Transport Instability Coordinates for Early-Warning Tests of Training Loss Spikes

Kim Sanghyun [1]

Abstract

Loss spikes in neural-network training are typically detected only after probe loss has entered an unstable regime. We frame future spike detection as a sequential testing problem in which row-wise, causal features act as test statistics for a finite non-NaN probe-loss spike event in a future lead window. Standard diagnostics such as loss trends, gradient norms, update norms, and adaptive-moment statistics measure scalar or parameter-space quantities, but they do not directly measure how the current stochastic optimizer transition moves the model's representations. We introduce Transport Instability Coordinates (TIC), a causal feature-space diagnostic based on ghost updates from a frozen optimizer state. At each measured step, TIC samples ghost microbatches, applies the corresponding temporary AdamW transitions, and records the induced changes in a fixed probe representation and probe loss. These measurements define trace, log-moment, whitened, modal-concentration, and loss-directed transport coordinates, which we use in a discrete-time hazard model for future non-NaN probe-loss spikes. In dense critical-learning-rate Transformer experiments, TIC improves over time, loss-trend, gradient-norm, update-norm, Adam-mismatch, trace-only, and raw-tabular baselines at a fixed false-alert budget. A secondary PolyPythias-410M audit shows that the same coordinates decompose known outlier windows into coordinate-specific transport expansion. The results support feature-space transport as an early-warning statistic for imminent training instability.

[1]Independent Researcher, South Korea. Correspondence to: Kim Sanghyun <tomousdc@gmail.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

1. Introduction

Loss spikes are usually recognized after they have already appeared in the loss curve. At that point the practitioner can lower the learning rate, reset an optimizer state, clip an update, or discard the run (Pascanu et al., 2013). Large-scale language-model reports also describe rare training instabilities and loss spikes that require operational intervention (Chowdhery et al., 2023). The diagnostic itself has arrived late. The telemetry available before a spike is mostly scalar: loss trend, gradient norm, update norm, update-to-parameter ratio, or statistics derived from AdamW's adaptive moments (Kingma & Ba, 2015; Loshchilov & Hutter, 2019). These quantities are useful. They do not, however, answer a more direct question: what distribution of representation changes does the current stochastic optimizer transition induce?

This paper treats loss-spike prediction as a problem of feature-space transport and sequential event detection. At a measured training state, we freeze the model parameters and AdamW optimizer state. We then draw ghost microbatches, compute the AdamW-style frozen-moment transition each microbatch would induce from that frozen state, apply each transition only to a temporary model copy, and measure the resulting change in a fixed probe representation. This procedure converts the current optimizer state into an empirical conditional distribution over feature increments. Transport Instability Coordinates are statistics of this distribution, used here as early-warning features for the alternative that a spike event will occur in a future lead window.

The claim is intentionally narrow. Scalar optimizer telemetry is not uninformative; in our experiments, update norm is a competitive scalar baseline. The purpose of TIC is to measure a different object. Total transport, diffusion trace, coherent drift, relative transport expansion, low-rank whitened drift, modal concentration, and loss-directed transport are all available before a future spike label is observed. They provide a coordinate system for asking not only whether an update is large, but how it moves the model's representations and probe loss under counterfactual one-step optimizer transitions, complementing stability views based on parameter-space optimization dynamics

(Cohen et al., 2021).

Our prediction setting is deliberately strict. At measurement row t, all features are functions of the training filtration at t together with row-t ghost randomness. The target is a future finite non-NaN probe-loss spike in a lead window, not a high current loss. Labels are defined from a causal rolling baseline. Thresholds and operating rules are selected on validation runs only. The held-out test split is evaluated once. The dense experiment is a fixed-protocol pilot rather than a deployment-calibrated alarm benchmark.

The raw held-out test archive contains 50 finite spike peaks. One terminal/boundary peak is retained for auditability but excluded from event-level operating metrics because it lacks the complete valid lead-window/evaluation convention. Thus event-level metrics use 49 evaluable events. Under the validation-selected transport risk rule, TIC detects 95.9% of evaluable held-out events, with median lead time 90 optimizer steps and 6.05 realized test false alerts per 1000 stable optimizer steps. The log-moment plane is the dominant transport component: it accounts for nearly all of the AUPRC gain over scalar baselines. The selected TIC rule gives the strongest event-level operating point, but it should not be read as a universal alarm system.

Contributions. We define Transport Instability Coordinates, a ghost-transition coordinate system for measuring stochastic optimizer motion in feature space. We introduce the log-moment transport plane, a whitened transport coefficient, modal concentration, and a loss-directed transport coordinate. Two elementary observations support the role of these coordinates: TIC is not determined by parameter-space update norm alone, and the log-moment plane retains a common-mode degree of freedom that trace ratios discard. We evaluate TIC in an expanded fixed-protocol dense Transformer benchmark for future finite non-NaN probe-loss spike prediction. We report a matched PolyPythias-410M post-hoc anatomy audit comparing two candidate outlier checkpoint windows with fourteen matched stable windows. We state the limits of the evidence and provide basic identities establishing filtration measurability, drift-diffusion decomposition, scale behavior, and the frozen evaluation protocol.

2. Problem Setup

2.1. Training Filtration and Prediction Target

Let $\mathcal{F}_t$ denote the training filtration immediately before row-t ghost measurements. It contains the parameters, AdamW optimizer state, scheduler state, observed training losses, observed probe losses, gradients, updates, and all train-fitted preprocessing quantities available at or before measurement row t. Row-t ghost microbatch randomness

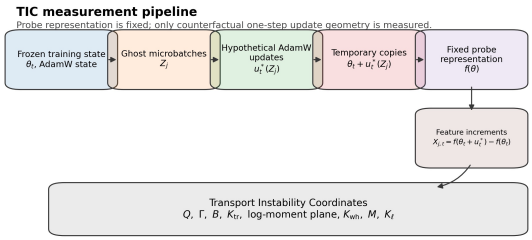

Figure 1. TIC measurement pipeline. The training state and AdamW state are frozen; ghost microbatches induce counterfactual AdamW-style frozen-moment transitions, whose effects on a fixed probe representation define transport coordinates.

is treated separately in Appendix D. The prediction target is

$$p_t^{(H)} = \Pr\left(\exists \tau \in [t + L_{\min}, t + H] \text{ such that } E_\tau = 1 \mid \mathcal{F}_t\right), \tag{1}$$

where E_τ is a finite probe-loss spike event. The future label is not a statement about whether the current loss is high. It asks whether a spike peak will occur after a minimum lead time and before a finite horizon.

Measurements are made every m optimizer steps. Internal formulas use measurement indices, but lead times are reported in optimizer steps. In the dense benchmark,

$$m = 10, \qquad L_{\min} = 2 \text{ measurement intervals}, \qquad H = 10 \text{ measurement i} \tag{2}$$

so the prediction window is 20 to 100 optimizer steps ahead.

2.2. Spike Definition

Let ℓ_t be the fixed probe loss at measurement row t. We define a causal robust baseline using only previous rows, with median/MAD-style scaling (Hampel et al., 1986; Rousseeuw & Croux, 1993):

$$b_t = \operatorname{median}\{\ell_i : i \in [t - W_b, t - L_{\text{gap}}]\}, \tag{3}$$

$$s_t = 1.4826 \operatorname{median}\{|\ell_i - b_t| : i \in [t - W_b, t - L_{\text{gap}}]\} + \epsilon. \tag{4}$$

Rows with too few baseline points are not eligible to be spike peaks. Prediction rows are excluded when their future lead-window/evaluation convention is incomplete. A spike peak at τ is

$$E_\tau = \mathbf{1}\{\ell_\tau - b_\tau \geq \max\{\delta_{\text{abs}}, \kappa s_\tau\}\}\mathbf{1}\{\ell_\tau \text{ is a local maximum}\} \times \mathbf{1}\{\ell_\tau < \infty\}\mathbf{1}\{\tau \text{ is outside the refractory window}\}. \tag{5}$$

The future label is

$$Y_t^{(H)} = \mathbf{1}\left\{\exists \tau \in [t + L_{\min}, t + H] \text{ such that } E_\tau = 1\right\}. \tag{6}$$

The event-search stage uses a pre-specified grid over δ_{abs} and κ to find a finite spike-producing regime. The search is predictor-blind: it uses event counts, nonfinite rates, label robustness, and time-only separability, but not TIC predictor performance. It does not use the final held-out test seeds 6–17, test predictor scores, or test operating metrics.

Frozen evaluation protocol. The train/validation/test seed-split rule is fixed before fitting any predictor or selecting any operating rule. After the predictor-blind event-search stage, the stress schedule, spike-label parameters, feature set, risk-model class, threshold grid, non-maximum suppression rule, validation selection criterion, and held-out test split are frozen. Held-out test predictor scores and held-out test operating metrics are not used to choose these objects. The held-out test labels affect only the final reported metrics.

3. Transport Instability Coordinates

3.1. Ghost Transition Distribution

At measurement row t, freeze the model parameters θ_t and AdamW optimizer state. For a ghost microbatch Z_j, let $u_t^*(Z_j)$ be the AdamW-style frozen-moment transition computed from the frozen state. The implementation applies the decoupled weight decay term and frozen first- and second-moment updates, without mutating the real model or optimizer state. With a fixed probe representation f, define the ghost feature increment

$$X_{j,t} = f(\theta_t + u_t^*(Z_j)) - f(\theta_t). \tag{7}$$

For k ghost microbatches, estimate

$$\mu_t = \frac{1}{k}\sum_j X_{j,t}, \tag{8}$$

$$Q_t = \frac{1}{k}\sum_j \|X_{j,t}\|_2^2, \tag{9}$$

$$\widehat{\Gamma}_t = \frac{1}{k-1}\sum_j \|X_{j,t} - \mu_t\|_2^2, \tag{10}$$

$$D_t = \frac{k-1}{k}\widehat{\Gamma}_t, \tag{11}$$

$$B_t = \|\mu_t\|_2^2. \tag{12}$$

Here $\widehat{\Gamma}_t$ is the sample-corrected diffusion-trace coordinate used in our reported features, while D_t is the denominator-k empirical dispersion. With this convention,

$$Q_t = D_t + B_t \tag{13}$$

exactly for the measured empirical ghost distribution. The result artifacts store the sample-corrected $\widehat{\Gamma}_t$ under the column name `Gamma`; the exact pathwise decomposition uses $D_t = (k-1)\widehat{\Gamma}_t/k$. Thus Q_t is total one-step mean-square feature transport, $\widehat{\Gamma}_t$ is the reported diffusion trace, and B_t is coherent drift energy.

3.2. Trace Transport and the Log-Moment Plane

The trace transport coefficient is

$$K_{\mathrm{tr},t} = \log\left(\frac{Q_t + \epsilon}{\widehat{\Gamma}_t + \epsilon}\right). \tag{14}$$

It is an isotropic drift-to-dispersion coordinate. It is useful, but it discards the separate scales of total transport and diffusion. Given causal running anchors $Q_{\mathrm{base},t}$ and $\widehat{\Gamma}_{\mathrm{base},t}$, define

$$\Delta_{Q,t} = \log(Q_t + \epsilon) - \log(Q_{\mathrm{base},t} + \epsilon), \tag{15}$$

$$\Delta_{\Gamma,t} = \log(\widehat{\Gamma}_t + \epsilon) - \log(\widehat{\Gamma}_{\mathrm{base},t} + \epsilon), \tag{16}$$

$$C_t = \frac{1}{2}(\Delta_{Q,t} + \Delta_{\Gamma,t}), \tag{17}$$

$$R_t = \Delta_{Q,t} - \Delta_{\Gamma,t}, \tag{18}$$

$$E_t = \max\{\Delta_{Q,t}, \Delta_{\Gamma,t}\}. \tag{19}$$

We call $(\Delta_Q, \Delta_\Gamma)$ the log-moment transport plane. C_t and E_t measure common-mode transport expansion. R_t measures whether the expansion is relatively more concentrated in total transport or diffusion. In the expanded dense benchmark, this plane is the dominant predictive transport component.

3.3. Whitened Transport

With k ghost samples, the empirical covariance has rank at most $k-1$. We therefore use a low-rank resolved subspace rather than a full covariance inverse. Let (e_i, λ_i) for $i = 1, \ldots, r$ be the top-r eigensystem of the centered ghost increments. The resolved whitened drift score is

$$W_t = \sum_{i=1}^{r} \frac{(e_i^\top \mu_t)^2}{\lambda_i + \lambda_{\mathrm{reg}}}. \tag{20}$$

The whitened transport coefficient is

$$K_{\mathrm{wh},t} = \log(1 + W_t). \tag{21}$$

This coordinate increases when coherent drift aligns with low-dispersion feature modes. It is an anisotropic extension of K_{tr} within the empirically resolved ghost subspace.

3.4. Modal Concentration

Define mode scores

$$s_i = \frac{(e_i^\top \mu_t)^2}{\lambda_i + \lambda_{\mathrm{reg}}}, \qquad S_t = \sum_{i=1}^{r} s_i. \tag{22}$$

If $S_t \le \epsilon_{\text{mode}}$, set $M_t = 0$. Otherwise define

$$\pi_i = \frac{s_i}{S_t}. \tag{23}$$

For $r > 1$, the modal concentration coordinate is

$$M_t = 1 - \frac{H(\pi)}{\log r}, \tag{24}$$

$$H(\pi) = -\sum_i \pi_i \log \pi_i. \tag{25}$$

For $r \le 1$, set $M_t = 1$ when $S_t > \epsilon_{\text{mode}}$ and $M_t = 0$ otherwise. This coordinate is near zero when resolved whitened drift is absent or spread across modes, and near one when nonzero resolved whitened drift is concentrated.

3.5. Loss-Directed Transport

Large representation transport is not necessarily dangerous. A candidate update that moves representations while lowering probe loss should not be counted in the same way as one that moves representations while increasing probe loss. For each ghost update, define

$$\delta\ell_{j,t} = \ell_{\text{probe}}(\theta_t + u_t^*(Z_j)) - \ell_{\text{probe}}(\theta_t). \tag{26}$$

Let $\bar{d}_t$ and v_t be the empirical mean and variance of these probe-loss changes. With small k, the variance denominator uses a fixed numerical floor. In the released implementation $\epsilon_\ell = \max\{10^{-8}, \epsilon\}$. The loss-directed transport coordinate is

$$K_{\ell,t} = \log\left(1 + \frac{\text{ReLU}(\bar{d}_t)^2}{v_t + \epsilon_\ell}\right). \tag{27}$$

This coordinate is causal in the filtration sense: it evaluates row-t candidate updates and does not use future losses.

3.6. Feature Vectors and Computational Cost

The transport TIC feature vector is

$$\Phi_t^{\text{tr}} = [E_t, C_t, R_t, K_{\text{tr},t}, K_{\text{wh},t}, K_{\ell,t}, M_t].$$

All transport coordinates are computed from the frozen row-t ghost distribution. We evaluate several feature families. Pure transport-family predictors use TIC coordinates only. The validation-selected TIC risk rule corresponds to the released artifact key `transport_full_hazard` and augments the full transport vector with optimizer step as a risk-time coordinate. All-feature logistic baselines additionally include scalar optimizer and loss telemetry. Table rows report the corresponding family.

A TIC measurement with k ghost microbatches requires k additional counterfactual optimizer transitions and probe evaluations at a measured row. Since measurements are made every m optimizer steps, the raw amortized frequency is approximately k/m ghost transitions per optimizer step, plus probe-evaluation cost. In the dense benchmark, $k = 4$ and $m = 10$, corresponding to roughly 0.4 ghost transitions per optimizer step before batching and implementation overheads. This cost is acceptable for a pilot diagnostic, but it is not a free training-time statistic.

4. Prediction Model

Stress schedules can confound spike prediction. We therefore report explicit time-only and learning-rate-only controls. For each predictor family, feature columns are imputed, centered, and scaled using training-split statistics only. The operational predictor is a fixed-horizon logistic risk model for $Y_t^{(H)}$:

$$s_t = \sigma\left(\alpha + \beta^\top \text{RN}(\Phi_t)\right). \tag{28}$$

The model is fit on train. Thresholds and non-maximum-suppression windows are selected on validation only, under the fixed false-alert selection rule. The held-out test split is evaluated once.

5. Experimental Design

5.1. Dense Critical-Learning-Rate Benchmark

The primary benchmark is a small byte-level GPT-style Transformer (Vaswani et al., 2017) trained on real WikiText-2 text (Merity et al., 2017) with AdamW (Loshchilov & Hutter, 2019). The model has hidden dimension 192, three Transformer blocks, sequence length 128, measurement interval 10 optimizer steps, and $\text{ghost_k} = 4$. The final frozen stress schedule is a linear learning-rate ramp from 0.03 to 0.30. The run uses held-out train, validation, and test seeds. The expanded fixed-protocol test split contains 50 raw finite spike peaks and 49 evaluable test spike events for the reported metrics. One terminal/boundary peak is retained for auditability but excluded from event-level operating metrics because it lacks the complete valid lead-window/evaluation convention.

The event-search stage initially considered several optimizer-stress regimes. Because non-spike settings cannot evaluate future spike prediction, the final regime was selected using only predictor-agnostic criteria: finite event count, nonfinite-run exclusion, label robustness under nearby spike thresholds, and rejection of settings that were nearly separable by time or learning rate alone. This stage did not use the final held-out test seeds 6–17, held-out test predictor scores, or held-out test operating metrics. After this stage, all protocol objects listed in Section 2.2 were frozen before held-out prediction.

Table 1. Held-out dense prediction results on the expanded fixed-protocol benchmark. The selected TIC row is the validation-selected transport risk rule. Metrics are reported on 49 evaluable test spike events; the raw held-out test archive contains 50 finite spike peaks, with one terminal/boundary peak retained for auditability but excluded from event-level operating metrics.

Predictor	AUROC	AUPRC	Brier	Event det.	Lead	False alerts	Precision	Recall	F1
Time	0.669	0.261	0.230	0.796	70	4.09	0.361	0.087	0.140
Loss trend	0.556	0.310	0.250	0.367	60	1.72	0.396	0.042	0.076
Update norm	0.716	0.413	0.210	0.571	80	3.32	0.500	0.124	0.199
Trace	0.574	0.249	0.239	0.367	65	5.28	0.168	0.040	0.065
Log-moment	0.860	0.542	0.155	0.918	70	3.62	0.479	0.124	0.198
Transport no-loss	0.867	0.544	0.156	0.918	60	3.26	0.505	0.124	0.200
Full transport	0.867	0.545	0.156	0.918	60	3.32	0.495	0.122	0.196
Selected TIC rule	0.868	0.546	0.157	0.959	90	6.05	0.512	0.238	0.325
Logistic all features	0.856	0.536	0.152	0.898	70	3.74	0.452	0.116	0.184
Raw tabular	0.766	0.434	0.165	0.469	70	3.86	0.261	0.051	0.086

5.2. Baselines

We compare the validation-selected TIC rule against time-only, learning-rate-only, loss-trend-only, gradient-norm-only, update-norm-only, update-to-parameter-ratio-only, Adam-mismatch-only, trace-only transport, log-moment-only transport, transport without loss-directed coordinates, full transport, raw tabular telemetry, raw tabular plus transport, and logistic all-feature models. Table 1 reports the representative rows; Table A2 gives the full baseline table. All thresholds are selected on validation under the same event-level non-maximum-suppression protocol.

5.3. Metrics

We report AUROC, AUPRC, Brier score, event detection rate, median lead time in optimizer steps, false alerts per 1000 stable optimizer steps, precision, recall, and F1. Event detection is event-level detection: an event is counted as detected if an alert occurs in its valid lead window after non-maximum suppression. Precision, recall, and F1 are row- or alert-level operating metrics and are therefore not identical to event detection. When a metric is undefined, it is left blank rather than forcing a value. The dense results are fixed-protocol pilot evidence, not deployment calibration.

6. Dense Prediction Results

The selected TIC rule detects 95.9% of the 49 evaluable held-out spike events $(47/49)$, with Wilson 95% interval 0.863–0.989, median lead time 90 optimizer steps, and 6.05 realized test false alerts per 1000 stable optimizer steps. It obtains the highest AUROC and AUPRC among the reported predictors, but the improvement over the log-moment plane is small: AUPRC is 0.546 for selected TIC rule and 0.542 for log-moment only. The log-moment rule detects 45/49 events, with Wilson 95% interval 0.808–0.968. The strongest scalar baseline is update norm, with AUPRC 0.413 and event detection 0.571 at 3.32 false alerts per 1000 stable optimizer steps.

The main empirical conclusion is therefore not that every

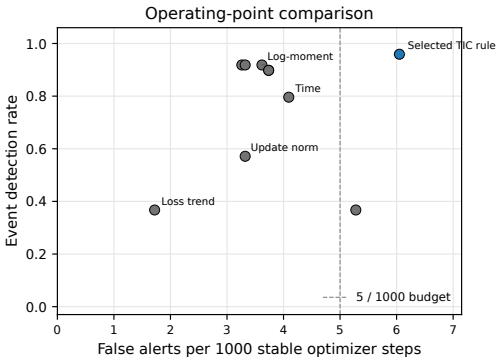

Figure 2. Dense operating-point comparison. Event detection is event-level detection on 49 evaluable held-out spike events; false alerts are realized held-out test rates per 1000 stable optimizer steps. The dashed line in the figure marks the validation selection target, not a constraint on realized test false-alert rates.

TIC coordinate contributes equally. Trace transport alone is weak. The log-moment plane is the dominant predictive component, showing that the important signal is not only the scalar drift-to-dispersion ratio, but the joint expansion of total transport and diffusion trace. The full and selected risk-rule variants provide the strongest overall operating point, especially for event-level detection and lead time, at the cost of a higher realized test false-alert rate.

The dense benchmark includes three control comparisons that constrain the interpretation of Table 1. First, the time-only and learning-rate-only rows test the most direct schedule explanation. Their AUPRC values are 0.261 and 0.259, respectively, compared with 0.542 for the log-moment plane and 0.546 for the selected TIC rule. Second, update norm is not a straw baseline: it is the strongest scalar row in Table 1, with AUPRC 0.413. The remaining gap indicates that a scalar parameter-space magnitude does not capture all of the signal carried by the feature-space transport moments. Third, trace-only transport has AUPRC 0.249, so the evidence is not explained by the ratio K_{tr} alone. Within this pilot regime, the signal is concentrated in the causal-running-anchor-relative expansion of Q and $\widehat{\Gamma}$. Appendix C reports an additional frozen-rule audit on six new held-out seeds (24 evaluable events), using the already selected protocol without re-selecting thresholds, features, labels, model class, or stress parameters.

7. PolyPythias-410M Matched-Window Anatomy Audit

PolyPythias checkpoints (van der Wal et al., 2025), built around the Pythia suite (Biderman et al., 2023), are sparse, so this experiment is not a temporal prediction benchmark. It is a post-hoc checkpoint-window anatomy audit. We compare two candidate outlier windows with fourteen

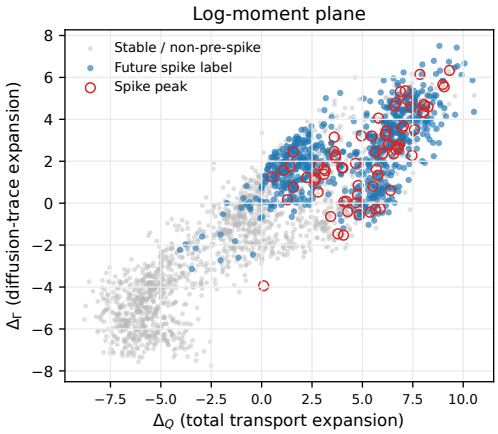

Figure 3. Log-moment transport plane. Future spike-labeled rows concentrate at high total transport expansion Δ_Q and high diffusion-trace expansion Δ_Γ. This supports the log-moment plane as the dominant predictive transport component.

Table 2. PolyPythias-410M matched-window audit summary.

Coordinate	Outlier median	Stable median	Contrast
E	4.226	0.043	+4.18 log units
K_{tr}	2.110	0.988	+1.12 log units
K_{wh}	1.054	0.286	+0.77 log units
Q	4305.47	61.04	70.5×
Γ	515.88	24.12	21.4×

matched stable windows across the same step ranges. The outlier windows are seed3, step 74000 to step 76000, and seed4, step 121000 to step 123000. Matched stable windows use seeds 1, 2, 5, 6, 7, 8, and 9 at the same step ranges. All 16 windows are processed successfully.

The outlier windows rank first and second under the transport envelope E, trace transport K_{tr}, common-mode expansion C, total transport Q, and diffusion trace Γ. For K_{wh}, seed4 ranks first and seed3 ranks fourth; the median outlier value is higher by 0.77 log units than the matched-stable median. As a descriptive rank sanity check, if the two outlier labels were placed uniformly at random among the 16 matched windows, the probability that both occupy the top two ranks for a coordinate is $1/\binom{16}{2} = 0.0083$. For K_{wh}, where the outliers rank first and fourth, the probability that both fall in the top four is $\binom{4}{2}/\binom{16}{2} = 0.05$. These calculations do not treat coordinates as independent tests. The audit supports an anatomical interpretation of TIC: known outlier checkpoint windows are not only high-loss windows, but windows with much larger feature-space transport expansion and diffusion than matched stable windows.

This audit does not establish lead-time prediction on PolyPythias. The checkpoint spacing is too coarse for that claim. Its role is to test whether the same coordinates that support dense prediction also decompose larger-scale 410M outlier windows into interpretable transport components.

8. Limitations

The dense benchmark is a controlled small-Transformer setting under critical learning-rate stress. It does not show universal prediction across all LLM training runs. The final dense result is still pilot-level: it has 49 evaluable test spike events and uses a validation-selected operating rule. It is not a deployment-calibrated alarm benchmark.

The stress schedule is artificial by design. It is useful for producing finite non-NaN spikes under a fixed protocol, but the result should not be read as evidence that the same thresholds or false-alert rates transfer to ordinary large-scale training runs. The selected TIC rule has the strongest event-detection operating point, but its realized test false-alert rate is higher than several transport ablations. The log-moment plane carries most of the predictive signal; the remaining coordinates are better interpreted as anatomical and modestly incremental in this pilot.

PolyPythias-410M is used only for matched-window anatomy, not temporal prediction. Ghost measurement is expensive, and scaling to frontier-size models will require fewer measurement points, more aggressive projection, amortized ghost evaluation, or selective triggering. Spike labels depend on the probe loss, robust baseline window, refractory rule, and finite-spike criterion. Strong scalar or tabular baselines may remain competitive in other regimes. TIC's value is therefore both predictive and anatomical, not a universal alarm guarantee.

9. Conclusion

Transport Instability Coordinates provide a feature-space view of stochastic optimizer transitions. In the dense benchmark, the log-moment plane carries most of the predictive signal for future finite probe-loss spikes, while the selected TIC rule gives the strongest event-level operating point. The matched PolyPythias-410M audit shows that the same coordinates separate larger-scale candidate outlier checkpoint windows from matched stable windows, supporting TIC as both a predictive and anatomical coordinate system for training instability.

References

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and van der Wal, O. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, 1986.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1310–1318, 2013.

Rousseeuw, P. J. and Croux, C. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

van der Wal, O., Lesci, P., Muller-Eberstein, M., Saphra, N., Schoelkopf, H., Zuidema, W., and Biderman, S. Polypythias: Stability and outliers across fifty language model pre-training runs. In *International Conference on Learning Representations*, 2025.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

A. Coordinate Identities and Local Motivation

A.1. Frozen Ballistic-Diffusive Identity

These calculations are not intended as a generative model of training dynamics. They justify the coordinate choices under local frozen-state approximations; the empirical claims are evaluated by the held-out dense benchmark and the PolyPythias matched-window audit.

Suppose local feature increments satisfy

$$X_s = b + \xi_s, \tag{29}$$

where b is a fixed drift vector, $\mathbb{E}[\xi_s] = 0$, $\mathrm{Cov}(\xi_s) = \Lambda$, and the ξ_s are uncorrelated across steps. For lag Δ, define

$$Y_\Delta = \sum_{s=1}^{\Delta} X_s. \tag{30}$$

Then

$$\mathbb{E}\|Y_\Delta\|^2 = \Delta^2\|b\|^2 + \Delta\ \mathrm{tr}(\Lambda). \tag{31}$$

Indeed, Y_Δ has mean Δb and covariance $\Delta\Lambda$, so the identity $\mathbb{E}\|Y\|^2 = \|\mathbb{E}Y\|^2 + \mathrm{tr}(\mathrm{Cov}(Y))$ gives the result. Dividing by Δ, the per-step mean-square transport is

$$S(\Delta) = \Delta\|b\|^2 + \mathrm{tr}(\Lambda). \tag{32}$$

The corresponding drift-to-diffusion coefficient is

$$\log\left(\frac{S(\Delta)}{\mathrm{tr}(\Lambda)}\right) = \log\left(1 + \frac{\Delta\|b\|^2}{\mathrm{tr}(\Lambda)}\right). \tag{33}$$

At $\Delta = 1$, under the empirical identification $B \approx \|b\|^2$, $\Gamma \approx \mathrm{tr}(\Lambda)$, and $Q \approx B + \Gamma$, this motivates the one-step empirical coefficient $\log(Q/\Gamma)$. It also explains why $\log(B/\Gamma)$ is not the correct one-step total transport coefficient: the one-step numerator contains both coherent drift energy and diffusion trace.

A.2. Loss-Directed Transport

Let a_t be the feature-space probe-loss gradient. Locally,

$$\delta\ell \approx a_t^\top X_t. \tag{34}$$

If X_t is Gaussian with mean μ_t and covariance Λ_t, then

$$a_t^\top X_t \sim \mathrm{Normal}(a_t^\top \mu_t, a_t^\top \Lambda_t a_t). \tag{35}$$

The relevant directional signal-to-noise ratio for loss increase is signed and loss-directed. This motivates K_ℓ: it counts transport as risky only when ghost candidate updates raise probe loss on average. This argument is a coordinate motivation, not a theorem about neural training dynamics.

A.3. Non-identifiability from Parameter-Space Update Norms

TIC coordinates are not functions of parameter-space update norm alone. For a small candidate update u, a first-order expansion of the probe representation gives

$$X(u) = J_t u + O(\|u\|^2), \tag{36}$$

where J_t is the local representation Jacobian at row t. If $J_t^\top J_t$ is not a scalar multiple of the identity, then there exist updates u and v with $\|u\| = \|v\|$ but $\|J_t u\| \neq \|J_t v\|$. Thus equal parameter-space update norms can induce different representation-space transports.

For a random candidate update U, the leading term in total transport is

$$Q_t \approx \mathbb{E}\|J_t U\|^2 = \mathrm{tr}\big(J_t^\top J_t\, \mathbb{E}[UU^\top]\big), \tag{37}$$

whereas the parameter-space update energy is

$$\mathbb{E}\|U\|^2 = \mathrm{tr}\big(\mathbb{E}[UU^\top]\big). \tag{38}$$

Update norm remains an informative scalar baseline, but it is not a substitute for representation-space transport.

A.4. Why the Log-Moment Plane Retains Information Lost by Ratios

The trace coefficient

$$K_{\mathrm{tr}} = \log(Q + \epsilon) - \log(\Gamma + \epsilon) \tag{39}$$

is invariant, away from the numerical floors, to common multiplicative expansion $(Q, \Gamma) \mapsto (aQ, a\Gamma)$. It therefore cannot distinguish quiet states from states in which total transport and diffusion both expand by the same factor.

The log-moment plane keeps this common-mode information by comparing each moment to a causal running anchor:

$$\Delta_Q = \log(Q + \epsilon) - \log(Q_{\mathrm{base}} + \epsilon), \tag{40}$$

$$\Delta_\Gamma = \log(\Gamma + \epsilon) - \log(\Gamma_{\mathrm{base}} + \epsilon). \tag{41}$$

The derived coordinates

$$C = \frac{\Delta_Q + \Delta_\Gamma}{2}, \qquad R = \Delta_Q - \Delta_\Gamma \tag{42}$$

separate common-mode expansion from relative imbalance. This explains why trace-only transport is a weak ablation and why the log-moment plane is the main predictive component in the dense benchmark.

B. Notation Summary

C. Additional Dense Baselines and Protocol Details

The dense run uses total steps 1800, measurement interval 10, ghost count 4, projection dimension 128, train

Table A1. Notation summary.

Symbol	Meaning
θ_t	model parameters at measurement row t
$\mathcal{F}_t$	training filtration up to row t
$\mathcal{G}_t$	row-t ghost microbatch randomness
Z_j	ghost microbatch
$u_t^*(Z_j)$	AdamW-style frozen-moment update
$f(\theta)$	fixed probe representation
$X_{j,t}$	ghost feature increment
μ_t	mean ghost feature increment
Q_t	total second moment
$\widehat{\Gamma}_t$	sample-corrected diffusion trace
D_t	denominator-k empirical dispersion
B_t	coherent drift energy
K_{tr}	trace transport coefficient
K_{wh}	whitened transport coefficient
K_ℓ	loss-directed transport
E, C, R	log-moment transport-plane coordinates
M	modal concentration
$p_t^{(H)}$	predicted probability of future spike within horizon H

seeds 0–3, validation seeds 4–5, and test seeds 6–17. The Transformer has hidden dimension 192, three layers, four heads, sequence length 128, dropout 0, and byte vocabulary size 256. AdamW uses decoupled weight decay 0.01. The frozen stress schedule is a linear learning-rate ramp from 0.03 to 0.30 with base ramp steps 300–900 and seed-dependent start jitter in $\{-100, -50, 0, 50, 100\}$. The selected operating rule has threshold 0.60347 and non-maximum-suppression window 20 optimizer steps. Spike labeling used $\delta_{\mathrm{abs}} = 0.05$, $\kappa = 4.0$, $W_b = 80$ measurement rows, $L_{\mathrm{gap}} = 2$, minimum baseline points 30, local-maximum radius 3, and refractory window 10 measurement rows. The predictor-blind event-search grid was $\delta_{\mathrm{abs}} \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$ and $\kappa \in \{3, 4, 5, 6\}$.

Additional frozen-rule held-out audit. As an appendix-only audit, we evaluated the completed frozen protocol on six new held-out seeds, 18–23. The audit used real WikiText-2, produced 24 raw finite spike peaks, and yielded 24 evaluable events. The original final metrics are unchanged. The new seeds were evaluation-only: no threshold, feature, label, model class, stress schedule, or validation selection was re-selected.

The audit preserves the main ranking pattern: log-moment and full-transport coordinates remain above time, learning-rate, update-norm, and trace-only controls in AUPRC. The selected TIC rule again gives the highest event-level detection, but at a higher realized false-alert rate.

Table A2. Full held-out dense baseline table. All rows use validation-selected thresholds and the same event-level non-maximum-suppression protocol.

Predictor	AUROC	AUPRC	Brier	Event det.	Lead	False alerts
Time	0.669	0.261	0.230	0.796	70	4.09
Learning rate	0.666	0.259	0.246	0.714	70	3.86
Loss trend	0.556	0.310	0.250	0.367	60	1.72
Gradient norm	0.369	0.173	0.252	0.245	65	5.34
Update norm	0.716	0.413	0.210	0.571	80	3.32
Update/parameter	0.734	0.388	0.193	0.469	50	3.14
Adam mismatch	0.531	0.192	0.251	0.000	–	0.00
Trace	0.574	0.249	0.239	0.367	65	5.28
Log-moment	0.860	0.542	0.155	0.918	70	3.62
Transport no-loss	0.867	0.544	0.156	0.918	60	3.26
Transport loss-only	0.665	0.402	0.218	0.551	80	3.74
Full transport	0.867	0.545	0.156	0.918	60	3.32
Raw tabular	0.766	0.434	0.165	0.469	70	3.86
Raw + transport	0.856	0.536	0.152	0.898	70	3.74
Selected TIC rule	0.868	0.546	0.157	0.959	90	6.05

Table A3. Additional frozen-rule held-out audit.

Predictor	AUROC	AUPRC	Brier	Event det.	Lead	False alerts
Time	0.679	0.256	0.228	0.833	70	3.99
Learning rate	0.673	0.247	0.246	0.750	70	3.76
Update norm	0.729	0.423	0.204	0.583	95	3.40
Trace	0.601	0.262	0.226	0.500	45	3.05
Log-moment	0.872	0.570	0.152	0.792	70	4.23
Full transport	0.882	0.577	0.173	0.917	80	4.11
Selected TIC rule	0.882	0.571	0.176	0.958	90	8.33

D. Basic TIC Identities and Protocol Invariants

D.1. Filtration Measurability of TIC Coordinates

Let $\mathcal{F}_t$ denote the training filtration immediately before the row-t ghost measurements. It contains the current parameters θ_t, the AdamW optimizer state, scheduler state, all losses and gradients observed at or before row t, the fixed probe batch, the fixed probe representation map f, and all preprocessing quantities fitted from the training split. Let $\mathcal{G}_t$ denote the sigma-field generated by the row-t ghost microbatch draws and any row-t randomness used only for those ghost evaluations. Then every TIC coordinate in Φ_t is $\mathcal{F}_t \vee \mathcal{G}_t$-measurable.

Indeed, conditional on $\mathcal{F}_t \vee \mathcal{G}_t$, the frozen model state θ_t, AdamW state, scheduler state, probe representation map, and ghost microbatches $Z_1, \ldots, Z_k$ are known. For each ghost microbatch Z_j, the ghost gradient and the AdamW-style frozen-moment transition $u_t^*(Z_j)$ are deterministic functions of the frozen state and Z_j. The real optimizer state is not mutated. Therefore the ghost feature increment

$$X_{j,t} = f(\theta_t + u_t^*(Z_j)) - f(\theta_t) \tag{43}$$

is $\mathcal{F}_t \vee \mathcal{G}_t$-measurable. The empirical mean μ_t, total transport Q_t, sample-corrected diffusion trace $\widehat{\Gamma}_t$, denominator-k dispersion D_t, coherent drift energy B_t, trace coefficient $K_{\mathrm{tr},t}$, log-moment coordinates E_t, C_t, R_t, whitened transport coefficient $K_{\mathrm{wh},t}$, modal concentration M_t, and loss-directed coordinate $K_{\ell,t}$ are deterministic functions of the

row-t ghost increments, row-t probe-loss evaluations, and train-fitted numerical floors or anchors. With a fixed deterministic convention for eigensystem tie-breaking in the low-rank covariance computation, these are all measurable functions of $\mathcal{F}_t \vee \mathcal{G}_t$. Hence Φ_t uses no future losses, future gradients, future spike labels, or held-out test-threshold information.

D.2. Empirical Drift-Diffusion Decomposition

For ghost increments $X_1, \ldots, X_k$, define

$$\mu = \frac{1}{k}\sum_{j=1}^{k} X_j, \tag{44}$$

$$Q = \frac{1}{k}\sum_{j=1}^{k} \|X_j\|^2, \tag{45}$$

$$\widehat{\Gamma} = \frac{1}{k-1}\sum_{j=1}^{k} \|X_j - \mu\|^2, \tag{46}$$

$$D = \frac{k-1}{k}\widehat{\Gamma}, \tag{47}$$

$$B = \|\mu\|^2. \tag{48}$$

Then $Q = D + B$.

To prove this, expand each increment around the empirical mean:

$$X_j = (X_j - \mu) + \mu. \tag{49}$$

Then

$$\sum_j \|X_j\|^2 = \sum_j \|X_j - \mu + \mu\|^2 \tag{50}$$

$$= \sum_j \|X_j - \mu\|^2 + 2\left\langle \sum_j (X_j - \mu), \mu \right\rangle + k\|\mu\|^2. \tag{51}$$

The cross term vanishes because $\sum_j (X_j - \mu) = 0$. Therefore

$$\frac{1}{k}\sum_j \|X_j\|^2 = \frac{1}{k}\sum_j \|X_j - \mu\|^2 + \|\mu\|^2 = D + B. \tag{52}$$

This identity is exact for the measured empirical ghost distribution; it is not a modeling approximation. The reported diffusion coordinate uses the sample-corrected trace $\widehat{\Gamma}$; the exact pathwise decomposition uses $D = (k-1)\widehat{\Gamma}/k$.

D.3. Conditional Moments and Finite-k Convention

The pathwise identity above does not require distributional assumptions. For interpretation, suppose additionally that, conditional on $\mathcal{F}_t$, the ghost increments $X_{1,t}, \ldots, X_{k,t}$ are iid with finite second moment, conditional mean $\mu_t^\star$, and conditional covariance Σ_t. Suppressing the row index and writing $Y_j = X_j - \mu^\star$, $\bar{Y} = k^{-1}\sum_{j=1}^{k} Y_j$, we have $\mu = \mu^\star + \bar{Y}$. Then

$$\mathbb{E}[Q \mid \mathcal{F}_t] = \mathrm{tr}(\Sigma_t) + \|\mu_t^\star\|_2^2, \tag{53}$$

$$\mathbb{E}[D \mid \mathcal{F}_t] = \left(1 - \frac{1}{k}\right)\mathrm{tr}(\Sigma_t), \tag{54}$$

$$\mathbb{E}[\widehat{\Gamma} \mid \mathcal{F}_t] = \mathrm{tr}(\Sigma_t), \tag{55}$$

$$\mathbb{E}[B \mid \mathcal{F}_t] = \|\mu_t^\star\|_2^2 + \frac{1}{k}\,\mathrm{tr}(\Sigma_t). \tag{56}$$

Indeed, $\mathbb{E}\|X_1\|^2 = \|\mu^\star\|^2 + \mathrm{tr}(\Sigma_t)$. Also

$$D = \frac{1}{k}\sum_{j=1}^{k} \|Y_j - \bar{Y}\|^2 = \frac{1}{k}\sum_{j=1}^{k} \|Y_j\|^2 - \|\bar{Y}\|^2, \tag{57}$$

and $\mathbb{E}\|\bar{Y}\|^2 = \mathrm{tr}(\Sigma_t)/k$. The formula for $B = \|\mu^\star + \bar{Y}\|^2$ follows from $\mathbb{E}[\bar{Y} \mid \mathcal{F}_t] = 0$. Thus the sample-corrected diffusion trace $\widehat{\Gamma} = kD/(k-1)$ is the usual unbiased covariance-trace estimator for iid ghost draws with $k > 1$, while D is the empirical dispersion entering the exact identity $Q = D + B$. The experiments report $\widehat{\Gamma}$ as the diffusion coordinate and use the corresponding causal running anchors in the log-moment plane.

D.4. Rank of the Empirical Ghost Covariance

Let

$$\widehat{\Sigma}_D = \frac{1}{k}\sum_{j=1}^{k}(X_j - \mu)(X_j - \mu)^\top = \frac{1}{k}AA^\top, \tag{58}$$

where $A = [X_1 - \mu, \ldots, X_k - \mu]$. Since $A\mathbf{1}_k = 0$, the centered columns are linearly dependent and $\mathrm{rank}(A) \le k-1$. Hence $\mathrm{rank}(\widehat{\Sigma}_D) \le k-1$. This is why TIC uses a resolved low-rank whitening subspace rather than a full covariance inverse; for $k = 4$, at most three nonzero empirical covariance directions can be resolved at a measurement row.

D.5. Scale Behavior of Transport Coordinates

Consider a common rescaling of the probe representation by a constant $a \neq 0$. That is, suppose the feature increments are transformed as

$$\widetilde{X}_j = aX_j. \tag{59}$$

Then the empirical mean rescales as

$$\widetilde{\mu} = \frac{1}{k}\sum_j \widetilde{X}_j = a\mu. \tag{60}$$

Consequently,

$$\widetilde{Q} = a^2 Q, \qquad \widetilde{\widehat{\Gamma}} = a^2\widehat{\Gamma}, \qquad \widetilde{D} = a^2 D, \qquad \widetilde{B} = a^2 B. \tag{61}$$

Therefore, away from the ϵ-dominated numerical floor, the trace coefficient is invariant to common representation rescaling:

$$\widetilde{K}_{\mathrm{tr}} = \log\left(\frac{\widetilde{Q}}{\widetilde{\widehat{\Gamma}}}\right) = \log\left(\frac{a^2 Q}{a^2 \widehat{\Gamma}}\right) = \log\left(\frac{Q}{\widehat{\Gamma}}\right) = K_{\mathrm{tr}}. \tag{62}$$

With the stabilizer included, exact invariance holds if ϵ is scaled consistently with representation units; otherwise the statement holds outside the floor-dominated regime.

The same argument applies to the log-moment coordinates. If the causal running anchors are computed in the same representation scale, then

$$\widetilde{Q}_{\mathrm{base}} = a^2 Q_{\mathrm{base}}, \qquad \widetilde{\widehat{\Gamma}}_{\mathrm{base}} = a^2 \widehat{\Gamma}_{\mathrm{base}}. \tag{63}$$

Ignoring the floor-dominated regime, or scaling ϵ consistently,

$$\widetilde{\Delta}_Q = \log(\widetilde{Q}) - \log(\widetilde{Q}_{\mathrm{base}}) = \log(a^2 Q) - \log(a^2 Q_{\mathrm{base}}) = \Delta_Q, \tag{64}$$

and similarly $\widetilde{\Delta}_\Gamma = \Delta_\Gamma$. Hence C, R, and E are invariant to common representation rescaling under the same convention. The log-moment plane is therefore anchor-relative: it measures expansion relative to a causal stable reference scale rather than the arbitrary units of the probe representation.

D.6. Modal Concentration Bounds

Let

$$s_i = \frac{(e_i^\top \mu)^2}{\lambda_i + \lambda_{\mathrm{reg}}}, \qquad S = \sum_{i=1}^{r} s_i. \tag{65}$$

If $S \le \epsilon_{\mathrm{mode}}$, TIC sets $M = 0$, assigning zero concentration when there is no resolved whitened drift to concentrate. If $S > \epsilon_{\mathrm{mode}}$, then $\pi_i = s_i/S$ is a probability vector. For $r > 1$,

$$M = 1 - \frac{H(\pi)}{\log r}, \qquad H(\pi) = -\sum_{i=1}^{r} \pi_i \log \pi_i. \tag{66}$$

Since $0 \le H(\pi) \le \log r$, it follows that $M \in [0, 1]$. Uniformly spread resolved whitened drift gives $H(\pi) = \log r$ and $M = 0$; concentration in a single mode gives $H(\pi) = 0$ and $M = 1$. The zero-drift convention avoids interpreting vanishing resolved drift as maximal concentration.

D.7. Frozen Evaluation Protocol

Let P denote the complete protocol consisting of the stress schedule, label definition, seed split, feature construction, model class, threshold grid, non-maximum suppression rule, and validation selection criterion. The event-search stage does not use the final held-out test seeds 6–17, test predictor scores, or test operating metrics. After event-search, P is fixed. The fitted predictor and alerting rule are functions only of the training data, validation data, and P:

$$\widehat{r} = \widehat{r}(D_{\mathrm{train}}, D_{\mathrm{val}}; P). \tag{67}$$

On the held-out test split, the reported metrics have the form

$$M(\widehat{r}(X_{\mathrm{test}}), Y_{\mathrm{test}}), \tag{68}$$

where X_{test} are test features and Y_{test} are test labels. Because Y_{test} is not an argument to $\widehat{r}$ or to P, held-out test labels affect only the final reported metrics. They do not affect the benchmark definition, feature set, model class, thresholds, operating rule, or validation selection criterion.

E. Additional Figures

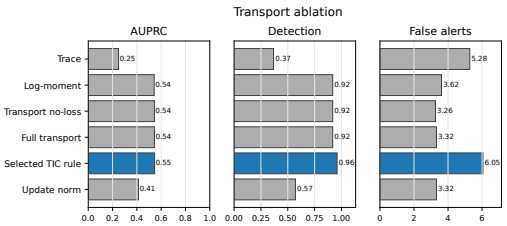

Figure A1. Transport ablation. AUPRC and event detection are higher-is-better; false alerts are lower-is-better. The figure shows the transition from trace-only transport to log-moment and full transport variants.

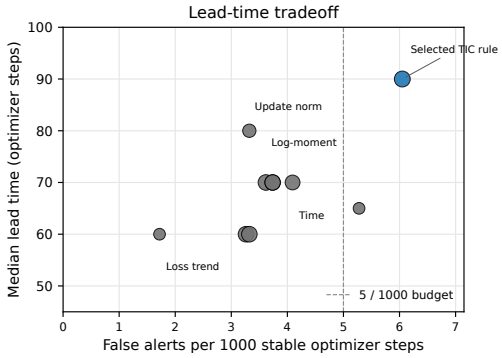

Figure A2. Lead-time tradeoff at validation-selected operating points; point size reflects test event detection rate.

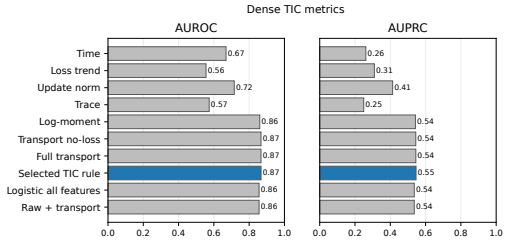

Figure A3. Dense TIC metrics summary.

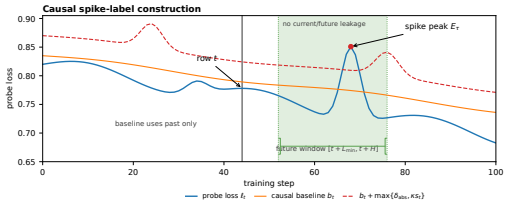

Figure A4. Causal spike-label construction.