SAME-CHECKPOINT INTERVENTION ON PRE- AND POST-CONTEXT STATES IN ENTITY-LOCAL TEMPORAL ENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Relational structure can shape a temporal encoder during training or be mixed directly into its deployed representation. We introduce a same-checkpoint diagnostic that separates these roles by exposing entity-local pre-context z^{pre} and message-passed post-context z^{post} states from the same trained system, and use a nonstationary equity panel as a stress test. Under a historical residual-correlation adjacency that ranks neighbors by $|\rho|$ and discards sign, switching retrieval from z^{pre} to z^{post} reduces rank correlation on 19/21 forward-state descriptors in the three-seed mean; 15 are negative in every seed and none is positive in every seed, with losses of 0.086–0.153 on eight highlighted descriptors. A sweep-selected, two-seed exploratory endpoint suggests that stronger graph-conditioned training can reallocate the unmixed state toward volatility/range geometry while degrading market-factor geometry, but the absence of a content-free shuffled-graph control and a third seed prevents attribution to relational edge content. The main result is therefore a controlled negative finding and a reusable diagnostic principle: training-time use of structure and inference-time relational propagation should be evaluated separately.

1 INTRODUCTION

Representation learning for temporal entities often faces two kinds of structure simultaneously. Each entity has a private history, while entities may also coexist in a relational environment. A common response is to place a graph neural network (GNN) on top of entity encoders, so that relations become part of the representation itself. This is attractive because message passing supplies an explicit relational inductive bias. It is also risky: if the downstream task depends on entity-specific state, neighborhood aggregation can erase precisely the idiosyncratic geometry that the temporal encoder was built to preserve.

We treat the encoder/graph boundary as a reusable diagnostic: compare states immediately before and after propagation from the *same checkpoint*. We use *entity-local* for a deployed state computed from one entity's history without cross-entity message passing. This separates inference-time mixing from graph-conditioned training; equities are a stress test for this general encoder–relation question, not the object of study itself.

This paper asks a simple question with a consequential distinction:

> *When relational context is informative, must it be mixed into the inference representation, or can it be more useful as a training signal that shapes otherwise entity-specific temporal embeddings?*

We study this question using daily equity histories as a stress test for nonstationary relational representation learning. The setting is deliberately difficult rather than economically complete: 499 present-day U.S. large-cap constituents, 12 historical features, and 4,149 aligned sessions. The universe is survivorship-biased and is not a point-in-time investable index; we make no trading-profitability claim. The scientific object is instead the geometry of temporal states. Given a history ending at date t, a useful embedding should place assets near one another when their *subsequent* 21-session regimes—volatility, range, drawdown, market-relative risk, and related descriptors—are

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.

similar. Future descriptors are used only for evaluation. They are never inputs to the representation or context losses.

Our architecture exposes two representations from the same trained system. A causal patch Transformer maps each asset history to a pre-context state $z_{i,t}^{\mathrm{pre}}$. A training-time graph branch constructs a historical residual-correlation graph and produces $z_{i,t}^{\mathrm{post}}$ through GCN or GAT message passing. The graph-neighbor objective is computed on the post-context branch, but its gradients flow into the temporal encoder. This creates a controlled separation between *context as representation* (z^{post} at inference) and *context as supervision* (train with the relational objective, deploy z^{pre}). Figure 1 summarizes this distinction.

The strongest result is paired: under the sign-agnostic historical residual-correlation adjacency, $z^{\mathrm{post}} - z^{\mathrm{pre}}$ is negative in mean on 19/21 descriptors across three graph-trained seeds; 15/21 are negative in every seed, 0/21 positive in every seed, and eight highlighted losses span 0.086–0.153 rank IC. Exploratorily, a seed-0 $\lambda_{\mathrm{ctx}} \in \{0, .01, .05, .15, .50\}$ sweep reallocates geometry toward volatility/range descriptors and away from market-factor descriptors, and a second $\lambda_{\mathrm{ctx}} = .50$ checkpoint reproduces the same four positive family-wise survivors. The endpoint remains sweep-selected, lacks a third seed and content-free graph, and therefore does not identify relational edge content. The key missing controls are sign-respecting propagation for E10 and a shuffled graph plus third $\lambda_{\mathrm{ctx}} = .50$ seed for the training-branch claim. Thus the two roles are empirically separable in this testbed.

Our contributions are deliberately specific:

- We introduce a reusable **same-checkpoint diagnostic** that exposes z^{pre} and z^{post} from one jointly trained system, isolating relational propagation at inference from graph-conditioned training without changing the encoder, checkpoint, or training trajectory.
- We pair that intervention with chronological retrieval, 21-date moving-block Romano–Wolf correction, synthetic-null calibration, and an explicit claim audit that separates supported, exploratory, and untested conclusions (Table 7).
- Under the sign-agnostic residual-correlation adjacency, three residual-GCN seeds show 19/21 negative mean deltas, 15/21 negative in every seed, 0/21 positive in every seed, and 0.086–0.153 rank-IC penalties on eight highlighted descriptors.
- Exploratorily, a seed-0 weight sweep suggests geometry reallocation toward a volatility-magnitude factor and a second $\lambda_{\mathrm{ctx}} = .50$ seed reproduces the same four positive family-wise survivors; confirmation requires the already-identified graph/content controls and fuller replication.

Table 7 audits claim status; Appendix H addresses likely objections.

2 Related Work

Time-series representation learning. Temporal Neighborhood Coding (TNC) learns nonstationary temporal states by discriminating local neighborhoods from distant windows (Tonekaboni et al., 2021). TS2Vec develops hierarchical contrastive representations over augmented contexts (Yue et al., 2022), while SoftCLT replaces hard pair identities with soft temporal and instance similarities (Lee et al., 2024). SimTS predicts a future latent representation from the past without negative pairs (Zheng et al., 2023), and TimeSiam pretrains Siamese encoders through past-to-current reconstruction (Dong et al., 2024). PatchTST demonstrates the value of patch-token temporal Transformers and channel-independent processing (Nie et al., 2023). TimeAlign explicitly aligns representations of past and future distributions (Hu et al., 2026). Our representation objective is closest in spirit to TNC: the base temporal loss uses history-only temporal neighborhoods, with a historical volatility-regime weighting. The contribution here is not a new universal time-series pretraining benchmark; it is the controlled role assigned to cross-entity relations.

Retrieval and predictive geometry. Retrieval-augmented forecasting methods such as RAFT retrieve similar historical patterns and use their observed continuations to support forecasting (Han et al., 2025). Our evaluation also uses retrieval, but asks a representation question: does neighborhood geometry at time t organize future *state descriptors* cross-sectionally? This is intentionally broader than point prediction and intentionally weaker than a claim of optimal forecasting.

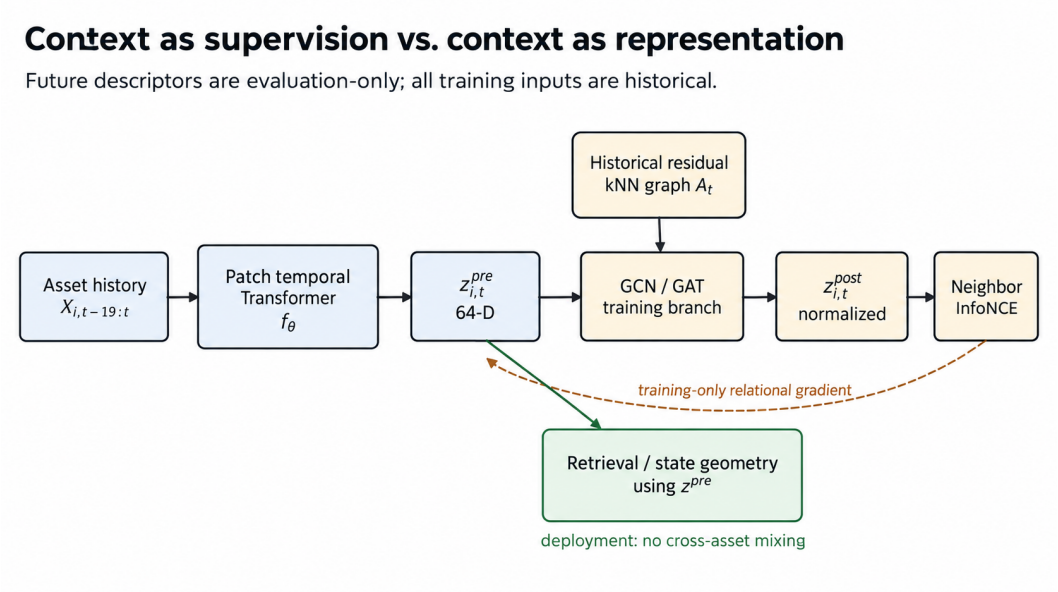

Figure 1: **Context as supervision versus context as representation.** The causal temporal encoder yields the entity-specific state z^{pre}. Historical cross-asset relations feed a graph-conditioned message-passing branch and graph-neighbor InfoNCE; gradients from that branch shape the temporal encoder, while deployment may retrieve either the pre- or post-context state. At deployment, the graph-free state z^{pre} can be retrieved without cross-asset mixing. Future descriptors are evaluation-only.

Relational learning in markets. Prior financial models exploit company relations through temporal relational ranking (Feng et al., 2019) and hierarchical graph attention (Kim et al., 2019). Other work constructs stock embeddings to transfer information across companies (Yi et al., 2022). These results motivate relational inductive bias but do not imply that message-passed states are universally preferable to entity-specific ones. Our study isolates that architectural choice under a history-only relation graph and compares the same trained system before and after context mixing.

Graph-supervised learning without inference-time message passing. The broad idea of using graph structure during training without requiring message passing at inference is established. GraphMLP uses adjacency-derived neighboring contrastive supervision with an MLP and does not require adjacency at test time (Hu et al., 2021); GSSC similarly treats structure as prior knowledge for constructing supervision signals rather than as a forward-propagation operator (Wu et al., 2025). A complementary line distills graph-aware teachers into graph-free students: GLNN (Zhang et al., 2022), Cold Brew (Zheng et al., 2022), and NOSMOG (Tian et al., 2023). We therefore do *not* claim graph-derived supervision without inference-time message passing as novel. Our contribution is a same-checkpoint intervention in a jointly trained temporal system: the shared encoder exposes states immediately before and after relational propagation, allowing us to measure the inference-time effect of propagation while holding the encoder, checkpoint, and training trajectory fixed, and separately to study how a graph-conditioned auxiliary objective alters the deployed entity-local representation. Distillation asks whether a separately trained graph-free student can match a graph-aware teacher; our diagnostic asks whether propagation itself helps or hurts relative to the *same model's own* pre-propagation state.

When graph mixing fails. GCN aggregation is closely related to Laplacian smoothing and can oversmooth node states (Li et al., 2018); repeated graph propagation can lose expressive power (Oono & Suzuki, 2020); and low-homophily settings can favor architectures that preserve ego and neighbor information separately (Zhu et al., 2020). We test oversmoothing-related explanations, but our evidence does not support a single scalar collapse metric as a complete account: changing graph type and adding projectors can substantially alter pairwise cosine structure without making inference-time graph mixing outperform the pre-context state.

3 PROBLEM SETUP

We observe N entities across time. For entity i at date t, let

$$X_{i,t} = [x_{i,t-L+1}, \ldots, x_{i,t}] \in \mathbb{R}^{F \times L} \tag{1}$$

be a strictly historical window. The temporal encoder produces an entity-specific state

$$z_{i,t}^{\mathrm{pre}} = f_\theta(X_{i,t}) \in \mathbb{R}^d. \tag{2}$$

A historical relation constructor $\mathcal{G}$ maps information available no later than t to an adjacency A_t. A context encoder g_ϕ may then form

$$z_{i,t}^{\mathrm{post}} = g_\phi(z_{1:N,t}^{\mathrm{pre}}, A_t)_i. \tag{3}$$

The central design choice is whether retrieval uses z^{post} (context is inside the deployed representation) or z^{pre} after training with a relational auxiliary loss (context is supervision only at inference).

Evaluation geometry. Let $y_{i,t}^m$ denote forward descriptor m measured over the next H sessions. For each query date, a representation retrieves k historical-bank states by cosine distance and predicts each descriptor from the corresponding historical outcomes. We summarize quality with date-wise cross-sectional Spearman rank correlation

$$\mathrm{RIC}_t^m = \mathrm{Spearman}\left(\widehat{y}_{1:N,t}^m, y_{1:N,t}^m\right), \qquad \overline{\mathrm{RIC}}^m = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \mathrm{RIC}_t^m\,. \tag{4}$$

Rank IC is not a claim about portfolio return. It tests whether representation neighborhoods organize entities according to their subsequent relative states on the same date. This makes it suitable for our scientific question and less sensitive to a market-wide level shift than pooled correlation.

4 METHOD

4.1 CAUSAL PATCH TEMPORAL ENCODER

The final controlled cohort uses $L = 20$ daily observations with $F = 12$ features. Per-channel normalization statistics are estimated on training bars and stored with the encoder. The time axis is split into patches of length and stride 8; incomplete trailing patches are right-padded. Each flattened patch is projected into a $d_{model} = 128$ token. A learned CLS token and positional embedding feed a 4-layer, 8-head Transformer with feed-forward width 512. The CLS output is projected to $d = 64$ to produce z^{pre}. This "base" temporal encoder is shared by the context-trained and context-free arms.

The 12 inputs consist of causally scaled OHLC features; one-day return; relative volume; log high–low range; log open–close gap; return relative to SPY; trailing realized volatility; SPY return; and trailing SPY realized volatility. The feature pipeline includes future-return arrays for legacy tasks, but the representation and context objective contracts explicitly reject any batch carrying future-return keys.

4.2 HISTORY-ONLY TEMPORAL REPRESENTATION LOSS

The base representation objective is a TNC-style discriminator with a history-only volatility-regime correction ("StatioCL" in the repository). For anchor state z_a, local temporal positive z_p, and distant negative z_n, a discriminator D is optimized with binary cross-entropy. Positive targets remain one. Distant negatives are downweighted when the historical realized-volatility regimes are similar:

$$s(v_a, v_n) = \exp\left[-\left(\frac{\log(v_a/v_n)}{\tau_v}\right)^2\right], \quad \tau_v = 0.5, \tag{5}$$

$$w_n = 1 - s(v_a, v_n). \tag{6}$$

Thus the objective does not force two temporally distant windows apart merely because they occur far apart in time if their historical volatility states are similar. No forward descriptor and no cross-asset correlation enters $\mathcal{L}_{\mathrm{rep}}$.

4.3 HISTORICAL CROSS-ASSET GRAPH

For the relational branch, 48 assets are sampled first. Residualization and neighbor construction are then performed *within that sampled cross-section*: the graph uses the preceding 120 sessions and $k = 8$ nearest neighbors by absolute correlation of market-residual returns, and the market proxy is the equal-weight return of those sampled assets rather than the full 499-name universe. Denote asset returns by $r_{i,s}$, the minibatch market proxy by $\bar{r}_s$, and their within-lookback means by μ_i and μ_m. The implementation centers each series, estimates an asset-specific loading on the centered minibatch market proxy, and projects that factor out:

$$\beta_i = \frac{\sum_s (r_{i,s} - \mu_i)(\bar{r}_s - \mu_m)}{\sum_s (\bar{r}_s - \mu_m)^2}, \tag{7}$$

$$\widetilde{r}_{i,s} = (r_{i,s} - \mu_i) - \beta_i(\bar{r}_s - \mu_m). \tag{8}$$

The adjacency retains the top-k neighbors by $|\rho_{ij}|$, is symmetrized, and includes self loops. The sign of ρ_{ij} is discarded, so strongly positive and strongly negative residual correlations both become nonnegative edges. This construction was introduced after the original percentile-correlation graph exhibited substantial isolation and degree alignment with the first market principal component; it was designed to reduce edge selection driven solely by the common market mode, not to model signed relations. Unsigned $|\rho|$ graphs are established in gene co-expression analysis (Langfelder & Horvath, 2008) and appear as financial-graph baselines (Baaijens et al., 2026); these precedents motivate, but do not validate, discarding sign here. A sign-respecting control is therefore an important unresolved test.

The primary context module is a two-layer GCN with hidden width 128 and dropout 0.2. An attention variant is evaluated as a mechanistic ablation. Let $G_\phi(z^{\mathrm{pre}}, A_t)$ be the graph output. The post-context state is a learned residual blend,

$$z^{\mathrm{post}} = \mathrm{LayerNorm}(\alpha z^{\mathrm{pre}} + (1-\alpha) G_\phi(z^{\mathrm{pre}}, A_t)), \quad \alpha = \sigma(a), \tag{9}$$

where a is initialized so that $\alpha = 0.5$.

4.4 CROSS-ASSET SUPERVISION

Cross-asset supervision is implemented as graph-neighbor InfoNCE on z^{post}. Embeddings are ℓ_2 normalized before similarities are computed; this normalization is essential because the earlier unnormalized implementation saturated logits at temperature 0.07. For each graph batch and anchor i,

$$\mathcal{L}_{\mathrm{ctx}} = -\frac{1}{N} \sum_i \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \log \frac{\exp(\bar{z}_i^\top \bar{z}_j / \tau_c)}{\sum_{q=1}^{N} \exp(\bar{z}_i^\top \bar{z}_q / \tau_c)}, \quad \tau_c = 0.07. \tag{10}$$

Here $\mathcal{N}_i$ is determined only by the historical adjacency and self-pairs are excluded from the positive average. The final joint objective used in the matched graph cohort and weight sweep is

$$\mathcal{L} = \mathcal{L}_{\mathrm{rep}} + \lambda_{\mathrm{ctx}} \mathcal{L}_{\mathrm{ctx}}, \tag{11}$$

with no forward-prediction head. Importantly, even when deployment uses z^{pre}, gradients from $\mathcal{L}_{\mathrm{ctx}}$ flow through z^{post}, the graph branch, and back into f_θ. We therefore describe the method as *cross-asset supervision without inference-time mixing*, not as graph-free training.

4.5 TWO DEPLOYMENT VIEWS

A single trained graph model exposes two retrieval views. **Context as representation** indexes z^{post} and therefore requires a contemporaneous cross-section and adjacency at inference. **Context as supervision** indexes z^{pre}; cross-asset relations were available during training but are absent from the deployed vector. The same-checkpoint z^{pre} versus z^{post} comparison gives the clean test of inference-time propagation; a separately trained context-free arm with the same temporal backbone and optimization budget supplies the $\lambda_{\mathrm{ctx}} = 0$ endpoint for studying the training signal.

5 EXPERIMENTAL SETUP

5.1 PANEL AND CHRONOLOGY

The frozen artifact contains 499 equities, 12 features, and 4,149 trading sessions. The first 2,904 bars form the training prefix and the next 622 bars extend through validation index 3,526; a 20-session embargo separates target availability at split boundaries. Evaluation uses 120 query dates, a training retrieval bank of 20,000 states, a validation retrieval set of 12,000 states for k selection, and 24,000 test queries. The forward horizon is $H = 21$ sessions. Retrieval evaluates $k \in \{5, 20, 50\}$ on validation and freezes the selected value before test scoring.

This is a **survivorship-biased present-day constituent panel**. The ticker set is a current hardcoded large-cap universe rather than historical point-in-time index membership; delisted and removed constituents are absent. Vendor price series are adjusted for corporate actions, but the dataset is not a CRSP-style total-return accounting panel. We therefore frame finance as a relational representation testbed, not a backtest. Survivorship bias limits investability claims and population generality, but not the chronology below.

5.2 FORWARD DESCRIPTORS AND BASELINES

The evaluator contains 21 correlated forward descriptors. They include forward volatility, high–low range, mean absolute return, downside deviation, drawdown, worst day, tail fraction, gap magnitude, future relative volume, beta to SPY, correlation to SPY, idiosyncratic volatility, and relative volatility. The main text emphasizes descriptors that diagnose the central hypotheses; the full multiplicity family is retained in statistical correction.

Baselines are intentionally difficult to dismiss. `joint_demeaned` subtracts the date-wise cross-sectional mean from the learned z^{pre} at zero learned-parameter cost. `trailing_beta` is a 120-session OLS summary of beta, residual volatility, and total volatility. Additional controls include current volatility, EWMA volatility, current liquidity, volatility plus liquidity, raw flattened windows, PCA-64, a random 64-D projection, and weighted degree in the residual graph. For each forward target, corrected inference compares the learned representation against that target's strongest available baseline, not a single weak universal control.

5.3 TRAINING COHORT AND STATISTICS

The decisive graph comparison (E10) uses the base Transformer, 64-D output, 250 epochs, 150 steps per epoch, batch size 128 for temporal samples, Monte Carlo sample size 4, cosine learning-rate schedule, patience set to the full 250 epochs, and $\lambda_{\mathrm{ctx}} = .05$; Appendix Table 3 lists the matched cohort. Three independently trained graph models use seeds 0, 1, and 2; each checkpoint is scored both as z^{pre} and z^{post}, so the inference-time comparison holds training trajectory and parameters fixed. Context-free models with the same schedule were also trained at seeds 0–2; seed 0 provides the $\lambda_{\mathrm{ctx}} = 0$ endpoint in E11. The graph model has 896,513 parameters and the context-free model 879,808 (about 1.9% fewer parameters in the context-free arm). At $\lambda_{\mathrm{ctx}} = 0$, the context module is physically absent (`context_type=none`); the training loop does not construct the graph batch, so an unused context dropout path does not consume RNG draws. This avoids an RNG-consumption confound, but the zero endpoint is not capacity-matched to the positive-λ_{ctx} arms. We report mean $\pm$ sample SD across the three paired graph checkpoints.

The auxiliary-weight sweep (E11) holds the seed and schedule fixed and evaluates $\{0, .01, .05, .15, .50\}$. It is a dose-response analysis, not a multi-seed estimate. We therefore separate two types of uncertainty: training-seed variability for the matched graph comparison and date-sampling uncertainty for the fixed-checkpoint target comparisons.

Because the 21 forward targets are correlated, a wall of unadjusted p-values would be misleading. For each target, the frozen evaluator selects the strongest available comparator by test-set rank IC, then forms a date-wise rank-IC difference against that comparator. It resamples query dates using a 21-session moving-block bootstrap with 2,000 resamples and applies Romano–Wolf stepdown jointly across the target family (Romano & Wolf, 2005). The common resample preserves cross-target dependence. We call a target a positive family-wise survivor only if its learned-minus-comparator

Table 1: Same-checkpoint comparison of post-context z^{post} and pre-context z^{pre} retrieval for three independently trained residual-GCN models. Values are cross-sectional rank IC, mean ± SD across three paired seeds; $\Delta = z^{post} - z^{pre}$. Eight highlighted descriptors spanning volatility/range, idiosyncratic-risk, drawdown, and market-factor families are shown; Appendix Table 4 reports all 21 targets.

Target	Pre-context	Post-context	Δ
HL range	0.672 ± 0.002	0.518 ± 0.035	-0.153
Mean —ret—	0.574 ± 0.003	0.433 ± 0.031	-0.141
Abs. gap	0.569 ± 0.002	0.434 ± 0.032	-0.135
Volatility	0.532 ± 0.002	0.398 ± 0.027	-0.135
Idio. vol	0.400 ± 0.003	0.285 ± 0.021	-0.116
Relative vol	0.514 ± 0.033	0.376 ± 0.034	-0.139
Max drawdown	0.297 ± 0.009	0.199 ± 0.015	-0.098
Beta	0.370 ± 0.011	0.284 ± 0.013	-0.086

delta is positive and its adjusted $p_{FW} < .05$. The correction is conditional on the selected comparator and checkpoint; it does not cover test-set comparator selection.

6 RESULTS

6.1 INFERENCE-TIME GRAPH PROPAGATION DEGRADES MOST FORWARD-STATE DESCRIPTORS

Table 1 is the highest-confidence empirical result. Within the same three graph-trained checkpoints, the message-passed representation z^{post} is worse than its own pre-context state z^{pre} on all eight displayed descriptors. The losses are not marginal: -0.153 rank IC for forward high–low range, -0.141 for mean absolute return, -0.135 for forward volatility, and -0.116 for idiosyncratic volatility. Because both views come from the same checkpoint, this isolates inference-time mixing rather than a different training run. Each seed supplies its own paired control, though three seeds still give limited across-training uncertainty.

Figure 2 plots eight highlighted paired penalties. Across the full 21-target family, the three-seed mean delta is negative on 19 targets and positive on two; 15 targets are negative in all three independently trained checkpoints, whereas no target is positive in all three. Appendix Table 4 reports every target and all seed-level deltas. We therefore avoid the broader phrase "consistently harmful." The result is still important because both a naive positive narrative ("relations should help related assets") and our repository-defined cross-asset-target hypothesis fail: in earlier controlled variants, graph gains were not larger for beta/correlation/relative-volatility targets than for single-asset targets. The effect also persists after replacing the original common-factor-dominated adjacency with residual and partial-correlation alternatives.

6.2 NO SINGLE COLLAPSE DIAGNOSTIC EXPLAINS THE PENALTY

Several obvious mechanisms were tested. The original unnormalized graph InfoNCE did create a numerical pathology: with $\tau_c = .07$, post-context dot products generated extremely large logits and effective rank collapsed. ℓ_2 normalization fixed the scale. Adding VICReg-style variance/covariance regularization (Bardes et al., 2022) increased effective rank in a diagnostic arm without producing wins over the control. Changing from fixed-weight GCN propagation to GAT reduced mean pairwise cosine similarity of z^{post} from 0.959 in a residual-GCN diagnostic to 0.770 in the GAT diagnostic, yet post-context forward-volatility rank IC still deteriorated sharply. Adding a projector can rescue a catastrophic GAT post-context result back near pre-context performance, but in the tracked single-seed diagnostic it does not make graph mixing better than the pre-context state.

These ablations argue against a simplistic explanation such as "the graph only fails because every embedding is identical." They do not establish a unique mechanism. A more conservative interpretation is that the relevant future geometry is partly entity-specific and that relation-based aggregation contracts or rotates it in directions misaligned with the evaluation descriptors. Appendix D reports the diagnostic geometry.

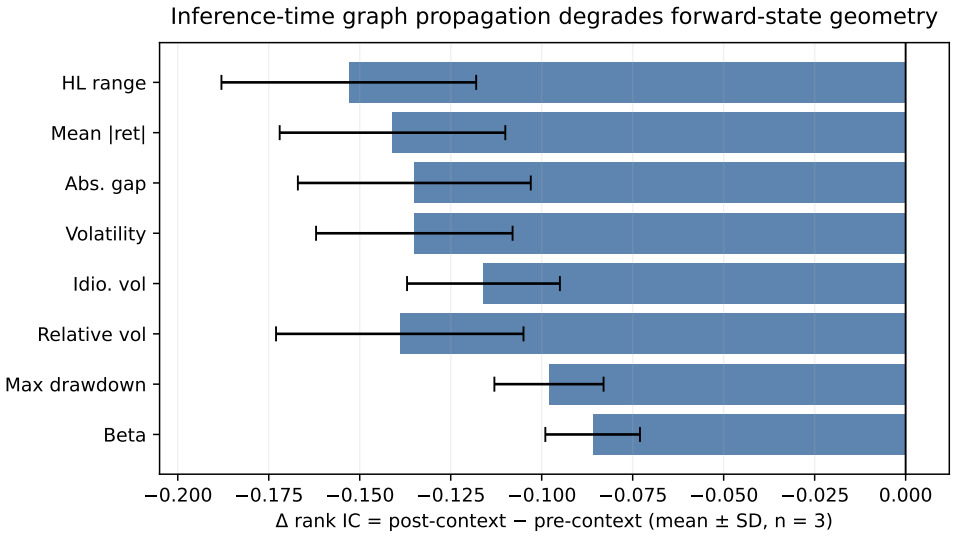

Figure 2: Post-context minus pre-context rank IC for eight highlighted descriptors in the matched E10 cohort. Error bars are sample SD of the paired seed differences ($n = 3$), not date-bootstrap confidence intervals. Across all 21 targets, 15 deltas are negative in every seed and none is positive in every seed; all seed-level values appear in Appendix Table 4.

Table 2: Exploratory seed-0 trajectory of the deployed pre-context embedding z^{pre} as auxiliary-branch weight varies. "+FWER" counts positive target deltas that survive Romano–Wolf correction against each target's strongest evaluated comparator over the target family. These adjusted values are conditional on each checkpoint and selected comparator; they do not correct test-set comparator selection or selection across the five λ_{ctx} values.

λ_{ctx}	Vol.	HL range	Mean $\lvert r\rvert$	Mkt. corr.	+FWER
0.00	0.526	0.666	0.567	+0.063	0
0.01	0.532	0.671	0.575	+0.030	0
0.05	0.533	0.673	0.576	-0.014	0
0.15	0.544	0.680	0.586	-0.050	0
0.50	0.548	0.687	0.593	-0.051	4

6.3 Exploratory evidence that the auxiliary branch reallocates z^{pre} geometry

The seed-0 E11 sweep is evidence about representation reallocation, not causal identification of graph content: as λ_{ctx} rises, several volatility/range descriptors improve while market-correlation geometry deteriorates. At $\lambda_{\text{ctx}} = .50$, high–low range, mean absolute return, absolute gap, and realized volatility beat their strongest evaluated comparators after Romano–Wolf correction, and an independent seed-1 checkpoint reproduces the same four survivors. We treat them as one correlated volatility-magnitude factor, not four independent discoveries; no target-level correlation matrix is available.

The endpoint remains exploratory: only two seeds are complete, it was selected after the seed-0 sweep, and no degree-matched shuffled/random graph was trained. Thus the data do not distinguish meaningful relation content from generic auxiliary-branch regularization.

Appendix Figure 4 visualizes the five-point seed-0 trajectory.

Appendix Table 6 reports both completed $\lambda_{\text{ctx}} = .50$ seeds: all four volatility-magnitude targets remain positive family-wise survivors in each, idiosyncratic volatility is positive but non-significant, and beta/market correlation remain strongly negative. These tests quantify query-date uncertainty conditional on each checkpoint and do not correct either sweep selection or test-set comparator selection.

6.4 SIMPLE FACTORS REMAIN THE RIGHT REPRESENTATION FOR SIMPLE FACTORS

The strongest baseline result is a useful constraint on the paper's scope. A 120-session rolling beta/residual-volatility summary reaches rank IC 0.522 on forward beta and 0.582 on forward market correlation. At $\lambda_{\text{ctx}} = .50$, the learned pre-context representation trails the strongest baseline by 0.153 and 0.630 rank IC on those two targets, respectively, with family-wise rejection in the negative direction. This is not an embarrassment to hide: it reveals that a representation can improve geometry for nonlinear volatility-regime descriptors while being inferior to a low-dimensional sufficient statistic for a known factor. Universal embedding quality should not be inferred from a single family of financial labels.

7 DISCUSSION

E10 shows that **inference-time propagation under the sign-agnostic historical residual-correlation construction used here** degrades most forward-state descriptors. The seed-0 sweep plus a second $\lambda_{\text{ctx}} = .50$ seed suggests z^{pre} geometry can be reallocated, but edge semantics remain unidentified without a content-free graph; PCA market-mode alignment and cross-domain generalization are also untested.

This negative result has independent methodological value: a plausible graph augmentation can damage the deployed state once propagation is isolated from its training signal. Unlike a graph-versus-no-graph model comparison, the same-checkpoint intervention separates changed optimization from the inference operator itself. The finding is domain-specific; the paired diagnostic, family-wise inference, and claim audit are reusable across jointly trained encoder–relation systems.

8 LIMITATIONS

E10 is limited to one survivorship-biased panel and one sign-agnostic residual-correlation construction: after 48-asset sampling, $|\rho|$ makes anti-correlated pairs positive neighbors. The choice was designed to reduce market-mode-driven selection and has unsigned-network precedents (Langfelder & Horvath, 2008; Baaijens et al., 2026), but a sign-respecting control is the highest-priority graph-construction test. A code audit also finds 48-name training/bank graphs versus up-to-200-name validation/test query graphs for z^{post}, plus historical bank states left at initialized zero when fewer than 120 prior sessions exist; the exact E10 affected share was not recomputed (Appendix B). Same-ticker histories remain eligible; exclusion retrieval, external representation baselines, and a second domain remain untested.

The exploratory training-branch claim has separate limitations: the two-seed $\lambda_{\text{ctx}} = .50$ endpoint is sweep-selected, lacks a content-free graph, and uses test-selected comparators, so its target-family correction is conditional on comparator selection. The $\lambda_{\text{ctx}} = 0$ model is about 1.9% smaller and the four positive targets are correlated. A shuffled graph and third $\lambda_{\text{ctx}} = .50$ seed therefore remain the key confirmation controls.

9 CONCLUSION

Training-time and inference-time roles of structure should be evaluated separately. Under the sign-agnostic historical residual-correlation adjacency used here, message passing degrades most retrieval descriptors in a paired three-seed comparison. A seed-0 sweep and independent second $\lambda_{\text{ctx}} = .50$ seed support exploratory geometry reallocation, but incomplete replication, sweep selection, test-set comparator selection, and no content-free graph prevent attribution to relational edge content. The same-checkpoint protocol turns the negative result into a reusable diagnostic: *separate what may shape optimization from what should be mixed into the deployed representation.*

REPRODUCIBILITY STATEMENT

The submission package contains the exact LaTeX source, generated tables, and figure files derived from the tracked evaluation artifacts. The headline graph result uses paired pre/post-context scores

from the same three stored E10 graph checkpoints, including the full 21-target seed-level audit in Appendix Table 4; the λ_{ctx} trajectory uses the stored seed-0 sweep, and the high-weight endpoint additionally uses the independently trained E13 seed-1 checkpoint. The accompanying refactored model interface independently enforces the representation contract used by the paper: z^{pre} is defined before context, the correlation adjacency is an explicit input, and the joint trainer passes only (z^{post}, A_t) to the context loss so that graph-loss gradients can reach the temporal encoder without redefining z^{pre}. Appendix B specifies data splits, model hyperparameters, objectives, retrieval, and statistical correction. We explicitly distinguish the replicated three-seed E10 result, the two-seed but sweep-selected high-weight endpoint, and fixed-checkpoint date-resampling inference.

Ethics Statement

The study uses market data and does not involve human subjects. The equity universe is survivorship-biased and is unsuitable for claims about historical investability or realized trading profitability; we state this limitation in the main text. The work could be misread as investment advice if representation metrics were equated with returns. We make no such claim and report no live or simulated profit metric.

AI Use Statement

Generative AI tools were used solely to polish and refine existing manuscript writing for clarity, grammar, and presentation. The authors remain fully responsible for all scientific content and the final manuscript.

References

Tein Baaijens, Çiçek Güven, and Gonzalo Nápoles. Graph learning on financial networks from pairwise similarity of firm-level characteristics. *Applied Network Science*, 11(1):31, 2026. doi: 10.1007/s41109-025-00755-2.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Jiaxiang Dong, Haixu Wu, Yuxuan Wang, Yun-Zhong Qiu, Li Zhang, Jianmin Wang, and Mingsheng Long. TimeSiam: A pre-training framework for siamese time-series modeling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11412–11436, 2024.

Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems*, 37(2):1–30, 2019. doi: 10.1145/3309547.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan O. Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21774–21797, 2025.

Yang Hu, Haoxuan You, Zhecan Wang, Zhicheng Wang, Erjin Zhou, and Yue Gao. Graph-MLP: Node classification without message passing in graph. *arXiv preprint arXiv:2106.04051*, 2021.

Yifan Hu, Jie Yang, Tian Zhou, Peiyuan Liu, Yujin Tang, Rong Jin, and Liang Sun. Bridging past and future: Distribution-aware alignment for time series forecasting. In *International Conference on Learning Representations*, 2026.

Raehyun Kim, Chan Ho So, Minbyul Jeong, Sanghoon Lee, Jinkyu Kim, and Jaewoo Kang. HATS: A hierarchical graph attention network for stock movement prediction. *arXiv preprint arXiv:1908.07999*, 2019.

Peter Langfelder and Steve Horvath. WGCNA: An r package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008. doi: 10.1186/1471-2105-9-559.

Seunghan Lee, Taeyoung Park, and Kibok Lee. Soft contrastive learning for time series. In *International Conference on Learning Representations*, 2024.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020.

Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005. doi: 10.1111/j.1468-0262.2005.00615.x.

Yijun Tian, Chuxu Zhang, Zhichun Guo, Xiangliang Zhang, and Nitesh V. Chawla. Learning mlps on graphs: A unified view of effectiveness, robustness, and efficiency. In *International Conference on Learning Representations*, 2023.

Sana Tonekaboni, Danny Eytan, and Anna Goldenberg. Unsupervised representation learning for time series with temporal neighborhood coding. In *International Conference on Learning Representations*, 2021.

Lirong Wu, Haitao Lin, Guojiang Zhao, Cheng Tan, and Stan Z. Li. Learning to model graph structural information on mlps via graph structure self-contrasting. *IEEE Transactions on Neural Networks and Learning Systems*, 36(7):12772–12783, 2025. doi: 10.1109/TNNLS.2024.3458405.

Ziruo Yi, Ting Xiao, Kaz-Onyeakazi Ijeoma, Ratnam Cheran, Yuvraj Baweja, and Phillip Nelson. Stock2Vec: An embedding to improve predictive models for companies. *arXiv preprint arXiv:2201.11290*, 2022.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. TS2Vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Shichang Zhang, Yozen Liu, Yizhou Sun, and Neil Shah. Graph-less neural networks: Teaching old mlps new tricks via distillation. In *International Conference on Learning Representations*, 2022.

Wenqing Zheng, Edward W. Huang, Nikhil Rao, Sumeet Katariya, Zhangyang Wang, and Karthik Subbian. Cold brew: Distilling graph node representations with incomplete or missing neighborhoods. In *International Conference on Learning Representations*, 2022.

Xiaochen Zheng, Xingyu Chen, Manuel Schurch, Amina Mollaysa, Ahmed Allam, and Michael Krauthammer. Simple contrastive representation learning for time series forecasting. *arXiv preprint arXiv:2303.18205*, 2023.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

A Repository audit and experiment genealogy

The repository contains two research generations: an older future-outcome-geometry scaffold on public time-series benchmarks and a newer finance experiment program centered on temporal encoders, context graphs, and cross-asset training losses. We did not merge their claims. The paper is based on the newer program because it has completed tracked evaluations, while the older public-benchmark tables are explicitly marked unexecuted in their scaffold.

The strongest comparable cohort is E10: matched 250-epoch graph and context-free training with seeds 0–2. The headline inference test is paired within each graph checkpoint, comparing its z^{pre}

and z^{post} views. E11 reuses that schedule and seed 0 to vary λ_{ctx}. E9 supplies graph-construction and mechanistic ablations. Early E1 evaluations are excluded from pooled comparison because they predate critical fixes: post-context evaluation was not guaranteed to run, context InfoNCE was unnormalized, and important shallow controls were absent. E5/E8 contain useful development evidence but are not needed for the headline matched result. E13 supplies one completed independent $\lambda_{\mathrm{ctx}} = .50$ replication checkpoint (seed 1), which is reported only as partial high-weight replication; seed 2 is not present in the supplied tracked slice.

B Exact protocol and information-flow audit

B.1 Feature availability

At representation time t, the encoder window contains only bars at or before t. OHLC channels use a 20-session causal min–max transform. Training normalization statistics are computed on the training prefix and stored in the model. Macro alignment in the data pipeline uses backward as-of merges with exact matches disallowed where macro features are enabled; the final E10/E11 inputs summarized here use the 12 daily channels listed in the main text.

Future-return arrays are stored in the processed bundle for other tasks, so leakage prevention cannot rely on file absence. Instead, the representation-objective and context-objective implementations declare forbidden future keys and raise if such keys enter their batches. The joint trainer passes only (z^{post}, A_t) to the context loss. The final E10/E11 objective has no prediction loss, so no future labels participate in optimization.

B.2 Split and target availability

The panel has 4,149 aligned bars. The stored E10/E11 split indices are train end 2,904 and validation end 3,526 with a 20-session embargo. Forward descriptors use $H = 21$. Retrieval banks enforce that both history and target horizons lie inside the appropriate earlier fold. Test query futures are used only after representation extraction, neighbor selection, and k selection have been frozen.

B.3 Retrieval

For each representation, cosine-nearest historical states are drawn from a frozen training bank sampled over eligible asset-date states. Candidate neighborhood sizes $k \in \{5, 20, 50\}$ are scored on a validation retrieval set; one k is then used for the test set. Neighbor outcomes are aggregated by an unweighted mean. The stored evaluation reports 20,000 bank states, 12,000 validation states, 24,000 test queries over 120 query dates (up to 200 live assets per date). In the frozen E10/E11 evaluation path, same-ticker historical states are not explicitly excluded; no same-ticker-excluded rerun is claimed here. For z^{post}, graph construction is cross-section dependent: training and bank encodings use 48-name graphs, while validation and test query dates can construct the residual graph over all 200 requested names on that date. In addition, `encode_joint_post_context` initializes its output to zero and leaves a requested state unchanged when the date has fewer than 120 prior sessions for adjacency construction; the exact fraction of E10 bank anchors affected by this behavior was not recomputed. Neither implementation detail affects z^{pre}; both are disclosed rather than corrected post hoc. Rank IC is computed within each date and averaged over dates.

C Training hyperparameters

The graph model contains 896,513 trainable parameters and the no-context model 879,808. The best checkpoint is selected along each completed training trajectory using the repository's validation objective; the schedule itself is not shortened by patience in E10/E11. A code audit additionally shows that the adjacency constructor inserts a unit self-loop and the GCN layer adds identity again before row normalization, so the propagated matrix has self weight 2 prior to normalization; this is the behavior of the stored checkpoints and is left unchanged.

Table 3: Hyperparameters of the matched E10/E11 cohort.

Component	Value
History window	20 sessions
Input features	12
Patch length / stride	8 / 8
Transformer	$d_{model} = 128$, 8 heads, 4 layers
Feed-forward width	512
Embedding dimension	64
Context batch assets	48
Residual-correlation lookback	120 sessions
Residual kNN degree	8
Context GCN	2 layers, hidden 128, dropout .2
Context temperature	.07, normalized embeddings
StatioCL volatility temperature	.5
Epochs / steps per epoch	250 / 150
Temporal batch size	128
Monte Carlo temporal samples	4
Learning-rate schedule	cosine
Early-stop patience	250 (full schedule)
E10 context weight	.05
E11 sweep	0, .01, .05, .15, .50

Table 4: Full E10 same-checkpoint audit over all 21 forward descriptors. Each entry is $\Delta = \mathrm{RIC}(z^{\mathrm{post}}) - \mathrm{RIC}(z^{\mathrm{pre}})$ for one independently trained residual-GCN checkpoint. Negative means inference-time propagation worsens retrieval geometry. Nineteen of 21 target means are negative; 15 targets are negative in all three seeds and none is positive in all three.

Target	Seed 0	Seed 1	Seed 2	Mean
HL range	-0.191	-0.151	-0.118	-0.153
Mean $\lvert r\rvert$	-0.173	-0.141	-0.111	-0.141
Relative vol.	-0.174	-0.109	-0.133	-0.139
Abs. gap	-0.171	-0.129	-0.105	-0.135
Volatility	-0.162	-0.132	-0.110	-0.135
Downside dev.	-0.152	-0.111	-0.102	-0.122
Idio. vol.	-0.142	-0.107	-0.097	-0.116
Worst day	-0.131	-0.104	-0.097	-0.111
Max drawdown	-0.111	-0.091	-0.093	-0.098
Beta	-0.088	-0.081	-0.088	-0.086
Tail frac.	-0.069	-0.067	-0.076	-0.071
Log-vol change	-0.083	-0.070	-0.058	-0.071
Relative volume	-0.038	-0.033	-0.026	-0.032
Log-volume change	-0.037	-0.019	-0.032	-0.029
Return	-0.007	-0.003	-0.033	-0.014
Up fraction	-0.023	+0.005	-0.019	-0.012
Time underwater	-0.015	+0.013	-0.027	-0.010
Volume spike	-0.006	-0.018	+0.009	-0.005
Skew	+0.004	-0.007	-0.009	-0.004
Recovered	+0.004	+0.017	-0.006	+0.005
Market corr.	+0.036	-0.008	+0.033	+0.020

D MECHANISTIC DIAGNOSTICS

The geometry diagnostics are single-seed ablations and should not be read as replicated causal identification. Their value is negative: interventions that substantially change pairwise cosine similarity or effective rank do not systematically convert post-context retrieval into a win. The residual-graph weighted-degree baseline is also mildly anti-predictive for key volatility/range descriptors in the tracked analysis, arguing against the explanation that the GNN merely learns a useful degree statistic.

Table 5: Mechanistic ablations from single-seed diagnostic arms. Cosine is mean pairwise cosine similarity; effective rank is out of 64. These diagnostics show that no single collapse/oversmoothing statistic explains the graph penalty.

Arm	z^{pre} cos.	z^{pre} rank	z^{post} cos.	z^{post} rank
partial GCN	.750	4.76	.939	13.32
partial attention	.746	3.85	.993	4.91
partial GCN + projector	.446	4.21	.449	4.52
residual GCN	.668	1.07	.959	7.75
residual GAT	.911	1.83	.770	2.94
residual GAT + projector	.767	1.70	.785	2.22

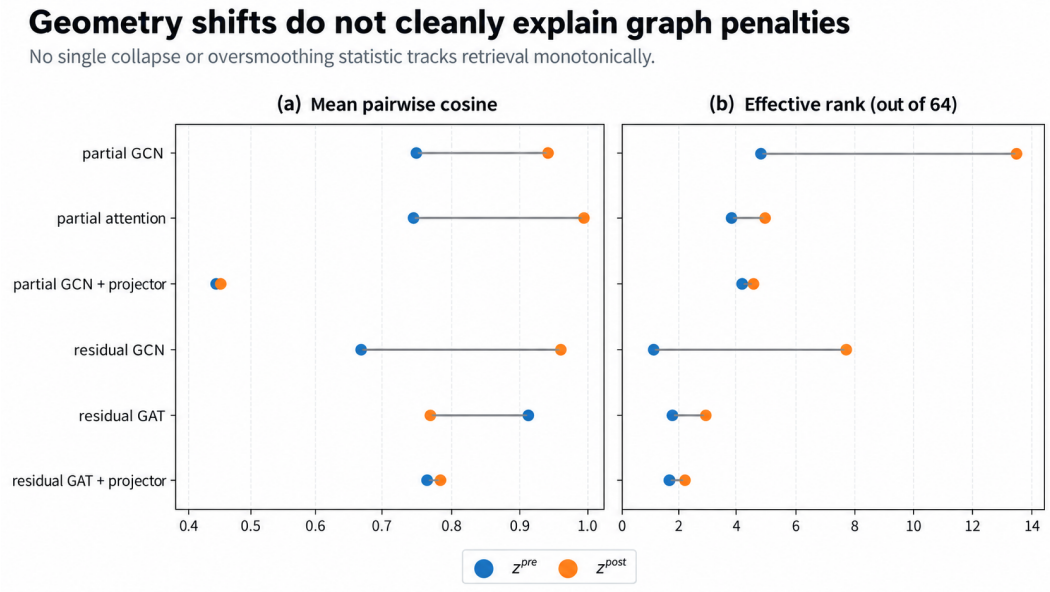

Figure 3: Paired geometry diagnostics for the single-seed mechanistic arms. Large changes in mean pairwise cosine and effective rank do not align monotonically with retrieval performance, so simple oversmoothing/collapse statistics are insufficient explanations of the graph penalty.

E MULTIPLE TESTING AND INFERENTIAL UNIT

For a fixed checkpoint and target m, the test statistic is the mean over query dates of the date-wise rank-IC difference relative to the strongest comparison baseline. Because adjacent query dates overlap in their 21-session future windows, IID query resampling is inappropriate. The evaluator therefore uses a moving block length of 21 dates and 2,000 bootstrap draws. Romano–Wolf stepdown resamples the entire target family with the same draw, preserving dependence among volatility, range, drawdown, and related outcomes. This controls family-wise error more appropriately than separately thresholding 21 nominal p-values, while retaining dependence information that Bonferroni discards (Romano & Wolf, 2005).

This inferential procedure addresses date-sampling uncertainty conditional on a trained checkpoint. It does *not* turn one training run into a multi-seed result. It controls multiplicity across targets only: the frozen evaluator first chooses each target's strongest comparator by test-set rank IC, so the adjusted values are conditional on that data-dependent comparator and do not correct comparator selection. Because $\lambda_{\text{ctx}} = .50$ was also highlighted after inspecting a five-point seed-0 sweep, they likewise do not correct sweep-level selection. The independent seed-1 endpoint reproduces the same four positive family-wise survivors, but this post-sweep replication does not retroactively remove selection or supply a third training seed. In repository-level synthetic-null calibration, 21 targets over 120 synthetic dates were generated as 0.7 times a shared Gaussian component plus 0.7 times target-specific Gaussian noise. Across 60 trials with 500 bootstrap draws per trial, naive uncorrected

Table 6: Exploratory two-seed $\lambda_{ctx} = 0.50$ comparison with the strongest evaluated comparator for each target. Entries are learned-minus-baseline rank-IC deltas with Romano–Wolf family-wise adjusted p_{FW} in parentheses, computed across targets conditional on each checkpoint. The endpoint was selected after the seed-0 sweep and each comparator is selected by test-set rank IC, so these values do not correct sweep-level or comparator-selection uncertainty.

Target	Seed 0 Δ (p_{FW})	Seed 1 Δ (p_{FW})
HL range	+0.028 (<.001)	+0.020 (<.001)
Mean $\lvert r \rvert$	+0.028 (<.001)	+0.022 (.001)
Abs. gap	+0.025 (<.001)	+0.018 (.001)
Volatility	+0.019 (.002)	+0.018 (.014)
Idio. vol.	+0.007 (.331)	+0.004 (.652)
Beta	-0.153 (<.001)	-0.156 (<.001)
Market corr.	-0.630 (<.001)	-0.633 (<.001)

target-wise testing produced at least one apparent winner in roughly 72% of trials, whereas the Romano–Wolf procedure did so in roughly 5%. The synthetic dates are IID, so this calibration checks cross-target family-wise multiplicity more directly than temporal block dependence; it motivates the stepdown procedure but does not remove the finite-sample limitation of 120 query dates and a 21-date block length. These distinctions are why the high-weight effect remains exploratory rather than confirmatory.

F HIGH-WEIGHT TARGET TRADE-OFF

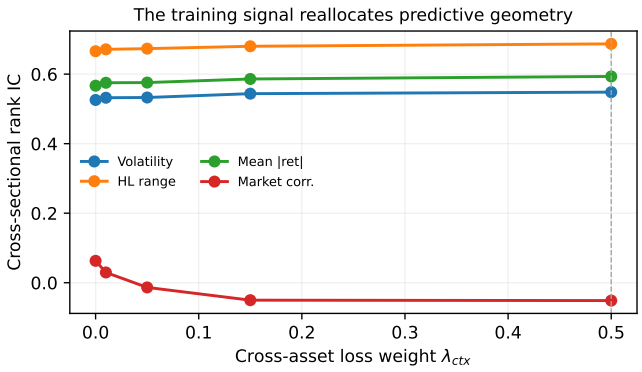

Figure 4: Seed-0 trajectory over the five evaluated λ_{ctx} values for deployed z^{pre}. Lines are visual guides, not a fitted dose response; $\lambda_{\text{ctx}} = .50$ was emphasized after inspecting the sweep, so target-family Romano–Wolf correction does not cover sweep selection.

The contrast is useful for interpreting representation capacity. The auxiliary branch does not merely increase every evaluation score; it changes which state dimensions dominate retrieval. The same four volatility-magnitude targets survive target-family correction in both completed $\lambda_{\text{ctx}} = .50$ seeds, while market-factor descriptors move in the opposite direction. These four descriptors are best interpreted as one underlying volatility-magnitude factor expressed through correlated range/return-magnitude measurements, not as four independent discoveries. A rolling regression remains the appropriate inductive bias for the market-factor targets. Whether the reallocation is caused by meaningful relation content or by branch-level regularization remains unresolved without a content-free graph control.

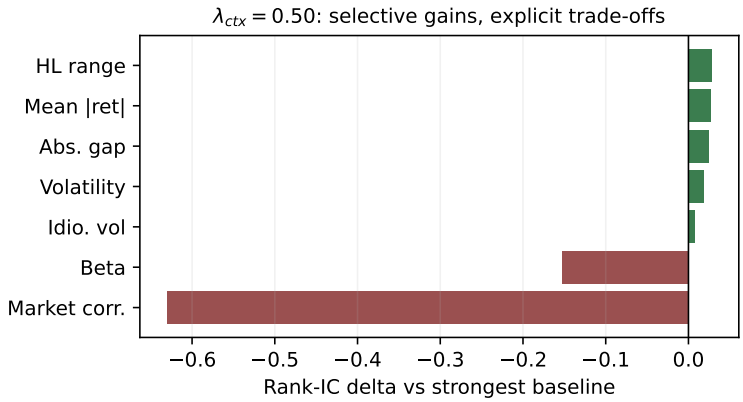

Figure 5: Exploratory seed-0 endpoint at $\lambda_{\mathrm{ctx}} = .50$. Volatility/range descriptors improve relative to their strongest evaluated comparators, while beta and market-correlation descriptors deteriorate sharply. Romano–Wolf correction is across targets conditional on this checkpoint; it does not correct test-set comparator selection or selection over the five-point λ_{ctx} sweep.

G Claim audit

Table 7: Status of headline claims in the submitted artifact. "Date-FWER" means family-wise inference over query dates for a fixed checkpoint, not training-seed replication.

Claim	Evidence	Status	Main qualification
Inference-time residual-GCN mixing degrades most forward-state descriptors	E10, 3 graph checkpoints	Supported	19/21 negative mean; 15/21 negative in all 3; 0/21 positive in all 3
Graph penalty is 0.086–0.153 on eight displayed descriptors	E10 raw eval JSONs	Supported	Paired across 3 seeds; descriptive, not target-wise FWER claim
Increasing auxiliary-branch weight reshapes z^{pre} toward volatility/range geometry	E11 seed-0 sweep + E13 seed-1 endpoint	Exploratory	Second $\lambda_{\mathrm{ctx}} = .50$ seed reproduces claim targets; no shuffled-graph control
At $\lambda_{\mathrm{ctx}} = .50$, a volatility-magnitude factor appears stronger than the strongest evaluated comparators	E11/E13, seeds 0–1, date-FWER	Exploratory	Same four survivors in both seeds; sweep- and comparator-selected; needs third seed/control
$\lambda_{\mathrm{ctx}} = .50$ is optimal	Not established	Untested	Sweep does not show a turnover
Specific relational edge content causes the high-weight gain	No content-free graph control	Untested	Random/shuffled adjacency is required
Unsigned $\lvert\rho\rvert$ propagation explains the E10 penalty	No sign-respecting control	Untested	Positive-only/signed graph required
Retrieval geometry is independent of persistent ticker identity	Same-ticker histories eligible	Untested	Exclusion rerun required
Trailing beta dominates learned states for forward beta/correlation	E10/E11 evals	Supported	Specific to those descriptors and panel
Graph failure is caused by oversmoothing alone	Diagnostics	Contradicted / unresolved	Geometry changes do not track retrieval monotonically
Results imply profitable trading	No valid trading protocol	Unsupported	Survivorship-biased universe; no cost-aware backtest
Generalizes to sensors/patients/fleets	No external-domain test	Untested	Presented only as hypothesis

H Adversarial reviewer checks

"Is this just supervised metric learning?" No forward state is supplied to the training losses. The temporal objective uses temporal neighborhoods and historical realized volatility; the context objective uses a historical adjacency. Future descriptors appear only in evaluation. The representation is therefore self-supervised/history-supervised in the operational sense used by the implementation, though the paper avoids novelty claims based on that label.

"Why use rank IC as representation quality?" The claim is explicitly about cross-sectional forward-state geometry, not universal representation quality. Rank IC tests whether retrieval neigh-

borhoods order entities by subsequent descriptors on the same date, and target-specific baselines reveal where that geometry is trivial or poorly learned.

"Is graph failure just oversmoothing?" The diagnostic data do not support that as a sufficient explanation. GAT reduces post-context pairwise cosine substantially while performing worse; projection can rescue a pathological GAT state to parity without producing a graph advantage. We therefore report oversmoothing as a candidate mechanism, not the conclusion.

"Did the graph receive an unfair schedule?" The headline E10 inference comparison is stronger than a between-model capacity comparison: each residual-GCN checkpoint is scored both before and after its own context block, so parameters and training trajectory are identical. Separately, the context-free $\lambda_{\mathrm{ctx}} = 0$ training arm uses the same base temporal encoder, 250 epochs, 150 steps/epoch, cosine schedule, split, and seeds; it is used to study the training signal rather than to define the post-context penalty.

"Does multiplicity erase the exploratory high-weight result?" At low λ_{ctx} there are no positive family-wise survivors. At the seed-0 $\lambda_{\mathrm{ctx}} = .50$ checkpoint, four correlated volatility-magnitude descriptors remain positive under the repository's 21-day moving-block Romano–Wolf procedure, and the independently trained seed-1 endpoint reproduces those same four positive survivors. This correction is across targets conditional on each checkpoint and on the comparator selected by test-set rank IC: it does not additionally correct comparator selection. Two completed seeds still do not constitute the three-seed evidence used for E10, and they do not correct the choice to emphasize $\lambda_{\mathrm{ctx}} = .50$ after inspecting a five-point seed-0 sweep.

"Could the graph penalty be an artifact of taking absolute correlation?" Yes, this remains a material alternative explanation. The E10 adjacency ranks neighbors by $|\rho|$, so anti-correlated residual histories become nonnegative edges. The magnitude-only construction was chosen to reduce market-mode-driven edge selection and has established precedents in unsigned correlation networks (Langfelder & Horvath, 2008; Baaijens et al., 2026); that motivation does not resolve whether sign should be preserved for this task. The three-seed result therefore establishes a penalty for *this residual-correlation propagation rule*, not for all graph propagation. A positive-only or signed graph is the highest-value missing control.

"Could same-ticker retrieval drive the result?" The frozen E10/E11 evaluator does not explicitly remove historical states from the query ticker. Persistent entity identity can therefore contribute to absolute retrieval quality. This is a limitation of the interpretation of the geometry scores, but it is shared exactly by z^{pre} and z^{post} within each checkpoint; it does not by itself explain the paired deterioration after propagation. A same-ticker-excluded rerun is still required for the stronger claim.

"Why is this an ICLR paper rather than a finance paper?" The central variable is where relational context enters representation learning: as inference-time message passing or as a training signal for an entity-specific temporal encoder. The finance panel supplies a difficult nonstationary relational testbed and strong factor baselines; the paper does not optimize or evaluate a trading strategy.