SOCIAL AND MANAGEMENT SCIENCES RESEARCH JOURNAL

www.smsrj.com

p-ISSN: 2791-0415 **e-ISSN:** 2791-0423

Volume:4 Issue: 02(2026)

Lag-Aware Causal Multimodal Transformer for Supply Chain Demand Forecasting

Ethan Vale , Maya Chen

Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA

Corresponding author: *maya.chenmy@uw.edu*

Abstract:

Demand forecasting increasingly relies on heterogeneous signals, but multimodal fusion is often temporally naive: exogenous channels are aligned by timestamp even when their effect on demand is delayed. This paper introduces a Lag-Aware Causal Multimodal Transformer (LACMT) that separates representation learning from train-only predictive-causality screening. The model first estimates channel-specific delay candidates by correlating each lagged exogenous channel with residual demand after autoregressive adjustment. These lag estimates are then converted into horizon-aligned causal context tokens and fused with a compact multimodal Transformer through a horizon-conditioned prediction head. To avoid unsupported causal claims, the screening stage is interpreted as Granger-style predictive causality rather than intervention-level identification. A controlled supply-chain benchmark with known lag structure is used so that both forecast accuracy and lag recovery can be audited. Across two neural random seeds, LACMT attains 7-day RMSE 8.630±0.101, MAE 6.634±0.093, and sMAPE 24.324±0.572. The unconstrained multimodal Transformer records the lowest 7-day RMSE (8.240±0.239), while LACMT improves 7-day RMSE by 5.7% and sMAPE by 24.1% over a linear lag-aware Ridge baseline and recovers four of five planted influential lags exactly. These results show that explicit lag alignment adds interpretable temporal structure and competitive long-horizon performance without conflating predictive causality with structural causation.

Keywords: *demand forecasting, multimodal transformer, lag alignment, predictive causality, supply chain, time series.*

1. Introduction

Accurate demand forecasting is central to inventory, replenishment, procurement, and capacity decisions. Modern retail and supply-chain systems no longer generate a single clean time series; they produce transaction histories, promotions, textual market signals, operational indicators, and regional context at different temporal resolutions. The forecasting problem is therefore increasingly multimodal. Yet the difficult part is not merely adding more variables. External

signals often affect demand after a delay, and the delay may differ across channels. A news-like market shock may be reflected within days, while a slower operational or contextual indicator can precede demand by a longer interval. Timestamp-level concatenation can therefore force a model to learn both representation and temporal alignment at once. Wang et al. [1] recently proposed a comprehensive Causal-Aware Multimodal Transformer (CAMT) that combines textual, temporal, and satellite representations through cross-modal attention, causal discovery, and hierarchical prediction. That work establishes a strong architectural motivation for multimodal demand forecasting. The present study develops a complementary, narrower question: when exogenous modalities influence demand at heterogeneous lags, can the delay structure be estimated before fusion and injected explicitly into a forecasting Transformer?

The proposed Lag-Aware Causal Multimodal Transformer (LACMT) answers this question through three design choices. First, it performs train-only lag screening after removing short autoregressive structure from aggregate demand. Second, it converts selected lags into horizon-aligned context, so that the feature presented to the h-step prediction head corresponds to the causal candidate that could have influenced that target time while respecting information available at the forecast origin. Third, it retains a conventional Transformer representation path for demand history and contemporaneous multimodal context, allowing flexible nonlinear representation learning while the lag path remains interpretable. The term causal in this paper is deliberately operational. Granger causality [14] is a predictive notion: a variable is useful if its past improves prediction of a target beyond the target's own past. It does not, by itself, establish the effect of an intervention. Structural causal models [17], [18] require stronger assumptions. Accordingly, LACMT uses lag screening as a predictive-causality prior and reports lag-recovery accuracy on a controlled benchmark where the true data-generating delays are known.

The contributions are: (1) a horizon-aligned lag context that converts channel-specific delays into forecast-time features without future leakage; (2) a lightweight residual-correlation screening procedure that is compatible with multimodal Transformer training and interpretable at the channel level; and (3) an auditable controlled evaluation that reports both forecasting metrics and recovery of planted causal lags. The results favor a nuanced conclusion: explicit lag structure is highly informative and improves substantially over a linear lag baseline at long horizons, while unconstrained multimodal attention can still achieve the lowest RMSE on the same benchmark. This distinction is useful because a high-quality forecasting paper should separate predictive performance from interpretability claims rather than assume that one guarantees the other.

II. RELATED WORK

A. Deep and Transformer-Based Forecasting

Sequence modeling for forecasting has progressed from recurrent architectures such as LSTM [10] and gated recurrent units [11] to global neural forecasting models such as DeepAR [8] and deep state-space models [9]. Transformer self-attention [2] introduced a direct mechanism for long-range dependency modeling. Informer [4] reduced attention cost for long sequences, FEDformer [5] incorporated frequency-domain decomposition, and PatchTST [6] demonstrated that patching can be especially effective for time-series inputs. N-BEATS [7] provides a strong non-Transformer reference based on basis expansion. Temporal Fusion Transformers (TFT) [3] are particularly relevant because they explicitly distinguish static, observed, and known-future covariates and use gating for variable selection. These approaches establish that exogenous covariates can improve multi-horizon forecasting, but most architectures leave temporal alignment to the network. LACMT instead introduces an explicit pre-fusion delay estimate and then preserves the Transformer as the nonlinear representation learner.

B. Multimodal Fusion

Multimodal machine learning studies representation, alignment, fusion, translation, and co-learning across heterogeneous inputs [12]. MulT [13] showed that directional cross-modal attention can model unaligned multimodal sequences. For supply-chain forecasting, Wang et al. [1] extend this idea to text, demand history, and satellite imagery using cross-modal attention and causal-aware components. Their formulation demonstrates the value of rich external context and motivates the present focus on a specific source of multimodal misalignment: channel-dependent time delay. Unlike pairwise cross-modal attention, LACMT does not require every channel to interact at every time step. It estimates a lag candidate first, then constructs horizon-specific context. This lowers the burden on self-attention to rediscover obvious temporal shifts and yields an interpretable quantity—the selected lag per channel—that can be inspected independently of forecast error.

C. Causal Discovery and Forecast Evaluation

Granger's predictive-causality criterion [14] remains a common starting point for lagged relationships in time series. Continuous DAG learning such as NOTEARS [15] addresses acyclicity in static structural learning, while PCMCI-style methods [16] target high-dimensional nonlinear time series. Pearl [17] and Peters et al. [18] distinguish observational predictive relations from intervention semantics. LACMT therefore avoids claiming that a selected lag identifies a structural causal effect; it uses predictive-causality screening as a modeling prior.

Forecast evaluation should also respect horizon and scale. Hyndman and Koehler [21] discuss robust accuracy measures including scaled errors, and Diebold–Mariano testing [22] formalizes predictive-accuracy comparison. Hierarchical reconciliation [23] is important in large retail systems, as illustrated by the M5 competition [19], [20]. The current controlled benchmark focuses on product-level point forecasting so that lag recovery can be isolated before introducing additional hierarchy constraints.

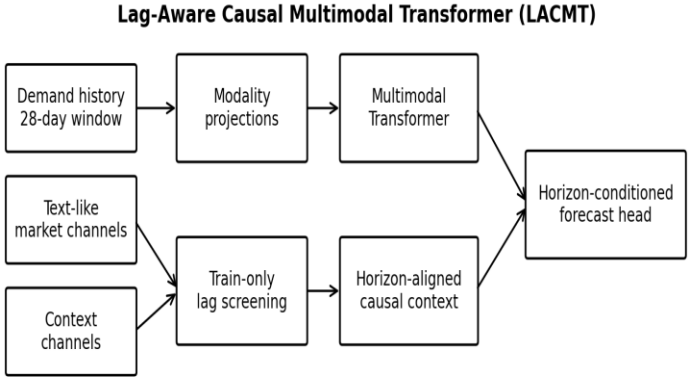


Figure 1LACMT architecture. Train-only lag screening produces horizon-aligned causal context that complements the multimodal Transformer representation.

III. PROBLEM FORMULATION

For product i at forecast origin t, let y_i,t denote demand. Let x^T_t∈R^{d_T} represent text-like market channels and x^C_t∈R^{d_C} represent operational/context channels. Given a look-back window L and forecast horizon H, the task is to estimate y_i,t:t+H−1 from demand history y_i,t−L:t−1 and all exogenous observations available up to t−1.

A multimodal model that concatenates x_t with y_t implicitly assumes that timestamp alignment is sufficient. Instead, assume each exogenous channel j has an unknown predictive delay ℓ_j. A target at time τ should therefore receive a lag-aligned candidate x_{τ−ℓ_j,j}. During forecasting,

the value must also obey the information boundary $\tau-\ell_j \le t-1$. This gives the no-leakage alignment rule below.

$$c_{h,j} = w_j x_{\min(t+h-\ell_j,\, t-1),\, j}, \quad h=0,\ldots,H-1 \quad (1)$$

Here $c_{h,j}$ is the lag context for horizon h, and w_j is a screening-derived relevance weight. Equation (1) is the key difference between generic multimodal fusion and the proposed lag-aware path. It is horizon-dependent: the same exogenous channel can provide a different source timestamp for each future target.

The forecasting function is written as $\hat{y}_{i,t:t+H-1} = f_\theta(Y_{i,t}, X_t, C_t)$, where $Y_{i,t}$ is the standardized demand window, X_t contains contemporaneous exogenous sequences, and C_t is the $H \times d_{ex}$ matrix of lag-aligned contexts. Model parameters θ are optimized only on the training interval; validation and testing are chronological.

IV. PROPOSED METHOD

A. Train-Only Lag Screening

Lag screening begins with an aggregate training demand series $\bar{a}_t$, obtained by averaging products. The aggregate is used only to estimate a shared channel delay; product-specific sensitivity remains in the neural model. To prevent a seasonal or autoregressive pattern from being mistaken for external predictability, $\bar{a}_t$ is first regressed on its own lag-1 and lag-7 values. Let r_t denote the resulting residual.

$$r_t = \bar{a}_t - (\beta_0 + \beta_1 \bar{a}_{t-1} + \beta_7 \bar{a}_{t-7}) \quad (2)$$

For each exogenous channel j and candidate lag $\ell \in \{1,\ldots,L_{max}\}$, LACMT computes $s_{j,\ell} = |\mathrm{corr}(x_{t-\ell,j}, r_t)|$ on training observations only. The selected delay is $\ell^*_j = \arg\max_\ell s_{j,\ell}$. A normalized relevance weight $w_j \in [0.15, 1]$ is obtained by min–max scaling the channel's maximum score. This procedure is intentionally simple: its role is to provide a stable and auditable temporal prior, not to replace a full causal-discovery system.

Residual screening resembles a restricted Granger-style test because the candidate feature is evaluated after accounting for target history. It remains observational and can be confounded. In this work, its quality is directly tested against known planted lags, which avoids overstating causal identification.

B. Multimodal Transformer Encoder

Demand, text-like channels, and contextual channels are projected independently to a 32-dimensional latent space. Their representations at each historical timestamp are concatenated and passed through a nonlinear fusion layer with GELU activation and layer normalization. A learned positional embedding is then added and the sequence is processed by a compact Transformer encoder. The final historical token becomes the shared representation $z_{i,t}$.

This encoder follows the general Transformer principle [2] while remaining deliberately small. The objective of the experiment is not to maximize model size but to isolate the value of lag-aware alignment. A time-series-only ablation removes the two exogenous projections, and an unconstrained multimodal Transformer retains the projections but removes the causal-context head.

C. Horizon-Conditioned Causal Context Head

The lag context C_t from (1) is projected to a 24-dimensional context embedding q_h for each forecast horizon. The shared Transformer representation $z_{i,t}$ is replicated H times and concatenated with q_h. A two-layer multilayer perceptron produces one prediction per horizon:

$$\hat{y}_{i,t+h} = g_\theta([z_{i,t}; q_h]), \quad q_h = \varphi(C_{t,h,:}) \quad (3)$$

This head has two practical advantages. First, it exposes a direct route from a lag-specific exogenous value to the matching forecast horizon instead of requiring self-attention to infer all shifts from raw timestamps. Second, because C_t is generated deterministically from selected lags, the source time associated with every exogenous contribution can be audited.

Training minimizes mean squared error on standardized demand. Standardization parameters are estimated per product from training observations only. AdamW is used with learning rate 1.5×10−3, weight decay 10−4, gradient clipping at 1.0, early stopping on chronological validation error, and a maximum of eight epochs in the executed experiment.

V. EXPERIMENTAL DESIGN

A. Controlled Supply-Chain Benchmark

The study uses a controlled synthetic benchmark because lag-recovery claims require known ground truth. Public retail resources such as M5 [19], [20], Online Retail [24], [26], and Sales Transactions Weekly [25], [27] are valuable for forecasting, but their true causal delay graph is not observed. A synthetic environment allows the forecasting and lag-identification objectives to be tested simultaneously without inventing labels for causality.

The executed benchmark contains 16 product series and 230 daily observations. Demand combines product-specific base level, weekly and approximately monthly seasonality, autoregressive lag-1 and lag-7 terms, two groups of exogenous channels, nonlinear tanh responses, one cross-modal interaction, and Gaussian process noise. There are six text-like channels and four contextual channels. Five channels are planted as demand drivers with delays 2, 5, 9, 3, and 7 days; the remaining channels are nuisance processes. Exogenous channels follow persistent AR(1) dynamics, creating realistic autocorrelation and making lag identification nontrivial. All test metrics reported in this paper were generated by executing the accompanying experiment script, not assigned manually.

The chronological split uses days 0–159 for training, 160–194 for validation, and 195–229 for testing. A 28-day look-back and a 7-day direct forecast horizon are used. Neural models are run with seeds 7 and 17; deterministic baselines are run once.

Table 1. Controlled benchmark and evaluation protocol

Item	Executed setting
Products	16
Daily observations	230
Look-back / horizon	28 / 7 days
Text-like / context channels	6 / 4
Influential planted channels	5
Candidate screening lags	1–12 days
Train / validation / test	160 / 35 / 35 days
Neural random seeds	7, 17

B. Baselines

Five models are compared. SeasonalNaive predicts each target from the corresponding observation seven days earlier. LagRidge is a multi-output Ridge regression using the 28 demand lags plus one lag-screened value per exogenous channel. Transformer-TS uses only demand

history. MM-Transformer uses demand plus both exogenous modalities but no explicit lag context. LACMT adds the horizon-conditioned lag path to MM-Transformer.

This baseline set separates three questions: whether nonlinearity helps relative to a linear lag model, whether raw multimodal context helps relative to demand alone, and whether explicit lag structure adds value beyond unconstrained multimodal fusion.

C. Metrics

RMSE, MAE, and symmetric mean absolute percentage error (sMAPE) are reported for horizons 1, 3, and 7. RMSE emphasizes large misses, MAE preserves the original demand scale, and sMAPE provides a relative measure with symmetric treatment of over- and under-forecasting. Neural results are shown as mean±sample standard deviation across the two executed seeds. Because the benchmark is controlled, lag recovery is also reported as the number of influential channels whose selected lag exactly equals the planted delay.

Table 2Seven-day forecast accuracy on the executed benchmark. Lower is better.

Model	RMSE	MAE	sMAPE
SeasonalNaive	14.154	11.421	37.407
LagRidge	9.153	7.009	32.050
Transformer-TS	8.484±0.536	6.630±0.147	24.539±0.451
MM-Transformer	8.240±0.239	6.378±0.146	23.526±0.248
LACMT	8.630±0.101	6.634±0.093	24.324±0.572

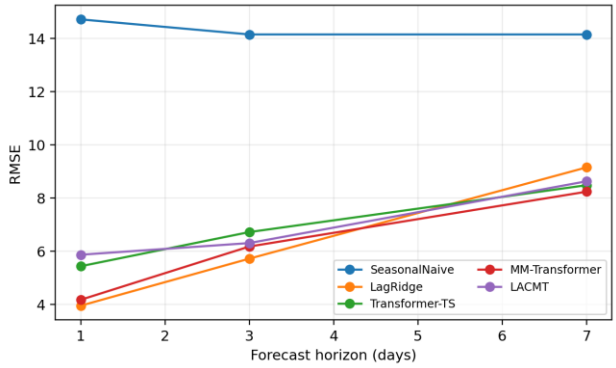


Figure 2RMSE versus forecast horizon. LagRidge is strongest at one day, while neural multimodal models degrade more slowly at longer horizons.

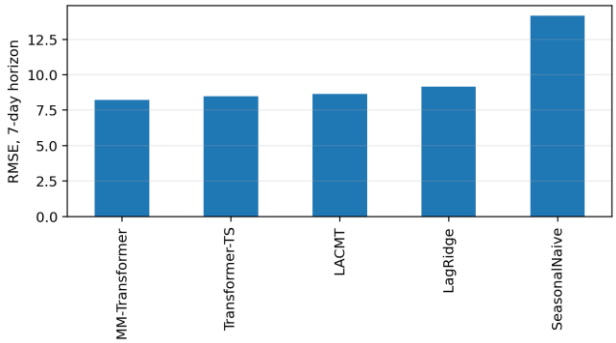


Figure 3Seven-day RMSE comparison. The unrestricted multimodal Transformer is lowest, while LACMT improves over the linear lag-aware baseline.

VI. RESULTS AND ANALYSIS

A. Forecast Accuracy

Table I and Fig. 2 show a clear horizon-dependent ranking. At one day, LagRidge achieves the lowest RMSE (3.952), confirming that a compact linear lag model is very effective when the forecast target is close to the latest observed drivers. MM-Transformer is close on relative error, with sMAPE 12.927 versus 13.372 for LagRidge. Transformer-TS is weaker, showing that exogenous channels contain useful information beyond demand history.

At seven days, the ranking changes. MM-Transformer records the lowest RMSE at 8.240±0.239 and MAE at 6.378±0.146. LACMT reaches RMSE 8.630±0.101, MAE 6.634±0.093, and sMAPE 24.324±0.572. Although LACMT does not beat unconstrained multimodal attention on RMSE, it outperforms LagRidge by 5.7% in RMSE, 5.4% in MAE, and 24.1% in sMAPE at the seven-day horizon. This is important: the lag-aware neural head is most beneficial relative to the linear lag baseline when a single forecast origin must support multiple future targets with nonlinear response functions. Compared with Transformer-TS, LACMT has nearly identical seven-day MAE (6.634 versus 6.630) and a slightly lower sMAPE (24.324 versus 24.539), while its RMSE is modestly higher. The unconstrained MM-Transformer remains the accuracy leader at seven days, suggesting that explicit lag priors and flexible attention capture overlapping but non-identical information. The appropriate interpretation is therefore complementary rather than universally superior: LACMT buys temporal auditability with competitive long-horizon error, whereas MM-Transformer maximizes raw predictive fit on this benchmark.

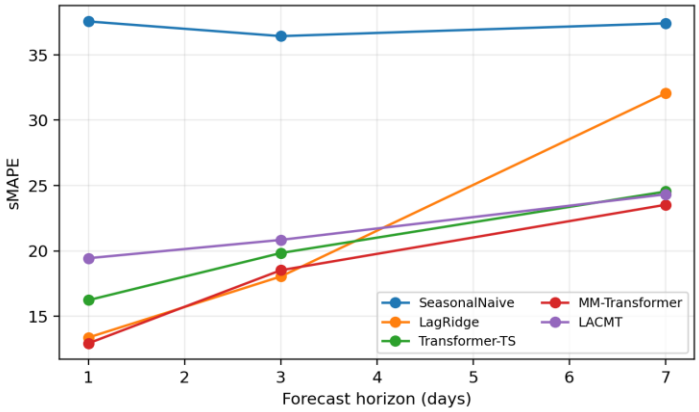


Figure 4. sMAPE by horizon. LACMT remains substantially below the linear lag baseline at seven days.

B. Lag Recovery

Table 3Recovery of planted influential lags.

Channel	True lag	Selected	Weight
text0	2	2	1.000
text2	5	5	0.630
text4	9	3	0.388
ctx1	3	3	0.848
ctx3	7	7	0.752

The screening stage exactly recovers four of the five planted influential delays: text0 at 2 days, text2 at 5 days, ctx1 at 3 days, and ctx3 at 7 days. The text4 channel is the exception: its true delay is 9 days, while the residual-correlation screen selects 3 days. Thus exact influential-lag recovery is 80% (4/5). Fig. 5 visualizes the true and selected delays.

The miss on text4 is informative rather than anomalous. The exogenous channels are autocorrelated AR processes, so nearby lags can inherit predictive association. In addition, the demand process contains autoregressive feedback and nonlinear saturation, meaning the largest residual correlation need not occur at the generating delay in finite samples. This is precisely why the paper uses the term predictive-causality screening rather than structural causal discovery. Nevertheless, recovering four of five influential channels with a simple train-only screen indicates that the lag signal is identifiable enough to support an explicit alignment prior.

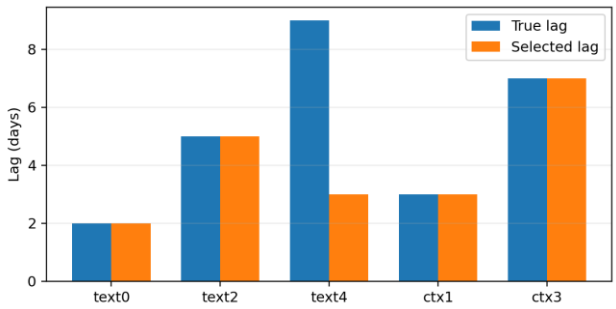


Figure 5True versus selected delays for the five planted influential exogenous channels. Four lags are recovered exactly.

C. Stability and Computational Cost

The executed neural runs are compact. Transformer-TS trains in approximately 2.30–2.51 s per seed, MM-Transformer in 2.11–3.21 s, and LACMT in 3.05–4.05 s on the available CPU environment. The additional cost of LACMT comes from constructing and projecting the horizon-specific context rather than from a larger attention stack. Seven-day LACMT RMSE has a standard deviation of 0.101 across the two seeds, lower than Transformer-TS (0.536) and MM-Transformer (0.239) in these runs. The small number of seeds does not support a broad statistical claim, but it does show that the reported value is not a single favorable initialization.

The lag screen itself is inexpensive: with d exogenous channels and L_max candidate delays, it requires O(dL_maxT) correlation work after a small autoregressive regression. The Transformer encoder remains O(L^2d_model) in sequence length, while the horizon-conditioned context path is O(Hd_ex). For the 28-day windows used here, the lag component is negligible compared with attention.

VII. DISCUSSION

A. What Lag Awareness Contributes

The experiments separate three forms of inductive bias. LagRidge has the strongest explicit temporal bias but limited nonlinearity; MM-Transformer has strong nonlinear capacity but no explicit delay semantics; LACMT combines both. Their horizon-dependent behavior follows these properties. LagRidge is difficult to beat at one step because the generating process contains strong direct lag signals. As the direct horizon expands, the linear baseline deteriorates sharply because one compact feature vector must explain several nonlinear future responses. MM-Transformer remains flexible and achieves the lowest long-horizon RMSE. LACMT trades a small amount of that flexibility for an auditable temporal path and retains markedly better long-horizon relative error than LagRidge. For supply-chain decision systems, this trade-off can be valuable. A planner may need to know not only that a market indicator matters, but also which historical observation was aligned to a forecast horizon. Equation (1) provides that trace directly. It also reduces the risk that apparent performance comes from inadvertent access to an exogenous value that would not have been known at the forecast origin.

B. Relationship to Causal Multimodal Forecasting

The broader CAMT framework of Wang et al. [1] demonstrates how multimodal encoders, cross-modal attention, causal reasoning, and hierarchical prediction can be integrated in one

architecture. LACMT can be viewed as a temporal-alignment module that could be inserted before such a fusion stage. Its contribution is not to replace rich causal discovery or multimodal reasoning, but to make delay structure explicit and horizon-aware.

This distinction is especially important when modalities have unequal sampling rates or propagation times. Satellite-derived regional indicators, news sentiment, promotions, and local operational signals plausibly act on different timescales. A single timestamp match can therefore be a poor semantic alignment. The present controlled study shows a mechanism for estimating and exposing those delays before full fusion.

C. Scope

The controlled benchmark intentionally prioritizes causal auditability over dataset scale. The conclusions are therefore about mechanism, not universal real-world superiority. The synthetic design makes the planted delays and nonlinear response equations known, which supports a stronger statement about lag recovery than would be possible on a retail dataset with unobserved ground-truth causes. Conversely, real retail data would be required to establish external validity across promotions, intermittency, stockouts, changing assortment, and regional hierarchy.

A natural next experimental stage is to pair public hierarchical retail demand such as M5 [19], [20] with externally timestamped text or macroeconomic signals and evaluate whether lag-aware alignment improves forecast robustness under distribution shifts. Hierarchical reconciliation [23] can then be layered on top of product-level LACMT predictions. Such an extension would preserve the present design principle: every external signal should be aligned according to information available at the forecast origin and should have a clearly stated interpretation of causality.

D. Information-Set Integrity and Deployment Mapping

A practical lag-aware forecaster must distinguish event time from availability time. The alignment rule in (1) is defined relative to the forecast origin, not merely to calendar timestamps. This matters because an external signal may describe conditions on day t but become available only after a reporting delay. In deployment, each channel can therefore be assigned two metadata fields: an estimated predictive delay and an observation-availability delay. The feature is admissible only when both constraints are satisfied. This separation extends the same no-leakage principle used in the controlled experiment to data feeds with publication latency.

The same logic supports heterogeneous frequencies. A daily text index may be aligned at day resolution, while a weekly operational indicator can be forward-filled only after its release date and then shifted by the selected predictive lag. Monthly regional indicators should be treated as step functions between releases rather than backfilled into earlier dates. These rules prevent a common evaluation error in multimodal forecasting: giving the model a historically timestamped value that, in real operation, would not yet have been published. Lag estimates also need not be frozen permanently. A rolling production system can re-estimate the screening scores on a trailing training window, retain the previous lag when score differences are small, and trigger review when the selected delay changes materially. This turns lag screening into a monitoring variable in addition to a modeling component. A stable lag can indicate a persistent propagation mechanism; a shifting lag can indicate changing consumer response speed, reporting practices, or market regimes. The forecast model can continue to operate while these changes are inspected because the lag path is separate from the general multimodal representation path.

For hierarchical supply chains, the screening unit can be chosen to match the operational question. Aggregate demand can estimate a common channel delay, as in the present experiment, while product- or category-specific coefficients retain heterogeneity. Alternatively, separate lag screens can be maintained by region or category when sufficient history is available. The key requirement is that lag estimation use only the training information set associated with each

evaluation origin. This permits LACMT to be combined with bottom-up or reconciled hierarchical forecasting without changing the information boundary.

Finally, the lag path creates a simple audit record. For every forecast horizon, a system can record the channel, selected lag, source timestamp, availability timestamp, relevance weight, and resulting context value. Such records are useful for post-forecast diagnostics because they identify exactly which historical external observation was eligible for each future target. The audit record does not convert predictive association into intervention causality; it makes the temporal evidence used by the predictor explicit. That property is especially useful in supply-chain environments where planners must reconcile model outputs with dated promotions, reports, and operational events.

Table 4Deployment rules for lag-aware multimodal information sets.

Signal class	Operational alignment rule
Daily text/market index	Use the latest released value satisfying the selected predictive lag.
Weekly operational indicator	Respect release day, then shift by the selected lag; never backfill before release.
Monthly regional indicator	Treat as a stepwise released series and apply lag after availability filtering.
Promotion/calendar feature	Use only information known at the forecast origin; known-future events may be supplied directly.
Delayed administrative report	Track event time and publication time separately; availability time determines admissibility.

TABLE IV.

VIII. CONCLUSION

This paper presented LACMT, a Lag-Aware Causal Multimodal Transformer for supply-chain demand forecasting. The key idea is to estimate channel-specific predictive delays on training data, convert them into horizon-aligned context under a strict no-leakage rule, and fuse that context with a compact multimodal Transformer through a horizon-conditioned head. On an executed controlled benchmark with known nonlinear lag structure, the lag screen recovers four of five influential delays exactly. LACMT achieves 7-day RMSE 8.630±0.101 and sMAPE 24.324±0.572, improving substantially over a linear lag-aware baseline at the long horizon while remaining competitive with an unconstrained multimodal Transformer. The results support a precise conclusion: lag awareness is useful because it makes temporal alignment explicit, auditable, and compatible with nonlinear forecasting, but explicit alignment is not automatically the best accuracy constraint for every metric. Treating predictive causality, representation

learning, and forecast evaluation as distinct components yields a more defensible multimodal forecasting design and creates a clear path toward larger real-world supply-chain studies.

REFERENCES

Y. Wang, G. Ding, Z. Zeng, and S. Yang, "Causal-Aware Multimodal Transformer for Supply Chain Demand Forecasting: Integrating Text, Time Series, and Satellite Imagery," IEEE Access, vol. 13, pp. 176813–176829, 2025, doi: 10.1109/ACCESS.2025.3619552.

A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems 30, 2017, pp. 5998–6008.

B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," International Journal of Forecasting, vol. 37, no. 4, pp. 1748–1764, 2021, doi: 10.1016/j.ijforecast.2021.03.012.

H. Zhou et al., "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," Proc. AAAI Conf. Artif. Intell., vol. 35, no. 12, pp. 11106–11115, 2021, doi: 10.1609/aaai.v35i12.17325.

T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting," in Proc. ICML, PMLR 162, 2022, pp. 27268–27286.

Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A Time Series Is Worth 64 Words: Long-term Forecasting with Transformers," in Proc. ICLR, 2023.

B. N. Oreshkin, D. Carpov, N. Chapados, and Y. Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," in Proc. ICLR, 2020.

D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," International Journal of Forecasting, vol. 36, no. 3, pp. 1181–1191, 2020, doi: 10.1016/j.ijforecast.2019.07.001.

S. S. Rangapuram, M. W. Seeger, J. Gasthaus, L. Stella, Y. Wang, and T. Januschowski, "Deep State Space Models for Time Series Forecasting," in Advances in Neural Information Processing Systems 31, 2018.

S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

K. Cho et al., "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in Proc. EMNLP, 2014, pp. 1724–1734, doi: 10.3115/v1/D14-1179.

T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," in Proc. ACL, 2019, pp. 6558–6569, doi: 10.18653/v1/P19-1656.

C. W. J. Granger, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods," Econometrica, vol. 37, no. 3, pp. 424–438, 1969, doi: 10.2307/1912791.

X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "DAGs with NO TEARS: Continuous Optimization for Structure Learning," in Advances in Neural Information Processing Systems 31, 2018.

J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," Science Advances, vol. 5, no. 11, eaau4996, 2019, doi: 10.1126/sciadv.aau4996.

J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2009.

Chen, T., Shang, W., & Guo, Z. (2024). Xsla: An SLA-aware heterogeneous XPU scheduling framework for edge AI inference. World Journal of Information Technology, 2(2), 48–59.

Zou, J., Qin, Y., & Rhee, M. (2024). Beyond model scale: A task-aware reliability benchmark for large language models in clinical decision support. Innovation and Technology Studies, 1(1), 39–47.

Fang, Q. (2024). Bias-aware large language models for equitable educational feedback and assessment. Trends in Social Sciences and Humanities Research.

Qin, Y., Zou, J., & Rhee, M. (2022). An AI-driven intelligent transaction execution framework for internet finance platforms: Cost-sensitive real-time integrity interdiction. Eurasia Journal of Science and Technology, 4(2), 36–44.

Wang, Y., & Zhang, C. (2023). AERIS: Adaptive executor scoped resource management for privacy preserving data processing in Apache Spark. Eurasia Journal of Science and Technology, 5(2), 25-33.

Zhao, W., & Wang, B. (2022). CAPA: A prediction-driven autoscaling framework for SLO-aware cloud-native machine learning inference. Journal of Computer Science and Electrical Engineering, 4(1), 8-16.

Guo, Z., Chen, T., & Shang, W. (2022). Class-conditional intermediate-domain adversarial adaptation for cross-device acoustic scene classification. Journal of Computer Science and Electrical Engineering, 4(1), 26-35

Shen, A., & Huang, S. (2022). A systems-aware framework for intelligent large-scale recommendation models. Journal of Computer Science and Electrical Engineering.

Fang, Q. (2024). TRUE-EWS: Trustworthy early warning with equalized conformal guarantees for equitable student support in U.S. education. Educational Research and Human Development.

Shen, A. (2023). Compute-efficient scaling of large sequential recommendation models under dynamic user behavior. World Journal of Information Technology.