VARIS: RELATIONAL MARKET-STATE LEARNING FOR FINANCIAL TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series foundation models have transformed forecasting from task-specific fitting into transferable temporal representation learning, yet this progress remains centered on individual sequences. Financial time-series foundation models further learn domain-specific asset dynamics from large-scale market data, but still largely treat each asset as an independent forecasting unit, leaving cross-asset structure implicit at prediction time. We introduce VARIS, a relational framework for financial time-series foundation models that represents each prediction time as a synchronized market cross-section and explicitly infers its time-varying cross-asset structure. VARIS separates latent common market components from asset-specific residual evidence, and performs reliability-aware lead–lag inference to jointly determine whether a useful cross-asset relation exists and, conditionally, its source asset and temporal lag. A state-conditioned dynamic layer router further assigns input-dependent weights to hierarchical temporal representations for individual, common-component, and relational reasoning. Across six markets and three forecasting horizons, VARIS improves mean RankIC by 46.8% and RankICIR by 52.3% over the strongest financial time-series foundation model baseline, with consistent gains in 16 of 18 evaluation settings. Controlled ablations, temporal negative controls, and unseen-asset evaluations show that these improvements arise from common-component residualization, relational abstention, and dynamic layer routing rather than increased parameterization or broad market co-movement. VARIS advances financial time-series foundation modeling from asset-level temporal representation toward prediction-time market-structure inference.

1 INTRODUCTION

Individual asset trajectories reflect both asset-specific dynamics and the changing structure of the surrounding market. A price increase may accompany broad market movement or follow changes in a few related stocks. The predictive relevance of these observations varies across target assets and forecast dates. Financial forecasting therefore calls for joint modeling of temporal dynamics and prediction-time cross-asset structure.

Time-series foundation models provide reusable temporal representations through large-scale pretraining (Das et al., 2024; Ansari et al., 2024; Woo et al., 2024). Kronos specializes this approach to financial K-line sequences with a tokenizer and an autoregressive backbone (Shi et al., 2026). Such pretraining pools experience across assets, but an asset-level forecasting query leaves the current cross-asset structure implicit. The modeling question is *how can financial time-series foundation models infer and use the changing market structure at prediction time?*

Cross-asset predictors already model shared concepts, stock interactions, and temporally displaced evidence (Xu et al., 2021; Li et al., 2024b; Zhao & Shen, 2024; Zhou et al., 2025). This structure contains both common movement and target-specific dependencies, whose predictive roles must be distinguished. Predicting an asset's own trajectory may require different representations from assessing its relevance to another stock. A relation's apparent relevance also depends on the representation and historical evidence used to assess it.

We introduce VARIS, a financial forecasting model that learns a relational market state over synchronized, variable-size asset sets. At each forecast date, this state represents common market

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.

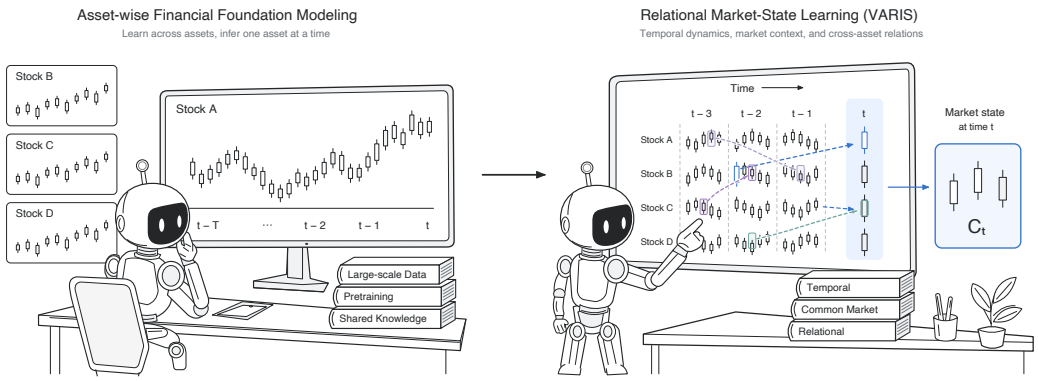


Figure 1: **From asset-wise modeling to relational market-state learning.** Left: financial foundation models learn across assets but forecast each asset from its own history. Right: VARIS jointly models temporal dynamics, common market context, and cross-asset relations for target-stock forecasting.

components and target-specific relations for stock ranking and return prediction. Its depth router selects representations for individual prediction, common-component estimation, and relation inference from a shared Kronos backbone. The relation path assesses information beyond the common component using current relevance and historical consistency. Training jointly optimizes these components and the prediction heads for target-stock forecasting.

VARIS couples representation selection with relation inference: depth weights shape the evidence used to infer relations, while relation weights determine its predictive contribution. The forecasting objective trains both decisions together, linking hierarchical temporal representations to the structure inferred at each market date.

Our evaluation spans Chinese A-shares, S&P 500 equities, and other national equity universes. Main comparisons assess VARIS against individual-asset foundation models and cross-asset predictors. Controlled ablations identify the contributions of common context, relation inference, and representation selection. Tests with unseen assets and changing source sets examine generalization beyond the training universe. Held-out-market transfer is evaluated separately from within-market forecasting.

The paper makes the following contributions:

- We introduce VARIS to extend asset-level temporal modeling to synchronized market cross-sections through a relational market-state model.
- We develop a Stable Lead-Lag Residual Adapter (SLRA) that jointly models relation acceptance and conditional source-lag selection. It combines common-component residualization and historical consistency with a null relation.
- We couple asset- and path-conditioned depth routing with relation inference to connect hierarchical temporal representations to market-state inference. This coupling learns representation mixtures suited to individual, common-component, and relational prediction.

2 Related Work

Pretrained temporal representations. TimesFM, Chronos, and Moirai learn transferable forecasting representations (Das et al., 2024; Ansari et al., 2024; Woo et al., 2024); Kronos pretrains on financial K-line sequences (Shi et al., 2026). Chronos-2 further supports grouped series and covariates (Ansari et al., 2025). These advances motivate a strong temporal backbone, while leaving open how a forecast should infer the current market cross-section. VARIS uses frozen temporal representations to learn a prediction-time relational market state.

Cross-asset prediction and lagged evidence. HIST and factor-based stock models capture shared information (Xu et al., 2021; Duan et al., 2022; 2025); MASTER models stock interactions (Li et al., 2024b). LIFT and DeltaLag establish the value of leading indicators and dynamic lags (Zhao & Shen, 2024; Zhou et al., 2025). VARIS joins these directions: common-component residualization distinguishes pairwise evidence from market-wide movement, and historical consistency plus a null option determine whether a source-lag relation should contribute at all.

Accessing intermediate representations. Attention Residuals aggregates information across Transformer depth according to content (Kimi Team et al., 2026). VARIS instead routes snapshots from a complete frozen-backbone pass separately for individual prediction, common-component estimation, and relation inference. This makes the representation used to judge a relation depend on the asset's role and current market state, without changing backbone execution.

3 VARIS

VARIS formulates stock ranking and return prediction through a relational market state inferred from the synchronized asset cross-section. Figure 2 shows its individual, common-component, and relational paths, trained under a shared forecasting objective. We first define the market-date input and the representations used by each path. We then describe common-context estimation, inference over lagged residual relations, and their integration into the final forecast.

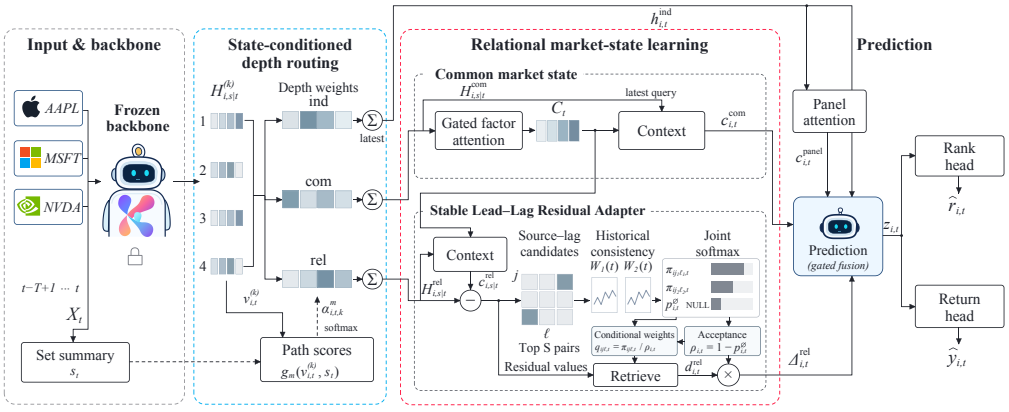


Figure 2: **VARIS mechanism.** A frozen backbone supplies layer snapshots for state-conditioned depth routing. Common market factors provide prediction context and the reference for relational residuals. The Stable Lead–Lag Residual Adapter selects source–lag pairs, assesses their historical consistency, and jointly compares them with a NULL relation. Conditional weights retrieve source residuals, while relation acceptance scales the resulting update before prediction.

3.1 Market-state inference at prediction time

At forecast date t, let $\mathcal{A}_t$ contain the assets with available input histories in one market. The input and targets are

$$X_t = \{X_{i,t-T+1:t}\}_{i\in\mathcal{A}_t}, \qquad y_{i,t}^{(h)} = p_{i,t+h}/p_{i,t} - 1, \qquad N_t = |\mathcal{A}_t|. \tag{1}$$

Each history contains open, high, low, close, volume, and amount channels. Dataset-specific adjustment and missing-field conventions accompany the input contract. An input mask defines the source universe; a separate label mask identifies assets with valid supervision. This separation preserves an observable source asset when its future label is unavailable. Each sample uses one market's trading-session axis. Parameters can be shared across markets through separate samples.

VARIS learns a latent state $Z_t = (C_t, R_t)$, where C_t is common context and R_t contains a distribution over source-lag pairs and a null relation for each target. This state is inferred from the current asset set and its histories. It determines the cross-asset correction to each asset's temporal representation. Shared encoders and masked set operations process the variable-size cross-section;

asset identifiers index observations and relations. The output is a ranking score and a return forecast for each valid target.

3.2 ASSET- AND PATH-CONDITIONED DEPTH ROUTING

Predicting an asset's return, estimating shared movement, and assessing a lagged source place different demands on its representation. VARIS therefore selects backbone depths separately for each asset and prediction path. A frozen backbone produces K layer snapshots $H^{(k)}_{i,s|t} \in \mathbb{R}^D$ from the input window ending at t. We use four depths distributed across the backbone. Path $m \in \{\mathrm{ind}, \mathrm{com}, \mathrm{rel}\}$ receives

$$\alpha^m_{i,t,k} = \mathrm{softmax}_k \, g_m(v^{(k)}_{i,t}, s_t),$$
$$H^m_{i,s|t} = \sum_{k=1}^{K} \alpha^m_{i,t,k} H^{(k)}_{i,s|t}, \qquad s \le t. \tag{2}$$

Here $v^{(k)}_{i,t}$ summarizes the valid history of asset i at depth k, and s_t is a permutation-invariant summary of the observed cross-section. The scorers g_m have separate parameters. Their weights have shape $[B, N, K]$ and apply across the history accessed at the current forecast date. The index $s \mid t$ denotes a historical position represented within the forecast information set at t, including the normalization and routing computed for that input window.

The individual path reads the latest valid representation, $h^{\mathrm{ind}}_{i,t} = H^{\mathrm{ind}}_{i,t|t}$. The common-component path supplies observations for estimating shared movement, while the relational path supplies queries, keys, and values for source-lag inference. Thus an asset can draw on different depths when making its own prediction and when supplying evidence to another asset. The backbone supplies reusable temporal features; the routers adapt their use to the current forecasting decision.

3.3 SEPARATING COMMON CONTEXT FROM RELATIONAL EVIDENCE

Shared market movement provides predictive context but can also dominate source-target similarity. VARIS gives this component its own prediction path and assesses lagged relations in the residual representation. A masked set-attention module first aggregates the latest individual representations into contemporaneous panel context $c^{\mathrm{panel}}_{i,t}$. In parallel, F learned factor queries attend to valid common-path asset-time representations. The mean of their preliminary factor states, $\bar{C}^{\mathrm{pre}}_t \in \mathbb{R}^D$, supplies a feature gate

$$g_t = D \, \mathrm{softmax}(W_g \bar{C}^{\mathrm{pre}}_t / \tau_f). \tag{3}$$

The gate rescales feature channels with mean weight one. A second factor extraction from the gated representations produces $C_t \in \mathbb{R}^{F \times D}$. Each asset-time representation queries these factors through a shared context projection. The latest common-path readout is $c^{\mathrm{com}}_{i,t} = \mathrm{Context}(H^{\mathrm{com}}_{i,t|t}, C_t)$. For the relational path, the same factor state defines

$$c^{\mathrm{rel}}_{i,s|t} = \mathrm{Context}(H^{\mathrm{rel}}_{i,s|t}, C_t),$$
$$r_{i,s|t} = H^{\mathrm{rel}}_{i,s|t} - c^{\mathrm{rel}}_{i,s|t}. \tag{4}$$

Both terms in Eq. 4 lie in the D-dimensional representation space. The shared C_t supplies common context for prediction and defines the reference for constructing residual evidence. SLRA uses these residuals for both relation scoring and retrieved values. Relation inference therefore assesses sources relative to the same common context used by the predictor.

3.4 SELECTING AND ACCEPTING STABLE SOURCE-LAG RELATIONS

SLRA answers two questions for each target: which lagged sources have useful residual evidence, and how much of that evidence should enter the forecast? It first projects the residual to a relation space of dimension d_r:

$$\tilde{r}_{i,s|t} = W_r \, \mathrm{LN}(r_{i,s|t}). \tag{5}$$

Current candidate selection. For target i, source $j \neq i$, and trading-session lag $\ell \in \mathcal{L}$, the current compatibility score is

$$u_{ij\ell,t} = \frac{Q(\tilde{r}_{i,t|t})^\top K(\tilde{r}_{j,t-\ell|t})}{\sqrt{d_r}} + b_\ell. \tag{6}$$

Candidate masks require the exact historical session and exclude unavailable observations and self-relations. We retain the S highest-scoring valid source-lag pairs, or all valid pairs when fewer are available, to form $\mathcal{S}_{i,t}$. The default configuration uses $\mathcal{L} = \{1, 2, 5\}$ and $S = 2$. Selection is over pairs, so a source can contribute at two different lags. This step concentrates the subsequent historical assessment on candidates relevant to the current target state.

Historical evidence refinement. A high current score can reflect a short-lived match. SLRA assesses each selected pair in two preceding windows $W_1(t)$ and $W_2(t)$. For window a, let $\bar{u}^{(a)}_{ij\ell,t}$ be the masked mean query-key compatibility over aligned pairs $(s, s-\ell)$. The adjusted score is

$$e_{ij\ell,t} = u_{ij\ell,t} + \eta \left[\frac{\bar{u}^{(1)}_{ij\ell,t} + \bar{u}^{(2)}_{ij\ell,t}}{2} - \kappa \left| \bar{u}^{(1)}_{ij\ell,t} - \bar{u}^{(2)}_{ij\ell,t} \right| \right]. \tag{7}$$

The historical mean rewards repeated support, while the discrepancy term discounts changes between windows. Current relevance determines which pairs receive this assessment; historical evidence then changes their relative support and their competition with NULL. The default cache separates eight active states, $t-7, \ldots, t$, from 21 preceding states, $t-28, \ldots, t-8$, used for historical assessment. Both use the forecast-conditioned representations in Eq. 2.

Joint source-lag and null inference. Ranking the retained candidates identifies a preferred source even when the entire set has weak evidence. SLRA instead places the selected set $\mathcal{S}_{i,t}$ and a learned null logit in one distribution:

$$\left(p^{\varnothing}_{i,t}, \{\pi_{ij\ell,t}\}_{(j,\ell)\in\mathcal{S}_{i,t}}\right) = \mathrm{softmax}\left([b_\varnothing, \{e_{ij\ell,t}\}]/\tau_r\right). \tag{8}$$

Define acceptance and the conditional relation distribution as

$$\rho_{i,t} = 1 - p^{\varnothing}_{i,t}, \qquad q_{ij\ell,t} = \pi_{ij\ell,t}/\rho_{i,t}. \tag{9}$$

For a nonempty candidate set, $q = \mathrm{softmax}_{(j,\ell)\in\mathcal{S}_{i,t}}(e/\tau_r)$, which computes the conditional distribution directly. The acceptance probability has the equivalent form

$$\rho_{i,t} = \sigma\left(\mathrm{LSE}_{(j,\ell)\in\mathcal{S}_{i,t}}(e_{ij\ell,t}/\tau_r) - b_\varnothing/\tau_r\right), \tag{10}$$

where LSE is log-sum-exp and σ is the sigmoid. Thus acceptance compares the aggregate evidence of the retained pairs with the null alternative, while q allocates their conditional contribution. With no valid candidate, $p^{\varnothing} = 1$, $\rho = 0$, and relation content is zero.

Conditional content and the accepted update are

$$d^{\mathrm{rel}}_{i,t} = W_o \sum_{(j,\ell)\in\mathcal{S}_{i,t}} q_{ij\ell,t} V \tilde{r}_{j,t-\ell|t}, \qquad \Delta^{\mathrm{rel}}_{i,t} = \rho_{i,t} d^{\mathrm{rel}}_{i,t}. \tag{11}$$

This factorization assigns distinct roles to relation selection and acceptance within a single prediction update. Setting $\rho = 1$ yields forced retrieval from the same candidate set; setting $\rho = 0$ removes the lagged residual correction while preserving individual and common-context prediction. The forecasting objective learns how source content and acceptance jointly contribute to the target.

3.5 Coupled adaptation and prediction

Representation selection and relation inference are coupled through the residual computation. Common-path depth weights change C_t and hence the reference used to residualize the relational path. Relational depth weights change candidate compatibility, historical support, and retrieved values. The relation distribution therefore depends on which representations the routers select, and the forecasting loss trains those routers through the resulting prediction. This coupling adapts the representation used to judge a source together with the decision to use that source.

Separate residual gates combine the paths:

$$z_{i,t} = h^{\mathrm{ind}}_{i,t} + \tanh(\gamma_p) c^{\mathrm{panel}}_{i,t} + \tanh(\gamma_c) c^{\mathrm{com}}_{i,t} + \tanh(\gamma_r) \Delta^{\mathrm{rel}}_{i,t}. \tag{12}$$

Linear heads map $z_{i,t}$ to a ranking score $\hat{r}_{i,t}$ and a return forecast $\hat{y}_{i,t}$. The global gate γ_r learns the scale of the relation branch; $\rho_{i,t}$ controls acceptance for a particular target and date. Zero-initializing γ_r preserves the individual and common-context predictor at the start of relation adaptation. The accepted relation update then enters through a single residual gate.

Training combines pairwise logistic ranking within a market-date cross-section and a masked Huber return loss:

$$\mathcal{L}_{\text{task}} = \frac{1}{|\mathcal{P}_t|} \sum_{(i,j)\in\mathcal{P}_t} \log\left(1 + \exp[-\operatorname{sgn}(y_i - y_j)(\hat{r}_i - \hat{r}_j)]\right) + \lambda_{\text{ret}}\mathcal{L}_{\text{Huber}}. \tag{13}$$

$\mathcal{P}_t$ contains valid asset pairs with distinct target values, and the Huber loss compares $\hat{y}_{i,t}$ with the observed return under the label mask. We set the ranking term to zero when $\mathcal{P}_t$ is empty. This objective supervises the final forecast, so common-context estimation, depth selection, and relation acceptance learn from the same prediction task.

Full-universe computation. The frozen backbone permits reuse of layer snapshots while the adaptation modules remain trainable. Current candidate scoring costs $O(N_t^2|\mathcal{L}|d_r)$; evaluating two windows of length w on S retained pairs costs $O(N_t S w d_r)$. Depth mixture construction costs $O(N_t K T_a D)$ for T_a accessed states. Total runtime also includes common-context attention and backbone or cache access. Target chunking bounds working memory while retaining the full eligible source set, so each target selects relations from the same market universe. Predictions are joined across target chunks before market-date evaluation.

4 Experiments

The experiments examine whether prediction-time market-state inference improves asset-level forecasting and how its rankings separate subsequent returns. We then identify the contributing relational information, examine its interaction with representation routing, and evaluate the complete model across later dates, held-out target assets, and reduced source availability.

4.1 Markets, tasks, and comparisons

Multi-market evaluation. The evaluation covers six equity universes: Chinese A-shares, S&P 500, Hang Seng Composite (HSCI), Nifty 500, Nikkei 225, and KRX 300. Each prediction uses a synchronized panel of stocks within one market. Universe composition and data coverage are documented in Appendix A.

Forecasting protocol. The main task predicts five-session returns from 128 trading sessions of history. All six markets share chronological training, validation, test, and out-of-time (OOT) boundaries, aligning evaluation periods across equity universes. Parameters are optimized on Train; hyperparameters and checkpoints are selected on Valid after training. Test provides the main comparisons, ablations, and aggregate improvements reported in the abstract. OOT assesses subsequent temporal generalization (Appendix C.2). Each market retains its trading calendar, with forward-return intervals contained within their assigned partition. Methods share preprocessing, input eligibility, and targets. The common timeline and detailed CN and S&P 500 sample counts appear in Appendix A.

Comparators and metrics. The main comparisons include Kronos with matched adaptation, a frozen-backbone probe, and the cross-asset predictors HIST, MASTER, LIFT, and DeltaLag. RankIC is the daily cross-sectional Spearman correlation between ranking scores and future returns; RankICIR is its temporal mean divided by its standard deviation, without annualization. Pearson IC measures linear association, while MSE and MAE evaluate return estimates. Metrics are computed within market-dates and averaged over dates. Paired runs and date-block bootstrap intervals quantify uncertainty.

4.2 Does market-state inference improve forecasting?

Market-level modeling should translate into predictive value across distinct equity universes. Table 1 compares ranking performance and temporal stability across markets; Table 2 resolves this comparison into ranking quality and return-estimation error in A-shares and S&P 500 equities. Both use the same five-session horizon.

Table 1: Five-session forecasting on the Test partitions across equity markets. Higher RankIC and RankICIR indicate stronger ranking performance and greater temporal stability, respectively.

	A-shares		S&P 500		HSCI	
Method	RankIC ↑	RankICIR ↑	RankIC ↑	RankICIR ↑	RankIC ↑	RankICIR ↑
Kronos-base (adapted)	0.0667	0.467	0.0386	0.279	0.0479	0.312
Kronos-base (probe)	0.0552	0.383	0.0309	0.215	0.0398	0.256
HIST	0.0685	0.481	0.0392	0.271	0.0531	0.349
MASTER	0.0746	0.529	0.0473	0.342	0.0568	0.376
LIFT	0.0718	0.517	0.0506	0.367	0.0584	0.394
DeltaLag	0.0789	0.558	0.0494	0.358	0.0609	0.408
VARIS	**0.0948**	**0.691**	**0.0607**	**0.398**	**0.0784**	**0.498**
	Nifty 500		Nikkei 225		KRX 300	
Method	RankIC ↑	RankICIR ↑	RankIC ↑	RankICIR ↑	RankIC ↑	RankICIR ↑
Kronos-base (adapted)	0.0514	0.356	0.0508	0.348	0.0603	0.406
Kronos-base (probe)	0.0451	0.299	0.0429	0.284	0.0506	0.335
HIST	0.0558	0.375	0.0509	0.349	0.0631	0.423
MASTER	0.0612	0.418	0.0596	0.417	0.0687	0.481
LIFT	0.0645	0.446	0.0628	**0.449**	0.0674	0.466
DeltaLag	0.0633	0.439	0.0617	0.433	0.0732	0.506
VARIS	**0.0692**	**0.501**	**0.0649**	0.403	**0.0973**	**0.752**

Table 2: Five-session Test results in A-shares and S&P 500 equities. Correlation metrics evaluate ranking scores; MSE and MAE evaluate predicted returns. MSE and MAE are multiplied by 10^3 and 10^2, respectively.

	A-shares				
Method	IC ↑	RankIC ↑	RankICIR ↑	MSE ↓	MAE ↓
Kronos-base (adapted)	0.0592	0.0667	0.467	2.385	3.567
HIST	0.0608	0.0685	0.481	2.382	3.562
MASTER	0.0651	0.0746	0.529	2.373	3.550
LIFT	0.0636	0.0718	0.517	2.367	**3.541**
DeltaLag	0.0692	0.0789	0.558	2.370	3.546
VARIS	**0.0825**	**0.0948**	**0.691**	**2.359**	3.543
	S&P 500				
Method	IC ↑	RankIC ↑	RankICIR ↑	MSE ↓	MAE ↓
Kronos-base (adapted)	0.0338	0.0386	0.279	1.216	2.501
HIST	0.0352	0.0392	0.271	1.218	2.502
MASTER	0.0418	0.0473	0.342	1.211	2.492
LIFT	0.0450	0.0506	0.367	1.207	**2.484**
DeltaLag	0.0437	0.0494	0.358	1.210	2.490
VARIS	**0.0529**	**0.0607**	**0.398**	**1.205**	2.486

VARIS improves RankIC and RankICIR over adapted Kronos-base in all six markets; LIFT retains the lowest MAE in Table 2. Figure 3 connects ranking performance to realized returns. Each model independently partitions the same eligible stocks into five score-ranked groups; returns are averaged within groups and equally across dates (Appendix C).

VARIS produces the largest Q5–Q1 separation, increasing it from DeltaLag's 1.219 to 1.499 percentage points in A-shares and from 0.662 to 0.777 in S&P 500. Intermediate groups remain close across methods. Endpoint separation complements RankIC: LIFT exceeds DeltaLag on S&P 500 RankIC but has smaller separation. We next examine the information and selection mechanisms underlying the forecasting gains.

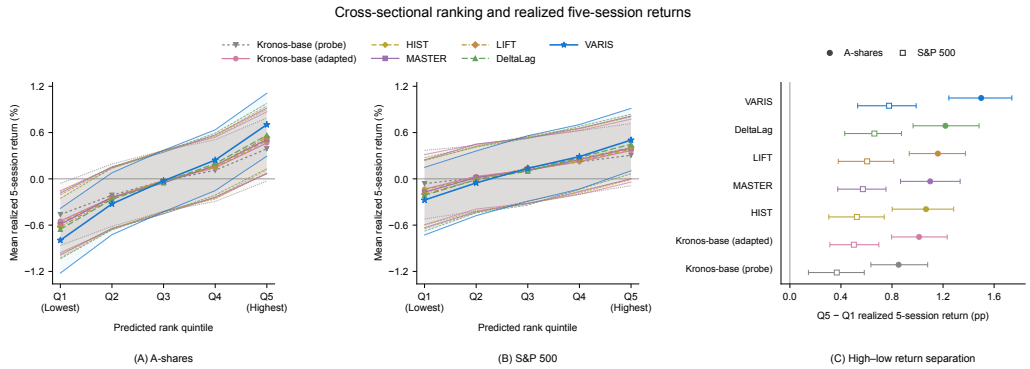


Figure 3: **Cross-sectional ranking and realized five-session returns.** (A,B) Date-averaged returns by model-specific score quintile; shading connects pointwise 95% interval endpoints. (C) Q5–Q1 separation in percentage points, before execution or costs, with 95% interval bars; circles denote A-shares and squares S&P 500.

4.3 What makes a relation useful?

The component controls separate shared information from relational evidence. Removing the relation path produces the largest RankIC decrease in Table 3; separate removals of common prediction, residualization, historical consistency, and NULL resolve their contributions to accuracy and stability.

Table 3: Mechanism contributions and routing–selection interaction on Test. Common-prediction removal retains the residual reference. Rows A–D form the routing–selection comparison.

	A-shares		S&P 500	
Variant	RankIC ↑	RankICIR ↑	RankIC ↑	RankICIR ↑
Without common prediction	0.0889	0.622	0.0541	0.381
Without relation path	0.0764	0.511	0.0447	0.310
Without residualization	0.0856	0.583	0.0496	0.336
Without historical consistency	0.0934	0.616	0.0584	0.371
Depth routing × relation selection				
A: Static / forced	0.0776	0.498	0.0462	0.318
B: Static / selective	0.0847	0.574	0.0528	0.346
C: Dynamic / forced (without NULL)	0.0831	0.559	0.0499	0.339
D: Dynamic / selective (VARIS)	**0.0948**	**0.691**	**0.0607**	**0.398**

Source identity and temporal alignment contribute to the relational gain. Preserving target histories and common prediction, source permutation, lag randomization, and past-window displacement yield A-shares RankIC of 0.0773, 0.0842, and 0.0798, versus 0.0948 for VARIS. These interventions test retrieved content and alignment (Appendix B).

Selective acceptance tracks predictive contribution: on matched targets, updates improve RankIC in high-acceptance bins and contribute little at high NULL probability. Appendix D provides source and temporal diagnostics.

4.4 How does depth routing interact with relation selection?

Dynamic routing gains more from selective than forced relation updates. In Table 3, its RankIC gain is 0.0101 versus 0.0055 in A-shares and 0.0079 versus 0.0037 in S&P 500. The positive interaction contrasts, 0.0046 and 0.0042, link representation choice to relation acceptance at the point-estimate level.

Asset- and path-conditioned routing improves prediction beyond shared depth selection. A-shares RankIC rises from 0.0827 with last-layer access to 0.0912 with shared dynamic routing and 0.0948 with asset- and path-conditioned routing (Appendix B). Together, these comparisons support coupling representation selection with relation inference.

4.5 Generalization across time, assets, and source availability

Having examined the contributing mechanisms, Figure 4 evaluates the complete model across later dates, held-out target stocks, and reduced source availability.

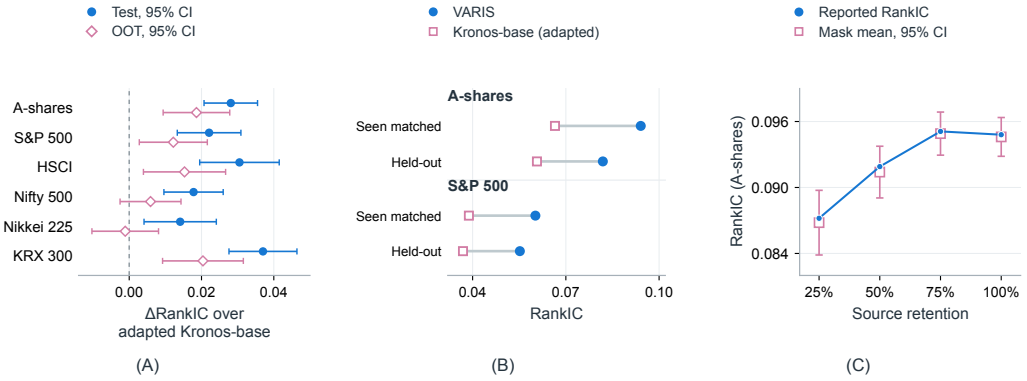


Figure 4: **Generalization across time, assets, and source availability.** (A) Test/OOT RankIC gains over adapted Kronos-base with 95% intervals. (B) Seen-matched versus adaptation-held-out stocks; segments connect methods. (C) A-shares source retention with fixed targets: circles show reported RankIC, squares show mask means, and bars show 95% intervals.

Temporal transfer retains market-dependent gains. With Valid-selected parameters and preprocessing fixed, OOT gains over adapted Kronos-base are smaller than Test gains in all six markets (Figure 4A). Four intervals remain above zero; Nifty 500 crosses zero, and Nikkei 225 has a slightly negative point estimate.

The advantage also extends to stocks excluded from adaptation. Held-out RankIC is 0.0819 in A-shares and 0.0552 in S&P 500 (Figure 4B): below the seen-matched values, but above adapted Kronos-base within each group.

Source coverage affects prediction on fixed targets. A-shares RankIC rises from 0.0872 at 25% retention to 0.0951 at 75%, with 0.0948 at full coverage (Figure 4C). These comparisons characterize transfer across time and assets alongside dependence on available sources. Appendix C gives detailed protocols and portfolio evaluation.

5 Conclusion

Pretraining learns reusable temporal representations across assets; a forecast must still decide which market evidence to trust. VARIS infers a latent market state at prediction time, combining asset histories, common context, and target-specific relations instead of fixed metadata.

Common movement provides context; residuals are checked across past windows to choose sources and how much to accept, including none. The forecast loss jointly trains representation choice and evidence use for each target on each date. Mean RankIC and RankICIR rise 46.8% and 52.3%, respectively, over adapted Kronos-base across six markets and three horizons. Source and lag controls and the routing–selection comparison support learning these choices together.

Out-of-time gains remain below Test gains and vary by market; learned links are predictive, not causal. These findings motivate regime adaptation, scalable retrieval, and evaluation on historical constituent universes. Financial foundation models should move from asset-wise temporal representation learning toward prediction-time relational market-state inference, jointly deciding which representations to use and what market evidence to trust.

AI USE STATEMENT

Generative AI tools assisted with experimental code design, revisions and debugging during experimentation, and code review. They also supported the refinement of experimental procedures and small-scale code runs to assess implementation feasibility. For manuscript preparation, AI assistance was used to polish the language and improve the organization and clarity of the text. The authors take responsibility for the experimental work, code, and final manuscript.

ETHICS STATEMENT

This work studies financial forecasting from security-level histories of prices, volumes, and trading amounts. Financial forecasting can influence capital allocation, and downstream use calls for attention to market coverage, changing conditions, transaction costs, and risk management. Applications should incorporate human oversight and evaluation appropriate to their deployment setting. Market data and pretrained checkpoints remain subject to their providers' access and redistribution terms.

REPRODUCIBILITY STATEMENT

The model formulation and prediction-time inference procedure are specified in Section 3 and Appendix G, including common-context estimation, source–lag selection, relation acceptance, depth routing, and the prediction heads. Appendix A documents the market universes, chronological partitions, input and label eligibility, the 128-session input window, preprocessing conventions, and the pretrained Kronos checkpoint and tokenizer. Appendix F defines the evaluation metrics, temporal aggregation rules, displayed units, and ranking-group analysis, while Appendices B and C describe the controlled mechanism comparisons, transfer settings, and uncertainty procedures. The two focal-market ranking-group intervals use joint circular date-block resampling with 10-session blocks and 1,000 resamples. Code and experiment configurations will be released after the review process. Market data and pretrained checkpoints remain subject to their providers' access and redistribution terms.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Ali Caner Türkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Bernie Wang. Chronos: Learning the Language of Time Series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From Univariate to Universal Forecasting, 2025. URL https://arxiv.org/abs/2510.15821.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 10148–10167, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Yitong Duan, Lei Wang, Qizhong Zhang, and Jian Li. FactorVAE: A probabilistic dynamic factor model based on variational autoencoder for predicting cross-sectional stock returns. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, volume 36, pp. 4468–4476, 2022. doi: 10.1609/aaai.v36i4.20369. URL https://ojs.aaai.org/index.php/AAAI/article/view/20369.

Yitong Duan, Weiran Wang, and Jian Li. FactorGCL: A hypergraph-based factor model with temporal residual contrastive learning for stock returns prediction. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, pp. 173–181, 2025. doi: 10.

1609/aaai.v39i1.31993. URL https://ojs.aaai.org/index.php/AAAI/article/view/31993.

Elliot L. Epstein, Apaar Sadhwani, and Kay Giesecke. A Set-Sequence Model for Time Series. In *Advances in Financial AI: Opportunities, Innovations, and Responsible AI Workshop at ICLR 2025*, 2025. URL https://iclr.cc/virtual/2025/33828.

Utku Evci, Vincent Dumoulin, Hugo Larochelle, and Michael C. Mozer. Head2Toe: Utilizing Intermediate Representations for Better Transfer Learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 6009–6033, 2022. URL https://proceedings.mlr.press/v162/evci22a.html.

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. ISSN 1046-8188. doi: 10.1145/3309547. URL https://doi.org/10.1145/3309547.

Xing Feng, Hongru Li, and Yinghua Yang. Time-lagged relation graph neural network for multivariate time series forecasting. *Engineering Applications of Artificial Intelligence*, 139:109530, 2025. ISSN 0952-1976. doi: 10.1016/j.engappai.2024.109530. URL https://doi.org/10.1016/j.engappai.2024.109530.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical Asset Pricing via Machine Learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020. ISSN 0893-9454. doi: 10.1093/rfs/hhaa009. URL https://doi.org/10.1093/rfs/hhaa009.

Jingyi Hou, Zhen Dong, Jiayu Zhou, and Zhijie Liu. Discovering Predictable Latent Factors for Time Series Forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(10):5106–5119, 2024. ISSN 1041-4347. doi: 10.1109/TKDE.2023.3335240. URL https://doi.org/10.1109/TKDE.2023.3335240.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cGDAkQo1C0p.

Kimi Team, Guangyu Chen, Yu Zhang, Jianlin Su, Weixin Xu, Siyuan Pan, Yaoyu Wang, Yucheng Wang, Guanduo Chen, Bohong Yin, Yutian Chen, Junjie Yan, Ming Wei, Y. Zhang, Fanqing Meng, Chao Hong, Xiaotong Xie, Shaowei Liu, Enzhe Lu, Yunpeng Tai, Yanru Chen, Xin Men, Haiqing Guo, Y. Charles, Haoyu Lu, Lin Sui, Jinguo Zhu, Zaida Zhou, Weiran He, Weixiao Huang, Xinran Xu, Yuzhi Wang, Guokun Lai, Yulun Du, Yuxin Wu, Zhilin Yang, and Xinyu Zhou. Attention Residuals, 2026. URL https://arxiv.org/abs/2603.15031.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3744–3753, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Shuqi Li, Yuebo Sun, Yuxin Lin, Xin Gao, Shuo Shang, and Rui Yan. CausalStock: Deep end-to-end causal discovery for news-driven multi-stock movement prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-1504. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/54d689d58fe54c92aee2d732fc49fca8-Abstract-Conference.html.

Tong Li, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, and Sen Huang. MASTER: Market-Guided Stock Transformer for Stock Price Forecasting. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, volume 38, pp. 162–170. Association for the

Advancement of Artificial Intelligence (AAAI), 2024b. doi: 10.1609/aaai.v38i1.27767. URL https://ojs.aaai.org/index.php/AAAI/article/view/27767.

Zhengze Li, Zilin Song, Tingting Yuan, and Xiaoming Fu. COGRASP: Co-Occurrence Graph Based Stock Price Forecasting. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 7527–7535. International Joint Conferences on Artificial Intelligence Organization, 2025. doi: 10.24963/ijcai.2025/837. URL https://www.ijcai.org/proceedings/2025/837.

Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-Tao Xia. TimeBridge: Non-stationarity matters for long-term time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39815–39840, 2025a. URL https://proceedings.mlr.press/v267/liu25cb.html.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4054556fcaa934b0bf76da52cf4f92cb-Abstract-Conference.html.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=JePfAI8fah.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative Pre-trained Transformers Are Large Time Series Models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 32369–32399, 2024b. URL https://proceedings.mlr.press/v235/liu24cb.html.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A Family of Highly Capable Time Series Foundation Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 39295–39317, 2025b. URL https://proceedings.mlr.press/v267/liu25be.html.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Hao Qian, Hongting Zhou, Qian Zhao, Hao Chen, Hongxiang Yao, Jingwei Wang, Ziqi Liu, Fei Yu, Zhiqiang Zhang, and Jun Zhou. MDGNN: Multi-Relational Dynamic Graph Neural Network for Comprehensive and Dynamic Stock Investment Prediction. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, volume 38, pp. 14642–14650. Association for the Advancement of Artificial Intelligence (AAAI), 2024. doi: 10.1609/aaai.v38i13.29381. URL https://ojs.aaai.org/index.php/AAAI/article/view/29381.

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, Tyler Derr, and Rajiv Ratn Shah. Stock Selection via Spatiotemporal Hypergraph Attention Network: A Learning to Rank Approach. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, volume 35, pp. 497–504. Association for the Advancement of Artificial Intelligence (AAAI), 2021. doi: 10.1609/aaai.v35i1.16127. URL https://ojs.aaai.org/index.php/AAAI/article/view/16127.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-Scale Time Series Foundation Models with Mixture of Experts. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tPFxnyNYUP.

Yu Shi, Zongliang Fu, Shuo Chen, Bohan Zhao, Wei Xu, Changshui Zhang, and Jian Li. Kronos: A Foundation Model for the Language of Financial Markets. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, volume 40, pp. 25366–25373. Association for the

Advancement of Artificial Intelligence (AAAI), 2026. doi: 10.1609/aaai.v40i30.39730. URL https://ojs.aaai.org/index.php/AAAI/article/view/39730.

Heyuan Wang, Shun Li, Tengjiao Wang, and Jiayi Zheng. Hierarchical Adaptive Temporal-Relational Modeling for Stock Trend Prediction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 3691–3698. International Joint Conferences on Artificial Intelligence Organization, 2021. doi: 10.24963/ijcai.2021/508. URL https://www.ijcai.org/proceedings/2021/508.

Mengyu Wang, Tiejun Ma, and Shay B. Cohen. Pre-training Time Series Models with Stock Data Customization. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, volume 2, pp. 3019–3030. ACM, 2025. doi: 10.1145/3711896.3737005. URL https://doi.org/10.1145/3711896.3737005.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0015. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0113ef4642264adc2e6924a3cbbdf532-Abstract-Conference.html.

Yuyang Wang, Alex Smola, Danielle C. Maddix, Jan Gasthaus, Dean Foster, and Tim Januschowski. Deep Factors for Forecasting. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 6607–6617, 2019. URL https://proceedings.mlr.press/v97/wang19k.html.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified Training of Universal Time Series Forecasting Transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 53140–53164, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Binqing Wu, Zongjiang Shang, Jianlong Huang, and Ling Chen. MillGNN: Learning Multi-Scale Lead-Lag Dependencies for Multi-Variate Time Series Forecasting. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 3344–3354. ACM, 2025. doi: 10.1145/3746252.3761173. URL https://doi.org/10.1145/3746252.3761173.

Wentao Xu, Weiqing Liu, Lewen Wang, Yingce Xia, Jiang Bian, Jian Yin, and Tie-Yan Liu. HIST: A Graph-based Framework for Stock Trend Forecasting via Mining Concept-Oriented Shared Information, 2021. URL https://arxiv.org/abs/2110.13716.

Yuqi Yuan, Xiong Luo, Qiaojuan Peng, and Wenbing Zhao. GraFT: Infusing pre-trained transformers with relational structure for time series forecasting. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence*, volume 40, pp. 28041–28049, 2026. doi: 10.1609/aaai.v40i33.40029. URL https://ojs.aaai.org/index.php/AAAI/article/view/40029.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, 2017. URL https://papers.nips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Lifan Zhao and Yanyan Shen. Rethinking Channel Dependence for Multivariate Time Series Forecasting: Learning from Leading Indicators. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JiTVtCUOpS.

Wanyun Zhou, Saizhuo Wang, Mihai Cucuringu, Zihao Zhang, Xiang Li, Jian Guo, Chao Zhang, and Xiaowen Chu. DeltaLag: Learning Dynamic Lead-Lag Patterns in Financial Markets. In *Proceedings of the 6th ACM International Conference on AI in Finance*, pp. 422–430. ACM, 2025. doi: 10.1145/3768292.3770421. URL https://doi.org/10.1145/3768292.3770421.

Zhuohang Zhu, Haodong Chen, Qiang Qu, and Vera Chung. FinCast: A Foundation Model for Financial Time-Series Forecasting. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 4539–4549. ACM, 2025. doi: 10.1145/3746252.3761261. URL https://doi.org/10.1145/3746252.3761261.

A DATA AND EXPERIMENTAL SETTINGS

Table 4 records the equity universes. Each market retains its own trading calendar. The CN panel contains 5,510 stocks and 16,027,476 daily records from 2000-01-04 to 2026-07-06. The S&P 500 panel contains 472 stocks with available histories in the constituent snapshot and 2,703,531 daily records from 2000-01-03 to 2026-02-20. Its membership is fixed by that snapshot rather than reconstructed from historical constituent changes.

Table 4: Equity universes and data coverage.

Market	Equity universe	Coverage	Membership source
Mainland China	A-shares	5,510 stocks	Exchange records
United States	S&P 500 equities	472-stock snapshot	S&P Dow Jones Indices
Hong Kong	Hang Seng Composite	Index constituents	Hang Seng Indexes
India	Nifty 500	Index constituents	NSE Indices
Japan	Nikkei 225	Index constituents	Nikkei Indexes
South Korea	KRX 300	Index constituents	KRX

Chronological partitions. All six markets use the same calendar boundaries: 2018-08-29 separates Train and Valid, 2022-04-01 separates Valid and Test, and 2024-05-23 defines the Test–OOT cutoff. Eligible anchors are determined on each market's trading calendar using the 128-session history and the forecast horizon. Forward-return intervals remain within a single partition; input histories include earlier observations available at the forecast date. Table 5 details the resulting five-session partitions for CN A-shares and S&P 500. Their final Train, Valid, and Test anchors are 2018-08-21, 2022-03-24, and 2024-05-16, respectively, and their first retained OOT anchor is 2024-05-24.

Table 5: Chronological partitions for five-session forecasting. Dates delimit retained forecast anchors. Stock–date counts measure input eligibility; label counts measure valid supervision.

Split	First anchor	Last anchor	Dates	Stock–dates	Labels
CN A-shares					
Train	2000-07-20	2018-08-21	4,388	2,842,091	2,809,049
Valid	2018-08-29	2022-03-24	865	2,912,875	2,909,242
Test	2022-04-01	2024-05-16	513	2,387,752	2,386,348
OOT	2024-05-24	2026-06-29	508	2,599,209	2,596,631
Total	2000-07-20	2026-06-29	6,274	10,741,927	10,701,270
S&P 500 snapshot-472					
Train	2000-07-05	2018-08-21	4,562	1,774,305	1,774,305
Valid	2018-08-29	2022-03-24	899	408,298	408,297
Test	2022-04-01	2024-05-16	534	248,669	248,669
OOT	2024-05-24	2026-02-12	431	202,947	202,942
Total	2000-07-05	2026-02-12	6,426	2,634,219	2,634,213

Cross-sectional coverage. Each anchor uses all eligible stocks on that date. The CN panel ranges from 1 to 5,206 eligible stocks, averaging approximately 1,712; the S&P 500 panel ranges from 323 to 472, averaging approximately 410. Input eligibility and label availability are tracked separately. Target chunks of 64 or 128 stocks partition computation while retaining the full eligible source universe.

Data preparation and training. The main configuration uses an eight-state active cross-sectional window, a 128-session input, and the open NeoQuasar/Kronos-base checkpoint with its base tokenizer. Price adjustment, missing-volume and amount conventions, input masks, and label availability are shared across methods. Return intervals crossing a split boundary are excluded. The primary horizon is five sessions, with one- and twenty-session evaluations in Appendix C.

Evaluation. Ranking scores determine IC and RankIC, while return estimates determine MSE and MAE. All comparisons use the same eligible targets within each market-date. MSE and MAE are multiplied by 10^3 and 10^2, respectively. Appendix C.1 specifies the paired uncertainty analysis.

B Detailed mechanism analysis

Component and interaction controls. Removing common prediction disables its direct forecast contribution while retaining the common reference for relation inference. Removing residualization uses total representations for relational keys, queries, and values. Forced relation updates set acceptance to one and retain conditional source-lag selection. Static routing uses learned, input-independent depth weights for each path; dynamic routing uses asset- and path-conditioned weights. Candidate budgets, lag sets, prediction-head capacity, and training budgets remain matched. Rows A–D of Table 3 give the interaction contrast $(D - C) - (B - A)$ for each metric.

Relational evidence. Source permutations and lag randomization perturb the relation path while preserving target histories and common prediction. The corresponding A-shares RankIC values are 0.0773, 0.0842, and 0.0798 for source permutation, lag randomization, and past-window displacement, compared with 0.0948 for VARIS. Acceptance bins compare the same targets with and without relational updates. Market-state stratification distinguishes periods of strong common movement from dispersed stock returns.

Temporal representations. Additional comparisons include last-layer access, uniform depth averaging, static mixing, shared dynamic routing, and asset- and path-conditioned routing. A-shares RankIC rises from 0.0827 with the last layer to 0.0912 with shared dynamic routing and 0.0948 with asset- and path-conditioned routing. Depth weights are examined by date and prediction path alongside selected lags and relation acceptance. Controlled layer removal measures the contribution of highly weighted representations.

C Extended evaluation

Ranking-group evaluation. Figure 3 evaluates five-session returns using each model's cross-sectional ranking scores. At each Test date, each model ranks the common eligible target universe and divides it into five approximately equal-sized groups, from Q1 (lowest score) to Q5 (highest score). Group membership is model-specific. The plotted quantity is the subsequent realized five-session return, with stocks weighted equally within each group and dates weighted equally in the reported mean. Ranking scores determine group membership; their magnitudes are not interpreted as returns. Q5–Q1 separation is computed within each date and then averaged across dates; its interval is computed from this paired difference rather than by subtracting marginal interval endpoints. For the two focal markets, the intervals use a joint circular block bootstrap with 10-session blocks and 1,000 resamples, sharing date resamples across models and groups within each market. This preserves the paired comparisons and accounts for local temporal dependence in the overlapping five-session labels. Group separation summarizes ranking extremes; RankIC evaluates ordering across the complete eligible cross-section. Appendix C.4 extends the plots to the remaining four markets and provides separate views of every model across all six markets.

Markets and forecast horizons. The extended comparison evaluates IC, RankIC, RankICIR, MSE, and MAE on each market's Test partition under the main evaluation protocol. Across 18 market–horizon settings, VARIS improves mean RankIC by 46.8% and mean RankICIR by 52.3% over Kronos-base, with gains in 16 settings for both measures. Relative gains are computed as the ratio of equal-weight macro-averages minus one.

C.1 Uncertainty in the relational gain

The incremental value of market structure is measured through paired differences on the same prediction targets. At the five-session horizon, let d_t denote the difference in daily RankIC between VARIS and matched-adaptation Kronos-base, and let $\Delta = \mathrm{mean}_t\, d_t$. This contrast connects the cross-market comparison in Table 1 to the precision with which each market's gain can be resolved.

Table 6: Paired Test RankIC gains over adapted Kronos-base at the five-session horizon. Confidence intervals use the normal approximation $\Delta \pm 1.96\,\mathrm{SE}$.

Market	ΔRankIC	SE	95% CI
A-shares	0.0281	0.0038	[0.0207, 0.0355]
S&P 500	0.0221	0.0045	[0.0133, 0.0309]
HSCI	0.0305	0.0056	[0.0195, 0.0415]
Nifty 500	0.0178	0.0042	[0.0096, 0.0260]
Nikkei 225	0.0141	0.0051	[0.0041, 0.0241]
KRX 300	0.0370	0.0048	[0.0276, 0.0464]

Table 6 complements the relative improvements with absolute paired gains and their uncertainty. The evaluation protocol resamples contiguous date blocks jointly for both models, preserving their paired predictions and temporal dependence. Block lengths account for overlapping return horizons and the autocorrelation of daily scores. RankICIR is recomputed within each resample, while variability across training seeds is reported separately.

For the aggregate comparison, each resample recomputes the ratio of equal-weight macro-averages across all 18 market–horizon settings. Shared calendar blocks preserve dependence across markets and horizons. This yields an interval for the same relative-gain statistic used in the abstract.

C.2 Transfer and computational evaluation

Out-of-time evaluation. Temporal generalization in Figure 4A uses the Valid-selected models on each market's OOT partition following the common 2024-05-23 cutoff. Eligible OOT anchors span 2024-05-24–2026-06-29 for CN A-shares and 2024-05-24–2026-02-12 for S&P 500. Model parameters, preprocessing, and forecasting settings remain fixed. OOT metrics are aggregated within each market over its eligible dates and reported separately from Test, measuring how prediction-time market-structure inference transfers to later market conditions.

Asset and market transfer. Held-out stocks in Figure 4B are excluded from adaptation training. Source-set experiments in Figure 4C retain the same prediction targets while varying available source assets; RankIC values for 25%, 50%, 75%, and 100% source retention are 0.0872, 0.0919, 0.0951, and 0.0948 in A-shares. Held-out-market transfer evaluates a separate change in market distribution.

Computation. Latency and peak memory are measured as the source universe grows, comparing dense evaluation with target chunking on identical source sets. Backbone cache construction and adaptation costs are reported separately for the 102.3M-parameter Kronos-base backbone and the trainable VARIS modules.

C.3 Portfolio evaluation

We further examine whether improvements in cross-sectional ranking translate into portfolio-level economic value under a common portfolio construction rule. At each prediction date t, we rank all eligible stocks by the ranking-head score $\hat{r}_{i,t}$ and select the top 10% to form an equally weighted long-only sleeve. For the primary five-session prediction horizon, we maintain five staggered sleeves: a new sleeve is initiated each trading day, allocated one-fifth of total capital, and held for five trading sessions. Signals generated using information available through date t are executed at the next trading day's open. Stock eligibility follows the screening rules used in the forecasting experiments and is identical across methods.

Portfolio wealth compounds realized returns. Turnover is computed from changes in actual portfolio weights between consecutive trading days, and transaction costs of 10 basis points (bps) are charged on the one-way traded notional. Figure C.1(a,b) reports cumulative net wealth under this common cost assumption, while panel (c) compares terminal net wealth across alternative cost levels. Together, the trajectories and cost-sensitivity analysis show how realized portfolio value accumulates over time and how it changes with implementation frictions.

Figure C.1(a,b) presents cumulative net wealth under the common portfolio construction rule. In both A-shares and S&P 500 equities, VARIS finishes with the highest terminal net wealth among the active forecasting models. The separation is more pronounced in A-shares, while the S&P 500 comparison is tighter. The trajectories nevertheless contain common drawdowns, temporary reversals, and periods in which the strongest comparator approaches or exceeds VARIS. The portfolio-level advantage therefore accumulates over the evaluation horizon rather than reflecting uniform period-by-period dominance. These results complement the forecasting metrics by showing that improvements in cross-sectional ranking can translate into higher realized portfolio value under a common allocation rule.

Figure C.1(c) evaluates transaction-cost assumptions of 0, 5, 10, 20, and 30 bps. Terminal net wealth declines as trading costs increase for all active forecasting models, confirming that the economic value of the ranking signal is sensitive to turnover and implementation frictions. VARIS retains the highest terminal wealth across the displayed cost range in both markets, although its absolute advantage narrows as costs increase. The sensitivity analysis therefore separates the predictive value of the ranking signal from the trading frictions required to realize it.

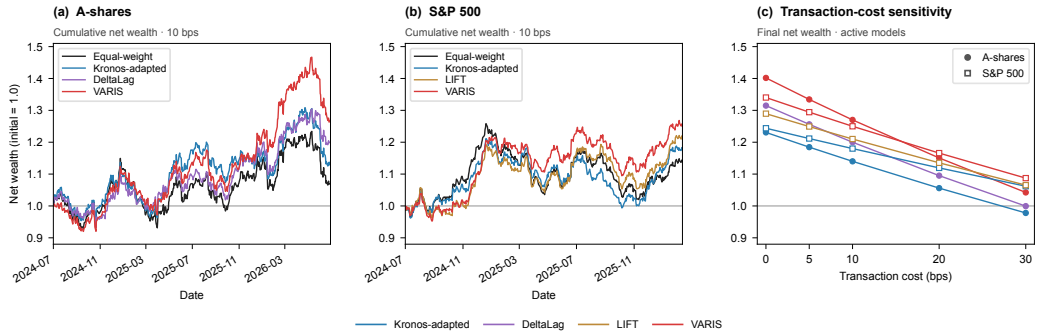


Figure C.1: **Portfolio-level economic evaluation.** Cumulative net wealth for (a) A-shares and (b) S&P 500 under transaction costs of 10 bps on the one-way traded notional. At each prediction date, stocks are ranked by the model's ranking scores, and the top 10% form an equally weighted long-only sleeve. Five staggered sleeves align portfolio construction with the five-session prediction horizon; signals generated from information available through date t are executed at the next trading day's open. Initial wealth is normalized to 1, and cumulative wealth compounds realized portfolio returns after transaction costs. (c) Terminal net wealth of active forecasting models under transaction costs of 0, 5, 10, 20, and 30 bps. Filled circles and open squares denote A-shares and S&P 500, respectively.

C.4 Ranking behavior across markets and models

The additional markets extend the score-stratification analysis in Figure 3 to all six evaluation universes. Figure C.2 compares all seven models on HSCI, Nifty 500, Nikkei 225, and KRX 300 using the ranking-group protocol in Appendix C. Figure C.3 summarizes the corresponding Q5–Q1 differences. These comparisons assess how model scores separate subsequent realized returns within each market; the ordinate measures five-session group returns rather than cumulative wealth.

VARIS has increasing mean returns from Q1 to Q5 in all six markets and the largest mean Q5–Q1 separation in five of them. In the four additional markets, its separation is 1.117, 1.020, 0.728, and 1.386 percentage points for HSCI, Nifty 500, Nikkei 225, and KRX 300, respectively. LIFT attains the largest separation in Nikkei 225 at 0.761 percentage points. Table 1 also assigns LIFT the higher RankICIR in this market, while VARIS has the higher mean RankIC. Together, the results connect the aggregate ranking gains to differentiation of realized returns while distinguishing full-cross-section ordering, temporal stability, and endpoint separation.

The following model-specific views separate overlapping intervals while retaining the same values as the combined comparisons. Figures C.4–C.10 each show one model across six markets. All model-specific plots and the four-market quintile overview use identical vertical limits, enabling direct comparison of return profiles. Each line connects five discrete group means, and the same-

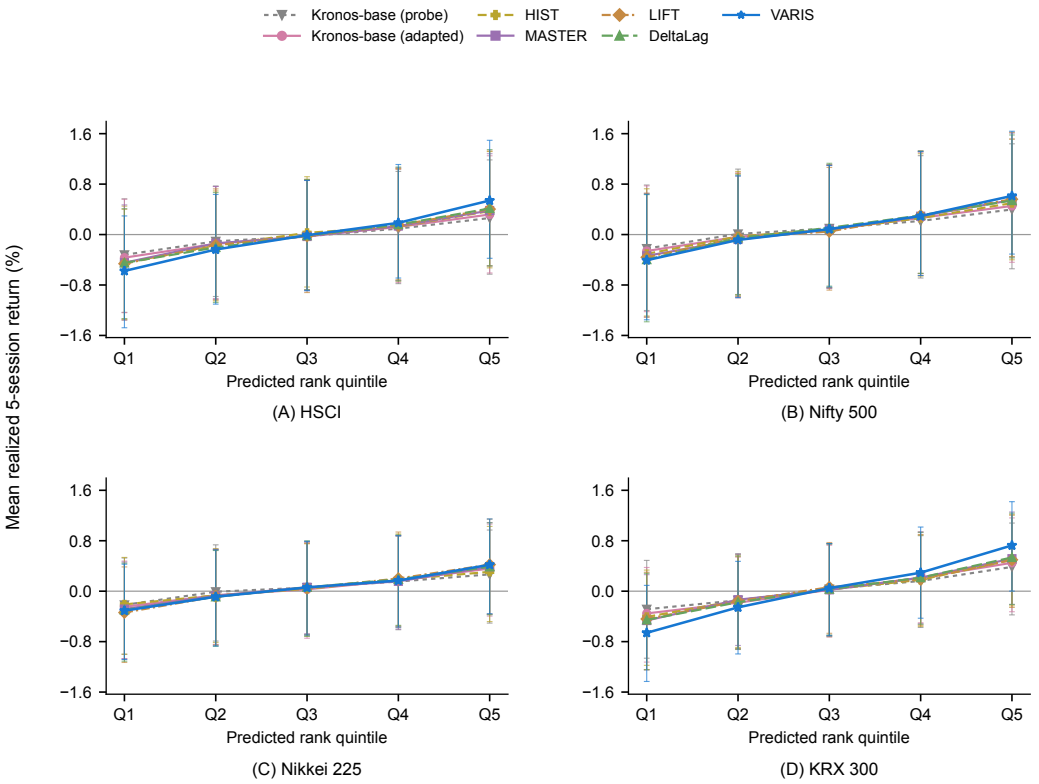


Figure C.2: **Cross-sectional ranking and realized five-session returns across additional markets.** (A) HSCI, (B) Nifty 500, (C) Nikkei 225, and (D) KRX 300. Each model independently forms five score-ranked groups within the common eligible target universe. Points show date-averaged realized five-session returns; error bars show 95% intervals. Model colors, markers, and line styles match Figure 3. Straight connections preserve the observed group means.

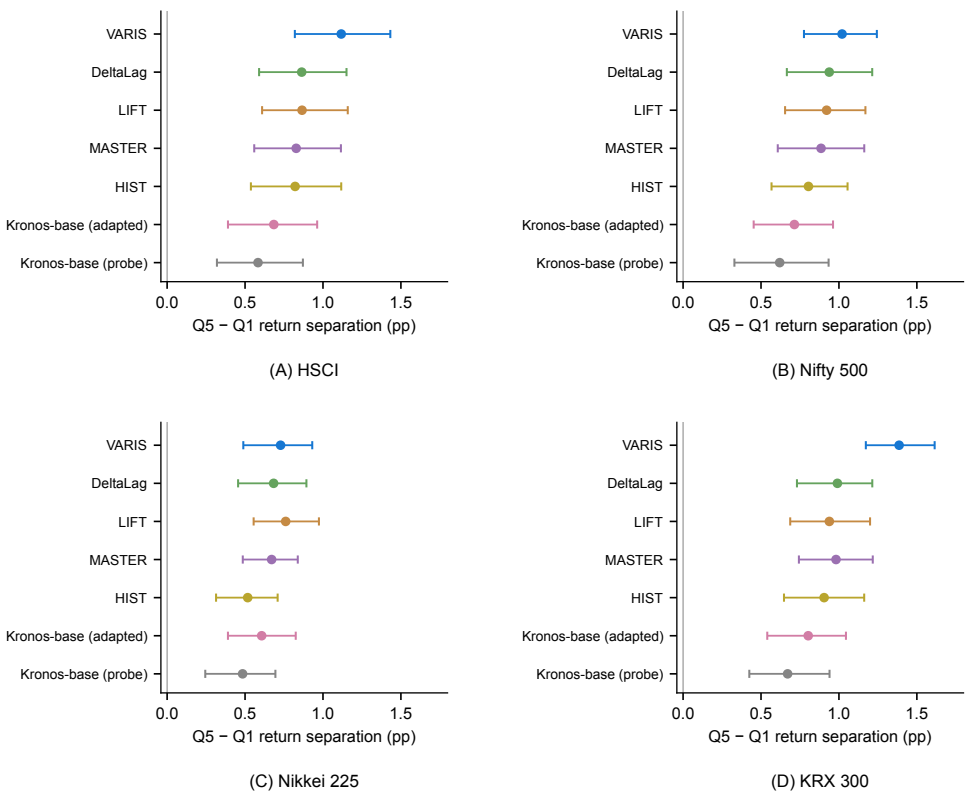


Figure C.3: **High–low realized-return separation across additional markets.** Points show mean Q5–Q1 realized five-session return differences in percentage points, with horizontal 95% interval bars. Panels follow Figure C.2; colors identify the same models. Separation summarizes the endpoints of each model's ranking and does not incorporate portfolio execution or trading costs.

color shading connects their pointwise 95% interval endpoints. The intervals describe uncertainty in individual estimates; comparisons of method differences require paired contrasts. These views make intermediate-group behavior and interval overlap visible without changing the underlying estimates.

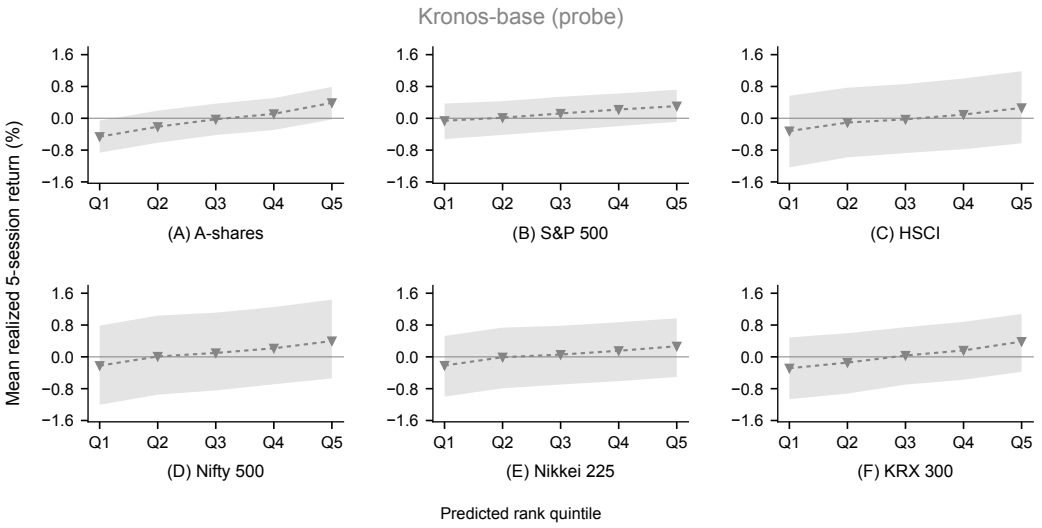


Figure C.4: **Kronos-base (probe): ranking and realized returns.** (A) A-shares, (B) S&P 500, (C) HSCI, (D) Nifty 500, (E) Nikkei 225, (F) KRX 300. Lines connect quintile means; same-color bands connect pointwise 95% interval endpoints. All model-specific figures share axis limits.

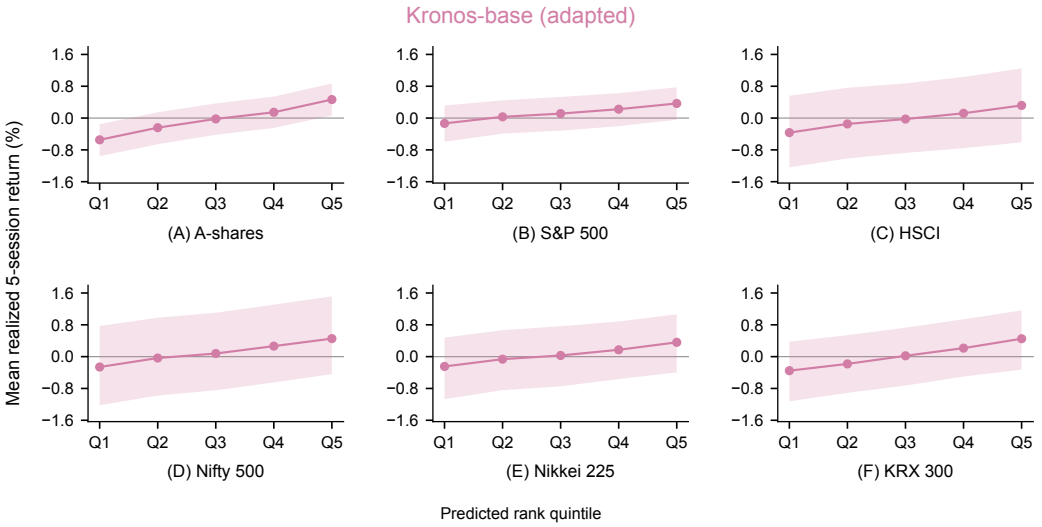


Figure C.5: **Kronos-base (adapted): ranking and realized returns.** (A) A-shares, (B) S&P 500, (C) HSCI, (D) Nifty 500, (E) Nikkei 225, (F) KRX 300. Lines connect quintile means; same-color bands connect pointwise 95% interval endpoints. All model-specific figures share axis limits.

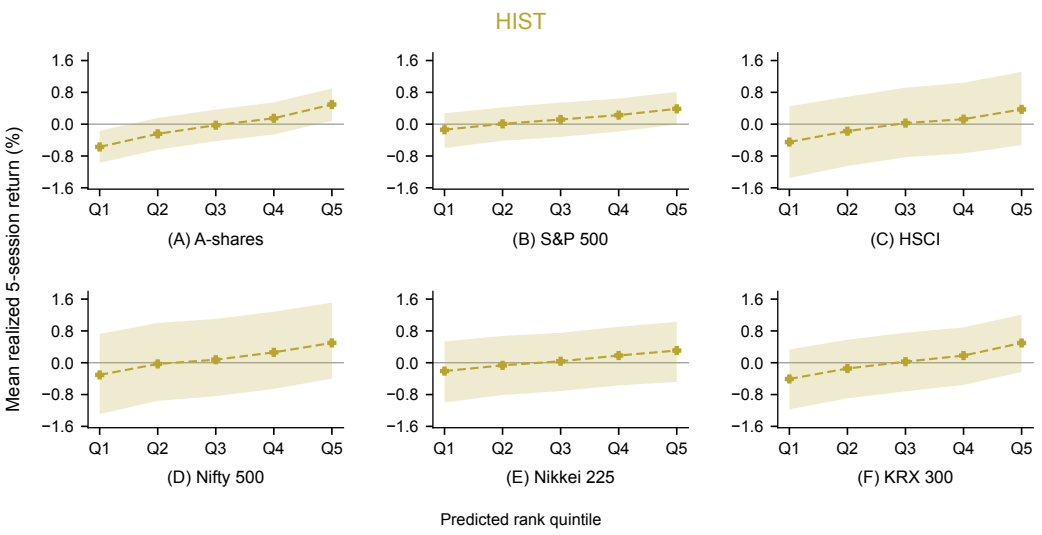


Figure C.6: **HIST: ranking and realized returns.** (A) A-shares, (B) S&P 500, (C) HSCI, (D) Nifty 500, (E) Nikkei 225, (F) KRX 300. Lines connect quintile means; same-color bands connect pointwise 95% interval endpoints. All model-specific figures share axis limits.

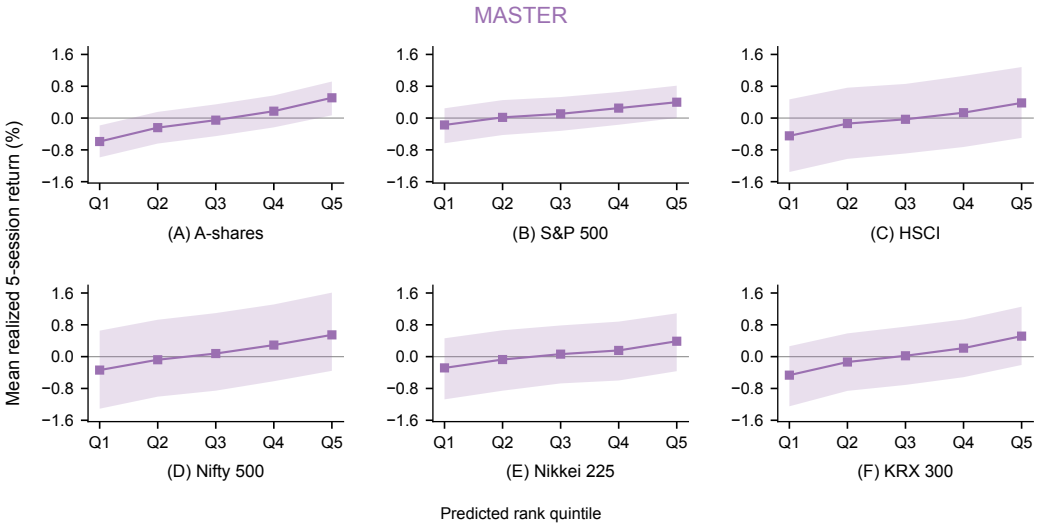


Figure C.7: **MASTER: ranking and realized returns.** (A) A-shares, (B) S&P 500, (C) HSCI, (D) Nifty 500, (E) Nikkei 225, (F) KRX 300. Lines connect quintile means; same-color bands connect pointwise 95% interval endpoints. All model-specific figures share axis limits.

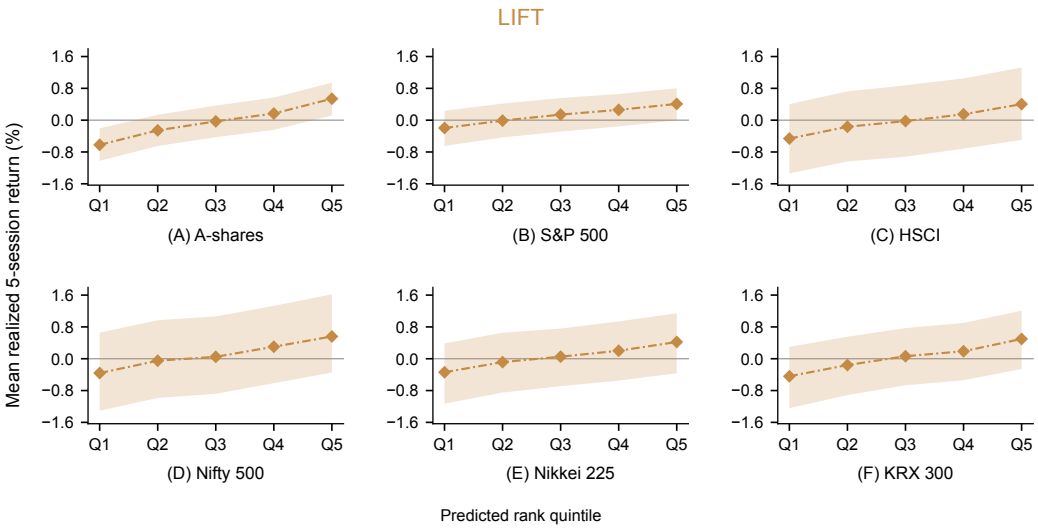


Figure C.8: **LIFT: ranking and realized returns.** (A) A-shares, (B) S&P 500, (C) HSCI, (D) Nifty 500, (E) Nikkei 225, (F) KRX 300. Lines connect quintile means; same-color bands connect pointwise 95% interval endpoints. All model-specific figures share axis limits.

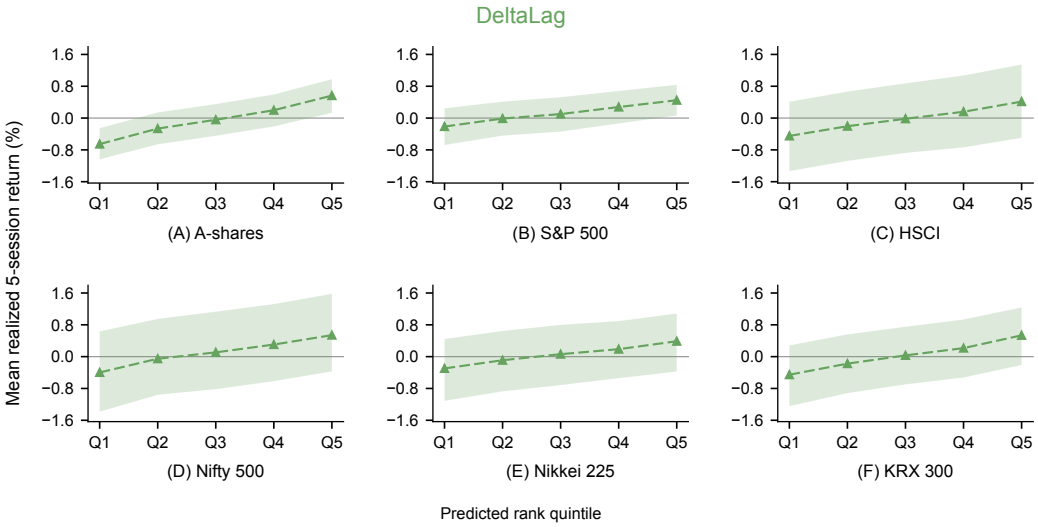


Figure C.9: **DeltaLag: ranking and realized returns.** (A) A-shares, (B) S&P 500, (C) HSCI, (D) Nifty 500, (E) Nikkei 225, (F) KRX 300. Lines connect quintile means; same-color bands connect pointwise 95% interval endpoints. All model-specific figures share axis limits.

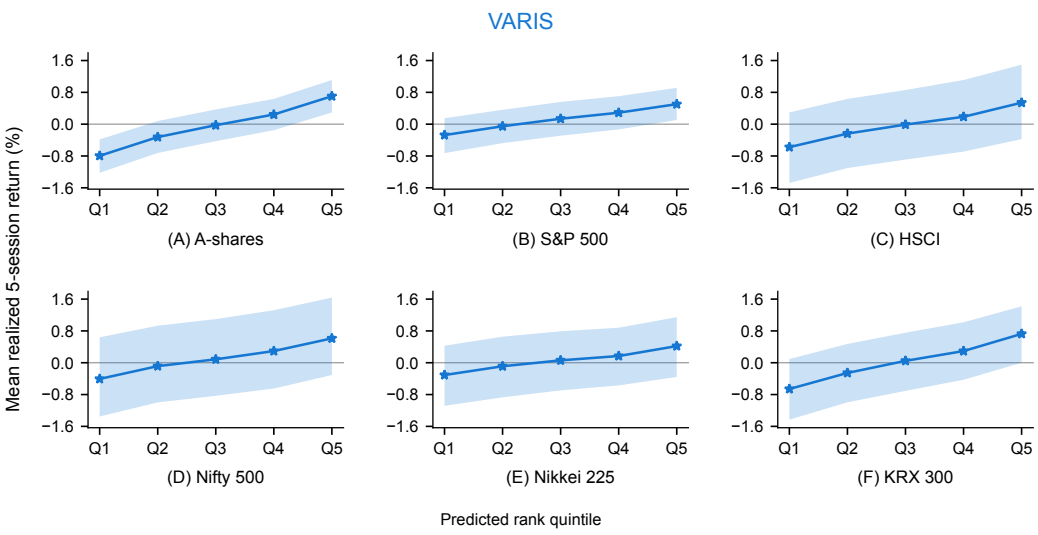


Figure C.10: **VARIS: ranking and realized returns.** (A) A-shares, (B) S&P 500, (C) HSCI, (D) Nifty 500, (E) Nikkei 225, (F) KRX 300. Lines connect quintile means; same-color bands connect pointwise 95% interval endpoints. All model-specific figures share axis limits.

D TARGET-CONDITIONED MARKET STRUCTURE ACROSS TIME

Market-state inference assigns a distinct role to each source asset and historical position in a target's forecast. The following cases examine these dimensions over 128 trading sessions, aligning source distributions with market-event timelines. Each market contains the same 64 sources for all three targets, with a fixed ordering and a shared color scale. This arrangement separates target identity from changes in the displayed universe.

Reading the representations. The left panels display a dense source diagnostic: query–key cosine compatibility is maximized over lags $\{1, 2, 5\}$ and normalized across eligible sources with temperature 0.34. The right panels show cosine compatibility for all strictly past source–target positions. These full-window diagnostics describe representation structure; the SLRA inference procedure in Section 3.4 uses its specified active states, candidate selection, historical assessment, and NULL competition. Historical positions are represented within the information set of the displayed window endpoint.

D.1 A-SHARES: MARKET POLICY AND TARGET-SPECIFIC STRUCTURE

The A-share window spans 14 March–15 September 2023, within the Test calendar interval. On 27 August, the Ministry of Finance and the State Taxation Administration announced a reduction of securities transaction stamp duty, effective 28 August.[1] Figure D.1 places this date against the source distributions for CITIC Securities, East Money, and ICBC.

Figure D.1 presents two levels of structure. Securities targets emphasize the broker group, while ICBC emphasizes bank sources. Within the broker group, weight concentrates on the CITIC–East Money pair toward the end of the window, following the marked policy date. Comparing the two directions distinguishes target-specific retrieval from a symmetric sector association. The bank target supplies a within-market comparison under the same calendar and source availability. This organization makes the paper's modeling question concrete: shared market conditions and target-specific historical evidence occupy different roles in a forecast.

[1] https://fgk.chinatax.gov.cn/zcfgk/c102416/c5211343/content.html

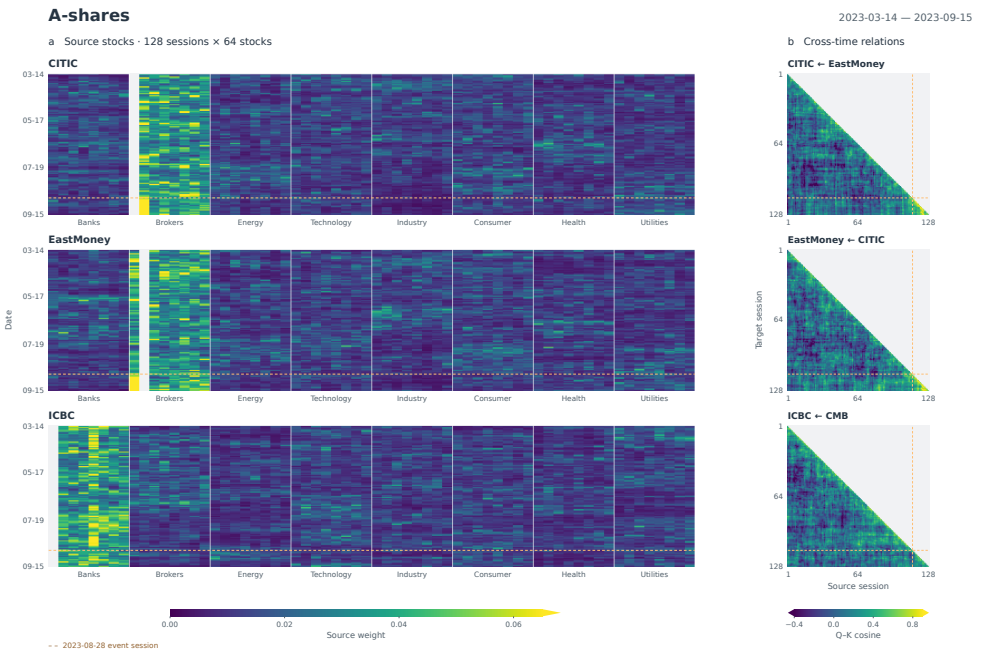


Figure D.1: A-share source and temporal structure, 14 March–15 September 2023. Left: source distributions for three targets over 128 sessions. Right: directional cross-time compatibility. The dashed line marks 28 August.

D.2 U.S. EQUITIES: COMPANY INFORMATION AND DIRECTIONAL RETRIEVAL

The U.S. window spans 13 December 2022–16 June 2023, also within the Test calendar interval. NVIDIA released its fiscal-year 2024 first-quarter results after the close on 24 May 2023; 25 May is the first subsequent trading session.[2] Figure D.2 compares AMD and NVIDIA with JPMorgan Chase under this timeline.

In Figure D.2, AMD's retrieval from NVIDIA is more concentrated after the marked earnings date than retrieval in the reverse direction. JPMorgan Chase instead emphasizes the financial group. The paired views distinguish three questions: which business group provides compatible representations, which individual source receives weight, and which historical positions support that compatibility. Their joint presentation clarifies why a target-conditioned source–lag state carries more information than an undirected stock grouping.

Connection to the empirical questions. These cases specify the structure that relational market-state modeling seeks to expose: target-dependent sources, temporally localized compatibility, and directional historical retrieval. The prediction comparisons and interventions in Appendix B address predictive value. Source permutations test source identity, lag interventions test temporal alignment, and paired removal of relational updates tests their contribution to the same targets. This connects the source and temporal dimensions of market structure to the forecasting objective.

D.3 SOURCE IDENTITIES

Figures D.3 and D.4 expand all source names from the corresponding overview panels. They use identical values, dates, and source ordering, allowing a business-group pattern to be traced to an individual stock. They are detailed views of the same two cases.

[2] https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-first-quarter-fiscal-2024

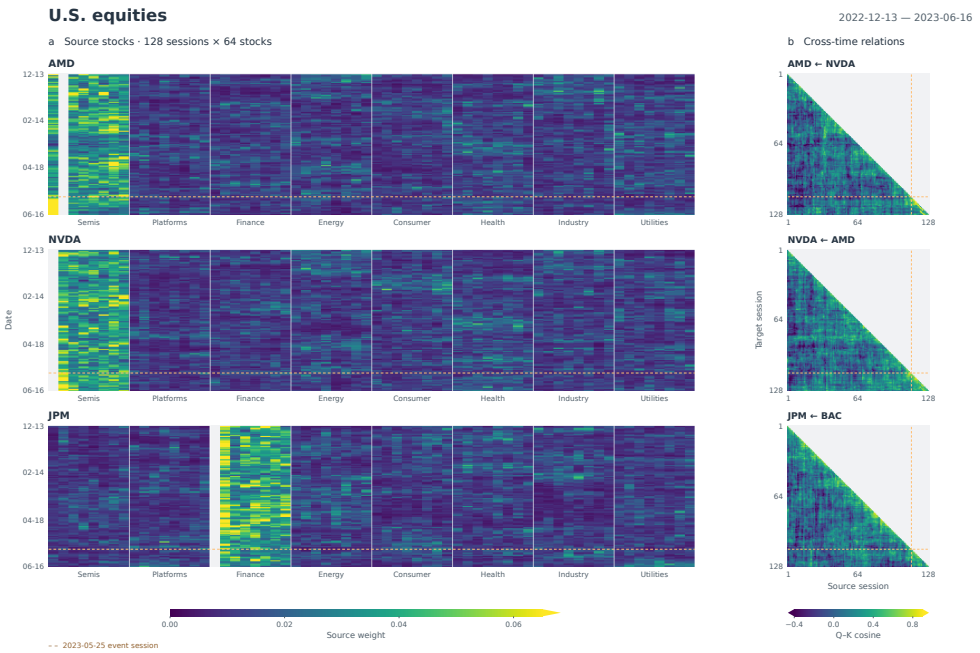


Figure D.2: U.S. source and temporal structure, 13 December 2022–16 June 2023. AMD and NVIDIA provide a directional pair; JPMorgan Chase provides a financial-sector comparison. The dashed line marks 25 May.

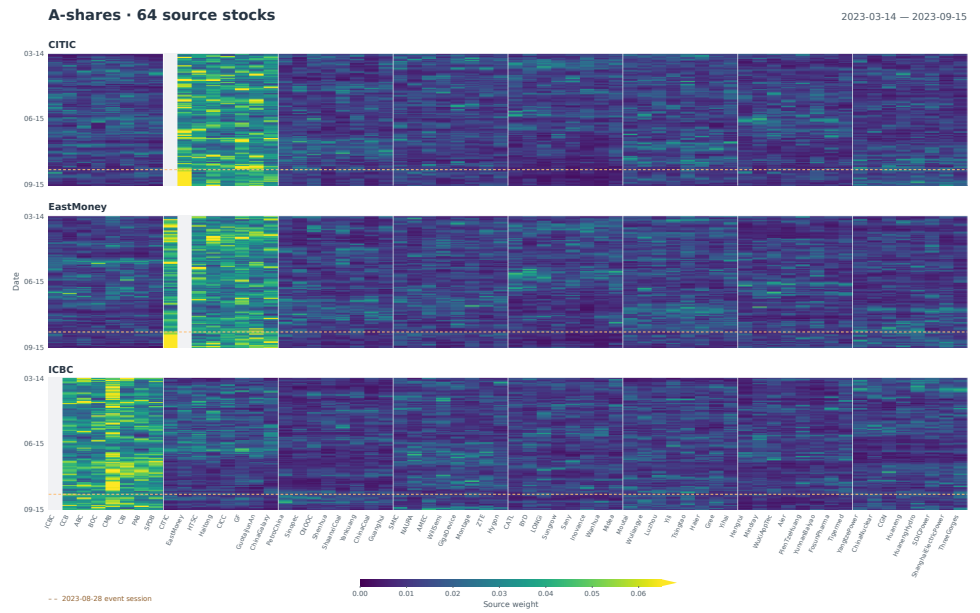


Figure D.3: Complete source labels for the A-share case in Figure D.1. Rows of panels correspond to CITIC Securities, East Money, and ICBC; all panels share the source order and color scale.

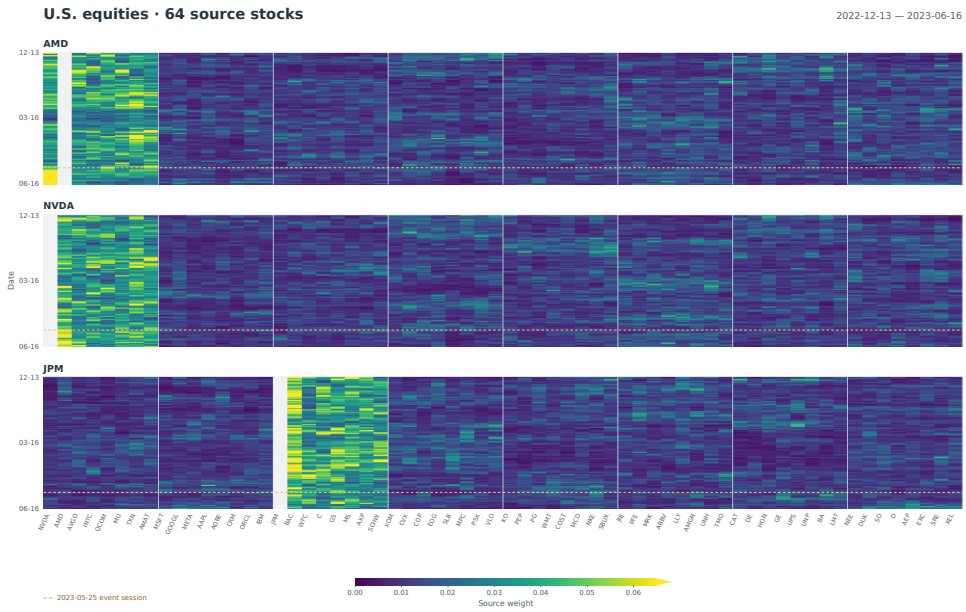


Figure D.4: Complete source labels for the U.S. case in Figure D.2. Rows of panels correspond to AMD, NVIDIA, and JPMorgan Chase, using the same values as the overview.

E Additional related work

VARIS connects transferable temporal representations with prediction-time inference of common market components and useful lagged sources. The following comparisons locate this combination within work on pretraining, cross-asset structure, temporal alignment, and representation selection.

E.1 Temporal pretraining and relational adaptation

TimesFM and Chronos learn transferable forecasting models from collections of time series (Das et al., 2024; Ansari et al., 2024). MOMENT broadens pretrained representation learning across tasks, while Timer develops generative pretraining for time-series analysis (Goswami et al., 2024; Liu et al., 2024b). Time-MoE explores sparse expert scaling, and Sundial extends large-scale generative forecasting (Shi et al., 2025; Liu et al., 2025b). Financial specialization includes Kronos's K-line tokenization, FinCast's financial pretraining, and SSPT's stock-specific pretraining tasks (Shi et al., 2026; Zhu et al., 2025; Wang et al., 2025).

Cross-series inference is also part of this landscape: Moirai accommodates an arbitrary number of variates, and Chronos-2 supports grouped series and covariates (Woo et al., 2024; Ansari et al., 2025). GraFT provides a closer structural precedent, injecting representations from a heterogeneous patch graph with static and adaptive edges into a pretrained Transformer's patch embeddings (Yuan et al., 2026). VARIS places its adaptation after a complete frozen-backbone pass. Its inferred market state separates common context from lagged residual evidence and controls the contribution of that evidence to each target forecast. This positions the contribution at the level of relational adaptation and evidence use, beyond the general combination of pretraining and cross-series information.

E.2 Shared factors and residual information

Gu et al. (2020) establish the predictive value of nonlinear interactions among return predictors. Factor models organize shared information more explicitly: Deep Factors combines global latent structure with local uncertainty, while predictable latent-factor learning decomposes observed series into forecastable components (Wang et al., 2019; Hou et al., 2024). In stock prediction, FactorVAE learns probabilistic dynamic factors, and HIST separates concept-oriented shared information from individual stock information (Duan et al., 2022; Xu et al., 2021). FactorGCL is especially relevant:

it extracts hidden factors from residuals after accounting for prior factors and uses temporal residual contrastive learning (Duan et al., 2025).

VARIS assigns its learned common state C_t two linked roles: supplying prediction context and defining the reference for relational residuals. Those residuals determine both source-lag scores and retrieved source values (Section 3.3). The methodological distinction lies in using the shared component to condition subsequent relation inference. The separate common-prediction and residualization controls in Section 4.3 examine these roles independently.

E.3 Cross-asset interactions and sources of relation information

Relational stock predictors differ in how they obtain and propagate structure. RSR incorporates temporal stock relations into ranking, while STHAN-SR models higher-order stock relations through a spatiotemporal hypergraph (Feng et al., 2019; Sawhney et al., 2021). HATR combines prior and adaptively learned graphs; MDGNN models multiple relation types as they evolve (Wang et al., 2021; Qian et al., 2024). MASTER learns interactions across stocks and time (Li et al., 2024b). COGRASP adds another source of evidence by constructing stock co-occurrence graphs from reports, newspapers, and social media (Li et al., 2025).

These approaches already establish cross-asset modeling and adaptive relations. VARIS infers its relation state from synchronized market histories and their pretrained representations. Its central design question is which target-specific evidence remains useful after accounting for common movement. This connects relation learning to both the residual reference and the representation used to assess each source.

E.4 Multivariate dependence, temporal alignment, and acceptance

PatchTST uses shared weights with channel-independent patches, iTransformer attends across variate tokens, and TimeXer incorporates exogenous variables into endogenous forecasting (Nie et al., 2023; Liu et al., 2024a; Wang et al., 2024). Explicit temporal alignment further refines cross-series dependence. LIFT estimates leading indicators and their leading steps, while DeltaLag uses sparse attention to select dynamic stock pairs and retrieve lag-aligned features (Zhao & Shen, 2024; Zhou et al., 2025). MillGNN learns lead-lag dependencies across grouping scales, and TLGNN models delayed relations with a learned graph (Wu et al., 2025; Feng et al., 2025). CausalStock combines lag-dependent temporal causal discovery with news information (Li et al., 2024a).

VARIS builds on this established role of dynamic lags. SLRA assesses candidates using residual evidence and consistency across preceding windows, then compares retained source-lag pairs with a NULL alternative in one distribution (Section 3.4). Its conditional weights q allocate source content, while acceptance ρ scales the relational update. This formulation couples temporal alignment to an explicit decision about the amount of relational evidence entering the forecast. The resulting weights describe predictive contribution; they do not establish causal transmission between stocks.

E.5 Variable asset sets and changing distributions

Deep Sets and Set Transformer provide architectures for permutation-invariant set processing (Zaheer et al., 2017; Lee et al., 2019). Set-Sequence brings this structure directly to financial panels: a learned cross-sectional summary conditions per-unit temporal prediction and supports a varying number of units (Epstein et al., 2025). VARIS combines common context with target-specific lagged retrieval, making the available source set part of relation inference.

Changing distributions create a related challenge. RevIN removes and restores instance statistics, while Non-stationary Transformers preserve dependency information that can be lost through stationarization (Kim et al., 2022; Liu et al., 2022). TimeBridge distinguishes short-term stationarity from longer-term cross-variate cointegration (Liu et al., 2025a). VARIS addresses changing evidence through market-conditioned routing and historical relation assessment. The later-date, held-out-asset, and source-retention evaluations in Section 4.5 examine complementary changes in the inputs available to this mechanism.

E.6 Intermediate representations and coupled relation inference

With Head2Toe, Evci et al. (2022) show how intermediate features from a frozen model support transfer beyond the final layer. Attention Residuals uses input-dependent attention over preceding layer outputs within a Transformer (Kimi Team et al., 2026). These precedents motivate selective reuse of depth, with different locations and purposes for aggregation.

In VARIS, asset- and path-conditioned routers mix snapshots from a complete backbone pass (Section 3.2). Common-path routing changes the residual reference; relational-path routing changes candidate scores, historical support, and retrieved values. The forecasting loss therefore couples the representation used to assess a source with the decision to use that source (Section 3.5). The routing–selection comparison in Section 4.4 examines this specific interaction. Together, these connections locate VARIS at the intersection of common-component modeling, selective lagged inference, and state-dependent reuse of pretrained representations.

F EVALUATION METRICS AND INTERPRETATION

The forecasting evaluation connects three quantities: the ordering of stocks, the accuracy of their predicted returns, and the realized returns associated with score-ranked groups. This section defines the corresponding metrics and explains how to read Tables 1–2 and Figure 3.

F.1 THE EVALUATION UNIT AND PREDICTION TARGETS

Each evaluation unit is a market-date cross-section at a specified forecast horizon h. For one market, let $\mathcal{E}_t^{(h)} \subseteq \mathcal{A}_t$ denote the eligible target stocks with available forward returns, and let $n_t = |\mathcal{E}_t^{(h)}|$. The return label is $y_{i,t}^{(h)} = p_{i,t+h}/p_{i,t} - 1$, where $t + h$ advances by h sessions on that market's trading calendar. A five-session label therefore describes the cumulative price change over five trading sessions. Its value is expressed as a decimal: 0.02 represents a return of 2%.

The ranking head produces $\hat{r}_{i,t}$, whose ordering determines stock preference. The return head produces $\hat{y}_{i,t}$ in the same units as $y_{i,t}^{(h)}$. Correlation metrics use the ranking-head score; error metrics use the return-head output. All methods are compared on the same eligible targets. The source universe $\mathcal{A}_t$ is defined by observable inputs, whereas $\mathcal{E}_t^{(h)}$ additionally requires a valid evaluation label. This distinction allows an asset with usable history to supply relational evidence even when its forward return is unavailable for scoring.

Target chunks are joined before evaluation so that each date contributes one complete cross-section. Stocks are compared within their market and forecast date, and the resulting daily statistics are averaged over dates. This gives each forecast date equal weight even when the number of eligible stocks changes.

F.2 IC AND RANKIC: LINEAR ASSOCIATION AND STOCK ORDERING

The information coefficient (IC) is the Pearson correlation between scores and realized returns within a date. Suppressing the horizon index on y for readability, its daily value is

$$\mathrm{IC}_t = \frac{\sum_{i\in\mathcal{E}_t}(\hat{r}_{i,t} - \bar{\hat{r}}_t)(y_{i,t} - \bar{y}_t)}{\sqrt{\sum_{i\in\mathcal{E}_t}(\hat{r}_{i,t} - \bar{\hat{r}}_t)^2}\sqrt{\sum_{i\in\mathcal{E}_t}(y_{i,t} - \bar{y}_t)^2}}, \tag{14}$$

where both means are taken over the eligible cross-section. Positive IC indicates that higher scores tend to accompany higher subsequent returns. IC uses the numerical distances between observations, so both stock ordering and the spacing of scores and returns influence its value.

RankIC is Spearman correlation, obtained by applying Pearson correlation to the two rank vectors:

$$\mathrm{RankIC}_t = \mathrm{corr}_{i\in\mathcal{E}_t}(\mathrm{rank}_t(\hat{r}_{i,t}), \mathrm{rank}_t(y_{i,t}))\,. \tag{15}$$

Ranks increase from low to high, and tied values receive their average rank. A strictly increasing transformation of the scores preserves RankIC because it preserves their ordering. IC and RankIC both lie in $[-1, 1]$: positive values indicate agreement, negative values indicate reversed association, and values near zero indicate weak association under the respective measure. Daily correlation is evaluated when at least two eligible observations and nonzero cross-sectional variation are present.

For a four-stock cross-section, score ranks $(1, 2, 3, 4)$ and realized-return ranks $(1, 3, 2, 4)$ exchange the middle two positions. The squared rank differences sum to 2, giving

$$\mathrm{RankIC}_t = 1 - \frac{6 \times 2}{4(4^2 - 1)} = 0.8. \tag{16}$$

This shortcut applies when there are no ties; Eq. 15 also handles average ranks. The calculation illustrates how RankIC measures agreement across the whole cross-section, including stocks outside the highest- and lowest-score groups.

F.3 Temporal aggregation, ICIR, and RankICIR

Let $\mathcal{T}_h$ contain the evaluable dates for one market and horizon. The table entries named IC and RankIC are temporal means of the daily coefficients:

$$\overline{\mathrm{IC}} = \frac{1}{|\mathcal{T}_h|} \sum_{t \in \mathcal{T}_h} \mathrm{IC}_t, \qquad \overline{\mathrm{RankIC}} = \frac{1}{|\mathcal{T}_h|} \sum_{t \in \mathcal{T}_h} \mathrm{RankIC}_t \,. \tag{17}$$

The associated information ratios divide each mean by the temporal standard deviation of its daily series:

$$\mathrm{ICIR} = \frac{\overline{\mathrm{IC}}}{\mathrm{sd}_{t \in \mathcal{T}_h}(\mathrm{IC}_t)}, \qquad \mathrm{RankICIR} = \frac{\overline{\mathrm{RankIC}}}{\mathrm{sd}_{t \in \mathcal{T}_h}(\mathrm{RankIC}_t)}. \tag{18}$$

These ratios summarize average predictive association relative to its fluctuation across dates. For a positive mean, a larger ratio can result from a higher mean, lower temporal dispersion, or both. ICIR uses the Pearson series; the main tables report RankICIR, which uses the Spearman series. Both definitions require a nonzero temporal standard deviation and apply no annualization factor.

Thus the A-shares RankIC of 0.0948 in Table 1 is an average cross-sectional rank correlation. Its RankICIR of 0.691 describes that average relative to daily variation. Neither statistic has return units. Realized return levels are summarized by the ranking-group profiles, while portfolio outcomes are evaluated through the allocation and execution protocol in Appendix C.3.

F.4 Return-estimation errors and displayed units

Mean squared error (MSE) and mean absolute error (MAE) assess the return head on each eligible cross-section:

$$\begin{aligned} \mathrm{MSE}_t &= \frac{1}{n_t} \sum_{i \in \mathcal{E}_t^{(h)}} \left(\hat{y}_{i,t} - y_{i,t}^{(h)} \right)^2, \\ \mathrm{MAE}_t &= \frac{1}{n_t} \sum_{i \in \mathcal{E}_t^{(h)}} \left| \hat{y}_{i,t} - y_{i,t}^{(h)} \right|. \end{aligned} \tag{19}$$

The reported error is the equal-date average of these daily quantities. MSE gives greater weight to large prediction errors through squaring; MAE measures average absolute error in return units. Both improve as they decrease. Because the two prediction heads serve different objectives, ordering quality and return-estimation accuracy provide complementary views of the forecast.

Table 2 multiplies MSE by 10^3 and MAE by 10^2 for display. For VARIS in A-shares, the displayed MSE of 2.359 corresponds to 0.002359 in squared decimal-return units. The displayed MAE of 3.543 corresponds to 0.03543 in decimal-return units, or an average absolute error of 3.543 percentage points. Correlation coefficients and information ratios retain their unscaled values.

F.5 Score quintiles and realized return separation

At date t, let $\mathcal{G}_{q,t}$ be the approximately equal-sized group with score quintile q, with Q1 containing the lowest scores and Q5 the highest. The realized group return and endpoint separation are

$$g_{q,t} = \frac{1}{|\mathcal{G}_{q,t}|} \sum_{i \in \mathcal{G}_{q,t}} y_{i,t}^{(h)}, \qquad d_t^Q = g_{5,t} - g_{1,t}. \tag{20}$$

Figure 3 reports the equal-date means of $g_{q,t}$ and d_t^Q. Each model forms its groups from its own scores on the common target universe. An increasing Q1–Q5 profile shows how score ordering separates subsequent realized returns; the endpoint difference focuses on the two ranking extremes.

Return separation has return units. The A-shares Q5–Q1 value of 1.499 percentage points corresponds to a difference of 0.01499 in decimal five-session returns. RankIC additionally measures ordering throughout the middle of the cross-section. The two summaries can therefore rank competing methods differently. The group statistic is computed directly from realized labels; Appendix C.3 specifies the separate mapping from scores to positions, realized portfolio returns, and transaction costs.

F.6 Absolute gains, relative gains, and uncertainty

An absolute metric gain is a subtraction in the metric's own units; a relative gain divides by the comparator. In A-shares, VARIS and adapted Kronos-base have RankIC values 0.0948 and 0.0667, respectively. Their absolute difference is 0.0281, and their relative difference is $(0.0948/0.0667 - 1) \times 100\% \approx 42.1\%$. These calculations concern the five-session A-shares setting in Table 1.

The abstract summarizes six markets and three horizons, giving $M = 18$ settings. For metric v, let v_a^{V} and v_a^{K} denote the VARIS and adapted Kronos-base values in setting a. The aggregate relative gain is

$$\mathrm{Gain}(v) = \left(\frac{M^{-1} \sum_{a=1}^{M} v_a^{\mathrm{V}}}{M^{-1} \sum_{a=1}^{M} v_a^{\mathrm{K}}} - 1 \right) \times 100\%. \tag{21}$$

Each market–horizon setting has equal weight. RankICIR is first computed within each setting and then included in this macro-average. This aggregation gives the reported 46.8% RankIC and 52.3% RankICIR improvements; it differs algebraically from averaging the 18 individual percentage improvements.

Paired uncertainty analysis retains the comparison at the date level. For RankIC, the daily contrast is $d_t = \mathrm{RankIC}_t^{\mathrm{V}} - \mathrm{RankIC}_t^{\mathrm{K}}$ on matched targets. Date-block resampling recomputes its mean while keeping predictions paired and retaining local temporal dependence. The same principle applies to Q5–Q1 differences, and RankICIR is recomputed from each resampled correlation series. Appendix C.1 describes the intervals for the model comparison; Appendix C gives the ranking-group bootstrap settings. The temporal standard deviation entering RankICIR and the uncertainty of a mean gain describe different statistical quantities.

G Understanding market-state inference

This section follows a VARIS forecast from its inputs through representation selection and relational inference to prediction. Table 7 summarizes the intermediate quantities.

Table 7: Notation for one market-date forecast. Vector dimensions refer to individual assets or asset-time positions; batching adds the corresponding leading axes.

Quantity	Role in inference
t, s, ℓ, h	Forecast date, historical position, source lag, and forward-return horizon, respectively.
$\mathcal{A}_t$, $\mathcal{E}_t^{(h)}$	Input-eligible source universe and label-eligible evaluation targets.
$H_{i,s\mid t}^{(k)} \in \mathbb{R}^D$	Backbone representation at depth k for position s in the input window ending at t.
$\alpha_{i,t,k}^{m}$	Depth weight for asset i and path m; weights sum to one over the K depths.
$C_t \in \mathbb{R}^{F \times D}$	Shared factor state used for common prediction and the relational residual reference.
$\mathcal{S}_{i,t}$	Retained source-lag pairs for target i, with at most S pairs.
$p_{i,t}^{\varnothing}$, $\rho_{i,t}$	NULL probability and total probability assigned to the retained relations, with $\rho = 1 - p^{\varnothing}$.
$q_{ij\ell,t}$, $\pi_{ij\ell,t}$	Conditional pair weight and joint pair probability, related by $\pi = \rho q$.
γ_r	Learned global residual-gate parameter; $\tanh(\gamma_r)$ scales the accepted relation branch.

G.1 Forecast time, source lag, and representation depth

The forecast horizon h determines the future return label. The source lag ℓ selects earlier evidence: candidate (j, ℓ) retrieves source j at $t - \ell$ for target i at t. The lag set $\{1, 2, 5\}$ can therefore accompany a five-session forecast, with lag and horizon controlling the evidence and target, respectively.

The 128-session input defines the backbone's historical information. Within it, the default cache allocates eight active states, $t - 7, \ldots, t$, and 21 preceding states, $t - 28, \ldots, t - 8$, to current

processing and historical assessment. The notation $s \mid t$ identifies position s as represented within the forecast window ending at t, including that window's normalization and routing. The forecast state uses observations available by t.

Four depth snapshots from the complete frozen backbone pass provide alternative features for each asset-time position. Each prediction path forms its own mixture, conditioned on the asset and market state, with weights shared across the accessed history. Temporal selection identifies the source position; depth selection determines its representation.

G.2 The sequence of operations for one prediction

The operations in Figure 2 proceed as follows.

1. **Form the input.** Align eligible histories and masks to the market calendar. Access the frozen backbone's layer snapshots and compute three path-specific depth mixtures using Eq. 2.
2. **Estimate shared context.** Compute panel context and C_t. Read common prediction features and construct relational residuals with the factor-conditioned reference in Eq. 4.
3. **Select candidates.** Score aligned target-source residuals, apply availability and self-relation masks, and retain up to $S = 2$ pairs. One source can occupy both slots at different lags.
4. **Refine evidence.** Assess retained pairs in two historical windows. Equation 7 rewards repeated support and discounts disagreement between windows.
5. **Allocate and accept content.** Normalize refined scores with NULL; compute q, ρ, and Δ^{rel} using Eqs. 9–11. An empty candidate set gives $\rho = 0$ and a zero update.
6. **Predict.** Fuse the paths through Eq. 12 and apply both prediction heads. Join target-chunk outputs before computing the market-date metrics in Appendix F.

Every target chunk retains the same eligible source pool and shared market context. Each target can select a source anywhere in that universe. Chunk size controls the working memory needed for candidate scoring while preserving this access.

G.3 Why the common state appears in two paths

The common state supplies both a predictive feature and a reference for relation inference. The common-path readout $c^{\mathrm{com}}_{i,t}$ enters the final predictor directly. The relational readout $c^{\mathrm{rel}}_{i,s|t}$ defines the decomposition

$$H^{\mathrm{rel}}_{i,s|t} = c^{\mathrm{rel}}_{i,s|t} + r_{i,s|t}. \tag{22}$$

All three terms are vectors in the same representation space. SLRA compares and retrieves the residual component, so a candidate's relevance is assessed relative to the learned common context. Common prediction can retain useful market-wide information while the relation path concentrates on the residual evidence associated with particular assets and lags.

The two controls in Table 3 act at different points in this computation. Removing common prediction disables the direct common contribution while preserving the residual reference. Removing residualization supplies total representations to relational scoring and retrieval. Reading these controls together distinguishes the value of predicting from common context from the value of using it to construct relational evidence.

G.4 Conditional selection, acceptance, and residual scaling

The relation distribution separates the source mixture from its total contribution. For a nonempty candidate set, q sums to one over retained pairs, while their joint probabilities sum to ρ. Substituting $\pi = \rho q$ into Eq. 11 gives

$$\Delta^{\mathrm{rel}}_{i,t} = W_o \sum_{(j,\ell)\in\mathcal{S}_{i,t}} \pi_{ij\ell,t} V \tilde{r}_{j,t-\ell|t}. \tag{23}$$

The NULL alternative receives probability mass and contributes a zero content vector. This identity makes the accepted update a mixture weighted by the joint probabilities, while the factorization into q and ρ exposes the two decisions for interpretation.

For example, joint probabilities $(p^{\varnothing}, \pi_1, \pi_2) = (0.75, 0.20, 0.05)$ give $\rho = 0.25$ and conditional weights $(q_1, q_2) = (0.8, 0.2)$. The first pair dominates the conditional mixture, but the complete relation branch receives only one quarter of that mixture's unscaled weight. Adding the same positive constant to both source logits preserves q and increases ρ relative to the unchanged NULL logit. Conditional preference and overall acceptance therefore respond to different aspects of the evidence.

Three sets of weights act on three distinct axes. Depth weights α select representations; conditional weights q allocate source-lag content; acceptance ρ controls the amount of that content for a particular target and date. The additional factor $\tanh(\gamma_r)$ in Eq. 12 sets the learned global scale and sign of the relation branch. In particular, forcing $\rho = 1$ retains source-lag selection while removing the target-specific acceptance decision.

G.5 How the two prediction objectives supervise the state

The ranking objective compares valid pairs within a market-date. For a pair whose realized returns satisfy $y_i > y_j$, its contribution is $\log(1 + \exp[-(\hat{r}_i - \hat{r}_j)])$. Increasing the correctly ordered score margin decreases this loss. Pairs with equal targets are excluded, and a cross-section with no valid ordered pair contributes zero ranking loss, as specified in the main objective.

The return objective supervises the numerical forecast through the Huber loss. For prediction error $u = \hat{y} - y$ and threshold $\delta > 0$, its per-target form is

$$\mathrm{Huber}_\delta(u) = \begin{cases} \frac{1}{2}u^2, & |u| \le \delta, \\ \delta\left(|u| - \frac{1}{2}\delta\right), & |u| > \delta. \end{cases} \tag{24}$$

It provides quadratic supervision near the target and linear growth for larger errors. The coefficient λ_{ret} balances this return objective with the ranking term. The ranking head consequently learns relative preference and the return head learns return magnitude, using the shared state $z_{i,t}$.

Both objectives supervise the adapted prediction state while the backbone remains frozen. Common-path routing changes the factors that define the residual reference; relational-path routing changes the representations used for compatibility, historical assessment, and retrieved values. These dependencies connect depth selection and relation inference through the final forecast, giving the routing–selection comparison its mechanistic interpretation.

G.6 Reading the routing–selection interaction

Rows A–D of Table 3 form a two-by-two comparison: static or dynamic depth routing, crossed with forced or selective relation updates. Let $v_A, \ldots, v_D$ denote one metric in these rows. Dynamic routing changes that metric by $v_C - v_A$ under forced updates and by $v_D - v_B$ under selective updates. Their interaction contrast is

$$I = (v_D - v_B) - (v_C - v_A) = (v_D - v_C) - (v_B - v_A). \tag{25}$$

The equivalent forms ask whether the routing gain changes with acceptance, or whether the acceptance gain changes with routing. A positive value indicates that the combined gain exceeds the sum of the two separate gains measured from the static/forced configuration.

For A-shares RankIC, the dynamic-routing gains are $0.0948 - 0.0847 = 0.0101$ with selective updates and $0.0831 - 0.0776 = 0.0055$ with forced updates, giving $I = 0.0046$. The same calculation gives $I = 0.0042$ in S&P 500. These absolute RankIC contrasts describe the coupling assessed in Section 4.4: adapting the representation used to judge a relation is more useful when the forecast also learns how much of that relation to accept.