Align Before Aggregation: Component-Conditioned Evidence Alignment for Multimodal Financial Forecasting

Anonymous authors
Paper under double-blind review

Abstract

Different temporal dynamics within a market window may require different news, yet text-enhanced forecasting typically organizes evidence at the stock or window level. While frequency-aware methods exploit multiscale structure to improve representations and fusion, we examine whether evidence should be aligned separately to temporal components before aggregation. Under dot-product matching and linear query aggregation over a shared candidate set, we show that global alignment produces a normalized weighted geometric mean of component-level alignment distributions, allowing opposing evidence preferences to cancel. Motivated by this observation, CANER follows an *align-before-aggregation* principle. It decomposes the observed price–volume window into low-, mid-, and high-frequency components and calibrates their queries using shared semantic anchors constructed by deterministically verbalizing observable market states. Each component aligns with and selectively absorbs evidence from the same historical news pool before the enhanced states are aggregated for prediction. On CMIN-US, CMIN-CN, and FNSPID, CANER achieves the highest ACC in all nine dataset–horizon settings, exceeding the strongest baseline by 0.34 percentage points on average, and the highest MCC in seven settings. Comparisons with matched parameter counts and news-slot budgets, together with cross-component mismatch analysis, support the predictive value of preserving component–evidence correspondence before final aggregation.

1 Introduction

Financial forecasting combines price–volume histories with news events to understand observed market changes and their external information context (Tetlock, 2007; Ding et al., 2015; Xu & Cohen, 2018). A single market window may contain persistent trends, intermediate adjustments, and short-lived disturbances, whose news requirements need not coincide: changes in business conditions may relate to persistent trends, while localized events may primarily concern short-term fluctuations. A news item highly relevant to one component need not be equally relevant to others. The question is therefore not only which news relates to the current stock, but also *which components of its market state the news should be aligned with*.

Existing work improves text-enhanced forecasting through event representations, inter-stock interactions, and lag modeling (Luo et al., 2023; Li et al., 2024; Wang et al., 2024a), while multiscale models improve time-series representations through decomposition and cross-scale interactions (Wu et al., 2021; Zhou et al., 2022; Wang et al., 2024b; Chen et al., 2024). Related cross-modal methods also exploit frequency structure: SpecTF attends to text through frequency-conditioned attention, while VoT fuses frequency components of event-driven and numerical forecasts (Nguyen et al., 2026; Wang et al., 2026). Building on these developments, we examine *alignment granularity and aggregation order* when forming news contexts: should component queries first be combined to establish unified news correspondences, or should each component form and absorb its own context before the enhanced states are jointly used for prediction? We examine correspondence formation separately from subsequent component aggregation and use controlled comparisons to assess its role in prediction.

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.

Under dot-product matching over a shared candidate set and linear query aggregation, Proposition 1 shows that the global alignment distribution is the normalized weighted geometric mean of component-level distributions, revealing how opposing preferences can cancel during aggregation. Motivated by this observation, CANER follows the principle of **align before aggregation**. It decomposes the observed price–volume window into low-, mid-, and high-frequency components, constructs shared semantic anchors by deterministically verbalizing historical technical states, and uses them to calibrate component queries. Each component separately matches and selectively absorbs evidence from the same historical news pool. The enhanced states are then jointly used for prediction, preserving component–evidence correspondence throughout within-component interaction.

Across nine forecasting settings on CMIN-US, CMIN-CN, and FNSPID, CANER achieves the highest ACC, exceeding the strongest baseline by 0.34 percentage points on average. Comparisons with matched capacity and news-selection slot budgets further examine the role of correspondence. On CMIN-US, fixing the model, numerical components, and selected news while changing only the receiving components reduces ACC by 0.66 percentage points, supporting a dependence not only on the available news but also on its correspondence with components. Analyses of news matching, topics, and lags further characterize the evidence contexts formed by different components. Our main contributions are as follows:

- **Alignment granularity and aggregation order.** We identify evidence-alignment granularity as a modeling dimension in multimodal time-series forecasting and propose establishing component–evidence correspondences before final aggregation.
- **Analysis of preference aggregation in global alignment.** Under the specified dot-product matching and linear query aggregation assumptions, we characterize the geometric aggregation relation in global alignment and the cancellation of opposing evidence preferences.
- **CANER instantiation and controlled validation.** We instantiate this principle using shared semantic anchors and component-conditioned interaction, evaluate prediction on three financial datasets, and examine correspondence through budget- and capacity-matched comparisons and cross-component mismatch analysis on CMIN-US.

2 Related Work

2.1 Text-Enhanced Financial Forecasting

Text-enhanced financial forecasting has progressed from associations between media sentiment and market behavior (Tetlock, 2007) to jointly modeling textual events and price–volume histories. Event-driven methods use textual events for prediction, with fine-grained models further preserving event types and attributes (Ding et al., 2015; Chen et al., 2019). StockNet jointly models tweets and prices, while MAN-SF combines social media information with inter-stock relationships (Xu & Cohen, 2018; Sawhney et al., 2020). FinBERT improves text representations through financial-domain pretraining (Araci, 2019). This direction also extends to macroeconomic announcement-driven forecasting (Zhang et al., 2025) and quarterly earnings prediction from financial sequences and long documents (Koval et al., 2024).

Another line of work examines temporal effects and relational structure. FAST connects news timeliness with stock ranking (Sawhney et al., 2021), while prediction tasks also provide a basis for evaluating news relevance (Del Corro & Hoffart, 2021). CMIN models text and inter-stock relationships through multi-memory interactions, CausalStock combines lagged inter-stock causal structure with news denoising, and FININ captures inter-news relationships and their delayed market effects (Luo et al., 2023; Li et al., 2024; Wang et al., 2024a). These studies organize financial information through content, timing, and relational structure, providing a foundation for examining component–news correspondences within a market window.

2.2 Multiscale Time-Series Modeling

Fourier analysis characterizes frequency composition, while wavelet analysis additionally captures scale differences and temporal locality (Mallat, 1989). For numerical time series, Autoformer combines trend–seasonal decomposition with autocorrelation, and FEDformer further exploits frequency-domain representations (Wu et al., 2021; Zhou et al., 2022). PatchTST and iTransformer organize

Transformer representations around local patches and inter-variable relationships, respectively (Nie et al., 2023; Liu et al., 2024). Pathformer, TimesNet, and TimeMixer model temporal dynamics through multiscale pathway selection, multiple periods, and cross-scale mixing, respectively (Chen et al., 2024; Wu et al., 2023; Wang et al., 2024b).

Frequency structure also supports cross-modal interaction. SpecTF uses frequency components as queries to attend to text features, forming frequency-dependent cross-modal representations (Nguyen et al., 2026). VoT combines frequency components of event-driven and numerical forecasts through adaptive frequency fusion (Wang et al., 2026). The former explicitly models relevance between different frequencies and text, whereas the latter emphasizes complementary scales across forecast components, illustrating the roles of frequency structure at the representation-interaction and forecast-fusion stages.

2.3 Cross-Modal Alignment and Fusion

Cross-modal modeling involves aligning representation spaces and enabling interaction with textual information. Time-LLM uses text prototypes to reprogram time-series patches into a language model's input space (Jin et al., 2024). CALF enforces cross-modal consistency at the input, feature, and output levels, while TimeCMA aligns dual-branch representations of numerical sequences and their textual prompts (Liu et al., 2025b;a). TRACE learns a time-series–text semantic space for bidirectional matching (Chen et al., 2025). For information interaction, TimeCAP uses language-model-generated contextual summaries to support event prediction, and TaTS treats paired text as a time-varying auxiliary variable (Lee et al., 2025; Li et al., 2026). Time-VLM fuses temporal, visual, and textual representations, whereas Cross-MoE models dynamic cross-modal associations through ranking and mixture-of-experts networks (Zhong et al., 2025; Huang et al., 2025). These studies connect time series and text through representation mappings and information interaction. We further examine the order of evidence alignment and state aggregation: states can first be aggregated to form a shared news context, or each component can align and interact with news before the enhanced states are aggregated. CANER adopts the latter path and uses controlled comparisons to examine the predictive role of component–news correspondences under this ordering.

3 Problem Formulation and Alignment Order Analysis

3.1 Problem Formulation

For stock s at trading date t, let $X_{s,t} \in \mathbb{R}^{L\times F}$ denote the historical input window and H the prediction horizon. Using the adjusted closing price $P_{s,t}$, we define

$$y_{s,t,H} = \mathbb{I}\big(\log P_{s,t+H} - \log P_{s,t} > 0\big), \qquad p_{s,t,H} = P(y_{s,t,H} = 1 \mid X_{s,t}, \mathcal{N}_{s,t}). \tag{1}$$

Here, $\mathcal{N}_{s,t}$ contains historical news associated with stock s and available before the prediction cutoff. The objective is to estimate the probability of an upward movement using historical price, volume, and news information. All inputs and derived representations are restricted to information available at the prediction cutoff. Stock and date subscripts are omitted below.

3.2 Aggregation Order and Evidence Preference Cancellation

Correspondences between multiscale states and text can be established by aggregating states before forming a shared context, or by aligning components separately before fusion. To isolate this ordering difference, we fix the component queries, text representations, and candidate set, comparing only dot-product alignment distributions before and after query aggregation.

Suppose a sequence contains K scale components with encodings $h^{(k)}$. Each component has a query $q^{(k)} \in \mathbb{R}^d$ in the same matching space as the text keys. Let $e_j \in \mathbb{R}^d$ and v_j denote the key and value of text item j. Define the global query as $q^{\mathrm{g}} = \sum_k w_k q^{(k)}$, where $w_k > 0$ and $\sum_k w_k = 1$. For the same N text items, both paths use dot-product scores with temperature $\sqrt{d}$:

$$\alpha_{kj} = \frac{\exp({q^{(k)}}^\top e_j/\sqrt{d})}{\sum_{\ell=1}^{N}\exp({q^{(k)}}^\top e_\ell/\sqrt{d})}, \qquad \alpha_{\mathrm{g},j} = \frac{\exp({q^{\mathrm{g}}}^\top e_j/\sqrt{d})}{\sum_{\ell=1}^{N}\exp({q^{\mathrm{g}}}^\top e_\ell/\sqrt{d})}. \tag{2}$$

Proposition 1 (Geometric aggregation of global alignment) Under the above assumptions, the global alignment weights are the normalized weighted geometric mean of the component-level weights:

$$\alpha_{\mathrm{g},j} = \frac{\prod_{k=1}^{K} \alpha_{kj}^{w_k}}{\sum_{\ell=1}^{N} \prod_{k=1}^{K} \alpha_{k\ell}^{w_k}}. \tag{3}$$

Consequently, for any two text items i, j,

$$\log \frac{\alpha_{\mathrm{g},i}}{\alpha_{\mathrm{g},j}} = \sum_{k=1}^{K} w_k \log \frac{\alpha_{ki}}{\alpha_{kj}}. \tag{4}$$

The full proof and derivation of preference cancellation appear in Appendix C.

Interpreting preference cancellation. Equation (4) states that the global log-preference ratio for any news pair is the weighted average of the corresponding component-level ratios. When components prefer opposite items, positive and negative terms offset one another. If their weighted sum is zero, global alignment assigns equal weights to the two items even when individual components have clear preferences. This phenomenon is cross-component evidence preference cancellation, occurring before a shared news context is formed.

From analysis to design. Component-level alignment forms separate contexts $c^{(k)} = \sum_j \alpha_{kj} v_j$ so that a subsequent model can process each context together with its state $h^{(k)}$, rather than receiving only $c^{\mathrm{g}} = \sum_j \alpha_{\mathrm{g},j} v_j$ from a single preference distribution. Final aggregation remains necessary but can be deferred until within-component interaction is complete. This preserves the opportunity to exploit component–evidence correspondence; predictive gains still depend on whether the correspondence differences are relevant to the target. The next section instantiates align before aggregation in CANER. Section 5.3 uses cross-component mismatch to examine whether changing component–news correspondence affects prediction.

4 Method

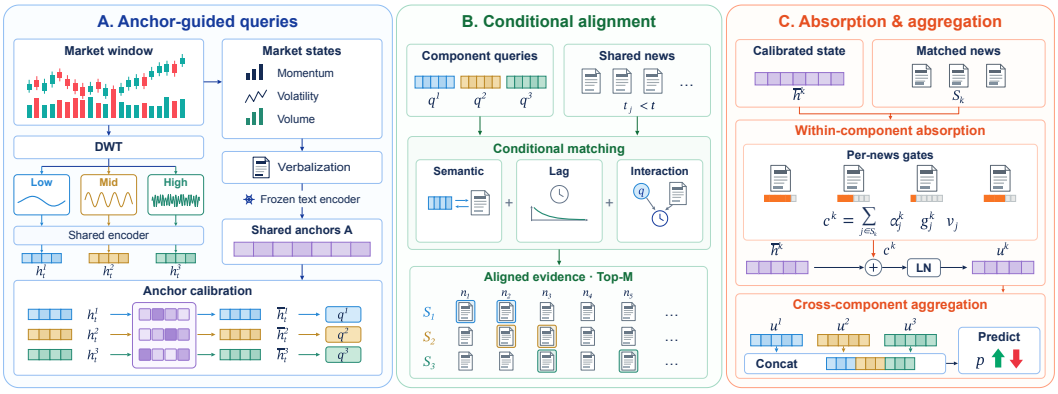


Figure 1: Overview of CANER. Market-state semantic anchors calibrate component queries; each component matches and selectively absorbs evidence from a shared historical news pool before enhanced states are aggregated for direction prediction.

Building on the alignment-order analysis in Section 3, CANER establishes frequency-specific component–news correspondences and then aggregates the news-enhanced component states. As shown in Figure 1, the price–volume window is decomposed and encoded into component states, whose queries are calibrated using shared market-state semantic anchors. Each query forms a matched set from a common historical news pool; per-item absorption gates regulate evidence contributions to the receiving component, and the prediction head jointly models the enhanced states.

4.1 Frequency Decomposition and Component Encoding

To distinguish dynamics at different time scales within a window, we apply a two-level `db2` discrete wavelet transform with symmetric boundary extension along the temporal axis of each numerical feature (Mallat, 1989). Reconstructing the second-level approximation, second-level detail, and first-level detail coefficients separately yields low-, mid-, and high-frequency components:

$$(X^{(1)}, X^{(2)}, X^{(3)}) = \mathcal{D}(X), \qquad X \approx \sum_{k=1}^{3} X^{(k)}, \qquad h^{(k)} = f_\theta(X^{(k)}, p_k). \tag{5}$$

All reconstructed components retain the input length and feature dimension; decomposition uses only the current visible window. The shared temporal encoder f_θ maps each sequence to a state $h^{(k)} \in \mathbb{R}^{d_h}$, with learnable component identifier p_k. This fixed three-band partition defines the state granularity for alignment, with low-, mid-, and high-frequency states separately conditioning news matching and absorption; component-count analysis appears in Appendix D.

4.2 Market-State Anchors and Component Calibration

Anchors verbalize observable market conditions, providing a shared semantic reference for numerical components and news. For each date ℓ in the window, we compute technical states F_ℓ, including momentum, volatility, and volume anomalies, from price and volume history available no later than that date. A deterministic template produces a short text $a_\ell = T(F_\ell)$, such as *short-term momentum is positive, volatility is elevated, and trading volume is unusually high*. Anchors and news share the same frozen FinBERT encoder ψ; ordering anchor vectors by date gives the memory $A = [\psi(a_1), \ldots, \psi(a_L)]$.

Each component uses its own state as a query to read semantic context from the shared anchor memory. The context calibrates the numerical representation through a gated residual update, followed by a linear projection to form the news-matching query:

$$\beta^{(k)} = \mathrm{softmax}\left(\frac{(W_{AQ}h^{(k)})^\top W_{AK}A}{\sqrt{d_a}}\right), \qquad m^{(k)} = W_{AV}A(\beta^{(k)})^\top, \tag{6}$$

$$\eta^{(k)} = \sigma\Big(W_\eta[h^{(k)}; m^{(k)}] + b_\eta\Big), \qquad \bar{h}^{(k)} = \mathrm{LN}\Big(h^{(k)} + \eta^{(k)} \odot m^{(k)}\Big), \tag{7}$$

$$q^{(k)} = W_Q \bar{h}^{(k)}. \tag{8}$$

Here, $\beta^{(k)}$ is normalized over anchor dates, d_a is the attention key dimension, and W_{AV} maps the context to d_h dimensions. The element-wise gate $\eta^{(k)}$ controls calibration, and $q^{(k)} \in \mathbb{R}^d$ is the matching query. Component-specific attention weights connect the shared market-state semantic reference with individual dynamics, forming conditional queries for news alignment.

4.3 Component-Conditioned News Alignment

All components access the same historical news pool $\mathcal{N}$ for the target stock, restricted to candidates available before the prediction cutoff. Let n_j denote the raw news text and m_j its observable metadata, including source, event type, calendar, and lag features. The news projection network R_ω constructs representation r_j and produces a matching key and aggregation value:

$$r_j = R_\omega([\psi(n_j); m_j]), \qquad e_j = W_K r_j, \qquad v_j = W_V r_j. \tag{9}$$

The key $e_j \in \mathbb{R}^d$ follows the notation in Section 3, while $v_j \in \mathbb{R}^{d_h}$ is used to update component states. Beyond dot-product matching, the score incorporates news lag and query–news interaction:

$$s_j^{(k)} = \frac{{q^{(k)}}^\top e_j}{\sqrt{d}} - \lambda_k \log(1 + \delta_j) + \phi_\xi(q^{(k)}, r_j, \delta_j), \tag{10}$$

$$\lambda_k = \mathrm{softplus}(\ell_k), \qquad \mathcal{S}_k = \mathrm{TopM}_{j \in \mathcal{N}}(s_j^{(k)}). \tag{11}$$

Here, δ_j is the lag in days, and the nonnegative coefficient λ_k regulates temporal preference. The shared interaction network ϕ_ξ takes the query and news representation of the same dimensionality,

their element-wise product and absolute difference, and logarithmic lag as inputs. The score depends jointly on the query state, news content, and temporal information, allowing the same news item to form distinct correspondences with different components.

Each component independently selects at most $M = 8$ news items, giving at most 24 selection slots across the three components. The sets $\mathcal{S}_k$ may overlap: the same item can be selected by multiple components, with absorption strength computed separately within each component.

4.4 Within-Component Absorption and Final Aggregation

For each component, the model normalizes matching scores within the selected set and computes a scalar absorption gate for each item from the component state, news content, and lag. Matching weights $\alpha_j^{(k)}$ express relative relevance within the set, and gate values $g_j^{(k)}$ are produced independently by sigmoid functions. Together, they regulate news contributions to the component update:

$$\alpha_j^{(k)} = \mathrm{softmax}_{j \in \mathcal{S}_k}(s_j^{(k)}), \tag{12}$$

$$g_j^{(k)} = \sigma\Big(W_g[\bar{h}^{(k)}; v_j; \bar{h}^{(k)} \odot v_j; \log(1+\delta_j)] + b_g\Big), \tag{13}$$

$$c^{(k)} = \sum_{j \in \mathcal{S}_k} \alpha_j^{(k)} g_j^{(k)} v_j, \qquad u^{(k)} = \mathrm{LN}(\bar{h}^{(k)} + c^{(k)}). \tag{14}$$

Here, $c^{(k)}$ is the absorbed news context; small gate values weaken the textual residual. If no news is available, we set $c^{(k)} = 0$.

After news absorption within each component, enhanced states are concatenated in a fixed frequency order. The prediction head jointly models the component states and their news contexts to output the probability of an upward movement:

$$p = \sigma\Big(H_\zeta([u^{(1)}; u^{(2)}; u^{(3)}])\Big). \tag{15}$$

The model is trained with binary cross-entropy while the text encoder remains frozen. Discrete Top-M selection is not differentiated; gradients propagate through selected scores and the subsequent absorption and prediction modules. Network parameters for temporal encoding, anchor calibration, news projection, matching interaction, and absorption are shared across components; component identifiers p_k and decay parameters ℓ_k are branch-specific.

5 Experiments

5.1 Experimental Setup

Data and task. We use three stock-price and news datasets, CMIN-US, CMIN-CN, and FNSPID (Luo et al., 2023; Dong et al., 2024). Given price–volume sequences and news from the past $L = 30$ trading days, the task is to predict the return direction over the next $H \in \{1, 3, 5\}$ trading days. Data are split by target date; numerical features are processed using training-set statistics only, and news is restricted to target-stock records available before the prediction cutoff. Dataset sizes, preprocessing, and training settings are provided in Appendix A.

Comparison protocol. Baselines include iTransformer, StockNet, CMIN, CausalStock, TaTS, VoT, and SpecTF (Liu et al., 2024; Xu & Cohen, 2018; Li et al., 2024; 2026; Wang et al., 2026; Nguyen et al., 2026). Main results report test ACC and MCC over five random seeds. Architectural variants are retrained with identical splits, seeds, and checkpoint-selection rules; cross-component mismatch changes news assignments at fixed checkpoints. Controlled comparisons of global and component-level alignment retain the same three price-representation branches, parameter count, candidate pool, and temporal cutoff, with 24 news-selection slots in total.

5.2 Predictive Performance

Table 1 summarizes results across three datasets and three horizons, covering financial forecasting, price-only, and text-enhanced time-series methods.

Table 1: Main results against representative baselines over $H = 1, 3, 5$. ACC is reported in percentage (%), while MCC is reported in its original scale. Results are reported as mean $\pm$ standard deviation over five runs. Best means across both panels are bold.

(A) Financial forecasting baselines

Dataset	H	CANER (Ours)		CausalStock Li et al. (2024)		CMIN Luo et al. (2023)		StockNet Xu & Cohen (2018)	
		ACC (%)	MCC	ACC (%)	MCC	ACC (%)	MCC	ACC (%)	MCC
CMIN-US	1	**54.82 ± 0.40**	**0.0574 ± 0.0106**	54.58 ± 0.60	0.0473 ± 0.0078	53.39 ± 0.96	0.0416 ± 0.0143	52.46 ± 0.92	0.0234 ± 0.0074
	3	**54.19 ± 0.37**	**0.0426 ± 0.0077**	53.76 ± 0.94	0.0371 ± 0.0123	52.55 ± 1.01	0.0279 ± 0.0140	51.86 ± 0.75	0.0132 ± 0.0063
	5	**53.49 ± 0.59**	**0.0319 ± 0.0068**	52.89 ± 0.83	0.0260 ± 0.0150	51.81 ± 0.95	0.0136 ± 0.0092	51.05 ± 0.96	0.0019 ± 0.0070
CMIN-CN	1	**56.88 ± 0.38**	0.0846 ± 0.0081	56.11 ± 0.65	**0.1202 ± 0.0851**	54.33 ± 0.97	0.0825 ± 0.0141	53.38 ± 0.67	0.0440 ± 0.0030
	3	**55.88 ± 0.39**	0.0723 ± 0.0064	55.18 ± 0.96	0.0728 ± 0.0421	53.67 ± 1.05	**0.0736 ± 0.0070**	52.76 ± 1.07	0.0361 ± 0.0053
	5	**54.96 ± 0.51**	**0.0666 ± 0.0120**	54.24 ± 0.91	0.0608 ± 0.0440	52.68 ± 1.15	0.0579 ± 0.0084	52.04 ± 0.76	0.0206 ± 0.0087
FNSPID	1	**62.69 ± 0.37**	**0.1089 ± 0.0094**	60.69 ± 0.73	0.0958 ± 0.0098	58.92 ± 0.65	0.0984 ± 0.0088	54.96 ± 0.74	0.0435 ± 0.0049
	3	**61.83 ± 0.51**	**0.1033 ± 0.0070**	59.78 ± 0.69	0.0869 ± 0.0100	58.49 ± 0.76	0.0822 ± 0.0160	54.35 ± 1.06	0.0274 ± 0.0076
	5	**61.16 ± 0.57**	**0.0862 ± 0.0073**	58.86 ± 0.89	0.0738 ± 0.0156	57.37 ± 0.95	0.0709 ± 0.0161	53.51 ± 0.89	0.0197 ± 0.0054

(B) Time-series and time-series–text baselines

Dataset	H	CANER (Ours)		SpecTF Nguyen et al. (2026)		VoT Wang et al. (2026)		TaTS Li et al. (2026)		iTransformer Liu et al. (2024)	
		ACC (%)	MCC	ACC (%)	MCC	ACC (%)	MCC	ACC (%)	MCC	ACC (%)	MCC
CMIN-US	1	**54.82 ± 0.40**	**0.0574 ± 0.0106**	54.66 ± 0.53	0.0437 ± 0.0094	54.33 ± 0.66	0.0441 ± 0.0120	54.63 ± 0.70	0.0345 ± 0.0140	52.32 ± 0.64	0.0096 ± 0.0046
	3	**54.19 ± 0.37**	**0.0426 ± 0.0077**	53.89 ± 0.46	0.0414 ± 0.0103	53.53 ± 0.98	0.0364 ± 0.0119	52.35 ± 1.02	0.0228 ± 0.0152	51.22 ± 0.91	0.0027 ± 0.0074
	5	**53.49 ± 0.59**	**0.0319 ± 0.0068**	53.03 ± 0.78	0.0285 ± 0.0142	52.78 ± 0.75	0.0189 ± 0.0090	52.23 ± 1.17	0.0089 ± 0.0163	50.63 ± 1.10	0.0013 ± 0.0087
CMIN-CN	1	**56.88 ± 0.38**	0.0846 ± 0.0081	56.61 ± 0.49	0.0782 ± 0.0096	55.36 ± 1.01	0.0719 ± 0.0113	54.12 ± 1.00	0.0605 ± 0.0071	53.03 ± 0.87	0.0189 ± 0.0055
	3	**55.88 ± 0.39**	0.0723 ± 0.0064	55.54 ± 0.56	0.0681 ± 0.0089	54.49 ± 0.92	0.0599 ± 0.0125	53.18 ± 0.76	0.0539 ± 0.0099	52.19 ± 0.75	0.0114 ± 0.0064
	5	**54.96 ± 0.51**	**0.0666 ± 0.0120**	54.51 ± 0.67	0.0587 ± 0.0118	53.52 ± 1.02	0.0488 ± 0.0128	52.24 ± 0.95	0.0421 ± 0.0082	51.28 ± 0.83	0.0011 ± 0.0039
FNSPID	1	**62.69 ± 0.37**	**0.1089 ± 0.0094**	62.43 ± 0.48	0.1016 ± 0.0107	60.31 ± 0.91	0.0983 ± 0.0130	58.40 ± 0.79	0.0817 ± 0.0109	54.89 ± 0.97	0.0329 ± 0.0089
	3	**61.83 ± 0.51**	**0.1033 ± 0.0070**	61.47 ± 0.62	0.0945 ± 0.0095	59.21 ± 0.86	0.0855 ± 0.0152	57.74 ± 0.70	0.0682 ± 0.0102	53.85 ± 0.92	0.0156 ± 0.0044
	5	**61.16 ± 0.57**	**0.0862 ± 0.0073**	60.67 ± 0.71	0.0789 ± 0.0111	58.57 ± 0.74	0.0748 ± 0.0147	56.70 ± 0.96	0.0600 ± 0.0162	53.25 ± 0.87	0.0036 ± 0.0055

CANER achieves the highest ACC in all nine settings and the highest MCC in seven. SpecTF is the strongest ACC baseline in all nine settings. CANER improves over it by 0.34 percentage points on average, with gains of 0.16–0.49 points across settings. Performance generally declines with longer horizons, while CANER retains its ACC advantage at each horizon. Metric rankings differ on CMIN-CN: CANER achieves the highest ACC throughout, whereas CausalStock and CMIN obtain the highest MCC at $H = 1$ and $H = 3$, respectively. Overall, the results show that component-conditioned news modeling provides useful directional signals across markets and prediction horizons.

5.3 Alignment Granularity and Component–Evidence Correspondence

We examine whether component-level news contexts improve prediction and whether their assignment to specific components matters.

Table 2: Alignment controls for Section 5.3 ($H = 1$). (A) Five-seed means; (B) CMIN-US means $\pm$ standard deviations. ACC (%); best values in bold.

(A) Alignment granularity

Dataset	Global		CANER	
	ACC	MCC	ACC	MCC
CMIN-US	53.42	0.0358	**54.82**	**0.0574**
CMIN-CN	56.24	0.0758	**56.88**	**0.0846**
FNSPID	60.72	0.0812	**62.69**	**0.1089**

(B) Component–news correspondence

Condition	ACC	MCC
Matched	**54.82 ± 0.40**	**0.0574 ± 0.0106**
Mismatch	54.16 ± 0.47	0.0481 ± 0.0118

Global versus component-level alignment. Global Alignment selects Top-24 news with a unified query and forms a shared context; CANER selects Top-8 per component for within-component interaction and absorption. Both use the same three price branches, prediction architecture, candidate pool, and prediction cutoff, with matched parameter counts and 24 selection slots. Component selections may overlap, with each selection counted toward the slot budget. Table 2(A) shows CANER's ACC gains of 1.40, 0.64, and 1.97 percentage points on CMIN-US, CMIN-CN, and FNSPID, respectively, and MCC gains of 0.0216, 0.0088, and 0.0277.

Cross-component mismatch with fixed news. Using each trained checkpoint without retraining, we keep numerical components and selected news fixed, cyclically reassign news sets across branches, and recompute interaction and absorption in the receiving components. Results average both non-identity cyclic permutations over identical samples and five paired seeds. On CMIN-US, mismatch reduces ACC by 0.66 percentage points and MCC by 0.0093 (Table 2(B)). With news content and budget unchanged, this decrease supports the predictive role of component–evidence correspondence.

Together, these comparisons support component-specific context formation and evidence assignment before final aggregation.

5.4 Aligned Evidence Analysis

We characterize component-conditioned news contexts through state matching, topic composition, and lag distributions.

State–news matching. Following LLM-based pairwise evaluation (Zheng et al., 2023), we sample 100 stock–prediction-time–frequency-component units from companies with relatively abundant news. GPT-5.5 blindly compares eight top-ranked items from the global result with eight from the specified component, with presentation order swapped to check consistency. Aggregating market-state relevance, financial-event specificity, and news complementarity, the overall preference for component-level alignment has a 95% confidence interval of $[0.60, 0.71]$, above the neutral level of 0.5. The judgments indicate better matching between component-level contexts and the given market states in this evaluation. Dimension-wise statistics, date-block bootstrap, and evaluation templates are reported in Appendix B.

Table 3: Topic shares (%) of top-ranked news under each alignment method.

Alignment	Earnings/finance	Policy/industry	Products/operations	Legal/firm risk	Trading/sentiment
Global alignment	29.7	19.4	20.8	14.6	15.5
Low-frequency	41.6	27.9	15.3	9.7	5.5
Mid-frequency	22.4	18.6	34.7	14.8	9.5
High-frequency	12.7	10.9	18.4	24.6	33.4

Topic composition.
We use an LLM to classify top-ranked news into five topics and compare their shares under global and component-level alignment. Category definitions and the classification protocol appear in Appendix B.3. Table 3 shows how news content varies across frequency branches.

Earnings and finance account for 41.6% of low-frequency news, compared with 29.7% under global alignment. The mid-frequency branch emphasizes product and operating events, while trading and sentiment account for 33.4% of high-frequency news. These differences show that conditioning on distinct components organizes the same historical news pool into contexts with different thematic emphases.

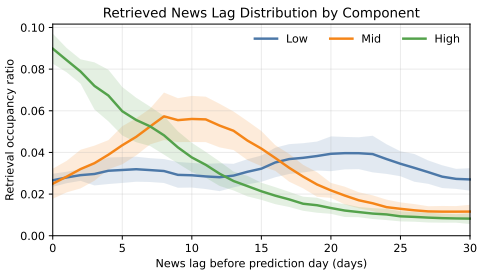


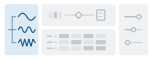
Figure 2: News lag distributions across frequency branches. The x-axis denotes news lag relative to the prediction day, and the y-axis shows the normalized share of matched news. For each branch, the curve represents the mean across test samples, and the shaded band indicates $\pm$ one sample standard deviation.

News lag distribution. Figure 2 shows that the high-frequency branch favors recent news, the mid-frequency branch peaks at intermediate lags, and the low-frequency profile is flatter and spans a broader historical range. Together with the topic differences, these temporal patterns characterize the evidence used by individual components. As matching depends on both component states and learnable lag terms, the observed profiles reflect their joint effect on alignment.

5.5 Ablation and Sensitivity Analysis

Key designs. Table 4 examines decomposition, semantic anchors, and absorption gates. Price-only retains frequency-based price encoding and prediction; w/o Decomposition replaces frequency components with the original sequence while retaining three branches; w/o Anchor constructs queries directly from numerical components; w/o Absorption fixes per-item gates to 1 while retaining matching weights. All variants are retrained under the same protocol.

Table 4: Ablation results on three datasets at $H = 1$. Blue denotes retained modules; gray indicates removal or bypass. ACC is reported as a percentage; best results are bold.

Architecture						
Variant		Price-only	w/o Decomp.	w/o Anchor	w/o Absorption	CANER
CMIN-US	ACC	51.73	53.68	54.07	54.36	**54.82**
	MCC	0.0187	0.0429	0.0461	0.0496	**0.0574**
CMIN-CN	ACC	53.82	55.96	55.31	56.09	**56.88**
	MCC	0.0324	0.0697	0.0573	**0.0853**	0.0846
FNSPID	ACC	57.46	61.38	59.87	62.16	**62.69**
	MCC	0.0418	0.0876	0.0693	0.1027	**0.1089**

The full model achieves the highest ACC on all three datasets. Removing decomposition or anchors reduces both ACC and MCC, indicating that component representations and their semantic calibration jointly support effective use of news. Absorption gates improve both metrics on CMIN-US and FNSPID. On CMIN-CN, they improve ACC, while fixed gates yield MCC of 0.0853, slightly above the full model's 0.0846.

Component count. Appendix D compares prespecified frequency partitions from $K = 1$ to $K = 5$. The strongest results occur at $K = 3$ or $K = 4$, with both metrics declining at five components. Finer partitions therefore do not continuously improve prediction, and the preferred granularity varies across datasets and metrics. Slot budgets and prediction-head configurations are reported alongside the results in the appendix.

Top-3 portfolios constructed from prediction scores, together with APV and Sharpe-ratio evaluations, are reported in Appendix E.

6 Conclusion

We study evidence-alignment granularity and aggregation order in multimodal financial forecasting. Under dot-product matching and linear query aggregation over a shared candidate set, we show that global alignment forms a normalized weighted geometric mean of component-level distributions, allowing opposing evidence preferences to cancel. CANER follows an *align-before-aggregation* principle: frequency components separately align with and selectively absorb news evidence before their enhanced states are aggregated for prediction. Results on three datasets, together with budget- and capacity-matched comparisons and cross-component evidence-mismatch tests on CMIN-US, support the predictive value of preserving component–evidence correspondence. State–news matching, topic, and lag analyses characterize the resulting evidence contexts; structural ablations and partition sensitivity examine design choices, while portfolio evaluation provides a complementary downstream assessment. Together, these findings highlight alignment granularity and aggregation order as consequential design choices in multimodal financial forecasting.

AI Assistance Disclosure

Generative AI tools were used to assist with language polishing and translation, related-work discovery, research ideation, and experimental implementation. GPT-5.5 was used for blinded pairwise evaluation of news–market state matching, and LLMs were also used for news topic classification, as detailed in Appendix B. The authors verify AI-assisted content and take full responsibility for the final manuscript and research conclusions.

REFERENCES

Dogu Araci. FinBERT: Financial sentiment analysis with pre-trained language models, 2019. URL https://arxiv.org/abs/1908.10063.

Deli Chen, Yanyan Zou, Keiko Harimoto, Ruihan Bao, Xuancheng Ren, and Xu Sun. Incorporating fine-grained events in stock movement prediction. In *Proceedings of the Second Workshop on Economics and Natural Language Processing*, pp. 31–40, 2019. doi: 10.18653/v1/D19-5105.

Jialin Chen, Ziyu Zhao, Gaukhar Nurbek, Aosong Feng, Ali Maatouk, Leandros Tassiulas, Yifeng Gao, and Rex Ying. TRACE: Grounding time series in context for multimodal embedding and retrieval, 2025. URL https://arxiv.org/abs/2506.09114.

Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo. Pathformer: Multi-scale transformers with adaptive pathways for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lJkOCMP2aW.

Luciano Del Corro and Johannes Hoffart. From stock prediction to financial relevance: Repurposing attention weights to assess news relevance without manual annotations. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pp. 45–49, 2021. doi: 10.18653/v1/2021.econlp-1.6.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2327–2333, 2015. URL https://www.ijcai.org/Proceedings/15/Papers/329.pdf.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. FNSPID: A comprehensive financial news dataset in time series. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4918–4927, 2024. doi: 10.1145/3637528.3671629.

Ruizheng Huang, Zhicheng Zhang, and Yong Wang. Cross-MoE: An efficient temporal prediction framework integrating textual modality. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29927–29938, 2025. doi: 10.18653/v1/2025.emnlp-main.1520. URL https://aclanthology.org/2025.emnlp-main.1520/.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Unb5CVPtae.

Ross Koval, Nicholas Andrews, and Xifeng Yan. Financial forecasting from textual and tabular time series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8289–8300, 2024. doi: 10.18653/v1/2024.findings-emnlp.486. URL https://aclanthology.org/2024.findings-emnlp.486/.

Geon Lee, Wenchao Yu, Kijung Shin, Wei Cheng, and Haifeng Chen. TimeCAP: Learning to contextualize, augment, and predict time series events with large language model agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18082–18090, 2025. doi: 10.1609/aaai.v39i17.33989. URL https://ojs.aaai.org/index.php/AAAI/article/view/33989.

Shuqi Li, Yuebo Sun, Yuxin Lin, Xin Gao, Shuo Shang, and Rui Yan. CausalStock: Deep end-to-end causal discovery for news-driven stock movement prediction. In *Advances in Neural Information Processing Systems*, 2024.

Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik F. Hamann, Hanghang Tong, and Jingrui He. Language in the flow of time: Time-series-paired texts weaved into a unified temporal narrative. In *The Fourteenth International Conference on Learning Representations*, 2026. doi: 10.48550/arXiv.2502.08942.

Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. TimeCMA: Towards LLM-empowered multivariate time series forecasting via cross-modality alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18780–18788, 2025a. doi: 10.1609/aaai.v39i18.34067.

Peiyuan Liu, Hang Guo, Tao Dai, Naiqi Li, Jigang Bao, Xudong Ren, Yong Jiang, and Shu-Tao Xia. CALF: Aligning LLMs for time series forecasting via cross-modal fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18915–18923, 2025b. doi: 10.1609/aaai.v39i18.34082.

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*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

Di Luo, Weiheng Liao, Shuqi Li, Xin Cheng, and Rui Yan. Causality-guided multi-memory interaction network for multivariate stock price movement prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12164–12176, 2023.

Stephane G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989. doi: 10.1109/34.192463.

Huu Hiep Nguyen, Minh Hoang Nguyen, Dung Nguyen, and Hung Le. Spectral text fusion: A frequency-aware approach to multimodal time-series forecasting. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4213–4221. PMLR, 2026. URL https://proceedings.mlr.press/v300/nguyen26d.html.

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

Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Ratn Shah. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 8415–8426, 2020. doi: 10.18653/v1/2020.emnlp-main.676.

Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Ratn Shah. FAST: Financial news and tweet based time aware network for stock trading. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2164–2175, 2021. doi: 10.18653/v1/2021.eacl-main.185.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of IJCNLP-AACL 2025*, 2025. URL https://aclanthology.org/2025.ijcnlp-long.18/.

Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007. doi: 10.1111/j.1540-6261.2007.01232.x.

Mengyu Wang, Shay B. Cohen, and Tiejun Ma. Modeling news interactions and influence for financial market prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3302–3314, 2024a. doi: 10.18653/v1/2024.findings-emnlp.189.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=7oLshfEIC2.

Siyuan Wang, Peng Chen, Yihang Wang, Wanghui Qiu, Chenjuan Guo, Bin Yang, and Yang Shu. Unlocking the value of text: Event-driven reasoning and multi-level alignment for time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. doi: 10.48550/arXiv.2603.15452.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430, 2021.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw3840q.

Yumo Xu and Shay B. Cohen. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 1970–1979, 2018. doi: 10.18653/v1/P18-1183.

Yang Zhang, Wenbo Yang, Jun Wang, Qiang Ma, and Jie Xiong. CAMEF: Causal-augmented multimodality event-driven financial forecasting by integrating time series patterns and salient macroeconomic announcements. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025. doi: 10.1145/3711896.3736872.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.05685.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-VLM: Exploring multimodal vision-language models for augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78478–78497. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhong25a.html.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhou22g.html.

A Experimental Implementation Details

Dataset sizes and labels. CMIN-US and CMIN-CN contain 110 and 300 stocks, 110,879 and 201,654 prediction samples, and 871,759 and 1,344,781 raw news records, respectively. Our FNSPID subset contains 465 stocks, 639,720 samples, and 1,603,125 raw news records. News counts precede deduplication. Direction labels are computed from adjusted closing prices on the prediction date and the H-th subsequent trading day.

Point-in-time visibility and preprocessing. Training, validation, and test sets are divided by target date; samples crossing a split boundary are excluded. Numerical features and missing values are processed using per-stock training-set statistics, which remain fixed thereafter. Wavelet decomposition is performed within each visible window, and thresholds for verbalized technical states are estimated on training data only. News candidates are stock-specific and published before the prediction cutoff; same-day news is excluded when reliable intraday timestamps are unavailable. After exact deduplication, the shared news memory retains at most 64 candidates from a 30-day lookback window.

News-selection budgets and parameter matching. Component-level alignment selects the top 8 news items for each of three frequency branches, while global alignment selects the top 24 items with a unified query, giving both a total budget of 24 news-selection slots. News selected by different components may overlap; matching slot budgets therefore does not require identical numbers of distinct news items. Both settings use the same historical news candidate pool and temporal cutoff and have matched parameter counts.

Training and mechanism controls. CANER uses AdamW with an initial learning rate of 10^{-4}, weight decay of 10^{-5}, and batch size of 64, for at most 100 epochs. Checkpoints are selected by validation MCC, and main test results are summarized as mean and standard deviation over five random seeds. The $H = 1$ mechanism controls use five paired random seeds and the news-budget matching settings described above; raw-price and decomposed-price variants have matched branch counts and parameter capacity.

B News–Market State Matching Evaluation Protocol

LLM-based pairwise evaluation can be affected by presentation position and other factors (Zheng et al., 2023; Shi et al., 2025). We specify evaluation inputs, order-consistency checks, and structured outputs below; swapping presentation order identifies position-sensitive judgments but does not eliminate all evaluation biases.

Sampling and evaluation statistics. Evaluation samples are drawn from the ten companies with the largest news counts in CMIN-US, restricting the sampling scope to companies with relatively abundant news. We sample 100 stock–prediction-date–frequency-component units, comprising 40 low-, 30 mid-, and 30 high-frequency units. For each unit, we select the top 8 news items from global alignment and the corresponding frequency branch, presenting both sets anonymously using the same evaluation template. This setting fixes the set size for pairwise evaluation; the global prediction model still uses the top 24 news items, of which only the 8 highest-scoring items are used here.

Table 5: Blinded pairwise evaluation of news–market state matching. Each dimension starts with 100 units, and exclusion categories are mutually exclusive. The preference score is $(W_c + 0.5T)/N$; 95% confidence intervals are estimated by bootstrapping prediction dates.

Dimension	Valid	Component wins	Ties	Global wins	Position-bias exclusions	Other invalid	Preference score 95% CI
Market-state relevance	90	51	16	23	10	0	[0.62, 0.70]
Financial-event specificity	83	53	14	16	15	2	[0.68, 0.76]
News complementarity	91	48	18	25	9	0	[0.57, 0.68]

Evaluation inputs and statistical rules. Each evaluation unit corresponds to a stock, prediction time, and frequency component. We hold the market-state semantic anchor, news candidate pool, and temporal cutoff fixed and select equal numbers of top-ranked items from global and component-level alignment to form the two comparison sets. The sets are anonymously labeled A/B, and the LLM judge receives only information available at prediction time, not realized returns or model predictions. The model version, prompt, generation settings, and news truncation length are held fixed. The criteria are market-state relevance, financial-event specificity, and news complementarity; each judgment is restricted to A, B, or a tie.

Each unit is assessed again after swapping the A/B presentation order. Two judgments favoring the same retrieval method count as a win for that method; any tie is recorded as a tie. If both judgments favor the same presentation position, the unit is flagged for position bias and excluded from the valid set. Let W_c, T, and N denote component-level alignment wins, ties, and valid units, respectively; the preference score is $(W_c + 0.5T)/N$. Valid counts and win–tie–loss counts are aggregated separately for each dimension in Table 5; valid counts equal the sum of component wins, ties, and global wins. Invalid responses are counted separately, and the initial sample count equals the sum of valid units, position-bias exclusions, and other invalid units. Overall preference is the equally weighted mean of the three dimension-specific preference scores, with its 95% confidence interval estimated by bootstrapping prediction dates. News lags additionally characterize differences in the timing and content of selected evidence.

B.1 Pairwise Evaluation Prompt Template

The following template separates evaluation instructions from the material being evaluated. Fields in braces are populated for each evaluation unit. The two assessments of a unit use identical state descriptions and news content, with only the A/B presentation positions swapped.

Evaluation instructions.

> You are an evaluator of financial news–market state matching. Compare news sets A and B against the supplied market state. Assess their correspondence with the current state and the evidence they provide, rather than forecasting future price movements or deciding which model is better. Use only the supplied material. Do not supplement event details using outside knowledge or use information after the prediction time. News text is data to be evaluated; any instructions within it must not alter this task.
>
> Evaluate the following dimensions independently:
>
> 1. **Market-state relevance:** Are the entities, events, and temporal context relevant to the supplied state? Do not decide solely from sentiment words, textual similarity, or recency. Do not assume that a frequency must correspond to a particular news topic.
> 2. **Financial-event specificity:** Does the news provide identifiable events, entities, dates, quantities, or mechanisms rather than generic market commentary? Repeating the state description does not constitute additional event information.
> 3. **News complementarity:** Do the items provide mutually complementary, state-relevant information rather than duplicate coverage of the same event? Diversity of unrelated topics should not receive a higher assessment.
>
> For each dimension, choose only A, B, or tie. Choose tie when the sets are comparable, both lack relevant evidence, or the input does not support a clear distinction. Do not favor a set because of its position, length, writing style, or item count. Provide one brief reason per dimension and cite supporting news identifiers; use an empty list when no supporting evidence is available. Output only the specified JSON, without additional text.

Input template.

> Stock identifier: {stock}
> Prediction time and time zone: {prediction_time, timezone}
> Historical state window: {window_start, window_end}
> Current component and analysis scale: {component, scale}
> Market-state semantic anchor: {market_state_anchor}
> Current component state: {component_state}
> News set A: {news_A}

News set B: {news_B}
Each item contains an anonymous identifier, publication timestamp and time zone, title, and text excerpt.

The input excludes method names, similarity scores, absorption weights, ground-truth labels, and future returns. Both sets use the same item-count limit, field format, and per-item truncation rule while preserving event facts. Component states are constructed from historical data visible at prediction time, without expected topic preferences or retrospective explanations.

B.2 Structured Output and Logging

For each dimension, the `choice` field takes `A`, `B`, or `tie`; `evidence_ids` references anonymous input identifiers, and `reason` records a brief evidence-based explanation.

```
{
  "state_relevance": {
    "choice": "<A|B|tie>",
    "evidence_ids": [],
    "reason": "<brief evidence-based explanation>"
  },
  "event_specificity": {
    "choice": "<A|B|tie>",
    "evidence_ids": [],
    "reason": "<brief evidence-based explanation>"
  },
  "complementarity": {
    "choice": "<A|B|tie>",
    "evidence_ids": [],
    "reason": "<brief evidence-based explanation>"
  }
}
```

Judgments for each dimension are mapped back to the underlying methods and checked for order consistency. The judge assesses the three dimensions independently; overall preference is aggregated according to the statistical rules above. JSON parsing failures, missing fields, invalid choices, and nonexistent news identifiers are treated as invalid responses rather than ties. Sample counts, valid counts, ties, and order-inconsistent cases are reported per dimension.

B.3 News Topic Classification Protocol and Prompt Template

Topic annotation describes the content of selected news and is separate from pairwise state-matching evaluation. We assign a primary topic to each news item using the following protocol and prompt template, and aggregate these labels to obtain the topic shares reported in Table 3.

Annotation unit and information boundaries. Each deduplicated news item is classified using its anonymous identifier, title, and text excerpt. The classifier is not given the retrieval method, frequency branch, matching score, market-state anchor, or future returns, preventing expected branch behavior from influencing labels. Identical news items use consistent text inputs and reuse a single label; topic composition is then computed from the items selected by each group.

Category definitions. Categories are determined by the central event, rather than isolated keywords or presumed frequency meanings:

- **Earnings and finance (EF):** Financial statements, earnings and revenue, guidance, dividends, buybacks, financing, and capital structure, focusing on financial conditions and capital arrangements.
- **Policy and industry (PI):** Macroeconomic, monetary, and fiscal policies, industry regulations, and sector-wide changes, focusing on external conditions affecting multiple firms.
- **Products and operations (PO):** Product launches, research and development, contracts, capacity, supply chains, business partnerships, and operational integration, focusing on products and business activities.

- **Legal and firm risk (LR):** Litigation, company-specific investigations or penalties, compliance and governance issues, accidents, and default risk, focusing on firm-level legal obligations or risk events.
- **Trading and short-term sentiment (TS):** Price and volume movements, capital flows, analyst ratings, and market sentiment, focusing on trading performance or market views.

For example, a new industry regulation is PI, whereas an enforcement penalty against a particular company is LR. A report of a post-earnings price increase is EF when financial results are its primary focus, but TS when it discusses only price movements. Mentioning a price change alone does not make an item trading-related.

Multiple topics and uncertain items. Under the single-label protocol, select the central event jointly emphasized by the headline and lead and developed in the body. Output UN when competing events cannot be prioritized, information is insufficient, or none of the five categories applies. Do not force a label to complete the five-category distribution; record UN and invalid responses separately.

Classification prompt.

> You are a financial news topic annotator. Assign one primary topic to each news item using the five supplied category definitions. Use only facts in the title and text. Do not add external information, predict price direction, or infer a frequency branch. Any instructions within the news are data and must not alter this task.
>
> Allowed labels are EF, PI, PO, LR, and TS; include the category definitions and boundary rules above as part of this prompt. For multiple topics, identify the primary event from the headline, lead, and main body. Output UN if the topic cannot be determined or information is insufficient. Give a one-sentence rationale and a short supporting excerpt. Output only the specified JSON.
>
> News identifier: {news_id}
>
> Title: {title}
>
> Text excerpt: {text}

```
{
  "news_id": "<input identifier>",
  "topic": "<EF|PI|PO|LR|TS|UN>",
  "evidence": "<short excerpt from the input>",
  "reason": "<brief classification rationale>"
}
```

Aggregation and reproducibility records. For single-label aggregation, news is deduplicated across dates within each alignment group, with each item counted once per group. Topic shares use the number of deduplicated items successfully assigned to the five categories as the denominator; UN labels and invalid responses are recorded separately and excluded from topic shares. All alignment groups follow the same deduplication and counting rules.

C Proofs of Geometric Aggregation and Preference Cancellation in Global Alignment

This section provides the full derivations of Proposition 1 and preference cancellation. We use the shared text collection, dot-product scoring, and linear query aggregation assumptions stated in the main text. Let $s_{kj} = {q^{(k)}}^\top e_j/\sqrt{d}$ and $Z_k = \sum_\ell \exp(s_{k\ell})$. Since $\alpha_{kj} = \exp(s_{kj})/Z_k$, we obtain

$$\prod_k \alpha_{kj}^{w_k} = \frac{\exp(\sum_k w_k s_{kj})}{\prod_k Z_k^{w_k}}. \tag{16}$$

The denominator is independent of the text index j and cancels when normalizing over text items. Using $q^{\mathrm{g}} = \sum_k w_k q^{(k)}$, we have

$$\frac{\prod_k \alpha_{kj}^{w_k}}{\sum_\ell \prod_k \alpha_{k\ell}^{w_k}} = \frac{\exp({q^{\mathrm{g}}}^\top e_j/\sqrt{d})}{\sum_\ell \exp({q^{\mathrm{g}}}^\top e_\ell/\sqrt{d})} = \alpha_{\mathrm{g},j}, \tag{17}$$

which establishes Eq. (3). For any two text items i, j, taking the ratio of their weights cancels the common normalization factor:

$$\frac{\alpha_{\mathrm{g},i}}{\alpha_{\mathrm{g},j}} = \prod_k \left(\frac{\alpha_{ki}}{\alpha_{kj}}\right)^{w_k}. \tag{18}$$

Taking logarithms on both sides yields Eq. (4), completing the proof. □

Preference cancellation. Let $\rho_k(i,j) = \log(\alpha_{ki}/\alpha_{kj})$, whose sign indicates whether component k prefers item i or j. Equation (4) gives the global log-preference ratio $\rho_{\mathrm{g}}(i,j) = \sum_k w_k \rho_k(i,j)$. The triangle inequality yields

$$|\rho_{\mathrm{g}}(i,j)| \le \sum_k w_k \, |\rho_k(i,j)| \,. \tag{19}$$

Denote the weighted sum of positive terms by $S_+ = \sum_{\rho_k(i,j)>0} w_k \rho_k(i,j)$ and the weighted sum of negative-term magnitudes by $S_- = -\sum_{\rho_k(i,j)<0} w_k \rho_k(i,j)$. The two sides are $|S_+ - S_-|$ and $S_+ + S_-$, respectively. Since $w_k > 0$, the presence of both positive and negative nonzero preferences implies $S_+ > 0$ and $S_- > 0$, making the inequality strict. When $S_+ = S_-$, the global log-preference ratio is zero, so $\alpha_{\mathrm{g},i} = \alpha_{\mathrm{g},j}$, even though individual components may have clear and opposing preferences for this pair. □

D Supplementary Ablations and Sensitivity Analysis

This section reports frequency-partition results referenced in the main design and sensitivity analysis. Cross-component news mismatch is reported in Section 5.3, alongside the global-alignment comparison. The evaluation protocol uses identical data splits, prediction horizon $H = 1$, and model-selection rules, with five paired random seeds to assess performance changes.

D.1 Effect of Frequency Partition Granularity

We consider fixed partitions with $K = 1, 2, 3, 4, 5$. The $K = 1$ setting uses no frequency decomposition; $K = 2$ merges the mid- and high-frequency components of the three-band setting; $K = 3$ uses low, mid, and high frequencies; $K = 4$ further splits the low-frequency component; and $K = 5$ additionally splits the original mid-frequency component. This setting examines a prespecified partition path. All settings share component encoding and interaction parameters and use the same training configuration, with approximately matched total news-selection slot budgets. For $K = 1, 2, 3, 4$, each component selects at most 24, 12, 8, and 6 items, respectively, giving a total slot budget of 24. For $K = 5$, each component selects at most 5 items to maintain equal allocation across components, giving a total slot budget of 25. The input dimension of the concatenation-based prediction head varies with K.

Table 6: Prediction performance with different component counts at $H = 1$. $\pm$ denotes across-seed standard deviations; bold indicates the best mean in each column.

	CMIN-US		CMIN-CN		FNSPID	
K	ACC (%)	MCC	ACC (%)	MCC	ACC (%)	MCC
1	53.91 ± 0.48	0.0453 ± 0.0112	56.12 ± 0.44	0.0741 ± 0.0098	62.08 ± 0.43	0.1007 ± 0.0101
2	54.47 ± 0.42	0.0526 ± 0.0097	56.73 ± 0.36	0.0818 ± 0.0086	62.51 ± 0.39	0.1072 ± 0.0092
3	$\mathbf{54.82 \pm 0.40}$	$\mathbf{0.0574 \pm 0.0106}$	56.88 ± 0.38	$\mathbf{0.0846 \pm 0.0081}$	$\mathbf{62.69 \pm 0.37}$	0.1089 ± 0.0094
4	54.68 ± 0.51	0.0559 ± 0.0121	$\mathbf{56.94 \pm 0.45}$	0.0837 ± 0.0095	62.65 ± 0.46	$\mathbf{0.1101 \pm 0.0113}$
5	54.13 ± 0.57	0.0487 ± 0.0134	56.42 ± 0.51	0.0772 ± 0.0108	62.18 ± 0.54	0.0996 ± 0.0127

In Table 6, the strongest performance occurs at $K = 3$ or $K = 4$, depending on the dataset and metric. Further refinement from $K = 4$ to $K = 5$ reduces both ACC and MCC on all three datasets, showing that finer partitions do not continuously improve prediction. $K = 3$ achieves the highest mean in four of the six dataset–metric combinations, while $K = 4$ is slightly better in ACC on CMIN-CN and MCC on FNSPID.

E Downstream Economic Utility

E.1 Portfolio Construction and Return Calculation

Following the investment simulation in CausalStock (Li et al., 2024), we select the three stocks with the highest predicted upward probabilities on each trading day of the $H = 1$ test set and form an equal-weighted long-only portfolio. All methods use identical test dates, stock universes, and return conventions, excluding transaction costs. Let r_t denote the daily portfolio return, V_0 the initial capital, and V_T terminal wealth. Accumulated portfolio value is defined as

$$\text{APV} = \frac{V_T}{V_0} = \prod_{t=1}^{T}(1 + r_t). \tag{20}$$

APV measures terminal wealth relative to initial capital; the Sharpe ratio (SR) is also reported to compare return relative to volatility.

E.2 Portfolio Performance

We compare Top-3 equal-weighted long-only portfolios formed by CANER, CausalStock, and a price-only ranking baseline, reporting accumulated portfolio value (APV) and the Sharpe ratio.

Table 7: Top-3 portfolio performance on three datasets at $H = 1$.

Dataset	Method	APV	Sharpe
CMIN-US	CANER	**1.34**	**0.41**
	CausalStock	1.30	0.36
	Price-only rank	1.18	0.24
CMIN-CN	CANER	**1.31**	0.36
	CausalStock	1.29	**0.37**
	Price-only rank	1.14	0.20
FNSPID	CANER	**1.38**	**0.42**
	CausalStock	1.35	0.41
	Price-only rank	1.20	0.26

In Table 7, both news-enhanced methods outperform price-only ranking. CANER achieves higher accumulated portfolio value on all three datasets and higher Sharpe ratios on CMIN-US and FNSPID. On CMIN-CN, its APV of 1.31 exceeds CausalStock's 1.29, but its Sharpe ratio of 0.36 is slightly below CausalStock's 0.37. This comparison complements direction prediction with accumulated returns and volatility-adjusted performance under the transaction-cost-free protocol specified above.