FROM TEMPORAL REPRESENTATION TO POINT PREDICTION: AUDITING SENSITIVITY TO EVENT–TIME REASSIGNMENT ACROSS PREDICTION STAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal point process (TPP) models are evaluated through predictive distributions, decoded point forecasts, and information accessible from hidden representations, but sensitivity at these stages need not agree. We introduce a same-target audit that reassigns historical inter-event times while preserving event identities and order, the interval multiset, total duration, and the future target. The audit separates raw-history predictability, representation accessibility, native distributional sensitivity, prediction movement, and decoded point-error response. Across five TPP architectures on three cascade datasets, SAHP and AttNHP show strong representation-access and positive native-likelihood responses, yet their operational point-MAE effects are weak or negative. Matched fixed-rate re-decoding shows that the endpoint depends strongly on the selected point functional: on Retweet, conditional-median and conditional-mean readouts restore positive responses, whereas on Cyberbullying and YouTube the conditional median remains negative while the conditional mean is positive. In contrast, S2P2 retains positive point-error responses under both its operational conditional-mean readout and a conditional-median control across all three datasets. These results show that temporal sensitivity can persist in learned representations and predictive distributions while being attenuated, reversed, or preserved at the final point readout. The effect of the point functional is itself model- and dataset-dependent, so temporal reliance should be audited across prediction stages rather than inferred from any single representation, likelihood, or point-error measure.

1 INTRODUCTION

Time is a defining component of event-sequence modeling. Neural temporal point processes (TPPs) explicitly encode timestamps or inter-event intervals (Du et al., 2016; Mei & Eisner, 2017; Zuo et al., 2020; Zhang et al., 2020; Mei et al., 2022). They are commonly evaluated using next-event likelihood, time-prediction error, and mark accuracy (Xue et al., 2024; Bosser & Ben Taieb, 2023), while learned representations can be examined with auxiliary probes (Hewitt & Liang, 2019; Elazar et al., 2021). However, good predictive performance or high temporal decodability does not establish how sensitivity to historical event–time assignment appears across a model's prediction pathway (Ravichander et al., 2021; Elazar et al., 2021).

We study this question as a four-stage audit spanning raw-history predictability, representation access, native distributional response, and decoded point-error response. These quantities are not equivalent: a prediction can move under intervention without increasing aggregate MAE. Our event–time reassignment preserves event identities and order, sequence length, the historical interval multiset, total duration, and the future target, and is evaluated through both a standardized same-target access head and the model's native prediction pathway.

On the Retweet benchmark (Xue et al., 2024), SAHP (Zhang et al., 2020) and AttNHP (Mei et al., 2022) show strong target-accessible and native-likelihood responses, while their original operational point decoders produce little or negative change in point-prediction mean absolute error (MAE). Their predictions nevertheless move under reassignment. THP (Zuo et al., 2020) and NHP (Mei & Eisner, 2017) retain positive decoded-MAE effects. We additionally evaluate S2P2 (Chang et al.,

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.

2025), a recent state-space TPP, and find positive representation-access, native-likelihood, and operational point-error responses on Retweet, Cyberbullying, and YouTube. Thus, the SAHP/AttNHP operational endpoint attenuation replicates across the cascade datasets but is not shared by this contemporary state-space architecture. Our MIMIC-IV clinical-event construction (Johnson et al., 2023; 2024) gives positive SAHP and AttNHP effects under the same operational protocol; this single construction does not establish a model-intrinsic domain effect.

To investigate the Retweet contrast, we cross THP and AttNHP backbones with Gaussian and Exponential time-prediction configurations. This factorial jointly changes output parameterization, objective, and point decoder, and switches the point-MAE endpoint while representation access remains comparatively stable. We then compare point functionals on matched canonical-recipe SAHP and AttNHP checkpoints, holding each seed's fitted representation and rates fixed. The results show that the final point-error response depends on the selected functional; cross-dataset estimates are reported in Sec. 5.3. A twelve-benchmark raw-history screen identifies settings with sufficient upstream temporal opportunity for the audit.

Contributions. Our contributions are threefold: (i) a same-target intervention audit that separates raw-history predictability, representation accessibility, native distributional sensitivity, prediction movement, and decoded point-error response under controlled event–time reassignment; (ii) a five-architecture, three-dataset evaluation showing a stage-wise dissociation in which temporal sensitivity can remain strong in learned representations and native predictive distributions while being strongly attenuated or reversed at the final point readout; and (iii) matched fixed-rate re-decoding and complementary decoder controls showing that readout dependence is itself model- and dataset-dependent: alternative point functionals reverse the SAHP/AttNHP endpoint on Retweet but do not fully resolve the Cyberbullying/YouTube pattern, whereas S2P2 remains positively sensitive under both its operational mean and conditional-median readouts.

2 Related Work

Neural temporal point processes. Neural TPPs learn representations of event history and conditional event-time distributions. Recurrent RMTPP and continuous-time NHP encode history differently (Du et al., 2016; Mei & Eisner, 2017); attention-based THP, SAHP, and AttNHP offer other encoders and intensity parameterizations (Zuo et al., 2020; Zhang et al., 2020; Mei et al., 2022). More recent state-space TPPs further broaden the architecture space; S2P2 (Chang et al., 2025) models marked event sequences with a continuous-time state-space formulation. We include S2P2 as a contemporary architecture reference in the same cascade audit. Intensity-free and flow-based models further broaden the choice of distributional form (Shchur et al., 2020a;b). Across these designs, prediction follows a pathway from historical representation to conditional distribution to a decoded point forecast. In our evaluated models, the native heads and decoders also differ (Sec. 4.1).

EasyTPP and comparative TPP studies standardize prediction benchmarks (Xue et al., 2024; Bosser & Ben Taieb, 2023). Forecast-evaluation theory distinguishes distributional scores from point-error scores (Gneiting & Raftery, 2007; Gneiting, 2011), including in point-process settings (Brehmer et al., 2024). These works motivate measuring different outputs, while our question is how one specified reassignment affects representation access, native likelihood, and decoded prediction error.

Representation probing and functional use. Probes test which attributes can be recovered from a learned representation using a restricted auxiliary predictor (Hewitt & Liang, 2019). Decodability, however, does not establish that the native model uses the decoded information (Ravichander et al., 2021; Elazar et al., 2021). This distinction matters for temporal prediction: a representation can support next-time prediction even when the same intervention has a different effect on native likelihood or decoded prediction error. Our access head therefore predicts the *same next-time target* as the native model, in the same error space, under the same paired intervention. Comparing S_{access} with native sensitivities compares the error response of an auxiliary readout with the model's own output responses; it does not measure generic decodability of temporal attributes.

Intervention-based and counterfactual analysis. Counterfactual TPP methods study event dynamics under alternative process intensities (Noorbakhsh & Gomez Rodriguez, 2022). More directly

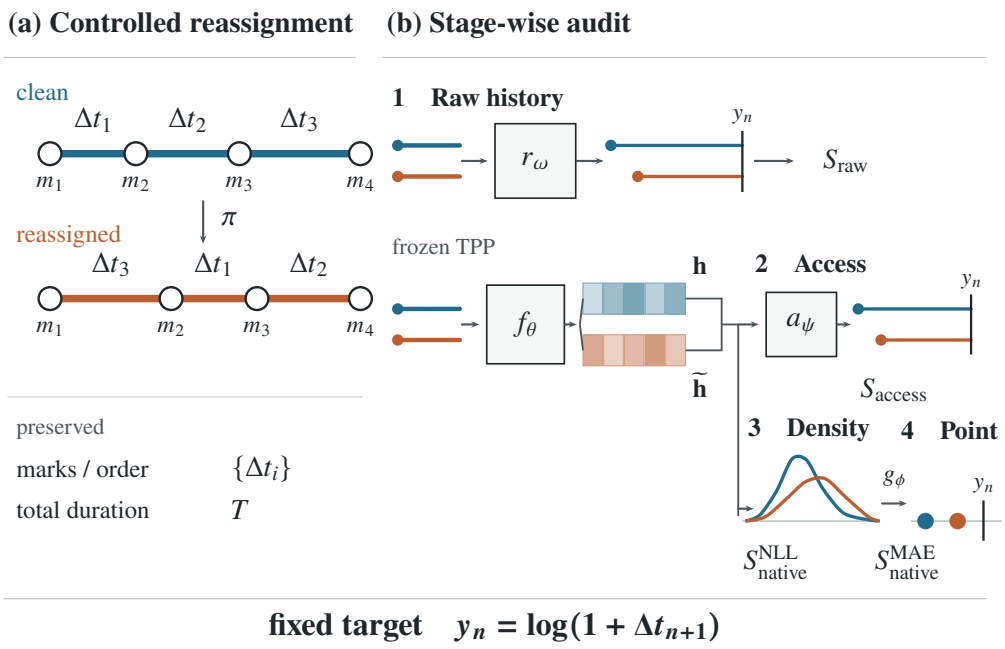

Figure 1: **Controlled event–time reassignment and stage-wise audit.** (a) Historical intervals are reassigned among event positions while preserving event identities and order, the interval multiset, total duration, and the future target. (b) A separate raw-history screen assesses upstream temporal opportunity. The paired histories then pass through the same frozen TPP encoder. A same-target access head reads out the representations; the native predictive distribution is assessed by NLL and its model-specific decoded point prediction by MAE. The colored feature, residual, distribution, and point glyphs are schematic and do not encode measured effect sizes.

related, factual–counterfactual explanations of marked TPP predictions identify small subsets of historical events that preserve predictive information (Liu et al., 2025). Other TPP work learns explicit temporal rules to explain events (Yang et al., 2024); permutation-based model reliance evaluates a model after input features are reassigned (Fisher et al., 2019). Time-aware attribution methods such as TIMING construct temporally informed attribution paths and highlight the risk of implausible temporal perturbations (Jang et al., 2025). TIMING attributes time-series predictions to input points, whereas our audit measures how a fixed-target interval reassignment changes successive stages of a marked TPP prediction. It characterizes responses to this specified perturbation rather than event-level causal attribution.

3 Auditing Responses to Event–Time Reassignment

Figure 1 summarizes the audit: the same paired clean and reassigned histories are traced through each stage of one trained prediction pathway while the future target is held fixed.

For history $\mathcal{H}_n = \{(m_1, t_1), \ldots, (m_n, t_n)\}$ with $\Delta t_i = t_i - t_{i-1}$, we study next-time prediction with target

$$y_{n+1} = \log(1 + \Delta t_{n+1}).$$

Event–time reassignment. For historical intervals $\boldsymbol{\Delta} = (\Delta t_1, \ldots, \Delta t_n)$ and a permutation π, we set $\widetilde{\boldsymbol{\Delta}} = (\Delta t_{\pi(1)}, \ldots, \Delta t_{\pi(n)})$ and reconstruct timestamps $\widetilde{t}_i = t_0 + \sum_{j=1}^{i} \widetilde{\Delta t}_j$. This preserves sequence length, event identities and order, the multiset of historical intervals, and total elapsed duration, while changing which interval is associated with each event position and hence the intermediate cumulative-time geometry. The future interval Δt_{n+1} is never permuted, so clean and intervened histories share exactly the same target; we therefore interpret the experiment as an intervention on historical event–time assignment rather than identification of a unique causal temporal

variable. For graded interventions, intervals are permuted only within contiguous blocks, with global reassignment as the limiting case (Appendix A.6).

For any point predictor q, intervention error sensitivity is

$$S(q) = \mathrm{MAE}_{\mathrm{int}}(q) - \mathrm{MAE}_{\mathrm{clean}}(q).$$

Positive S indicates greater error under reassignment; near-zero S means little net aggregate MAE change and does not imply unchanged predictions. Because reassignment changes the joint input distribution, we interpret S as an error response to this specified intervention.

Why compare access, likelihood, and point-prediction error? A large S_{access} indicates that the frozen representation supports prediction of the fixed target in a manner whose error is affected by reassignment. A positive $S^{\mathrm{NLL}}_{\mathrm{native}}$ indicates that the native predictive distribution assigns lower likelihood to the fixed target after the same intervention. Finally, $S^{\mathrm{MAE}}_{\mathrm{point}}(g)$ measures the change in error of a point prediction produced by functional g of the fitted distribution. These quantities measure distinct responses: access, native likelihood, and point-prediction error under the same intervention.

To measure representation accessibility, we freeze each trained encoder and fit the same 65-parameter linear predictor to the same next-time target:

$$S_{\mathrm{access}} = \mathrm{MAE}^{\mathrm{access}}_{\mathrm{int}} - \mathrm{MAE}^{\mathrm{access}}_{\mathrm{clean}}.$$

Same-target access control. The access predictor is deliberately lightweight and standardized: 65 trainable parameters, trained on clean frozen representations toward the same next-time target as the native model, then applied to paired clean and reassigned representations at test time. It therefore does not ask whether an arbitrary temporal attribute is decodable, but whether the representation exposes information useful for the model's actual next-time target under the intervention used to evaluate the native pathway. For a point functional g applied to the fitted native distribution, decoded point-prediction MAE sensitivity is

$$S^{\mathrm{MAE}}_{\mathrm{point}}(g) = \mathrm{MAE}_{\mathrm{int}}(g) - \mathrm{MAE}_{\mathrm{clean}}(g),$$

where both errors use the same fixed targets y_j. The original tables retain $S^{\mathrm{MAE}}_{\mathrm{native}}$ as shorthand for $S^{\mathrm{MAE}}_{\mathrm{point}}(g_0)$, with g_0 the model-specific operational decoder defined in Sec. 4.1. Their access–native gaps are likewise conditional on g_0. To distinguish net error change from output movement, we define paired prediction drift over N evaluated positions as

$$D_{\mathrm{point}}(g) = \frac{1}{N}\sum_{j=1}^{N} \left| g(\theta_{j,\mathrm{int}}) - g(\theta_{j,\mathrm{clean}}) \right|,$$

where θ denotes native distribution parameters. The existing D_{native} values equal $D_{\mathrm{point}}(g_0)$. Native distributional sensitivity is

$$S^{\mathrm{NLL}}_{\mathrm{native}} = \mathrm{NLL}^{\mathrm{native}}_{\mathrm{int}} - \mathrm{NLL}^{\mathrm{native}}_{\mathrm{clean}}. \tag{1}$$

Because the models use different native likelihood formulations, $S^{\mathrm{NLL}}_{\mathrm{native}}$ is interpreted only within model and is not used for cross-family magnitude ranking.

The twelve-benchmark raw-history screen is used as an exploratory upstream-opportunity analysis and for dataset selection; Cyberbullying and YouTube are separate within-family replication datasets.

4 Experimental Setup

We evaluate THP (Zuo et al., 2020), SAHP (Zhang et al., 2020), AttNHP (Mei et al., 2022), NHP (Mei & Eisner, 2017), and S2P2 (Chang et al., 2025) on three retweet-cascade datasets: the EasyTPP Retweet benchmark (Zhao et al., 2015; Xue et al., 2024) and the Cyberbullying and YouTube domains from ICWSM'20 (Wu et al., 2020). All model–dataset combinations use the same next-time target, event–time reassignment, representation-access protocol, training budget, and five-seed evaluation. Exact dataset construction, splits, and preprocessing are reported in Appendix A.9.

Replication protocol. Retweet is the primary setting for the detailed output-space and downstream controls; Cyberbullying and YouTube are targeted external replications in the same cascade process family. We retain cascades with at least nine retweets and reuse the same three follower-count mark categories as Retweet, avoiding dataset-specific redefinition of the mark space.

4.1 Prediction Pathways and Point Functionals

All point errors use $y_{i+1} = \log(1 + \Delta t_{i+1})$. THP reports the mean, also the median, of its Gaussian output over y. For SAHP and AttNHP, $\lambda_i^{(y)} > 0$ is the Exponential rate over y_{i+1}. Their time loss and reported operational point decoder are

$$\mathcal{L}_{\mathrm{Exp},i} = -\log \lambda_i^{(y)} + \lambda_i^{(y)} y_{i+1}, \qquad \hat{y}_{i+1}^{\mathrm{op}} = \log\left(1 + \frac{1}{\lambda_i^{(y)}}\right).$$

We re-decode the same trained rates without changing the encoder or fitted distribution, using

$$g_{\mathrm{op}}(\lambda) = \log(1 + 1/\lambda), \qquad g_{\mathrm{med}}(\lambda) = \frac{\log 2}{\lambda}, \qquad g_{\mathrm{mean}}(\lambda) = \frac{1}{\lambda}.$$

The operational readout g_{op} differs from the conditional median and mean of the fitted Exponential distribution, g_{med} and g_{mean}, respectively. The raw-time control uses the operational readout's algebraic inverse $1/\lambda_i^{(y)}$, without treating this as a raw-time Exponential rate (Appendix A.6).

For NHP, $\bar{F}_i(u)$ denotes the prospective waiting-time survival function from the causal post-event state. The reported decoder is $\hat{y}_{i+1}^{\mathrm{op}} = \log(1 + \int_0^\infty \bar{F}_i(u)\, du)$, a transformed conditional *mean* waiting time. The corresponding prospective conditional median decoder is $\log(1 + m_i)$ with $\bar{F}_i(m_i) = 1/2$; it was not evaluated. Conditional- median readouts in y-space for SAHP and AttNHP are evaluated on all three cascade datasets. Cross-model MAE comparisons are conditional on the stated decoders. Prospective quadrature, leakage checks, and intensity definitions are in Appendix A.12. For S2P2, the native output is a mark-specific conditional intensity over the rescaled raw inter-event-time axis rather than a scalar time head. Following the EasyTPP prediction pathway, its operational point prediction is the thinning Monte-Carlo conditional mean waiting time, rescaled to raw time and then mapped to the evaluation space by $\log(1 + \Delta t)$. Because the intensity defines a full conditional waiting-time distribution, we additionally examine the conditional median from the same accepted thinning samples. Its causal post-event representation is evaluated with the same standardized same-target access protocol. Access heads are identical across models and trained only on clean frozen representations; training details are in Appendix A.8.

Clinical cross-domain evaluation. For cross-domain evaluation, we retain THP, SAHP, AttNHP, and NHP on MIMIC-IV clinical event sequences (Johnson et al., 2023; 2024): 92,315 sequences covering the first 24 hours of each ICU stay, with clinical events bundled at one-minute resolution into a 31-mark vocabulary under an 80/10/10 patient-disjoint split. Construction details are in Appendix A.3. Event-based and point-process models have previously addressed irregular clinical records and EHR representations (Enguehard et al., 2020; Karami et al., 2024; Xiao et al., 2025). Here MIMIC provides a cross-domain check under the same operational evaluation protocol. Decoder-specific re-evaluation is not available for this construction.

Evaluation protocol. All clean–intervention comparisons are paired: the same target event is evaluated before and after reassignment under a fixed intervention manifest. The matched decoder audit uses a separately trained five-seed canonical-recipe checkpoint set; its estimates are therefore not paired with the headline audit across tables. Within each matched checkpoint, all alternative point functionals are evaluated on the same fitted rates (Appendix A.2). Auxiliary controls use their experiment-specific paired procedures. Native NLL is interpreted only within model because the architectures use different likelihoods.

5 RESULTS

5.1 TEMPORAL SENSITIVITY CAN DISSOCIATE ACROSS PREDICTION STAGES

We first ask whether sensitivity to historical event–time assignment is preserved as information passes from the learned representation to the native predictive distribution and finally to a decoded point forecast. The answer is model-dependent. SAHP and AttNHP remain strongly sensitive at the representation level and show positive native-likelihood responses, yet this sensitivity nearly disappears under their operational point readouts. THP, NHP, and S2P2 do not show the same collapse. Table 1 quantifies this stage-wise dissociation on Retweet.

Table 1: **Retweet intervention audit under the operational point readouts.** Five-seed means with 95% Student-t intervals. S_{access} measures same-target representation sensitivity, $S^{\text{NLL}}_{\text{native}}$ measures within-model native distributional sensitivity, and $S^{\text{MAE}}_{\text{point}}(g_0)$ measures decoded point-error sensitivity under the model-specific operational readout. Native NLL magnitudes are interpreted only within model and are not used for cross-family ranking.

Model	Native clean MAE	S_{access}	$S^{\text{NLL}}_{\text{native}}$	$S^{\text{MAE}}_{\text{point}}(g_0)$
THP	1.041 [1.037,1.046]	0.634 [0.614,0.655]	0.830 [0.786,0.873]	0.847 [0.814,0.880]
NHP	1.102 [1.092,1.112]	1.066 [1.034,1.098]	1.329 [1.255,1.403]	0.766 [0.747,0.786]
SAHP	2.473 [2.457,2.489]	0.910 [0.873,0.948]	0.134 [0.128,0.140]	-0.010 [-0.020,0.001]
AttNHP	2.478 [2.464,2.492]	0.828 [0.798,0.857]	0.128 [0.125,0.131]	-0.017 [-0.022,-0.013]
S2P2	1.4876 [1.4680,1.5073]	1.0342 [0.9791,1.0893]	1.1920 [1.0931,1.2908]	1.1170 [1.0344,1.1995]

The stage-wise contrast is sharpest for SAHP and AttNHP: their access responses remain large ($0.910/0.828$) and their native-NLL responses are positive ($0.134/0.128$), yet their operational point-MAE effects collapse to $-0.010/-0.017$. THP and NHP instead retain positive point-error effects ($0.847/0.766$), while S2P2 is positive at all three audited stages. S2P2 is therefore an architecture reference showing that the attenuation is not universal, rather than a controlled state-space ablation; its output parameterization and native prediction pathway differ from SAHP/AttNHP.

5.2 THE POINT READOUT EXPLAINS MUCH OF THE RETWEET ATTENUATION

The preceding dissociation raises a natural question: does the weak SAHP/AttNHP endpoint reflect a loss of temporal information, or the way their fitted distributions are converted into point predictions? We test this without retraining the representation or rate predictor by re-decoding the same fitted rates with alternative point functionals.

The result is striking on Retweet. For both SAHP and AttNHP, the signed response increases from approximately zero under the operational readout to a clearly positive response under the conditional median and to a much larger response under the conditional mean (Table 2; Figure 2).

Table 2: **Retweet matched canonical-recipe point-functional audit.** All errors use $y = \log(1 + \Delta t)$. Each seed's matched rates are re-evaluated with all three point functionals. Brackets are 95% Student-t intervals across seeds for the signed point-MAE effect; the THP row is protocol-matched to this audit and distinct from the original THP row in Table 1.

Model	Point functional g	Clean MAE	$S^{\text{MAE}}_{\text{point}}(g)$ [95% CI]	Std. ratio	Pearson r
THP-Gauss	μ	1.0394	$+0.8549$ $[+0.8074, +0.9024]$	0.8622	0.8398
SAHP	g_{op}	2.4733	-0.0098 $[-0.0203, +0.0008]$	0.1829	0.8159
SAHP	g_{med}	1.4782	$+0.3473$ $[+0.2954, +0.3992]$	0.5975	0.8395
SAHP	g_{mean}	1.0418	$+0.8566$ $[+0.7948, +0.9185]$	0.8620	0.8395
AttNHP	g_{op}	2.4779	-0.0173 $[-0.0219, -0.0126]$	0.1819	0.7994
AttNHP	g_{med}	1.5044	$+0.3544$ $[+0.3217, +0.3870]$	0.5890	0.8211
AttNHP	g_{mean}	1.0926	$+0.8603$ $[+0.8232, +0.8975]$	0.8497	0.8211

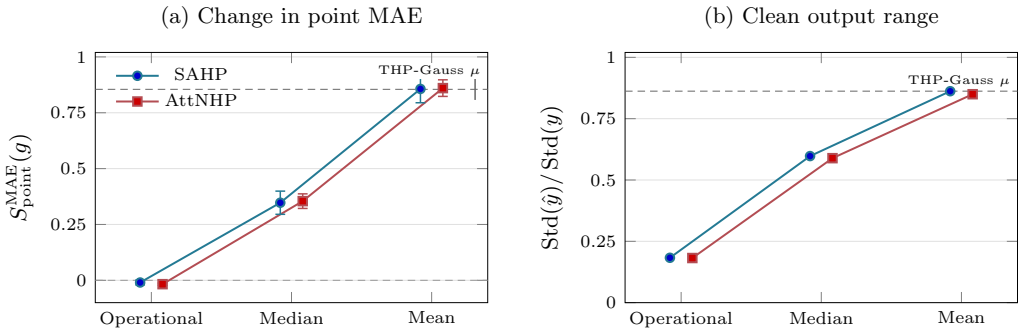

Figure 2: **Retweet matched point-functional audit.** (a) Signed point-MAE effects with five-seed 95% Student-t intervals; the dashed THP-Gauss μ reference is $+0.8549$ $[+0.8074, +0.9024]$. (b) Clean prediction-to-target standard-deviation ratios; no uncertainty intervals were available for these ratios. Within each SAHP or AttNHP checkpoint, representation, trained rate predictor, event positions, and intervention are fixed; only the point functional changes. The dashed THP-Gauss ratio is 0.8622.

The corresponding prediction-range ratios rise from about 0.18 under the operational readout to 0.59 under the median and 0.85 under the mean, showing that the readout changes both the scale of prediction movement and its aggregate MAE consequence (Appendix A.2).

5.3 Readout dependence is model- and dataset-dependent

The Retweet sign reversal does not generalize as a universal decoder effect. On Cyberbullying and YouTube, the conditional mean restores a positive SAHP/AttNHP response, but the conditional median remains negative. Thus, changing the point functional matters, but it does not by itself explain the cross-dataset attenuation.

The model dependence is equally important. S2P2 remains positive at the representation, likelihood, and operational point-error stages on all three cascade datasets, and its conditional-median control is also positive throughout. NHP likewise retains positive responses under two prospective decoding schemes (Appendix Table 23). Together, these results show that the consequence of point decoding depends on both the fitted model and the dataset rather than following a universal ordering.

Table 3: **S2P2 stage-wise audit across the three cascade datasets.** Five-seed means with 95% Student-t intervals for the three intervention effects. Std. ratio denotes the decoded prediction-to-target standard-deviation ratio.

Dataset	S_{access}	$S^{\mathrm{NLL}}_{\mathrm{native}}$	$S^{\mathrm{MAE}}_{\mathrm{point}}$	Std. ratio
Retweet	1.0342 [0.9791,1.0893]	1.1920 [1.0931,1.2908]	1.1170 [1.0344,1.1995]	0.780
Cyberbullying	0.2572 [0.1732,0.3412]	0.2269 [0.2052,0.2486]	0.3995 [0.3575,0.4415]	0.804
YouTube	0.3863 [0.3401,0.4325]	0.3134 [0.2728,0.3540]	0.5030 [0.4654,0.5406]	0.819

Across all three cascades, S2P2 remains positive at access, likelihood, and operational point-error stages while maintaining broad decoded output ranges (std.-ratio 0.780–0.819). As an additional readout control, the conditional median of the same sampled waiting-time distribution also retains positive signed point-MAE effects on Retweet, Cyberbullying, and YouTube ($+0.724$, $+0.219$, and $+0.284$, respectively). Thus, the positive S2P2 endpoint is not specific to its operational mean readout.

Figure 3 confirms that readout dependence is not a universal sign reversal: the conditional mean is positive for SAHP/AttNHP on all three datasets, whereas the conditional median remains negative on Cyberbullying and YouTube. Together with the positive S2P2 and NHP controls, this shows that the point-functional effect depends on both model and dataset.

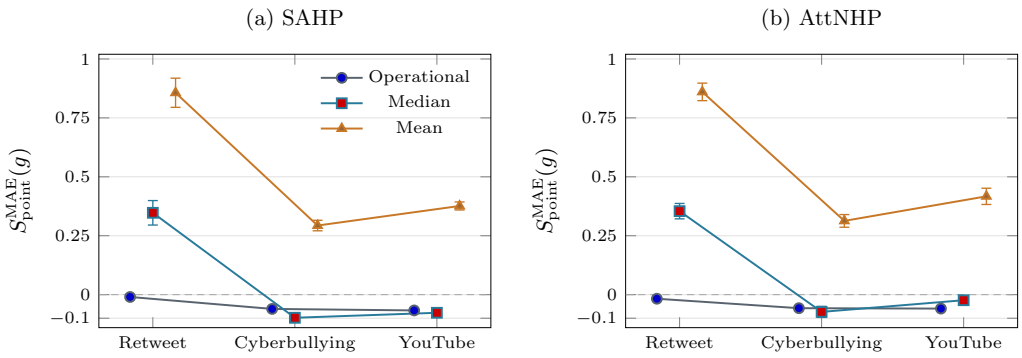

Figure 3: **Decoder-family dependence across the cascade datasets.** Each panel shows five-seed signed point-MAE effects with 95% Student-t intervals for one model. Retweet uses the matched results in Table 2; Cyberbullying and YouTube use the matched decoder-family estimates in Appendix Table 5.

5.4 A CONSTRUCTION-SPECIFIC CLINICAL CROSS-DOMAIN CHECK

Under the original operational protocol, the 31-mark MIMIC-IV construction has positive raw-history sensitivity and positive access, native-likelihood, and decoded-MAE responses for SAHP and AttNHP (Appendix Table 10). Its NHP predictor performs worse than the constant-median baseline and has a near-zero decoded-MAE effect; we therefore interpret that result cautiously. Since alternative decoders were not evaluated, MIMIC is a construction-specific cross-domain check, not evidence of a decoder-independent domain effect.

5.5 CONTROLS SUPPORT THE STAGE-WISE INTERPRETATION

Under graded reassignment, THP develops sensitivity in both pathways, whereas SAHP and AttNHP gain access sensitivity but retain small operational MAE effects (Figure 4); this pattern is specific to the operational readout.

End-to-end native-pathway factorial. To test the prediction pathway during representation learning, we train the four THP/AttNHP by Gaussian/Exponential combinations under matched splits, budgets, interventions, and access protocols (Appendix Table 11). On Retweet, switching THP from Gaussian to Exponential changes the point-MAE effect by $-0.860\,[-0.911, -0.809]$; switching AttNHP to Gaussian changes it by $+0.916\,[+0.893, +0.939]$. The interaction is small ($+0.056\,[-0.005, +0.117]$), while access remains similar within backbone. This associates the endpoint contrast with the joint native pathway, but the factorial cannot separate its parameterization, objective, and decoder. Re-decoding its fitted AttNHP Exponential checkpoint changes the operational effect from -0.025 to $+0.883$, near the separately trained Gaussian effect ($+0.891$), localizing a substantial contribution to the point readout.

Complementary frozen-representation and supervision controls. Frozen AttNHP representations support sensitive predictions under Gaussian and Exponential heads ($S_{\mathrm{MAE}} = 0.778$ and 1.125). Clean-output distillation recovers a larger response ($0.658\,[0.641, 0.675]$); the compact control is reported in Appendix A.5.

Evaluation robustness. The operational pattern persists across probe, metric-space, normalization, and training controls. Under same-mark reassignment, SAHP/AttNHP remain negative operationally ($-0.0114/-0.0175$) but positive under the conditional median ($+0.3249/+0.3373$) and conditional mean ($+0.8163/+0.8255$), retaining about 94–96% of the corresponding global effects. Per-example error changes nearly cancel despite nonzero prediction movement (Appendix Table 14).

Audit scope. The twelve-benchmark screen finds large raw-history sensitivity on Retweet ($S_{\mathrm{raw}} = 0.715$), a smaller effect on Volcano (0.139), and $|S_{\mathrm{raw}}| \leq 0.012$ on the remaining ten. Volcano also lacks reliable representation access despite positive raw sensitivity: upstream opportunity does not guarantee preservation (Appendix A.11).

6 DISCUSSION

Event–time reassignment affects learned representations and native predictive distributions, while its point-error consequence depends on the selected point functional, prediction pathway, and dataset. The operational SAHP/AttNHP attenuation replicates across all three cascade datasets but is not shared by THP, NHP, or S2P2. On Retweet, changing only the point functional of fixed fitted rates reverses the signed MAE response, and the same ordering persists under same-mark reassignment. Across Cyberbullying and YouTube, however, the conditional median remains negative while the conditional mean is positive. Decoder choice therefore explains a substantial part of the Retweet endpoint contrast but does not by itself account for the cross-dataset pattern; model calibration and dataset-dependent distributional differences remain plausible contributors. S2P2 further shows that this functional dependence is not universal: both its operational mean and conditional-median readouts remain positive across the three cascades. Per-example analysis further shows that weak operational aggregate responses can reflect cancellation of positive and negative error changes rather than prediction invariance. The factorial independently associates the endpoint contrast with the joint native pathway. MIMIC retains its separate operational-protocol interpretation.

Scope and limitations. The results concern this potentially out-of-distribution reassignment audit (Jang et al., 2025) and the stated point readouts, not general or causal temporal dependence. Re-decoding with the conditional median in y-space is reported for SAHP and AttNHP on the cascade datasets. The prospective NHP median could not be evaluated because the retained artifacts do not contain the fitted model states required to reconstruct the conditional survival function. Epoch-wise validation curves are not reported. We do not separately identify whether the remaining cross-dataset differences in conditional median responses arise from conditional-density misspecification, calibration, or dataset-specific dynamics.

7 CONCLUSION

Our audit separates responses to event–time reassignment across raw-history prediction, representation access, native predictive distributions, prediction movement, and point-prediction error. Across three cascade datasets, SAHP and AttNHP retain representation and distributional sensitivity while showing weak or negative responses under their operational point readouts, whereas THP, NHP, and S2P2 retain positive operational responses. Matched fixed-rate re-decoding shows that the selected point functional can substantially alter the signed endpoint response and reverse it on Retweet, while the persistence of negative conditional-median effects on Cyberbullying and YouTube shows that decoder choice is not a complete cross-dataset explanation. S2P2 further remains positive under both its operational mean and conditional-median readouts. Thus, point-functional dependence is itself model- and dataset-dependent, and representation access, distributional response, prediction movement, and point-error response should be distinguished when auditing temporal reliance.

AI USE STATEMENT

The experimental design was developed by the authors. We used generative AI tools to assist with drafting and revising the manuscript, generating and editing code that implements the author-designed experiments, and analyzing and interpreting experimental results. We also used these tools to identify candidate literature and dataset-source references, check bibliographic details, and integrate and format citations and the reference list; OpenAI Codex was used for this citation work. The authors take responsibility for the final manuscript, code, analyses, results, and references, including material developed with AI assistance.

REFERENCES

Alibaba Cloud Tianchi. Taobao User Behavior Dataset. Online dataset, n.d. URL https://tianchi.aliyun.com/dataset/649. Accessed 2026-09-24.

Wonho Bae, Mohamed Osama Ahmed, Frederick Tung, and Gabriel L. Oliveira. Meta temporal point processes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=QZfdDpTX1uM.

Tanguy Bosser and Souhaib Ben Taieb. On the predictive accuracy of neural temporal point process models for continuous-time event data. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=3OSISBQPrM.

Jonas R. Brehmer, Tilmann Gneiting, Marcus Herrmann, Warner Marzocchi, Martin Schlather, and Kirstin Strokorb. Comparative evaluation of point process forecasts. *Annals of the Institute of Statistical Mathematics*, 76:47–71, 2024. doi: 10.1007/s10463-023-00875-5. URL https://doi.org/10.1007/s10463-023-00875-5.

Yuxin Chang, Alex James Boyd, Cao Xiao, Taha Kass-Hout, Parminder Bhatia, Padhraic Smyth, and Andrew Warrington. Deep continuous-time state-space models for marked event sequences. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/ee39348acc798915d2d15a8bbbd417b8-Abstract-Conference.html.

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, 2016. doi: 10.1145/2939672.2939875. URL https://doi.org/10.1145/2939672.2939875.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Joseph Enguehard, Dan Busbridge, Adam Bozson, Claire Woodcock, and Nils Hammerla. Neural temporal point processes for modelling electronic health records. In *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 136 of *Proceedings of Machine Learning Research*, pp. 85–113, 2020. URL https://proceedings.mlr.press/v136/enguehard20a.html.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019. URL https://jmlr.org/papers/v20/18-760.html.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011. doi: 10.1198/jasa.2011.r10138. URL https://doi.org/10.1198/jasa.2011.r10138.

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. URL https://doi.org/10.1198/016214506000001437.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Hyeongwon Jang, Changhun Kim, and Eunho Yang. TIMING: Temporality-aware integrated gradients for time series explanation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26877–26895, 2025. URL https://proceedings.mlr.press/v267/jang25a.html.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV, version 3.1. PhysioNet, 2024. URL https://physionet.org/content/mimiciv/3.1/.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x. URL https://www.nature.com/articles/s41597-022-01899-x.

Hojjat Karami, David Atienza, and Anisoara Ionescu. TEE4EHR: Transformer event encoder for better representation learning in electronic health records. *Artificial Intelligence in Medicine*, 154: 102903, 2024. doi: 10.1016/j.artmed.2024.102903. URL https://doi.org/10.1016/j.artmed.2024.102903.

Michael Kechinov. eCommerce Events History in Cosmetics Shop. Kaggle dataset, sourced from REES46 Marketing Platform, 2019. URL https://www.kaggle.com/datasets/mkechinov/ecommerce-events-history-in-cosmetics-shop. Data collected by the Open CDP project and REES46 Marketing Platform.

Sishun Liu, Ke Deng, Xiuzhen Zhang, and Yan Wang. Learning marked temporal point process explanations based on counterfactual and factual reasoning. In *Proceedings of the 28th European Conference on Artificial Intelligence*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 1841–1848, 2025. doi: 10.3233/FAIA251016. URL https://doi.org/10.3233/FAIA251016.

Hongyuan Mei and Jason Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/6463c88460bd63bbe256e495c63aa40b-Abstract.html.

Hongyuan Mei, Chenghao Yang, and Jason Eisner. Transformer embeddings of irregularly spaced events and their participants. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Rty5g9imm7H.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 188–197, 2019. doi: 10.18653/v1/D19-1018. URL https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/. The dataset host recommends citing this paper for the Amazon Review Data (2018).

Kimia Noorbakhsh and Manuel Gomez Rodriguez. Counterfactual temporal point processes. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24810–24823, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d3faa41886997cfc2128b930077fa49-Abstract.html.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3363–3377, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. Intensity-free learning of temporal point processes. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=HygOjhEYDH.

Oleksandr Shchur, Nicholas Gao, Marin Biloš, and Stephan Günnemann. Fast and flexible temporal point processes with triangular maps. In *Advances in Neural Information Processing Systems*, volume 33, 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/00ac8ed3b4327bdd4ebbebcb2ba10a00-Abstract.html.

Smithsonian Institution Global Volcanism Program. Volcanoes of the world. Online volcano and eruption database, n.d. URL https://volcano.si.edu/. Accessed 2026-09-24.

U.S. Geological Survey. Earthquake data. Earthquake Hazards Program online data portal, n.d. URL https://www.usgs.gov/programs/earthquake-hazards/science/earthquake-data. Accessed 2026-09-24.

Boudewijn F. van Dongen. BPI Challenge 2012. Eindhoven University of Technology, dataset, 2012. URL https://doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f.

Chris Whong. FOILing NYC's Taxi Trip Data. Data release and accompanying article, 2014. URL https://chriswhong.com/open-data/foil_nyc_taxi/.

Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Variation across scales: Measurement fidelity under twitter data sampling. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):715–725, 2020. doi: 10.1609/icwsm.v14i1.7337. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7337.

Tingsong Xiao, Zelin Xu, Wenchong He, Zhengkun Xiao, Yupu Zhang, Zibo Liu, Shigang Chen, My T. Thai, Jiang Bian, Parisa Rashidi, and Zhe Jiang. XTSFormer: Cross-temporal-scale transformer for irregular-time event prediction in clinical applications. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28502–28510, 2025. doi: 10.1609/aaai.v39i27.35073. URL https://ojs.aaai.org/index.php/AAAI/article/view/35073.

Siqiao Xue, Xiaoming Shi, James Y. Zhang, and Hongyuan Mei. HYPRO: A hybridly normalized probabilistic model for long-horizon prediction of event sequences. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/dfbb3d1807b21dadee735eb75069ada4-Abstract-Conference.html.

Siqiao Xue, Xiaoming Shi, Zhixuan Chu, Yan Wang, Hongyan Hao, Fan Zhou, Caigao Jiang, Chen Pan, James Y. Zhang, Qingsong Wen, Jun Zhou, and Hongyuan Mei. EasyTPP: Towards open benchmarking temporal point processes. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e66233a208ef32f56df6312263239fa0-Abstract-Conference.html.

Yang Yang, Chao Yang, Boyang Li, Yinghao Fu, and Shuang Li. Neuro-symbolic temporal point processes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56665–56680, 2024. URL https://proceedings.mlr.press/v235/yang24ag.html.

Qiang Zhang, Aldo Lipani, Omer Kirnap, and Emine Yilmaz. Self-attentive hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11183–11193, 2020. URL https://proceedings.mlr.press/v119/zhang20q.html.

Qingyuan Zhao, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1513–1522, 2015. doi: 10.1145/2783258.2783401. URL https://snap.stanford.edu/seismic/.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11692–11702, 2020. URL https://proceedings.mlr.press/v119/zuo20a.html.

A APPENDIX

A.1 FULL CROSS-DATASET AUDIT

The full seed-level intervals for the cross-dataset operational-readout audit are reported below.

Table 4: **Cross-model and cross-dataset audit under operational point readouts (full seed-level intervals).** All five architectures are evaluated on all three retweet-cascade datasets using the same target and event–time reassignment protocol. S_{access} measures representation-level same-target sensitivity, $S^{\text{MAE}}_{\text{native}}$ measures decoded native point-prediction MAE sensitivity under each model-specific operational readout g_0, and $S^{\text{NLL}}_{\text{native}}$ measures native likelihood-space sensitivity. $S^{\text{NLL}}_{\text{native}}$ is interpreted only as a within-model intervention contrast and is not used for cross-family magnitude comparison. Brackets denote 95% Student-t confidence intervals across five seeds. NHP rows use the prospective conditional-mean decoder (Appendix A.12); NHP and S2P2 std-ratio entries are reported without confidence intervals.

Dataset	Model	S_{access}	$S^{\text{MAE}}_{\text{native}}$	$S^{\text{NLL}}_{\text{native}}$	$\text{Std}(\hat{y})/\text{Std}(y)$
Retweet	THP	0.634 [0.614,0.655]	0.847 [0.814,0.880]	0.830 [0.786,0.873]	0.861 [0.844,0.878]
	NHP	1.066 [1.034,1.098]	0.766 [0.747,0.786]	1.329 [1.255,1.403]	0.722
	SAHP	0.910 [0.873,0.948]	-0.010 [-0.020,0.001]	0.134 [0.128,0.140]	0.183 [0.176,0.190]
	AttNHP	0.828 [0.798,0.857]	-0.017 [-0.022,-0.013]	0.128 [0.125,0.131]	0.182 [0.178,0.186]
	S2P2	1.0342 [0.9791,1.0893]	1.1170 [1.0344,1.1995]	1.1920 [1.0931,1.2908]	0.780
Cyberbullying	THP	0.201 [0.190,0.211]	0.291 [0.276,0.306]	0.226 [0.213,0.239]	0.809 [0.798,0.821]
	NHP	0.251 [0.239,0.264]	0.132 [0.116,0.148]	0.321 [0.315,0.328]	0.606
	SAHP	0.238 [0.224,0.252]	-0.063 [-0.066,-0.060]	0.021 [0.020,0.023]	0.149 [0.147,0.151]
	AttNHP	0.305 [0.289,0.321]	-0.058 [-0.063,-0.052]	0.020 [0.019,0.022]	0.147 [0.143,0.150]
	S2P2	0.2572 [0.1732,0.3412]	0.3995 [0.3575,0.4415]	0.2269 [0.2052,0.2486]	0.804
YouTube	THP	0.311 [0.297,0.325]	0.360 [0.344,0.376]	0.291 [0.281,0.302]	0.845 [0.836,0.853]
	NHP	0.376 [0.356,0.396]	0.173 [0.162,0.183]	0.412 [0.392,0.432]	0.640
	SAHP	0.368 [0.356,0.380]	-0.067 [-0.072,-0.062]	0.028 [0.026,0.030]	0.168 [0.166,0.170]
	AttNHP	0.371 [0.344,0.399]	-0.059 [-0.063,-0.055]	0.029 [0.028,0.030]	0.166 [0.163,0.170]
	S2P2	0.3863 [0.3401,0.4325]	0.5030 [0.4654,0.5406]	0.3134 [0.2728,0.3540]	0.819

Matched decoder-family audit on Cyberbullying and YouTube. The following two datasets use matched canonical-recipe regenerated checkpoints, with each of five seeds evaluated under all three point functionals. Brackets give 95% Student-t intervals across seeds. These regenerated-checkpoint rows are distinct from the canonical operational-readout rows for SAHP and AttNHP in Appendix Table 4.

Table 5: **Matched point-MAE results on Cyberbullying and YouTube.** Entries are five-seed means with 95% Student-t intervals.

Dataset	Model / decoder	Clean MAE	Reassigned MAE	$S^{\text{MAE}}_{\text{point}}$
Cyberbullying	SAHP, operational	$4.3609\,[4.3481, 4.3736]$	$4.3004\,[4.2910, 4.3097]$	$-0.0605\,[-0.0650, -0.0561]$
Cyberbullying	SAHP, median	$2.2057\,[2.1540, 2.2574]$	$2.1077\,[2.0735, 2.1418]$	$-0.0980\,[-0.1180, -0.0781]$
Cyberbullying	SAHP, mean	$1.2905\,[1.2819, 1.2991]$	$1.5838\,[1.5697, 1.5979]$	$+0.2933\,[+0.2710, +0.3156]$
Cyberbullying	AttNHP, operational	$4.3641\,[4.3516, 4.3766]$	$4.3067\,[4.2981, 4.3153]$	$-0.0574\,[-0.0631, -0.0516]$
Cyberbullying	AttNHP, median	$2.2229\,[2.1750, 2.2707]$	$2.1498\,[2.1215, 2.1780]$	$-0.0731\,[-0.0965, -0.0496]$
Cyberbullying	AttNHP, mean	$1.3177\,[1.3095, 1.3258]$	$1.6307\,[1.6098, 1.6516]$	$+0.3130\,[+0.2862, +0.3398]$
YouTube	SAHP, operational	$4.2511\,[4.2456, 4.2566]$	$4.1842\,[4.1786, 4.1898]$	$-0.0669\,[-0.0715, -0.0623]$
YouTube	SAHP, median	$2.1535\,[2.1293, 2.1776]$	$2.0766\,[2.0555, 2.0978]$	$-0.0769\,[-0.0929, -0.0608]$
YouTube	SAHP, mean	$1.2751\,[1.2726, 1.2777]$	$1.6514\,[1.6366, 1.6663]$	$+0.3763\,[+0.3593, +0.3933]$
YouTube	AttNHP, operational	$4.2423\,[4.2218, 4.2629]$	$4.1833\,[4.1624, 4.2041]$	$-0.0591\,[-0.0620, -0.0562]$
YouTube	AttNHP, median	$2.1349\,[2.0509, 2.2188]$	$2.1113\,[2.0441, 2.1784]$	$-0.0236\,[-0.0427, -0.0045]$
YouTube	AttNHP, mean	$1.3097\,[1.2996, 1.3197]$	$1.7268\,[1.7002, 1.7535]$	$+0.4172\,[+0.3827, +0.4516]$

Table 6: **Matched output-range and correlation diagnostics.** Entries are five-seed means with 95% Student-t intervals.

Dataset	Model / decoder	Std. ratio	Pearson r
Cyberbullying	SAHP, operational	$0.1479\,[0.1464, 0.1494]$	$0.7690\,[0.7648, 0.7731]$
Cyberbullying	SAHP, median	$0.5585\,[0.5464, 0.5707]$	$0.7968\,[0.7964, 0.7972]$
Cyberbullying	SAHP, mean	$0.8058\,[0.7882, 0.8233]$	$0.7968\,[0.7964, 0.7972]$
Cyberbullying	AttNHP, operational	$0.1474\,[0.1452, 0.1496]$	$0.7586\,[0.7528, 0.7644]$
Cyberbullying	AttNHP, median	$0.5544\,[0.5397, 0.5692]$	$0.7850\,[0.7841, 0.7859]$
Cyberbullying	AttNHP, mean	$0.7999\,[0.7787, 0.8211]$	$0.7850\,[0.7841, 0.7859]$
YouTube	SAHP, operational	$0.1678\,[0.1659, 0.1696]$	$0.7964\,[0.7921, 0.8008]$
YouTube	SAHP, median	$0.5930\,[0.5803, 0.6058]$	$0.8254\,[0.8251, 0.8258]$
YouTube	SAHP, mean	$0.8556\,[0.8372, 0.8739]$	$0.8254\,[0.8251, 0.8258]$
YouTube	AttNHP, operational	$0.1661\,[0.1623, 0.1700]$	$0.7817\,[0.7780, 0.7853]$
YouTube	AttNHP, median	$0.5848\,[0.5637, 0.6060]$	$0.8124\,[0.8105, 0.8144]$
YouTube	AttNHP, mean	$0.8438\,[0.8132, 0.8743]$	$0.8124\,[0.8105, 0.8144]$

A.2 Decoder Semantics, Calibration, and Provenance

For an Exponential rate $\lambda^{(y)}$ over $y = \log(1+\Delta t)$, the operational readout is $g_{\text{op}} = \log(1+1/\lambda^{(y)})$, the conditional median is $g_{\text{med}} = (\log 2)/\lambda^{(y)}$, and the conditional mean is $g_{\text{mean}} = 1/\lambda^{(y)}$. These functionals answer different point-prediction objectives; the matched Retweet comparison is reported in the main text.

Calibration and provenance. A training-fitted scale $g_c = c/\lambda^{(y)}$ yields $c^\star$ values near one across the evaluated datasets, while the Exponential conditional median remains $(\log 2)/\lambda^{(y)}$. This scalar diagnostic does not resolve whether the remaining cross-dataset differences arise from calibration, conditional-density misspecification, or dataset-specific dynamics. The matched decoder audit regenerated SAHP and AttNHP checkpoints because the original canonical weights were unavailable; its checkpoint set is distinct from the headline audit. One Cyberbullying SAHP matched-run cell marginally exceeded the predefined operational regression tolerance, with an absolute discrepancy of 2.12×10^{-3} against a 2.00×10^{-3} threshold. We retain this cell for transparency; the discrepancy is comparable to its cross-seed variation and does not alter the qualitative decoder-family conclusion.

Table 7: **Training-fitted scale $c^\star$ in $g_c = c/\lambda^{(y)}$.**

Model	Retweet	Cyberbullying	YouTube
SAHP	0.9984	1.0001	1.0014
AttNHP	0.9897	0.9960	0.9857

Audit provenance. Table 8 summarizes the checkpoint families and pairing structure used by the principal analyses.

A.3 MIMIC Construction and Audit

Construction. We construct event sequences from the first 24 hours of each MIMIC-IV ICU stay using charted vital signs, laboratory measurements, inputs, outputs, and procedures. Events are assigned to stays using the corresponding stay/admission identifiers and timestamps. Events occurring within the same minute are bundled into a multi-category mark, yielding a 31-mark vocabulary. We use an 80/10/10 patient-disjoint split.

Attrition and splitting. Starting from 94,458 ICU stays from 65,366 subjects, 94,455 stays have at least one event in the first 24 hours; removing sequences with fewer than 20 bundles discards 2,140 stays, and truncating sequences longer than 512 bundles affects 1 stay, resulting in 92,315 retained sequences with mean length 65 and median length 61. We use an 80/10/10 patient-disjoint

Table 8: **Checkpoint families and pairing structure for principal audits.**

Analysis	Checkpoint family	Seeds	Intervention	Uncertainty unit	Cross-table relation
Headline audit	Canonical headline	5	Global	Seed-level Student-t	Reference
Matched decoder	Canonical-recipe regenerated	5	Global	Seed-level Student-t	Not paired with headline
Same-mark decoder	Same matched decoder family	5	Same-mark	Seed-level Student-t	Paired within audit
Factorial	Dedicated matched factorial	5/cell	Global	Seed-level Student-t	Within-factorial only
S2P2	EasyTPP, matched budget	5	Global	Seed-level Student-t	Architecture reference

split by `subject_id`, giving 73,928 training, 9,340 validation, and 9,047 test sequences, with no subject shared across splits.

The sequences are converted to the same event-sequence representation used by the retweet-cascade datasets and evaluated with the same next-time target, event–time reassignment, representation-access protocol, native evaluation, and five-seed uncertainty estimation.

Table 9: **MIMIC clinical event-sequence construction.** Sequences contain minute-bundled clinical events from the first 24 hours of each ICU stay. Splits are patient-disjoint by `subject_id`.

Statistic	Value
Source ICU stays	94,458
Source subjects	65,366
Stays with ≥ 1 event in window	94,455
Removed (< 20 bundles)	2,140
Truncated (> 512 bundles)	1
Retained sequences	92,315
Event-mark vocabulary K	31
Observation window	First 24 h
Minimum sequence length	20 bundles
Maximum sequence length	512 bundles
Mean retained length	65
Median retained length	61
Train sequences	73,928
Validation sequences	9,340
Test sequences	9,047
Split	80/10/10, subject-disjoint

Raw-history gate. The MIMIC construction is screened with the same fixed-capacity mark-aware raw-history diagnostic used for the twelve-benchmark suite, under the same target, event–time reassignment, and predictor protocol.

The screen yields $S_{\mathrm{raw}} = 0.018345$ with a hash-averaged seed interval of $[0.018251, 0.018438]$. Against the constant-median baseline defined in Appendix A.10, the raw predictive margin is $G_{\mathrm{raw}} = 0.022351$, with raw clean MAE 0.178302 and constant-median test MAE 0.200652. The corresponding relative degradation is $R_{\mathrm{raw}} = 0.102887$.

The result is stable across the three independent hash realizations: S_{raw} ranges from 0.017794 to 0.018956, and G_{raw} from 0.022065 to 0.022618. For each hash realization, predictor-seed uncertainty is summarized with a 95% Student-t interval over five predictor seeds. The reported point estimate and interval endpoints are then averaged across the three hash realizations. We refer to the resulting quantity as a *hash-averaged seed interval*; it summarizes seed-level uncertainty and hash stability but is not interpreted as a joint 95% confidence interval over both sources of variation.

Thus, reassignment-sensitive temporal opportunity is detectably present upstream in the MIMIC construction.

Operational-decoder audit. Table 10 preserves the original five-seed cross-domain result under the stated model-specific decoders. Alternative Exponential point functionals were not evaluated on this construction.

Table 10: **Cross-domain audit on MIMIC clinical event sequences under the original operational decoders.** Five-seed means with 95% Student-t intervals ($df = 4$). Bold entries indicate reliably positive decoded point-prediction MAE sensitivity. NHP's decoded MAE response is reported, but its interpretation is qualified by substantially weaker clean-task predictive quality than the other three models.

Model	Clean MAE	S_{access}	$S^{\text{MAE}}_{\text{native}}$	$S^{\text{NLL}}_{\text{native}}$	Pearson r
THP	0.1613 [0.1586,0.1639]	0.0433 [0.0426,0.0439]	**0.0468 [0.0454,0.0481]**	0.4183 [0.3696,0.4670]	0.6037 [0.6013,0.6061]
SAHP	0.1680 [0.1669,0.1692]	0.0432 [0.0422,0.0443]	**0.0379 [0.0367,0.0391]**	0.1906 [0.1808,0.2004]	0.6064 [0.6036,0.6092]
AttNHP	0.1709 [0.1692,0.1726]	0.0321 [0.0295,0.0348]	**0.0341 [0.0326,0.0356]**	0.1944 [0.1857,0.2030]	0.5844 [0.5820,0.5869]
NHP	0.2835 [0.2819,0.2851]	0.0469 [0.0444,0.0493]	-0.0033 [-0.0052,-0.0013]	1.3712 [1.3041,1.4383]	0.0498 [0.0306,0.0691]

A.4 Native-Pathway Factorial on Retweet

Table 11 preserves the original matched factorial under the original pathway-specific decoders. The same-checkpoint mean re-decoding of its AttNHP Exponential cell is reported in Sec. 5.5.

Table 11: **End-to-end 2×2 native-pathway factorial on Retweet.** The Gaussian and Exponential native pathway configurations are crossed with the THP and AttNHP backbones; each configuration jointly sets output parameterization, training objective, and original point decoder g_0. S_{access} uses the same standardized frozen readout as the main audit.

Backbone	Native pathway	Clean MAE	S_{access}	$S^{\text{MAE}}_{\text{native}}$	Std. ratio
THP	Gaussian	1.039	0.641	0.855	0.862
THP	Exponential	2.482	0.637	-0.005	0.184
AttNHP	Gaussian	1.077	0.900	0.891	0.848
AttNHP	Exponential	2.472	0.874	-0.025	0.182

A.5 Frozen-Representation and Supervision Controls

Frozen AttNHP representations support intervention-sensitive point prediction under both Gaussian and Exponential heads. On Retweet, the matched head-family control gives $S_{\text{MAE}} = 0.778$ $[0.759, 0.797]$ for the Gaussian head and 1.125 $[1.080, 1.170]$ for the Exponential head. Thus, the representation and output family can support a strong response under the same operational decoder.

As a complementary recoverability control, clean-output distillation with the encoder frozen and native heads reinitialized yields $S_{\text{global}} = 0.658$ $[0.641, 0.675]$ under the native Exponential pathway. This establishes that a sensitive native solution is recoverable from the same representation; it does not identify which part of the modified supervision causes the recovery. Additional decoder-reset and intervention-conditioned access-head controls were consistent with this qualitative conclusion; we omit their implementation-level diagnostics here to keep the appendix focused on the controls directly supporting the main claims.

A.6 Intervention Robustness

Global permutation provides a strong perturbation, but conclusions based on a single extreme intervention may be sensitive to distribution shift. We therefore construct a graded family of block-wise event–time reassignments. For block size k, the historical interval sequence is divided into contiguous blocks $B_1, B_2, \ldots$, and interval assignments are permuted only within each block. This construction preserves the global interval multiset as well as the interval multiset within each block. Small values of k produce highly local reassignment; as k increases, intervals can move progres-

sively farther from their original positions, and the global shuffle corresponds to the limiting case in which intervals may be reassigned across the entire observed history.

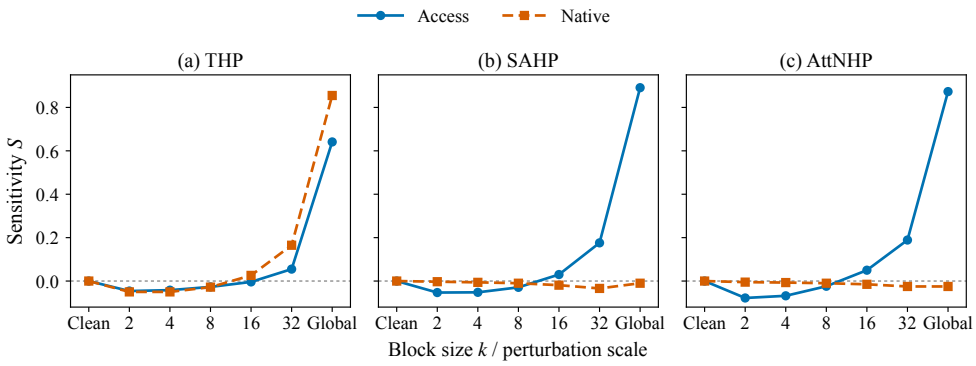

Figure 4: **Graded local-to-global MAE response on Retweet under the original operational decoders.** As reassignment becomes more non-local, THP develops sensitivity in both pathways, whereas SAHP and AttNHP develop increasing access sensitivity without a corresponding increase in decoded point-prediction MAE under the original point readouts. The curves do not establish a decoder-invariant stage effect.

We quantify perturbation magnitude using mean positional displacement

$$D_{\text{disp}} = \frac{1}{n} \sum_{i=1}^{n} |\pi(i) - i| \,,$$

and for each block size k, temporal sensitivity is

$$S(k) = \text{MAE}_k - \text{MAE}_{\text{clean}},$$

computed separately for the access and native prediction pathways. Figure 4 and Table 12 show that the access and operationally decoded point-prediction MAE responses separate as reassignment becomes broader. For AttNHP, the paired 95% bootstrap interval for $\Delta_k = S_{\text{access}}(k) - S^{\text{MAE}}_{\text{native}}(k)$ moves from $\Delta_2 = -0.074\,[-0.075, -0.072]$ to $\Delta_{32} = +0.214\,[+0.211, +0.216]$ and $\Delta_{\text{global}} = +0.903\,[+0.896, +0.910]$. These are error effects under the operational decoders and do not establish decoder-invariant prediction movement.

Table 12: **Selected local-to-global MAE intervention effects on Retweet.** Entries are taken from the dedicated graded-intervention experiment and report representation-level access sensitivity $S_{\text{access}}(k)$ and decoded native point-prediction MAE sensitivity $S^{\text{MAE}}_{\text{native}}(k)$. The canonical global-reassignment effects are reported separately in Table 1.

	THP		SAHP		AttNHP	
Perturbation	S_{access}	$S^{\text{MAE}}_{\text{native}}$	S_{access}	$S^{\text{MAE}}_{\text{native}}$	S_{access}	$S^{\text{MAE}}_{\text{native}}$
$k = 4$	-0.042	-0.050	-0.052	-0.006	-0.068	-0.007
$k = 16$	-0.004	$+0.026$	$+0.030$	-0.019	$+0.050$	-0.015
$k = 32$	$+0.055$	$+0.165$	$+0.176$	-0.034	$+0.189$	-0.025

Same-mark reassignment control. The same-mark global control reports $P_{\text{moved}} = 1.000$ and retains the following fraction of each model's global native-NLL response:

Table 13: **Same-mark reassignment and native-NLL response retention.**

Model	S_{NLL} retention	P_{moved}
THP	97.20%	1.000
SAHP	95.67%	1.000
AttNHP	95.90%	1.000

Table 14: **Same-mark point-functional audit on Retweet.** Five-seed means with 95% Student-t intervals. Same-mark reassignment restricts interval permutations within mark classes. D_{point} denotes mean absolute clean–reassigned prediction movement.

Model	Readout	Clean MAE	$S^{\mathrm{MAE}}_{\mathrm{point}}$ [95% CI]	D_{point}
SAHP	g_{op}	2.4733	$-0.0114\,[-0.0211, -0.0016]$	0.312
SAHP	g_{med}	1.4782	$+0.3249\,[+0.2743, +0.3755]$	1.038
SAHP	g_{mean}	1.0418	$+0.8163\,[+0.7609, +0.8717]$	1.497
AttNHP	g_{op}	2.4779	$-0.0175\,[-0.0218, -0.0132]$	0.326
AttNHP	g_{med}	1.5044	$+0.3373\,[+0.3052, +0.3694]$	1.091
AttNHP	g_{mean}	1.0926	$+0.8255\,[+0.7873, +0.8637]$	1.574

The same-mark intervention preserves the decoder ordering observed under global reassignment. Relative to the matched global effects, the direct median retains approximately 93.6% and 95.2% of the SAHP and AttNHP effects, respectively, while the conditional mean retains approximately 95.3% and 96.0%. The operational effects remain slightly negative. Thus, restricting interval reassignment within mark classes substantially preserves the point-functional contrast.

The retention percentages compare matched global and same-mark reassignment under the same checkpoint family. An operational-MAE retention ratio is not reported because both effects are near zero. This control narrows mark–time mismatch but does not establish that the reassigned histories lie on the data support.

Per-example error cancellation. For each evaluated target, we define

$$\delta e_i = |\hat{y}_{i,\mathrm{int}} - y_i| - |\hat{y}_{i,\mathrm{clean}} - y_i|.$$

Under the operational readout, predictions remain non-invariant, with mean absolute clean–intervened movement of approximately 0.32. Nevertheless, the directions of the per-example error changes are nearly balanced: $P(\delta e_i > 0) = 0.482$ and $P(\delta e_i < 0) = 0.506$, yielding a mean aggregate error change near zero. Under the conditional-mean readout, $P(\delta e_i > 0) \approx 0.69$, producing a strongly positive aggregate point-MAE response. Thus, the weak operational endpoint reflects cancellation of per-example error increases and decreases rather than a lack of prediction movement.

Raw-time metric-space audit. This appendix reports the full metric-space diagnostics underlying the raw-time control in Sec. 5.5. All results use an independently repeated run family (tail_run) and are not mixed with the canonical unified-run audit.

For each test sequence i, model, and seed, define the per-sequence raw-time error change

$$\delta_i = |\widehat{\Delta t}_i^{\,\mathrm{int}} - \Delta t_i| - |\widehat{\Delta t}_i^{\,\mathrm{clean}} - \Delta t_i|,$$

where $\widehat{\Delta t}^{\,\mathrm{op}} = 1/\lambda^{(y)}$ is the algebraic inverse transformation of the canonical point decoder and Δt_i is the true next interval. We partition sequences into bins by the percentile of the true target Δt_i, defined per dataset and shared across models and seeds. For bin b,

$$C_b = \frac{1}{N}\sum_{i \in b} \delta_i, \qquad S^{\mathrm{MAE}}_{\Delta t} = \sum_b C_b,$$

where N is the total number of test sequences. The top-1% bin corresponds to the 99–100 percentile of Δt_i.

Table 15: **Raw-time metric-space audit for SAHP and AttNHP.** $S_{\Delta t}^{\mathrm{MAE}}$ measures MAE degradation in raw inter-event time ($\widehat{\Delta t}^{\,\mathrm{op}} = 1/\lambda^{(y)}$) under global reassignment. C_b for the top-1% bin is the weighted contribution of the largest true-target sequences to $S_{\Delta t}^{\mathrm{MAE}}$, not a fraction. Brackets denote 95% Student-t confidence intervals across five seeds. All results use the independently repeated tail_run family.

Dataset	Model	$S_{\Delta t}^{\mathrm{MAE}}$	Top-1% C_b
Retweet	SAHP	$+0.0765$ $[+0.0249, +0.1281]$	$+0.0337$ $[+0.0314, +0.0359]$
Retweet	AttNHP	$+0.0369$ $[+0.0158, +0.0581]$	$+0.0295$ $[+0.0284, +0.0306]$
Cyberbullying	SAHP	-0.3515 $[-0.3722, -0.3308]$	$+0.0112$ $[+0.0105, +0.0120]$
Cyberbullying	AttNHP	-0.3375 $[-0.3632, -0.3118]$	$+0.0107$ $[+0.0098, +0.0116]$
YouTube	SAHP	-0.3444 $[-0.3771, -0.3117]$	$+0.0137$ $[+0.0128, +0.0147]$
YouTube	AttNHP	-0.3398 $[-0.3689, -0.3107]$	$+0.0124$ $[+0.0118, +0.0131]$

On Retweet, $S_{\Delta t}^{\mathrm{MAE}}$ is positive for both SAHP and AttNHP, while on Cyberbullying and YouTube the aggregate raw-space effects are negative. The top-1% target-interval bin nevertheless makes a positive contribution for all six model–dataset combinations. Because $S_{\Delta t}^{\mathrm{MAE}}$ is the sum of the bin-wise contributions, the remaining target-interval bins on Cyberbullying and YouTube collectively contribute a larger negative change that outweighs this positive tail contribution. The aggregate sign therefore depends on how intervention-induced error changes are distributed across the target range. We treat this as a metric-dependence observation rather than as evidence that reassignment improves temporal modeling. Most importantly, moving from $\log(1 + \Delta t)$ evaluation to raw inter-event time does not recover a consistently positive decoded-MAE response. The transformed evaluation space is therefore not a sufficient explanation for the main cross-dataset pattern.

Error-scale robustness. The primary analysis uses the absolute intervention effect

$$S = \mathrm{MAE}_{\mathrm{intervention}} - \mathrm{MAE}_{\mathrm{clean}},$$

because it directly measures the change in prediction error under a paired intervention. Since clean MAE differs across models and prediction pathways, we additionally evaluate a relative-degradation statistic

$$R = \frac{\mathrm{MAE}_{\mathrm{intervention}} - \mathrm{MAE}_{\mathrm{clean}}}{\mathrm{MAE}_{\mathrm{clean}}} = \frac{S}{\mathrm{MAE}_{\mathrm{clean}}}.$$

The denominator is always the clean MAE of the *same prediction pathway*; in particular, access effects are normalized by access clean MAE rather than native clean MAE. Confidence intervals are obtained with the same paired bootstrap protocol as the main analysis, recomputing both numerator and denominator within each bootstrap replicate.

The reported relative effect is estimated within paired bootstrap replicates rather than computed as the ratio of separately aggregated summary statistics. For bootstrap replicate b, we compute

$$R^{(b)} = \frac{\mathrm{MAE}^{(b)}_{\mathrm{intervention}} - \mathrm{MAE}^{(b)}_{\mathrm{clean}}}{\mathrm{MAE}^{(b)}_{\mathrm{clean}}}.$$

We then summarize the bootstrap distribution of $R^{(b)}$. Consequently, the reported R need not equal the ratio between rounded aggregate S and aggregate clean-MAE values shown elsewhere.

Table 16: **Relative degradation under global event–time reassignment on Retweet.** Normalizing each pathway by its own clean MAE preserves the main access–decoded-error contrast. $R^{\mathrm{MAE}}_{\mathrm{native}}$ measures relative degradation of the decoded native point prediction. Relative effects and confidence intervals are obtained from the paired bootstrap distribution, with numerator and denominator recomputed within each replicate.

Model	Access MAE	R_{access}	Native MAE	$R^{\mathrm{MAE}}_{\mathrm{native}}$
THP	1.143	$0.555\ [0.532, 0.578]$	1.041	$0.813\ [0.778, 0.848]$
SAHP	1.070	$0.851\ [0.813, 0.888]$	2.473	$-0.004\ [-0.008, 0.000]$
AttNHP	1.127	$0.735\ [0.703, 0.766]$	2.478	$-0.007\ [-0.009, -0.005]$

The scale-normalized analysis therefore reproduces the access–decoded-error contrast in MAE space. THP exhibits substantial relative degradation in both pathways ($R_{\mathrm{access}} = 0.555$, $R^{\mathrm{MAE}}_{\mathrm{native}} = 0.813$), whereas SAHP and AttNHP retain large access effects (0.851 and 0.735) but near-zero relative decoded MAE effects (-0.004 and -0.007). Therefore, the observed error-level contrast is not solely a consequence of the larger clean native MAEs of SAHP and AttNHP. This analysis does not characterize the separate native likelihood sensitivity analyzed in Sec. 5.3.

Access-head training-subset robustness. The canonical access-head protocol uses the first 800 training sequences. To test whether the reported effects depend on this particular subset, we repeat access-head training using three independently sampled 800-sequence subsets for each of the five encoder seeds, producing 15 runs per model.

Across these random subsets, the access sensitivities are

$$S^{\mathrm{THP}}_{\mathrm{access}} = 0.583 \pm 0.034,$$

$$S^{\mathrm{SAHP}}_{\mathrm{access}} = 0.909 \pm 0.030,$$

and

$$S^{\mathrm{AttNHP}}_{\mathrm{access}} = 0.854 \pm 0.024.$$

For comparison, the canonical first-800 estimates are 0.641, 0.891, and 0.873, respectively. These first-800 estimates are from this robustness experiment and need not equal the canonical headline values in Table 1. SAHP and AttNHP closely reproduce their access effects. THP exhibits a moderately smaller access effect under random subset sampling, but remains strongly intervention-sensitive. The qualitative three-model pattern is therefore unchanged: THP exhibits substantial MAE sensitivity through both access and decoded native point-prediction pathways, whereas SAHP and AttNHP retain large access effects with much smaller decoded native MAE effects.

Table 17: **Robustness to access-head training-subset selection.** Random-800 results aggregate three subset-sampling seeds for each of five encoder seeds. The first-800 column reports the corresponding first-800 estimates from this robustness experiment, which need not equal the canonical headline values in Table 1.

Model	First-800	Random-800	Range
THP	0.641	0.583 ± 0.034	[0.531,0.653]
SAHP	0.891	0.909 ± 0.030	[0.861,0.962]
AttNHP	0.873	0.854 ± 0.024	[0.798,0.885]

A.7 TPP Model Configurations

Table 18 summarizes the architecture and core training configuration for each TPP model. S2P2 is evaluated using the EasyTPP implementation under the same cascade training budget and five-seed audit protocol.

Parameter counts and mark-embedding details are implementation-specific and are reported in the released configuration files.

Table 18: **Model architectures, core training configuration, and point readouts.**

Setting	THP	SAHP	AttNHP	NHP	S2P2
State/hidden dimension	64	64	64	64	64
Architecture	Transformer	Transformer	Temporal attention	CT-LSTM	State-space
Layers	2	2	2	—	—
Attention heads	4	4	4	—	—
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW
Learning rate	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
Epochs	10	10	10	10	10
Seeds	0–4	0–4	0–4	0–4	0–4
Point readout	Gaussian mean	g_{op}	g_{op}	Prospective mean	EasyTPP native next-time prediction

A.8 Access-Head and Statistical Procedures

All clean–intervention comparisons are paired using the same test examples and intervention manifest. Raw-history sensitivity is element-pooled, whereas native TPP sensitivity is aggregated by sequence; these are different estimands.

Access head. For each frozen representation, the same 65-parameter linear head predicts the shared target $y_{i+1} = \log(1 + \Delta t_{i+1})$. It is trained only on clean representations using a fixed training subset, then evaluated on paired clean and reassigned representations with the target held fixed. Five seeds are used, paired with the corresponding encoder seeds.

Uncertainty and factorial contrasts. Headline results report means and 95% Student-t intervals across five seeds. Paired bootstrap intervals are used for single-model intervention contrasts and factorial controls; the resampling unit and interval type are reported with each analysis. The end-to-end factorial interaction is

$$I = \left[S^{\mathrm{MAE}}_{\mathrm{THP,Exp}} - S^{\mathrm{MAE}}_{\mathrm{THP,Gauss}}\right] - \left[S^{\mathrm{MAE}}_{\mathrm{AttNHP,Exp}} - S^{\mathrm{MAE}}_{\mathrm{AttNHP,Gauss}}\right].$$

The four cells use a dedicated matched run family and support within-factorial contrasts; the canonical headline estimates remain in Table 1.

A.9 Dataset Sources and Preprocessing

The twelve benchmarks are drawn from three public data sources: EasyTPP (Xue et al., 2024), Meta-TPP (Bae et al., 2023), and external public event-log repositories. Table 20 summarizes the source, split protocol, sequence counts, mark vocabulary size, and sequence-length statistics for each dataset.

For the seven EasyTPP benchmarks (Retweet, StackOverflow, Earthquake, Taobao, Amazon, Taxi, and Volcano), we use the official EasyTPP train/validation/test splits without repartitioning the data. The Retweet benchmark traces to the Twitter cascade data released with SEISMIC (Zhao et al., 2015); the StackOverflow badge-event benchmark was introduced by Du et al. (2016); and the Taxi records trace to the New York City taxi data release documented by Whong (2014). Taobao uses Alibaba's user-behavior data (Alibaba Cloud Tianchi, n.d.) and the event-sequence construction described by Xue et al. (2022). Amazon uses the 2018 Amazon Review Data (Ni et al., 2019). EasyTPP's Earthquake and Volcano event-sequence files are processed from the U.S. Geological Survey earthquake data (U.S. Geological Survey, n.d.) and the Smithsonian Global Volcanism Program (Smithsonian Institution Global Volcanism Program, n.d.), respectively.

The remaining five benchmarks do not provide an official TPP split under the preprocessing used here. For Reddit, MOOC, Wiki, BPI 2012, and eCommerce, we therefore perform a sequence-level 80/10/10 split with fixed random seed 0. The split is performed at the sequence level, so events from the same sequence never appear in multiple partitions.

Reddit, MOOC, and Wiki are obtained from the Meta-TPP release (Bae et al., 2023), whose processed NPZ files contain the complete set of sequences without predefined train/validation/test partitions. BPI 2012 is parsed from the original XES event log (van Dongen, 2012). For eCommerce, user-level event sequences are constructed from the REES46 cosmetics-shop CSV data (Kechinov, 2019) before applying the sequence-level split.

All datasets are converted to the common representation used by the unified raw-history diagnostic, where each event is represented by a categorical mark and an inter-event interval Δt. No common chronological split is imposed across datasets.

External retweet-cascade replication datasets. Cyberbullying and YouTube are drawn from the ICWSM'20 Twitter cascade collection (Wu et al., 2020). Unlike the twelve datasets used for broad exploratory screening, these two datasets are used specifically to test cross-dataset replication within the retweet-cascade process family.

We retain cascades containing at least nine retweets and then perform a sequence-level 80/10/10 train/validation/test split with fixed random seed 0. The split is performed after the length filter, and no cascade occurs in more than one partition. Retweeters are mapped to the same three follower-count mark categories as in the Retweet benchmark: small (< 120 followers), medium (120–1362), and large (≥ 1363). No dataset-specific re-binning is performed. Table 19 summarizes the exact processed datasets used in the experiments.

Table 19: **External retweet-cascade replication datasets.** Counts refer to the final processed sequences used by the experiments after retaining cascades with at least nine retweets. Splits are performed at the sequence level after filtering using fixed random seed 0.

Dataset	K	#Train	#Val	#Test	Mean len.	Median len.
Cyberbullying	3	353,163	44,145	44,146	94.4	21
YouTube	3	216,424	27,053	27,053	85.3	20

Clinical cross-domain dataset (MIMIC). The MIMIC clinical event-sequence construction contains 31 event marks and is used as an out-of-family cross-domain test. It is converted to the same event-sequence representation and evaluated with the same target, intervention, access protocol, and five-seed procedure as the other datasets. Construction details, event-mark definition, split counts, and sequence statistics are reported in Appendix A.3.

Table 20: **Dataset sources, splits, and sequence statistics.** K denotes the number of distinct event marks. Sequence counts refer to the final partitions used in our experiments. EasyTPP datasets use their official predefined splits. The five external datasets use a sequence-level 80/10/10 split with fixed random seed 0.

Dataset	Source	K	#Train	#Val	#Test	Split protocol	Mean len.	Median len.
Retweet	EasyTPP	3	20,000	2,000	2,000	Official EasyTPP split	93.0	90
StackOverflow	EasyTPP	22	1,401	401	401	Official EasyTPP split	64.6	57
Earthquake	EasyTPP	7	3,000	400	900	Official EasyTPP split	16.5	16
Taobao	EasyTPP	17	1,300	200	500	Official EasyTPP split	56.5	61
Amazon	EasyTPP	16	6,454	922	1,851	Official EasyTPP split	44.7	42
Taxi	EasyTPP	10	1,400	200	400	Official EasyTPP split	37.0	38
Volcano	EasyTPP	1	400	50	181	Official EasyTPP split	14.0	5
Reddit	Meta-TPP	985	8,000	1,000	1,000	Seq.-level 80/10/10 (seed 0)	65.3	60
MOOC	Meta-TPP	98	5,637	704	706	Seq.-level 80/10/10 (seed 0)	68.1	65.5
Wiki	Meta-TPP	8,227	800	100	100	Seq.-level 80/10/10 (seed 0)	114.6	125
BPI 2012	4TU.ResearchData	36	7,726	966	966	Seq.-level 80/10/10 (seed 0)	26.1	24
eCommerce	REES46 / Kaggle	4	82,768	10,346	10,347	Seq.-level 80/10/10 (seed 0)	32.1	16

A.10 RAW-HISTORY SCREENING PROTOCOL

The screen compares a fixed-capacity raw-history predictor with a constant training-median baseline in the transformed target $y = \log(1 + \Delta t)$. For N test targets, define

$$\mathrm{MAE}_{\mathrm{const,test}} = \frac{1}{N}\sum_{i=1}^{N} |y_i - m|,$$
$$G_{\mathrm{raw}} = \mathrm{MAE}_{\mathrm{const,test}} - \mathrm{MAE}^{\mathrm{raw}}_{\mathrm{clean}},$$
$$S_{\mathrm{raw}} = \mathrm{MAE}^{\mathrm{raw}}_{\mathrm{int}} - \mathrm{MAE}^{\mathrm{raw}}_{\mathrm{clean}}, \qquad R_{\mathrm{raw}} = \frac{S_{\mathrm{raw}}}{\mathrm{MAE}^{\mathrm{raw}}_{\mathrm{clean}}}.$$

Here m is the training-target median. To keep predictor capacity comparable across datasets with different mark vocabularies, the predictor uses a fixed-dimensional, non-trainable mark hash and a fixed history length, followed by a linear predictor. The same target and intervention are used across all twelve benchmarks; results are averaged across three hash realizations and five predictor seeds. The cross-benchmark table below reports the complete screening results. Increasing the hash dimension on Wiki and Reddit leaves their qualitative conclusions unchanged. Synthetic positive and negative controls confirm that the audit distinguishes interval-independent from metric-time-dependent targets.

A.11 CROSS-BENCHMARK RAW-SCREENING RESULTS

On Volcano, THP, SAHP, and AttNHP show no reliable representation-level access sensitivity, while the raw-history screen remains positive; the cross-benchmark estimates are reported in Table 21.

Cross-benchmark screening intervals. Table 21 reports the full cross-benchmark raw-screening statistics with hash-averaged seed intervals for S_{raw}, computed as described in Appendix A.8. The hash-seed stability column reports the min–max range across the three hash realizations.

Table 21: **Cross-benchmark raw-screening statistics with hash-averaged seed intervals.** For each hash realization, predictor-seed uncertainty is summarized with a 95% Student-t interval over five predictor seeds (df=4). The reported point estimate and interval endpoints are then averaged across the three hash realizations. We refer to the resulting quantity as a *hash-averaged seed interval*; it summarizes seed-level uncertainty and hash stability but is not interpreted as a joint 95% confidence interval over both sources of variation. Hash-seed stability reports the min–max range across the three hash realizations.

Dataset	G_{raw}	S_{raw}	Hash-avg. seed interval	Hash stability
Retweet	+0.93	+0.715	$[+0.712, +0.718]$	$[+0.7149, +0.7154]$
Volcano	+0.21	+0.139	$[+0.131, +0.147]$	$[+0.1389, +0.1392]$
Taxi	+0.007	+0.0005	$[+0.0004, +0.0006]$	$[+0.0004, +0.0005]$
StackOverflow	+0.014	-0.0001	$[-0.0002, +0.0000]$	$[-0.0002, -0.00005]$
Earthquake	+0.085	-0.012	$[-0.013, -0.011]$	$[-0.0124, -0.0124]$
Taobao	-0.005	$+0.00002$	$[+0.00001, +0.00003]$	$[+0.000015, +0.000020]$
Amazon	-0.001	-0.0002	$[-0.0003, -0.0001]$	$[-0.0002, -0.0001]$
Reddit	+0.001	-0.00003	$[-0.00004, -0.00002]$	$[-0.00005, -0.00002]$
MOOC	-0.0005	-0.000006	$[-0.000007, -0.000005]$	$[-0.000008, -0.000004]$
Wiki	≈ 0	$+0.0002$	$[+0.0001, +0.0003]$	$[+0.000148, +0.000164]$
BPI 2012	+0.244	$+0.0005$	$[+0.0004, +0.0006]$	$[0.0004, 0.0006]$
eCommerce	-0.004	$+0.0001$	$[+0.0000, +0.0002]$	$[0.0000, 0.0002]$

Table 22: **Prospective NHP intervention audit across the three retweet-cascade datasets (conditional-mean decoder).** Brackets denote 95% Student-t confidence intervals across five seeds. $G_{\mathrm{A-N}} = S_{\mathrm{access}} - S^{\mathrm{MAE}}_{\mathrm{native}}$ is computed per paired seed before forming the five-seed Student-t confidence interval.

Dataset	Native clean MAE	S_{access}	$S^{\mathrm{MAE}}_{\mathrm{native}}$	$G_{\mathrm{A-N}}$	$S^{\mathrm{NLL}}_{\mathrm{native}}$	Std. ratio	Pearson r
Retweet	1.102 [1.092,1.112]	1.066 [1.034,1.098]	0.766 [0.747,0.786]	0.300 [0.250,0.349]	1.329 [1.255,1.403]	0.722	0.823
Cyberbullying	1.489 [1.462,1.516]	0.251 [0.239,0.264]	0.132 [0.116,0.148]	0.120 [0.111,0.128]	0.321 [0.315,0.328]	0.606	0.783
YouTube	1.483 [1.442,1.525]	0.376 [0.356,0.396]	0.173 [0.162,0.183]	0.203 [0.177,0.230]	0.412 [0.392,0.432]	0.640	0.809

Table 23: **Decoder robustness for prospective NHP point prediction.** Decoder choice changes the magnitude of NHP's decoded intervention response and the relative access–native gap, but both strictly prospective decoders yield positive native point-prediction MAE sensitivity on all three datasets.

Dataset	Cond.-mean $S^{\mathrm{MAE}}_{\mathrm{native}}$	Plug-in $S^{\mathrm{MAE}}_{\mathrm{native}}$	Δ (CM $-$ plug-in)	CM std ratio	Plug-in std ratio
Retweet	0.766	1.718	-0.952 $[-1.105, -0.798]$	0.722	1.328
Cyberbullying	0.132	0.403	-0.271 $[-0.303, -0.239]$	0.606	0.827
YouTube	0.173	0.537	-0.365 $[-0.409, -0.321]$	0.640	1.014

A.12 Prospective NHP Decoder

Conditional-mean decoder. From the causal post-event state, NHP predicts the conditional mean waiting time

$$\mathbb{E}[\Delta t_{i+1} \mid \mathcal{H}_i] = \int_0^\infty \bar{F}_i(u)\, du, \qquad \bar{F}_i(u) = \exp\left[- \int_0^u \lambda_{\mathrm{tot}}(t_i + s \mid \mathcal{H}_i)\, ds \right].$$

The reported transformed point prediction is $\log(1 + \mathbb{E}[\Delta t_{i+1} \mid \mathcal{H}_i])$. The prospective conditional-median decoder $\log(1 + m_i)$, where $\bar{F}_i(m_i) = 1/2$, was not evaluated.

Causality and numerical evaluation. The representation and prediction use only the post-event history $\mathcal{H}_i$; changing the stored future interval while holding that history fixed leaves both unchanged. The survival integral is evaluated by adaptive quadrature until the survival tail is negligible. A prospective instantaneous-rate plug-in decoder is included as a robustness control.

The conditional-mean decoder produces smaller intervention effects than the instantaneous-rate plug-in on all three datasets, but the qualitative result is unchanged. Decoder semantics affect both the magnitude of the decoded intervention response and its relation to the standardized access readout, including the sign of $G_{\mathrm{A-N}}$: the paired gap flips from -0.652 to $+0.300$ on Retweet, -0.152 to $+0.120$ on Cyberbullying, and -0.161 to $+0.203$ on YouTube. They do not, however, change the endpoint conclusion: NHP retains reliably positive decoded point-prediction MAE sensitivity under both strictly prospective decoders.

A.13 Training Robustness

This appendix reports training-quality diagnostics for the 15 canonical attention-model Retweet model runs (three architectures $\times$ five seeds) used in the main analysis. All runs were trained using each model's native objective on clean histories only.

Table 24: **Training-quality summary for canonical Retweet model runs.** Each row aggregates five independent training seeds. "Instability" counts runs with NaN loss or divergence. "Stable" records the available validation summary within the ten-epoch budget; it does not establish convergence.

Model	Seeds	Mark Acc.	Time NLL	Point MAE	Instability /Validation status
THP	5	0.534	1.633	1.041	0 / Stable
SAHP	5	0.534	2.126	2.473	0 / Stable
AttNHP	5	0.534	2.132	2.478	0 / Stable

Across all 15 canonical runs, no run exhibited NaN loss, divergence, or termination due to numerical instability. The "Stable" entries describe the available within-budget validation summaries only; they do not by themselves establish convergence. Time NLL values are native-model training-quality diagnostics and are not used for cross-model likelihood ranking because the architectures employ different native likelihood parameterizations.

SAHP and AttNHP have substantially higher clean point MAE (2.473 and 2.478) than THP (1.041), and their decoded prediction ranges are narrow (Appendix A.14) under the operational readout. The 100-epoch endpoint summaries in Table 25 preserve operational std-ratios near 0.18, operational effects near zero, and conditional-mean effects around $+0.86$–$+0.88$. This weakens a simple ten-epoch stopping explanation. Epoch-wise validation curves and seed-level intervals were not supplied, so convergence cannot be fully assessed.

Table 25: **Retweet endpoint summaries after extending training to 100 epochs.** Epoch-10 rows provide within-model comparison baselines for the 100-epoch summaries. These endpoint summaries are separate from the matched Retweet decoder-family audit.

Model	Epochs	Operational std. ratio	S_{op}	S_{mean}	Validation NLL
SAHP	10	0.1829	-0.0098	$+0.8566$	2.1844
SAHP	100	0.1842	-0.0083	$+0.8648$	2.1871
AttNHP	10	0.1819	-0.0171	$+0.8612$	2.1907
AttNHP	100	0.1832	-0.0278	$+0.8752$	2.1869

A.14 Native Output-Space Diagnostics

This appendix reports the detailed native likelihood intervention statistics underlying $S^{\text{NLL}}_{\text{native}}$ (Eq. 1). All NLL values refer to the *time* component of the native predictive distribution only; mark cross-entropy is excluded.

For THP, SAHP, AttNHP, and S2P2, we report the five-seed clean NLL, reassigned NLL, and paired intervention effect. For NHP, the frozen audit records retain the paired per-seed NLL differences for all five seeds, but complete clean and reassigned component values are unavailable for seeds 1–4. We therefore report the valid five-seed paired contrast without constructing unsupported component-wise summaries.

Table 26: **Native time-likelihood intervention effects on Retweet.** $S^{\mathrm{NLL}}_{\mathrm{native}}$ is the per-model increase in native time NLL under global event–time reassignment. Because the evaluated models use different native likelihood formulations and target parameterizations, absolute values of $S^{\mathrm{NLL}}_{\mathrm{native}}$ are interpreted only within model and are not used for cross-model magnitude ranking. For NHP, only the paired five-seed intervention contrast was retained consistently across all seeds; component-wise clean and reassigned NLL summaries are therefore left unreported.

Model	Clean NLL	Reassigned NLL	$S^{\mathrm{NLL}}_{\mathrm{native}}$
THP	1.633	2.463	$0.830\ [0.786, 0.873]$
NHP	—	—	$1.329\ [1.255, 1.403]$
SAHP	2.126	2.260	$0.134\ [0.128, 0.140]$
AttNHP	2.132	2.260	$0.128\ [0.125, 0.131]$
S2P2	-1.3255	-0.1335	$1.1920\ [1.0931, 1.2908]$

All five evaluated models exhibit reliably positive native likelihood degradation under reassignment; for NHP, only the paired intervention contrast is reported because complete component-wise clean and reassigned summaries were not retained across all five seeds. The within-model contrast between $S^{\mathrm{NLL}}_{\mathrm{native}} > 0$ and small $S^{\mathrm{MAE}}_{\mathrm{native}}$ for SAHP and AttNHP is the key observation motivating the output-space decomposition in Sec. 5.3.

Output-space diagnostics. This table combines the Retweet native-likelihood and output-space diagnostics used in Sec. 5.3.

Table 27: **Native likelihood and output-space diagnostics on Retweet.** The within-model NLL change is not used for cross-model ranking. $\mathrm{Std}(\hat{y})/\mathrm{Std}(y)$ measures decoded range relative to the target, and drift is distinct from MAE change. All quantities are computed over the same five canonical model seeds and paired reassignment manifests used in the unified Retweet audit. Brackets denote 95% Student-t confidence intervals across seeds. NHP statistics are computed under the prospective conditional-mean decoder (Appendix A.12). Its prediction drift was not recomputed under this decoder, and $\mathbb{E}|\Delta \log \lambda^{(y)}|$ is reported only for the SAHP/AttNHP Exponential-head analysis; unavailable entries are denoted by "—".

Model	Within-model $S^{\mathrm{NLL}}_{\mathrm{native}}$	$\frac{\mathrm{Std}(\hat{y})}{\mathrm{Std}(y)}$	Pearson r	D_{native}	$\mathbb{E}\lvert\Delta \log \lambda^{(y)}\rvert$
THP	$0.830\ [0.786, 0.873]$	$0.861\ [0.844, 0.878]$	$0.839\ [0.839, 0.840]$	$1.554\ [1.516, 1.593]$	—
NHP	$1.329\ [1.255, 1.403]$	0.722	0.823	—	—
SAHP	$0.134\ [0.128, 0.140]$	$0.183\ [0.176, 0.190]$	$0.816\ [0.814, 0.818]$	$0.322\ [0.305, 0.338]$	$0.429\ [0.407, 0.450]$
AttNHP	$0.128\ [0.125, 0.131]$	$0.182\ [0.178, 0.186]$	$0.799\ [0.798, 0.801]$	$0.336\ [0.327, 0.345]$	$0.447\ [0.434, 0.459]$
S2P2	$1.1920\ [1.0931, 1.2908]$	$0.7801\ [0.7481, 0.8121]$	$0.7368\ [0.7339, 0.7397]$	$1.8468\ [1.7584, 1.9352]$	—

The clean correlations and nonzero output ranges in the table rule out a constant-output account. Clean and reassigned log rates remain moderately correlated ($r = 0.561\ [0.541, 0.581]$ for SAHP and $0.520\ [0.515, 0.525]$ for AttNHP), despite the mean absolute parameter changes shown in the table.