder-sentry-horizon settings. The same direction occurs in $67/80$ PENS and $21/24$ Amazon settings. Regularity and Transition affect event detection, warning lead, and false alarms differently, and their joint use does not systematically dominate the individual views. Source-only warnings retain 86.4% of self-warning discrimination on MovieLens and 83.6% on PENS when applied to held-out encoders. These results establish History-State evidence as a prospective warning source beyond recent utility while showing that finer structured effects are more context-dependent.

1 INTRODUCTION

Personalization systems use a user's past interactions to decide what to recommend or generate next. A movie recommender may use previously watched films to rank new candidates, while a text generator may use earlier reading history to adapt its response. These systems encode the growing history into a learned hidden *preference state* that guides future outputs, although no gold state trajectory directly establishes its correctness (Kang & McAuley, 2018; An et al., 2019; Ao et al., 2021). We usually monitor them through output quality, such as ranking accuracy or whether generated content matches the user's preference. This signal arrives only after the hidden state has already influenced the downstream output. Future task-specific utility can therefore decline while immediate performance still appears normal. We call such a future decline degradation. This leads to our central question: *can the evolving relation between user interactions and encoder states warn of future degradation before it becomes visible in output quality?*

Related methods answer nearby questions. Sequential and personalized models learn preference states from evolving histories to improve recommendation or generation (Kang & McAuley, 2018; An et al., 2019; Li et al., 2020; Du et al., 2023). Failure prediction and uncertainty quantification assess whether a current output should be trusted (Corbière et al., 2019; Vlachou, 2025). Temporal event-prediction and time-to-event methods use past observations to predict later events (Du et al., 2016; Lee et al., 2020). Representation analyses compare internal states across models or test whether they can be functionally interchanged (Kornblith et al., 2019; Bansal et al., 2021). What remains largely unexplored is whether a user's evolving interaction history together with the corresponding

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.

encoder states produced carries prospective evidence of later personalization degradation that recent utility misses. A change in user preference is not necessarily degradation, and a warning about current output quality does not necessarily warn of a future decline. We call this relation ***Temporal State-History Matching (TSHM)***.

We represent TSHM as an ordered joint record without requiring user interactions and hidden states to share a vector space or be compared through a direct distance. The theory considers a local change in the causal user history that affects expected future utility, while recent utility is initially insensitive to that change. If the observed encoder states also change in response to that same underlying history perturbation, the History-State record can expose the change before immediate utility reacts comparably. Under these conditions, a perturbation of size $\epsilon \to 0$ can change the observed state component and expected future utility at first order in ϵ, while recent utility changes only at second order or less. Hence, History-State evidence can reveal a future-quality-relevant change before its effect becomes clearly visible in immediate quality. Whether a learned monitor can exploit this theoretical separation remains an empirical question. We study it through CoTRAC (**Co**rrespondence of **T**emporal **R**epresentations **A**nd **C**ausal History), which monitors already-trained, frozen personalization encoders. CoTRAC defines the History-State evidence exposed to a pluggable learned monitor, or sentry, rather than fixing one warning model. The sentry converts this evidence into an early-warning score, allowing different monitor designs to use the same frozen host. Its broad History-State (H) signal captures how interaction history and encoder states evolve together. Regularity (R) checks whether nearby projected-state updates remain locally organized. Transition (T) checks whether the latest projected-state update agrees with what its causal past predicts.

We evaluate CoTRAC on MovieLens-1M for movie recommendation, PENS for news recommendation and personalized headline generation, and Amazon Product Reviews for product recommendation. These settings cover multiple domains and both recommendation and generation. We monitor 13 frozen personalization encoders (hosts) using dataset-specific subsets of eight pluggable sentry designs spanning different temporal and risk biases. We first compare History-State warning against matched recent-utility controls. On MovieLens-1M, H improves mean normalized AUPRC (nAUPRC) by 0.11 and outperforms the utility control in all 160 matched encoder-sentry-horizon settings. The advantage also recurs on PENS and Amazon, indicating warning evidence that recent quality misses. We next examine what R and T add to this broad warning. We measure event coverage, warning lead, and false alarms separately because these properties need not improve together. On PENS, the refinements increase event coverage on average while warning lead changes little or slightly decreases. On Amazon, they warn earlier on average and also raise more false alarms. Their joint RT view does not systematically dominate the individual views. Finally, we test reuse across encoders and tasks. Source-only warning rankings retain 86.4% of self-warning discrimination on held-out unseen MovieLens encoders and 83.6% on PENS encoders. With the preference backbone fixed across recommendation and generation, H exceeds the task-specific utility control in nine of ten comparisons. Three tested open-weight large language models also exhibit positive warning discrimination before their first generated token.

Contributions. We formulate TSHM and establish conditions under which future-quality-relevant change can remain visible in the observed History-State record while recent utility responds more weakly. We develop CoTRAC as a common monitoring framework for frozen personalization hosts through a pluggable sentry and H, R, and T views, without requiring user interactions and hidden states to share coordinates or retraining the host. Across recommendation and personalized generation, we show warning beyond recent utility, distinct coverage, earliness, and false-alarm behavior, and substantial reuse across unseen encoders and downstream consumers.

2 Related Work

Preference Evolution And Drift. Personalized systems model evolving histories through recurrent, self-attentive, time-aware, and frequency-sensitive architectures (Hidasi et al., 2016; Kang & McAuley, 2018; Li et al., 2020; Du et al., 2023). Related work captures preference change through long- and short-term states, dynamic embeddings, shift modeling, and drift-aware representations (An et al., 2019; Kumar et al., 2019; Wang et al., 2024; Lin et al., 2025). Personalized generation conditions outputs on learned preference histories and injects them during decoding (Ao et al., 2021; Song et al., 2023; Salemi et al., 2024; Chatterjee et al., 2026a), while change-point and temporal-

graph methods characterize behavioral shifts and disengagement (Eskandanian & Mobasher, 2018; Baek & Lee, 2026). Preference change or state irregularity does not itself establish future degradation in a trained model's task utility.

Failure Prediction and Internal-State Monitoring. Reliability methods estimate output correctness or uncertainty through calibration, confidence, failure prediction, and selective prediction (Guo et al., 2017; Hendrycks & Gimpel, 2017; Corbière et al., 2019; Geifman & El-Yaniv, 2017; Lakshminarayanan et al., 2017; Zhu et al., 2024). Recommender studies examine undesirable behavior, delayed-outcome surrogates, and conversational failures (Chia et al., 2022; Wang et al., 2022; Vlachou, 2025). Internal-state probes anticipate language-model behavior or hallucination before generation (Ashok & May, 2025; Alnuhait et al., 2025). We instead test whether a frozen personalization model's evolving interaction-state relation warns of a later decline in its own task utility beyond recent utility.

Prospective Outcome Prediction. Survival and early-event methods predict delayed outcomes from longitudinal evidence (Lee et al., 2018; Ren et al., 2019; Maystre & Russo, 2022; Yèche et al., 2023), and early-warning research identifies observable changes preceding adverse transitions (Scheffer et al., 2009). These methods motivate prospective prediction without establishing the additional warning value of evolving hidden preference states over recent output utility.

Representation Analysis And Warning Reuse. Representation analyses compare internal spaces through correlation, kernel alignment, probing, and causal interventions (Raghu et al., 2017; Kornblith et al., 2019; Hewitt & Manning, 2019; Geiger et al., 2021). Recommendation research examines transferable representations through pretraining, continual learning, shift robustness, and invariance (Hou et al., 2022; Wang et al., 2024; Choi et al., 2025; Liao et al., 2025). Neither line establishes whether prospective warning rankings remain informative for an unseen encoder or task.

3 Temporal State-History Matching And Early Warning

Personalization interface and warning-time information. Personalization systems produce user-specific outputs by encoding preferences from the evolving history. At each timestep t, an action a_t on content c_t forms interaction $b_t = \langle a_t, c_t \rangle$ and causal history $\mathcal{H}_{\leq t} = \langle b_1, \ldots, b_t \rangle$. A frozen preference encoder produces the hidden state $z_t = \mathcal{M}_\theta(\mathcal{H}_{\leq t})$, and the task head uses z_t together with the downstream query or task context q_{t+1} to produce $\widehat{y}_{t+1} = \mathcal{G}_\phi(z_t, q_{t+1})$. The output is evaluated by task-specific utility $u_{t+1} = \mathcal{U}(\widehat{y}_{t+1}, y^\star_{t+1})$. A future decline in utility over horizon h ($u_{t+1:h}$) is defined as degradation $Y_{t,h} = \mathcal{D}_h(U_{>t})$. A warning issued at t may use only information available through t; future utility supplies supervision and evaluation outcomes, not warning-time evidence.

Temporal State-History Matching. Rather than monitoring the utility produced at each timestep, we monitor how the interaction history and the corresponding hidden states evolve together over time. This relation provides the basis for prospective degradation warning. Over the W observations ending at t, the ordered interaction-state record is

$$\mathcal{C}_t^{(W)} = \langle (b_\tau, z_\tau) \rangle_{\tau=t-W+1}^{t}, \tag{1}$$

with truncation for shorter histories. We call this relation *Temporal State-History Matching* (TSHM). The host-produced state is learned, not a gold representation of user preference; TSHM therefore treats the temporally paired interaction–state response as the monitoring object rather than either marginal alone, without requiring shared coordinates or a direct distance (Appendix A.4).

Prospective separation from recent utility. Although recent utility can be predictive, the interaction-state record may expose a future-relevant change that recent utility does not yet reveal. Formally, consider a reference causal history and a nearby admissible variation of size ϵ, both processed by the same frozen model. Under the local smoothness and utility-blind-direction conditions of Theorem D.2, and the state-response condition of Corollary D.3, there exists a local direction

$$\left\| \mathbf{u}_{t,\epsilon}^{(m)} - \mathbf{u}_{t,0}^{(m)} \right\|_2 \leq \frac{L_U}{2}\epsilon^2, \quad \bar{u}_{t,h}^{(0)} - \bar{u}_{t,h}^{(\epsilon)} \geq \frac{a}{2}\epsilon, \quad d_{\mathcal{C}}\left(\mathcal{C}_{t,\epsilon}^{(W)}, \mathcal{C}_{t,0}^{(W)}\right) \geq \lambda_Z c_Z \epsilon. \tag{2}$$

Thus, along the stated direction, *recent utility changes at most to second order, while expected future utility decreases at first order and the observed state component separates at first order*. This conditional local observability result neither guarantees such a direction for every trajectory nor implies that the interaction-state relation is incorrect, that interactions add information beyond states, or that a learned warning predicts $Y_{t,h}$. Appendix D.2 gives the assumptions and proofs.

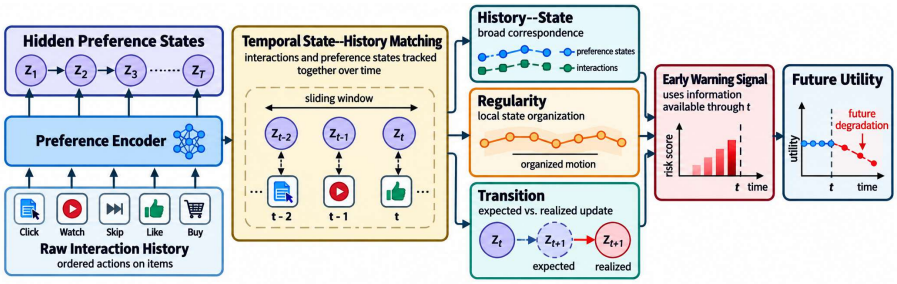


Figure 1: **Overview of CoTRAC.** A frozen personalization encoder maps the evolving user history to hidden preference states. CoTRAC learns a prospective History-State warning (H) from temporal relations, then refines that warning through local state-motion Regularity (R) and agreement of the newest update with its causal prediction (Transition, T). Future task utility defines the degradation outcome and is unavailable at warning time. The full implementation appears in Figure 4.

Three monitoring views. We decompose prospective warning via TSHM into three core components. History-State (H) asks whether the evolving interaction-state record carries broad prospective warning. Regularity (R) asks whether the realized local host-state trajectory remains organized, while Transition (T) asks whether the newest host-state update agrees with a causal prediction formed before that update is observed. Unlike H, both R and T operate only on the host-state trajectory and are used to refine the broad History-State warning. Appendix A specifies the complete frozen-host interface and causal monitoring boundary. Section 4 implements R and T on a common projected monitoring path, while TSHM itself remains defined on the host interaction-state record. Figure 1 summarizes these relationships in the CoTRAC pipeline.

4 Instantiating TSHM With CoTRAC

CoTRAC tests whether a learned sentry can convert the causal History-State evidence in Section 3 into prospective warning for a frozen personalization model. History-State (H) learns the broad warning, while Regularity (R) and Transition (T) refine it through structured views of state evolution, separately or through their combined interaction RT.

History-State Sentry (Monitor). The sentry H converts the available interaction history and recent host states into an internal warning representation, which is then mapped to horizon-specific warning scores. For $Z_t^{(W)} = \langle z_\tau \rangle_{\tau=t-W+1}^{t}$ being the recent host encoder-state window, each sentry receives only the causal interaction history and corresponding states already produced by the frozen host. Architecture-specific projections, masks, and temporal organization belong to the warning model and do not require interactions and states to share coordinates. For sentry architecture m,

$$
\begin{aligned}
r_t^{H,m} &= \mathcal{B}_m\Big(\Pi_m(\mathcal{H}_{\le t}, Z_t^{(W)})\Big), \\
\ell_t^{H,m} &= \Gamma_m\Big(\mathcal{R}_m(r_t^{H,m})\Big), \qquad s_{t,h}^{H,m} = \sigma(\ell_{t,h}^{H,m}).
\end{aligned}
\tag{3}
$$

The learned sentry representation $r_t^{H,m}$ is distinct from the host preference state z_t being monitored. The eight sentries differ in temporal processing and risk parameterization while preserving the same causal evidence and degradation targets. Appendix E gives their exact realizations.

Structured state evidence. To characterize how the host states evolve locally, Regularity summarizes whether the realized state trajectory remains organized, whereas Transition measures whether the newest update agrees with what its preceding trajectory predicts. Both use a common 128-dimensional projected monitoring state ζ_t derived from z_t. This projection standardizes their dimensional interface rather than aligning hosts coordinate by coordinate. Regularity measures speed, directional, acceleration, and discrete-jerk deviations in the realized projected trajectory. Its precursor predicts four bounded satisfaction quantities whose conjunction $s_t^R = \prod_{k=1}^{4} p_{k,t}^R$ is high only when all four conditions remain well satisfied. Transition instead predicts the newest projected-state displacement from earlier displacements while holding that newest update out of the causal predictor. After

the update is observed, its uncertainty-normalized prediction residual r_t^T gives $s_t^T = \exp(-r_t^T)$. Neither Regularity nor Transition is itself degradation. Appendix B.2 and Appendix B.3 give the implementation-exact features, precursor objectives, scaling, and validity rules.

Joint evidence & refinement. Regularity and Transition are projected separately to $\widetilde{\mathbf{R}}_t, \widetilde{\mathbf{T}}_t \in \mathbb{R}^{32}$. Their joint representation retains both views, simultaneous activation, and disagreement:

$$\mathbf{f}_t^{RT} = \left[\widetilde{\mathbf{R}}_t; \widetilde{\mathbf{T}}_t; \widetilde{\mathbf{R}}_t \odot \widetilde{\mathbf{T}}_t; \left|\widetilde{\mathbf{R}}_t - \widetilde{\mathbf{T}}_t\right|\right], \qquad \mathbf{f}_t^R = \widetilde{\mathbf{R}}_t, \quad \mathbf{f}_t^T = \widetilde{\mathbf{T}}_t. \tag{4}$$

Thus, RT is an interaction between R and T, not a fourth state property. After H is trained, it remains fixed. For $A \in \{R, T, RT\}$, an adapter combines the frozen sentry representation with structured evidence:

$$\begin{aligned}(\boldsymbol{\Delta}_t^A, g_t^A) &= \mathcal{A}_A(r_t^H, \mathbf{f}_t^A), \qquad \alpha_t^A = \sigma(g_t^A), \\ \ell_{t,h}^{H+A} &= \ell_{t,h}^H + \alpha_t^A \Delta_{t,h}^A.\end{aligned} \tag{5}$$

The scalar gate controls correction strength and the signed residual determines its horizon-specific direction, so irregularity or transition surprise is not assigned a fixed harmful direction. Appendices B.5– B.7 give the exact construction.

Training hierarchy. The host encoder and task head remain frozen throughout. Each H sentry first learns prospective degradation from causal prefixes and is then frozen. Regularity and Transition are trained separately from causal projected-state trajectories without future utility or future-degradation labels, using distinct precursor objectives $\mathcal{L}_{\mathrm{pre},R}$ and $\mathcal{L}_{\mathrm{pre},T}$, and are then frozen. The final adapters learn from the common prospective warning objective together with mechanism-aware gate supervision $\mathcal{L}_R$ and $\mathcal{L}_T$ and an anchor to the frozen H objective. These gate losses are distinct from the precursor losses. The resulting sequence is therefore H warning training $\rightarrow$ R/T precursor training without future-degradation supervision $\rightarrow$ freezing $H, R, T \rightarrow$ structured refinement. Appendices B.8–B.11 give the complete objectives and parameter-freezing specification.

5 Evaluation

Tasks and datasets. We evaluate movie recommendation on MovieLens-1M (Harper & Konstan, 2015), news recommendation and personalized headline generation on PENS (Ao et al., 2021), and product recommendation on Amazon Reviews (Hou et al., 2024). Utility is nDCG for MovieLens, MRR for PENS recommendation and Amazon, and ROUGE-L for PENS generation. RQ1 and RQ2 use the recommendation settings, while RQ3 studies source-only encoder reuse, the PENS recommendation-to-generation task change, and a separate pre-generation interface using Mistral-7B (Jiang et al., 2023), Phi-4-mini (Abouelenin et al., 2025), and Qwen3-8B (Yang et al., 2025).

Training, test, and leakage boundary. Each personalization host is first trained on its task-training partition and then frozen. Warning training never updates the host encoder or task head. Frozen hosts then produce separate warning-training and held-out warning pools with disjoint users: MovieLens uses 35K/15K trajectory instances, and PENS and Amazon each use 10K/2K. Warning parameters are fitted only on warning-training trajectories, and operating thresholds are fixed from pre-test information before TEST evaluation. A warning at time t uses only interactions and states available through t. Future utility is revealed only afterward to form supervision or held-out outcomes. Regularity and Transition are learned from training-side projected-state trajectories without future-utility or future-degradation labels and are frozen before structured warning refinement. Appendix F gives the complete split, eligibility, calibration, and checkpoint protocol.

Frozen encoders and sentries. The primary MovieLens hosts are FEARec (Du et al., 2023), GRU4Rec (Hidasi et al., 2016), LRURec (Yue et al., 2024), Mamba4Rec (Liu et al., 2024), and SASRec (Kang & McAuley, 2018). PENS uses GTP (Song et al., 2023), IMPerSumm (Chatterjee et al., 2026c), PENS-NAML-T1 (Wu et al., 2019a), PerDucer (Chatterjee et al., 2026b), and Walk2Pers (Chatterjee et al., 2026a), with the same five backbones later examined under recommendation and generation. Amazon uses Caser (Tang & Wang, 2018), GRU4Rec, and SR-GNN (Wu et al., 2019b). MovieLens uses eight sentries, namely Prefix-MLP, TCN-ELECTS-EW, LS-EEP, TCSA, TCSR, DRSA-style, DeepHit-style, and Temporal-Prior. PENS uses LS-EEP, Prefix-MLP, TCN-ELECTS-EW, and TCSR, while Amazon uses Prefix-MLP and TCN-ELECTS-EW. Sentries monitor frozen hosts rather than replace them. Appendix E gives their exact realizations.

Table 1: **RQ1: History-State warning beyond utility.** H and U share degradation targets and held-out observations. Results average over architectures and horizons. MovieLens and PENS report mean $\pm$ SD from paired user-level bootstraps. "$H > U$" counts architecture-horizon comparisons with higher H AUPRC.

Dataset / Task	Preference model	U nAUPRC	H nAUPRC	Δ nAUPRC	Δ AUPRC	$H > U$
MovieLens-1M *Movie recommendation*	FEARec	0.016±0.007	0.071±0.005	+0.055	+0.048	32/32
	GRU4Rec	0.107±0.004	0.236±0.008	+0.129	+0.097	32/32
	LRURec	0.062±0.005	0.181±0.011	+0.119	+0.095	32/32
	Mamba4Rec	0.081±0.003	0.216±0.006	+0.135	+0.107	32/32
	SASRec	0.047±0.004	0.141±0.015	+0.095	+0.083	32/32
	MovieLens mean	**0.062**	**0.169**	**+0.106**	**+0.086**	**160/160**
PENS *News recommendation*	GTP	0.014±0.003	0.019±0.01	+0.005	+0.003	11/16
	IMPerSumm	0.029±0.004	0.057±0.007	+0.027	+0.017	16/16
	PENS-NAML-T1	0.023±0.006	0.037±0.005	+0.014	+0.007	14/16
	PerDucer	0.020±0.002	0.024±0.005	+0.004	+0.003	10/16
	Walk2Pers	0.028±0.004	0.056±0.003	+0.028	+0.016	16/16
	PENS mean	**0.023**	**0.039**	**+0.016**	**+0.009**	**67/80**
Amazon *Product recommendation*	Caser	-0.026	-0.011	+0.015	+0.004	8/8
	GRU4Rec	-0.004	0.005	+0.009	+0.004	6/8
	SR-GNN	-0.040	-0.020	+0.020	+0.009	7/8
	Amazon mean	**-0.023**	**-0.009**	**+0.015**	**+0.006**	**21/24**

Prospective targets and matched RQ1 controls. We evaluate horizons $h \in \{1, 2, 3, 5\}$. MovieLens defines degradation from a future nDCG drop of at least 0.1 with one guarded timestep before the earliest credited onset. Amazon uses the analogous guarded MRR drop. PENS uses task-specific future outcomes from MRR or ROUGE-L. Within every RQ1 cell, History-State warning H and recent-utility warning U use the same frozen host, target, horizon, eligible held-out positions, and validity mask. MovieLens and Amazon use architecture-matched H and U sentries. PENS uses one common utility-warning architecture selected from four variants using MovieLens TEST performance, fixed before testing on PENS, and retrained from scratch on PENS. MovieLens TEST was not used for any other warning-model, threshold, hyperparameter, or checkpoint selection.

Structured and source-only comparisons. RQ2 compares each $H+R$, $H+T$, and $H+RT$ warning only with its frozen H parent under the same host, sentry, horizon, and held-out population. RQ3 uses leave-one-model-out source-only reuse: the remaining encoders' warning scores are independently rank-normalized on common evaluation indices and averaged. The held-out encoder contributes no warning parameters, hidden-state coordinates, or warning scores to this consensus. Its own warning is used only as the self-warning reference. In the PENS task-change experiment, the preference backbone is fixed while recommendation and generation use their own task-specific outcomes. In the separate open-weight LLM setting, the monitored state is read after the available user and task context has been processed and before the first autoregressive output token.

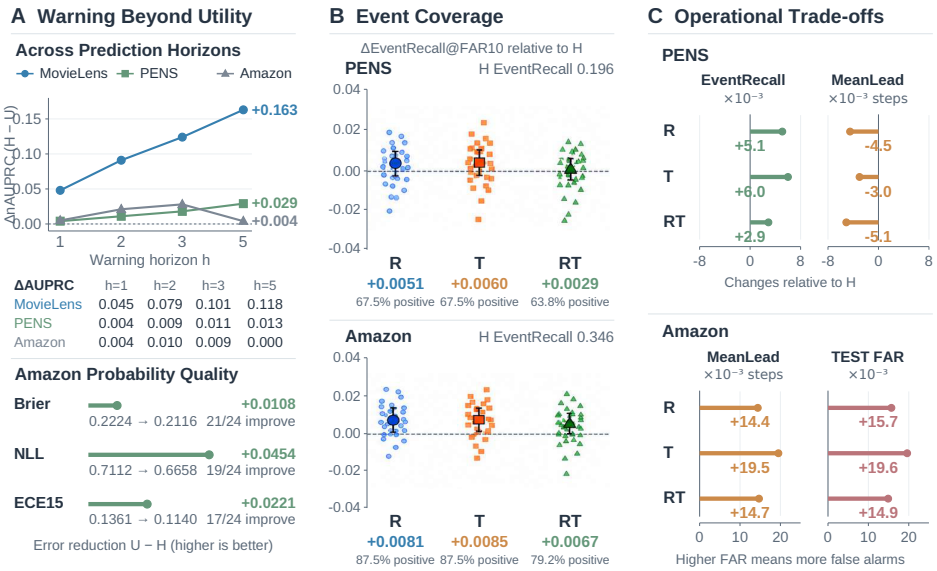
Warning criteria. Discrimination is measured by area under the precision-recall curve (AUPRC) and $\mathrm{nAUPRC} = (\mathrm{AUPRC} - \pi)/(1 - \pi)$, where π is degradation prevalence in the matched evaluation population. EventRecall@FAR10 measures the fraction of degradation episodes receiving a valid pre-onset alarm, MeanLead@FAR10 measures advance notice among detected episodes, and realized TEST FAR records the actual false-alarm burden at the pre-test operating point. We additionally report Brier score, binary negative log-likelihood, and ECE15 where probability quality is analyzed. All method differences are formed within matched cells before aggregation. Paired user-level bootstraps quantify held-out-user uncertainty where reported. Appendix F gives exact metric denominators and aggregation rules.

6 Results & Analysis

The empirical findings address three questions. RQ1: Does History-State H evidence add prospective information beyond recent utility? RQ2: How do Regularity R and Transition T change an established History-State warning? RQ3: What remains reusable when the encoder, downstream task, or consumer changes? Host, sentry, horizon, and operating-point results are reported in Appendix G.

6.1 RQ1. History-State Warning Beyond Recent Utility.

History-State evidence adds prospective discrimination beyond recent utility. Under matched hosts, targets, horizons, and held-out



(D) $H > U$ By Sentry

Dataset	Evaluated Sentry	$H > U$
MovieLens	All eight sentries	**20/20 each**
PENS	Prefix-MLP	14/20
	TCN-ELECTS-EW	20/20
	LS-EEP	18/20
	TCSR	15/20
Amazon	Prefix-MLP	12/12
	TCN-ELECTS-EW	9/12

(E) Joint Refinement By Horizon

Horizon	$RT > \max(R, T)$	Joint Gap $\times 10^3$
$h = 1$	12/40	-1.429
$h = 2$	13/40	-1.090
$h = 3$	16/40	-1.107
$h = 5$	16/40	-0.929
All	57/160	-1.139

Figure 2: **History-State warning and structured refinement.** **(A)** Horizon-wise $H - U$ and Amazon probability-quality reductions. **(B)** EventRecall changes, including individual cells and means. **(C)** PENS coverage/lead and Amazon lead/FAR changes. Higher FAR is worse. **(D)** Positive $H > U$ counts across hosts and horizons for evaluated sentries. The MovieLens row summarizes $20/20$ for each of eight sentries. **(E)** MovieLens $RT - \max(R, T)$ nAUPRC changes, averaged across cells.

positions, MovieLens mean nAUPRC rises from 0.06 for the recent-utility control U to 0.17 for H, a gain of 0.11 (Table 1). All $160/160$ host-sentry-horizon comparisons favor H. The same direction recurs on PENS in $67/80$ comparisons and on Amazon in $21/24$, with mean gains of approximately 0.02 in both domains.

The sentry breakdown in Figure 2D locates the exceptions. TCN-ELECTS-EW has $20/20$ positive PENS comparisons, whereas all three Amazon exceptions occur with that sentry. The advantage also persists across future horizons. In Figure 2A, mean $H - U$ remains positive at every evaluated horizon. MovieLens increases from 0.05 nAUPRC at $h = 1$ to 0.16 at $h = 5$, PENS from 0.004 to 0.029, while Amazon peaks at 0.028 at $h = 3$ before falling to 0.004 at $h = 5$. Thus, the prospective advantage is not confined to the nearest admissible event, although its horizon profile is not universal.

Additional information does not imply uniformly strong absolute prediction. Amazon has mean H nAUPRC of -0.009 versus -0.023 for U. Thus, H improves relative discrimination while remaining below the prevalence reference on average. The matched Amazon probability-quality results in Figure 2A also reduce Brier score, negative log-likelihood, and ECE15 by 0.011, 0.045, and 0.022. The RQ1 claim is therefore additional prospective information rather than universally strong degradation prediction.

Appendix D.2 gives conditions under which future-relevant state evidence can remain locally visible while recent utility is less responsive. The experiments do not verify those assumptions causally. They establish that learned History-State warnings repeatedly exploit information missed by the utility-based warning under matched architectures. Appendix G.2 gives the full matched evidence and low-support sensitivity cases.

Table 2: **Non-additive Regularity and Transition effects.** Mean nAUPRC changes relative to frozen H, scaled by 10^3 ($H = 0$). Bold marks the strongest of R, T, and RT. Joint gain is $\Delta RT - \max(\Delta R, \Delta T)$. Best Overall includes H.

Dataset	Preference Model	ΔR	ΔT	ΔRT	$\Delta RT - \max(\Delta R, \Delta T)$	Best Structured	Best Overall
MovieLens	FEARec	**+2.449**	+0.059	+1.397	-1.052	R	R
	GRU4Rec	+0.790	+0.558	**+2.480**	+1.690	RT	RT
	LRURec	-0.534	+0.376	**+0.433**	+0.057	RT	RT
	Mamba4Rec	+1.756	+1.064	**+1.940**	+0.184	RT	RT
	SASRec	**+2.086**	-0.538	+0.556	-1.530	R	R
PENS	GTP	+4.534	**+4.667**	+4.603	-0.064	T	T
	IMPerSumm	+2.075	**+2.121**	+1.719	-0.402	T	T
	PENS-NAML-T1	**+2.708**	+2.226	+2.437	-0.271	R	R
	PerDucer	+0.374	**+0.951**	+0.300	-0.651	T	T
	Walk2Pers	+3.404	**+3.535**	+2.535	-1.000	T	T
Amazon	Caser	-5.710	**-2.252**	-9.274	-7.021	T	**H**
	GRU4Rec	**+1.225**	-2.192	+1.078	-0.147	R	R
	SR-GNN	**+2.916**	+1.503	+0.181	-2.735	R	R

6.2 RQ2. Conditional Effects Of Regularity And Transition. Structured evidence changes warning behavior conditionally rather than defining a universal winner. Relative to each model's frozen H parent, all three refinements increase mean EventRecall on PENS and Amazon (Figure 2B). The largest mean gains come from T, at 6.0×10^{-3} on PENS and 8.5×10^{-3} on Amazon, with positive changes in a clear majority of matched cells.

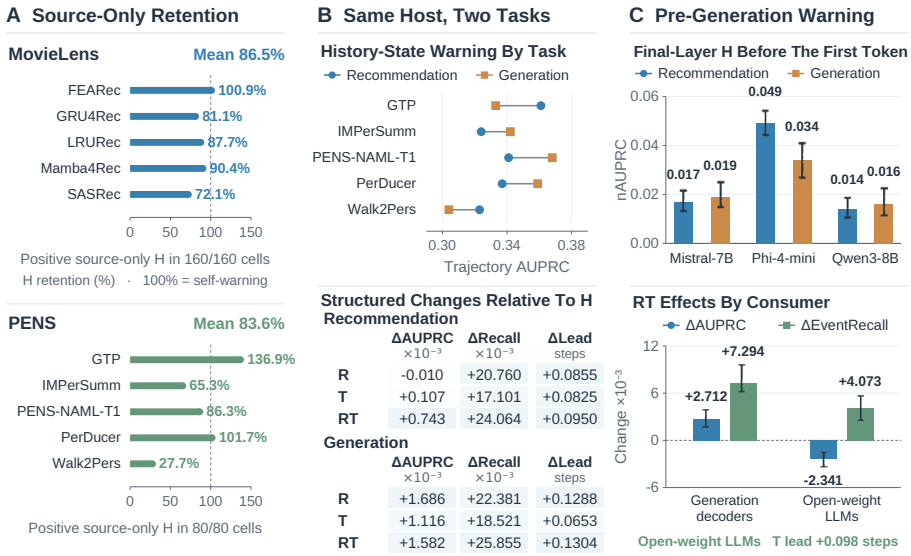
Event coverage, warning lead, and false-alarm burden separate across domains. PENS coverage gains accompany slightly negative mean lead changes of roughly $(3 \text{ to } 5) \times 10^{-3}$ steps in magnitude. On Amazon, MeanLead increases by approximately 0.01 to 0.02 steps while realized TEST FAR rises by a similar amount (Figure 2C). Detecting additional events therefore need not mean detecting them earlier, and earlier warning can carry greater alarm burden, consistent with Proposition D.10.

Regularity and Transition are not uniformly additive. In Table 2, RT is strongest for three of five MovieLens hosts. It exceeds both individual refinements in only $57/160$ MovieLens cells and is 1.139×10^{-3} nAUPRC below the stronger individual view on average. The within-cell joint gap remains negative at every MovieLens horizon (Figure 2E). On PENS, T leads for four backbones, R for one, and RT for none. On Amazon, unmodified H remains best overall for Caser, while GRU4Rec and SR-GNN favor R. Structured evidence is therefore useful conditionally, and its joint contribution need not exceed the individual views. This is the empirical counterpart of the conditional-refinement and non-additivity results in Appendix D.3. Appendix G.3 reports the full matched results.

6.3 RQ3. Reuse Across Encoders, Tasks, And Consumers. Since TSHM does not require hidden-state coordinates to align across hosts, we next ask whether the resulting broad warning remains informative when the encoder changes. We find that broad History-State warning (H) remains informative for unseen encoders. Source-only reuse uses only the remaining encoders' rank-normalized warning scores. The held-out encoder contributes no warning score, parameter, or hidden-state coordinate. In Figure 3A, transferred H nAUPRC is positive in all $160/160$ MovieLens cells and all $80/80$ PENS cells, retaining 86.45% and 83.58% of self-warning nAUPRC on average. Retention above 100% for some targets means that the source consensus ranks those outcomes better than the target's own warning. It is not a fraction of model parameters transferred.

The finer structured effects are less stable across encoders. All three transferred effects have negative means and fewer than half positive cells on MovieLens, compared with positive means and approximately two-thirds positive cells on PENS (Figure 3D). Broad-warning reuse and structured-effect reuse are therefore distinct empirical properties. Proposition D.17 and Corollary D.18 give one representation-level reason for this distinction without establishing coordinate geometry as its cause.

The warning profile also changes with the downstream task or consumer. With the same five PENS preference backbones under recommendation and headline generation, H exceeds its task-specific utility control in nine of ten backbone-task comparisons (Figure 3B). Walk2Pers generation is the only exception, with H and U AUPRC of 0.304 and 0.306. Under generation, all three structured variants increase AUPRC, EventRecall, and MeanLead, whereas HR and HT change recommendation AUPRC only marginally. In Figure 3C, all three tested open-weight LLMs have positive pre-generation H nAUPRC under both tasks, while structured criteria again separate. Among



(D) Structured Reuse Relative To Source-Only H

Domain	R	T	RT
MovieLens	-1.737 (43.8%)	-2.005 (39.4%)	-1.483 (47.5%)
PENS	$+1.473$ (66.2%)	$+1.400$ (70.0%)	$+1.649$ (68.8%)

Figure 3: **Warning reuse.** **(A)** Source-only H retention. **(B)** Task changes with five fixed PENS backbones. **(C)** Pre-token warning and consumer-specific refinements, with lead stated separately in steps. **(D)** Mean structured changes relative to source-only H, in nAUPRC $\times 10^3$, with positive-cell percentages in parentheses. MovieLens and PENS use 160 and 80 target-sentry-horizon cells.

pre-test-admitted OpenLLM RT runs, AUPRC decreases by 2.341×10^{-3} while EventRecall increases by 4.073×10^{-3}, and Transition increases MeanLead by 0.098 steps. Broad H therefore remains observable across these interfaces, while finer structured effects are more context-dependent.

Theorem D.13 in Appendix D.5 provides a conditional bound for nested warning signatures under a common degradation target. It does not apply directly to the task- or consumer-change experiments, predict the observed retention percentages, or determine the sign of transferred structured effects. Appendix G.4 reports the complete analyses.

7 Limitations And Future Work

The theory gives local, conditional observability results and does not guarantee accurate warning for every trajectory. The empirical findings establish prospective association rather than a causal mechanism linking state evolution to later degradation. `CoTRAC` is also host-relative: information discarded by the frozen preference representation cannot be recovered by its monitor. Regularity and Transition capture specific local properties of projected state motion rather than a complete account of preference evolution, consistent with their conditional and task-dependent effects. Degradation is defined through task-specific utility, so warning behavior can change with the downstream consumer even when the preference backbone is fixed. Broader model families, tasks, and intervention settings remain to be studied.

8 Conclusion

In this paper, we study whether the evolving interaction-state relation provides prospective warning beyond recent utility. History-State evidence improves matched discrimination, structured refinements alter warning behavior without universal dominance, and broad warning transfers more consistently than finer structure. The theory gives a conditional observability separation. Together, these findings support History-State evidence as prospective monitoring rather than causal diagnosis.

Reproducibility Statement

The appendices specify the monitored interface, causal information boundary, sentry architectures, structured features, objectives, training order, user-disjoint splits, degradation targets, calibration rules, warning metrics, uncertainty procedure, and complete result tables used in the paper. The theorem assumptions and proofs are also included. We additionally provide supplementary ZIP archives containing implementation code for the included experiments. The archives do not contain raw experiment outputs, and reproducing the reported numerical results requires the corresponding datasets and computational resources. We distinguish the availability of implementation code from a complete release of all experimental artifacts.

Ethics Statement

CoTRAC monitors user-interaction histories and internal preference states, which may encode sensitive behavioral information. A deployment should therefore apply the same data-minimization, access-control, retention, and privacy protections used for the underlying personalization system. Warnings are predictive signals rather than causal diagnoses. Acting on them without validation could create unnecessary interventions or unequal service. The reported experiments use existing benchmark data and do not introduce new human-subject data collection. The paper does not claim that the evaluated operating points are suitable for consequential deployment.

AI Use Statement

We used AI to identify relevant literature and edit and polish portions and figures of the manuscript. AI-assisted feedback also supported checks of clarity. The authors take full responsibility for all mathematical arguments, experimental protocols, numerical results, and bibliographic references against the available research materials and primary sources.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, H. Awadalla, Nguyen Bach, Jianmin Bao, A. Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dongdong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Mingcheng Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yun-Meng Li, Chen Liang, Xihui Lin, Zeqi Lin, Meng-Jie Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, V. Mazalov, Ali Mousavi, Anh Nguyen, Jing Pan, D. Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xiaocheng Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, I. Zabir, Jianwen Zhang, L. Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *ArXiv*, March 2025. URL https://www.microsoft.com/en-us/research/publication/phi-4-mini-technical-report-compact-yet-powerful-multimodal-language-models-via-mixture-of-loras/.

Deema Alnuhait, Neeraja Kirtane, Muhammad Khalifa, and Hao Peng. FACTCHECKMATE: Preemptively detecting and mitigating hallucinations in LMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12413–12428. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.663. URL https://aclanthology.org/2025.findings-emnlp.663/.

Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. Neural news recommendation with long- and short-term user representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 336–345. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1033.

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. Pens: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 82–92. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.7.

Dhananjay Ashok and Jonathan May. Language models can predict their own behavior. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2102. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5a9c1af5f76da0bd37903b6f23e96c74-Abstract-Conference.html.

Heon Baek and Eui-Bang Lee. Soft clustering-based detection of behavioral disengagement in temporal graph recommendation. *Expert Systems with Applications*, 2026. doi: 10.1016/j.eswa.2026.132617.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 225–236, 2021.

Parthiv Chatterjee, Asish Batha, Tashvi Patel, Sourish Dasgupta, and Tanmoy Chakraborty. Beyond markovian drifts: Action-biased geometric walks with memory for personalized summarization. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8c696956ee2bc75c51d789562ec39930-Abstract-Conference.html.

Parthiv Chatterjee, Asish Joel Batha, Sourish Dasgupta, and Tanmoy Chakraborty. PerDucer: Keyphrase-driven personalization inducer for summarization from user histories. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20651–20677, San Diego, California, United States, July 2026b. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1035. URL https://aclanthology.org/2026.findings-acl.1035/.

Parthiv Chatterjee, Dhara Jhaveri, Sourish Dasgupta, and Tanmoy Chakraborty. IMPerSumm: Information-modulated user preference modeling for personalized text summarization. In *Proceedings of the ACM Web Conference 2026*, WWW '26, pp. 6774–6785, New York, NY, USA, 2026c. Association for Computing Machinery. ISBN 9798400723070. doi: 10.1145/3774904.3792642. URL https://doi.org/10.1145/3774904.3792642.

Patrick John Chia, Jacopo Tagliabue, Federico Bianchi, Chloe He, and Brian Ko. Beyond NDCG: Behavioral testing of recommender systems with RecList. In *Companion Proceedings of the Web Conference 2022*, WWW, pp. 99–104. Association for Computing Machinery, 2022. doi: 10.1145/3487553.3524215. URL https://doi.org/10.1145/3487553.3524215.

Seungyoon Choi, Sein Kim, Hongseok Kang, Wonjoong Kim, and Chanyoung Park. Dynamic time-aware continual user representation learning. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pp. 740–749. Association for Computing Machinery, 2025. doi: 10.1145/3726302.3729959. URL https://doi.org/10.1145/3726302.3729959.

Charles Corbière, Nicolas THOME, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/757f843a169cc678064d9530d12a1881-Paper.pdf.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2 edition, 2006. doi: 10.1002/047174882X.

Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1555–1564. ACM, 2016. doi: 10.1145/2939672.2939875.

Xinyu Du, Huanhuan Yuan, Pengpeng Zhao, Jianfeng Qu, Fuzhen Zhuang, Guanfeng Liu, Yanchi Liu, and Victor S. Sheng. Frequency enhanced hybrid attention network for sequential recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 78–88, 2023. doi: 10.1145/3539618.3591689.

Farzad Eskandanian and Bamshad Mobasher. Detecting changes in user preferences using hidden markov models for sequential recommendation tasks. In *Proceedings of the Workshop on Intelligent Recommender Systems by Knowledge Transfer and Learning*, 2018. URL https://arxiv.org/abs/1810.00272.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30 of *NeurIPS*, pp. 4878–4887. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

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 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4), 2015. doi: 10.1145/2827872.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, ICLR, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In *International Conference on Learning Representations*, ICLR, 2016. URL https://arxiv.org/abs/1511.06939.

Yupeng Hou, Shanlei Mu, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. Towards universal sequence representation learning for recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD, pp. 585–593. Association for Computing Machinery, 2022. doi: 10.1145/3534678.3539381. URL https://doi.org/10.1145/3534678.3539381.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Wang-Cheng Kang and Julian J. McAuley. Self-attentive sequential recommendation. In *Proceedings of the 2018 IEEE International Conference on Data Mining*, ICDM, pp. 197–206. IEEE, 2018. doi: 10.1109/ICDM.2018.00035. URL https://doi.org/10.1109/ICDM.2018.00035.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, pp. 1269–1278. Association for Computing Machinery, 2019. doi: 10.1145/3292500.3330895. URL https://doi.org/10.1145/3292500.3330895.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30 of *NeurIPS*, pp. 6402–6413. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 2314–2321, 2018. doi: 10.1609/aaai.v32i1.11842.

Changhee Lee, Jinsung Yoon, and Mihaela van der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1):122–133, 2020. doi: 10.1109/TBME.2019.2909027.

Jiacheng Li, Yujie Wang, and Julian McAuley. Time interval aware self-attention for sequential recommendation. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, WSDM, pp. 322–330. Association for Computing Machinery, 2020. doi: 10.1145/3336191.3371786. URL https://doi.org/10.1145/3336191.3371786.

Yuxin Liao, Yonghui Yang, Min Hou, Le Wu, Hefei Xu, and Hao Liu. Mitigating distribution shifts in sequential recommendation: An invariance perspective. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pp. 1603–1613. Association for Computing Machinery, 2025. doi: 10.1145/3726302.3730036. URL https://doi.org/10.1145/3726302.3730036.

Xiaolin Lin, Weike Pan, and Zhong Ming. Towards interest drift-driven user representation learning in sequential recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, pp. 1541–1551. Association for Computing Machinery, 2025. doi: 10.1145/3726302.3730099. URL https://doi.org/10.1145/3726302.3730099.

Chengkai Liu, Jianghao Lin, Jianling Wang, Hanzhou Liu, and James Caverlee. Mamba4rec: Towards efficient sequential recommendation with selective state space models. *arXiv preprint arXiv:2403.03900*, 2024. doi: 10.48550/arXiv.2403.03900.

Lucas Maystre and Daniel Russo. Temporally-consistent survival analysis. In *Advances in Neural Information Processing Systems*, volume 35 of *NeurIPS*, pp. 10671–10683. Curran Associates, Inc., 2022. doi: 10.52202/068431-0775. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/455e1e30edf721bd7fa334fffabdcad8-Abstract-Conference.html.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30 of *NeurIPS*, pp. 6076–6085. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. Deep recurrent survival analysis. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pp. 4798–4805, 2019. doi: 10.1609/aaai.v33i01.33014798.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, Bangkok, Thailand, aug 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.399. URL https://aclanthology.org/2024.acl-long.399/.

Marten Scheffer, Jordi Bascompte, William A. Brock, Victor Brovkin, Stephen R. Carpenter, Vasilis Dakos, Hermann Held, Egbert H. van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461:53–59, 2009. doi: 10.1038/nature08227. URL https://doi.org/10.1038/nature08227.

Yun-Zhu Song, Yi-Syuan Chen, Lu Wang, and Hong-Han Shuai. General then personal: Decoupling and pre-training for personalized headline generation. *Transactions of the Association for Computational Linguistics*, 11:1588–1607, 2023. doi: 10.1162/tacl_a_00621.

Jiaxi Tang and Ke Wang. Personalized top-n sequential recommendation via convolutional sequence embedding. In Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek (eds.), *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pp. 565–573. ACM, 2018. doi: 10.1145/3159652.3159656. URL https://doi.org/10.1145/3159652.3159656.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9. URL https://doi.org/10.1007/978-3-540-71050-9.

Maria Vlachou. Failure prediction in conversational recommendation systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, RecSys, pp. 599–604. Association for Computing Machinery, 2025. doi: 10.1145/3705328.3748043. URL https://doi.org/10.1145/3705328.3748043.

Wenjie Wang, Xinyu Lin, Liuhui Wang, Fuli Feng, Yunshan Ma, and Tat-Seng Chua. Causal disentangled recommendation against user preference shifts. *ACM Transactions on Information Systems*, 42(1):12:1–12:27, 2024. doi: 10.1145/3593022. URL https://doi.org/10.1145/3593022.

Yuyan Wang, Mohit Sharma, Can Xu, Sriraj Badam, Qian Sun, Lee Richardson, Lisa Chung, Ed H. Chi, and Minmin Chen. Surrogate for long-term user experience in recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD, pp. 4100–4109. Association for Computing Machinery, 2022. doi: 10.1145/3534678.3539073. URL https://doi.org/10.1145/3534678.3539073.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. Neural news recommendation with attentive multi-view learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 3863–3869, 2019a. doi: 10.24963/ijcai.2019/536.

Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. Session-based recommendation with graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):346–353, 2019b. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.3301346. URL http://dx.doi.org/10.1609/aaai.v33i01.3301346.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Hugo Yèche, Alizée Pace, Gunnar Rätsch, and Rita Kuznetsova. Temporal label smoothing for early event prediction. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39913–39938. PMLR, 2023. URL https://proceedings.mlr.press/v202/yeche23a.html.

Zhenrui Yue, Yueqi Wang, Zhankui He, Huimin Zeng, Julian McAuley, and Dong Wang. Linear recurrent units for sequential recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pp. 930–938, 2024. doi: 10.1145/3616855.3635760.

Fei Zhu, Xu-Yao Zhang, Zhen Cheng, and Cheng-Lin Liu. Revisiting confidence estimation: Towards reliable failure prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (5):3370–3387, May 2024. doi: 10.1109/TPAMI.2023.3342285. URL https://doi.org/10.1109/TPAMI.2023.3342285.

APPENDIX CONTENTS

A Notation And Host-State Monitoring Interface **18**

A.1 Interaction History And Frozen Host 18

A.2 Frozen Host-State Extraction 18

A.3 Causal Monitoring Window 19

A.4 History Access Without A Second Preference Encoder 19

A.5 Prospective Outcomes And Structured Views 19

B Exact Construction Of Structured State-History Refinements **20**

B.1 Structured Sentry Interface 20

B.2 Regularity Evidence 21

B.3 Transition Evidence 22

B.4 Regularity And Transition As Distinct Views 23

B.5 Projected Structured Evidence 23

B.6 Gated Residual Refinement 24

B.7 Mechanism-Aware Gate Supervision 24

B.8 Prospective Degradation Objective 24

B.9 Composite Refinement Objectives 25

B.10 History-State Anchoring 25

B.11 Frozen And Learned Parameters 25

B.12 Causal Boundary 25

B.13 Relation To The Main-Paper Formulation 26

C Auxiliary Properties Of The CoTRAC Construction **26**

C.1 History-State Recovery And Bounded Refinement 26

C.2 Interpretation Of The Mechanism-Aware Gate 26

C.3 Rank-Normalized Source-Only Transfer 27

C.4 Interpretation Of Prevalence-Normalized AUPRC 28

C.5 Relation To The Formal Theory And Empirical Findings 28

D Theoretical Foundations Of Prospective Warning **29**

D.1 Setup And Non-Anticipation 29

D.2 Recent Utility And Prospective Incompleteness 29

D.3 Regularity And Transition As Conditional Refinements 33

D.4 Calibration-Conditioned Non-Equivalence Of Warning Criteria 35

D.5 Portability Of Broad And Structured Warning Signatures 37

E Instantiations Of Temporal Monitors **40**

E.1 Common Causal Input Contract 40

E.2 Common Sentry Factorization 41
E.3 Prefix-MLP: Weak Temporal Inductive Bias 42
E.4 TCN-ELECTS-EW: Local Causal Temporal Patterns 42
E.5 LS-EEP: Causal Attention Over Past States 43
E.6 TCSA: Recency-Weighted Causal Attention 43
E.7 TCSR: Recurrent Evidence With Discrete Hazard Risk 44
E.8 DRSA-Style: Recurrent Future-Time Hazard Evolution 44
E.9 DeepHit-Style: Discrete Event-Time Distribution 45
E.10 Temporal-Prior: Recurrent Evidence With Causal Progress Modulation 45
E.11 Three Families Of Risk Parameterization 45
E.12 Shared Prospective Supervision 46
E.13 What Is Shared And What Is Not 46
E.14 Relation To Temporal State-History Matching 47

F Evaluation Protocol And Reporting **47**
F.1 Training And Frozen-State Extraction 47
F.2 Evaluation Metrics And Reporting 48
F.3 Results Protocol 50

G Additional Experimental Results **51**
G.1 Evaluation Settings And Result Coverage 51
G.2 RQ1: Future-Degradation Warning Beyond Recent Utility 52
G.3 RQ2: Regularity And Transition As Refinements Of An Established Warning . . . 53
G.4 RQ3: Reuse Across Encoders, Tasks, and Consumers 53

Table 3: Notation used in the main paper and appendices. H and U denote warning models, while $\mathcal{U}$ denotes task-specific output utility.

Symbol	Meaning
b_t	Observed interaction at time t
$\mathcal{H}_{\leq t}$	Causal user-interaction history
z_t	Preference state produced by the frozen host
ζ_t	Fixed 128-D projected monitoring state used by R/T
$\mathcal{C}_t^{(W)}$	Ordered interaction-state pairs over the causal window
$h \in \{1, 2, 3, 5\}$	Evaluated future warning horizon
$Y_{t,h}$	Future utility-degradation outcome, unavailable at warning time
U	Recent-utility warning control, distinct from task utility $\mathcal{U}$
H	Learned broad History-State warning
R, T, RT	Regularity, Transition, and their joint structured refinement
$s_{t,h}^H$	Prospective H warning probability
s_t^R, s_t^T	Regularity and Transition consistency, not degradation probabilities
$\alpha_t^A, \Delta_{t,h}^A$	Scalar refinement gate and signed horizon-wise correction
π	Degradation prevalence in the matched evaluation population
nAUPRC	$(\text{AUPRC} - \pi)/(1 - \pi)$
TEST FAR	Realized held-out false-alarm rate, not the nominal target

A Notation And Host-State Monitoring Interface

This appendix fixes the notation and monitoring interface used throughout the paper. It makes explicit what belongs to the frozen personalization host, what information is available to a sentry at warning time, and what is reserved for future evaluation. No additional modeling assumption is introduced here. The symbols used in the main paper and subsequent appendices are collected in Table 3.

A.1 Interaction History And Frozen Host

A personalization system receives an ordered sequence of user interactions. At timestep t, an interaction is $b_t = \langle a_t, c_t \rangle$, where a_t is the observed user action and c_t is the associated item, document, content object, or task context. The complete causal interaction history available through timestep t is

$$\mathcal{H}_{\leq t} = \langle b_1, \ldots, b_t \rangle. \tag{6}$$

We represent a trained personalization host by the same encoder and downstream task interface as in Section 3:

$$\begin{aligned} z_t &= \mathcal{M}_\theta(\mathcal{H}_{\leq t}), \\ \widehat{y}_{t+1} &= \mathcal{G}_\phi(z_t, q_{t+1}), \\ u_{t+1} &= \mathcal{U}(\widehat{y}_{t+1}, y_{t+1}^\star). \end{aligned} \tag{7}$$

Here, $z_t \in \mathbb{R}^{d_z}$ is the hidden preference state exposed by the host after history through t, q_{t+1} is the downstream task input when required, and u_{t+1} is task-specific output utility. For recommendation, $\widehat{y}_{t+1}$ denotes candidate scores or a ranking. For personalized generation, it denotes the generated output or output distribution conditioned on the corresponding task context.

No gold preference-state trajectory is available to establish whether the learned z_t values are correct. `CoTRAC` therefore evaluates their prospective warning value rather than their correctness. Different hosts need not expose states with the same dimension, coordinate system, or encoder architecture. The monitored state need only belong to the frozen personalization computation whose future utility is being evaluated.

A.2 Frozen Host-State Extraction

The personalization host is trained before warning monitoring begins. During sentry training and evaluation, the parameters of $\mathcal{M}_\theta$ and $\mathcal{G}_\phi$ remain fixed. For each observed history prefix, the host produces $\langle z_1, \ldots, z_t \rangle$, which `CoTRAC` retains as monitoring evidence. Warning gradients are never propagated into the host, and warning optimization does not modify the recommendation or generation outputs used to define utility. Warning performance therefore measures what can be recovered from

a frozen preference representation rather than what can be obtained after adapting the host to the warning objective.

Different hosts can expose differently parameterized state spaces. Cross-host comparisons are therefore made through sentry outputs or their induced warning orderings rather than coordinate-wise comparison of hidden states. In particular, the source-only reuse experiment in RQ3 does not assume aligned latent coordinates across encoders.

A.3 Causal Monitoring Window

At observation time t, a sentry is restricted to information that has already been realized. For a monitoring window of maximum length W, let $\tau_t = \max(1, t - W + 1)$ and define

$$\mathcal{H}_t^{(W)} = \langle b_{\tau_t}, \ldots, b_t \rangle, \qquad \mathcal{Z}_t^{(W)} = \langle z_{\tau_t}, \ldots, z_t \rangle. \tag{8}$$

The complete causal information used for monitoring is

$$\mathcal{I}_t = (\mathcal{H}_{\leq t}, \mathcal{Z}_t^{(W)}), \qquad \mathscr{F}_t = \sigma(\mathcal{I}_t). \tag{9}$$

Any warning issued at t must be measurable with respect to $\mathscr{F}_t$. Information occurring after t is reserved for defining and evaluating the future outcome.

The operational Temporal State-History Matching record is

$$\mathcal{C}_t^{(W)} = \langle (b_\tau, z_\tau) \rangle_{\tau=\tau_t}^{t}. \tag{10}$$

Each element pairs an observed interaction with the host state produced after the corresponding causal prefix. The ordering retains how the user-side observations and host states develop together over time.

A History-State sentry maps this causal evidence to a prospective warning:

$$\ell_{t,h}^{H} = \mathcal{F}_{\psi,h}(\mathcal{C}_t^{(W)}), \qquad s_{t,h}^{H} = \sigma(\ell_{t,h}^{H}). \tag{11}$$

The operator $\mathcal{F}_{\psi,h}$ is architecture dependent, while the causal evidence contract is shared across the evaluated sentries. Appendix E gives the architecture-specific realizations.

A.4 History Access Without A Second Preference Encoder

Temporal State-History Matching does not imply direct vector similarity between b_t and z_t. Raw interactions and hidden preference states need not lie in the same space, and `CoTRAC` does not require quantities such as $\|b_t - z_t\|$ to be defined. Matching refers to the ordered temporal and causal relation between observed interactions and the host states produced as that history is absorbed.

This distinction also separates relational monitoring from marginal prediction. Let $\mathcal{F}_t^H = \sigma(\mathcal{H}_{\leq t})$, $\mathcal{F}_t^Z = \sigma(Z_t^{(W)})$, and $\mathcal{F}_t^C = \sigma(\mathcal{H}_{\leq t}, Z_t^{(W)})$. A history-only warning is measurable with respect to $\mathcal{F}_t^H$, while a state-only warning is measurable with respect to $\mathcal{F}_t^Z$. Either marginal may carry prospective predictive information, but neither by itself represents the monitored interaction-state relation. TSHM retains the realized user-side history together with the host-side state response produced as that history evolves. It therefore defines a relational monitoring object rather than claiming that the joint record must contain strictly more predictive information than either marginal.

Access to $\mathcal{H}_{\leq t}$ also does not introduce a second preference encoder. `CoTRAC` does not learn another mapping $\widetilde{\mathcal{M}}(\mathcal{H}_{\leq t})$ whose output is treated as a competing user representation and then compared with z_t. A sentry may project interaction fields or host states as part of its warning architecture. Those projections are not consumed by the downstream task head and do not replace the host representation.

A.5 Prospective Outcomes And Structured Views

Future utility remains outside the definition of TSHM. Let $h > 0$ denote a prospective horizon and let $\mathcal{D}_h$ be the task-specific degradation operator. The future warning target is

$$Y_{t,h} = \mathcal{D}_h(U_{>t}), \tag{12}$$

where $U_{>t}$ denotes the future utility trace required by the corresponding evaluation protocol. At warning time t, $Y_{t,h}$ is unavailable. It is used afterward to determine whether the causal evidence in $\mathcal{I}_t$ was prospectively informative.

The broad History-State view operates on $\mathcal{C}_t^{(W)}$. The structured Regularity and Transition mechanisms use a common projected monitoring path. Let $\zeta_t \in \mathbb{R}^{128}$ denote the fixed projection of host state z_t used by those mechanisms. This projection standardizes their dimensional interface without aligning different hosts coordinate by coordinate. With projected displacement $d_t = \zeta_t - \zeta_{t-1}$, Regularity characterizes local organization of realized projected-state motion, while Transition predicts the newest projected displacement from earlier displacements and compares that prediction with the realized update. Appendix B gives the exact construction.

For $A \in \{R, T, RT\}$, structured refinement uses a scalar gate α_t^A and a signed horizon-wise correction $\Delta_{t,h}^A$:

$$\ell_{t,h}^{H+A} = \ell_{t,h}^{H} + \alpha_t^A \Delta_{t,h}^A. \tag{13}$$

Here, R, T, and RT denote Regularity, Transition, and their joint interaction. The joint construction is not treated as a fourth independent state property.

This notation separates three objects that remain distinct throughout the paper. The causal History-State record specifies what is observable at warning time, the sentry converts that evidence into prospective risk, and future utility determines whether the warning was informative. TSHM therefore defines a monitored relation rather than an objectively correct state or a causal mechanism of degradation.

B Exact Construction Of Structured State-History Refinements

Section 4 introduces Regularity and Transition as structured refinements of the History-State warning. This appendix gives their exact construction. The same Regularity, Transition, and RT mechanisms are used across datasets and sentry architectures. Dataset-specific host encoders feed a common monitoring path, after which the structured mechanisms below are unchanged. Regularity and Transition are trained from causal state trajectories without future-degradation supervision, frozen, and then used by separately trained refinement adapters.

For the structured mechanisms, let $\zeta_1, \ldots, \zeta_L \in \mathbb{R}^{128}$ denote the valid, chronologically ordered monitoring states in the window ending at observation time t, where $L \leq W = 8$. These are the frozen host states after a fixed projection to the common 128-dimensional monitoring path. This projection reconciles dimensionality for the monitoring mechanism without aligning different hosts coordinate by coordinate. Define

$$d_j = \zeta_j - \zeta_{j-1}, \qquad j = 2, \ldots, L, \qquad a_j = d_j - d_{j-1}, \qquad j = 3, \ldots, L. \tag{14}$$

B.1 Structured Sentry Interface

Let $\mathcal{C}_t^{(W)}$ be the causal correspondence record defined in Appendix A. A trained History-State sentry produces a multi-horizon warning logit and an internal sentry representation,

$$\left(\boldsymbol{\ell}_t^H, \mathbf{r}_t^H\right) = \mathcal{F}_\psi\left(\mathcal{C}_t^{(W)}\right), \qquad \boldsymbol{\ell}_t^H = \left[\ell_{t,h}^H\right]_{h \in \mathcal{K}}. \tag{15}$$

Here, $\mathbf{r}_t^H$ is learned by the sentry for prospective warning. It is distinct from the host preference state z_t and from its projected monitoring state ζ_t. The host state is evidence being monitored, while $\mathbf{r}_t^H$ is the sentry representation learned from that evidence.

For each trained History-State sentry, `CoTRAC` constructs three separate structured sentries,

$$H + R, \qquad H + T, \qquad H + RT. \tag{16}$$

They share the same trained History-State parent and are optimized separately. The parent sentry remains fixed during structured-refinement training. The evaluated object is the complete composite sentry $H + A$ for $A \in \{R, T, RT\}$.

B.2 REGULARITY EVIDENCE

Regularity asks whether the realized projected-state trajectory remains locally coherent. It first converts the current causal state window into four raw violation features. For $L \geq 4$, let $\epsilon = 10^{-6}$ and define

$$\begin{aligned} v_{1,t}^R &= \frac{|\|d_L\|_2 - \mathrm{med}_{2\leq j<L}\,\|d_j\|_2|}{\mathrm{med}_{2\leq j<L}\,\|d_j\|_2 + \epsilon}, \\ v_{2,t}^R &= 1 - \frac{d_{L-1}^\top d_L}{\|d_{L-1}\|_2\|d_L\|_2 + \epsilon}, \\ v_{3,t}^R &= \begin{cases} \|a_L\|_2, & L = 4, \\ \dfrac{\|a_L\|_2}{\mathrm{med}_{3\leq j<L}\,\|a_j\|_2 + \epsilon}, & L \geq 5, \end{cases} \\ v_{4,t}^R &= \|a_L - a_{L-1}\|_2. \end{aligned} \tag{17}$$

The four coordinates respectively measure speed deviation from earlier movement in the same window, directional deviation from the preceding displacement, acceleration deviation from earlier acceleration, and discrete jerk magnitude. The within-window medians exclude the latest displacement or acceleration being evaluated. For exactly four valid states, no earlier acceleration set is available, so the third violation reduces to the magnitude of the latest acceleration. The fourth quantity is the norm of a second difference of displacements rather than an explicit curvature calculation. For $L < 4$, the raw vector is set to zero.

The raw violation vector $\mathbf{v}_t^R = [v_{1,t}^R, v_{2,t}^R, v_{3,t}^R, v_{4,t}^R]$ is the input to the $4 \to 64 \to 32$ Regularity precursor. The soft supervision targets are constructed separately from eligible training windows and are not network inputs. For each component k, let $\mathcal{E}_R$ denote the training rows used to fit the Regularity mechanism and define

$$q_{25,k}^R = \mathrm{Quantile}_{0.25}\{v_{k,i}^R : i \in \mathcal{E}_R\}, \qquad q_{75,k}^R = \mathrm{Quantile}_{0.75}\{v_{k,i}^R : i \in \mathcal{E}_R\}. \tag{18}$$

The component scale is

$$I_k^R = \max\left(q_{75,k}^R - q_{25,k}^R, 10^{-4}\right). \tag{19}$$

Each raw violation is converted into a training-derived satisfaction target by

$$\widetilde{p}_{k,t}^R = \sigma\left(\mathrm{clip}\left(\frac{q_{75,k}^R - v_{k,t}^R}{I_k^R}, -30, 30\right)\right), \qquad k = 1, \ldots, 4. \tag{20}$$

The same mapping is used for all four components. A violation equal to its training 75th percentile maps to 0.5. Smaller violations map above 0.5, while larger violations map below 0.5. These quantities are trajectory-derived supervision targets rather than probabilities of avoiding future degradation.

Let $\mathbf{h}_t^R \in \mathbb{R}^{32}$ be the learned precursor representation. The $32 \to 4$ satisfaction head applies an elementwise sigmoid,

$$p_{k,t}^R = \sigma\left([\mathrm{Head}_R(\mathbf{h}_t^R)]_k\right), \qquad k = 1, \ldots, 4. \tag{21}$$

The four outputs are bounded independently rather than normalized by a softmax. The precursor is trained with component-wise binary cross-entropy,

$$\mathcal{L}_{\mathrm{pre},R} = \frac{1}{4|\mathcal{E}_R|} \sum_{t\in\mathcal{E}_R} \sum_{k=1}^{4} \mathrm{BCE}\left(p_{k,t}^R, \widetilde{p}_{k,t}^R\right). \tag{22}$$

No future utility or future-degradation label appears in this precursor objective.

The learned satisfaction values are

$$\mathbf{p}_t^R = [p_{1,t}^R, p_{2,t}^R, p_{3,t}^R, p_{4,t}^R] \in (0,1)^4, \tag{23}$$

and their conjunction is

$$s_t^R = \prod_{k=1}^{4} p_{k,t}^R. \tag{24}$$

The product requires all four learned Regularity conditions to remain well satisfied, so one low component cannot be hidden by averaging it with the others. The Regularity evidence is

$$\mathbf{R}_t = [\mathbf{p}_t^R; s_t^R; \mathbf{h}_t^R] \in \mathbb{R}^{37}. \tag{25}$$

The precursor is frozen before structured-refinement training. Regularity therefore supplies a fixed description of local projected-state motion learned without future-degradation supervision. Low Regularity does not itself imply future degradation.

B.3 Transition Evidence

Transition asks whether the newest projected-state displacement agrees with a prediction formed from earlier displacements. It is trained as a causal self-supervised prediction mechanism. For a valid window with $L \geq 3$, define

$$n = \min(L-2, W-2). \tag{26}$$

Let κ_T denote the training-derived displacement scale, defined as the median of the positive norms of the latest displacement among valid training rows and floored at 10^{-5}. The causal GRU input and held-out prediction target are

$$\mathbf{X}_t = \frac{1}{\kappa_T}\begin{bmatrix} d_{L-n} \\ \vdots \\ d_{L-1} \end{bmatrix} \in \mathbb{R}^{n\times 128}, \qquad \mathbf{y}_t = \frac{d_L}{\kappa_T} \in \mathbb{R}^{128}. \tag{27}$$

The latest displacement d_L is the self-supervised target and never enters the GRU input for that window.

Let $\mathbf{g}_t^T \in \mathbb{R}^{64}$ be the causal GRU state. The Transition precursor produces a predicted mean, a clipped coordinate-wise log variance, and a 32-dimensional learned representation,

$$\begin{aligned} \mathbf{g}_t^T &= \mathrm{GRU}_T(\mathbf{X}_t; n), \\ \boldsymbol{\mu}_t &= \mathrm{Head}_\mu(\mathbf{g}_t^T) \in \mathbb{R}^{128}, \\ \boldsymbol{\lambda}_t &= \mathrm{clip}\big(\mathrm{Head}_\lambda(\mathbf{g}_t^T), -8, 6\big) \in \mathbb{R}^{128}, \\ \mathbf{h}_t^T &= \mathrm{Proj}_T(\mathbf{g}_t^T) \in \mathbb{R}^{32}. \end{aligned} \tag{28}$$

Writing

$$V_{t,j} = \exp(\lambda_{t,j}), \tag{29}$$

the precursor is trained on valid Transition rows with the diagonal heteroscedastic Gaussian objective

$$\mathcal{L}_{\mathrm{pre},T} = \frac{1}{|\mathcal{E}_T|}\sum_{t\in\mathcal{E}_T}\frac{1}{2\cdot 128}\sum_{j=1}^{128}\left[(y_{t,j}-\mu_{t,j})^2\exp(-\lambda_{t,j}) + \lambda_{t,j}\right]. \tag{30}$$

Direction and magnitude are not separate precursor-loss terms. They are computed only after the causal prediction has been formed.

Once the held-out displacement $\mathbf{y}_t$ is available, define the standardized Transition residual

$$r_t^T = \sqrt{\frac{1}{128}\sum_{j=1}^{128}\frac{(y_{t,j}-\mu_{t,j})^2}{V_{t,j}+10^{-6}}}. \tag{31}$$

The bounded Transition-alignment quantity used for later gate supervision is

$$s_t^T = \exp(-r_t^T). \tag{32}$$

A smaller standardized residual therefore produces stronger Transition consistency.

The Transition evidence retains four additional scalar views of the same realized-versus-predicted update. With $\epsilon = 10^{-6}$ and

$$\widetilde{\boldsymbol{\mu}}_t = \frac{\boldsymbol{\mu}_t}{\|\boldsymbol{\mu}_t\|_2 + \epsilon}, \tag{33}$$

define

$$\begin{aligned}\mathrm{Dir}_t &= 1 - \mathrm{clip}\left(\frac{\mathbf{y}_t^\top \boldsymbol{\mu}_t}{\|\mathbf{y}_t\|_2\|\boldsymbol{\mu}_t\|_2 + \epsilon}, -1, 1\right),\\ \mathrm{Mag}_t &= \frac{|\|\mathbf{y}_t\|_2 - \|\boldsymbol{\mu}_t\|_2|}{\|\mathbf{y}_t\|_2 + \epsilon},\\ \mathrm{Unc}_t &= \sqrt{\frac{1}{128}\sum_{j=1}^{128} V_{t,j}},\\ \mathrm{Orth}_t &= \left\|\mathbf{y}_t - (\mathbf{y}_t^\top \widetilde{\boldsymbol{\mu}}_t)\widetilde{\boldsymbol{\mu}}_t\right\|_2.\end{aligned} \tag{34}$$

The direction feature clips the stabilized cosine to $[-1, 1]$. The magnitude feature is a relative absolute norm difference. Predictive uncertainty is the square root of mean predictive variance. The orthogonal feature uses the stabilized $\widetilde{\boldsymbol{\mu}}_t$ in Eq. 33.

The Transition evidence is

$$\mathbf{T}_t = [s_t^T; r_t^T; \mathrm{Dir}_t; \mathrm{Mag}_t; \mathrm{Unc}_t; \mathrm{Orth}_t; \mathbf{h}_t^T] \in \mathbb{R}^{38}. \tag{35}$$

Thus, $\mathbf{T}_t$ retains alignment, standardized residual, four interpretable mismatch or uncertainty scalars, and the learned 32-dimensional representation.

A Transition window is valid when $L \geq 3$, which supplies at least one past displacement and one held-out target displacement. Let $m_t^T \in \{0, 1\}$ denote this validity indicator. Invalid rows have $m_t^T = 0$ and do not contribute to Transition precursor or gate supervision. Additional task-level eligibility restrictions are specified by the evaluation protocol.

The Transition precursor is frozen before structured-refinement training. Transition surprise is therefore a property of causal state evolution, not a future-degradation label.

B.4 Regularity And Transition As Distinct Views

Regularity and Transition use the same projected state evolution but answer different questions. Regularity describes the organization of the realized local trajectory. Transition evaluates the newest displacement against a causal prediction formed without that displacement. A trajectory can remain smooth in speed and direction while producing an update that is unexpected from its preceding motion. Conversely, a locally sharp update can be consistent with what the causal predictor expected. The two evidence streams are therefore retained separately rather than collapsed into one trajectory-change statistic.

B.5 Projected Structured Evidence

Regularity and Transition have different native dimensions. Before refinement, they are projected separately into a common 32-dimensional space,

$$\widetilde{\mathbf{R}}_t = \phi_R(\mathbf{R}_t), \qquad \widetilde{\mathbf{T}}_t = \phi_T(\mathbf{T}_t), \tag{36}$$

where

$$\phi_R : \mathbb{R}^{37} \to \mathbb{R}^{32}, \qquad \phi_T : \mathbb{R}^{38} \to \mathbb{R}^{32} \tag{37}$$

are trainable refinement projections. These projections belong to the warning model and do not modify the frozen host states.

The single-view structured representations are

$$\mathbf{f}_t^R = \widetilde{\mathbf{R}}_t, \qquad \mathbf{f}_t^T = \widetilde{\mathbf{T}}_t. \tag{38}$$

For the joint condition, CoTRAC explicitly retains both agreement and disagreement,

$$\mathbf{f}_t^{RT} = \left[\widetilde{\mathbf{R}}_t; \widetilde{\mathbf{T}}_t; \widetilde{\mathbf{R}}_t \odot \widetilde{\mathbf{T}}_t; \left|\widetilde{\mathbf{R}}_t - \widetilde{\mathbf{T}}_t\right|\right] \in \mathbb{R}^{128}. \tag{39}$$

The elementwise product represents simultaneous activation of the two projected views, while the absolute difference represents their disagreement. RT is therefore an interaction construction rather than a third structured mechanism.

B.6 Gated Residual Refinement

Structured evidence refines the frozen History-State warning rather than replacing it. For $A \in \{R, T, RT\}$, define

$$\mathbf{x}_t^A = [\mathbf{r}_t^H; \mathbf{f}_t^A]. \tag{40}$$

A residual branch predicts a horizon-wise signed correction,

$$\boldsymbol{\delta}_t^A = D_A(\mathbf{x}_t^A) \in \mathbb{R}^{|\mathcal{K}|}, \tag{41}$$

while a separate scalar gate gives

$$g_t^A = G_A(\mathbf{x}_t^A), \qquad \alpha_t^A = \sigma(g_t^A) \in (0, 1). \tag{42}$$

The composite warning is

$$\boxed{\boldsymbol{\ell}_t^{H+A} = \boldsymbol{\ell}_t^H + \alpha_t^A \boldsymbol{\delta}_t^A}, \qquad A \in \{R, T, RT\}. \tag{43}$$

The gate controls how strongly the structured view can modify H, while the signed residual determines the horizon-wise direction and magnitude of that modification. Neither low Regularity nor high Transition surprise is assigned a fixed degradation direction.

The final residual layer is initialized at zero. Therefore

$$\boldsymbol{\delta}_t^A = \mathbf{0} \quad \Longrightarrow \quad \boldsymbol{\ell}_t^{H+A} = \boldsymbol{\ell}_t^H \tag{44}$$

at the start of refinement training. Every structured sentry begins from its corresponding trained History-State solution.

B.7 Mechanism-Aware Gate Supervision

The complementary gate quantity

$$1 - \alpha_t^A = \sigma(-g_t^A) \tag{45}$$

represents the stay-with-History-State side of the adapter. Structured supervision acts on this quantity. It does not directly determine the sign of the degradation correction.

For an adapter whose gate is supervised by Regularity, the auxiliary loss is

$$\mathcal{L}_R(g^A) = \mathrm{BCEWithLogits}\left(-g_t^A, s_t^R\right). \tag{46}$$

High conjunction s_t^R favors retaining the parent warning. Lower Regularity permits a larger correction without asserting that irregularity itself predicts degradation.

For an adapter whose gate is supervised by Transition, the corresponding valid-row loss is

$$\mathcal{L}_T(g^A) = \frac{\sum_t m_t^T \,\mathrm{BCEWithLogits}\left(-g_t^A, s_t^T\right)}{\sum_t m_t^T}. \tag{47}$$

A small standardized residual makes s_t^T large and favors retaining H. Greater transition surprise permits stronger refinement. Invalid Transition rows do not contribute to this loss.

The $H + R$ adapter uses $\mathcal{L}_R(g^R)$, the $H + T$ adapter uses $\mathcal{L}_T(g^T)$, and the joint $H + RT$ gate is supervised by both criteria using its own gate logit g^{RT}. This does not constrain RT to dominate either individual view. The joint adapter instead learns whether their agreement or disagreement warrants changing the established H warning.

B.8 Prospective Degradation Objective

History-State and all structured composite sentries predict the same prospective degradation outcome. For observation i and horizon h, let $y_{i,h} \in \{0, 1\}$ be the degradation label, let $v_{i,h} \in \{0, 1\}$ indicate whether that horizon is evaluable, and let $\widehat{r}_{i,h} = \sigma(\ell_{i,h})$. With $\Omega = \{(i, h) : v_{i,h} = 1\}$, the masked binary cross-entropy is

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{|\Omega|} \sum_{(i,h)\in\Omega} \left[y_{i,h} \log \widehat{r}_{i,h} + (1 - y_{i,h}) \log(1 - \widehat{r}_{i,h})\right]. \tag{48}$$

The prediction objective also contains the pairwise ranking term $\mathcal{L}_{\text{rank}}$ and horizon-monotonicity term $\mathcal{L}_{\text{mono}}$ defined by the warning protocol. The common objective is

$$\mathcal{L}_{\text{pred}} = \mathcal{L}_{\text{BCE}} + 0.1\mathcal{L}_{\text{rank}} + 0.025\mathcal{L}_{\text{mono}}. \tag{49}$$

The same prospective objective is used for H and the complete $H + A$ sentries. Regularity and Transition therefore refine one warning problem rather than defining separate degradation targets.

B.9 Composite Refinement Objectives

After H, R, and T have been trained and frozen, the three structured adapters are optimized with

$$\mathcal{L}_{H+R} = 0.7\mathcal{L}_{\text{pred}}^{H+R} + 0.3\mathcal{L}_R(g^R) + \lambda_D\mathcal{L}_D, \tag{50}$$

$$\mathcal{L}_{H+T} = 0.7\mathcal{L}_{\text{pred}}^{H+T} + 0.3\mathcal{L}_T(g^T) + \lambda_D\mathcal{L}_D, \tag{51}$$

and

$$\mathcal{L}_{H+RT} = 0.4\mathcal{L}_{\text{pred}}^{H+RT} + 0.3\mathcal{L}_R(g^{RT}) + 0.3\mathcal{L}_T(g^{RT}) + \lambda_D\mathcal{L}_D. \tag{52}$$

The semantic weights sum to one in each composite objective. The anchoring term $\mathcal{L}_D$ is an additional conservative regularizer rather than part of that normalized partition. The precursor losses $\mathcal{L}_{\text{pre},R}$ and $\mathcal{L}_{\text{pre},T}$ do not appear here because both precursors are already frozen.

B.10 History-State Anchoring

Structured refinement retains an explicit reference to the frozen History-State parent. On the same training batch, let $\mathcal{L}_{\text{pred}}^{H}$ be the parent prediction loss and $\mathcal{L}_{\text{pred}}^{H+A}$ the composite prediction loss. The anchoring term is

$$\mathcal{L}_D = \left[\mathcal{L}_{\text{pred}}^{H+A} - \mathcal{L}_{\text{pred}}^{H} + m_D\right]_+, \tag{53}$$

where $[x]_+ = \max(0, x)$. In the evaluated configuration, $m_D = 0.002$ and $\lambda_D = 1$. The penalty vanishes when the composite loss is at least m_D below its parent loss on that training batch. Otherwise, it penalizes the shortfall from this margin. This is a training anchor, not a pointwise warning constraint or a guarantee of held-out improvement.

B.11 Frozen And Learned Parameters

The structured construction has three stages. First, the History-State sentry parameters θ_H are trained on prospective degradation and then fixed. Second, the Regularity and Transition precursor parameters θ_R and θ_T are trained from causal state trajectories using $\mathcal{L}_{\text{pre},R}$ and $\mathcal{L}_{\text{pre},T}$, without future-degradation supervision, and then fixed. Third, each adapter parameter set θ_A is trained while the preceding components remain frozen. During this final stage,

$$\theta_H, \theta_R, \theta_T \quad \text{are fixed,} \qquad \theta_A \quad \text{is optimized.} \tag{54}$$

The adapter parameters include the structured projection, residual network, and scalar gate. Future degradation supervises the composite warning through $\mathcal{L}_{\text{pred}}$ during this stage. It never changes the already trained Regularity or Transition mechanisms.

B.12 Causal Boundary

Regularity and Transition preserve the same warning-time causal boundary as the History-State sentry. Regularity uses projected states already realized by t. Transition predicts the latest displacement from earlier projected displacements and withholds d_L from its GRU input until the prediction has been formed. After the state at t is observed, the held-out displacement is used to materialize Transition evidence. Hence

$$\mathbf{R}_t, \mathbf{T}_t, \boldsymbol{\ell}_t^{H+A} \in \mathscr{F}_t. \tag{55}$$

Future utility $U_{>t}$ and degradation outcomes $Y_{t,h}$ remain outside this information set at warning time. The precursor mechanisms therefore characterize causal state evolution, while future outcomes determine whether those characteristics improve prospective warning.

B.13 Relation To The Main-Paper Formulation

The abstract refinement in Section 4 maps directly to the implementation. For $A \in \{R, T, RT\}$,

$$\Delta_{t,h}^{A} \equiv \delta_{t,h}^{A}, \qquad \alpha_t^A = \sigma\big(G_A([\mathbf{r}_t^H; \mathbf{f}_t^A])\big)\,, \tag{56}$$

and

$$\ell_{t,h}^{H+A} = \ell_{t,h}^{H} + \alpha_t^A \delta_{t,h}^{A}. \tag{57}$$

The three views can therefore be separated by the question they answer. H asks whether the evolving interaction-state record contains prospective warning beyond recent utility. R asks whether the realized projected-state trajectory remains locally coherent. T asks whether the newest projected-state update agrees with a causal prediction formed from earlier updates. RT represents their interaction when simultaneous activation or disagreement is useful. None of these constructions assumes that structured refinement must improve every warning criterion. Their empirical contribution is therefore evaluated separately for discrimination, event coverage, warning lead, and false-alarm burden.

C Auxiliary Properties Of The CoTRAC Construction

Appendix B gives the exact structured construction, while Appendix D gives the formal theory of prospective warning. This section records the identities needed to connect those two parts to the empirical analysis. It establishes how structured sentries retain History-State warning as a reference, how Regularity and Transition control later refinement, what is transferred in source-only encoder reuse, and how prevalence-normalized AUPRC relates to ordinary AUPRC. These properties describe the construction and its interpretation. They do not imply improved held-out warning performance.

C.1 History-State Recovery And Bounded Refinement

For each structured view $A \in \{R, T, RT\}$, the implemented warning logit is

$$\ell_{t,h}^{H+A} = \ell_{t,h}^{H} + \alpha_t^A \delta_{t,h}^{A}, \qquad \alpha_t^A \in (0, 1). \tag{58}$$

Setting the structured residual to zero exactly recovers the parent warning:

$$\delta_{t,h}^{A} = 0 \quad \Longrightarrow \quad \ell_{t,h}^{H+A} = \ell_{t,h}^{H}. \tag{59}$$

The final residual layer is initialized at zero, so every structured sentry begins refinement training from its corresponding History-State solution. Whenever the residual branch can represent the zero function, that parent solution remains available to the structured sentry after training begins. This makes the held-out H versus $H + A$ comparisons in RQ2 interpretable as departures from a common trained parent rather than comparisons between unrelated warning models.

The gate also limits the immediate effect of the structured correction:

$$\left|\ell_{t,h}^{H+A} - \ell_{t,h}^{H}\right| = \alpha_t^A \left|\delta_{t,h}^{A}\right| \le \left|\delta_{t,h}^{A}\right|. \tag{60}$$

Since the sigmoid is $1/4$-Lipschitz,

$$\left|s_{t,h}^{H+A} - s_{t,h}^{H}\right| \le \frac{1}{4}\alpha_t^A \left|\delta_{t,h}^{A}\right| \le \frac{1}{4}\left|\delta_{t,h}^{A}\right|. \tag{61}$$

The structured sentry therefore learns a gated departure from a fixed History-State reference. These bounds describe the architecture only. They do not guarantee better discrimination, event coverage, warning lead, false-alarm burden, or any other held-out criterion.

C.2 Interpretation Of The Mechanism-Aware Gate

Regularity and Transition are trained from causal projected-state trajectories before structured-refinement training and are then fixed. Their precursor objectives do not use future utility or future-degradation labels. Regularity materializes $\mathbf{R}_t$ from four learned satisfaction quantities, their conjunction s_t^R, and a learned representation. Transition materializes $\mathbf{T}_t$ from causal prediction consistency s_t^T, its standardized residual, four auxiliary mismatch or uncertainty scalars, and a

learned representation. During refinement, s_t^R and s_t^T supervise the complementary gate value $1 - \alpha_t^A = \sigma(-g_t^A)$, which represents the stay-with-History-State side of the gate. The signed degradation correction remains learned from the prospective objective.

For Regularity,

$$s_t^R = \prod_{j=1}^{4} p_{j,t}^R. \tag{62}$$

The isolated gate objective then favors

$$\alpha_t^R \approx 1 - s_t^R. \tag{63}$$

High joint Regularity therefore favors retaining the History-State warning. Reduced Regularity permits a larger correction without determining whether prospective risk should increase or decrease.

For Transition,

$$s_t^T = \exp(-r_t^T). \tag{64}$$

The corresponding isolated gate objective then favors

$$\alpha_t^T \approx 1 - \exp(-r_t^T). \tag{65}$$

A newest update consistent with its causal prediction favors retaining the History-State warning. Greater standardized transition surprise permits stronger refinement without fixing the direction of that refinement.

In the complete structured sentry, the gate losses are optimized together with the prospective prediction and History-State anchoring objectives. Equations 63 and 65 therefore explain the isolated effect of gate supervision rather than prescribing the learned gate deterministically. Regularity and Transition also remain distinct. Regularity describes the organization of realized projected-state motion, while Transition compares the newest projected-state displacement with a prediction formed from earlier displacements. Appendix D.3 formalizes their conditional and potentially non-additive role relative to the frozen History-State representation.

C.3 Rank-Normalized Source-Only Transfer

RQ3 asks what can be reused when preference encoders need not share hidden-state coordinates or calibrated warning scales. The source-only experiment therefore transfers warning orderings. Let $\Omega = \{1, \ldots, n\}$ be the common evaluation indices and let $s_j(i)$ be the warning score contributed by source encoder j for index i. For each source independently, define

$$\rho_j(i) = \frac{\mathrm{rank}_j(i) - 1}{n - 1}, \tag{66}$$

The source-only consensus is

$$s_{\mathrm{src}}(i) = \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \rho_j(i), \tag{67}$$

where $\mathcal{S}$ contains only source encoders and excludes the held-out target encoder.

Any strictly increasing source-specific transformation $s_j \mapsto g_j \circ s_j$ preserves the source ordering and therefore leaves both ρ_j and s_{src} unchanged. Source-only reuse is consequently invariant to monotone recalibration of an individual source score. What crosses encoders is the ordering induced by source warnings, not their absolute probability scales or coordinate-wise alignment of their hidden states.

This invariance has a narrow scope. It does not cover non-monotone score transformations, changes in degradation definitions, or distribution shifts that alter the warning ordering itself. It also does not establish the performance of rank averaging on a held-out encoder. Appendix D.5 gives a separate distributional analysis of broad and structured warning signatures, while RQ3 measures source-only reuse empirically.

C.4 Interpretation Of Prevalence-Normalized AUPRC

Degradation prevalence varies across evaluation cells, so nAUPRC expresses discrimination relative to the prevalence reference within each matched population. For $0 < \pi < 1$,

$$\mathrm{nAUPRC} = \frac{\mathrm{AUPRC} - \pi}{1 - \pi}. \tag{68}$$

For fixed π, this is a strictly increasing affine transformation:

$$\Delta\mathrm{nAUPRC} = \frac{\Delta\mathrm{AUPRC}}{1 - \pi}. \tag{69}$$

Two methods evaluated on the same matched population therefore have the same ordering under AUPRC and nAUPRC. The prevalence reference $\mathrm{AUPRC} = \pi$ maps to $\mathrm{nAUPRC} = 0$, while perfect ranking maps to 1. A negative nAUPRC indicates discrimination below the prevalence reference in that population. The normalization does not make populations with different prevalences statistically equivalent.

C.5 Relation To The Formal Theory And Empirical Findings

Appendix D gives the formal results corresponding to the empirical claims. It establishes five conditional properties. First, every `CoTRAC` warning is restricted to information available by its observation time. Second, under the stated local smoothness, utility-blind-direction, and state-response conditions, recent utility can be second-order flat along a future-quality-relevant direction while the observed state component of the History-State record changes at first order. Third, Regularity and Transition can contribute only through information that remains useful after the frozen History-State representation is already known, and their contributions need not combine additively. Fourth, discrimination, event coverage, warning lead, and false-alarm burden have no universal monotone relationship after a fixed pre-test calibration rule. Fifth, under the stated compatible product metrics, projecting a richer structured signature onto its shared History-State coordinates cannot increase source-target Wasserstein distance, which gives the broad signature a no-looser worst-case Lipschitz transfer bound.

The RQ1 results exhibit the empirical pattern for which the prospective-separation theory is relevant. On the five supported MovieLens hosts, H exceeds the matched utility warning by $+0.106$ mean nAUPRC and is better in all $160/160$ encoder-sentry-horizon comparisons. The same direction recurs on PENS with $+0.016$ mean nAUPRC and $67/80$ positive comparisons, and on Amazon with $+0.015$ and $21/24$. The mean $H - U$ difference is positive at every tested horizon in all three domains. The horizon profile is not universal. MovieLens grows from $+0.048$ at $h = 1$ to $+0.163$ at $h = 5$, while Amazon peaks at $h = 3$ and declines by $h = 5$. Amazon also retains slightly negative absolute mean H nAUPRC despite the positive $H - U$ difference. The empirical insight is therefore not that History-State evidence always yields a strong warning. It is that the learned joint record can contain prospective information beyond recent utility even when absolute predictability remains limited. Section 6.1 reports the primary result, and Appendix G.2 gives the host, sentry, horizon, and probability-quality breakdowns.

The RQ2 results exhibit the conditional and non-additive behavior formalized for structured evidence. On PENS, R, T, and RT increase event coverage on average while their mean warning-lead changes remain slightly negative. On Amazon, the refinements increase warning lead by approximately 0.014 to 0.020 and increase realized false-alarm burden by approximately 0.015 to 0.020. The joint RT view also does not systematically dominate its components. Across the 160 MovieLens host-sentry-horizon cells, RT exceeds both individual refinements in only $57/160$ cells and is lower than the stronger individual refinement by 1.139×10^{-3} nAUPRC on average. It is strongest for none of the five PENS hosts and remains below the stronger individual refinement for all three Amazon hosts. Structured evidence can therefore alter which events are detected, when they are detected, and how much alarm burden is incurred without producing one universally superior refinement. Section 6.2 reports the primary findings, and Appendix G.3 gives the matched host, horizon, and operating-criterion analyses.

The RQ3 results expose the distinction that motivates the portability analysis. Broad source-only H warning remains positively discriminative in all $160/160$ MovieLens and $80/80$ PENS transfer

cells, retaining 86.45% and 83.58% of self-warning discrimination on average. The finer structured effects do not transfer in the same way. Relative to transferred H, mean R, T, and RT effects are $(-1.737, -2.005, -1.483) \times 10^{-3}$ nAUPRC on MovieLens and $(+1.473, +1.400, +1.649) \times 10^{-3}$ on PENS. The same PENS preference backbones also produce different structured warning behavior under recommendation and generation, even though the state producer is held fixed. The pre-generation OpenLLM experiments retain positive H discrimination while structured changes again depend on the consumer and metric. The portability theory therefore motivates separating broad-warning reuse from reuse of finer state-dynamics structure. The task and consumer comparisons further show that representation geometry alone cannot explain every structured difference. Section 6.3 reports the primary findings, and Appendix G gives the complete source-only, task-change, and pre-generation analyses.

These empirical patterns do not validate the assumptions of the formal results by observation. The theory identifies behaviors that are mathematically possible under explicit conditions. The experiments establish which corresponding patterns occur in the evaluated systems. Together, they make the empirical differences interpretable without turning them into causal conclusions or theorem verification.

D Theoretical Foundations Of Prospective Warning

This appendix gives the mathematical basis for the empirical questions in Section 3. The theory serves four distinct purposes. It first fixes the causal information boundary of a valid warning. It then identifies conditions under which a future-quality-relevant change can be locally visible in the observed History-State record while recent utility remains comparatively flat. It next formalizes Regularity and Transition as conditional refinements of an already learned History-State representation and explains why their contributions need not add. Finally, it separates the warning criteria used in RQ2 and establishes a distributional portability bound for the broad and structured signatures studied in RQ3. Every result is conditional on its stated assumptions. None substitutes for the corresponding empirical test.

D.1 Setup And Non-Anticipation

The first requirement is causal rather than predictive. A warning must be computable from information already available when it is issued. From Section 3, an interaction is $b_t = \langle a_t, c_t \rangle$, the causal history is $\mathcal{H}_{\leq t} = \langle b_1, \ldots, b_t \rangle$, and the frozen personalization system produces

$$\begin{aligned} z_t &= \mathcal{M}_\theta(\mathcal{H}_{\leq t}), \\ \hat{y}_{t+1} &= \mathcal{G}_\phi(z_t, q_{t+1}), \\ u_{t+1} &= \mathcal{U}(\hat{y}_{t+1}, y^\star_{t+1}). \end{aligned} \tag{70}$$

Under this indexing, u_j evaluates the output produced from z_{j-1}. Let $\mathcal{I}_t = (\mathcal{H}_{\leq t}, Z_t^{(W)})$ denote the causal monitoring information through t, with the usual truncation for short prefixes, and let $\mathcal{F}_t = \sigma(\mathcal{I}_t)$. The History-State record $\mathcal{C}_t^{(W)} = \langle (b_\tau, z_\tau) \rangle_{\tau=t-W+1}^{t}$ uses observations through t. The structured mechanisms apply a fixed seeded projection to the observed host states, so each projected monitoring state ζ_τ is also $\mathcal{F}_t$-measurable. Regularity uses projected states through t. Transition forms its prediction from earlier projected displacements and withholds the newest displacement until that prediction has been formed. With trained parameters fixed, every warning $s_{t,h}^{H}$ or $s_{t,h}^{H+A}$, $A \in \{R, T, RT\}$, is therefore $\mathcal{F}_t$-measurable. The evaluation target $Y_{t,h} = \mathcal{D}_h(U_{>t})$ is defined from utility after t and is never supplied as warning-time evidence. The theory below concerns what can be inferred inside this boundary.

D.2 Recent Utility And Prospective Incompleteness

Recent utility is a meaningful warning source and should not be treated as a vacuous control. Let

$$\mathbf{u}_t^{(m)} = (u_{t-m+1}, \ldots, u_t)^\top \tag{71}$$

collect the last m realized utility values. The first result states why this history can be predictive whenever utility persists over time.

Proposition D.1 (Predictive value of utility persistence)**.** *Assume u_{t+h} and $\mathbf{u}_t^{(m)}$ have finite second moments. Let Σ_u be the covariance matrix of $\mathbf{u}_t^{(m)}$, and let $c_h = \mathrm{Cov}(\mathbf{u}_t^{(m)}, u_{t+h})$. If Σ_u is positive definite, the smallest mean-squared error among affine predictors based on recent utility is*

$$\mathrm{MSE}_U^\star = \mathrm{Var}(u_{t+h}) - c_h^\top \Sigma_u^{-1} c_h. \tag{72}$$

When $c_h \neq 0$, this error is strictly smaller than that of the best constant predictor.

Proof. Center the variables, with the affine intercept absorbed by their means. For a linear coefficient w, the centered prediction error is

$$\mathrm{Var}(u_{t+h}) - 2w^\top c_h + w^\top \Sigma_u w. \tag{73}$$

Completing the square gives the minimizer $w = \Sigma_u^{-1} c_h$ and Eq. 72. Positive definiteness implies $c_h^\top \Sigma_u^{-1} c_h > 0$ whenever $c_h \neq 0$. □

Proposition D.1 establishes that recent utility can carry genuine prospective information. This matches the empirical role of U as a substantive control rather than a deliberately weak baseline. The central question is narrower. Does recent utility necessarily expose every local change that matters for future utility?

To make that question precise, consider a reference causal history and a nearby admissible variation. Let $\mathbf{x}_t \in \mathbb{R}^D$ provide local coordinates for a family of causal histories $\mathcal{H}_{\leq j}(\mathbf{x})$ that includes every prefix required to compute $u_{t-m+1}, \ldots, u_t$. The same frozen personalization model processes every member of this family:

$$\begin{aligned} z_j(\mathbf{x}) &= \mathcal{M}_\theta(\mathcal{H}_{\leq j}(\mathbf{x})), \\ \hat{y}_{j+1}(\mathbf{x}) &= \mathcal{G}_\phi(z_j(\mathbf{x}), q_{j+1}(\mathbf{x})), \\ u_{j+1}(\mathbf{x}) &= \mathcal{U}(\hat{y}_{j+1}(\mathbf{x}), y^\star_{j+1}(\mathbf{x})). \end{aligned} \tag{74}$$

The task context q and target $y^\star$ follow the same task and evaluation rule throughout the family. Their future realizations are not warning inputs. Define

$$\begin{aligned} \Psi_t(\mathbf{x}) &:= \mathbf{u}_t^{(m)}(\mathbf{x}), \\ \mathcal{C}_t^{(W)}(\mathbf{x}) &:= \langle (b_\tau(\mathbf{x}), z_\tau(\mathbf{x})) \rangle_{\tau=t-W+1}^{t}, \\ \mathbf{Z}_t(\mathbf{x}) &:= \mathrm{vec}\, \langle z_\tau(\mathbf{x}) \rangle_{\tau=t-W+1}^{t}, \\ \mu_{t,h}(\mathbf{x}) &:= \mathbb{E}[u_{t+h} \mid \mathbf{x}_t = \mathbf{x}]. \end{aligned} \tag{75}$$

Here, Ψ_t is the recent-utility trace, $\mathbf{Z}_t$ is the state component of the History-State record, and $\mu_{t,h}$ is expected future utility under the specified continuation and task-evaluation process. The Euclidean coordinates are only a local analytical device. A discrete collection of histories need not admit such a smooth family globally, and the comparison below applies only where both $\mathbf{x}_t$ and $\mathbf{x}_t + \epsilon v$ correspond to admissible histories.

The key condition is that expected future utility changes in at least one local direction that the recent-utility Jacobian does not detect to first order. Under this condition, the different orders in ϵ follow directly.

Theorem D.2 (Bounded prospective observability gap)**.** *Fix $\mathbf{x}_t$ and horizon h. Suppose an admissible neighborhood contains the ball $B_r(\mathbf{x}_t)$, Ψ_t and $\mu_{t,h}$ are continuously differentiable on this ball, their Jacobian and gradient are respectively L_U- and L_F-Lipschitz, and*

$$\nabla \mu_{t,h}(\mathbf{x}_t) \notin \mathrm{Row}(J_{\Psi_t}(\mathbf{x}_t)). \tag{76}$$

Then there exists a unit direction v for which recent utility has zero first-order change while expected future utility decreases. In particular,

$$J_{\Psi_t}(\mathbf{x}_t) v = 0, \qquad a := -\nabla \mu_{t,h}(\mathbf{x}_t)^\top v > 0. \tag{77}$$

For every $0 < \epsilon \leq r$ satisfying $L_F \epsilon \leq a$,

$$\|\Psi_t(\mathbf{x}_t + \epsilon v) - \Psi_t(\mathbf{x}_t)\|_2 \leq \frac{L_U}{2} \epsilon^2, \tag{78}$$

while

$$\mu_{t,h}(\mathbf{x}_t) - \mu_{t,h}(\mathbf{x}_t + \epsilon v) \geq \frac{a}{2} \epsilon. \tag{79}$$

Proof. Write $J = J_{\Psi_t}(\mathbf{x}_t)$ and $g = \nabla \mu_{t,h}(\mathbf{x}_t)$. Since $\mathrm{Row}(J)^\perp = \ker(J)$, condition 76 implies that the orthogonal projection $g_\perp$ of g onto $\ker(J)$ is nonzero. Set

$$v = -\frac{g_\perp}{\|g_\perp\|_2}. \tag{80}$$

Then $\|v\|_2 = 1$, $Jv = 0$, and $g^\top v = -\|g_\perp\|_2$. Hence Eq. 77 holds with $a = \|g_\perp\|_2 > 0$.

Because $Jv = 0$, the fundamental theorem of calculus gives

$$\begin{aligned} \Psi_t(\mathbf{x}_t + \epsilon v) - \Psi_t(\mathbf{x}_t) &= \int_0^\epsilon J_{\Psi_t}(\mathbf{x}_t + rv) v \, dr \\ &= \int_0^\epsilon \left[J_{\Psi_t}(\mathbf{x}_t + rv) - J_{\Psi_t}(\mathbf{x}_t) \right] v \, dr. \end{aligned} \tag{81}$$

The L_U-Lipschitz Jacobian therefore yields

$$\|\Psi_t(\mathbf{x}_t + \epsilon v) - \Psi_t(\mathbf{x}_t)\|_2 \le \int_0^\epsilon L_U r \, dr = \frac{L_U}{2}\epsilon^2. \tag{82}$$

For expected future utility,

$$\begin{aligned} \mu_{t,h}(\mathbf{x}_t + \epsilon v) - \mu_{t,h}(\mathbf{x}_t) &= \int_0^\epsilon \nabla \mu_{t,h}(\mathbf{x}_t + rv)^\top v \, dr \\ &\le -a\epsilon + \int_0^\epsilon L_F r \, dr \\ &= -a\epsilon + \frac{L_F}{2}\epsilon^2. \end{aligned} \tag{83}$$

Rearranging and using $L_F \epsilon \le a$ gives Eq. 79. □

Theorem D.2 concerns recent utility and expected future utility. To connect the same local direction to the monitored History-State evidence, the state component must also respond to that direction.

Corollary D.3 (History-State visibility of utility-hidden change)**.** *Under Theorem D.2, assume $\mathbf{Z}_t$ is differentiable at $\mathbf{x}_t$ and*

$$D\mathbf{Z}_t(\mathbf{x}_t) v \neq 0. \tag{84}$$

Let $d_{\mathcal{C}}$ be a product metric on the paired History-State record that dominates its state component with weight $\lambda_Z > 0$, meaning

$$d_{\mathcal{C}}\left(\mathcal{C}_t^{(W)}(\mathbf{x}), \mathcal{C}_t^{(W)}(\mathbf{x}') \right) \ge \lambda_Z \left\| \mathbf{Z}_t(\mathbf{x}) - \mathbf{Z}_t(\mathbf{x}') \right\|_2. \tag{85}$$

Then, for all sufficiently small admissible $\epsilon > 0$, there exists $c_Z > 0$ such that

$$d_{\mathcal{C}}\left(\mathcal{C}_t^{(W)}(\mathbf{x}_t + \epsilon v), \mathcal{C}_t^{(W)}(\mathbf{x}_t) \right) \ge \lambda_Z c_Z \epsilon. \tag{86}$$

For the same variation, recent utility obeys Eq. 78 while expected future utility falls by at least $a\epsilon/2$.

Proof. Differentiability gives

$$\mathbf{Z}_t(\mathbf{x}_t + \epsilon v) - \mathbf{Z}_t(\mathbf{x}_t) = \epsilon D\mathbf{Z}_t(\mathbf{x}_t) v + o(\epsilon). \tag{87}$$

Since $D\mathbf{Z}_t(\mathbf{x}_t) v \neq 0$, for all sufficiently small positive ϵ,

$$\|\mathbf{Z}_t(\mathbf{x}_t + \epsilon v) - \mathbf{Z}_t(\mathbf{x}_t)\|_2 \ge c_Z \epsilon, \qquad c_Z = \frac{1}{2}\|D\mathbf{Z}_t(\mathbf{x}_t) v\|_2 > 0. \tag{88}$$

Equation 85 then gives Eq. 86. The recent-utility and expected-future-utility bounds follow from Theorem D.2. □

Corollary D.3 gives the three scales summarized in Section 3. With $\mathcal{C}_{t,\epsilon}^{(W)} = \mathcal{C}_t^{(W)}(\mathbf{x}_t + \epsilon v)$, $\mathbf{u}_{t,\epsilon}^{(m)} = \Psi_t(\mathbf{x}_t + \epsilon v)$, and $\bar{u}_{t,h}^{(\epsilon)} = \mu_{t,h}(\mathbf{x}_t + \epsilon v)$, the monitored state record changes at least to first order, expected future utility decreases at least to first order, and recent utility changes at most to second order.

Corollary D.4 (Relative local sensitivity)**.** *Under Theorem D.2, if $L_U > 0$ and the recent-utility difference is nonzero,*

$$\frac{\mu_{t,h}(\mathbf{x}_t) - \mu_{t,h}(\mathbf{x}_t + \epsilon v)}{\|\Psi_t(\mathbf{x}_t + \epsilon v) - \Psi_t(\mathbf{x}_t)\|_2} \geq \frac{a}{L_U \epsilon}. \tag{89}$$

Proof. The numerator is at least $a\epsilon/2$ by Eq. 79, while the positive denominator is at most $L_U\epsilon^2/2$ by Eq. 78. Dividing the two bounds gives Eq. 89. If the recent-utility difference is zero, utility is exactly unchanged along the comparison and no ratio is required. □

Ranking metrics provide a concrete reason that immediate utility can be locally flat even when the hidden state moves. A sufficiently small state change cannot alter a ranking whose pairwise score margins remain open.

Corollary D.5 (Local invariance of ranking utility)**.** *For a fixed candidate set and task context q_{t+1}, let $s(z) \in \mathbb{R}^N$ be the item-score vector produced by $\mathcal{G}_\phi(z, q_{t+1})$. Suppose the reference scores have minimum pairwise margin*

$$\gamma_t := \min_{i \neq j} |s_i(z_t) - s_j(z_t)| > 0 \tag{90}$$

and

$$\|s(z) - s(z')\|_\infty \leq L_s \|z - z'\|_2 \tag{91}$$

for $L_s > 0$. If

$$\|z' - z_t\|_2 < \frac{\gamma_t}{2L_s}, \tag{92}$$

the candidate ranking is unchanged. Under fixed relevance labels, ranking utilities depending only on that order, including MRR and nDCG, are therefore unchanged.

Proof. Every item score changes by less than $\gamma_t/2$. For any pair, the combined perturbation of its score difference is therefore smaller than the original pairwise margin. No pair can reverse order, so the full ranking and every utility depending only on that ranking remain unchanged. □

Interpretation for RQ1 and the observed separation. Theorem D.2 and Corollary D.3 identify a specific local regime in which recent utility can react much less strongly than a future-quality-relevant change visible in the observed state record. The conclusion is conditional. Equation 76 must provide a direction that recent utility does not detect to first order, and the observed states must respond along that same direction. The result concerns expected future utility rather than the binary degradation target $Y_{t,h}$, and it does not prove that the interaction component adds information beyond the observed states.

The RQ1 results exhibit the empirical pattern for which this separation is relevant. On the supported MovieLens models, learned History-State warning exceeds the matched recent-utility warning by $+0.106$ mean nAUPRC and does so in every one of the $160/160$ encoder-sentry-horizon comparisons. The direction also recurs on PENS with $67/80$ positive comparisons and on Amazon with $21/24$, and the mean $H - U$ advantage remains positive at every evaluated horizon in all three domains. Amazon also provides an important boundary. Its absolute mean H nAUPRC remains slightly negative even though $H - U$ is positive. The experiment therefore does not say that theoretically visible evidence must be easy to predict from. It shows that learned History-State warning can exploit prospective information that a matched utility-history warning misses even in a setting where absolute warning remains difficult.

Corollary D.5 gives a concrete mechanism for one part of this possibility in ranking systems. A hidden-state change can occur without changing MRR or nDCG while score margins preserve the current ranking. The experiments do not establish that ranking plateaus cause the observed $H - U$ gains, and Corollary D.4 does not predict the empirical horizon curve because warning horizon h is not perturbation size ϵ. The theory and evidence together support the narrower conclusion that recent output quality need not exhaust the prospective information exposed by the evolving frozen host. Section 6.1 reports the primary findings, and Appendix G.2 gives the complete matched evidence.

D.3 Regularity And Transition As Conditional Refinements

RQ2 begins only after the broad History-State sentry and the Regularity and Transition precursors have been trained. The precursors use causal projected-state trajectories without future-degradation supervision and are then fixed. The remaining theoretical question is conditional. Given the representation already learned by H, what prospective information can structured evidence still contribute?

Let $V_t^H := r_t^H$ denote the frozen History-State sentry representation used by the structured adapter, and let A_t denote $\mathbf{R}_t$, $\mathbf{T}_t$, or their joint pair $(\mathbf{R}_t, \mathbf{T}_t)$. To contain every causal quantity needed by these objects, write

$$X_t := (\mathcal{I}_t, \mathcal{I}_{t-1}). \tag{93}$$

With all precursor and parent parameters fixed, both V_t^H and A_t are deterministic functions of X_t. The future target remains $Y_{t,h}$.

Theorem D.6 (Conditional value of structured state evidence)**.** *Fix t, h and abbreviate $X = X_t$, $V = V_t^H$, $A = A_t$, and $Y = Y_{t,h}$. Assume the required regular conditional distributions exist and that the displayed odds and likelihood ratios are defined on the support under consideration. Then*

$$\mathrm{I}(Y; A) \le \mathrm{I}(Y; X), \qquad \mathrm{I}(Y; A \mid X) = 0, \tag{94}$$

and

$$0 \le \mathrm{I}(Y; A \mid V) \le \mathrm{I}(Y; X \mid V). \tag{95}$$

Define the Bayes-optimal logits

$$\ell_V^\star(v) = \log \frac{\mathbb{P}(Y = 1 \mid V = v)}{\mathbb{P}(Y = 0 \mid V = v)} \tag{96}$$

and

$$\ell_{V+A}^\star(v, a) = \log \frac{\mathbb{P}(Y = 1 \mid V = v, A = a)}{\mathbb{P}(Y = 0 \mid V = v, A = a)}. \tag{97}$$

Then

$$\ell_{V+A}^\star(v, a) = \ell_V^\star(v) + \Delta_A^\star(v, a), \tag{98}$$

where

$$\Delta_A^\star(v, a) = \log \frac{p(a \mid Y = 1, V = v)}{p(a \mid Y = 0, V = v)}. \tag{99}$$

For the smallest achievable population log-loss $\mathcal{R}_{\log}^\star$ using the indicated evidence,

$$\mathcal{R}_{\log}^\star(V) - \mathcal{R}_{\log}^\star(V, A) = \mathrm{I}(Y; A \mid V) \le \mathrm{I}(Y; X \mid V). \tag{100}$$

Proof. Since A is a deterministic function of X once the trained precursor is fixed, $Y \to X \to A$ forms a Markov chain under the data-generating process. Hence $\mathrm{I}(Y; A \mid X) = 0$, and the data-processing inequality gives $\mathrm{I}(Y; A) \le \mathrm{I}(Y; X)$ (Cover & Thomas, 2006). Because V and A are both functions of X,

$$\mathrm{I}(Y; X, A \mid V) = \mathrm{I}(Y; X \mid V). \tag{101}$$

Applying the conditional chain rule in the other order gives

$$\mathrm{I}(Y; X \mid V) = \mathrm{I}(Y; A \mid V) + \mathrm{I}(Y; X \mid V, A). \tag{102}$$

The final term is nonnegative, which proves Eq. 95.

For the odds decomposition, Bayes' rule gives

$$\frac{\mathbb{P}(Y = 1 \mid V = v, A = a)}{\mathbb{P}(Y = 0 \mid V = v, A = a)} = \frac{\mathbb{P}(Y = 1 \mid V = v)\, p(a \mid Y = 1, V = v)}{\mathbb{P}(Y = 0 \mid V = v)\, p(a \mid Y = 0, V = v)}. \tag{103}$$

Taking logarithms yields Eqs. 98 and 99. Finally, the minimum expected log-loss given evidence Z equals $\mathrm{H}(Y \mid Z)$ because any alternative conditional prediction adds a nonnegative expected KL divergence (Gneiting & Raftery, 2007). Therefore

$$\mathcal{R}_{\log}^\star(V) - \mathcal{R}_{\log}^\star(V, A) = \mathrm{H}(Y \mid V) - \mathrm{H}(Y \mid V, A) = \mathrm{I}(Y; A \mid V), \tag{104}$$

which completes the proof. □

The theorem gives an exact population interpretation of incremental structured information. Once the frozen H representation is known, only the part of A that remains informative about Y can improve the Bayes-optimal log-loss.

Corollary D.7 (No guaranteed contribution from a structured view)**.** *If*

$$Y \perp A \mid V, \tag{105}$$

then

$$\mathrm{I}(Y; A \mid V) = 0 \tag{106}$$

and

$$\Delta_A^\star(V, A) = 0 \tag{107}$$

almost surely wherever the conditional odds are defined.

Proof. Conditional independence makes the two class-conditional laws of A given V identical on the relevant support. Their likelihood ratio is one, its logarithm is zero, and the conditional mutual information vanishes. □

Regularity and Transition can also carry overlapping, redundant, or interaction-specific information. Their separate contributions therefore need not sum to their joint contribution.

Corollary D.8 (Regularity and Transition are not generally additive)**.** *Let $R = \mathbf{R}_t$ and $T = \mathbf{T}_t$. Then*

$$\mathrm{I}(Y; R, T \mid V) = \mathrm{I}(Y; R \mid V) + \mathrm{I}(Y; T \mid V, R). \tag{108}$$

The Bayes-optimal joint correction satisfies

$$\begin{aligned} \Delta_{RT}^\star(v, r, t) &= \log \frac{p(r, t \mid Y = 1, V = v)}{p(r, t \mid Y = 0, V = v)} \\ &= \Delta_R^\star(v, r) + \log \frac{p(t \mid r, Y = 1, V = v)}{p(t \mid r, Y = 0, V = v)}. \end{aligned} \tag{109}$$

The second term generally differs from the correction obtained from T without conditioning on R. Separate R and T corrections therefore need not add to the optimal joint correction.

Proof. Equation 108 is the conditional mutual-information chain rule. For the correction, factor

$$p(r, t \mid Y, V) = p(r \mid Y, V) p(t \mid r, Y, V) \tag{110}$$

inside the class-conditional likelihood ratio and take logarithms. The remaining T term conditions on R, which is precisely why it need not equal a separately fitted T correction. □

Corollary D.9 (Interpretation of staged `CoTRAC` training)**.** *The implemented structured warning*

$$\ell_{t,h}^{H+A} = \ell_{t,h}^{H} + \alpha_t^A \Delta_{t,h}^A, \qquad A \in \{R, T, RT\}, \tag{111}$$

uses the same baseline-plus-correction structure as Eq. 98. The correspondence is architectural rather than an equality between learned and Bayes-optimal quantities. The parent H is finite-sample trained and then frozen, the adapter receives (V_t^H, A_t), the scalar gate restricts the expressed correction, and finite optimization need not recover $\Delta_A^\star$. The RT adapter can model interactions between Regularity and Transition rather than imposing additive separate corrections.

Remark D.1 (Scope of the refinement result). Theorem D.6 isolates the population information available in A after conditioning on the full frozen sentry representation V. The implemented $H + A$ adapter also receives V, while the empirical H baseline uses its already trained risk head. Consequently, a finite-sample $H + A$ versus H difference is an effect of the complete structured-refinement architecture and is not a direct estimator of $\mathrm{I}(Y; A \mid V)$. The theorem also provides no ordering of finite-sample AUPRC, EventRecall, MeanLead, or false-alarm rates.

Interpretation for RQ2 and the observed structured effects. Theorem D.6 gives structured evidence a deliberately conditional role. Once the frozen History-State representation V is already known, the population log-loss value remaining in a structured view after conditioning on V is governed by $\mathrm{I}(Y; A \mid V)$. Corollary D.7 therefore allows a structured view to add nothing, while Corollary D.8 shows why separately useful Regularity and Transition evidence need not become more useful when combined.

The held-out results exhibit both kinds of behavior. On PENS, Transition gives the largest structured discrimination change for four of the five hosts and Regularity for the remaining host, while RT is best for none. On MovieLens, RT is strongest for some hosts and an individual view is strongest for others. Across all 160 MovieLens host-sentry-horizon cells, RT exceeds both individual views in only $57/160$ cells and is lower than the stronger individual refinement by 1.139×10^{-3} nAUPRC on average. On Amazon, the unmodified H warning remains best overall for Caser, while Regularity is favored for GRU4Rec and SR-GNN. The joint representation is therefore not a monotone more-evidence-is-better construction in practice.

The operational results sharpen the same point. PENS refinements increase event coverage on average without increasing mean warning lead, while Amazon refinements increase warning lead together with false-alarm burden. Conditional usefulness is therefore both host-dependent and criterion-dependent. The theorem does not predict which structured view will help a particular host or metric, and the empirical comparisons do not directly estimate conditional mutual information. Section 6.2 reports the primary findings, and Appendix G.3 gives the complete matched comparisons.

D.4 Calibration-Conditioned Non-Equivalence Of Warning Criteria

RQ2 evaluates more than discrimination. A useful warning may detect additional degradation episodes, arrive earlier for already detected episodes, or change false-alarm burden. These quantities depend on different aspects of the score sequence and therefore need not improve together.

For a warning method $M \in \{H, H+R, H+T, H+RT\}$, let pre-test calibration data $\mathcal{D}_{\mathrm{cal}}$ and a fixed calibration rule produce a finite threshold

$$\tau_M = \kappa_M(\mathcal{D}_{\mathrm{cal}}). \tag{112}$$

After calibration is complete, the held-out alarm at eligible position i is

$$a_i^M = \mathbf{1}\{s_i^M \geq \tau_M\}. \tag{113}$$

Let $N_E > 0$ be the number of degradation episodes. A detected episode has a first valid pre-onset alarm and lead $\ell_e^M \in [0, L_{\max}]$. Let $N_0 > 0$ denote the number of eligible non-event positions.

Proposition D.10 (Calibration-rule-agnostic non-equivalence of warning criteria)**.** *For any fixed finite pre-test thresholds, there is no universal monotone relationship among AUPRC, EventRecall, MeanLead, and realized false-alarm rate.*

Event coverage and lead. *Suppose a warning detects $k > 0$ episodes with mean lead L. If a refinement preserves the first-alarm times of those episodes and additionally detects $m > 0$ episodes with mean lead $\bar{\ell}_{\mathrm{new}}$, then*

$$\Delta\mathrm{EventRecall} = \frac{m}{N_E} > 0, \tag{114}$$

while

$$\Delta\mathrm{MeanLead} = \frac{m}{k+m}\left(\bar{\ell}_{\mathrm{new}} - L\right). \tag{115}$$

Consequently,

$$|\Delta\mathrm{MeanLead}| \leq \frac{m}{k+m} L_{\max}. \tag{116}$$

Lead and alarm burden. *If the first alarms of the same k detected episodes move earlier by $\delta_1, \ldots, \delta_k$, then*

$$\Delta\mathrm{MeanLead} = \frac{1}{k}\sum_{j=1}^{k} \delta_j. \tag{117}$$

If F of the N_0 eligible non-event positions raise false alarms, then $\mathrm{FAR} = F/N_0$. *Adding f false alarms changes it by*

$$\Delta\mathrm{FAR} = \frac{f}{N_0}. \tag{118}$$

Discrimination and thresholded alarms. *There exist finite score configurations with identical AUPRC and different event detection, warning lead, or false-alarm burden. There also exist finite configurations with identical alarm indicators and different AUPRC. Hence none of AUPRC, EventRecall, MeanLead, and realized FAR determines the others.*

Proof. The original k detected episodes contribute total lead kL. After adding m newly detected episodes, total lead becomes $kL + m\bar{\ell}_{\mathrm{new}}$ and the number of detected episodes becomes $k + m$. Subtracting L from the resulting mean gives Eq. 115. Since both L and $\bar{\ell}_{\mathrm{new}}$ lie in $[0, L_{\max}]$, Eq. 116 follows. Advancing the first alarm of an already detected episode adds exactly its advance to that episode's lead, which gives Eq. 117. False alarms are counted over a different set of eligible positions, so Eq. 118 follows independently.

For discrimination versus event detection, take two positive-labelled eligible positions belonging to different degradation episodes, with one score above the fixed threshold and the other below it. Exchange their scores. Since both exchanged ranks remain positive-labelled, the ranked binary label sequence and therefore AUPRC are unchanged. Arrange an additional valid alarm for the first episode but none for the second. Before the exchange only the first episode is detected. After the exchange the first remains detected through its additional alarm and the second becomes detected, so EventRecall changes while AUPRC does not.

For discrimination versus lead, use two positive-labelled positions within the same detected episode, one earlier and one later in its admissible pre-onset window, with exactly one score above the threshold. Exchanging those two scores preserves the ranked binary label sequence and AUPRC while moving the first alarm and therefore MeanLead. For discrimination versus false alarms, place a negative position below all positives and move its score across the fixed threshold while keeping it below every positive score. The ranked positive-negative sequence and AUPRC remain unchanged while FAR changes. Conversely, place one positive and one negative position on the same side of the threshold and exchange their score ordering without crossing that threshold. Their alarm indicators remain unchanged while the ranked positive-negative sequence and AUPRC can change. These constructions can be placed relative to any finite fixed threshold, so pre-test calibration does not impose a monotone relationship among the four criteria. □

Corollary D.11 (Non-equivalence for normalized AUPRC)**.** *For a fixed held-out population with $0 \le \pi < 1$,*

$$\mathrm{nAUPRC} = \frac{\mathrm{AUPRC} - \pi}{1 - \pi} \tag{119}$$

and

$$\Delta\mathrm{nAUPRC} = \frac{\Delta\mathrm{AUPRC}}{1 - \pi}. \tag{120}$$

The discrimination constructions in Proposition D.10 therefore carry over unchanged to nAUPRC.

Interpretation for the observed warning trade-offs. Proposition D.10 matters because the empirical results display the kind of separation it permits. On PENS, R, T, and RT increase EventRecall on average while MeanLead changes slightly in the opposite direction. Detecting additional degradation episodes therefore does not imply that the newly detected episodes are found earlier. On Amazon, structured warning increases MeanLead by approximately 0.014 to 0.020, while realized TEST FAR also rises by approximately 0.015 to 0.020. Earlier warning can therefore come with greater alarm burden.

The same separation appears outside the main RQ2 setting. Among the $26/48$ OpenLLM RT runs admitted by the pre-test calibration procedure, RT decreases AUPRC by 2.341×10^{-3} on average while increasing EventRecall by 4.073×10^{-3} and MeanLead by 0.093. This is a direct example of why discrimination, event coverage, and earliness cannot be collapsed into one notion of better warning. Proposition D.10 does not explain why these empirical changes occur. It establishes why observing them is mathematically coherent and why each criterion must be reported separately.

D.5 Portability Of Broad And Structured Warning Signatures

RQ3 studies source-only reuse by combining rank-normalized source warnings. The result below addresses a related distributional question: how source-target shift changes when a broad warning signature is augmented with structured evidence. It does not analyze the rank-averaging procedure itself and assumes a common degradation target across the source and target settings.

Let the common broad coordinate be the learned History-State warning

$$S^H := s^H_{t,h} \in [0,1]. \tag{121}$$

The scalar warning lies in a common numerical range across encoders even when their hidden states differ in dimension or coordinate system. This does not require raw source probabilities to be equally calibrated, which is why the empirical source-only protocol rank-normalizes them. Let

$$A \in \{\mathbf{R}_t, \mathbf{T}_t, (\mathbf{R}_t, \mathbf{T}_t)\} \tag{122}$$

denote additional structured evidence and let $Y := Y_{t,h} \in \{0,1\}$. For source and target settings $d \in \{s,t\}$, define

$$Q^H_d := \mathcal{L}_d(S^H, Y), \qquad Q^{H,A}_d := \mathcal{L}_d(S^H, A, Y). \tag{123}$$

Within each setting, Q^H_d is exactly the (S^H, Y) marginal of $Q^{H,A}_d$. Use compatible metrics

$$d_H\big((s,y),(s',y')\big) := d_S(s,s') + \lambda_Y |y - y'| \tag{124}$$

and

$$\begin{aligned} d_{H,A}\big((s,a,y),(s',a',y')\big) := d_S(s,s') + \lambda_A d_A(a,a') \\ + \lambda_Y |y - y'|, \end{aligned} \tag{125}$$

where $\lambda_A, \lambda_Y > 0$ and all laws have finite first moments. Projection that removes A is then 1-Lipschitz.

Lemma D.12 (Wasserstein contraction under a Lipschitz map)**.** *Let P and Q be probability laws with finite first moments, and let f be L_f-Lipschitz. Then*

$$W_1(f_\# P, f_\# Q) \le L_f W_1(P,Q), \tag{126}$$

where $f_\# P$ denotes the pushforward law of $f(X)$ for $X \sim P$.

Proof. Take any coupling γ of P and Q. Applying f to both components produces a coupling of $f_\# P$ and $f_\# Q$ whose expected transport cost is at most

$$L_f \int d(x,x')\, d\gamma(x,x'). \tag{127}$$

Taking the infimum over all couplings proves Eq. 126. □

The portability comparison is now a direct consequence of projection. Under compatible product metrics, adding structured coordinates cannot make the Wasserstein distance smaller than that of the retained broad marginal.

Theorem D.13 (Wasserstein portability hierarchy of nested warning signatures)**.** *For every richer signature retaining exactly the same S^H and Y,*

$$W_1(Q^H_s, Q^H_t) \le W_1(Q^{H,A}_s, Q^{H,A}_t). \tag{128}$$

If a joint signature retains separate Regularity and Transition coordinates under compatible additive product metrics, define the corresponding laws $Q^{H,R}_d$, $Q^{H,T}_d$, and $Q^{H,R,T}_d$. Then

$$\begin{aligned} W_1(Q^H_s, Q^H_t) &\le W_1(Q^{H,R}_s, Q^{H,R}_t) \\ &\le W_1(Q^{H,R,T}_s, Q^{H,R,T}_t), \end{aligned} \tag{129}$$

and

$$\begin{aligned} W_1(Q^H_s, Q^H_t) &\le W_1(Q^{H,T}_s, Q^{H,T}_t) \\ &\le W_1(Q^{H,R,T}_s, Q^{H,R,T}_t). \end{aligned} \tag{130}$$

Proof. The projection $\pi_H(s, a, y) = (s, y)$ is 1-Lipschitz under Eqs. 124 and 125. Moreover,

$$Q_d^H = (\pi_H)_\# Q_d^{H,A} \tag{131}$$

for each $d \in \{s, t\}$. Lemma D.12 therefore gives Eq. 128. The remaining inequalities follow by projecting (S^H, R, T, Y) onto (S^H, R, Y) or (S^H, T, Y) and applying the same contraction argument. □

The theorem compares distributional shift in nested signatures. Its immediate consequence is a corresponding ordering of worst-case discrepancies over Lipschitz test functions.

Corollary D.14 (Worst-case transfer discrepancy)**.** *For probability laws P, Q, define*

$$\Gamma_K(P, Q) := \sup_{\mathrm{Lip}(\varphi) \le K} |\mathbb{E}_P \varphi - \mathbb{E}_Q \varphi| \,. \tag{132}$$

Under finite first moments, Kantorovich–Rubinstein duality gives

$$\Gamma_K(P, Q) = K W_1(P, Q) \tag{133}$$

for $K > 0$ (Villani, 2009). Hence

$$\Gamma_K(Q_s^H, Q_t^H) \le \Gamma_K(Q_s^{H,A}, Q_t^{H,A}). \tag{134}$$

Proof. Rescale the 1-Lipschitz test functions in the Kantorovich–Rubinstein dual representation by K and apply Theorem D.13. □

The same comparison yields a bound on changes in expected loss when the corresponding loss-composed predictors satisfy a common Lipschitz condition.

Corollary D.15 (No-looser worst-case transfer-risk bound for the broad signature)**.** *Let f_H act on S^H and let $f_{H,A}$ act on (S^H, A). Define*

$$\varphi_H(s, y) = \mathcal{L}(f_H(s), y) \tag{135}$$

and

$$\varphi_{H,A}(s, a, y) = \mathcal{L}(f_{H,A}(s, a), y). \tag{136}$$

If both functions are K-Lipschitz under their respective metrics, then

$$\begin{aligned} |\mathcal{R}_t(f_H) - \mathcal{R}_s(f_H)| &\le K W_1(Q_s^H, Q_t^H), \\ |\mathcal{R}_t(f_{H,A}) - \mathcal{R}_s(f_{H,A})| &\le K W_1(Q_s^{H,A}, Q_t^{H,A}), \end{aligned} \tag{137}$$

where each $\mathcal{R}_d$ is the expected loss under the corresponding domain law. The displayed upper bound for the broad signature is therefore no larger than the displayed bound for the richer nested signature.

Proof. Apply Corollary D.14 to φ_H and $\varphi_{H,A}$ separately, then compare their Wasserstein terms using Theorem D.13. □

The conclusion concerns the size of these worst-case bounds. It does not order realized target risks, AUPRC values, or any particular source-only ensemble.

Corollary D.16 (Sufficient condition for a strict portability gap)**.** *Let*

$$Q_d^A := \mathcal{L}_d(A, Y) \tag{138}$$

use the corresponding marginal product metric. If

$$W_1(Q_s^A, Q_t^A) > W_1(Q_s^H, Q_t^H), \tag{139}$$

then

$$W_1(Q_s^{H,A}, Q_t^{H,A}) > W_1(Q_s^H, Q_t^H). \tag{140}$$

Proof. Projection from (S^H, A, Y) to (A, Y) is also 1-Lipschitz, so

$$W_1(Q_s^A, Q_t^A) \le W_1(Q_s^{H,A}, Q_t^{H,A}). \tag{141}$$

Combining this inequality with the assumed strict inequality gives the result. □

The preceding results compare nested warning signatures without requiring the structured coordinates themselves to be representation invariant. Regularity and Transition have different invariance properties, which makes this distinction relevant.

Proposition D.17 (Coordinate sensitivity of structured geometry)**.** *Suppose two monitoring coordinate systems describe the same local projected-state trajectory through*

$$\zeta'_j = B\zeta_j + c, \tag{142}$$

where B is invertible. Then the corresponding displacements, accelerations, and discrete jerks satisfy

$$d'_j = Bd_j, \qquad a'_j = Ba_j, \qquad a'_j - a'_{j-1} = B(a_j - a_{j-1}). \tag{143}$$

Consequently, the Euclidean-norm and directional quantities used by Regularity generally depend on $B^\top B$. If $B = Q$ is orthogonal, the raw Regularity quantities are unchanged, and applying the same Q to the training trajectories also preserves the training-derived Regularity targets. For a general invertible B, no such invariance is guaranteed.

Proof. Translation cancels from every finite difference. Under a general linear transformation,

$$\|Bd\|_2^2 = d^\top B^\top Bd, \tag{144}$$

and

$$(Bd_i)^\top (Bd_j) = d_i^\top B^\top Bd_j. \tag{145}$$

Hence the speed, directional, acceleration, and jerk quantities can change under a general invertible transformation. When $B = Q$ is orthogonal, $Q^\top Q = I$, so Euclidean norms and inner products are preserved. The resulting Regularity quantities, and therefore the training-derived targets constructed from them, remain unchanged. □

Transition has a narrower invariance because predictive uncertainty is represented coordinate-wise. Let $e_t = y_t - \mu_t$ and $D_t = \mathrm{diag}(V_{t,1}, \ldots, V_{t,128})$. Its standardized residual can be written as

$$(r_t^T)^2 = \frac{1}{128} e_t^\top (D_t + \epsilon I)^{-1} e_t. \tag{146}$$

Corollary D.18 (Invariance boundary of Transition scalars)**.** *Let P be a signed permutation matrix, and suppose the realized displacement, predicted mean, and diagonal predictive variance transform as*

$$y'_t = Py_t, \qquad \mu'_t = P\mu_t, \qquad D'_t = PD_tP^\top. \tag{147}$$

Then the standardized residual r_t^T, alignment s_t^T, directional mismatch, magnitude mismatch, predictive uncertainty, and orthogonal residual are unchanged. More generally, the directional, magnitude, and orthogonal quantities are invariant under orthogonal transformations of y_t and μ_t. The standardized residual is not generally invariant under an arbitrary rotation when the transformed covariance cannot be represented by the diagonal uncertainty model.

Proof. A signed permutation is orthogonal and therefore preserves Euclidean norms and inner products. Moreover,

$$D'_t + \epsilon I = P(D_t + \epsilon I)P^\top. \tag{148}$$

Hence

$$(e'_t)^\top (D'_t + \epsilon I)^{-1} e'_t = e_t^\top (D_t + \epsilon I)^{-1} e_t, \tag{149}$$

which preserves r_t^T and therefore s_t^T. Signed permutations also preserve the mean predictive variance, while the same norm and inner-product argument preserves the directional, magnitude, and orthogonal quantities.

The result concerns the Transition quantities conditional on the predictor outputs transforming as stated. It does not establish equivariance of the trained predictor or its learned representation. Because predictive uncertainty is represented diagonally, a general rotation of a non-isotropic covariance need not remain representable in the same parameterization. □

Interpretation for RQ3 and the observed portability split. Theorem D.13 makes a deliberately limited statement. Under compatible nested metrics and a common target definition, discarding additional structured coordinates cannot increase source-target Wasserstein distance. The corresponding worst-case Lipschitz transfer bound for the broad signature is therefore no looser. The theorem does not imply lower realized target risk, higher target AUPRC, or superior rank-averaged transfer for H.

The source-only results make the broad-versus-structured distinction concrete. Broad H warning remains positively discriminative in all $160/160$ MovieLens and $80/80$ PENS source-only cells, retaining 86.45% and 83.58% of self-warning discrimination on average. This reuse requires neither aligned hidden coordinates nor shared source calibration because the empirical protocol transfers rank-normalized warning orderings. Finer state-dynamics effects are less uniform: relative to transferred H, the mean transferred R, T, and RT effects are $(-1.737, -2.005, -1.483) \times 10^{-3}$ nAUPRC on MovieLens and $(+1.473, +1.400, +1.649) \times 10^{-3}$ on PENS.

Proposition D.17 and Corollary D.18 explain why broad-warning portability and structured-mechanism portability should be treated separately. Regularity is preserved under orthogonal coordinate changes but not under arbitrary invertible reparameterizations, while Transition combines orthogonally stable geometric quantities with a diagonal uncertainty normalization whose invariance is narrower. The empirical results also show that coordinate sensitivity is not a complete explanation. When the same PENS preference backbone is held fixed and only the downstream task changes from recommendation to generation, structured warning behavior changes. Pre-generation OpenLLMs likewise retain positive broad H discrimination while exhibiting different structured effects from the conventional generation decoders. These task and consumer comparisons fall outside the common-target Wasserstein result and indicate that structured portability can also depend on how the states are consumed and how degradation is defined.

The theory therefore provides a boundary rather than a numerical forecast. Projection gives the broad signature a no-looser worst-case Wasserstein bound under the stated conditions, while the experiments show substantial reuse of broad warning orderings and more context-dependent reuse of finer structured effects. Section 6.3 reports the primary findings, and Appendix G gives the complete source-only, task-change, and pre-generation evidence.

The prospective-separation result explains how recent utility can remain locally insensitive while future-relevant state evidence changes; RQ1 tests the corresponding empirical pattern. The conditional-refinement results explain why Regularity and Transition need not help uniformly or combine additively; RQ2 exhibits both behaviors. The metric result explains why discrimination, event coverage, warning lead, and false alarms must remain separate. Finally, the portability analysis separates reuse of the broad warning from reuse of finer structured evidence, matching the distinction observed in RQ3.

Remark. None of these empirical findings verifies the assumptions of the corresponding formal result. The theory identifies distinct behaviors permitted under explicit conditions, while the experiments establish which of those behaviors occur in the evaluated systems.

E Instantiations Of Temporal Monitors

Sections 3 and 4 define the scientific monitoring object independently of a particular early-warning architecture. We now show how each of the eight evaluated History-State sentries instantiates that object. The important invariant is the *information that is observed by each sentry monitors*. At observation time t, every sentry receives only two sources of host-side evidence: the causal interaction history available through t and the hidden preference states already produced by the frozen host from that history. The eight sentries differ in how this evidence is organized, temporally summarized, and mapped to prospective degradation risk. They do not receive future utility, future hidden states, or an independently reconstructed preference state.

E.1 Common Causal Input Contract

Let $\mathcal{H}_{\leq t} = \langle b_1, \ldots, b_t \rangle$ denote the causal interaction history and let $\mathcal{Z}_t^{(W)} = \langle z_{\tau_t}, \ldots, z_t \rangle$ denote the corresponding recent host-state trajectory, with $\tau_t = \max(1, t - W + 1)$. These are the same two evidence streams introduced in Appendix A.

Table 4: Eight History-State sentry instantiations of the common Temporal State-History Matching interface. These MovieLens sentries observe only causal interaction history and host states available through the observation time. Differences lie in the temporal backbone and prospective-risk parameterization.

Sentry	Causal Evidence Organization	Temporal Backbone	Risk Parameterization	Primary Inductive Bias
Prefix-MLP	Fixed prefix summary	MLP	Direct multi-horizon	Weak temporal structure
TCN-ELECTS-EW	Ordered recent trace + current evidence	Causal temporal CNN	Direct multi-horizon	Local temporal motifs
LS-EEP	Ordered causal state trace	Causal Transformer	Direct multi-horizon	Long-range past-state relations
TCSA	Recency-weighted causal trace	Causal Transformer	Direct multi-horizon	Recency-sensitive relations
TCSR	Ordered causal trace + current evidence	GRU	Discrete conditional hazard	Cumulative temporal risk
DRSA-style	Ordered causal trace + current evidence	GRU + future-time GRUCell	Recurrent conditional hazard	Sequential future-risk evolution
DeepHit-style	Ordered causal trace + current evidence	GRU	Discrete event-time distribution	Explicit future event-time mass
Temporal-Prior	Ordered trace + causal progress features	Time-gated GRU	Direct multi-horizon	Trajectory-position modulation

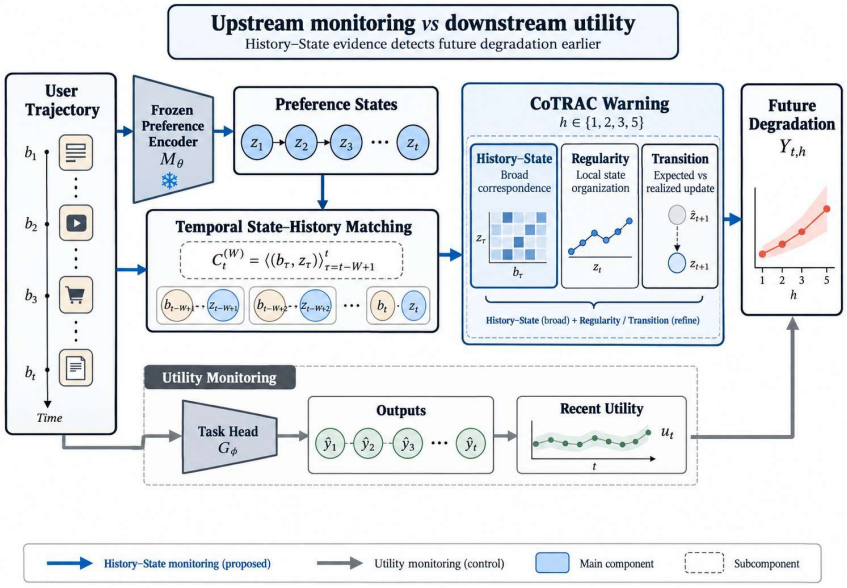


Figure 4: Instantiations of the common temporal analysis layer in CoTRAC. All three structured views operate on the same causal History-State evidence. History-State captures broad temporal correspondence, Regularity resolves local state-trajectory organization, and Transition compares expected and realized state evolution.

The host state z_t remains the preference representation produced by the frozen personalization host. A sentry may embed interaction fields, project state vectors, apply masks, or derive deterministic properties such as prefix length or timestep age. These operations belong to the early-warning sentry and do not define a second preference encoder.

For sentry architecture m, let Π_m denote this architecture-specific causal input organization. We write

$$X_t^{(m)} = \Pi_m \left(\mathcal{H}_{\le t}, \mathcal{Z}_t^{(W)} \right). \tag{150}$$

The map Π_m may project the observed fields, apply causal masks, or derive deterministic temporal coordinates, while remaining restricted to the same causal history and host-state evidence.

In the evaluated configuration, the recent state window has maximum length $W = 8$, with shorter prefixes masked accordingly. The window length and architecture-specific input organization are sentry design choices rather than assumptions of Temporal State-History Matching.

E.2 Common Sentry Factorization

Every evaluated sentry can be written as the composition

$$\mathcal{F}_m = \Gamma_m \circ R_m \circ B_m \circ \Pi_m, \tag{151}$$

where Π_m organizes the causal evidence, B_m is the architecture-specific temporal backbone, R_m is its native prospective-risk head, and Γ_m converts that native output into the common horizon-wise warning representation.

More explicitly,

$$\begin{aligned} X_t^{(m)} &= \Pi_m\left(\mathcal{H}_{\le t}, \mathcal{Z}_t^{(W)}\right), \\ \mathbf{q}_t^{(m)} &= B_m\left(X_t^{(m)}\right), \\ \boldsymbol{\xi}_t^{(m)} &= R_m\left(\mathbf{q}_t^{(m)}\right), \\ \boldsymbol{\ell}_t^{H,m} &= \Gamma_m\left(\boldsymbol{\xi}_t^{(m)}\right). \end{aligned} \tag{152}$$

Here, $\mathbf{q}_t^{(m)}$ is the internal representation of the early-warning sentry and should not be confused with the host preference state z_t. The canonical sentries expose a 128-dimensional sentry representation before the risk head.

For direct multi-horizon sentries, Γ_m is the identity map and $\boldsymbol{\xi}_t^{(m)}$ already contains the logits for the evaluated horizons. For survival- or event-time-based sentries, R_m first predicts a native hazard or event-time representation and Γ_m converts it to cumulative degradation risk at the common horizons.

Thus, all eight models satisfy the common interface

$$\boxed{\boldsymbol{\ell}_t^{H,m} = \mathcal{F}_m\left(\mathcal{H}_{\le t}, \mathcal{Z}_t^{(W)}\right)} \qquad m \in \mathcal{M}, \tag{153}$$

Equation 153 is the concrete architecture-level realization of the History-State abstraction. The sentries differ in Π_m, B_m, R_m, and Γ_m. They do not differ in what future information is made available.

E.3 Prefix-MLP: Weak Temporal Inductive Bias

Prefix-MLP provides the least structurally constrained neural realization in the sentry panel. Its purpose is to test whether a fixed-dimensional nonlinear summary of the available history-state evidence already contains prospective degradation information.

The causal evidence is materialized as a fixed-dimensional prefix vector:

$$\mathbf{x}_t^{\mathrm{PMLP}} = \Pi_{\mathrm{PMLP}}\left(\mathcal{H}_{\le t}, \mathcal{Z}_t^{(W)}\right). \tag{154}$$

The canonical backbone maps this vector through a feed-forward network with hidden width 256, LayerNorm and GELU transformations, and a residual feed-forward block, producing the common 128-dimensional sentry representation:

$$\mathbf{q}_t^{\mathrm{PMLP}} = B_{\mathrm{MLP}}\left(\mathbf{x}_t^{\mathrm{PMLP}}\right) \in \mathbb{R}^{128}. \tag{155}$$

Its prediction head directly produces the horizon-wise warning logits,

$$\boldsymbol{\ell}_t^{H,\mathrm{PMLP}} = W_{\mathrm{PMLP}}\mathbf{q}_t^{\mathrm{PMLP}} + \mathbf{b}_{\mathrm{PMLP}}. \tag{156}$$

Prefix-MLP therefore asks whether the available correspondence evidence is predictive even without convolutional, attention-based, recurrent, or survival-specific temporal structure. Its inclusion prevents the central state-history result from depending on a highly specialized sequential architecture.

E.4 TCN-ELECTS-EW: Local Causal Temporal Patterns

TCN-ELECTS-EW introduces an explicit local temporal inductive bias. The recent state trajectory is retained as an ordered sequence and processed with causal one-dimensional convolutions.

Let $\mathbf{S}_t$ denote the projected and masked state window produced from $\mathcal{Z}_t^{(W)}$. The canonical temporal path applies successive causal convolutions, including dilation 2 in the second temporal block:

$$\begin{aligned} \mathbf{H}_t^{(1)} &= \mathrm{CConv}_1\left(\mathbf{S}_t\right), \\ \mathbf{H}_t^{(2)} &= \mathrm{CConv}_2\left(\mathbf{H}_t^{(1)}\right). \end{aligned} \tag{157}$$

Causal padding ensures that the representation at any sequence position depends only on that position and earlier observed states.

The temporal endpoint representation is combined with a projection of the current causal evidence:

$$\mathbf{q}_t^{\mathrm{TCN}} = \phi_{\mathrm{TCN}}\left(\left[\mathbf{q}_t^{\mathrm{temp}}; \mathbf{q}_t^{\mathrm{current}}\right]\right), \tag{158}$$

where both terms are obtained solely from $\mathcal{H}_{\leq t}$ and $\mathcal{Z}_t^{(W)}$.

The final sentry representation is mapped to prospective warning through the canonical multi-horizon risk head:

$$\boldsymbol{\ell}_t^{H,\mathrm{TCN}} = R_{\mathrm{TCN}}\left(\mathbf{q}_t^{\mathrm{TCN}}\right). \tag{159}$$

This sentry tests whether short-range temporal motifs and local changes in the host-state trajectory provide a sufficient realization of state-history warning. The name TCN-ELECTS-EW denotes an ELECTS-inspired early-warning adaptation rather than an exact reproduction of the original ELECTS model.

E.5 LS-EEP: Causal Attention Over Past States

LS-EEP replaces local temporal convolution with causal self-attention. Its purpose is to test whether prospective warning depends on relationships among different portions of the already observed state trajectory.

Each state in the causal window is projected to the common attention width:

$$\mathbf{u}_\tau = W_Z z_\tau + \mathbf{b}_Z, \qquad \tau \leq t. \tag{160}$$

The canonical implementation uses a two-layer Transformer encoder with four attention heads, model width 128, and feed-forward width 256. A causal attention mask enforces

$$A_{\tau,\tau'} = -\infty \qquad \text{for } \tau' > \tau. \tag{161}$$

This mask prevents any sequence position from attending to a future state.

Writing CTr for this causal Transformer,

$$\mathbf{Q}_t^{\mathrm{LS}} = \mathrm{CTr}\left(\mathbf{u}_{\tau_t:t}\right), \tag{162}$$

and the final valid causal representation is fused with the remaining history-side evidence through Π_{LS} to obtain $\mathbf{q}_t^{\mathrm{LS}} \in \mathbb{R}^{128}$.

The prediction head is direct:

$$\boldsymbol{\ell}_t^{H,\mathrm{LS}} = R_{\mathrm{direct}}\left(\mathbf{q}_t^{\mathrm{LS}}\right). \tag{163}$$

LS-EEP therefore tests whether non-local relationships among previously realized preference states carry warning that need not be captured by local convolution alone. LS-EEP is a project-defined causal-attention sentry. The Transformer architecture supplies its backbone rather than defining a separate preference model.

E.6 TCSA: Recency-Weighted Causal Attention

TCSA shares the causal-attention family with LS-EEP but explicitly introduces recency before attention. For a state at lag $k \geq 0$, its projected representation is attenuated as

$$\widetilde{\mathbf{u}}_{t-k} = \mathbf{u}_{t-k}\exp(-\lambda k), \qquad \lambda \geq 0, \tag{164}$$

where the decay rate is learned.

The recency-weighted sequence is passed through the same form of causal self-attention:

$$\mathbf{Q}_t^{\mathrm{TCSA}} = \mathrm{CTr}\left(\widetilde{\mathbf{u}}_{\tau_t:t}\right), \tag{165}$$

and the resulting causal representation is fused with the corresponding history-side evidence to give $\mathbf{q}_t^{\mathrm{TCSA}}$.

Its direct warning head is

$$\ell_t^{H,\mathrm{TCSA}} = R_{\mathrm{direct}}\left(\mathbf{q}_t^{\mathrm{TCSA}}\right). \tag{166}$$

The learned decay changes only how already-observed evidence is weighted. It does not introduce an additional source of information. TCSA therefore asks whether the temporal state-history signal is better expressed when recent parts of the causal trajectory are allowed to contribute more strongly than older parts.

E.7 TCSR: Recurrent Evidence With Discrete Hazard Risk

TCSR changes both the temporal backbone and the semantics of the warning head. The causal state trajectory is summarized recurrently:

$$\mathbf{q}_t^{\mathrm{GRU}} = \mathrm{GRU}\left(\Pi_{\mathrm{state}}(\mathcal{Z}_t^{(W)})\right). \tag{167}$$

This representation is fused with the projected current history-side evidence and mapped to the common sentry representation $\mathbf{q}_t^{\mathrm{TCSR}} \in \mathbb{R}^{128}$.

Unlike a direct multi-horizon head, TCSR predicts discrete conditional hazards. Let $\lambda_{t,k} \in (0,1)$ be the predicted degradation hazard in future bin k. The cumulative degradation probability through horizon h is

$$F_t^{\mathrm{TCSR}}(h) = 1 - \prod_{k=1}^{h} (1 - \lambda_{t,k}). \tag{168}$$

The common CoTRAC warning logit is then

$$\ell_{t,h}^{H,\mathrm{TCSR}} = \mathrm{logit}_\varepsilon\left(F_t^{\mathrm{TCSR}}(h)\right), \tag{169}$$

where $\mathrm{logit}_\varepsilon$ denotes the numerically clipped logit transformation.

TCSR therefore instantiates the same causal History-State evidence through a recurrent temporal representation while imposing cumulative event-risk structure on the output. The architectural lineage is related to temporally-consistent survival modeling (Maystre & Russo, 2022). The present sentry is adapted to the common CoTRAC degradation horizons.

E.8 DRSA-Style: Recurrent Future-Time Hazard Evolution

The DRSA-style sentry also uses a recurrent representation of causal state-history evidence:

$$\mathbf{q}_t^{\mathrm{DRSA}} = B_{\mathrm{GRU}}\left(\Pi_{\mathrm{DRSA}}(\mathcal{H}_{\le t}, \mathcal{Z}_t^{(W)})\right). \tag{170}$$

Its distinctive component is the risk head. Starting from the sentry representation, a second recurrent process evolves over prospective event-time bins. Let $\mathbf{e}_k$ be the learned embedding of future bin k. Then

$$\mathbf{s}_{t,k} = \mathrm{GRUCell}\left(\mathbf{e}_k, \mathbf{s}_{t,k-1}\right), \tag{171}$$

with the initial future-risk state determined by $\mathbf{q}_t^{\mathrm{DRSA}}$. At each future step,

$$\lambda_{t,k} = \sigma\left(\mathbf{w}_\lambda^\top \mathbf{s}_{t,k} + b_\lambda\right) \tag{172}$$

is the conditional degradation hazard.

Cumulative risk is

$$F_t^{\mathrm{DRSA}}(h) = 1 - \prod_{k=1}^{h} (1 - \lambda_{t,k}), \tag{173}$$

and the corresponding common warning logit is

$$\ell_{t,h}^{H,\mathrm{DRSA}} = \mathrm{logit}_\varepsilon\left(F_t^{\mathrm{DRSA}}(h)\right). \tag{174}$$

The DRSA-style sentry adapts the recurrent discrete-survival structure of DRSA (Ren et al., 2019) to the state-history evidence and degradation targets defined by the CoTRAC evaluation protocol.

E.9 DeepHit-Style: Discrete Event-Time Distribution

DeepHit-style uses the same class of recurrent causal evidence representation,

$$\mathbf{q}_t^{\mathrm{DH}} = B_{\mathrm{GRU}}\left(\Pi_{\mathrm{DH}}(\mathcal{H}_{\le t}, \mathcal{Z}_t^{(W)})\right), \tag{175}$$

while changing how prospective event time is parameterized.

Its risk head predicts one normalized distribution over the future event bins and the no-event-within-window outcome:

$$[p_{t,1}, \ldots, p_{t,K}, p_{t,>K}] = \mathrm{softmax}\left(R_{\mathrm{DH}}(\mathbf{q}_t^{\mathrm{DH}})\right). \tag{176}$$

The degradation probability by horizon h is

$$F_t^{\mathrm{DH}}(h) = \sum_{k=1}^{h} p_{t,k}. \tag{177}$$

The common warning logit is therefore

$$\ell_{t,h}^{H,\mathrm{DH}} = \mathrm{logit}_\varepsilon\left(F_t^{\mathrm{DH}}(h)\right). \tag{178}$$

This sentry follows the discrete event-time modeling principle of DeepHit (Lee et al., 2018), adapted to a single future degradation event and the common CoTRAC horizon set. It differs from TCSR and DRSA-style in where temporal structure is imposed: TCSR and DRSA-style parameterize conditional hazards, whereas DeepHit-style parameterizes a normalized event-time distribution directly.

E.10 Temporal-Prior: Recurrent Evidence With Causal Progress Modulation

Temporal-Prior tests whether the same causal state-history evidence should be interpreted differently depending on where the current observation lies in the already observed trajectory.

The causal evidence is first summarized recurrently:

$$\mathbf{q}_t^{\mathrm{TP}} = B_{\mathrm{GRU}}\left(\Pi_{\mathrm{TP}}(\mathcal{H}_{\le t}, \mathcal{Z}_t^{(W)})\right). \tag{179}$$

Let $\mathbf{c}_t \in \mathbb{R}^6$ denote the canonical progress features derived deterministically from the causal prefix. A small network maps these features to a 128-dimensional modulation:

$$\mathbf{g}_t = G_{\mathrm{time}}(\mathbf{c}_t), \qquad G_{\mathrm{time}} : 6 \to 64 \to 128. \tag{180}$$

The sentry representation is modulated as

$$\widetilde{\mathbf{q}}_t^{\mathrm{TP}} = \mathbf{q}_t^{\mathrm{TP}} \odot (0.5 + \mathbf{g}_t). \tag{181}$$

The final warning head is direct:

$$\boldsymbol{\ell}_t^{H,\mathrm{TP}} = R_{\mathrm{direct}}\left(\widetilde{\mathbf{q}}_t^{\mathrm{TP}}\right). \tag{182}$$

The progress features are derived from the prefix already observed at time t. They therefore remain within the information set $\mathscr{F}_t$ and do not reveal future utility or future hidden states. Temporal-Prior differs from the other recurrent sentries by allowing the meaning of a given state-history configuration to depend explicitly on causal trajectory progress.

E.11 Three Families Of Risk Parameterization

The eight sentries differ not only in their temporal backbones but also in how prospective risk is represented. Their heads fall into three broad families.

Direct multi-horizon risk. Prefix-MLP, TCN-ELECTS-EW, LS-EEP, TCSA, and Temporal-Prior directly produce one logit for each evaluated horizon:

$$\boldsymbol{\ell}_t = W\mathbf{q}_t + \mathbf{b}. \tag{183}$$

Conditional-hazard risk. TCSR and DRSA-style produce future-step hazards and convert them to cumulative risk:

$$F_t(h) = 1 - \prod_{k=1}^{h} (1 - \lambda_{t,k}). \tag{184}$$

Discrete event-time risk. DeepHit-style predicts a normalized event-time mass and obtains cumulative risk through

$$F_t(h) = \sum_{k=1}^{h} p_{t,k}. \tag{185}$$

These differences are useful for the scientific test. A positive state-history result that persists across the panel cannot be attributed only to one assumption about whether future degradation should be represented as independent horizon logits, cumulative hazards, or a discrete event-time distribution.

E.12 Shared Prospective Supervision

Although their internal parameterizations differ, all eight History-State sentries solve the same prospective prediction problem. For the evaluated horizons $\mathcal{K} = \{1, 2, 3, 5\}$, each observation has the target vector

$$\mathbf{Y}_t = [Y_{t,1}, Y_{t,2}, Y_{t,3}, Y_{t,5}], \tag{186}$$

where $Y_{t,h} = 1$ indicates that the corresponding prospective degradation criterion is satisfied within horizon h.

The common History-State training objective is the prediction objective defined in Appendix B:

$$\mathcal{L}_H = \mathcal{L}_{\mathrm{BCE}} + 0.1\,\mathcal{L}_{\mathrm{rank}} + 0.025\,\mathcal{L}_{\mathrm{mono}}. \tag{187}$$

The first term fits valid future-degradation labels, the ranking term encourages higher risk for degrading than stable examples, and the monotonicity term discourages cumulative risk from decreasing as the prediction horizon expands.

Thus, architecture changes are not accompanied by changes in the scientific prediction target. The panel varies the *sentry inductive bias* while preserving the information source and prospective supervision.

E.13 What Is Shared And What Is Not

The eight sentries should not be interpreted as eight independent experimental replications. They are eight architecture-level realizations of the same monitoring problem. Their role is to test whether the paper's central finding depends on one particular way of reading state-history evidence.

Across the panel, the following quantities are shared:

$$\left\{\mathcal{H}_{\leq t}, \mathcal{Z}_t^{(W)}, Y_{t,h}, \mathcal{K}\right\}. \tag{188}$$

That is, they observe the same causal evidence, target the same future degradation outcomes, and are evaluated at the same horizons.

The following components vary:

$$\{\Pi_m, B_m, R_m, \Gamma_m\}. \tag{189}$$

The sentries therefore differ in how history-state evidence is represented, which temporal dependencies are favored, and how prospective event time is parameterized.

This distinction is central to the interpretation of the empirical results. Directional consistency across the sentry panel supports robustness to sentry inductive bias. It does not create eight statistically independent copies of the underlying host trajectory.

E.14 Relation To Temporal State-History Matching

The eight constructions can now be placed directly under the abstraction from Section 3. Temporal State-History Matching asks whether the ordered evolution of the host's hidden preference states, interpreted together with the causal interaction evidence that produced them, contains systematic prospective warning structure.

Every History-State sentry considered here has the form

$$\boxed{\left(\mathcal{H}_{\leq t}, \mathcal{Z}_t^{(W)}\right) \longrightarrow \mathcal{F}_m \longrightarrow \boldsymbol{\ell}_t^{H,m}} \tag{190}$$

with no future host state and no future utility on the left-hand side.

Prefix-MLP asks whether a nonlinear causal summary is sufficient. TCN-ELECTS-EW asks whether local temporal motifs are sufficient. LS-EEP asks whether causal long-range state relationships matter. TCSA additionally tests explicit recency weighting. TCSR asks whether cumulative hazard structure is useful. DRSA-style asks whether prospective hazard itself should evolve recurrently. DeepHit-style asks whether degradation time should be represented as a discrete event-time distribution. Temporal-Prior asks whether causal trajectory progress should modulate the meaning of the observed state-history evidence.

These are different computational hypotheses about the same scientific object. Consequently, persistence of prospective warning across this panel supports a stronger claim than success under one bespoke predictor: joint History-State evidence remains informative under multiple ways of organizing and reading causal preference-state evolution.

F Evaluation Protocol And Reporting

This appendix follows the order of the experiment: train the personalization model, extract its hidden-state trajectory, train `CoTRAC` on that fixed trajectory, and evaluate warnings on held-out future outcomes. It then defines the warning metrics and states how matched results are assembled for each research question. The personalization interface and the three monitoring resolutions are those of Section 3. No additional preference model is introduced here.

F.1 Training And Frozen-State Extraction

Task-trained preference models. For each dataset and preference model, we first obtain its task-trained checkpoint using the corresponding training partition. At an observed timestep t, the trained encoder and task head produce $z_t = \mathcal{M}_\theta(\mathcal{H}_{\leq t})$, $\hat{y}_{t+1} = \mathcal{G}_\phi(z_t, q_{t+1})$, and the subsequently observed task utility $u_{t+1} = \mathcal{U}(\hat{y}_{t+1}, y^\star_{t+1})$, exactly as in Section 3. MovieLens uses nDCG as the downstream utility, PENS uses MRR for news recommendation and ROUGE-L for headline generation, and Amazon uses MRR for product recommendation. The three open-weight LLMs (OpenLLMs) form a separate pretrained, frozen consumer-interface setting: their monitored state is read after the available user and task context has been processed and before the first autoregressive output token. A host checkpoint is fixed before any warning model is fitted, so learning a warning never updates $\mathcal{M}_\theta$ or $\mathcal{G}_\phi$.

Causal state and utility records. We process each permitted training or held-out trajectory in its observed order. After interaction b_τ has been absorbed, we retain the host state z_τ and pair it with that interaction. The warning available at t is built from $\mathcal{C}_t^{(W)} = \langle (b_\tau, z_\tau) \rangle_{\tau=t-W+1}^{t}$, with the window truncated at the start of a shorter history. Its utility-only counterpart receives $\mathbf{u}_t^{(m)} = \langle u_\tau \rangle_{\tau=t-m+1}^{t}$. Both records are constructed from the same frozen host and the same eligible warning times. Future utility supplies the supervised target $Y_{t,h} = \mathcal{D}_h(U_{>t})$ only after the observation window has ended. We construct training and held-out records within their respective partitions. No future held-out utility enters state extraction, warning optimization, or threshold selection.

Warning training and refinement. The broad History-State sentry H learns a horizon-wise degradation score from $\mathcal{C}_t^{(W)}$, while the utility control U learns from $\mathbf{u}_t^{(m)}$ under the same future target, warning horizon, and eligible training positions. For PENS, we compare four utility-warning architectures using MovieLens TEST results, select the strongest utility architecture, fix that choice before

PENS evaluation, and retrain it from scratch on PENS training data. No PENS TEST result is used to select the utility-control architecture. For the valid training pairs $\Omega = \{(t,h) : Y_{t,h} \text{ is defined}\}$, let $s_{t,h} = \sigma(\ell_{t,h})$. The common prospective-warning objective is

$$\begin{aligned} \mathcal{L}_{\mathrm{BCE}} &= -\frac{1}{|\Omega|} \sum_{(t,h)\in\Omega} \big[Y_{t,h} \log s_{t,h} + (1 - Y_{t,h}) \log(1 - s_{t,h})\big], \\ \mathcal{L}_{\mathrm{pred}} &= \mathcal{L}_{\mathrm{BCE}} + 0.1\mathcal{L}_{\mathrm{rank}} + 0.025\mathcal{L}_{\mathrm{mono}}. \end{aligned} \tag{191}$$

Here, $\mathcal{L}_{\mathrm{BCE}}$ fits the future-degradation target, $\mathcal{L}_{\mathrm{rank}}$ encourages higher warning scores for higher-risk observations, and $\mathcal{L}_{\mathrm{mono}}$ discourages cumulative predicted risk from decreasing as the horizon expands. After training, H is frozen and becomes the common warning reference for structured refinement.

Regularity and Transition are trained separately from training-side projected-state trajectories before structured refinement. The Regularity precursor receives four raw violation features for speed deviation, directional deviation, acceleration deviation, and discrete jerk, and learns training-quartile-derived satisfaction targets through $\mathcal{L}_{\mathrm{pre},R}$. The Transition precursor uses earlier projected displacements to predict the held-out newest displacement and its coordinate-wise uncertainty through $\mathcal{L}_{\mathrm{pre},T}$. Neither precursor uses $Y_{t,h}$, future utility, or any future-degradation label. Directional disagreement, magnitude disagreement, predictive uncertainty, and orthogonal residual are materialized Transition features rather than additional precursor-loss terms. Appendix B gives the exact precursor targets, objectives, dimensions, scaling, and validity rules. Once trained, both precursors are frozen.

Each of HR, HT, and HRT then learns a gated correction to the frozen H logit,

$$\ell_{t,h}^{H+A} = \ell_{t,h}^{H} + \alpha_t^A \Delta_{t,h}^A, \qquad A \in \{R, T, RT\}. \tag{192}$$

For $A \in \{R, T\}$, the already learned consistency score s_t^A supervises the stay-with-H side of the refinement gate through $\mathcal{L}_A = \mathrm{BCE}(s_t^A, \sigma(-g_t^A))$. Thus, $\mathcal{L}_R$ and $\mathcal{L}_T$ are refinement-gate losses and are distinct from the precursor-training losses $\mathcal{L}_{\mathrm{pre},R}$ and $\mathcal{L}_{\mathrm{pre},T}$. The structured objectives are

$$\begin{aligned} \mathcal{L}_{H+A} &= 0.7\mathcal{L}_{\mathrm{pred}} + 0.3\mathcal{L}_A + \lambda_D \mathcal{L}_D, \qquad A \in \{R, T\}, \\ \mathcal{L}_{H+RT} &= 0.4\mathcal{L}_{\mathrm{pred}} + 0.3\mathcal{L}_R + 0.3\mathcal{L}_T + \lambda_D \mathcal{L}_D. \end{aligned} \tag{193}$$

Here, $\mathcal{L}_D$ penalizes unsupported deterioration of the prospective objective relative to the frozen H reference. During structured-refinement training, H and the Regularity and Transition precursors remain fixed, and only the refinement parameters are optimized. The resulting HR, HT, and HRT sentries therefore learn whether and how structured state evidence should modify the established warning. They do not modify the frozen host state or its recommendation or generation output.

Splits and operating thresholds. The evaluated horizons are $\mathcal{K} = \{1, 2, 3, 5\}$. MovieLens uses the fixed ordered split of 35K warning-training and 15K held-out instances. PENS and Amazon use their corresponding disjoint warning-training and held-out pools (10K/2K). A trajectory may yield several eligible warning times, so these instance counts are not the numbers of positive labels or distinct degradation episodes. We choose an operating threshold for each warning method using designated pre-test information and freeze it before TEST evaluation. For conventional PENS models, calibration determines operating thresholds without filtering otherwise eligible results. The OpenLLM setting separately applies pre-test calibration-based run admission, as specified in Appendix G.

F.2 Evaluation Metrics And Reporting

Downstream task utility. The degradation labels are constructed from the original host's output-quality sequence, not from the warning score. For a ranking instance i with one relevant next item at rank r_i, its reciprocal-rank utility is $1/r_i$ and its nDCG utility at cutoff K is $\mathbf{1}\{r_i \le K\}/\log_2(r_i+1)$. The corresponding mean recommendation metrics over N_{task} instances are

$$\begin{aligned} \mathrm{MRR} &= \frac{1}{N_{\mathrm{task}}} \sum_i \frac{1}{r_i}, \\ \mathrm{nDCG}\,@K &= \frac{1}{N_{\mathrm{task}}} \sum_i \frac{\mathbf{1}\{r_i \le K\}}{\log_2(r_i + 1)}. \end{aligned} \tag{194}$$

For headline generation, ROUGE-L is the task utility. With L the longest common subsequence length between an output and its reference, $P_L = L/|\hat{y}|$ and $R_L = L/|y^\star|$, its F-score form is

$$\text{ROUGE-L} = \frac{(1+\beta^2)P_L R_L}{R_L + \beta^2 P_L},$$

using the same fixed β convention across all comparisons. These are the underlying output-utility measurements. AUPRC and the alarm metrics below instead evaluate predictions of their *future degradation*.

Evaluation population and prevalence. Fix one preference model, warning architecture, horizon h, and its set $\mathcal{V}_h$ of held-out times with valid future outcomes. Write $n = |\mathcal{V}_h|$, $y_i = Y_{t_i,h} \in \{0,1\}$, $s_i = s_{t_i,h} \in [0,1]$, and $n_+ = \sum_{i=1}^n y_i$. The event rate, or positive-label prevalence, is $\pi = n_+/n$. Ranking metrics are computed within this cell before averaging across models, architectures, or horizons. AUPRC and its normalization require $n_+ > 0$, and normalized AUPRC additionally requires $\pi < 1$. Cells without both label classes are not assigned a discrimination score.

Discrimination: AUPRC and nAUPRC. Sort warning scores from highest to lowest and collect tied scores at the same threshold. At each distinct threshold j, let TP_j and FP_j denote the cumulative numbers of positive and negative labels above that threshold. With precision $P_j = \text{TP}_j/(\text{TP}_j+\text{FP}_j)$, recall $R_j = \text{TP}_j/n_+$, and $R_0 = 0$, the stepwise precision-recall area is

$$\begin{aligned} \text{AUPRC} &= \sum_j (R_j - R_{j-1})P_j, \\ \text{nAUPRC} &= \frac{\text{AUPRC} - \pi}{1-\pi}. \end{aligned} \tag{195}$$

AUPRC measures whether warnings rank future-degrading observations above non-degrading ones. Its prevalence reference is π. nAUPRC is zero at that reference, one for perfect ranking, and can be negative. For methods evaluated on the same valid times, $\Delta\,\text{nAUPRC} = \Delta\,\text{AUPRC}/(1-\pi)$, so normalization preserves their AUPRC ordering. Differences between cells with different event rates are normalized within each cell before averaging. The paired comparison uses $\Delta_{H-U}\mathcal{M} = \mathcal{M}(H) - \mathcal{M}(U)$ for either discrimination metric $\mathcal{M}$.

Probability quality. Where probability-quality results are reported, the same held-out y_i and predicted degradation probabilities s_i give the Brier score and binary negative log-likelihood (NLL):

$$\begin{aligned} \text{Brier} &= \frac{1}{n}\sum_{i=1}^n (s_i - y_i)^2, \\ \text{NLL} &= -\frac{1}{n}\sum_{i=1}^n \big[y_i \log s_i + (1-y_i)\log(1-s_i)\big]. \end{aligned} \tag{196}$$

Predicted probabilities are clipped only for numerical stability when evaluating NLL. For expected calibration error with 15 probability bins $B_1, \ldots, B_{15}$, write $\text{conf}(B_j) = |B_j|^{-1}\sum_{i\in B_j} s_i$ and $\text{freq}(B_j) = |B_j|^{-1}\sum_{i\in B_j} y_i$. Then

$$\text{ECE}_{15} = \sum_{j:|B_j|>0} \frac{|B_j|}{n}\left|\text{freq}(B_j) - \text{conf}(B_j)\right|. \tag{197}$$

Smaller Brier, NLL, and ECE15 indicate better probability quality under their respective definitions. Amazon's reported reductions use the paired difference $\mathcal{M}(U) - \mathcal{M}(H)$ on the same model-sentry-horizon cell, so a positive reduction denotes lower error for H. AUPRC concerns score ordering, whereas these probability metrics also depend on the predicted probability values.

Event detection, warning lead, and false alarms. For a warning method M, let τ_M be its frozen pre-test operating threshold and $A_t^M = \mathbf{1}\{s_{t,h}^M \geq \tau_M\}$ its alarm at an eligible time. Let $\mathcal{E}$ be the set of eligible degradation episodes, κ_e the onset time of episode e, and $\mathcal{P}_e$ the admissible pre-onset warning times for that episode under its task's guard and horizon. An episode is detected when $D_e^M = \mathbf{1}\{\exists t \in \mathcal{P}_e : A_t^M = 1\} = 1$. Its first valid alarm time is $t_e^M = \min\{t \in \mathcal{P}_e : A_t^M = 1\}$. The

operational metrics are

$$\begin{aligned}\text{EventRecall}^M &= \frac{\sum_{e\in\mathcal{E}} D_e^M}{|\mathcal{E}|},\\ \text{MeanLead}^M &= \frac{\sum_{e\in\mathcal{E}:D_e^M=1}(\kappa_e - t_e^M)}{\sum_{e\in\mathcal{E}} D_e^M},\\ \text{FAR}^M_{\text{TEST}} &= \frac{\sum_{t\in\mathcal{V}_-} A_t^M}{|\mathcal{V}_-|},\end{aligned} \tag{198}$$

where $\mathcal{V}_-$ is the eligible non-event exposure set. EventRecall counts degradation episodes receiving a valid pre-onset alarm, MeanLead counts timesteps before onset among detected episodes, and held-out FAR counts alarms per eligible non-event opportunity. MeanLead is undefined when no episode is detected, and an operational rate is undefined when its denominator is zero. The suffix *@FAR10* identifies the nominal 0.10 pre-test false-alarm operating point. TEST FAR is measured on held-out observations and need not equal 0.10. Detection and lead are evaluated after threshold selection and cannot be reconstructed from AUPRC alone.

Paired uncertainty and display scales. For user-clustered uncertainty, each bootstrap replicate samples held-out users with replacement, retains all eligible observations belonging to each sampled user, and applies the *same sampled users* to both members of a matched comparison. After $B = 10{,}000$ replicates, the reported bootstrap standard deviation of a paired difference $\Delta^{(b)}$ is

$$\sqrt{\frac{\sum_{b=1}^{B}(\Delta^{(b)} - \bar{\Delta})^2}{B-1}}.$$

Percentile bounds, where reported, use the corresponding bootstrap quantiles. These intervals quantify sampling variation across held-out users rather than variability from independently retraining the preference model or warning system.

Deployment reporting. For each measured preference-model–monitor configuration c, the incremental latency is $\delta_c^{\text{lat}} = L_c^{\text{model+CoTRAC}} - L_c^{\text{model}}$, with relative overhead $100\delta_c^{\text{lat}}/L_c^{\text{model}}$ percent. Added parameters and relative parameter overhead are defined analogously from matched parameter counts. Absolute costs and configuration-wise percentage overheads are summarized separately, while throughput is reported from measured requests-per-second values. Deployment measurements are reported separately from warning-discrimination cells.

F.3 Results Protocol

Prospective targets and admissible positions. Every warning at t is evaluated using the future utility of the *same frozen preference model*. For the MovieLens guarded target, an onset at valid step k is $e_k = \mathbf{1}\{u_k \le u_{k-1} - 0.1\}$ and

$$Y_{t,h} = \mathbf{1}\{\exists k \in \{t+2, \dots, t+h+1\} : e_k = 1\}. \tag{199}$$

Thus the first admissible onset is one guarded step beyond $t+1$. Amazon likewise uses a guarded MRR drop of at least 0.1 with guard length one. PENS uses its task-specific prospective degradation construction. The PENS task-change analysis evaluates each task's low-utility outcome with its own event rate and valid positions. A warning time contributes to a horizon only when the required future utility is available, so valid position sets and episode counts may differ across horizons and evaluation populations. Within any reported paired cell, both methods use the same validity mask, future labels, and output-utility trajectory. Episode matching, eligible pre-onset intervals, and non-event exposure follow the corresponding task-specific evaluation protocol. A positive horizon label and a distinct degradation episode are different counting units.

RQ1: matched utility and History-State warning. On MovieLens and Amazon, we fit U and H for the corresponding sentry architectures. On PENS, the utility-warning architecture selected from the MovieLens TEST results serves as a common U control for the four H sentries. Each comparison uses the same future degradation target, held-out positions, and horizon. We subtract discrimination metrics within each comparison before averaging and count $H > U$ only when H achieves strictly higher AUPRC. MovieLens primary analysis covers five hosts, eight sentry architectures, and four horizons (160 matched cells). PENS covers five hosts, four sentries, and four horizons (80 cells).

Amazon covers three hosts, two sentries, and four horizons (24 cells). Model-level means average their available sentry-horizon cells, and domain means weight the displayed preference models equally. Horizon profiles instead average the paired differences at fixed h. Amazon's Brier, NLL, and ECE15 comparisons likewise pair U and H by model, sentry, and horizon.

RQ2: refinements of the same frozen parent. The comparison is H versus HR, HT, or HRT, with each structured score matched to its corresponding frozen H parent at the same model, sentry, horizon, and held-out positions. For any metric $\mathcal{M}$, set $\Delta_A\mathcal{M} = \mathcal{M}(H + A) - \mathcal{M}(H)$, where $A \in \{R, T, RT\}$. The main structured table averages the cell-level nAUPRC differences within each model. Its joint-gain column is $\Delta_{RT} - \max(\Delta_R, \Delta_T)$, and the unchanged parent supplies the zero-change reference when comparing all four alternatives. EventRecall, MeanLead, realized FAR, and probability-quality changes are each calculated from their corresponding matched observations before aggregation. Every structured comparison uses the H parent from the same RQ2 evaluation population rather than substituting an H score from RQ1.

RQ3: reuse across encoders, tasks, and consumers. In source-only leave-one-model-out (LOMO) reuse, exclude one preference model, rank-normalize the remaining source models' warning scores on common evaluation indices, and average those source rankings. Rank normalization uses only source warning scores available before the corresponding future degradation outcomes are revealed. The resulting consensus is evaluated against the held-out model's future degradation outcomes. The target model's warning scores are used only to form the matched self-warning reference. No target warning weights are trained or transferred, and the hidden-state coordinates of different models are not aligned. For cells with a positive self-warning nAUPRC, retention is $100\,\mathrm{nAUPRC}_{\text{source only}} \,/\, \mathrm{nAUPRC}_{\text{self}}$. Reported retention is the mean of these cell-level ratios, not a ratio of pooled means. The unadapted MovieLens experiment contains 160 cells and the PENS experiment contains 80.

PENS task change uses the same five frozen preference backbones separately for recommendation and headline generation, evaluates each task against its own low-utility labels, and matches H, HR, HT, and HRT within that task. The three OpenLLMs form a separate pre-generation cohort. Their H nAUPRC is computed for each model-task-sentry-horizon cell using its corresponding prevalence before averaging the 32 cells for each model and task.

G Additional Experimental Results

This appendix gives the complete evaluation behind the three research questions. We begin with the evaluated models, warning architectures, degradation targets, and operating points. We then report the full utility-versus-History-State comparisons, the changes produced by Regularity and Transition, and warning reuse across encoders, tasks, and pre-generation language-model states. Table 4 in Appendix E specifies the sentry architectures, while Table 6 in Appendix G lists the preference models and evaluated sentries.

G.1 Evaluation Settings And Result Coverage

What Each Warning Metric Measures. The criteria used throughout the paper are collected in Table 7. AUPRC measures how well warning scores rank future-degrading positions above stable positions, and nAUPRC places that ranking relative to the degradation prevalence in the matched evaluation population. EventRecall@FAR10 counts degradation episodes with a valid pre-onset alarm, while MeanLead@FAR10 measures the average advance notice among detected episodes. The operating threshold is selected before held-out evaluation. TEST FAR records the false-alarm rate actually realized on that population. Encoder-reuse retention compares source-only and self-warning nAUPRC within the same target cell, whereas the task-change tables report trajectory-level AUPRC under each task's low-utility outcome.

Degradation Targets And Operating Points. The task-specific outcome and threshold source for each domain are listed in Table 8. MovieLens uses a guarded nDCG drop of at least 0.1 with a one-step guard. Amazon uses a guarded MRR drop of at least 0.1 and fixes its operating threshold at the 90th percentile of training-negative scores from the final checkpoint. PENS uses task-specific degradation targets and a separate pre-test calibration pool. These definitions fix what constitutes a future event and how a score becomes an alarm before any held-out comparison is made.

G.2 RQ1: Future-Degradation Warning Beyond Recent Utility

Event Support Across Preference Models. The positive-label counts and degradation prevalence for all seven audited MovieLens models appear in Table 10. The five supported models contribute between $4{,}484$ and $36{,}171$ positive positions across the four horizons. Each yields $32/32$ positive matched $H - U$ comparisons. DiffRec and DreamRec additionally show how the same evaluation behaves with only 60 to 267 and 2 to 10 positive positions, respectively. Their complete sensitivity results remain visible alongside the supported-five analysis, allowing outcome support to be read together with warning score.

PENS Warning Populations. The five PENS backbones across horizons $h \in \{1, 2, 3, 5\}$ are summarized in Table 11. The RQ1 held-out population decreases from $81{,}578$ eligible positions at $h = 1$ to $71{,}573$ at $h = 5$, while future-degradation prevalence rises as the prediction window expands. The structured RQ2 evaluation separately reports its positive positions and degradation episodes. Keeping those counts attached to their respective experiments makes each subsequent comparison reproducible at its actual unit of evaluation.

Results Across Domains And Experimental Stages. The RQ1 U/H comparisons and RQ2 structured changes share one model roster in Table 12. The MovieLens RQ1 scores come from its guarded degradation experiment. MovieLens RQ2 changes are measured relative to the H parent of its matched structured-refinement experiment. PENS RQ1 and RQ2 likewise retain their own H parents. Each reported R, T, or RT change is therefore paired with the H score trained and evaluated in that same experiment. Across the five supported MovieLens hosts, $H - U$ ranges from $+0.055$ to $+0.135$ nAUPRC. PENS and Amazon retain their corresponding within-domain differences.

MovieLens: All Hosts And Horizons. Architecture-matched utility and History-State AUPRC, nAUPRC, and win counts for all seven MovieLens hosts are reported in Table 13. The supported five account for the $160/160$ positive matched comparisons emphasized in the main paper, while the two low-support hosts retain their measured outcomes in the same record. The horizon-specific comparison appears in Table 14. Across the supported five, mean $H-U$ nAUPRC is $+0.048$, $+0.091$, $+0.124$, and $+0.163$ at horizons 1, 2, 3, and 5, with $40/40$ positive architecture-host comparisons at each horizon. The underlying counts and paired means for every supported host and horizon are listed in Table 15. For example, GRU4Rec increases from $+0.0541$ at $h = 1$ to $+0.1975$ at $h = 5$, while all five hosts maintain a positive $H - U$ difference throughout the evaluated horizon set.

PENS Recommendation: Model And Warning-Architecture Coverage. The four-horizon comparison for GTP, IMPerSumm, PENS-NAML-T1, PerDucer, and Walk2Pers is reported in Table 16. Their mean $H - U$ nAUPRC advantages across horizons are $+0.005$, $+0.027$, $+0.014$, $+0.004$, and $+0.028$, respectively. The corresponding sentry-level values appear in Tables 19 through 23. These five tables contain all 80 matched sentry-horizon cells and yield $H > U$ in $11/16$, $16/16$, $14/16$, $10/16$, and $16/16$ cells by backbone. The per-cell results locate the mean advantage across warning architectures and future horizons, including the small negative differences in individual GTP and PerDucer cells.

Amazon Recommendation: Horizon And Architecture Dependence. The matched U/H comparison for Caser, GRU4Rec, and SR-GNN at all four horizons is reported in Table 17. The largest mean differences occur at $h = 3$ for all three hosts: $+0.0229$, $+0.0273$, and $+0.0351$ nAUPRC, respectively. All 24 sentry-horizon cells appear in Table 24. The $21/24$ positive comparisons include both Prefix-MLP and TCN-ELECTS-EW. The three negative cells occur for GRU4Rec with TCN-ELECTS-EW at $h = 1$ and $h = 5$, and SR-GNN with TCN-ELECTS-EW at $h = 5$. Absolute H nAUPRC is retained beside each difference so the strength of the resulting warning remains visible at every operating horizon.

Quality Of Predicted Degradation Probabilities. Matched Amazon Brier score, negative log-likelihood, and ECE15 for U and H are listed in Table 18. The mean reductions summarized in the main paper are 0.0108 in Brier score, 0.0454 in negative log-likelihood, and 0.0221 in ECE15. At $h = 2$, for example, GRU4Rec reduces Brier score from 0.3245 to 0.2816, NLL from 0.9099 to 0.7647, and ECE15 from 0.2359 to 0.1519. Rows with increased errors are also retained, showing where improved overall warning discrimination coincides with weaker probability quality.

G.3 RQ2: Regularity And Transition As Refinements Of An Established Warning

MovieLens Structured Effects Across Horizons. In Table 25, $H + R$, $H + T$, and $H + RT$ are compared against the same frozen H parent at every horizon. Regularity produces positive mean nAUPRC changes at all four horizons, from $+1.277$ to $+1.403$ in units of 10^{-3}. The joint refinement reaches $+2.039 \times 10^{-3}$ at $h = 5$, while Transition has small negative mean changes at $h = 3$ and $h = 5$. At the preference-model level (Table 26), FEARec and SASRec obtain their largest mean structured nAUPRC changes with Regularity, while GRU4Rec, LRURec, and Mamba4Rec obtain theirs with RT. The matched RT-versus-individual comparison in Table 27 contains all 160 host-sentry-horizon cells. RT exceeds both individual refinements in $57/160$ cells, and its mean difference from the stronger individual refinement is -1.139×10^{-3} nAUPRC. These comparisons identify the models and horizons in which the two views work well together.

PENS Structured Effects And Host Variation. Across the 80 matched PENS cells in Table 28, Regularity, Transition, and RT change AUPRC by $+1.639$, $+1.686$, and $+1.455$ in units of 10^{-3}, while their EventRecall changes are $+0.00509$, $+0.00596$, and $+0.00290$. Their mean lead changes are -0.00450, -0.00299, and -0.00509. The host-level pattern in Table 29 explains these averages. All three refinements improve GTP recall and lead. PENS-NAML-T1 gains discrimination alongside small recall or lead reductions, while PerDucer shows positive discrimination changes alongside lower event recall and lead. These measurements connect the structured evidence to its effect on the established warning for each preference representation.

Absolute Warning Behavior And Evaluated Exposure. Absolute AUPRC, nAUPRC, EventRecall, MeanLead, TEST FAR, Brier score, NLL, and ECE15 for PENS and Amazon appear in Table 30. On PENS, H detects 0.1959 of degradation episodes and HT detects 0.2019, with mean leads of 3.2294 and 3.2264. On Amazon, H has mean lead 3.0462 and TEST FAR 0.3231, while HT has mean lead 3.0657 and TEST FAR 0.3427. The associated evaluation exposure is listed in Table 31. Each PENS realization spans 80 cells, five hosts, and four sentries, while each Amazon realization spans 24 cells, three hosts, and two sentries. These absolute metrics and support counts give the operational scale of the paired changes.

Detection, lead, and alarm burden by horizon. Table 32 traces the four warning realizations over each future horizon. On PENS at $h = 3$, EventRecall increases from 0.2047 for H to 0.2161 for HT, with leads of 3.4307 and 3.4281 and TEST FAR of 0.1056 and 0.1096. On Amazon at $h = 5$, HT increases recall from 0.7064 to 0.7145 and lead from 4.4955 to 4.5607, together with TEST FAR increasing from 0.6931 to 0.7433. Table 33 quantifies how frequently the PENS metrics improve in the same matched cell: HT improves AUPRC in 95.0% of cells and event recall in 67.5%, while a strictly positive improvement in AUPRC, recall, and lead occurs in 16.2%. The measured changes give separate, concrete descriptions of how structured evidence changes each operating criterion.

Amazon Host-Specific Refinements. Absolute discrimination, recall, lead, false-alarm rate, and probability-quality scores for each Amazon host are reported in Table 34. Caser has H nAUPRC -0.0112 and reaches recall 0.3556 with HR versus 0.3455 with H. GRU4Rec increases nAUPRC from 0.0054 to 0.0067 under HR, while SR-GNN increases it from -0.0200 to -0.0171. The horizon-specific changes in Table 35 locate the largest Amazon lead increases at $h = 5$, alongside higher TEST FAR and Brier errors in the matched cells. The matched PENS horizon analysis in Table 36 has positive mean nAUPRC change for every refinement at every horizon, with the largest EventRecall gains at $h = 3$ for HR and HT. Together, these host- and horizon-level measurements explain the distinct discrimination and operational patterns summarized in RQ2.

G.4 RQ3: Reuse Across Encoders, Tasks, and Consumers

Source-Only Reuse Across MovieLens Encoders. For each held-out encoder, source-only warning is constructed by ranking the same evaluation indices separately under the remaining source encoders and averaging their normalized ranks. The combined ranking is paired with the held-out encoder's own H warning in Table 37. All 160 source-only H cells have positive nAUPRC, and mean matched-cell retention is 86.45%. Retention varies from 72.13% for SASRec to 100.93% for FEARec. For all eight warning architectures (Table 38), source-only warning is positive in all 20 cells, with mean retention ranging from 80.75% for TCSA to 90.22% for DRSA-style. Across the four future horizons

(Table 39), retention ranges from 85.15% at $h = 1$ to 88.55% at $h = 5$. These measurements locate the reuse effect across held-out hosts, warning architectures, and prediction horizons.

PENS Encoder Reuse And Structured Source Warnings. Across the five held-out PENS encoders in Table 40, source-only H nAUPRC is positive in all 80 matched cells. Mean retention is 83.58%, ranging from 27.69% for Walk2Pers to 136.90% for GTP. Across horizons (Table 41), mean retention is 89.21% at $h = 1$ and 76.42% at $h = 5$. The transferred structured variants are compared with the same source-only H consensus in Table 42. On MovieLens, the mean R, T, and RT changes are -1.737, -2.005, and -1.483 in units of 10^{-3} nAUPRC. On PENS, they are $+1.473$, $+1.400$, and $+1.649$. The broad H warning is positively discriminative across all tested source-only cells, while the structured corrections follow different patterns across the two domains.

Supplementary Source-Only Adaptation. Tables 43 through 45 examine a supplementary MovieLens adaptation built from the source warning variants. At every horizon (Table 43), mean adapted-minus-source H nAUPRC is positive and reaches $+1.808 \times 10^{-4}$ at $h = 5$. Saved bootstrap intervals are reported beside each change. Across the five held-out models (Table 44), mean changes range from $+0.488 \times 10^{-4}$ for LRURec to $+1.533 \times 10^{-4}$ for Mamba4Rec. The fitted convex combination in Table 45 assigns an average 95.58% weight to the broad H warning, with the remaining weight distributed across R, T, and RT. The adaptation refines the source-only consensus using source-side warning information. The unadapted source-only results above provide its reference.

The Same Preference Backbone Under Two Tasks. The same five PENS preference backbones are compared under news recommendation and personalized headline generation in Table 46. Each task uses its own low-utility outcome. For GTP, H AUPRC is 0.361 in recommendation and 0.333 in generation. For PENS-NAML-T1, it is 0.341 and 0.368. Both tasks' low-utility rates are listed alongside their warning scores, allowing the result to be read relative to event frequency as well as across models. PerDucer generation, for example, obtains H AUPRC 0.359 at a low-utility rate of 0.353. The other nine backbone-task cells have H AUPRC below the corresponding low-utility rate. Across horizons (Table 47), mean H AUPRC changes from 0.340 to 0.334 for recommendation and from 0.345 to 0.337 for generation. Across the four warning architectures (Table 48), broad H scores range from 0.333 to 0.341 in recommendation and 0.334 to 0.345 in generation. These measurements show how the same backbone yields different prospective warning scores under the two downstream utility targets.

Task-Specific Structured Behavior. Recommendation and generation are paired within each of the five backbones in Table 49. Generation H AUPRC is higher for IMPerSumm, PENS-NAML-T1, and PerDucer and lower for GTP and Walk2Pers. The accompanying low-utility rates show how the task's event frequency changes at the same time. Within each task, HR, HT, and HRT are paired with that task's H parent in Table 50. In recommendation, HRT changes trajectory AUPRC by $+0.000743$, low-utility recall by $+0.02406$, and lead by $+0.09500$ steps. In generation, HRT changes the corresponding quantities by $+0.001582$, $+0.02585$, and $+0.13041$. Changes in flag rate are reported alongside these quantities, making the increased detection and warning activity visible together.

Pre-Generation Open-Weight LLM States And Consumer Families. The main paper measures H warning from the final transformer-layer state after the causal user and task context has been processed and before the first generated token. Structured changes for recommendation and headline generation among runs admitted by the pre-test calibration procedure are reported in Table 51. In headline generation, RT changes AUPRC by -3.160×10^{-3} while increasing recall by 5.294×10^{-3} and mean lead by 0.121. In news recommendation, its six admitted runs have a $+0.388 \times 10^{-3}$ AUPRC change. The calibrated runs are separated by consumer family in Table 52. Among the $18/32$ admitted conventional-generation-decoder RT runs, mean AUPRC and recall changes are $+2.712 \times 10^{-3}$ and $+7.294 \times 10^{-3}$. Among the $26/48$ admitted open-weight LLM RT runs, the corresponding changes are -2.341×10^{-3} and $+4.073 \times 10^{-3}$, with mean lead change $+0.093$. These measurements characterize how the two structured views affect a warning that is available before autoregressive decoding across task and consumer interfaces.

Table 5: Evaluation coverage across datasets, tasks, frozen hosts, and warning questions.

Dataset	Task / Interface	Frozen Hosts	Evaluation Role	Utility for Warning Target
MovieLens-1M	Movie recommendation	FEARec, GRU4Rec, LRURec, Mamba4Rec, SASRec	RQ1: matched U vs. H RQ2: $R/T/RT$ specialization RQ3: source-only encoder reuse	nDCG ranking utility guarded future degradation
PENS	News recommendation	GTP, IMPerSumm, PENS-NAML-T1, PerDucer, Walk2Pers	RQ1: matched U vs. H RQ2: structured warning RQ3: source-only encoder reuse	MRR recommendation utility task-specific future degradation
PENS	Headline generation	GTP, IMPerSumm, PENS-NAML-T1, PerDucer, Walk2Pers	RQ3: same-backbone task change and structured task profile	ROUGE-L generation utility task-specific future degradation
PENS	Pre-generation open-weight LLM interface	Mistral-7B, Phi-4-mini, Qwen3-8B	RQ3: pre-generation warning and consumer-family boundary	Task-specific recommendation / generation degradation targets
Amazon Product Reviews	Product recommendation	Caser, GRU4Rec, SR-GNN	RQ1: matched U vs. H RQ2: structured warning and specialization	MRR ranking utility guarded future degradation

Table 6: Preference-model and sentry roster.

Domain	Preference model	Task	Sentries	Supported RQs
MovieLens	FEARec	Movie rec.	8	RQ1/RQ2/RQ3
MovieLens	GRU4Rec	Movie rec.	8	RQ1/RQ2/RQ3
MovieLens	LRURec	Movie rec.	8	RQ1/RQ2/RQ3
MovieLens	Mamba4Rec	Movie rec.	8	RQ1/RQ2/RQ3
MovieLens	SASRec	Movie rec.	8	RQ1/RQ2/RQ3
PENS	GTP	News rec. + headline gen.	4	RQ1/RQ2/RQ3
PENS	IMPerSumm	News rec. + headline gen.	4	RQ1/RQ2/RQ3
PENS	PENS-NAML-T1	News rec. + headline gen.	4	RQ1/RQ2/RQ3
PENS	PerDucer	News rec. + headline gen.	4	RQ1/RQ2/RQ3
PENS	Walk2Pers	News rec. + headline gen.	4	RQ1/RQ2/RQ3
Amazon	Caser	Product rec.	2	RQ1/RQ2
Amazon	GRU4Rec	Product rec.	2	RQ1/RQ2
Amazon	SR-GNN	Product rec.	2	RQ1/RQ2

Table 7: Warning evaluation criteria. FAR10 names the pre-test calibrated operating-point target, not a guarantee of 10% held-out FAR. Amazon calibrates from training negatives. PENS and MovieLens use their corresponding pre-test protocols.

Criterion	Metric	Question	Direction
Discrimination	AUPRC / nAUPRC	Order future degradation relative to stable timestamps	Higher
Event detection	EventRecall@FAR10	Valid pre-onset alarms among degradation episodes	Higher
Warning lead	MeanLead@FAR10	Advance notice among detected episodes	Higher
Alarm burden	Realized TEST FAR	False alarms on held-out non-event exposure	Lower
Encoder reuse	Mean cell-wise retention	Source-only versus self H normalized AUPRC	Higher
Task change	Trajectory AUPRC	Low-utility outcome under each PENS task	Higher

Table 8: Degradation targets and operating-point provenance across datasets. Amazon uses the 90th percentile of training-negative scores from the final checkpoint with no separate calibration split or admission filter. Realized TEST FAR is reported in the operational tables.

Domain	Degradation target	Guard	Threshold source	Evaluation cohort
MovieLens	Guarded onset, $\Delta u \geq 0.1$ drop	Guard $g = 1$	Pre-test threshold protocol	Five supported hosts; eight sentries
PENS	Task-specific future degradation	Task-dependent	Separate pre-test calibration	Five recommendation backbones; four sentries
Amazon	MRR drop ≥ 0.1	Guard $g = 1$	training-negative 90th percentile	Three hosts; two sentries

Table 9: CoTRAC hyperparameters and neural dimensions. Settings are shared across preference models unless otherwise specified; dimensions indicate input-to-output widths.

Component	Hyperparameter / layer	Setting / dimension
Shared monitoring and training		
History-State monitoring	State-window length W	8
Prospective warning	Horizons $\mathcal{K}$	$\{1, 2, 3, 5\}$
History-State sentry	Input projection	$256 \to 128$
History-State sentry	Representation width d_H	128
History-State sentry	Warning output	$128 \to 4$
Sentry optimization	Learning rate / batch size	10^{-3} / 256
Sentry optimization	Maximum epochs / configured patience	10 / 4
Reproducibility	Random seed	42
History-State sentry architectures		
Prefix-MLP	Feed-forward dimensions	$128 \to 256 \to 128$
TCN-ELECTS-EW	Channels / kernel / dilations	128 / 3 / $\{1, 2\}$
LS-EEP, TCSA	Transformer layers / heads	2 / 4
LS-EEP, TCSA	Model width / FFN width	128 / 256
TCSR, DRSA-style	History GRU input / hidden width	128 / 128
DeepHit-style	History GRU input / hidden width	128 / 128
DRSA-style	Future-risk GRUCell width	128
Temporal-Prior	Progress-feature projection	$6 \to 64 \to 128$
Regularity and Transition precursor mechanisms		
Regularity (R)	Raw-violation encoder	$4 \to 64 \to 32$
Regularity (R)	Satisfaction head	$32 \to 4$
Transition (T)	Causal GRU input / hidden width	128 / 64
Transition (T)	Learned representation projection	$64 \to 32$
Transition (T)	Mean / log-variance heads	$64 \to 128$ each
Structured warning refinement		
Regularity (R)	Evidence projection ϕ_R	$37 \to 32$
Transition (T)	Evidence projection ϕ_T	$38 \to 32$
HR, HT	Adapter input $[r^H; f^A]$	$128 + 32 = 160$
HRT	Adapter input $[r^H; f^{RT}]$	$128 + 4(32) = 256$
HR, HT	Gate / residual networks	$160 \to 64 \to \{1, 4\}$
HRT	Gate / residual networks	$256 \to 64 \to \{1, 4\}$
Training objectives and weights		
History-State H	BCE / ranking / monotonicity weights	1.0 / 0.1 / 0.025
Regularity precursor	Objective $\mathcal{L}_{\mathrm{pre},R}$	BCE to training-quartile satisfaction targets
Transition precursor	Objective $\mathcal{L}_{\mathrm{pre},T}$	Heteroscedastic Gaussian loss
HR, HT refinement	Prediction / gate-supervision weights	0.7 / 0.3
HRT refinement	Prediction / R-gate / T-gate weights	0.4 / 0.3 / 0.3
Structured refinement	Anchoring weight λ_D	1.0

Table 10: MovieLens support and prevalence ranges span the four guarded horizons. DiffRec and DreamRec are shown as low-support sensitivity cases and are excluded from the five-host headline.

Host	Support	TEST positives	Prevalence	U nAUPRC	H nAUPRC	ΔnAUPRC	H>U
DiffRec	Low support	60–267	0.0005–0.0026	-0.000007	-0.000008	-0.000001	16/32
DreamRec	Degenerate	2–10	0.0000–0.0001	0.000	0.000062	+0.000062	24/32
FEARec	Supported	4,484–18,146	0.0387–0.1747	0.016	0.071	+0.055	32/32
GRU4Rec	Supported	10,132–36,171	0.0874–0.3482	0.107	0.236	+0.129	32/32
LRURec	Supported	7,984–29,856	0.0689–0.2874	0.062	0.181	+0.119	32/32
Mamba4Rec	Supported	8,587–31,513	0.0741–0.3033	0.081	0.216	+0.135	32/32
SASRec	Supported	4,854–18,789	0.0419–0.1808	0.047	0.141	+0.095	32/32

Table 11: PENS recommendation support for matched RQ1 and structured RQ2.

Backbone	h	RQ1 TEST N	RQ1 prev.	RQ2 positives	RQ2 events	Sentries
GTP	1	81,578	0.1313	10,715	10,715	4
GTP	2	79,022	0.2563	20,250	10,715	4
GTP	3	76,501	0.3635	27,810	10,715	4
GTP	5	71,573	0.5319	38,073	10,868	4
IMPerSumm	1	81,578	0.1253	10,218	10,218	4
IMPerSumm	2	79,022	0.2442	19,301	10,218	4
IMPerSumm	3	76,501	0.3460	26,472	10,218	4
IMPerSumm	5	71,573	0.5076	36,330	10,362	4
PENS-NAML-T1	1	81,578	0.1455	11,872	11,872	4
PENS-NAML-T1	2	79,022	0.2826	22,331	11,872	4
PENS-NAML-T1	3	76,501	0.3977	30,428	11,872	4
PENS-NAML-T1	5	71,573	0.5732	41,026	12,043	4
PerDucer	1	81,578	0.1300	10,609	10,609	4
PerDucer	2	79,022	0.2545	20,113	10,609	4
PerDucer	3	76,501	0.3606	27,588	10,609	4
PerDucer	5	71,573	0.5286	37,830	10,742	4
Walk2Pers	1	81,578	0.1339	10,927	10,927	4
Walk2Pers	2	79,022	0.2613	20,651	10,927	4
Walk2Pers	3	76,501	0.3695	28,268	10,927	4
Walk2Pers	5	71,573	0.5384	38,534	11,096	4

Table 12: Cross-domain overview of the RQ1 U/H comparison and the separate RQ2 structured effects. RQ1 and RQ2 summarize their respective evaluation populations, and each structured change is computed relative to its matched H parent.

Domain	Host	RQ1 U nAUPRC	RQ1 H nAUPRC	R $\times 10^3$	T $\times 10^3$	RT $\times 10^3$	Comparison
MovieLens	FEARec	0.016	0.071	+2.450	+0.058	+1.397	guarded RQ1 / structured RQ2
MovieLens	GRU4Rec	0.107	0.236	+0.790	+0.558	+2.480	guarded RQ1 / structured RQ2
MovieLens	LRURec	0.062	0.181	-0.534	+0.376	+0.433	guarded RQ1 / structured RQ2
MovieLens	Mamba4Rec	0.081	0.216	+1.756	+1.064	+1.940	guarded RQ1 / structured RQ2
MovieLens	SASRec	0.047	0.141	+2.086	-0.539	+0.556	guarded RQ1 / structured RQ2
PENS	GTP	0.014	0.019	+4.534	+4.667	+4.603	matched RQ1 / RQ2
PENS	IMPerSumm	0.029	0.057	+2.075	+2.121	+1.719	matched RQ1 / RQ2
PENS	PENS-NAML-T1	0.023	0.037	+2.708	+2.226	+2.437	matched RQ1 / RQ2
PENS	PerDucer	0.020	0.024	+0.374	+0.951	+0.300	matched RQ1 / RQ2
PENS	Walk2Pers	0.028	0.056	+3.404	+3.535	+2.535	matched RQ1 / RQ2
Amazon	Caser	-0.026	-0.011	-5.710	-2.252	-9.274	matched RQ1 / RQ2
Amazon	GRU4Rec	-0.004	0.005	+1.225	-2.192	+1.078	matched RQ1 / RQ2
Amazon	SR-GNN	-0.040	-0.020	+2.916	+1.503	+0.181	matched RQ1 / RQ2

Table 13: MovieLens architecture-matched utility versus History–State controls.

Host	U AUPRC	H AUPRC	ΔAUPRC	U nAUPRC	H nAUPRC	ΔnAUPRC	Wins
DiffRec	0.001	0.001	-0.000001	-0.000007	-0.000008	-0.000001	16/32
DreamRec	0.000051	0.000112	+0.000062	0.000	0.000062	+0.000062	24/32
FEARec	0.114	0.162	+0.048	0.016	0.071	+0.055	32/32
GRU4Rec	0.292	0.388	+0.097	0.107	0.236	+0.129	32/32
LRURec	0.220	0.315	+0.095	0.062	0.181	+0.119	32/32
Mamba4Rec	0.245	0.352	+0.107	0.081	0.216	+0.135	32/32
SASRec	0.147	0.229	+0.083	0.047	0.141	+0.095	32/32

Table 14: MovieLens RQ1 horizon sensitivity under matched degradation settings.

Scope	h	Prevalence	U nAUPRC	H nAUPRC	ΔnAUPRC	H>U
All 7	1	0.045	0.023	0.057	+0.034	51/56
All 7	2	0.087	0.040	0.106	+0.065	51/56
All 7	3	0.124	0.051	0.140	+0.088	49/56
All 7	5	0.185	0.064	0.180	+0.117	49/56
Supported 5	1	0.062	0.032	0.080	+0.048	40/40
Supported 5	2	0.121	0.057	0.148	+0.091	40/40
Supported 5	3	0.173	0.072	0.195	+0.124	40/40
Supported 5	5	0.259	0.089	0.253	+0.163	40/40

Table 15: Five MovieLens guarded RQ1 means by host and horizon. The shared utility-history reference is sentry architecture matched.

Host	h	N	Pos.	Prev.	U nAUPRC	H nAUPRC	ΔnAUPRC	Wins
FEARec	1	115,895	4,484	0.0387	0.0094	0.0320	+0.0226	8/8
FEARec	2	112,895	8,606	0.0762	0.0148	0.0575	+0.0428	8/8
FEARec	3	109,895	12,208	0.1111	0.0177	0.0791	+0.0613	8/8
FEARec	5	103,895	18,146	0.1747	0.0212	0.1147	+0.0935	8/8
GRU4Rec	1	115,895	10,132	0.0874	0.0524	0.1065	+0.0541	8/8
GRU4Rec	2	112,895	19,148	0.1696	0.0985	0.2093	+0.1108	8/8
GRU4Rec	3	109,895	26,213	0.2385	0.1248	0.2771	+0.1523	8/8
GRU4Rec	5	103,895	36,171	0.3482	0.1519	0.3494	+0.1975	8/8
LRURec	1	115,895	7,984	0.0689	0.0322	0.0890	+0.0567	8/8
LRURec	2	112,895	15,142	0.1341	0.0544	0.1572	+0.1028	8/8
LRURec	3	109,895	21,087	0.1919	0.0705	0.2073	+0.1368	8/8
LRURec	5	103,895	29,856	0.2874	0.0905	0.2690	+0.1785	8/8
Mamba4Rec	1	115,895	8,587	0.0741	0.0414	0.1039	+0.0626	8/8
Mamba4Rec	2	112,895	16,226	0.1437	0.0715	0.1922	+0.1207	8/8
Mamba4Rec	3	109,895	22,473	0.2045	0.0924	0.2514	+0.1590	8/8
Mamba4Rec	5	103,895	31,513	0.3033	0.1174	0.3164	+0.1990	8/8
SASRec	1	115,895	4,854	0.0419	0.0233	0.0669	+0.0436	8/8
SASRec	2	112,895	9,263	0.0820	0.0437	0.1230	+0.0793	8/8
SASRec	3	109,895	12,974	0.1181	0.0536	0.1619	+0.1083	8/8
SASRec	5	103,895	18,789	0.1808	0.0661	0.2130	+0.1469	8/8

Table 16: PENS recommendation RQ1 H versus U comparison by backbone and horizon. Each row averages four H sentries. The matched causal utility score is shared within that backbone-horizon.

Backbone	h	Prev.	U nAUPRC	H nAUPRC	ΔnAUPRC	ΔAUPRC	Wins
GTP	1	0.1313	0.0064	0.0071	+0.0006	+0.0006	2/4
GTP	2	0.2563	0.0134	0.0147	+0.0013	+0.0010	2/4
GTP	3	0.3635	0.0182	0.0224	+0.0043	+0.0027	3/4
GTP	5	0.5319	0.0186	0.0335	+0.0149	+0.0070	4/4
IMPerSumm	1	0.1253	0.0109	0.0215	+0.0106	+0.0093	4/4
IMPerSumm	2	0.2442	0.0237	0.0453	+0.0216	+0.0163	4/4
IMPerSumm	3	0.3460	0.0342	0.0646	+0.0304	+0.0199	4/4
IMPerSumm	5	0.5076	0.0480	0.0952	+0.0473	+0.0233	4/4
PENS-NAML-T1	1	0.1455	0.0106	0.0125	+0.0019	+0.0017	3/4
PENS-NAML-T1	2	0.2826	0.0195	0.0292	+0.0097	+0.0070	3/4
PENS-NAML-T1	3	0.3977	0.0255	0.0424	+0.0169	+0.0102	4/4
PENS-NAML-T1	5	0.5732	0.0379	0.0637	+0.0258	+0.0110	4/4
PerDucer	1	0.1300	0.0075	0.0084	+0.0008	+0.0007	2/4
PerDucer	2	0.2545	0.0164	0.0196	+0.0032	+0.0024	2/4
PerDucer	3	0.3606	0.0227	0.0278	+0.0051	+0.0033	3/4
PerDucer	5	0.5286	0.0319	0.0406	+0.0087	+0.0041	3/4
Walk2Pers	1	0.1339	0.0125	0.0209	+0.0084	+0.0073	4/4
Walk2Pers	2	0.2613	0.0233	0.0447	+0.0214	+0.0158	4/4
Walk2Pers	3	0.3695	0.0316	0.0640	+0.0324	+0.0204	4/4
Walk2Pers	5	0.5384	0.0462	0.0940	+0.0478	+0.0221	4/4

Table 17: Full Amazon product-recommendation RQ1 H versus U by host and horizon. Two sentries per row use matched warning targets and held-out populations.

Host	h	Prev.	U nAUPRC	H nAUPRC	ΔnAUPRC	ΔAUPRC	Wins
Caser	1	0.3165	-0.0050	-0.0020	+0.0030	+0.0020	2/2
Caser	2	0.5848	-0.0242	-0.0092	+0.0151	+0.0063	2/2
Caser	3	0.7550	-0.0234	-0.0006	+0.0229	+0.0056	2/2
Caser	5	0.9104	-0.0511	-0.0332	+0.0179	+0.0016	2/2
GRU4Rec	1	0.2528	0.0058	0.0042	-0.0016	-0.0012	1/2
GRU4Rec	2	0.4753	-0.0009	0.0129	+0.0138	+0.0073	2/2
GRU4Rec	3	0.6337	-0.0042	0.0231	+0.0273	+0.0100	2/2
GRU4Rec	5	0.8180	-0.0165	-0.0185	-0.0019	-0.0004	1/2
SR-GNN	1	0.2800	-0.0095	0.0054	+0.0149	+0.0107	2/2
SR-GNN	2	0.5288	-0.0346	0.0004	+0.0349	+0.0165	2/2
SR-GNN	3	0.6974	-0.0477	-0.0127	+0.0351	+0.0106	2/2
SR-GNN	5	0.8689	-0.0685	-0.0733	-0.0048	-0.0006	1/2

Table 18: Amazon probability quality. Positive ΔBrier denotes a reduction in error from U to H. NLL and ECE15 are reported as absolute U/H scores so negative changes remain observable.

Host	h	U Brier	H Brier	ΔBrier	U NLL	H NLL	U ECE	H ECE
Caser	1	0.2264	0.2216	+0.0048	0.6464	0.6361	0.0767	0.0565
Caser	2	0.2615	0.2600	+0.0016	0.7253	0.7205	0.1174	0.1112
Caser	3	0.1995	0.2014	-0.0020	0.6181	0.6266	0.1059	0.1198
Caser	5	0.0877	0.0871	+0.0005	0.3957	0.4151	0.0725	0.0719
GRU4Rec	1	0.2127	0.1938	+0.0189	0.6315	0.5770	0.1060	0.0540
GRU4Rec	2	0.3245	0.2816	+0.0429	0.9099	0.7647	0.2359	0.1519
GRU4Rec	3	0.2921	0.2778	+0.0143	0.9253	0.8165	0.2240	0.2022
GRU4Rec	5	0.1749	0.1735	+0.0015	0.8472	0.7627	0.1549	0.1531
SR-GNN	1	0.2264	0.2067	+0.0197	0.6526	0.6046	0.1102	0.0562
SR-GNN	2	0.2941	0.2704	+0.0237	0.8194	0.7404	0.1717	0.1245
SR-GNN	3	0.2426	0.2391	+0.0035	0.7571	0.7209	0.1531	0.1558
SR-GNN	5	0.1267	0.1267	-0.0001	0.6060	0.6047	0.1053	0.1103

Table 19: PENS RQ1 sentry-horizon cells, GTP (part 1/5). Together these five tables report all 80 matched cells. U is the same causal control for the sentries of a given backbone-horizon.

Backbone	Sentry	h	U nAUPRC	H nAUPRC	ΔnAUPRC
GTP	ls-eep	1	0.0064	0.0073	+0.0009
GTP	ls-eep	2	0.0134	0.0163	+0.0029
GTP	ls-eep	3	0.0182	0.0200	+0.0018
GTP	ls-eep	5	0.0186	0.0271	+0.0085
GTP	prefix-mlp	1	0.0064	0.0052	-0.0013
GTP	prefix-mlp	2	0.0134	0.0131	-0.0003
GTP	prefix-mlp	3	0.0182	0.0251	+0.0069
GTP	prefix-mlp	5	0.0186	0.0388	+0.0202
GTP	tcn-elects-ew	1	0.0064	0.0095	+0.0031
GTP	tcn-elects-ew	2	0.0134	0.0214	+0.0080
GTP	tcn-elects-ew	3	0.0182	0.0293	+0.0112
GTP	tcn-elects-ew	5	0.0186	0.0457	+0.0271
GTP	tcsr	1	0.0064	0.0062	-0.0002
GTP	tcsr	2	0.0134	0.0081	-0.0053
GTP	tcsr	3	0.0182	0.0154	-0.0028
GTP	tcsr	5	0.0186	0.0223	+0.0037

Table 20: PENS RQ1 sentry-horizon cells, IMPerSumm (part 2/5). Together these five tables report all 80 matched cells. U is the same causal control for the sentries of a given backbone-horizon.

Backbone	Sentry	h	U nAUPRC	H nAUPRC	ΔnAUPRC
IMPerSumm	ls-eep	1	0.0109	0.0216	+0.0107
IMPerSumm	ls-eep	2	0.0237	0.0450	+0.0213
IMPerSumm	ls-eep	3	0.0342	0.0644	+0.0302
IMPerSumm	ls-eep	5	0.0480	0.0980	+0.0500
IMPerSumm	prefix-mlp	1	0.0109	0.0212	+0.0104
IMPerSumm	prefix-mlp	2	0.0237	0.0448	+0.0211
IMPerSumm	prefix-mlp	3	0.0342	0.0634	+0.0293
IMPerSumm	prefix-mlp	5	0.0480	0.0955	+0.0475
IMPerSumm	tcn-elects-ew	1	0.0109	0.0223	+0.0114
IMPerSumm	tcn-elects-ew	2	0.0237	0.0470	+0.0233
IMPerSumm	tcn-elects-ew	3	0.0342	0.0675	+0.0333
IMPerSumm	tcn-elects-ew	5	0.0480	0.0962	+0.0482
IMPerSumm	tcsr	1	0.0109	0.0209	+0.0101
IMPerSumm	tcsr	2	0.0237	0.0445	+0.0208
IMPerSumm	tcsr	3	0.0342	0.0631	+0.0289
IMPerSumm	tcsr	5	0.0480	0.0912	+0.0433

Table 21: PENS RQ1 sentry-horizon cells, PENS-NAML-T1 (part 3/5). Together these five tables report all 80 matched cells. U is the same causal control for the sentries of a given backbone-horizon.

Backbone	Sentry	h	U nAUPRC	H nAUPRC	ΔnAUPRC
PENS-NAML-T1	ls-eep	1	0.0106	0.0111	+0.0005
PENS-NAML-T1	ls-eep	2	0.0195	0.0265	+0.0070
PENS-NAML-T1	ls-eep	3	0.0255	0.0399	+0.0144
PENS-NAML-T1	ls-eep	5	0.0379	0.0652	+0.0274
PENS-NAML-T1	prefix-mlp	1	0.0106	0.0166	+0.0060
PENS-NAML-T1	prefix-mlp	2	0.0195	0.0395	+0.0200
PENS-NAML-T1	prefix-mlp	3	0.0255	0.0551	+0.0296
PENS-NAML-T1	prefix-mlp	5	0.0379	0.0835	+0.0456
PENS-NAML-T1	tcn-elects-ew	1	0.0106	0.0141	+0.0035
PENS-NAML-T1	tcn-elects-ew	2	0.0195	0.0320	+0.0125
PENS-NAML-T1	tcn-elects-ew	3	0.0255	0.0466	+0.0211
PENS-NAML-T1	tcn-elects-ew	5	0.0379	0.0619	+0.0241
PENS-NAML-T1	tcsr	1	0.0106	0.0083	-0.0023
PENS-NAML-T1	tcsr	2	0.0195	0.0188	-0.0007
PENS-NAML-T1	tcsr	3	0.0255	0.0279	+0.0024
PENS-NAML-T1	tcsr	5	0.0379	0.0440	+0.0061

Table 22: PENS RQ1 sentry-horizon cells, PerDucer (part 4/5). Together these five tables report all 80 matched cells. U is the same causal control for the sentries of a given backbone-horizon.

Backbone	Sentry	h	U nAUPRC	H nAUPRC	ΔnAUPRC
PerDucer	ls-eep	1	0.0075	0.0073	-0.0002
PerDucer	ls-eep	2	0.0164	0.0163	-0.000021
PerDucer	ls-eep	3	0.0227	0.0234	+0.0008
PerDucer	ls-eep	5	0.0319	0.0336	+0.0017
PerDucer	prefix-mlp	1	0.0075	0.0063	-0.0012
PerDucer	prefix-mlp	2	0.0164	0.0138	-0.0026
PerDucer	prefix-mlp	3	0.0227	0.0205	-0.0021
PerDucer	prefix-mlp	5	0.0319	0.0306	-0.0013
PerDucer	tcn-elects-ew	1	0.0075	0.0121	+0.0046
PerDucer	tcn-elects-ew	2	0.0164	0.0287	+0.0124
PerDucer	tcn-elects-ew	3	0.0227	0.0406	+0.0179
PerDucer	tcn-elects-ew	5	0.0319	0.0608	+0.0289
PerDucer	tcsr	1	0.0075	0.0077	+0.0001
PerDucer	tcsr	2	0.0164	0.0196	+0.0032
PerDucer	tcsr	3	0.0227	0.0267	+0.0040
PerDucer	tcsr	5	0.0319	0.0373	+0.0054

Table 23: PENS RQ1 sentry-horizon cells, Walk2Pers (part 5/5). Together these five tables report all 80 matched cells. U is the same causal control for the sentries of a given backbone-horizon.

Backbone	Sentry	h	U nAUPRC	H nAUPRC	ΔnAUPRC
Walk2Pers	ls-eep	1	0.0125	0.0227	+0.0102
Walk2Pers	ls-eep	2	0.0233	0.0480	+0.0247
Walk2Pers	ls-eep	3	0.0316	0.0682	+0.0366
Walk2Pers	ls-eep	5	0.0462	0.0980	+0.0517
Walk2Pers	prefix-mlp	1	0.0125	0.0236	+0.0111
Walk2Pers	prefix-mlp	2	0.0233	0.0491	+0.0258
Walk2Pers	prefix-mlp	3	0.0316	0.0698	+0.0382
Walk2Pers	prefix-mlp	5	0.0462	0.1012	+0.0550
Walk2Pers	tcn-elects-ew	1	0.0125	0.0226	+0.0101
Walk2Pers	tcn-elects-ew	2	0.0233	0.0497	+0.0265
Walk2Pers	tcn-elects-ew	3	0.0316	0.0706	+0.0390
Walk2Pers	tcn-elects-ew	5	0.0462	0.1021	+0.0558
Walk2Pers	tcsr	1	0.0125	0.0147	+0.0022
Walk2Pers	tcsr	2	0.0233	0.0318	+0.0086
Walk2Pers	tcsr	3	0.0316	0.0474	+0.0159
Walk2Pers	tcsr	5	0.0462	0.0749	+0.0286

Table 24: Amazon RQ1 sentry-horizon cells, preserving the exact two-sentry comparison and its below-prevalence absolute H cases.

Host	Sentry	h	U nAUPRC	H nAUPRC	ΔnAUPRC
Caser	prefix-mlp	1	-0.0056	-0.0038	+0.0019
Caser	prefix-mlp	2	-0.0151	-0.0062	+0.0089
Caser	prefix-mlp	3	-0.0189	0.0185	+0.0374
Caser	prefix-mlp	5	-0.0439	-0.0397	+0.0042
Caser	tcn-elects-ew	1	-0.0043	-0.0002	+0.0041
Caser	tcn-elects-ew	2	-0.0334	-0.0121	+0.0213
Caser	tcn-elects-ew	3	-0.0280	-0.0196	+0.0084
Caser	tcn-elects-ew	5	-0.0583	-0.0267	+0.0316
GRU4Rec	prefix-mlp	1	-0.0012	0.0065	+0.0077
GRU4Rec	prefix-mlp	2	0.0007	0.0104	+0.0097
GRU4Rec	prefix-mlp	3	-0.0147	0.0187	+0.0335
GRU4Rec	prefix-mlp	5	-0.0749	-0.0236	+0.0512
GRU4Rec	tcn-elects-ew	1	0.0128	0.0020	-0.0109
GRU4Rec	tcn-elects-ew	2	-0.0025	0.0154	+0.0179
GRU4Rec	tcn-elects-ew	3	0.0063	0.0275	+0.0212
GRU4Rec	tcn-elects-ew	5	0.0418	-0.0133	-0.0551
SR-GNN	prefix-mlp	1	-0.0088	0.0024	+0.0113
SR-GNN	prefix-mlp	2	-0.0315	-0.0084	+0.0232
SR-GNN	prefix-mlp	3	-0.0469	-0.0218	+0.0250
SR-GNN	prefix-mlp	5	-0.0816	-0.0765	+0.0051
SR-GNN	tcn-elects-ew	1	-0.0101	0.0085	+0.0185
SR-GNN	tcn-elects-ew	2	-0.0376	0.0091	+0.0467
SR-GNN	tcn-elects-ew	3	-0.0486	-0.0035	+0.0452
SR-GNN	tcn-elects-ew	5	-0.0554	-0.0701	-0.0147

Table 25: MovieLens structured effects relative to the same frozen H score, by horizon. All five supported hosts and eight sentries are retained. Values are scaled by 10^3. No confidence intervals are inferred from aggregate rows.

h	View	ΔAUPRC $\times 10^3$	ΔnAUPRC $\times 10^3$	Median ΔnAUPRC	Positive cells	Cells
1	R	+1.197	+1.277	+0.852	60.0%	40
1	T	+0.686	+0.739	+0.107	55.0%	40
1	RT	+0.651	+0.702	-0.169	47.5%	40
2	R	+1.188	+1.344	+0.857	60.0%	40
2	T	+0.435	+0.513	+0.655	52.5%	40
2	RT	+1.127	+1.302	+1.069	57.5%	40
3	R	+1.050	+1.214	+0.938	60.0%	40
3	T	-0.040	-0.024	+0.101	50.0%	40
3	RT	+1.132	+1.402	+0.930	65.0%	40
5	R	+1.138	+1.403	+1.379	62.5%	40
5	T	-0.043	-0.013	+0.013	50.0%	40
5	RT	+1.461	+2.039	+1.960	57.5%	40

Table 26: MovieLens R/T/RT changes from frozen H, averaged across the eight sentries and four horizons. These figures come from the matched five-host structured-refinement experiment and are distinct from the guarded RQ1 comparison.

Host	View	ΔAUPRC $\times 10^3$	ΔnAUPRC $\times 10^3$	Positive cells	Positive horizons	Cells
FEARec	R	+2.165	+2.450	87.5%	4/4	32
FEARec	T	+0.080	+0.058	56.2%	3/4	32
FEARec	RT	+1.239	+1.397	65.6%	4/4	32
GRU4Rec	R	+0.606	+0.790	37.5%	4/4	32
GRU4Rec	T	+0.395	+0.558	50.0%	3/4	32
GRU4Rec	RT	+1.835	+2.480	62.5%	4/4	32
LRURec	R	-0.372	-0.534	37.5%	2/4	32
LRURec	T	+0.374	+0.376	50.0%	2/4	32
LRURec	RT	+0.332	+0.433	37.5%	4/4	32
Mamba4Rec	R	+1.508	+1.756	78.1%	4/4	32
Mamba4Rec	T	+0.936	+1.064	56.2%	4/4	32
Mamba4Rec	RT	+1.592	+1.940	71.9%	4/4	32
SASRec	R	+1.809	+2.086	62.5%	4/4	32
SASRec	T	-0.490	-0.539	46.9%	1/4	32
SASRec	RT	+0.466	+0.556	46.9%	3/4	32

Table 27: MovieLens matched joint-refinement comparisons. $RT - \max(R, T)$ is computed within each host-sentry-horizon cell, not from the maximum of domain means. No additive benefit is assumed.

h	RT−R $\times 10^3$	RT>R	RT−T $\times 10^3$	RT>T	RT−best $\times 10^3$	RT>both
1	-0.575	40.0%	-0.037	52.5%	-1.429	12/40
2	-0.042	55.0%	+0.789	47.5%	-1.090	13/40
3	+0.188	55.0%	+1.426	57.5%	-1.107	16/40
5	+0.636	52.5%	+2.052	60.0%	-0.929	16/40
All	+0.052	50.6%	+1.058	54.4%	-1.139	57/160

Table 28: PENS recommendation: complete matched structured effects from H.

View	ΔAUPRC $\times 10^3$	ΔnAUPRC $\times 10^3$	AUPRC win	ΔEventRecall	ΔMeanLead	ΔTEST FAR	Cells
HR	+1.639	+2.619	93.8%	+0.00509	-0.00450	+0.00104	80
HT	+1.686	+2.700	95.0%	+0.00596	-0.00299	+0.00123	80
HRT	+1.455	+2.319	91.2%	+0.00290	-0.00509	-0.00024	80

Table 29: PENS recommendation structured behavior by preference backbone.

Backbone	View	ΔnAUPRC $\times 10^3$	ΔRecall	ΔLead	ΔTEST FAR	Positive cells
GTP	HR	+4.534	+0.0333	+0.0156	+0.0213	16/16
GTP	HT	+4.667	+0.0323	+0.0259	+0.0215	16/16
GTP	HRT	+4.603	+0.0312	+0.0201	+0.0207	16/16
IMPerSumm	HR	+2.075	+0.0056	-0.0009	+0.0008	16/16
IMPerSumm	HT	+2.121	+0.0036	-0.0021	-0.0003	16/16
IMPerSumm	HRT	+1.719	+0.0047	+0.0008	+0.0010	15/16
PENS-NAML-T1	HR	+2.708	+0.0003	-0.0051	-0.0022	16/16
PENS-NAML-T1	HT	+2.226	-0.0025	-0.0051	-0.0051	15/16
PENS-NAML-T1	HRT	+2.437	-0.0004	-0.0095	-0.0040	15/16
PerDucer	HR	+0.374	-0.0226	-0.0179	-0.0168	11/16
PerDucer	HT	+0.951	-0.0123	-0.0160	-0.0117	13/16
PerDucer	HRT	+0.300	-0.0286	-0.0220	-0.0209	11/16
Walk2Pers	HR	+3.404	+0.0088	-0.0142	+0.0022	16/16
Walk2Pers	HT	+3.535	+0.0088	-0.0177	+0.0018	16/16
Walk2Pers	HRT	+2.535	+0.0076	-0.0148	+0.0020	16/16

Table 30: RQ2 absolute structured-warning metrics for PENS recommendation and Amazon product recommendation.

Domain	View	AUPRC	nAUPRC	Recall	Lead	TEST FAR	Brier	NLL	ECE15
PENS	H	0.347	0.039	0.1959	3.2294	0.1103	0.1967	0.5773	0.0126
PENS	HR	0.349	0.041	0.2010	3.2249	0.1113	0.1970	0.5780	0.0128
PENS	HT	0.349	0.041	0.2019	3.2264	0.1115	0.1969	0.5776	0.0120
PENS	HRT	0.349	0.041	0.1988	3.2243	0.1100	0.1969	0.5779	0.0130
Amazon	H	0.593	-0.009	0.3455	3.0462	0.3231	0.2116	0.6658	0.1140
Amazon	HR	0.594	-0.009	0.3535	3.0606	0.3388	0.2166	0.7105	0.1292
Amazon	HT	0.593	-0.010	0.3540	3.0657	0.3427	0.2172	0.7070	0.1301
Amazon	HRT	0.593	-0.011	0.3522	3.0609	0.3380	0.2174	0.7101	0.1308

Table 31: Structured-evaluation coverage. FAR is realized on held-out evaluation.

Domain	View	Cells	Hosts	Sentries	Horizons	Positives range	TEST FAR
PENS	H	80	5	4	4	10,218–41,026	0.1103
PENS	HR	80	5	4	4	10,218–41,026	0.1113
PENS	HT	80	5	4	4	10,218–41,026	0.1115
PENS	HRT	80	5	4	4	10,218–41,026	0.1100
Amazon	H	24	3	2	4	3,882–8,574	0.3231
Amazon	HR	24	3	2	4	3,882–8,574	0.3388
Amazon	HT	24	3	2	4	3,882–8,574	0.3427
Amazon	HRT	24	3	2	4	3,882–8,574	0.3380

Table 32: PENS/Amazon operational behavior across horizons. Thresholds are fixed before held-out evaluation, and realized false-alarm burden is reported separately.

Domain	h	View	Cells	Event recall	Mean lead	TEST FAR
PENS	1	H	20	0.1412	2.0000	0.1172
PENS	1	HR	20	0.1420	2.0000	0.1173
PENS	1	HT	20	0.1421	2.0000	0.1168
PENS	1	HRT	20	0.1404	2.0000	0.1154
PENS	2	H	20	0.1822	2.7428	0.1128
PENS	2	HR	20	0.1876	2.7386	0.1134
PENS	2	HT	20	0.1870	2.7372	0.1133
PENS	2	HRT	20	0.1853	2.7355	0.1121
PENS	3	H	20	0.2047	3.4307	0.1056
PENS	3	HR	20	0.2144	3.4211	0.1087
PENS	3	HT	20	0.2161	3.4281	0.1096
PENS	3	HRT	20	0.2112	3.4243	0.1075
PENS	5	H	20	0.2557	4.7441	0.1054
PENS	5	HR	20	0.2601	4.7399	0.1059
PENS	5	HT	20	0.2623	4.7404	0.1062
PENS	5	HRT	20	0.2583	4.7374	0.1051
Amazon	1	H	6	0.1015	2.0000	0.1357
Amazon	1	HR	6	0.1040	2.0000	0.1386
Amazon	1	HT	6	0.1057	2.0000	0.1407
Amazon	1	HRT	6	0.1057	2.0000	0.1402
Amazon	2	H	6	0.2209	2.5614	0.1923
Amazon	2	HR	6	0.2290	2.5634	0.2006
Amazon	2	HT	6	0.2288	2.5647	0.2010
Amazon	2	HRT	6	0.2262	2.5670	0.1991
Amazon	3	H	6	0.3532	3.1280	0.2713
Amazon	3	HR	6	0.3669	3.1301	0.2841
Amazon	3	HT	6	0.3670	3.1373	0.2858
Amazon	3	HRT	6	0.3635	3.1290	0.2788
Amazon	5	H	6	0.7064	4.4955	0.6931
Amazon	5	HR	6	0.7143	4.5487	0.7322
Amazon	5	HT	6	0.7145	4.5607	0.7433
Amazon	5	HRT	6	0.7134	4.5475	0.7339

Table 33: PENS recommendation matched-cell multi-metric consistency. Balanced requires positive AUPRC change and nonnegative recall and lead changes. Strict triple requires all three to be strictly positive.

View	Cells	AUPRC win	Recall>0	Lead>0	Balanced	Strict triple	ΔTEST FAR
HR	80	93.8%	67.5%	26.2%	33.8%	18.8%	+0.00104
HT	80	95.0%	67.5%	25.0%	31.2%	16.2%	+0.00123
HRT	80	91.2%	63.7%	26.2%	31.2%	16.2%	-0.00024

Table 34: Amazon absolute metrics by host and structured realization.

Host	View	nAUPRC	Recall	Lead	TEST FAR	Brier	NLL	ECE15
Caser	H	-0.0112	0.3455	3.0279	0.3186	0.1925	0.5996	0.0899
Caser	HR	-0.0169	0.3556	3.0492	0.3374	0.1970	0.6261	0.1049
Caser	HT	-0.0135	0.3547	3.0568	0.3410	0.1980	0.6319	0.1082
Caser	HRT	-0.0205	0.3546	3.0524	0.3400	0.1973	0.6289	0.1049
GRU4Rec	H	0.0054	0.3516	3.0561	0.3274	0.2317	0.7302	0.1403
GRU4Rec	HR	0.0067	0.3628	3.0719	0.3436	0.2357	0.7855	0.1501
GRU4Rec	HT	0.0032	0.3627	3.0765	0.3464	0.2375	0.7826	0.1548
GRU4Rec	HRT	0.0065	0.3616	3.0695	0.3405	0.2387	0.7882	0.1580
SR-GNN	H	-0.0200	0.3392	3.0547	0.3234	0.2107	0.6677	0.1117
SR-GNN	HR	-0.0171	0.3422	3.0605	0.3356	0.2172	0.7197	0.1325
SR-GNN	HT	-0.0185	0.3445	3.0637	0.3406	0.2162	0.7067	0.1274
SR-GNN	HRT	-0.0199	0.3404	3.0606	0.3335	0.2163	0.7132	0.1293

Table 35: Amazon matched structured changes relative to H, by host and horizon. Positive Brier improvement means smaller Brier error. Positive TEST FAR means greater held-out alarm burden.

Host	h	View	ΔnAUPRC $\times 10^3$	ΔRecall	ΔLead	ΔFAR	ΔBrier gain
Caser	1	HR	-1.336	+0.0000	+0.0000	+0.0026	-0.0059
Caser	1	HT	-0.019	+0.0029	+0.0000	+0.0039	-0.0066
Caser	1	HRT	-0.322	+0.0021	+0.0000	+0.0040	-0.0065
Caser	2	HR	-0.298	+0.0080	+0.0018	+0.0060	-0.0064
Caser	2	HT	+0.433	+0.0066	+0.0079	+0.0035	-0.0083
Caser	2	HRT	-1.615	+0.0060	+0.0058	+0.0060	-0.0069
Caser	3	HR	-4.553	+0.0230	+0.0103	+0.0226	-0.0042
Caser	3	HT	-7.411	+0.0209	+0.0243	+0.0259	-0.0055
Caser	3	HRT	-7.040	+0.0205	+0.0078	+0.0199	-0.0042
Caser	5	HR	-16.652	+0.0093	+0.0730	+0.0438	-0.0013
Caser	5	HT	-2.014	+0.0062	+0.0833	+0.0564	-0.0015
Caser	5	HRT	-28.119	+0.0079	+0.0844	+0.0557	-0.0014
GRU4Rec	1	HR	+3.394	+0.0067	+0.0000	+0.0035	-0.0043
GRU4Rec	1	HT	+2.711	+0.0069	+0.0000	+0.0042	-0.0062
GRU4Rec	1	HRT	+3.118	+0.0073	+0.0000	+0.0041	-0.0068
GRU4Rec	2	HR	+0.131	+0.0109	+0.0107	+0.0146	-0.0035
GRU4Rec	2	HT	-0.848	+0.0102	+0.0090	+0.0150	-0.0077
GRU4Rec	2	HRT	+0.420	+0.0088	+0.0086	+0.0113	-0.0104
GRU4Rec	3	HR	+2.044	+0.0177	+0.0007	+0.0102	-0.0047
GRU4Rec	3	HT	-0.404	+0.0157	+0.0062	+0.0082	-0.0055
GRU4Rec	3	HRT	+2.256	+0.0142	-0.0006	+0.0036	-0.0074
GRU4Rec	5	HR	-0.667	+0.0092	+0.0522	+0.0366	-0.0037
GRU4Rec	5	HT	-10.227	+0.0117	+0.0665	+0.0486	-0.0038
GRU4Rec	5	HRT	-1.483	+0.0094	+0.0459	+0.0336	-0.0036
SR-GNN	1	HR	+3.872	+0.0009	+0.0000	+0.0026	-0.0074
SR-GNN	1	HT	+1.999	+0.0031	+0.0000	+0.0068	-0.0068
SR-GNN	1	HRT	+2.747	+0.0033	+0.0000	+0.0056	-0.0047
SR-GNN	2	HR	+4.046	+0.0055	-0.0065	+0.0041	-0.0092
SR-GNN	2	HT	+3.273	+0.0069	-0.0071	+0.0075	-0.0078
SR-GNN	2	HRT	+2.116	+0.0014	+0.0022	+0.0031	-0.0088
SR-GNN	3	HR	-1.831	+0.0006	-0.0047	+0.0054	-0.0071
SR-GNN	3	HT	-0.108	+0.0048	-0.0025	+0.0093	-0.0054
SR-GNN	3	HRT	-3.721	-0.0037	-0.0044	-0.0012	-0.0068
SR-GNN	5	HR	+5.578	+0.0051	+0.0344	+0.0367	-0.0021
SR-GNN	5	HT	+0.847	+0.0065	+0.0456	+0.0453	-0.0020
SR-GNN	5	HRT	-0.418	+0.0037	+0.0257	+0.0331	-0.0020

Table 36: PENS recommendation matched R/T/RT changes from frozen H, averaged across five backbones and four sentries per horizon.

h	View	ΔnAUPRC $\times 10^3$	ΔRecall	ΔLead	ΔTEST FAR	nAUPRC wins
1	HR	+1.358	+0.00086	+0.00000	+0.00004	20/20
1	HT	+1.435	+0.00099	+0.00000	-0.00039	20/20
1	HRT	+1.263	-0.00074	+0.00000	-0.00183	20/20
2	HR	+2.589	+0.00541	-0.00420	+0.00057	19/20
2	HT	+2.487	+0.00480	-0.00564	+0.00048	20/20
2	HRT	+2.327	+0.00317	-0.00733	-0.00069	20/20
3	HR	+2.712	+0.00968	-0.00960	+0.00305	18/20
3	HT	+2.816	+0.01142	-0.00262	+0.00403	18/20
3	HRT	+2.299	+0.00651	-0.00639	+0.00184	16/20
5	HR	+3.817	+0.00441	-0.00419	+0.00051	18/20
5	HT	+4.061	+0.00663	-0.00371	+0.00080	18/20
5	HRT	+3.388	+0.00266	-0.00664	-0.00029	17/20

Table 37: MovieLens source-only H reuse by held-out encoder. Retention is the mean of matched cell-wise nAUPRC ratios, not the ratio of aggregate means. No held-out warning scores are used to construct the source consensus.

Held-out model	Source AUPRC	Self AUPRC	Source nAUPRC	Self nAUPRC	Retention	Positive
FEARec	0.166	0.165	0.075	0.073	100.93%	32/32
GRU4Rec	0.361	0.396	0.196	0.242	81.12%	32/32
LRURec	0.309	0.326	0.170	0.191	87.70%	32/32
Mamba4Rec	0.343	0.360	0.202	0.224	90.35%	32/32
SASRec	0.211	0.251	0.117	0.163	72.13%	32/32

Table 38: MovieLens source-only H warning by sentry architecture over five held-out hosts and four horizons. Sentry-specific retention uses matched cell ratios.

Sentry	Source nAUPRC	Self nAUPRC	Retention	Source>0	Source>self
deephit-style	0.152	0.178	86.60%	20/20	3/20
drsa-style	0.160	0.180	90.22%	20/20	3/20
ls-eep	0.144	0.178	82.38%	20/20	2/20
prefix-mlp	0.152	0.175	89.17%	20/20	4/20
tcn-elects-ew	0.155	0.177	88.22%	20/20	1/20
tcsa	0.144	0.180	80.75%	20/20	1/20
tcsr	0.158	0.181	88.56%	20/20	3/20
temporal-prior	0.151	0.180	85.66%	20/20	3/20

Table 39: MovieLens source-only H retention distribution. Each horizon contains 40 held-out host-sentry comparisons. Mean retention is 86.446% (86.4% at one decimal).

h	Source nAUPRC	Self nAUPRC	Mean ret.	Median ret.	Min ret.	Max ret.	Cells
1	0.070	0.084	85.15%	84.80%	68.97%	105.00%	40
2	0.132	0.156	86.02%	85.59%	67.08%	112.68%	40
3	0.176	0.209	86.06%	86.32%	66.15%	111.17%	40
5	0.231	0.266	88.55%	89.65%	67.93%	110.93%	40
All	0.152	0.179	86.45%	85.72%	66.15%	112.68%	160

Table 40: PENS source-only transfer by held-out encoder. Retention is the mean matched-cell ratio of transferred to self nAUPRC. Values over 100% are possible.

Held-out	Source AUPRC	Self AUPRC	Source nAUPRC	Self nAUPRC	Retention	Positive
GTP	0.336	0.333	0.024	0.019	136.90%	16/16
IMPerSumm	0.329	0.341	0.036	0.057	65.25%	16/16
PENS-NAML-T1	0.367	0.371	0.031	0.037	86.31%	16/16
PerDucer	0.332	0.333	0.023	0.024	101.74%	16/16
Walk2Pers	0.335	0.360	0.014	0.056	27.69%	16/16

Table 41: PENS source-only H reuse by horizon. Retention is averaged over 20 matched target-sentry cells per horizon.

h	Source nAUPRC	Self nAUPRC	Mean ret.	Median ret.	Cells
1	0.0103	0.0141	89.21%	85.83%	20
2	0.0217	0.0307	88.29%	80.59%	20
3	0.0293	0.0442	80.39%	76.72%	20
5	0.0418	0.0654	76.42%	75.21%	20
All	0.0258	0.0386	83.58%	79.54%	80

Table 42: Structured source-only reuse relative to the source-only H consensus. The transferred correction is measured within each matched target-sentry-horizon cell.

Domain	View	ΔnAUPRC $\times 10^3$	Positive cells	Median ΔnAUPRC	Cells
MovieLens	R	-1.737	43.8%	-0.00077	160
MovieLens	T	-2.005	39.4%	-0.00088	160
MovieLens	RT	-1.483	47.5%	-0.00010	160
PENS	R	+1.473	66.2%	+0.00082	80
PENS	T	+1.400	70.0%	+0.00073	80
PENS	RT	+1.649	68.8%	+0.00112	80

Table 43: MovieLens source-only adaptation, reported separately from the unadapted experiment. Bootstrap intervals are computed for the adapted-minus-source H nAUPRC difference.

h	Source nAUPRC	Adapted nAUPRC	ΔnAUPRC $\times 10^4$	Wins	Saved 95% CI	Source ret.	Adapted ret.
1	0.07029	0.07032	+0.298	26/40	[+0.059, +0.627]	85.15%	85.19%
2	0.13160	0.13169	+0.889	30/40	[+0.640, +1.184]	86.02%	86.09%
3	0.17561	0.17570	+0.893	28/40	[+0.540, +1.247]	86.06%	86.12%
5	0.23058	0.23076	+1.808	29/40	[+1.271, +2.345]	88.55%	88.63%

Table 44: Supplementary MovieLens source-only adaptation by held-out host. The adaptation uses neither target labels nor target warning scores and is reported separately from the unadapted main transfer estimates.

Held-out	Source nAUPRC	Adapted nAUPRC	ΔnAUPRC $\times 10^4$	Wins	Source ret.	Adapted ret.
FEARec	0.07486	0.07496	+1.040	31/32	100.93%	101.06%
GRU4Rec	0.19586	0.19592	+0.636	22/32	81.12%	81.15%
LRURec	0.17014	0.17018	+0.488	19/32	87.70%	87.72%
Mamba4Rec	0.20218	0.20233	+1.533	18/32	90.35%	90.42%
SASRec	0.11706	0.11718	+1.163	23/32	72.13%	72.19%

Table 45: Mean convex adaptation weights across folds for five source-only held-out encoders. w_H denotes the broad H warning; the remaining weights correspond to the R, T, and RT warning variants.

h	w_H	w_R	w_T	w_{RT}
1	95.80%	0.73%	1.92%	1.56%
2	95.76%	0.70%	2.14%	1.40%
3	95.57%	0.78%	2.22%	1.43%
5	95.19%	1.02%	1.99%	1.81%
Mean	95.58%	0.81%	2.06%	1.55%

Table 46: PENS five-backbone, two-task trajectory-level warning.

Backbone	Task	U AUPRC	H AUPRC	HR AUPRC	HT AUPRC	HRT AUPRC	Low-u rate
GTP	News rec.	0.261	0.361	0.369	0.365	0.367	0.407
GTP	Headline gen.	0.311	0.333	0.340	0.339	0.339	0.342
IMPerSumm	News rec.	0.247	0.324	0.323	0.322	0.324	0.389
IMPerSumm	Headline gen.	0.316	0.342	0.344	0.342	0.342	0.347
PENS-NAML-T1	News rec.	0.233	0.341	0.338	0.338	0.339	0.367
PENS-NAML-T1	Headline gen.	0.330	0.368	0.369	0.366	0.368	0.376
PerDucer	News rec.	0.254	0.337	0.337	0.341	0.339	0.402
PerDucer	Headline gen.	0.317	0.359	0.356	0.358	0.357	0.353
Walk2Pers	News rec.	0.242	0.323	0.319	0.321	0.321	0.381
Walk2Pers	Headline gen.	0.306	0.304	0.305	0.306	0.307	0.341

Table 47: PENS trajectory-level warning by task and prospective horizon. Means span the same five backbones and four warning sentries in each task.

Task	h	U AUPRC	H AUPRC	HR AUPRC	HT AUPRC	HRT AUPRC	Low-u rate
News rec.	1	0.247	0.340	0.340	0.339	0.340	0.389
News rec.	2	0.247	0.338	0.338	0.339	0.339	0.389
News rec.	3	0.247	0.337	0.337	0.337	0.338	0.389
News rec.	5	0.247	0.334	0.334	0.334	0.335	0.389
Headline gen.	1	0.316	0.345	0.346	0.346	0.346	0.352
Headline gen.	2	0.316	0.342	0.344	0.344	0.344	0.352
Headline gen.	3	0.316	0.341	0.342	0.342	0.342	0.352
Headline gen.	5	0.316	0.337	0.339	0.338	0.339	0.352

Table 48: PENS trajectory-level warning by sentry and task. Each sentry-task row averages five matched backbones and four horizons.

Task	Sentry	U AUPRC	H AUPRC	HR AUPRC	HT AUPRC	HRT AUPRC	Cells
News rec.	ls-eep	0.247	0.338	0.341	0.340	0.340	20
News rec.	prefix-mlp	0.247	0.341	0.340	0.339	0.340	20
News rec.	tcn-elects-ew	0.247	0.333	0.330	0.330	0.330	20
News rec.	tcsr	0.247	0.337	0.338	0.340	0.342	20
Headline gen.	ls-eep	0.316	0.343	0.346	0.344	0.346	20
Headline gen.	prefix-mlp	0.316	0.345	0.346	0.345	0.345	20
Headline gen.	tcn-elects-ew	0.316	0.343	0.341	0.340	0.340	20
Headline gen.	tcsr	0.316	0.334	0.339	0.341	0.340	20

Table 49: Five-backbone task comparison: trajectory-level H warning under recommendation versus generation. Task-specific low-utility prevalences differ. Differences describe within-model task effects, not zero-shot cross-task transfer.

Backbone	Rec rate	Rec H AUPRC	Gen rate	Gen H AUPRC	Gen−Rec AUPRC	Gen RT−H	Cells rec/gen
GTP	0.4067	0.361	0.3417	0.333	-0.0275	+0.00594	16;16
IMPerSumm	0.3887	0.324	0.3472	0.342	+0.0177	+0.00004	16;16
PENS-NAML-T1	0.3672	0.341	0.3757	0.368	+0.0275	-0.00022	16;16
PerDucer	0.4017	0.337	0.3532	0.359	+0.0216	-0.00125	16;16
Walk2Pers	0.3807	0.323	0.3407	0.304	-0.0195	+0.00339	16;16

Table 50: Same five PENS backbones under recommendation and generation. Changes are from the H warning for the corresponding task and use trajectory-level AUPRC, low-utility recall, mean lead in steps, and flag rate.

Task	View	Δtraj. AUPRC	Δrecall	Δlead steps	Δflag rate	Cells
News rec.	HR	-0.000010	+0.02076	+0.08549	+0.02075	80
News rec.	HT	+0.000107	+0.01710	+0.08248	+0.01683	80
News rec.	HRT	+0.000743	+0.02406	+0.09500	+0.02267	80
Headline gen.	HR	+0.001686	+0.02238	+0.12879	+0.02075	80
Headline gen.	HT	+0.001116	+0.01852	+0.06533	+0.01683	80
Headline gen.	HRT	+0.001582	+0.02585	+0.13041	+0.02267	80

Table 51: Open-weight LLM residual effects separated by recommendation and headline generation among pre-test-calibration-admitted runs. These are within-setting matched effects, not frozen cross-model transfer. The negative RT AUPRC effect is concentrated in open-weight LLM headline generation.

Task	Residual	Runs	Models	ΔAUPRC $\times 10^3$	ΔRecall $\times 10^3$	ΔLead	AUPRC win	Balanced	Strict triple
Headline Gen.	R	19	3	+0.103	+2.739	+0.058	42.1%	42.1%	5.3%
Headline Gen.	RT	20	3	-3.160	+5.294	+0.121	40.0%	35.0%	0.0%
Headline Gen.	T	20	3	+0.358	+0.907	+0.158	40.0%	40.0%	0.0%
News Rec.	R	23	3	+2.380	-2.083	-0.015	69.6%	56.5%	0.0%
News Rec.	RT	6	3	+0.388	0.000	0.000	66.7%	66.7%	0.0%
News Rec.	T	12	3	+0.158	0.000	0.000	58.3%	58.3%	0.0%

Table 52: Consumer-family structured comparison.

Consumer	View	ΔAUPRC $\times 10^3$	Median ΔAUPRC	AUPRC win	ΔRecall	Lead	FAR	Balanced	
Generation decoder	R	+0.815	+0.771	93.8%	+1.107	-0.500	+0.302	62.5%	56.2%
Generation decoder	T	+1.376	+0.112	100.0%	+2.052	-0.158	+0.019	63.2%	42.1%
Generation decoder	RT	+2.712	+1.995	94.4%	+7.294	+0.190	+0.006	55.6%	44.4%
OpenLLM	R	+1.350	+0.068	57.1%	+0.098	+0.018	+0.573	50.0%	2.4%
OpenLLM	T	+0.283	-0.014	46.9%	+0.567	+0.098	+0.361	46.9%	0.0%
OpenLLM	RT	-2.341	-0.055	46.2%	+4.073	+0.093	-0.718	42.3%	0.0%

Table 53: **Deployment cost of CoTRAC.** Averages across 48 measured configurations on MovieLens and PENS at batch size one. CoTRAC adds 0.79 ms per request and 0.139M parameters on average. Mean configuration-wise overheads are 37.85% for latency and 1.19% for parameters.

Metric	Preference model	Model + CoTRAC	Added cost
Inference latency (ms/request)	21.93	22.72	+0.79
Throughput (requests/s)	506.60	398.23	−108.37
Parameters (M)	13.84	13.98	+0.139
FP32 parameter size (MiB)	52.79	53.32	+0.53