

Multimodal Learning for Time Series Forecasting: Architectures, Challenges, and New Directions

Dao Thi Phuong Anh

Faculty of Information Technology, Hanoi University of Natural Resources and Environment
41A Phu Dien, Phu Dien Ward, Hanoi, Vietnam
Corresponding Author: Dao Thi Phuong Anh

ABSTRACT

Time series forecasting increasingly exploits information that is not itself a numerical series, such as news text, imagery, contextual metadata, and relational structure among series. The term "multimodal" is nonetheless used without an operational definition, so that genuinely cross-modal models are pooled with multivariate and multi-view models that raise different problems. This paper proposes an operational criterion, together with a four-question procedure, that separates multimodal forecasting from multivariate, multi-view, and cross-modal model-reuse settings. On this basis it develops a five-axis taxonomy covering modality type, temporal alignment granularity, fusion mechanism, model family, and output form with uncertainty representation, and presents a structured synthesis of 33 published studies across eight application domains. Applying the criterion consistently shows that only 14 studies fully satisfy the definition. Notably, five large-language-model-based forecasters presented as series-text alignment methods use only text generated from the input series itself, which cannot add information about the future. Combined with published ablation evidence, this indicates that the benefit of an auxiliary modality is conditional rather than universal, and that evaluation practice, rather than architecture, is the binding constraint of the field. Generalisation under a shifting context-target relationship — the very condition auxiliary context is expected to address — remains essentially untested. The paper analyses seven groups of open challenges and proposes five research directions, notably multimodal self-supervised learning and offline semantic feature extraction with large language models, together with an eight-item minimum evaluation protocol intended to make the contribution of an auxiliary modality verifiable.

KEYWORDS: Multimodal learning, time series forecasting, modality fusion, temporal alignment, endogenous text, large language models.

Date of Submission: 09-10-2026

Date of acceptance: 21-10-2026

I. INTRODUCTION

1.1 Motivation

Time series forecasting underpins decision making in energy, transport, public health, finance, and industrial operations. For most of its history the problem has been posed in unimodal form: the model observes the numerical series and nothing else. Deep architectures built on the attention mechanism [1], in particular patch-based representations and the inverted-variate design [2], [3], have substantially improved the modelling of long-range dependence, while the methodological literature on assessing forecast accuracy has developed in parallel [4].

Yet the events that move a time series frequently leave no numerical trace at the moment they matter. A monetary policy decision affects an exchange rate before it registers in any indicator; a newly launched product has an image and a description but no demand history; an industrial machine approaches failure in a way visible on an inspection photograph before it appears in vibration statistics. These observations motivate multimodal time series forecasting (MM-TSF): supplying the forecaster with text, imagery, metadata, or relational structure alongside the numerical series. The direction has accelerated sharply since large pretrained language and vision models made non-numerical modalities cheap to encode, and it is precisely this rapid expansion that creates the need for systematisation.

1.2 Related Surveys

Jiang et al. present a tutorial and survey of multimodal time series analysis, organising methods through a cross-modal interaction framework spanning fusion, alignment, and transference [5]. Liu et al. survey cross-modality modelling for time series analytics in the era of large language models [6]; an earlier survey addresses the first wave of language-model applications to time series [7]; and a comprehensive survey covers renewable energy systems specifically [8]. General background is provided by surveys of multimodal learning at large [9]–[12], by a review of multimodal clustering [13], and by the multimedia fusion literature, where the early–intermediate–late vocabulary originates [14], [15].

1.3 Research in Vietnam

Published domestic work on time series forecasting is predominantly unimodal. In the energy domain, a model combining multi-kernel convolutional blocks with an LSTM encoder has been proposed for weekend long-term load forecasting in Tay Ninh province and benchmarked against Transformer, Reformer, and Informer baselines [16]; the same journal reports comparisons of optimisation algorithms for WaveNet [17] and a stacking ensemble for short-term load forecasting [18]. In finance, a hybrid LSTM–GRU model has been evaluated for forecasting the VN-Index over 2008–2023 [19]. Notably, that study itself observes that equity prices respond to policy decisions and to crowd sentiment, while its model uses price history alone, and proposes exploiting such factors as future work [19] – a direct illustration of the multimodal gap this paper addresses. To the best of our knowledge, no MM-TSF dataset or benchmark has been constructed for Vietnamese data.

1.4 Research Gap and Contributions

Three gaps persist across the surveys above.

- No operational inclusion criterion. The boundary of "multimodal" is left implicit, so that models fusing news with prices, models combining several sensors of the same kind, and models reusing a language backbone on purely numerical input are listed side by side, although they differ in alignment difficulty, leakage exposure, and in what constitutes a fair baseline.
 - Organising schemes that conflate independent design choices. Neither the early–intermediate–late trichotomy nor the fusion–alignment–transference axis separates the granularity of temporal alignment from the stage of fusion, and neither treats the deployment condition under which a model was evaluated as a descriptive attribute.
 - Reported gains recorded without evaluation provenance, even though a controlled ablation has shown that removing the language-model component from several prominent forecasters does not degrade accuracy [20].
- The paper makes four contributions: (i) an operational definition with a four-question procedure that distinguishes exogenous from endogenous text, and a three-layer organisation of the corpus into core studies, a control group, and boundary studies (Sections 2–3); (ii) a five-axis taxonomy (Section 4); (iii) an analysis of representative architectures across eight domains together with datasets and leakage risks (Sections 5–6); and (iv) an analysis of seven groups of challenges, five research directions, and an eight-item minimum evaluation protocol (Sections 7–8). Section 9 concludes and states the limitations of this review.

II. REVIEW METHODOLOGY

This paper reports a structured review: research questions are fixed in advance, sources and query strings are recorded, inclusion and exclusion criteria are defined beforehand, and every retained study is coded on the same extraction form. We do not claim a systematic review in the PRISMA sense, because record counts at each screening stage were not retained in full; instead, transparency is secured by publishing the criteria and the classification rationale for every study in Table 5.

2.1 Research Questions

TABLE 1

Research Questions

ID	Research question
RQ1	Which modality types are used in MM-TSF, and how are they distributed across model families and application domains?
RQ2	At what temporal granularity are modalities aligned, and at what stage are they fused?
RQ3	Which datasets, forecasting tasks, and evaluation metrics are used?
RQ4	Under what conditions does an auxiliary modality genuinely improve forecasting relative to a matched unimodal baseline, and at what cost?
RQ5	How are models evaluated under missing or noisy modalities, information leakage, cold start, and distribution shift?
RQ6	What gaps remain for self-supervised learning, large language model exploitation, and generalisation in MM-TSF?

2.2 Sources and Query Strings

Seven sources were searched: Scopus, Web of Science Core Collection, IEEE Xplore, the ACM Digital Library, ScienceDirect, SpringerLink, and the official proceedings of NeurIPS, ICML, ICLR, AAAI, IJCAI, KDD, and ACM Multimedia. Proceedings were searched directly because several venues central to this field are indexed unevenly in the commercial databases. The window runs from January 2015 to the date of the search. The base query string is:

("multimodal" OR "multi-modal" OR "cross-modal") AND ("time series" OR "temporal") AND ("forecast*" OR "predict*") AND ("text" OR "language model" OR "image" OR "vision" OR "graph" OR "metadata" OR "context")

For each domain the final clause was replaced by domain vocabulary ("load", "photovoltaic" for energy; "traffic", "mobility" for transport; "clinical", "EHR" for healthcare; "remaining useful life", "prognostics" for industry), and Vietnamese-language sources were searched in domestic ISSN journals. Backward and forward citation chasing was applied to the surveys [5]–[8] and to every retained study.

2.3 Criteria and Corpus Structure

TABLE 2
Inclusion and Exclusion Criteria

Inclusion (all must hold)	Exclusion (any one suffices)
I1. At least one input is a time series.	E1. Task is only classification, anomaly detection, imputation, or retrieval.
I2. An auxiliary modality satisfying Definition 1 (Section 3.1) is present.	E2. Only multiple numerical channels of the same kind, with no cross-modal interaction.
I3. The task is forecasting future values, trajectories, or events.	E3. Preprint with no confirmed accepting venue.
I4. Quantitative experiments are reported.	E4. No quantitative experiment.
I5. Peer reviewed and officially published (Scopus/WoS journal or DOI-bearing proceedings).	E5. Duplicate or superseded version.

The corpus of 33 studies is organised into three layers so that the criteria are applied transparently and consistently. The core layer comprises 14 studies satisfying all of I1–I5. The control group comprises 12 studies of family H4 (Section 3.3) — eleven models and one ablation study — which violate I2 but are retained deliberately, because they are routinely cited as multimodal forecasting and are the direct object of RQ4. The boundary layer comprises seven studies that violate E1 or E2; they are retained to illustrate where the definition bites in each domain and are never used in comparisons. Two datasets currently available only as preprints [21], [22] are mentioned in Section 6 as benchmark context and support no conclusion.

2.4 Coding and Synthesis

Each study was coded on: venue; layer and family; domain; task; provenance of the auxiliary modality (exogenous or endogenous); alignment granularity; fusion mechanism; output form and uncertainty representation; presence of a matched unimodal baseline; computational cost; and tests for missing modalities, leakage, cold start, and out-of-distribution behaviour. Fields coded "not reported" are treated as findings. Synthesis is qualitative by design: because studies differ in datasets, look-back windows, normalisation, baseline strength, and metric definitions, no pooling and no meta-analysis is performed.

III. FOUNDATIONS AND SCOPE

3.1 Multimodal, Multivariate, and Multi-view Forecasting

Definition 1. A study belongs to multimodal time series forecasting when the model consumes at least two sources of information that differ substantially in representational structure or acquisition mechanism; at least one input is a time series; and the auxiliary modality is capable of supplying information that cannot be reduced to a function of that same series.

Multivariate forecasting enlarges the input with channels of the same kind, acquired by the same mechanism: adding a second temperature sensor changes the dimensionality of the input, not its modality. Multi-view forecasting supplies several transformations of one underlying signal — a spectrum, a wavelet decomposition, a heatmap rendered from the series — which is a change of representation rather than an addition of independently acquired information. Both are excluded by the third clause. The distinction matters because a gain reported by a multi-view model is a statement about the inductive bias of the architecture, whereas a gain reported by a multimodal model is a statement about added information; the two require different baselines to substantiate.

3.2 Problem Formulation

Let $X_{1:t}$ denote the observed history of the numerical series and h the forecast horizon. Unimodal forecasting learns the mapping

$$f: X_{1:t} \mapsto X_{t+1:t+h} \quad (1)$$

In the multimodal extension the model additionally observes auxiliary streams $M = \{m^{(1)}, \dots, m^{(K)}\}$, each a sequence of observations in a modality-specific space (text tokens, images, graphs, feature vectors) carrying its own timestamps. Forecasting becomes

$$\hat{X}_{t+1:t+h} = f(X_{1:t}, A(M, t)) \quad (2)$$

where A is an alignment operator that places the auxiliary streams onto the temporal support of the forecast, and A may depend only on observations timestamped at or before t . Making A explicit is the point of the formulation: that constraint is where leakage is blocked; A is where irregular sampling is handled, since auxiliary streams are typically sparser than the series; and A is the component most often left unspecified in the literature. Fig. 1 summarises the general framework and the position of the five taxonomic axes of Section 4. In event forecasting the target is a categorical or point-process object, so metrics are classification-style; in probabilistic forecasting the model returns a predictive distribution rather than a point.

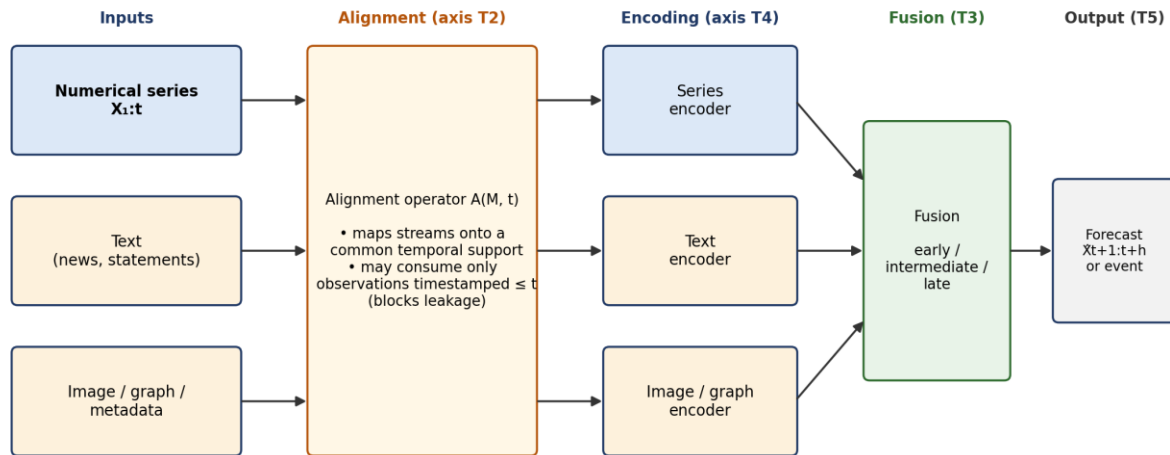


Fig. 1. General framework of MM-TSF and the position of taxonomic axes T2–T5. The alignment operator A is the locus of leakage control: it may consume no observation timestamped after t .

3.3 Exogenous and Endogenous Text

The third clause of Definition 1 has an important consequence for the family of models built on large language models. We call text exogenous when it is acquired independently of the series — news items, policy statements, reports, clinical notes — and endogenous when it is generated from the input window itself: descriptive statistics verbalised into sentences, series values wrapped into a prompt template, or prototypes drawn from the vocabulary table of a language model. Endogenous text is a deterministic function g of the series, so by a basic property of mutual information, for every such g ,

$$I(X_{t+1:t+h}; X_{1:t}, g(X_{1:t})) = I(X_{t+1:t+h}; X_{1:t}) \quad (3)$$

In other words, adding endogenous text cannot contribute any information about the future; any observed gain must come from the inductive bias of the language encoder or from prior knowledge embedded in the backbone, not from a new observation. This is not a criticism of the quality of such models — several of their designs are valuable as representation-learning techniques — but a requirement on how their evidence is read: they must be compared against a unimodal baseline of comparable capacity and cannot be treated as evidence for a multimodal benefit. Fig. 2 operationalises Definition 1 as four questions; Table 3 gives the resulting five families.

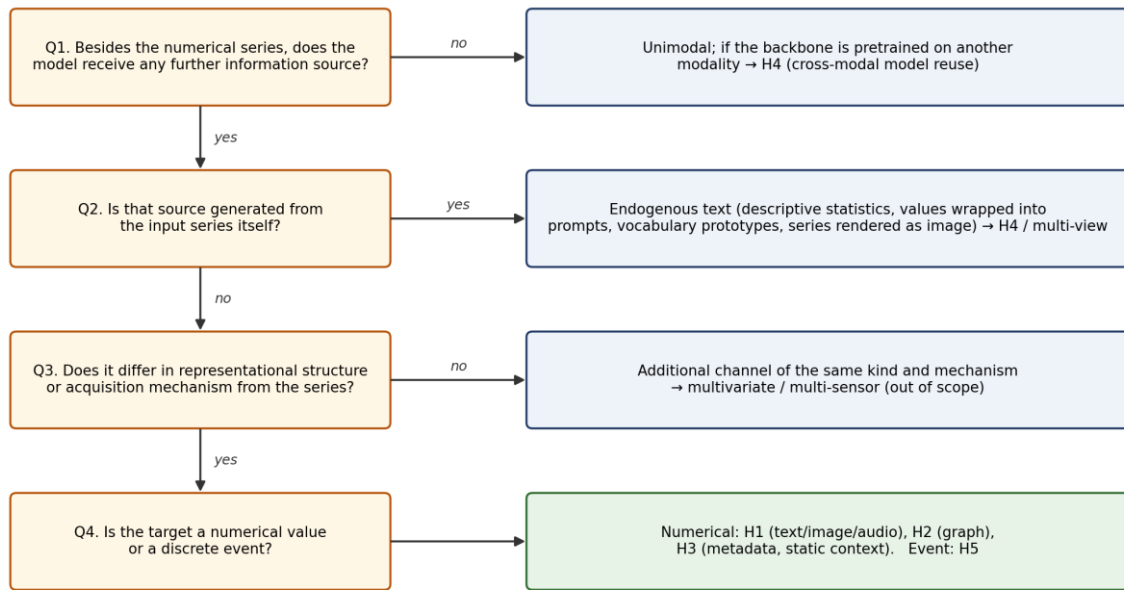


Fig. 2. Four-question procedure operationalising Definition 1. The second question separates exogenous from endogenous auxiliary content and is the step most often skipped in the literature.

TABLE 3
Five Families of Studies under Definition 1

Family	Name	Criterion	Representative studies (see Table 5)
H1	Strict MM-TSF	Series with exogenous text, imagery, or audio, having its own encoder and acquisition channel.	ChatTime; GPT4MTS; Modality-aware Transformer; CAMEF; From News to Forecast; MTFT; Wang et al. (2023)
H2	Relational MM-TSF	Series with an explicit graph or knowledge structure not derivable from the series.	MAGNN; ST-LLM+
H3	Context-aware	Series with exogenous metadata, covariates, or static context.	UniTime; UrbanMind; MTFT (tabular component)
H4	Cross-modal reuse / endogenous text	No exogenous modality: either numerical input with a backbone pretrained on another modality, or text/imagery generated from the series itself.	Time-LLM; One Fits All; AutoTimes; PromptCast; CALF; TimeCMA; BALM-TSF; TEST; Time-VLM; ST-LLM; TimeCAP
H5	Event forecasting	Target is the occurrence or type of a discrete event, with an exogenous auxiliary modality.	MM-Forecast; Sawhney et al. (2020); MONOPOLY

3.4 Modality Availability and Alignment

Training corpora are typically assembled so that every window carries every modality, whereas deployment rarely guarantees this: a news feed has quiet days, a product photograph may be missing for a new item, a clinical note may not be written. Three regimes should be distinguished: complete availability; intermittent availability; and structurally sparse availability, in which the auxiliary modality is present for a small minority of windows by the nature of the problem — the normal case for event-driven text. Alignment logically precedes fusion, and its granularity governs leakage directly: aligning a monthly summary to the month it describes injects end-of-month information into a forecast issued at the start of that month, violating the constraint on A in (2).

3.5 Forecasting Tasks and Evaluation Metrics

Point forecasting is dominated by MSE and MAE on normalised series. Both are scale dependent, so averaging across datasets is uninformative [4], and every comparison in this paper is made within a dataset. Probabilistic forecasting requires proper scoring rules such as CRPS or the pinball loss, which reward calibration as well as sharpness [23]. Trajectory forecasting in meteorology uses RMSE and anomaly correlation against an operational numerical baseline [24]. Event forecasting uses F1 and AUC, so results of family H5 are never placed beside those of H1–H4.

IV. A TAXONOMY OF MM-TSF METHODS

The early–intermediate–late trichotomy [14], [15] correctly describes the fusion stage but is a single axis. We describe every study on five independent axes (Table 4); the classical trichotomy is retained as a secondary attribute of axis T3.

- T1 — Modality type. Rendering a series as an image is a multi-view transformation rather than an added

modality, and is coded as such even when the original paper calls it multimodal. Configurations with more than two exogenous modalities remain rare.

- T2 — Temporal alignment granularity, from exact timestamp correspondence to learned latent correspondence. This is the least reported axis: many papers do not state their granularity and it must be inferred from the data-preparation description, yet it is the axis most tightly coupled to leakage and to irregular sampling.
- T3 — Fusion mechanism, with a stage label. Intermediate cross-attention is the modal choice in family H1; prompt verbalisation dominates family H4.
- T4 — Model family. Baseline strength varies by family, so an improvement over a recurrent network and an improvement over a time series foundation model are not the same claim.
- T5 — Output form and uncertainty representation. Across the corpus the modal value of the uncertainty attribute is "none".

TABLE 4
The Five Taxonomic Axes

Axis	Name	Categories
T1	Modality type	series + exogenous text; series + image; series + metadata; series + graph; multi-sensor; image-rendered series; endogenous text; more than two modalities
T2	Alignment granularity	exact timestamp; window-level; event-to-window; coarse document-level; asynchronous; learned latent
T3	Fusion mechanism	concatenation; gating; cross-attention; co-attention; graph message passing; mixture of experts; prompt verbalisation; decision fusion — each labelled early / intermediate / late
T4	Model family	statistical and shallow; recurrent-convolutional; Transformer; graph neural network; foundation model (language / vision / vision-language / time series)
T5	Output uncertainty and	point; quantile; trajectory; event; single- or multi-horizon — uncertainty: none; quantile; distributional; ensemble; conformal

V. REPRESENTATIVE ARCHITECTURES AND FUSION STRATEGIES BY DOMAIN

The domain determines which auxiliary modality is realistically available, how tight the alignment can be, and what counts as a useful forecast. This section begins with the group of cross-domain foundation models — the largest group, and the one where applying Definition 1 makes the greatest difference — and then proceeds through seven specific domains. Table 5 lists the whole corpus; Table 6 records the architectural coding.

5.1 Cross-domain Foundation Models and the Endogenous Text Problem

Among models using exogenous text, ChatTime builds a unified foundation model with a bimodal interface between numbers and text, supporting zero-shot forecasting and context-guided forecasting [25]; GPT4MTS collects news summaries from the GDELT event database, encodes them with BERT, and fuses them with the patched series through the prompt layers of GPT-2 [26]. UniTime uses static domain instructions in textual form to help a single unified model distinguish data from different domains [27]; because that text is exogenous but time-invariant, the study is placed in family H3.

On the other side is a larger group whose mechanisms are presented as series-text alignment or fusion, but whose text is entirely endogenous. CALF aligns the series representation with a low-dimensional projection of the language model vocabulary embedding through cross-attention [28]; the language branch of TimeCMA wraps the input series itself into a prompt template and retains only the embedding of the last token to reduce cost [29]; BALM-TSF feeds descriptive statistics of the raw series (minimum, maximum, median, trend) into a frozen language model together with a learnable prompt, then balances the two branches by variance scaling and a contrastive objective [30]; TEST aligns series embeddings with text prototypes [31]; and Time-VLM renders the series as an image for its vision branch while generating the text branch mainly from statistics of that same series together with a static dataset description [32]. The same family also contains reprogramming and prompting models that use no external text at all: Time-LLM [33], One Fits All [34], AutoTimes [35], and PromptCast [36].

The consequence of this classification is confirmed independently by experiment. Tan et al. ablate the language-model component from One Fits All, Time-LLM, and CALF — all three of family H4 under Table 3 — and find that across 13 datasets and two metrics the ablated variants outperform Time-LLM in 26 of 26 cases, CALF in 22 of 26, and One Fits All in 19 of 26; pretrained language models also fail to outperform models trained from scratch and do not help in few-shot settings [20]. This is exactly what (3) predicts. It does not negate the value of techniques such as variance balancing or token compression, but it shows that the gains reported by family H4 cannot be attributed to multimodality.

5.2 Finance

Finance is the oldest and densest domain, because price movements react to information that is natively textual and precisely timestamped. The Modality-aware Transformer fuses report text with numerical economic indicators through feature-level, intra-modal, and inter-modal attention layers to forecast US interest rates [37]. MAGNN represents assets as nodes of a heterogeneous graph with edges drawn from a financial knowledge graph [38]. CAMEF integrates the text of six categories of macroeconomic announcement with high-frequency price

series under a causal learning mechanism, augments training with counterfactual events generated by a large language model, and validates the causal component by ablation [39]; it is the only study in the corpus to model the causal structure between context and target explicitly. From News to Forecast employs language agents to select relevant news with a reflection mechanism [40]. The work of Lee and Yoo compares three fusion stages between domestic and foreign index series [41]; since both inputs are numerical series of the same kind, it belongs to the boundary layer (multivariate) despite its historical importance.

5.3 Demand and Retail

Retail faces cold start: for a new product the auxiliary modality is the only information that exists. MTFT combines a convolutional encoder for product imagery, a graph encoder for the geographical delivery structure, descriptive text, and tabular covariates such as seasonality and promotional events, emitting quantiles per product, warehouse, and delivery period, with experiments on more than 50 million demand data points [42]. It is the clearest instance in the corpus of both more-than-two exogenous modalities and a non-trivial uncertainty representation.

5.4 Energy

Wang et al. build a CNN-LSTM model fusing textual data with the electricity demand series for medium- and long-term forecasting of Chinese power demand, and report that the fused model outperforms single models such as ARIMA, CNN, and LSTM [43]. Renewable generation forecasting is dominated by sky and satellite imagery and numerical weather prediction fields alongside the generation series, and has been surveyed separately [8]. Two domain properties stand out: the auxiliary modality is often a physical observation, hence regularly sampled and rarely missing; and forecast error carries asymmetric cost, which argues for probabilistic outputs. Domestic studies of electricity load [16]–[18] remain unimodal, while event calendars, weather bulletins, and dispatch notices are readily available exogenous sources.

5.5 Weather and Climate

GraphCast learns an encode–process–decode graph network on an icosahedral mesh and matches or exceeds the operational high-resolution ECMWF model on most verification targets [24]; its relational machinery rests on inductive message passing [44]. Because the input is a single numerical modality organised relationally, the study belongs to the boundary layer. A genuinely multimodal extension appears with Terra, which pairs climate series with geographical imagery and textual descriptions at global scale [45]. The domain is distinctive in evaluating against a strong physical baseline, which makes its improvement claims easier to interpret than most of the corpus.

5.6 Traffic and Urban Mobility

ST-LLM treats the timesteps at each location as tokens, learns a spatio-temporal embedding, and passes it through a partially frozen language model [46]; since the input is entirely numerical, the study belongs to family H4. Its successor ST-LLM+ integrates the road-network graph explicitly — an exogenous structure — and reports robustness in few-shot and zero-shot settings [47], placing it in family H2. UrbanMind encodes spatio-temporal context into prompts and uses test-time adaptation to handle distribution shift between training and test periods [48], placing it in family H3. The domain retains unexploited multimodal potential: incident reports, event calendars, and social-media signals are precisely timestamped and causally upstream of traffic volume.

5.7 Healthcare and Industry: Two Domains with No Core Study

In healthcare, multimodality is least optional: physiological series are irregularly sampled, clinical notes appear unpredictably, and imaging is occasional. Work on irregular multimodal electronic health records combines irregularly sampled series and notes through time-attentive fusion [49]; MedFuse fuses clinical series with chest radiographs [50]; and MFNet fuses medical codes, series, notes, and demographics [51]. All three address outcome-prediction tasks of a classification nature and therefore fall in the boundary layer under E1, although the asynchronous alignment mechanism of [49] is directly relevant to MM-TSF.

In industrial prognostics, "multimodal" usually means multi-sensor or multi-view. MIASTN constructs a multi-scale degradation indicator space by decomposing the original vibration signal and fusing those views [52]; work on aero-engines performs self-adaptive fusion of data from multiple sensors [53]. Both fall in the boundary layer under E2. Genuine multimodality here would combine degradation series with maintenance logs, inspection imagery, or operator reports — sources that exist in industrial practice but have scarcely been exploited. The two domains with the clearest need for multimodality therefore contain no core study in the corpus: this is a gap in the literature rather than in the data.

5.8 Event-driven Forecasting

MM-Forecast determines whether an accompanying image plays a highlighting or a complementary role for an event, verbalises that role, and passes it to a language-model forecaster [54]. An earlier multimedia lineage forecasts financial risk and volatility from conference audio and text [55] and from monetary policy conference

video [56]. TimeCAP employs two language-model agents, one generating a textual context for the series and one predicting the event [57]; because that context is generated from the series itself, the study is assigned to the control group. Results of family H5 use classification metrics and are never pooled with the regression-metric domains.

TABLE 5
The Corpus of 33 Studies by Layer, Family, and Provenance of Auxiliary Information

ID	Study	Year	Venue	Domain	Auxiliary information (provenance)	Family
Core layer — satisfying I1–I5 (14 studies)						
C1	ChatTime [25]	2025	AAAI	cross-domain	contextual (exogenous) text	H1
C2	GPT4MTS [26]	2024	AAAI	cross-domain	GDELT summaries (exogenous) news	H1
C3	UniTime [27]	2024	WWW	cross-domain	static domain instructions textual	H3
C4	Modality-aware Transf. [37]	2024	ACM ICAIF	finance	report economic indicators text	+ H1
C5	MAGNN [38]	2022	Pattern Recognit.	finance	financial knowledge graph	H2
C6	CAMEF [39]	2025	KDD	finance	macroeconomic announcements (exogenous)	H1
C7	From News to Forecast [40]	2024	NeurIPS	cross-domain	news (exogenous) items	H1
C8	MTFT [42]	2024	IEEE MultiMedia	retail	imagery + text + tabular + geography	H1/H3
C9	Wang et al. [43]	2023	Energy	energy	textual (exogenous) data	H1
C10	ST-LLM+ [47]	2025	IEEE TKDE	traffic	road-network graph	H2
C11	UrbanMind [48]	2025	KDD	urban	spatio-temporal context as prompts	H3
C12	MM-Forecast [54]	2024	ACM MM	events	event imagery text	+ H5
C13	Sawhney et al. [55]	2020	ACM MM	finance	conference + text audio	H5
C14	MONOPOLY [56]	2022	ACM MM	finance	video + audio + text	H5
Control group — family H4, violating I2, retained for RQ4 (12 studies)						
C15	Time-LLM [33]	2024	ICLR	cross-domain	none; LLM reprogramming	H4
C16	One Fits All [34]	2023	NeurIPS	cross-domain	none; LM backbone	H4
C17	AutoTimes [35]	2024	NeurIPS	cross-domain	none; LLM backbone	H4
C18	PromptCast [36]	2024	IEEE TKDE	cross-domain	series verbalised (endogenous)	H4
C19	CALF [28]	2025	AAAI	cross-domain	LLM vocabulary embeddings (endogenous)	H4
C20	TimeCMA [29]	2025	AAAI	cross-domain	series wrapped into prompts (endogenous)	H4

ID	Study	Year	Venue	Domain	Auxiliary information (provenance)	Family
C21	BALM-TSF [30]	2025	CIKM	cross-domain	descriptive statistics of the series (endogenous)	H4
C22	TEST [31]	2024	ICLR	cross-domain	text prototypes (endogenous)	H4
C23	Time-VLM [32]	2025	ICML	cross-domain	series rendered as image + statistics text	H4
C24	ST-LLM [46]	2024	IEEE MDM	traffic	none; LLM backbone	H4
C25	TimeCAP [57]	2025	AAAI	events	context generated by an agent from the series	H4
C26	Tan et al. [20]	2024	NeurIPS	cross-domain	ablation study on C15, C16, C19	—
Boundary layer — violating E1 or E2, never used in comparisons (7 studies)						
C27	Lee & Yoo [41]	2020	J. Supercomput.	finance	foreign index series (multivariate)	E2
C28	GraphCast [24]	2023	Science	meteorology	numerical mesh as graph (single modality)	E2
C29	Irregular EHR [49]	MM 2023	ICML	healthcare	series + notes; classification task	E1
C30	MedFuse [50]	2022	MLHC	healthcare	series + radiographs; classification task	E1
C31	MFNet [51]	2024	Neural Networks	healthcare	codes, series, notes; outcome classification	E1
C32	MASTN [52]	2025	Expert Appl.	Syst. industry	multi-scale decomposition of a vibration signal (multi-view)	E2
C33	Aero-engine fusion [53]	2023	Reliab. Syst. Saf.	Eng. industry	multi-sensor	E2

TABLE 6
Architectural Coding on Axes T2–T5 (Core Layer and Selected Control Studies)

ID	Study	T2 alignment	T3 fusion (stage)	T4 model family	T5 uncertainty
C1	ChatTime	window-level	prompt verbalisation (early)	language foundation model	none
C2	GPT4MTS	window-level	embedding concatenation (intermediate)	Transformer + LM	none
C3	UniTime	coarse level document-	concatenation (early)	Transformer + LM	none
C4	Modality-aware Transf.	window-level	intra- and inter-modal attention (intermediate)	Transformer	none
C5	MAGNN	window-level	graph message passing (intermediate)	graph neural network	none
C6	CAMEF	event-to-window	cross-attention + causal contrastive (intermediate)	TS foundation model + LM	none
C7	From News to Forecast	event-to-window	prompt verbalisation + agent selection (early)	LLM agents	none
C8	MTFT	window-level	gating + attention (intermediate)	Transformer + CNN + GNN	quantile
C9	Wang et al.	window-level	feature concatenation (intermediate)	CNN-LSTM	none

ID	Study	T2 alignment	T3 fusion (stage)	T4 model family	T5 uncertainty
C10	ST-LLM+	exact timestamp	graph-enhanced (intermediate)	fusion GNN + LLM	none
C12	MM-Forecast	event-to-window	verbalised image role (early)	LLM + MLLM	none
C19	CALF (control)	not applicable	cross-attention with vocabulary embeddings	LLM	none
C20	TimeCMA (control)	not applicable	cross-branch (intermediate)	retrieval Transformer + LLM	none
C21	BALM-TSF (control)	not applicable	concatenation after contrastive alignment	Transformer + LLM	none
C23	Time-VLM (control)	not applicable	vision-language co-attention	vision-language model	none

For the control group, axis T2 is recorded as "not applicable" because there is no exogenous auxiliary stream to align. Alignment granularity for the core layer was inferred from the data-preparation description where the original paper does not state it explicitly.

VI. DATASETS AND BENCHMARKS

Benchmark design has been the binding constraint of this field. The authors of Time-MMD motivate their work by three deficiencies of earlier data: concentration in few domains, coarse text-series alignment, and contamination where the accompanying text already discloses the outcome to be forecast; Time-MMD spans nine domains with fine-grained alignment and explicit contamination control [58]. Context-is-Key supplies seventy-one tasks in which textual context is necessary, so that a model ignoring the text cannot succeed [59]. MoTime evaluates the usefulness of the auxiliary modality under two separate scenarios, one of which is cold start [21]; TimeText Corpus pairs series with text in climate and healthcare [22].

TABLE 7
Datasets and Benchmarks

Dataset	Venue	Modalities	Domains	Distinctive property	Leakage control
Time-MMD [58]	NeurIPS D&B 2024	series + text	9 domains	fine-grained alignment	descriptive predictive separated vs. text
Context-is-Key [59]	ICML 2025	series + text	multiple	71 tasks requiring context	by construction
Terra [45]	NeurIPS 2024	series + imagery + text	climate, geospatial	global coverage	not stated
CAMEF corpus [39]	KDD 2025	text + high-frequency prices	finance	6 macroeconomic event types	anchored to event time
MoTime [21]	preprint 2025	series + text + tabular + imagery	multiple	cold-start scenario	not stated
TimeText [22]	Corpus preprint 2024	series + text	climate, healthcare	time-aligned	not stated
ETT, Electricity, Traffic, Weather	various	series only	energy, transport, meteorology	standard benchmarks	unimodal not applicable

The unimodal benchmarks are listed because most of the control group reports its headline results on them. A benchmark containing no auxiliary modality cannot demonstrate a multimodal benefit; it can only show that an architecture does not lose accuracy when its auxiliary channel carries nothing.

Dataset-level alignment and information leakage

The native resolution of a dataset bounds the achievable alignment granularity: if the text is timestamped only by month, the effective granularity remains document-level whatever the architecture. Three leakage mechanisms recur: retrospective description (a summary written after the period it covers), publication-time misattribution (associating an article with the event date rather than the publication date), and pretraining contamination (a foundation model having seen the evaluation period). The first is addressed by Time-MMD through the separation of descriptive from predictive text [58]; the third is almost never audited in the corpus.

Licensing, reproducibility, and benchmark limitations

Text and imagery — precisely what makes a corpus multimodal — frequently carry licences prohibiting redistribution, so many corpora are released as retrieval scripts; content drifts over time, making exact reproduction difficult even with released code. Domain coverage also remains concentrated in finance and general cross-domain settings; industry has no genuinely multimodal benchmark; and no benchmark supplies a controlled shift split in which the context-target relationship itself changes.

VII. OPEN CHALLENGES AND LIMITATIONS

7.1 Temporal Alignment across Modalities

Within the core layer, window-level alignment predominates, event-to-window alignment appears only in the event-driven lineage, and asynchronous alignment is addressed almost exclusively in the healthcare literature [49], which belongs to the boundary layer. A general mechanism for aligning irregularly sampled auxiliary modalities onto a regular forecast support remains open.

7.2 Usefulness of the Auxiliary Modality

This is where the field is least settled. For the control group, the evidence in Section 5.1 and equation (3) show that reported gains cannot be attributed to multimodality. For the core layer, gains are reported widely but under heterogeneous protocols, frequently without an identically trained unimodal baseline and almost never with significance testing. Baseline strength also varies greatly: from long-horizon Transformer forecasters [2], [3], [60], [61], through the linear models shown to match or exceed them [62], to time series foundation models [63], [64]. A multimodal model that beats a weak Transformer but loses to a linear baseline has demonstrated nothing about multimodality.

7.3 Missing and Noisy Modalities

Within the surveyed corpus, no core study reports a controlled modality-withholding experiment at inference time, although structurally sparse availability is the normal case for event-driven text. The general multimodal learning literature treats this as an open problem [9], [13]; that framing has not transferred into forecasting practice. Robustness to noisy or misleading modalities is likewise unevaluated.

7.4 Information Leakage

Of the three leakage mechanisms in Section 6, only retrospective description is handled systematically, and only at the dataset level [58], [59]. Pretraining contamination affects every study using a foundation-model backbone on a historical evaluation period, yet is almost never audited.

7.5 Generalisation and Distribution Shift

Domain shift receives partial attention through cross-domain designs [25], [27], test-time adaptation [48], and few-shot or zero-shot protocols [47]. But concept shift — where the relationship between context and target itself changes — is not systematically evaluated anywhere in the corpus. CAMEF comes closest through counterfactual event augmentation [39], but does not evaluate under an explicit reversal of that relationship. The omission is striking because the motivating argument for MM-TSF is precisely that context carries information about regime changes.

7.6 Interpretability and Uncertainty

Agent-based designs [40], [57] generate natural-language explanations, but an explanation produced alongside a forecast is not evidence that the forecast depended on the cited context; methods establishing counterfactual dependence are needed. As for uncertainty, quantile output appears only in retail [42] and ensembles in operational meteorology [24], while a modality that is sometimes present, sometimes noisy, and sometimes irrelevant should change predictive uncertainty — which a point forecast cannot express. Proper scoring rules [23] are seldom used.

7.7 Computational Cost and Scalability

Three cost regimes are distinguishable: lightweight adaptation, freezing the auxiliary encoder and training only an alignment module [30], [39]; prompt-based approaches, incurring inference cost proportional to prompt length, which is why TimeCMA compresses information into a final token and caches its embedding [29]; and full multimodal training [42]. Agent designs that call a language model repeatedly per forecast occupy a substantially higher cost regime. The ablation result [20] carries a cost corollary: if removing the language model does not reduce accuracy, the entire cost of retaining it has bought nothing. Architectures tied to a fixed modality set further cause cost to grow with the number of supported configurations.

VIII. RESEARCH DIRECTIONS AND EVALUATION RECOMMENDATIONS

8.1 Multimodal Self-supervised Learning

Hierarchical contrastive learning over augmented context views produces timestamp-level representations good enough that a linear regression on top competes with specialised forecasters [65]; time–frequency consistency provides a further self-supervised signal [66]; and multimodal self-supervised learning in general has been surveyed separately [12]. These two lines have scarcely met in MM-TSF. Three pretext tasks can be envisaged: discriminating text–window pairs that share a timestamp from temporally mismatched pairs; predicting the relative temporal order of a textual event and a movement in the series; and reconstructing a masked series window from exogenous text of the same period. These generate supervision without forecast labels while forcing the model to learn the alignment relation itself. The direction suits domestic conditions particularly well, where

labelled data are scarce but raw paired data — load series and news bulletins, stock indices and financial reporting — are relatively abundant.

8.2 Verifiable Exploitation of Large Language Models

Large language models currently play three roles that the literature rarely distinguishes: forecasting backbone (family H4), encoder for exogenous text (family H1), and instrument for knowledge extraction or data augmentation, as in the counterfactual event generation of CAMEF [39]. The ablation evidence [20] and equation (3) weaken the first role but say nothing about the other two. The direction proposed here is to use large language models offline to extract from exogenous text a structured, timestamped set of semantic features — event type, polarity, severity, scope of impact — which are then fed to the forecaster. This separates inference cost from the forecasting loop, makes the contribution of the text attributable, and permits leakage control at the feature level.

8.3 Modality-flexible Architectures and Adaptive Modality Selection

Almost every architecture in Table 6 assumes a fixed modality set. Architectures that remain well defined as modalities appear and disappear — through modality-agnostic encoders into a shared space, explicit absence tokens, or training that randomly withholds modalities — would make robustness to missing modalities a property of the design. A complementary direction is adaptive modality selection: the model decides per window whether to consult the auxiliary modality, routing to a cheap unimodal branch when the text carries no signal, thereby addressing usefulness (Section 7.2) and cost (Section 7.7) simultaneously.

8.4 Regime-aware and Causally Robust Forecasting

To close the gap in Section 7.5, two complementary directions are available. First, conditioning the model on latent regimes, with the auxiliary modality serving as the regime-detection source, since a policy change is described in text before it appears in the series. Second, reducing dependence on unstable context features in the spirit of invariant learning [67] and gradient-reversal adversarial training [68], grounded in the structural causal model framework [69]. These are proposed directions, not verified conclusions. The verification protocol is straightforward: construct paired training and evaluation splits in which a measurable context feature carries a controlled correlation with the target, high during training and weakened or reversed at evaluation; then compare the multimodal model, the matched unimodal baseline, and the multimodal model under a dependence-reducing constraint on identical splits.

8.5 Contamination-aware Benchmarks

Benchmarks should separate descriptive from predictive text as Time-MMD does [58], audit and report pretraining contamination, attach retrieval dates and content checksums for corpora distributed as scripts, and supply at least one controlled context–target shift split. In the domestic context, an MM-TSF benchmark for Vietnamese data — pairing load series or stock indices with Vietnamese news and announcements carrying publication timestamps — is a feasible and currently vacant contribution.

8.6 A Minimum Evaluation Protocol

Table 8 lists eight items that an MM-TSF publication should report for the contribution of the auxiliary modality to be attributable. Item 1 is new relative to existing recommendations and follows directly from the analysis in Section 3.3

TABLE 8
Recommended Minimum Evaluation Protocol for an MM-TSF Publication

#	Item to report	Rationale and verification
1	Provenance of the auxiliary modality	State whether text or imagery is exogenous or generated from the series (Fig. 2). Endogenous text cannot add information about the future, by (3).
2	Matched unimodal baseline	Same architecture, data, training procedure, and hyperparameter selection, differing only in the absence of the auxiliary modality; plus at least one strong linear baseline.
3	Temporal alignment granularity	State it on axis T2 and check it against the native resolution of the dataset.
4	Leakage-control argument	Timestamping rules guaranteeing the constraint on A in (2); separation of descriptive from predictive text; a pretraining-contamination check where a foundation model is used.
5	Results under modality withholding	Measure degradation when each modality is removed in turn at inference time.
6	Dispersion and significance	Mean with standard deviation over at least three random seeds; significance testing where feasible.
7	Training and inference cost	Trainable parameter count, training time, and inference cost per forecast.
8	Deployment conditions evaluated	Complete, missing, and noisy modalities; domain shift; regime shift; cold start — and an explicit statement of which conditions were not evaluated.

IX. CONCLUSION

This paper has fixed the scope of multimodal time series forecasting with an operational definition and a four-question procedure, separated it explicitly from multivariate, multi-view, and cross-modal model-reuse settings, and organised the literature on five axes that disentangle design choices the classical fusion trichotomy conflates. Applying the definition consistently to 33 studies shows that only 14 fully satisfy it; a substantial group of language-model-based forecasters uses endogenous text only, and published ablation evidence is consistent with the conclusion that their reported gains cannot be attributed to multimodality. Healthcare and industry — the two domains with the clearest need for multimodality — contain no core study in the corpus.

Three conclusions stand out. The benefit of an auxiliary modality is conditional. Evaluation practice, not architecture, is the binding constraint of the field. And generalisation under a shifting context–target relationship, the very condition auxiliary context is expected to address, remains untested. In the domestic context, the two most feasible short-term directions are multimodal self-supervised learning and the offline use of large language models to extract semantic features from Vietnamese text, together with the construction of an MM-TSF benchmark for Vietnamese data.

This review has the following limitations. It is a structured rather than a PRISMA-compliant systematic review; screening and coding were performed by a single reviewer, so the application of Definition 1 in boundary cases may be uneven; criterion E3 favours verifiability over currency, so the most recent results available only on arXiv are not included; and coverage of Vietnamese-language literature remains limited. Coding of axis T2 rests on the methodological description where the original paper does not state the granularity explicitly.

ACKNOWLEDGEMENTS

This research was carried out within the framework of the 2026 institutional science and technology project of Hanoi University of Natural Resources and Environment, grant number HUNRE.2026.11.05.

REFERENCES

- [1] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.
- [2] Y. Nie, N. H. Nguyen, P. Sinthong, J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2023.
- [3] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, M. Long, "iTransformer: Inverted transformers are effective for time series forecasting," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2024.
- [4] R. J. Hyndman, A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679–688, 2006.
- [5] Y. Jiang, K. Ning, Z. Pan, X. Shen, J. Ni, W. Yu, A. Schneider, H. Chen, Y. Nevmyvaka, D. Song, "Multi-modal time series analysis: A tutorial and survey," in *Proc. 31st ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD)*, 2025, pp. 6043–6053.
- [6] C. Liu, S. Zhou, Q. Xu, H. Miao, C. Long, Z. Li, R. Zhao, "Towards cross-modality modeling for time series analytics: A survey in the LLM era," in *Proc. 34th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2025.
- [7] Y. Jiang, Z. Pan, X. Zhang, S. Garg, A. Schneider, Y. Nevmyvaka, D. Song, "Empowering time series analysis with large language models: A survey," in *Proc. 33rd Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2024.
- [8] M. Mansouri, K. Attouri, S. S. Refaat, "Multimodal learning techniques for time series forecasting in renewable energy systems: A comprehensive survey," *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3602914.
- [9] T. Baltrušaitis, C. Ahuja, L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019.
- [10] P. P. Liang, A. Zadeh, L.-P. Morency, "Foundations and trends in multimodal machine learning: Principles, challenges, and open questions," *ACM Computing Surveys*, vol. 56, no. 10, 2024.
- [11] P. Xu, X. Zhu, D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12113–12132, 2023.
- [12] Y. Zong, O. Mac Aodha, T. M. Hospedales, "Self-supervised multimodal learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 7, pp. 5299–5318, 2025.
- [13] S. Raya, M. Orabi, I. Afyouni, A. M. Khedr, "Multi-modal data clustering using deep learning: A systematic review," *Neurocomputing*, vol. 607, p. 128348, 2024.
- [14] P. K. Atrey, M. A. Hossain, A. El Saddik, M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Systems*, vol. 16, pp. 345–379, 2010.
- [15] S. Y. Boulahia, A. Amamra, M. R. Madi, S. Daikh, "Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition," *Machine Vision and Applications*, vol. 32, no. 6, 2021.
- [16] B. Q. Hoan, V. V. Ba, T. X. V. Nghia, V. N. Dieu, and L. T. Kien, "Ứng dụng học sâu trong xây dựng mô hình dự báo phụ tải dài hạn tại các thời điểm cuối tuần tại Tây Ninh" ["Applying deep learning to build a long-term weekend load forecasting model in Tay Ninh"], *TNU Journal of Science and Technology*, vol. 230, no. 15, 2025, doi: 10.34238/tnu-jst.13702. (in Vietnamese)
- [17] D. N. Hung, N. T. Linh, and N. M. Tam, "So sánh thuật toán tối ưu của mạng Wavenet trong bài toán dự báo phụ tải điện" ["Comparison of optimisation algorithms for WaveNet in electricity load forecasting"], *TNU Journal of Science and Technology*, vol. 228, no. 07, pp. 37–45, 2023, doi: 10.34238/tnu-jst.6956. (in Vietnamese)
- [18] N. A. Tuan, N. M. Duc, N. T. Ngan, and T. T. Ngoc, "Nghiên cứu và ứng dụng mô hình tổ hợp Stacking Ensemble trong dự báo phụ tải ngắn hạn" ["Research and application of the stacking ensemble model in short-term load forecasting"], *TNU Journal of Science and Technology*, vol. 230, no. 06, pp. 148–156, 2025, doi: 10.34238/tnu-jst.11886. (in Vietnamese)
- [19] T. D. Tuyen, "Đánh giá hiệu suất mô hình phức hợp LSTM-GRU: Nghiên cứu điển hình về dự báo chỉ số đo lường xu hướng biến động giá cổ phiếu trên sàn giao dịch chứng khoán Hồ Chí Minh" ["Performance evaluation of a hybrid LSTM-GRU model: a case study on forecasting the stock price trend index of the Ho Chi Minh Stock Exchange"], *Can Tho University Journal of Science*, vol. 60, no. 1D, pp. 235–249, 2024, doi: 10.22144/ctujos.2023.232. (in Vietnamese)
- [20] M. Tan, M. A. Merrill, V. Gupta, T. Althoff, T. Hartvigsen, "Are language models actually useful for time series forecasting?," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, 2024, pp. 60162–60191.
- [21] X. Zhou, W. Wang, F. J. Baldán, W. Buntine, C. Bergmeir, "MoTime: A dataset suite for multimodal time series forecasting," *arXiv preprint arXiv:2505.15072*, 2025.
- [22] K. Kim, H. Tsai, R. Sen, A. Das, Z. Zhou, A. Tanpure, M. Luo, R. Yu, "Multi-modal forecaster: Jointly predicting time series and textual data," *arXiv preprint arXiv:2411.06735*, 2024.
- [23] T. Gneiting, A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.
- [24] R. Lam et al., "Learning skillful medium-range global weather forecasting," *Science*, vol. 382, no. 6677, pp. 1416–1421, 2023.
- [25] C. Wang, Q. Qi, J. Wang, H. Sun, Z. Zhuang, J. Wu, L. Zhang, J. Liao, "ChatTime: A unified multimodal time series foundation model bridging numerical and textual data," in *Proc. AAAI Conf. on Artificial Intelligence*, vol. 39, no. 12, 2025, pp. 12694–12702.
- [26] F. Jia, K. Wang, Y. Zheng, D. Cao, Y. Liu, "GPT4MTS: Prompt-based large language model for multimodal time-series forecasting," in

- Proc. AAAI Conf. on Artificial Intelligence, vol. 38, no. 21, 2024, pp. 23343–23351.
- [27] X. Liu, J. Hu, Y. Li, S. Diao, Y. Liang, B. Hooi, R. Zimmermann, "UniTime: A language-empowered unified model for cross-domain time series forecasting," in Proc. ACM Web Conference (WWW), 2024, pp. 4095–4106.
- [28] P. Liu, H. Guo, T. Dai, N. Li, J. Bao, X. Ren, Y. Jiang, S.-T. Xia, "CALF: Aligning LLMs for time series forecasting via cross-modal fine-tuning," in Proc. AAAI Conf. on Artificial Intelligence, vol. 39, no. 18, 2025, pp. 18915–18923.
- [29] C. Liu, Q. Xu, H. Miao, S. Yang, L. Zhang, C. Long, Z. Li, R. Zhao, "TimeCMA: Towards LLM-empowered multivariate time series forecasting via cross-modality alignment," in Proc. AAAI Conf. on Artificial Intelligence, vol. 39, no. 18, 2025, pp. 18780–18788.
- [30] S. Zhou, H. Schöner, H. Lyu, É. Fouché, S. Wang, "BALM-TSF: Balanced multimodal alignment for LLM-based time series forecasting," in Proc. 34th ACM Int. Conf. on Information and Knowledge Management (CIKM), 2025, pp. 4498–4508.
- [31] C. Sun, H. Li, Y. Li, S. Hong, "TEST: Text prototype aligned embedding to activate LLM's ability for time series," in Proc. Int. Conf. on Learning Representations (ICLR), 2024.
- [32] S. Zhong, W. Ruan, M. Jin, H. Li, Q. Wen, Y. Liang, "Time-VLM: Exploring multimodal vision-language models for augmented time series forecasting," in Proc. 42nd Int. Conf. on Machine Learning (ICML), PMLR vol. 267, 2025, pp. 78478–78497.
- [33] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, Q. Wen, "Time-LLM: Time series forecasting by reprogramming large language models," in Proc. Int. Conf. on Learning Representations (ICLR), 2024.
- [34] T. Zhou, P. Niu, X. Wang, L. Sun, R. Jin, "One fits all: Power general time series analysis by pretrained LM," in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, 2023.
- [35] Y. Liu, G. Qin, X. Huang, J. Wang, M. Long, "AutoTimes: Autoregressive time series forecasters via large language models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 37, 2024.
- [36] H. Xue, F. D. Salim, "PromptCast: A new prompt-based learning paradigm for time series forecasting," IEEE Trans. Knowledge and Data Engineering, vol. 36, no. 11, pp. 6851–6864, 2024.
- [37] H. Emami Gohari, X.-H. Dang, S. Y. Shah, P. Zefros, "Modality-aware Transformer for financial time series forecasting," in Proc. 5th ACM Int. Conf. on AI in Finance (ICAIF), 2024, pp. 677–685.
- [38] D. Cheng, F. Yang, S. Xiang, J. Liu, "Financial time series forecasting with multi-modality graph neural network," Pattern Recognition, vol. 121, p. 108218, 2022.
- [39] Y. Zhang, W. Yang, J. Wang, Q. Ma, J. Xiong, "CAMEF: Causal-augmented multi-modality event-driven financial forecasting by integrating time series patterns and salient macroeconomic announcements," in Proc. 31st ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD), 2025.
- [40] X. Wang, M. Feng, J. Qiu, J. Gu, J. Zhao, "From news to forecast: Integrating event analysis in LLM-based time series forecasting with reflection," in Advances in Neural Information Processing Systems (NeurIPS), vol. 37, 2024, pp. 58118–58153.
- [41] S. I. Lee, S. J. Yoo, "Multimodal deep learning for finance: Integrating and forecasting international stock markets," The Journal of Supercomputing, vol. 76, no. 10, pp. 8294–8312, 2020.
- [42] M. Sukel, S. Rudinac, M. Worring, "Multimodal Temporal Fusion Transformers are good product demand forecasters," IEEE MultiMedia, vol. 31, no. 2, pp. 48–60, 2024.
- [43] D. Wang, J. Gan, J. Mao, F. Chen, L. Yu, "Forecasting power demand in China with a CNN-LSTM model including multimodal information," Energy, vol. 263, p. 126012, 2023.
- [44] W. L. Hamilton, R. Ying, J. Leskovec, "Inductive representation learning on large graphs," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
- [45] W. Chen, X. Hao, Y. Wu, Y. Liang, "Terra: A multimodal spatio-temporal dataset spanning the Earth," in Advances in Neural Information Processing Systems (NeurIPS), vol. 37, 2024.
- [46] C. Liu, S. Yang, Q. Xu, Z. Li, C. Long, Z. Li, R. Zhao, "Spatial-temporal large language model for traffic prediction," in Proc. 25th IEEE Int. Conf. on Mobile Data Management (MDM), 2024.
- [47] C. Liu, K. H. Hettige, Q. Xu, C. Long, S. Xiang, G. Cong, Z. Li, R. Zhao, "ST-LLM+: Graph enhanced spatio-temporal large language models for traffic prediction," IEEE Trans. Knowledge and Data Engineering, 2025.
- [48] Y. Liu, Y. Zhang, X. Zhang, L. Tian, Y. Li, J. Luo, "UrbanMind: Urban dynamics prediction with multifaceted spatial-temporal large language models," in Proc. 31st ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD), vol. 2, 2025, pp. 1951–1962.
- [49] X. Zhang, S. Li, Z. Chen, X. Yan, L. Petzold, "Improving medical predictions by irregular multimodal electronic health records modeling," in Proc. 40th Int. Conf. on Machine Learning (ICML), PMLR vol. 202, 2023.
- [50] N. Hayat, K. J. Geras, F. E. Shamout, "MedFuse: Multi-modal fusion with clinical time-series data and chest X-ray images," in Proc. Machine Learning for Healthcare Conference (MLHC), PMLR vol. 182, 2022, pp. 479–503.
- [51] C. Wang, X. Yang, M. Sun, Y. Gu, J. Niu, W. Zhang, "Multimodal fusion network for ICU patient outcome prediction," Neural Networks, vol. 180, p. 106672, 2024.
- [52] Y. Zhou, H. Wang, H. Jin, Y. Liu, X. Liu, Z. Cao, "Remaining useful life prediction for machinery using multimodal interactive attention spatial-temporal networks with deep ensembles," Expert Systems with Applications, vol. 263, p. 125808, 2025.
- [53] J. Chen, D. Li, R. Huang, Z. Chen, W. Li, "Aero-engine remaining useful life prediction method with self-adaptive multimodal data fusion and cluster-ensemble transfer regression," Reliability Engineering & System Safety, vol. 234, 2023.
- [54] H. Li, Z. Yang, Y. Ma, Y. Bin, Y. Yang, T.-S. Chua, "MM-Forecast: A multimodal approach to temporal event forecasting with large language models," in Proc. 32nd ACM Int. Conf. on Multimedia (ACM MM), 2024, pp. 2776–2785.
- [55] R. Sawhney, P. Mathur, A. Mangal, P. Khanna, R. R. Shah, R. Zimmermann, "Multimodal multi-task financial risk forecasting," in Proc. 28th ACM Int. Conf. on Multimedia (ACM MM), 2020.
- [56] P. Mathur, A. Neerkaje, M. Chhibber, R. Sawhney, F. Guo, F. Démoncourt, S. Dutta, D. Manocha, "MONOPOLY: Financial prediction from monetary policy conference videos using multimodal cues," in Proc. 30th ACM Int. Conf. on Multimedia (ACM MM), 2022.
- [57] G. Lee, W. Yu, K. Shin, W. Cheng, H. Chen, "TimeCAP: Learning to contextualize, augment, and predict time series events with large language model agents," in Proc. AAAI Conf. on Artificial Intelligence, vol. 39, no. 17, 2025, pp. 18082–18090.
- [58] H. Liu, S. Xu, Z. Zhao, L. Kong, H. Kamarthi, A. B. Sasanur, M. Sharma, J. Cui, Q. Wen, C. Zhang, B. A. Prakash, "Time-MMD: Multi-domain multimodal dataset for time series analysis," in Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, vol. 37, 2024, pp. 77888–77933.
- [59] A. R. Williams, A. Ashok, É. Marcotte, V. Zantedeschi, J. Subramanian, R. Riachi, J. Requeima, A. Lacoste, I. Rish, N. Chapados, A. Drouin, "Context is Key: A benchmark for forecasting with essential textual information," in Proc. 42nd Int. Conf. on Machine Learning (ICML), 2025.
- [60] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in Proc. AAAI Conf. on Artificial Intelligence, vol. 35, no. 12, 2021, pp. 11106–11115.
- [61] H. Wu, J. Xu, J. Wang, M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, 2021, pp. 22419–22430.
- [62] A. Zeng, M. Chen, L. Zhang, Q. Xu, "Are transformers effective for time series forecasting?," in Proc. AAAI Conf. on Artificial Intelligence, vol. 37, no. 9, 2023, pp. 11121–11128.
- [63] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, D. Sahoo, "Unified training of universal time series forecasting transformers," in Proc. 41st Int. Conf. on Machine Learning (ICML), PMLR vol. 235, 2024.
- [64] X. Shi, S. Wang, Y. Nie, D. Li, Z. Ye, Q. Wen, M. Jin, "Time-MoE: Billion-scale time series foundation models with mixture of experts," in Proc. Int. Conf. on Learning Representations (ICLR), 2025.
- [65] Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong, B. Xu, "TS2Vec: Towards universal representation of time series," in Proc. AAAI Conf. on Artificial Intelligence, vol. 36, no. 8, 2022, pp. 8980–8987.
- [66] X. Zhang, Z. Zhao, T. Tsiligkaridis, M. Zitnik, "Self-supervised contrastive pre-training for time series via time-frequency consistency," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, 2022, pp. 3988–4003.
- [67] Q. Wu, H. Zhang, J. Yan, D. Wipf, "Handling distribution shifts on graphs: An invariance perspective," in Proc. Int. Conf. on Learning Representations (ICLR), 2022.
- [68] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, "Domain-adversarial training of neural networks," Journal of Machine Learning Research, vol. 17, no. 59, pp. 1–35, 2016.
- [69] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2009.