MULTIMODAL INFLUENCE LEARNING FOR URBAN SPATIO-TEMPORAL FORECASTING: A CROSS-CITY FOUNDATION MODEL AND DATASET

Anonymous authors
Paper under double-blind review

ABSTRACT

Urban spatio-temporal foundation models learn reusable forecasting patterns by pretraining across cities, yet they rely mainly on single-modal spatiotemporal observational data. Multimodal information such as roads, remote sensing, weather, and events provides predictive evidence beyond these observations, but transferring the learned use of multimodal information across cities remains an open problem. A further unaddressed challenge is that modalities contain heterogeneous content, and individually beneficial modalities can interfere when used jointly. To this end, we propose MILU, a multimodal urban spatio-temporal foundation model that learns each modality's predictive influence rather than directly mixing modality content. MILU encodes each modality separately, retrieves historical spatio-temporal context conditioned on both the forecasting query and modality information, and composes the resulting influences in a shared influence space to modulate forecasting states under bounded scaling. MILU is jointly pretrained across cities by comparing forecasting errors with and without multimodal information, so that it learns how to use multimodal information under varying availability. This transfer also requires pretraining data with diverse modalities, ample observations, and broad geographic coverage, which existing urban foundation model resources do not yet provide. We therefore construct the CityModal Dataset, covering 87 regions across four continents with approximately 19.21 billion valid spatio-temporal data points, and associate urban observations with road, remote-sensing, weather, and event information. Experiments show that, without target training, MILU achieves the lowest MAE on 28 of 30 short-term and 27 of 30 long-term forecasting tasks, and achieves increasing gains as more modalities are used jointly. With one tenth of the target training data, it outperforms the strongest task-specific baseline trained on the full target set in aggregate performance. Controlled comparisons further show that multimodal utilization transfers to new scenarios, extends to modality combinations unseen during training, and improves with limited target-data adaptation.

1 INTRODUCTION

Modeling urban spatio-temporal dynamics, including traffic and travel demand, is essential for transportation planning and city management. **Urban spatio-temporal foundation models**, including OpenCity (Li et al., 2025), UniST (Yuan et al., 2024), and UrbanFM (Chen et al., 2026a), learn reusable forecasting patterns by pretraining on spatio-temporal observations across cities. However, these observations mainly capture unimodal historical spatiotemporal dynamics, while **multimodal information provides additional clues about environmental conditions, infrastructure, and external disruptions**. Harnessing this information is therefore a natural next step toward more capable and transferable urban spatio-temporal foundation models.

Following this direction, **task-specific multimodal models** exploit such information for individual forecasting tasks. CKG-GNN (Zhang et al., 2025) incorporates roads, points of interest (POIs), and weather for traffic speed forecasting, while E^2-CSTP (Huang et al., 2026) combines spatio-temporal observations with text and imagery. *These methods, however, primarily learn multimodal relationships within individual datasets.* An urban spatio-temporal foundation model must instead learn

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

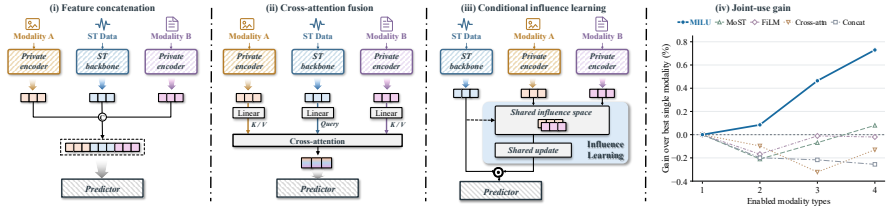

Figure 1: Multimodal fusion paradigms and joint-use gains over the best individual modality.

reusable ways of using heterogeneous information across cities and tasks, enabling transfer without relearning for each city. This leads to the first challenge (**C1**): **learning reusable multimodal spatio-temporal modeling capabilities across cities through pretraining.**

MoST (Xu et al., 2026) takes a step toward this goal by targeting traffic prediction in unseen cities, selecting task-relevant modalities, and using them to guide spatial experts. Yet even when useful information is identified, a fundamental challenge remains: **modalities contain heterogeneous content, and individually beneficial modalities can interfere when used jointly.** Conventional fusion mechanisms provide a simple way to combine them. Concatenation (Fig. 1(i)), exemplified by T3 (Han et al., 2024), directly joins modality representations, whereas cross-attention (Fig. 1(ii)), as in FUSE-Traffic (Yu et al., 2025), uses spatio-temporal features to attend to modality content. As illustrated in Fig. 1(iv), we progressively enable one to four modality types and compare concatenation (Concat), cross-attention (Cross-attn), feature modulation (FiLM) (Perez et al., 2018), MoST, and our MILU. For each modality set, we measure the *relative multimodal gain* with respect to the best individual modality, so that positive values indicate that joint use improves over the strongest single modality. MILU achieves positive and increasing gains as more modalities are enabled, whereas the compared methods do not consistently preserve individual-modality benefits. This observation reveals a second challenge (**C2**): **coordinating individually useful heterogeneous modalities so that their joint use does not undermine their individual benefits.**

Addressing the above two challenges at foundation-model scale requires diverse modalities, sufficient observations, and broad geographic coverage. However, the available public data resources such as BjTT (Zhang et al., 2024a), TraffiDent (Gou et al., 2026), and IUTF (Lin et al., 2025) vary in modality coverage, leaving gaps for multimodal urban spatio-temporal pretraining. Without a modality in the pretraining corpus, the model cannot learn its relationship with urban spatio-temporal dynamics. Moreover, multimodal data volume and geographic coverage remain limited for large-scale cross-city pretraining. This leads to the third challenge (**C3**): **building a pretraining corpus with diverse modalities, ample observations, and broad geographic coverage.**

In this paper, we propose **MILU**, a multimodal urban spatio-temporal foundation model, together with **CityModal**, a large-scale multimodal pretraining corpus. MILU learns *predictive influence* from multimodal information rather than directly mixing heterogeneous modalities. Cross-city pretraining over multiple data sources develops reusable multimodal spatio-temporal capabilities in a shared model. Specifically, *relative multimodal gain supervision* explicitly trains for forecasting improvements over the same model using spatio-temporal observations alone. Staged pretraining first learns multimodal utilization with a frozen pretrained backbone, then jointly refines the multimodal modules and selected backbone parameters. Modality-subset supervision exposes the model to varying information availability during pretraining, supporting forecasting with subsets of modalities.

To reduce joint-use interference, MILU does not directly combine modality content. Instead, it preserves modality-specific representations and retrieves historical spatio-temporal context conditioned on the forecasting query and modality information and forms *predictive influence* representations in a shared influence space. The aggregated influences modulate the forecasting state through bounded channel scaling (Fig. 1(iii)), controlling the adjustment. We further provide a theoretical analysis under a common-response model, showing that shared influence learning can reduce the repeated response-estimation error of concatenation and the aggregation error of fixed-weight content attention. Under suitable conditions on representation error, the resulting expected prediction risk is lower than that of both alternatives. This analysis isolates the statistical benefit of shared influence learning and motivates its use in MILU.

To support large-scale multimodal pretraining, we construct **CityModal**, linking spatio-temporal data across cities, years, and forecasting tasks in Europe, North America, Asia, and Oceania with road, remote-sensing, weather, and event information. This substantially expands modality coverage and geographic diversity for urban spatio-temporal pretraining. CityModal contains approximately 19.21 billion valid spatio-temporal data points from heterogeneous sources, organized into unified

samples, aligned in space and time, and matched with multimodal information. These properties support MILU pretraining and evaluation of multimodal spatio-temporal modeling capabilities.

Overall, this paper makes the following contributions:

- We propose **MILU**, a multimodal urban spatio-temporal foundation model that learns reusable multimodal modeling capabilities across cities through shared pretraining, staged optimization, and relative gain supervision explicitly targeting predictive gains from multimodal information.
- We introduce **multimodal influence learning** in a shared influence space. Rather than directly fusing heterogeneous modality content, MILU preserves modality-specific representations and retrieves historical context conditioned on the forecasting context and modality information to form composable predictive influences. These influences modulate forecasting states through bounded channel scaling, reducing joint-use interference among heterogeneous modalities.
- We construct **CityModal**, to the best of our knowledge, the largest multimodal pretraining corpus for urban spatio-temporal foundation models. It links urban spatio-temporal observations across regions and years with road, remote-sensing, weather, and event information, expanding modality coverage, valid data volume, and geographic diversity for pretraining and evaluation under varying modality availability.
- Cross-city experiments on zero-shot forecasting, few-shot adaptation, multimodal gain transfer, and data and model scaling show that MILU leads on most zero-shot tasks and remains competitive with foundation and task-specific models using one tenth of the target training data.

2 Related Work

2.1 Spatio-Temporal Foundation Models

Pretraining across datasets shifts spatio-temporal modeling from dataset-specific fitting toward reusable forecasting and representations. General time-series foundation models, including TimesFM (Das et al., 2024), Chronos (Ansari et al., 2024), Moirai (Woo et al., 2024), and MOMENT (Goswami et al., 2024), learn transferable patterns across domains. Urban spatio-temporal modeling also requires capturing spatial dependencies and cross-city distribution shifts. OpenCity (Li et al., 2025) and UniST (Yuan et al., 2024) pretrain across cities to learn shared spatio-temporal patterns, while FactoST (Zhong et al., 2026) and STUNet (Chen et al., 2026b) disentangle temporal patterns from spatial representations for transfer across traffic networks. Other studies broaden the structures and tasks of shared models. UniFlow (Yuan et al., 2026a) unifies grid-based and graph-based flow prediction, while UrbanDiT (Yuan et al., 2026b) supports multiple urban tasks, including forecasting and imputation. UrbanFM (Chen et al., 2026a) scales urban spatio-temporal pretraining to larger datasets and more cities. ***These studies primarily learn reusable representations and forecasting from spatio-temporal observations under spatial and temporal distribution shifts. In contrast, MILU extends urban spatio-temporal pretraining to learn to use heterogeneous multimodal information across cities.***

2.2 Multimodal Spatio-Temporal Models

Multimodal spatio-temporal forecasting uses information beyond historical observations to capture environmental conditions, context, and external disruptions. Early methods, including DeepST (Zhang et al., 2016) and ST-ResNet (Zhang et al., 2017), jointly model spatio-temporal dependencies with calendar and weather information for urban crowd-flow prediction. Recent methods add richer semantic and event information. T3 (Han et al., 2024), FUSE-Traffic (Yu et al., 2025), and STORM (Hong et al., 2024) use text or event context to capture traffic and crowd-flow changes that historical observations alone may not explain. ***They primarily learn task- or dataset-specific multimodal relationships rather than pretraining reusable multimodal capabilities across cities.***

Recent work studies how sources interact and aid prediction. MoSSL (Deng et al., 2024) explores multimodal self-supervised learning, while E^2-CSTP (Huang et al., 2026) and CausalX (Zhang et al., 2026) model cross-modal dependencies through causal structures and interventions, respectively. Aurora (Wu et al., 2026) extends time-series foundation modeling to multimodal inputs for zero-shot forecasting conditioned on heterogeneous modalities. ***These studies go beyond spatio-temporal***

histories but do not specifically address heterogeneous modality coordination for transferable urban spatio-temporal forecasting across cities.

MoST (Xu et al., 2026) is particularly relevant to our setting. It predicts traffic in unseen cities by selecting relevant modalities and guiding spatial experts with multimodal context. MILU addresses a complementary problem: *how useful modalities can jointly influence forecasting without mutual interference*. Rather than selecting or directly fusing modality representations, **MILU preserves modality-specific content and maps each modality into an influence in a shared influence space**. Composed influences modulate the forecasting state, while relative gain supervision explicitly targets forecasting gains from multimodal information. Cross-city pretraining makes this multimodal utilization reusable and composable.

MILU complements multimodal spatio-temporal models: while prior methods incorporate, select, align, or causally model multimodal information, MILU learns *transferable predictive influence* and *joint-use coordination* as reusable capabilities of an urban spatio-temporal foundation model.

3 PRELIMINARY

Definition 1 (Urban Spatio-Temporal Data.) It comprises observations varying over time across sensors, stations, regular grid cells, or irregular regions, measuring flow, speed, occupancy, or travel demand. For N spatial units and C observed variables, $X^{\text{hist}} \in \mathbb{R}^{T\times N\times C}$ represents T historical steps, and $Y \in \mathbb{R}^{H\times N\times C}$ represents the next H steps over the same units and variables. Spatial information $\mathcal{S}$ describes unit locations or adjacency relations. Datasets share these axes but differ in spatial-unit counts, structures, and available spatial information.

Problem Formulation. The primary spatio-temporal modality is the modality being forecast, with historical observations X^{hist} and future targets Y. Let $\mathcal{Z}_{\text{avail}}$ denote a sample's multimodal information available before the forecast origin. Modalities retain their data forms; availability varies across samples and datasets. A forecasting model f_θ, with parameters θ, uses these inputs to estimate Y: $\widehat{Y} = f_\theta(X^{\text{hist}}, \mathcal{S}, \mathcal{Z}_{\text{avail}}), \quad \widehat{Y} \in \mathbb{R}^{H\times N\times C}$. Multimodal information provides context without changing target variables. When $\mathcal{Z}_{\text{avail}} = \varnothing$, forecasting uses spatio-temporal inputs alone. Cross-city pretraining learns a shared model across cities and sources for transfer to held-out settings differing in spatial structure, observation distribution, or available multimodal information. Zero-shot forecasting uses pretrained weights; few-shot transfer adapts them on limited target data.

4 METHOD

4.1 OVERALL ARCHITECTURE

As shown in Fig. 2(b), MILU's spatio-temporal backbone produces historical representations and forecasting states, while modality-specific encoders extract multimodal content. A shared influence module uses these representations to form influences and modulate forecasting states before the prediction head.

Spatio-temporal encoding. A Transformer encoder-decoder (Vaswani et al., 2017), shared across datasets, processes each target variable separately. Core and surrounding halo units form groups for local spatial attention with cost independent of dataset size. Normalization aligns dataset and variable scales using visible historical core statistics, unchanged at inference. The backbone and prediction head use token-level masked reconstruction: query tokens hide observations while retaining temporal and spatial information.

Each encoder and decoder block successively applies temporal attention per spatial unit, within-group spatial attention, and a position-wise feed-forward network (FFN). Temporal and spatial positional encodings allow attention to distinguish the relative positions of observations and queries. The encoder maps masked input $\widetilde{X}$ and spatial relations $\mathcal{S}$ to spatio-temporal representations Z. A shared learnable query embedding marks query positions, which the decoder refines through temporal and spatial query attention over encoded context and current query states.

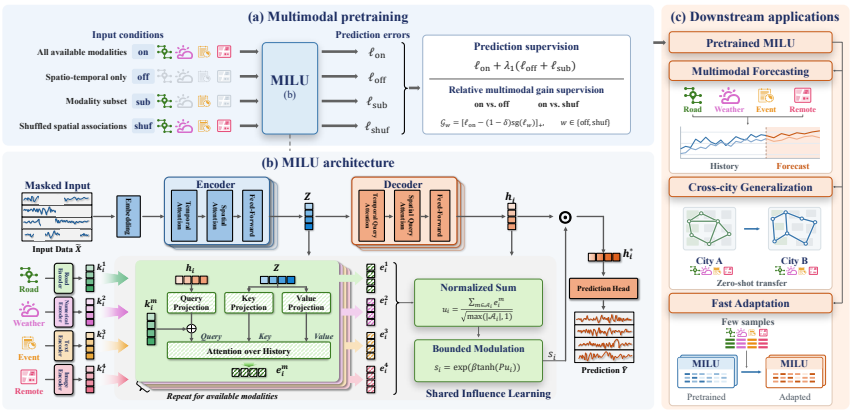

Figure 2: MILU overview. (a) Multimodal pretraining with prediction and relative multimodal gain supervision. (b) Modality-conditioned historical retrieval and shared influence modulation. (c) Multimodal forecasting, zero-shot transfer, and few-shot adaptation.

Forecasting queries span the future window, producing a forecasting state $h_i \in \mathbb{R}^d$ at each position i, with hidden width d. After shared influence modulation using multimodal information, the prediction head $\mathcal{R}$ applies layer normalization and a linear map to produce a normalized scalar forecast.

Multimodal encoding. Each modality-specific encoder E_m maps records $\mathcal{C}_i^m$ at position i to d_c-dimensional content representations. The four modalities comprise road network attributes and local directional structure, remote sensing imagery, numerical weather fields, and event text records. Frozen pretrained encoders process imagery and text; lightweight trainable modules adapt all modality features into content representations. **Architectural details** are provided in **Appendix A**.

4.2 Shared Influence Learning

Existing work uses multimodal information for spatio-temporal forecasting via modality selection or content fusion (Xu et al., 2026; Yu et al., 2025). The predictive role of environmental or event conditions depends on current spatio-temporal dynamics. For example, the same rainfall may accompany different speed changes in free-flowing traffic and worsening congestion. MILU represents each modality by its **prediction-dependent influence on the forecasting state**, rather than directly fusing its content. Modality content and the forecasting state jointly condition historical retrieval at the same spatial unit, forming influences aggregated in the shared influence space.

Modality-conditioned retrieval. Modality-specific layer normalization LN_m and projection P_m map encoded content c_i^m to a modality key $k_i^m = P_m \,\mathrm{LN}_m(c_i^m)$. Each event candidate retains a separate key with scope, time, and distance information relative to i. When multiple candidates are available, retrieval is performed separately for each key.

Forecasting state h_i supplies the query, while historical representations $z_{t,n(i)} = Z_{t,n(i)}$ at the same spatial unit $n(i)$ supply keys and values:

$$q_i = W_q \,\mathrm{LN}_s(h_i), \qquad K_{i,t} = W_K \,\mathrm{LN}_s(z_{t,n(i)}), \qquad V_{i,t} = W_V \,\mathrm{LN}_s(z_{t,n(i)}). \tag{1}$$

The three projections and layer normalization are shared across modalities, with outputs of width r. For a branch with valid content, let $\mathcal{V}_i$ denote the nonempty set of valid visible historical times at this spatial unit. The modality key and forecasting query jointly determine how shared historical values contribute to the influence representation:

$$\alpha_{i,t}^m = \mathrm{softmax}_{t\in\mathcal{V}_i}\left(\frac{(q_i + k_i^m)^\top K_{i,t}}{\sqrt{r}}\right), \qquad e_i^m = \sum_{t\in\mathcal{V}_i} \alpha_{i,t}^m W_V \,\mathrm{LN}_s(z_{t,n(i)}). \tag{2}$$

Query q_i and modality key k_i^m jointly select historical evidence for the current forecast. Unlike cross-attention over modality content, each modality assigns its own weights $\alpha_{i,t}^m$ to shared values $V_{i,t}$ from spatio-temporal history. The shared projection $W_V \in \mathbb{R}^{r\times d}$ places these values and their weighted summaries e_i^m in a common r-dimensional representation space, which we call the shared influence space. With $r < d$, this space provides compact shared coordinates for predictive modu-

lation rather than semantic alignment of modality content. For a modality with multiple candidate records, the forecasting query pools their influences into the modality-level representation e_i^m before cross-modality aggregation.

Influence aggregation and modulation. To let one model use the modality combinations across cities, we aggregate modality influences as a set in the shared influence space. Let $\mathcal{A}_i$ contain enabled modalities with valid content, and let $n_i = |\mathcal{A}_i|$. The aggregate $u_i \in \mathbb{R}^r$ is $u_i = \frac{\sum_{m \in \mathcal{A}_i} e_i^m}{\sqrt{\max(n_i, 1)}}$.

Summation makes aggregation invariant to modality order, while square-root normalization moderates magnitude growth with modality count. While W_V constructs shared influence representations, a shared projection $P \in \mathbb{R}^{d \times r}$ maps their aggregate to the d channels of the forecasting state. Following conditional feature modulation (Perez et al., 2018), we convert the aggregated modality-conditioned historical influences into bounded channel-wise scaling of the forecasting state:

$$s_i = \exp\big(\beta \tanh(P u_i)\big), \qquad h_i^\star = h_i \odot s_i. \tag{3}$$

Here, $\odot$ denotes channel-wise multiplication, and $\exp$ and $\tanh$ act elementwise. Parameter $\beta > 0$ bounds the scaling coefficients within $(e^{-\beta}, e^{\beta})$. The prediction head maps the modulated state to the forecast, $\widehat{y}_i = \mathcal{R}(h_i^\star)$. An empty available set gives $u_i = 0$, $s_i = \mathbf{1}$, and naturally recovers prediction from spatio-temporal observations alone (ST-only).

4.3 Multimodal Pretraining

We pretrain MILU across cities and data sources under varying modality combinations (Fig. 2(a)), explicitly supervising the relative forecasting gains from multimodal information.

For each sample, the same model uses four input conditions with identical historical observations and targets. The on condition supplies all available multimodal information, off uses the ST-only condition, sub uses a nonempty subset of the available modalities, and shuf shuffles the associations between multimodal content and spatial positions.

Prediction supervision. The on, off, and sub conditions receive prediction supervision on valid Core query targets. We describe the MAE terms, computed in sample-normalized observation space to reduce variable-specific scale differences. For condition v, the loss is $\ell_v = E_v / b$, where E_v is its MAE and b is the frozen pretrained spatio-temporal model's MAE on the same sample, with a positive floor. This shared scale measures error relative to baseline forecasting performance while preventing excessive weights near zero baseline error.

Relative Multimodal Gain Supervision. We explicitly supervise the relative error reduction of on over the same model's ST-only prediction (off), making the predictive gain from multimodal information a training objective. Comparing on with shuf further encourages gains from correctly associated content. Each pair uses common valid supervision positions: on/off uses positions with valid multimodal information, and on/shuf further requires the content associations to change. For comparison condition w, either off or shuf, we compute $\mathcal{G}_w = \big[\ell_{\text{on}} - (1-\delta)\,\text{sg}(\ell_w)\big]_+$.

Both errors use the same sample-specific scale b. Here, $[x]_+ = \max(x, 0)$ and sg denotes stop-gradient. This term encourages a relative error reduction of at least $\delta \in (0, 1)$ for on. Stopping gradients through the comparison error directs gain supervision through the on prediction, while off retains its own prediction supervision. The resulting objective is

$$\mathcal{L} = \ell_{\text{on}} + \lambda_1(\ell_{\text{off}} + \ell_{\text{sub}}) + \lambda_2(\mathcal{G}_{\text{off}} + \mathcal{G}_{\text{shuf}}). \tag{4}$$

Weights λ_1 and λ_2 control prediction supervision on off and sub and relative gain supervision, respectively. Training averages the loss equally over valid samples.

Staged optimization. Starting from the pretrained spatio-temporal backbone and prediction head, we freeze both to train the trainable parts of the modality encoders and the shared influence module. We then include the spatio-temporal encoder, decoder, and prediction head in joint optimization.

4.4 Theoretical Analysis of Shared Influences

Under a common-response regime, our linear analysis explains how sharing predictive responses can improve statistical efficiency over estimating modality-specific responses separately, motivating

MILU's representation-sharing principle. The **formal setup**, Theorem 1, Corollary 1, and **proofs** are in **Appendix D**.

5 CITYMODAL DATASET CONSTRUCTION

Construction. Across regions and years, we standardize spatial identifiers and variable units and align timestamps to a common interval. We filter invalid values, fill short gaps in flow, speed, and occupancy sequences, and match observations with road, remote-sensing, weather, and event information by location and time window. A common format stores spatio-temporal observations, matched multimodal information, locations, and modality-availability indicators for shared pretraining and modality-conditioned retrieval. **Sources, processing, and counting rules**: **Appendix B**.

Dataset comparison. Table 1 compares CityModal with BjTT (Zhang et al., 2024a), TraffiDent (Gou et al., 2026), IUTF (Lin et al., 2025), and WorldST, the pretraining corpus of UrbanFM (Chen et al., 2026a). CityModal combines multi-year traffic records, diverse urban environments, and four types of multimodal information to support learning shared predictive influences across cities.

Table 1: Comparison of urban spatio-temporal datasets.

Dataset	Region	Year	Node	Valid points (M)	Modalities
BjTT	1	2022	1,260	81.65	Event, Road
TraffiDent	42	2022–2024	16,972	14,667.65	Event, Road
IUTF	38	2008–2018	17,188	176.30	Road, Weather
WorldST	39	2008–2018	15,732	1,171.12	–
CityModal	87	2008–2024	46,285	19,205.01	Event, Road, Weather, Remote

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

Datasets. We evaluate on 19 datasets: eight subsets of CityModal and 11 public datasets, namely Bike-CHI and Taxi-CHI (Fang et al., 2024); METR-LA and PEMS-BAY (Li et al., 2017); PEMS03, PEMS04, PEMS07, and PEMS08 (Song et al., 2020); BjTT (Zhang et al., 2024a); Taxi-BJ (Zhang et al., 2017); and Traffic-SH (Li et al., 2025). The 30 forecasting tasks comprise 19 with multimodal information and 11 with spatio-temporal observations alone. Pretraining uses CityModal supplemented with public datasets.

Baselines. We compare against spatio-temporal foundation models MoST (Xu et al., 2026), UrbanFM (Chen et al., 2026a), and OpenCity (Li et al., 2025); time-series foundation models Chronos-2 (Ansari et al., 2025), TimesFM (Das et al., 2024), and Moirai (Liu et al., 2025); task-specific models E^2-CSTP (Huang et al., 2026), T3 (Han et al., 2024), STAEformer (Liu et al., 2023a), STID (Shao et al., 2022), and PatchSTG (Fang et al., 2025); and Historical Average (HA).

At five-minute intervals, Short and Long forecasting predict 12 and 24 future steps from equally long histories, respectively. GeoMAE and GeoRMSE are the geometric means of task-level MAE and RMSE, denoted as MAE and RMSE below; lower values are better. **Data, training, evaluation, and baseline settings and hyperparameter analysis** are provided in **Appendix C**.

6.2 CROSS-SCENARIO FORECASTING AND TARGET-DATA EFFICIENCY

Tables 2 and 3 compare zero-shot forecasting and few-shot adaptation across scenarios. Single, Multi, and Overall cover 11 ST-only tasks, 19 tasks with multimodal information, and all 30 tasks, respectively; S/L denote Short/Long forecasting. Each method uses its supported inputs. Best and second-best results are bold and underlined, respectively.

Zero-shot forecasting. Table 2 compares foundation models without target-task parameter updates. MILU achieves the lowest aggregate MAE and RMSE in both task groups at both horizons. Relative to the strongest baseline for each overall metric, TimesFM for MAE and UrbanFM for RMSE, MILU reduces MAE/RMSE by 12.2%/9.4% for Short and 16.5%/11.8% for Long. These results demonstrate cross-scenario forecasting across input conditions and horizons.

Table 2: Zero-shot forecasting performance of foundation models.

Method	Single-S		Single-L		Multi-S		Multi-L		Overall-S		Overall-L	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MoST	5.094	9.142	6.428	11.417	10.381	18.028	13.531	23.379	7.996	14.054	10.299	17.976
UrbanFM	3.902	7.467	5.069	9.646	8.857	16.358	11.375	20.828	6.558	12.270	8.458	15.706
OpenCity	6.069	11.135	7.113	13.095	12.175	21.748	14.046	24.949	9.432	17.015	10.945	19.697
Chronos	3.702	7.902	4.827	10.304	8.736	17.273	11.162	21.716	6.377	12.967	8.208	16.522
TimesFM	3.606	7.675	4.811	10.227	8.545	17.000	11.056	21.601	6.228	12.700	8.149	16.421
Moirai	3.903	8.148	5.266	10.904	8.896	17.084	11.966	22.309	6.576	13.022	8.856	17.159
MILU	**3.186**	**6.703**	**3.991**	**8.501**	**7.478**	**14.909**	**9.262**	**18.373**	**5.469**	**11.121**	**6.802**	**13.850**

Few-shot adaptation and target-data efficiency. In Table 3, foundation models use 10% of target training data and task-specific models and HA use 100%, all evaluated on identical test samples. Compared with UrbanFM under the same data budget, MILU reduces overall MAE/RMSE by 21.2%/16.8% for Short and 25.6%/21.5% for Long. MILU also outperforms PatchSTG, the strongest fully supervised task-specific baseline in overall errors, by 1.4%/1.3% for Short and 0.9%/0.9% for Long. MILU thus achieves lower overall forecasting errors with one tenth of the target training data. **Task-level results** are in **Appendix C.3**.

Table 3: Few-shot foundation models versus fully supervised baselines.

Method	Single-S		Single-L		Multi-S		Multi-L		Overall-S		Overall-L	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
HA	4.251	8.806	4.248	8.801	11.340	21.272	11.747	22.044	7.914	15.395	8.090	15.742
E^2-CSTP	2.485	5.215	2.850	5.865	7.359	13.914	8.247	15.341	4.943	9.709	5.586	10.783
T3	**2.086**	4.861	**2.373**	5.376	7.278	13.808	8.054	15.740	4.603	9.416	5.145	10.615
STAEformer	2.205	5.058	2.450	5.564	7.080	13.762	7.777	15.343	4.616	9.535	5.091	10.578
STID	2.197	4.954	2.595	5.691	6.751	13.473	7.529	14.913	4.473	9.335	5.095	10.475
PatchSTG	2.097	4.843	2.383	5.344	6.811	13.476	7.491	14.836	4.422	9.260	4.922	10.202
MoST	3.151	5.888	3.720	7.002	8.816	15.660	10.518	18.387	6.046	10.940	7.185	12.905
UrbanFM	3.062	6.480	3.657	7.596	7.790	14.912	9.204	17.497	5.532	10.985	6.562	12.885
OpenCity	3.151	6.709	3.620	7.794	8.792	16.320	9.895	18.256	6.035	11.780	6.844	13.361
Chronos	3.424	7.704	4.121	9.603	8.380	17.085	9.993	20.769	6.036	12.758	7.222	15.652
TimesFM	3.317	7.387	4.029	9.152	8.217	16.578	9.956	20.026	5.892	12.325	7.146	15.028
Moirai	3.112	7.078	3.751	8.404	7.973	15.996	9.592	18.994	5.647	11.863	6.798	14.086
MILU	2.131	**4.828**	2.391	**5.343**	**6.600**	**13.218**	**7.374**	**14.630**	**4.360**	**9.137**	**4.879**	**10.112**

6.3 Transfer of Multimodal Utilization Across Scenarios

On 19 held-out tasks, we compare all available multimodal information (on) with ST-only (off), using identical weights, histories, and test targets to isolate the benefit of multimodal information. We report relative error reductions from off to on with pretrained weights and weights adapted using 10% of target training data.

Results. Multimodal utilization transfers to new scenarios at both horizons, with target adaptation increasing overall MAE gains from 3.362% to 7.427% for Short and from 3.999% to 8.092% for Long. All four prediction-variable groups benefit in MAE under both zero-shot and few-shot evaluation; Figure 3 shows Long results. **Full results** are in **Appendix C.7**.

6.4 Generalization to Unseen Modality Combinations

We test whether learned modality influences remain useful under new combinations on the 19 held-out multimodal tasks. A, B, and C denote Road, Remote, and Weather. With the main experiment's data splits, we train on A+C and B+C without jointly enabling A and B, then test unseen A+B and A+B+C with fixed weights.

Results. The unseen Road+Remote pair yields lower MAE than either modality alone; adding Weather gives the lowest MAE across evaluated combinations. Both findings hold at both horizons (Long in Figure 4), showing that learned modality influences remain useful in unseen combinations. **Individual-modality contributions, full protocols, and results** are in **Appendix C.6**.

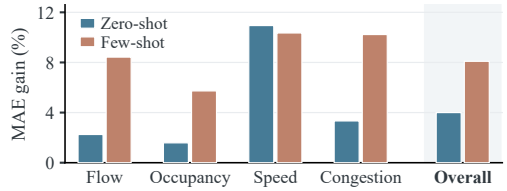

Figure 3: Multimodal transfer gains.



Figure 4: Unseen modality combinations.

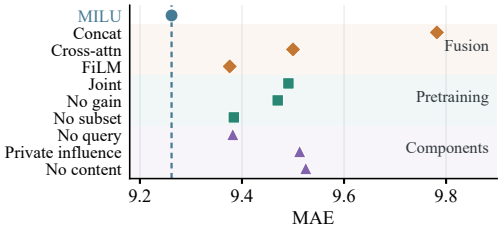

Figure 5: MILU design ablations.

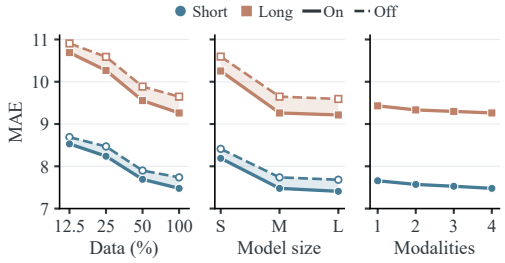

Figure 6: Data, modality, and model scaling.

6.5 Fusion, Pretraining, and Component Ablations

Figure 5 compares nine variants on zero-shot Long MAE across the 19 multimodal tasks.

Fusion. With identical multimodal inputs, MILU's shared influence pathway outperforms concatenation (Concat), cross-attention (Cross-attn), and feature modulation (FiLM), supporting the benefit of learning predictive influence rather than directly fusing modality representations.

Pretraining. Joint, No gain, and No subset remove the initial frozen-backbone phase, relative gain supervision, and modality-subset supervision, respectively. Each removal increases forecasting error, indicating that staged pretraining and supervision of multimodal gains and modality availability all contribute to performance.

Components. No query, Private influence, and No content ablate the forecasting query in historical retrieval, the shared influence projections, and modality content, respectively. The results support a shared influence space and response, with historical retrieval conditioned on both modality content and forecasting context. **Variant definitions and full results** are provided in **Appendix C.5**.

6.6 Data, Modality, and Model Scaling

Figure 6 evaluates zero-shot data, modality, and model scaling on the 19 multimodal tasks. Data and model scaling fix the number of training updates, while modality scaling fixes the pretrained weights. Solid and dashed curves denote multimodal information on and off, respectively.

Data scaling. Increasing pretraining data from 12.5% to 100% reduces MAE by 12.334% for Short and 13.353% for Long, with a larger multimodal gain at the latter endpoint.

Modality scaling. Increasing the number of modalities from one to four reduces MAE from 7.658 to 7.478 for Short and from 9.430 to 9.262 for Long, with improvements at every evaluated step.

Model scaling. Scaling from S to M and L reduces MAE at both horizons while preserving the multimodal gain. The largest improvement occurs from S to M, which motivates M as the default configuration. **Full settings and results** are provided in **Appendix C.4**.

Model efficiency. MILU is equipped with 2.311M forecasting parameters. Enabling multimodal information reduces MAE by 4.0% with an additional 7.44ms of inference latency per batch. **Parameter count, latency, and memory usage** are compared in **Appendix C.9**.

7 Conclusion

MILU organizes heterogeneous modality content into a shared influence space through history-conditioned retrieval and modulates spatio-temporal states via bounded channel scaling. Staged pretraining uses relative gain and modality-subset supervision to support varying information conditions. Across 30 tasks at each horizon, MILU outperforms the strongest baseline for each metric in both zero-shot and few-shot settings, with larger multimodal gains after target adaptation. Further scaling of pretraining data and model capacity may improve zero-shot transfer.

AI USE STATEMENT

We used generative AI tools to assist with literature search and synthesis, code implementation of our method, and language polishing. The authors reviewed and tested the AI-assisted implementation, checked citations against the original sources, and verified figures and reported results against experimental records. All AI-assisted text was reviewed and revised by the authors. The authors take responsibility for the final text, claims, citations, and artifacts.

REPRODUCIBILITY STATEMENT

Appendix A details the model architecture and pretraining procedure, **Appendix B** describes dataset construction and preprocessing, and **Appendix C** specifies the experimental protocols and training settings. Code is available at `https://anonymous.4open.science/r/MILU-01FD`. The CityModal dataset constructed in this work will be released publicly upon acceptance.

ETHICS STATEMENT

CityModal will be released in accordance with the licenses and access restrictions of its data sources. For assets that cannot be redistributed, we will provide acquisition instructions. We have applied privacy protection measures during dataset construction.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

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

Wei Chen, Yuqian Wu, Junle Chen, Xiaofang Zhou, and Yuxuan Liang. UrbanFM: Scaling urban spatio-temporal foundation models. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 10635–10645, 2026a.

Yujun Chen, Shihao Tu, Wenyue Ding, Yicheng Lu, Qingkai Ren, Yangjie Zheng, and Yang Yang. Unified spatio-temporal tokens are bases for generalizable traffic forecasting. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 508–519, 2026b. doi: 10.1145/3770855.3817746. URL `https://doi.org/10.1145/3770855.3817746`.

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

Jiewen Deng, Renhe Jiang, Jiaqi Zhang, and Xuan Song. Multi-modality spatio-temporal forecasting via self-supervised learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 2018–2026, 2024. doi: 10.24963/ijcai.2024/223. URL `https://www.ijcai.org/proceedings/2024/223`.

Jiangyi Fang, Liyue Chen, Di Chai, Yayao Hong, Xiuhuai Xie, Longbiao Chen, and Leye Wang. UCTB: An urban computing tool box for building spatiotemporal prediction services. In *2024 IEEE International Conference on Software Services Engineering (SSE)*, pp. 54–65. IEEE, 2024.

Yuchen Fang, Yuxuan Liang, Bo Hui, Zezhi Shao, Liwei Deng, Xu Liu, Xinke Jiang, and Kai Zheng. Efficient large-scale traffic forecasting with transformers: A spatial data management perspective. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 1*, pp. 307–317, 2025.

Philip J. Fleming and John J. Wallace. How not to lie with statistics: The correct way to summarize benchmark results. *Communications of the ACM*, 29(3):218–221, 1986. doi: 10.1145/5666.5673.

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

Xiaochuan Gou, Ziyue Li, Tian Lan, Junpeng Lin, Bingyu Zhao, Chen Zhang, Di Wang, Xiangliang Zhang, et al. TraffiDent: A dataset for understanding the interplay between traffic dynamics and incidents. *Advances in Neural Information Processing Systems*, 38, 2026.

Xiao Han, Zhenduo Zhang, Yiling Wu, Xinfeng Zhang, and Zhe Wu. Event traffic forecasting with sparse multimodal data. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8855–8864, 2024. doi: 10.1145/3664647.3680706. URL https://doi.org/10.1145/3664647.3680706.

Yayao Hong, Hang Zhu, Tieqi Shou, Zeyu Wang, Liyue Chen, Leye Wang, Cheng Wang, and Longbiao Chen. STORM: A spatio-temporal context-aware model for predicting event-triggered abnormal crowd traffic. *IEEE Transactions on Intelligent Transportation Systems*, 25(10):13051–13066, 2024. doi: 10.1109/TITS.2024.3390185. URL https://doi.org/10.1109/TITS.2024.3390185.

Yuting Huang, Ziquan Fang, Zhihao Zeng, Lu Chen, and Yunjun Gao. Causal spatio-temporal prediction: an effective and efficient multi-modal approach. *Advances in Neural Information Processing Systems*, 38:125786–125812, 2026.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

Zhonghang Li, Long Xia, Lei Shi, Yong Xu, Dawei Yin, and Chao Huang. Open spatio-temporal foundation models for traffic prediction. *ACM Transactions on Intelligent Systems and Technology*, 2025.

Xuhui Lin, Qiuchen Lu, Long Chen, Jack Chin Pang Cheng, Jiayi Yan, Jingke Hong, and Pengjun Zhao. IUTF dataset: Enabling cross-border resource for analysing the impact of rainfall on urban transportation. *Scientific Data*, 13(1):22, 2025.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025.

Hangchen Liu, Zheng Dong, Renhe Jiang, Jiewen Deng, Jinliang Deng, Quanjun Chen, and Xuan Song. Spatio-temporal adaptive embedding makes vanilla transformer SOTA for traffic forecasting. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4125–4129, 2023a. doi: 10.1145/3583780.3615160. URL https://doi.org/10.1145/3583780.3615160.

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. LargeST: A benchmark dataset for large-scale traffic forecasting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 75354–75371, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ee57cd73a76bd927ffca3dda1dc3b9d4-Abstract-Datasets_and_Benchmarks.html.

Allister Loder, Lukas Ambühl, Monica Menendez, and Kay W Axhausen. Understanding traffic capacity of urban networks. *Scientific reports*, 9(1):16283, 2019.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016. URL https://www.jmlr.org/papers/v17/15-242.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 4454–4458, 2022.

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 914–921, 2020. doi: 10.1609/aaai.v34i01.5438. URL https://ojs.aaai.org/index.php/AAAI/article/view/5438.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Chenjuan Guo. Aurora: Towards universal generative multimodal time series forecasting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VVJ6Ck9JBl.

Ronghui Xu, Jihao Chen, Jindong Tian, Chenjuan Guo, and Bin Yang. MoST: A foundation model for multi-modality spatio-temporal traffic prediction. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 1716–1727, 2026. doi: 10.1145/3770854.3780162. URL https://dl.acm.org/doi/10.1145/3770854.3780162.

Chenyang Yu, Xinpeng Xie, Yan Huang, and Chenxi Qiu. FUSE-Traffic: Fusion of unstructured and structured data for event-aware traffic forecasting. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, pp. 620–630, 2025. doi: 10.1145/3748636.3762776. URL https://doi.org/10.1145/3748636.3762776.

Yuan Yuan, Jingtao Ding, Jie Feng, Depeng Jin, and Yong Li. UniST: A prompt-empowered universal model for urban spatio-temporal prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4095–4106, 2024. doi: 10.1145/3637528.3671662. URL https://doi.org/10.1145/3637528.3671662.

Yuan Yuan, Jingtao Ding, Chonghua Han, Zhi Sheng, Depeng Jin, and Yong Li. UniFlow: A foundation model for unified urban spatio-temporal flow prediction. *IEEE Transactions on Mobile Computing*, 25(8):13034–13046, 2026a. doi: 10.1109/TMC.2026.3673205. URL https://doi.org/10.1109/TMC.2026.3673205.

Yuan Yuan, Chonghua Han, Jingtao Ding, Guozhen Zhang, Depeng Jin, and Yong Li. Diffusion transformers as open-world spatiotemporal foundation models. *Advances in Neural Information Processing Systems*, 38:30518–30545, 2026b.

Chengyang Zhang, Yong Zhang, Qitan Shao, Jiangtao Feng, Bo Li, Yisheng Lv, Xinglin Piao, and Baocai Yin. BjTT: A large-scale multimodal dataset for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):18992–19003, 2024a. doi: 10.1109/TITS.2024.3440650. URL https://doi.org/10.1109/TITS.2024.3440650.

Junbo Zhang, Yu Zheng, Dekang Qi, Ruiyuan Li, and Xiuwen Yi. DNN-based prediction model for spatio-temporal data. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 1–4, 2016. doi: 10.1145/2996913.2997016. URL https://doi.org/10.1145/2996913.2997016.

Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. doi: 10.1609/aaai.v31i1.10735. URL https://ojs.aaai.org/index.php/AAAI/article/view/10735.

Shiqi Zhang, Pan Mu, Hanting Yan, Yuchao Zhu, Cong Bai, et al. CausalX: A unified and causally-interpretable plug-and-play model for multi-modal spatio-temporal forecasting. In *Forty-third International Conference on Machine Learning*, 2026.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 conference on empirical methods in natural language processing: industry track*, pp. 1393–1412, 2024b.

Yatao Zhang, Yi Wang, Song Gao, and Martin Raubal. Context-aware knowledge graph framework for traffic speed forecasting using graph neural network. *IEEE Transactions on Intelligent Transportation Systems*, 26(3):3885–3902, 2025. doi: 10.1109/TITS.2024.3520511. URL https://ieeexplore.ieee.org/document/10819256/.

Siru Zhong, Junjie Qiu, Yangyu Wu, Xingchen Zou, Zhongwen Rao, Bin Yang, Chenjuan Guo, Hao Xu, and Yuxuan Liang. Learning to factorize spatio-temporal foundation models. *Advances in Neural Information Processing Systems*, 38:88232–88260, 2026.

A Supporting Backbone and Training Details

A.1 Inputs, normalization, and spatio-temporal backbone

To control spatial attention costs across datasets of different sizes, we partition spatial units by proximity into disjoint core groups (Core) of at most 16 units. Since partitioning cuts some neighborhood connections across group boundaries, we augment each Core with nearby external units, forming a buffer (Halo) that restores the lost spatial context. Core and Halo units jointly participate in within-group attention, while reconstruction and forecasting losses are evaluated only at queried Core positions with valid targets. Halo units may be shared across groups. Each group accommodates up to 32 spatial units, with a standard full group containing 16 Core and 16 Halo units. All groups share the same backbone, keeping the parameter count independent of dataset size.

The M model predicts 24 future steps from 24 historical steps at five-minute intervals, using 32 spatial positions per group. Each time-location pair is represented by a token with four channels encoding normalized observations, time of day, day of week, and observation visibility. Calendar features are normalized by 1,440 minutes and 6, respectively.

Each sample contains one target variable over one spatial group and time window. Let $\mathcal{V}$ contain genuinely visible historical observations at Core units. To align numerical scales across datasets and variables, we compute the normalization statistics

$$\mu = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} x_i, \qquad \sigma = \sqrt{\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}(x_i-\mu)^2} + 10^{-5}. \tag{5}$$

Observations and targets are transformed by $(x-\mu)/\sigma$, with future observation channels set to zero. Predictions return to original units through $\sigma\widehat{y}+\mu$. Statistics use valid historical Core observations, excluding Halo context, padding, and masked values. Since these statistics depend only on the observed input, the same normalization applies to new datasets without parameter updates. Requests determine forecasting positions; target validity is used only for supervision. For variables aligned by interpolation, inputs use native sampling anchors, while evaluation retains interpolated targets.

Starting from $Z^0 = \mathrm{Embed}(\widetilde{X})$, encoder block ℓ applies temporal attention, spatial attention, and a feed-forward transformation:

$$Z^\ell = \mathcal{F}_\ell(\mathcal{S}_\ell(\mathcal{T}_\ell(Z^{\ell-1}))), \qquad \ell = 1,\dots,N_{\mathrm{enc}}. \tag{6}$$

Operators $\mathcal{T}_\ell, \mathcal{S}_\ell, \mathcal{F}_\ell$ denote temporal attention, spatial attention, and feed-forward sublayers, respectively, each using layer normalization, residual connections, and padding masks. The M model has $d = 128, N_{\mathrm{enc}} = 6$, eight attention heads, and feed-forward width 512. Temporal rotary positions use a 600-second scale; two-dimensional spatial rotary positions use a 2,000-meter scale. Spatial attention also incorporates topology biases derived from spatial relations. The encoded output is $Z = Z^{N_{\mathrm{enc}}}$.

The decoder marks forecasting positions with a learned query embedding e_{Q} and applies temporal query attention, spatial query attention, and feed-forward updates:

$$\begin{aligned} D^0 &= Z + M_{\mathrm{Q}} \odot e_{\mathrm{Q}}, \qquad D^\ell = \mathcal{D}_\ell(D^{\ell-1}, Z), \\ h_i &= [Z + M_{\mathrm{Q}} \odot W_{\mathrm{D}} D^{N_{\mathrm{dec}}}]_i, \qquad \mathcal{R}(h) = W_o\,\mathrm{LN}_o(h). \end{aligned} \tag{7}$$

Here, M_{Q} marks forecasting positions, $\mathcal{D}_\ell$ is a decoder block, and W_{D}, W_o are affine maps of widths $d \to d$ and $d \to 1$. Each query-attention sublayer uses current forecasting states as queries, with keys and values formed from current states at forecasting positions and encoded representations elsewhere. Future positions are updated jointly without an autoregressive temporal mask. Feed-forward transformations act on valid positions. The M model uses one decoder block, influence dimension $r = 32$, and modulation bound $\beta = 0.2$. The shared influence module selects history from genuinely visible, non-query, non-padding positions at the same spatial unit.

A.2 MODALITY ENCODERS AND EVENT-CANDIDATE RETRIEVAL

Each modality-specific encoder E_m maps records $\mathcal{C}_i^m$ associated with position i into content representations of width $d_c = 128$. Modality-specific layer normalization and bias-free projections map content to 32-dimensional keys. Road, Remote, and Weather produce spatial-unit summaries, while Event retains a content representation for each candidate.

Road jointly encodes road attributes, neighborhood statistics at three scales, and directional structure. Continuous attributes use numerical projections, categorical attributes use embeddings, and field identifiers guide aggregation. Neighborhood statistics use field and scale embeddings with attention pooling, while a local semantic map supplies directional structure. A multilayer perceptron combines these three representations and their validity indicators into road content.

Remote uses frozen DINOv2 ViT-B/14 (Oquab et al., 2023) to encode a 1 km square image centered on each spatial unit, retaining one CLS representation and a 4×4 grid of spatial patch representations. Layer normalization, a 128-dimensional projection, and two-dimensional sinusoidal positions transform the 768-dimensional visual features. A learned query then pools them into a content summary.

Weather jointly encodes 12 numerical fields and their validity indicators from the hour containing the last historical timestamp. Missing fields are set to zero before layer normalization and a $12 \to 128 \to 128$ GELU network, and a bias-free linear projection adds their validity information. When all fields are missing, the branch outputs zero and is excluded from influence composition. Otherwise, its hourly content representation serves all future forecasting positions.

Event uses the English gte-base-en-v1.5 model of the frozen GTE encoder (Zhang et al., 2024b), processing at most 512 text tokens and retaining a 768-dimensional CLS representation. Each forecasting position has up to eight candidates. Layer normalization and a 128-dimensional projection produce text content c_j. Candidate j combines content with its relation to forecasting position i to form

$$k_{ij} = P_E \,\mathrm{LN}_E(c_j) + W_{\mathrm{scope}}[l_{ij}; g_{ij}] + f_{\mathrm{rel}}(r_{ij}). \tag{8}$$

Here, P_E, LN_E are Event-specific projection and layer normalization, and $[\cdot;\cdot]$ denotes concatenation. Indicators $l_{ij}, g_{ij} \in \{0, 1\}$ mark local and regional scope matches, which W_{scope} maps to 32 dimensions. Relation vector r_{ij} contains log-scaled event age (forecasting time minus record availability time) and geographic distance with their validity indicators. A two-layer SiLU network f_{rel} maps these relations into the modality-key space.

Let $\mathcal{E}_i$ contain candidates with valid text and scope matches. Each candidate reads history with $q_i + k_{ij}$ using Section 4.2, producing $e_{ij} \in \mathbb{R}^r$. The forecasting query then pools these influences:

$$\omega_{ij} = \frac{\exp(q_i^\top k_{ij}/\sqrt{r})}{\sum_{j' \in \mathcal{E}_i} \exp(q_i^\top k_{ij'}/\sqrt{r})}, \qquad e_i^{\mathrm{Event}} = \sum_{j \in \mathcal{E}_i} \omega_{ij} e_{ij}. \tag{9}$$

Candidate-specific retrieval preserves the dynamics associated with different events, while query-conditioned pooling selects influences relevant to the current forecast. The branch does not enter composition when candidates or visible history are absent.

A.3 TRAINING OBJECTIVE

The four input conditions follow Section 4.3. We detail the MAE component here. Let $\mathcal{Q}$ contain the valid Core query targets of one sample, p_v denote predictions from the current model under condition v, and p_0 denote predictions from the frozen pretrained spatio-temporal model. For a nonempty position set A, define

$$E(p; A) = \frac{1}{|A|} \sum_{i \in A} |p_i - y_i|, \qquad b = \max(E(p_0; \mathcal{Q}), 10^{-3}). \tag{10}$$

Errors are measured in sample-normalized observation units. The scale b is shared across all input conditions and comparison sets for the sample and receives no gradients. Define $\ell_v(A) = E(p_v; A)/b$, using $\mathcal{Q}$ when A is omitted. For $w = \text{off}$, let A_w contain supervised positions with valid multimodal information; for $w = \text{shuf}$, it additionally requires the content associations to change. The relative multimodal gain term is

$$\mathcal{G}_w = \left[\ell_{\text{on}}(A_w) - (1-\delta)\,\text{sg}(\ell_w(A_w))\right]_+. \tag{11}$$

Here, $[x]_+ = \max(x, 0)$ and sg denotes stop-gradient. The off and shuf errors are detached within the gain term; off retains its own prediction supervision. When at least two modality types are enabled, sub samples a nonempty proper subset; otherwise its additional prediction loss is zero. An empty comparison set contributes zero to $\mathcal{G}_w$. With $\lambda_1 = 0.2$, $\lambda_2 = 1$, and $\delta = 0.01$, the MAE component is

$$\mathcal{L} = \ell_{\text{on}} + 0.2(\ell_{\text{off}} + \ell_{\text{sub}}) + \mathcal{G}_{\text{off}} + \mathcal{G}_{\text{shuf}}. \tag{12}$$

Training additionally includes the analogous MSE prediction and gain terms with weight 0.2, normalized by the frozen model's sample-wise MSE with a floor of 10^{-5}, with $(1-\delta)^2$ replacing $(1-\delta)$ in separate MSE hinges. Errors are averaged over supervised positions within each sample, and the combined objective is averaged equally over valid samples.

A.4 Pretraining procedure

We pretrain the backbone and prediction head on spatio-temporal observations through token-level masked reconstruction. Random-point, random-node, random-time-slice, and future-window masks hide individual tokens, sequences at selected spatial units, entire time slices, and the future window, respectively. Reconstructing these values from the remaining observations encourages the model to learn both temporal evolution and spatial dependencies. Training minimizes mean absolute error over valid prediction positions. Multimodal pretraining consistently predicts the next 24 steps from the previous 24 using the objective above.

Multimodal training first freezes the backbone and prediction head to learn private encoders and the shared influence module, then jointly updates the multimodal modules, final three encoder blocks, query-decoder blocks, and prediction head. The input projection, first three encoder blocks, and query embedding remain frozen. A separate fixed initialization model provides reference error scales.

B Data Construction and Availability

B.1 Spatial subset inventory and coverage statistics

CityModal standardizes heterogeneous spatio-temporal observations and associates them with available road, remote-sensing, weather, and event information through spatial and temporal alignment. The inventory below covers training, validation, and testing and comprises 87 geographic subsets: seven regional partitions, 42 counties, and 38 cities. Different periods and variables within each subset are counted together.

LargeST spatial subsets. LargeST (Liu et al., 2023b)[1] supplies the observations for these subsets. GLA denotes Greater Los Angeles, GBA the Greater Bay Area, and SD San Diego. Remaining constructed nodes are listed by their native PeMS districts. The seven disjoint subsets in Table 4 contain 8,600 sensors. Each subset provides flow observations and road, remote-sensing, weather, and event information.

TraffiDent spatial subsets. TraffiDent (Gou et al., 2026)[2] supplies the observations for these subsets. All 16,961 nodes with valid coordinates match the original metadata through station IDs. The County field covers all nodes across 42 counties, providing the partition summarized in Table 5.

[1] https://github.com/liuxu77/LargeST

[2] https://www.kaggle.com/datasets/gpxlcj/xtraffic

Table 4: LargeST regional partitions in CityModal.

Spatial subset	PeMS districts	Sensors
GLA	7, 8, 12	3,834
GBA	4	2,352
SD	11	716
PeMS District 3	3	480
PeMS District 5	5	211
PeMS District 6	6	484
PeMS District 10	10	523

Each subset provides flow, speed, and occupancy, with road, remote-sensing, weather, and event information.

Table 5: TraffiDent county partitions in CityModal.

County subset	Nodes	County subset	Nodes
Alameda	897	Placer	213
Alpine	8	Riverside	825
Amador	32	Sacramento	835
Butte	42	San Benito	26
Calaveras	16	San Bernardino	893
Colusa	10	San Francisco	76
Contra Costa	773	San Joaquin	687
El Dorado	141	San Luis Obispo	158
Fresno	334	San Mateo	400
Glenn	22	Santa Barbara	132
Kern	177	Santa Clara	992
Kings	27	Santa Cruz	95
Los Angeles	4,575	Solano	375
Madera	60	Sonoma	312
Marin	212	Stanislaus	178
Mariposa	6	Sutter	48
Merced	240	Tulare	72
Monterey	166	Tuolumne	34
Napa	22	Ventura	277
Nevada	50	Yolo	95
Orange	2,409	Yuba	19

UTD19 spatial subsets. After data filtering and processing, the UTD19 (Loder et al., 2019)[3] component of CityModal covers the following 38 cities and 20,724 detectors, including all retained cities used for training, validation, and testing. F, O, and S denote flow, occupancy, and speed, respectively. Table 6 lists variables present in each city; individual observations retain per-variable validity indicators. Each city is associated with road, remote-sensing, and weather information, with no event modality.

Spatio-temporal observation counts enumerate valid time, spatial-unit, and variable records after processing. Availability is recorded for each spatial unit for roads, each unit-year pair for remote sensing, and each unit-hour-field combination for weather. Events retain source-specific identifiers, spatial scopes, and validity intervals; event records and their spatial matches are distinguished.

The sum across separately counted sources is 19,205,010,258 available scalar records. Exact source-level counts are 4,406,085,187 for LargeST from 2017-01-01 to 2021-12-31, 14,658,826,054 for TraffiDent from 2022-01-01 to 2024-12-31, and 140,099,017 for UTD19 from 2008-05-16 to 2018-02-27. Dates denote the earliest and latest local dates with valid records. Toulouse contributes

[3] https://utd19.ethz.ch/

Table 6: UTD19 city statistics in CityModal.

City subset	Detectors	Variables
Augsburg	713	F/O
Basel	77	F/O
Bern	755	F/O
Birmingham	52	F/S
Bolton	166	F/O/S
Bordeaux	425	F/O
Bremen	548	F/O
Cagliari	122	F/O
Constance	122	F/O/S
Darmstadt	392	F/O
Essen	36	F/O/S
Frankfurt	112	F/O
Graz	300	F/O
Groningen	55	F/O/S
Hamburg	418	F/O
Innsbruck	16	F/S
Kassel	505	F/O
London	5,719	F/O
Los Angeles	1,722	F/O
Luzern	158	F/O
Madrid	1,116	F/O
Manchester	181	F/O/S
Marseille	169	F/O
Melbourne	954	F
Munich	520	F/O
Rotterdam	259	F/O/S
Santander	219	F/O
Speyer	184	F/O
Strasbourg	142	F/O
Stuttgart	298	F/O
Taipeh	445	F/O
Torino	399	F/O/S
Toronto	188	F/O
Toulouse	476	F/O
Utrecht	1,027	F
Vilnius	581	F/O
Wolfsburg	133	F/O
Zurich	1,020	F/O

UTD19's earliest records and Melbourne its latest; observation intervals differ across cities. Counts include all retained training, validation, and test data. All three sources count available values after processing, including accepted interpolations and gap fills. UTD19 comprises 131,558,927 cleaned original observations and 8,540,090 additional available records. Grouped arrays are deduplicated by node identity to exclude Core-Halo repeats and padding. Different variables at the same time and location are counted separately. UTD19 cities have a median of 13 days of stored five-minute time steps, with 32 cities at or below 31 days; coverage across years can consist of discontinuous periods.

B.2 Modality association and temporal semantics

All multimodal information used in each forecasting sample is available before the forecast origin.

Road information uses native node attributes and OpenStreetMap snapshots[4]. Neighborhood processing extracts road classes, lengths, densities, and orientation distributions and generates local road rasters. Node-to-road identity, neighborhood content, and native spatial relations are recorded separately. LargeST and TraffiDent neighborhood road information is constructed around valid observation coordinates. UTD19 combines native node-to-road relations with neighborhood road geometry. Snapshot versions, spatial identifiers, attribute validity, and association results are retained.

Sentinel-2 data are acquired through Google Earth Engine[5], Microsoft Planetary Computer[6], and Earth Search[7], with additional imagery from IGN orthophotos. We organize imagery around observation locations and construct model inputs through spatial cropping, projection, and raster resampling, retaining pixel validity, image sources, and acquisition years. The resulting imagery provides static information about the surrounding environment.

Weather uses historical HRRR, ERA5, ERA5-Land, and CERRA products. For the US data, HRRR comes from public forecast Zarr archives[8], while ERA5 and ERA5-Land come from Open-Meteo public OM archives. UTD19 uses the Open-Meteo Historical Weather API[9] to obtain CERRA, ERA5, and ERA5-Land. We associate hourly weather grids with observation locations, aggregate spatially within each source, and combine sources using field-specific priorities, retaining physical units, source identity, and per-field validity. Each forecasting window uses weather for the hour containing the last historical observation; missing hours and fields are represented by masks. Weather is aligned with traffic observations by its valid time.

Events use local or regional spatial scopes. Local matching takes the union of 5 km anchor and full-geometry neighborhoods, with road and direction matching where needed; regional matching uses explicit region geometry. Source-specific short and long windows are 6 and 24 hours, respectively, with trusted declared expiry taking precedence. Matching follows record availability times and validity intervals. UTD19 has no event modality.

B.3 Spatio-temporal data filtering and processing details

We standardize spatial identifiers, variable units, and timestamps, aligning observations to a five-minute grid. Resampling follows variable semantics: counts and demand use mass-preserving allocation, while continuous state variables use temporal aggregation or interpolation. Processing retains per-variable validity indicators and distinguishes original observations from accepted interpolations and short-gap fills. Remaining missing values are represented by masks, while genuine zeros remain valid observations.

Spatio-temporal data retain the observation identities provided by each original source. Node retention is recorded separately from subsequent training use. Original UTD19 Paris data have a 60-minute interval and fall outside the current unified five-minute processing scope; Tokyo has static metadata but no spatio-temporal records. Toulouse detector 262 refers to two locations and road identifiers, preventing reliable assignment of its observation values, so the detector is excluded. The original TraffiDent dataset has 16,972 nodes (Table 1); CityModal retains 16,961 with valid coordinates, aligned by station identifiers.

C Experimental Protocols and Detailed Results

Evaluation and reporting. For each of the K forecasting tasks, let $\mathcal{V}_k$ be the set of valid test targets, with original-scale observations y_j and predictions $\hat{y}_j$. We compute

$$\mathrm{MAE}_k = \frac{1}{|\mathcal{V}_k|} \sum_{j \in \mathcal{V}_k} |y_j - \hat{y}_j|, \qquad \mathrm{RMSE}_k = \sqrt{\frac{1}{|\mathcal{V}_k|} \sum_{j \in \mathcal{V}_k} (y_j - \hat{y}_j)^2}. \tag{13}$$

[4] https://www.openstreetmap.org/copyright
[5] https://developers.google.com/earth-engine
[6] https://planetarycomputer.microsoft.com/
[7] https://element84.com/earth-search/
[8] https://mesowest.utah.edu/html/hrrr/zarr_documentation/
[9] https://open-meteo.com/en/docs/historical-weather-api

We aggregate each metric $E \in \{\mathrm{MAE}, \mathrm{RMSE}\}$ as $\mathrm{Geo}E = (\prod_{k=1}^{K} E_k)^{1/K}$, giving each task equal weight; both metrics are lower-is-better. Aggregate GeoMAE and GeoRMSE are denoted by MAE and RMSE in result tables, figures, and discussion. Task-level MAE and RMSE retain their original-scale definitions. Short and Long results are aggregated separately. All methods use the same test origins and valid targets, and overall errors are computed directly over all 30 tasks rather than by averaging group scores. Foundation-model results come from a single training run; task-specific learning models are trained independently three times and their mean results are reported. HA is deterministic. Fixed random seeds are used for data partitioning and sampling across comparative experiments. Experiment-specific settings follow below.

C.1 Data roles, tasks, and inputs

Forecasting subsets. Each dataset-prediction-variable pair defines one forecasting subset and one equally weighted evaluation unit. Flow and speed prediction in the same region therefore count as two subsets. The test pool contains 30 subsets from 19 datasets: eight CityModal subsets and 11 external public datasets, evaluated separately at the Short and Long horizons.

Data partitioning. We hold out the test regions and periods listed below and use the remaining data as pretraining candidates. Spatio-temporal alignment removes candidate records that overlap the test set. This filtering also applies to baselines trained on the common corpus. Within each retained pretraining subset, time points are ordered chronologically: the first 85% are used for training and the final 15% for validation. The UTD19 component of CityModal contributes 32 cities to this pool, with six cities reserved for testing. Table 7 summarizes the source roles.

Table 7: Data sources used for pretraining and evaluation.

Source / pool	Pretraining role	Evaluation role
LargeST	GLA/GBA, 2017–2021	San Diego regional holdout
TraffiDent	2022–2023	California 2024 temporal holdout
UTD19	32 non-test cities	Six held-out cities; 13 tasks
Extra single-modal sources	Taxi-NYC; Bike-NYC/DC; Traffic-HZ/ZZ/CD/JN; Paris	Separate targets listed below
BjTT and extended targets	Excluded from MILU pretraining	Target adaptation / test

Table 8 lists the datasets and prediction variables for every task in the main pool. Spatio-temporal resampling follows the variable-specific rules in Appendix B.3. Congestion-level error is computed using the numerical congestion codes. Five-minute count/demand targets use native counts or count-preserving redistribution; state variables use their respective aggregation or interpolation rules.

Modality availability is recorded per sample.

Task groups. Held-out CityModal regions, periods, and cities, together with BjTT, form the multi-modal group: nine datasets and 19 tasks. Other external public targets form the single-modal group: 10 datasets and 11 tasks. Membership follows the evaluation input protocol and does not change with model input support or sample-level missingness. The groups are disjoint and together contain all 30 tasks.

Forecasting and adaptation. Evaluation uses a five-minute time grid. Short uses 12 historical steps to predict 12 future steps; Long uses 24 to predict 24, corresponding to one and two hours. Target test partitions follow the fixed benchmark schedules. The remaining target data are split chronologically, with the first 85% of time points used for training and the final 15% for validation. Windows do not cross natural time gaps or split boundaries. All few-shot foundation models use the same fixed random subset of 10% of the valid 24-to-24 training windows. Normalization statistics are computed only from training observations in the selected windows. Each selected window is divided into two consecutive, non-overlapping 12-to-12 windows for Short. Models share the target subset and test origins for each setting, with checkpoints selected by validation error. Zero-shot performs no target-task parameter updates.

Table 8: Test datasets and their prediction variables.

Dataset	Variables
San Diego	flow
California 2024	flow, occupancy, speed
Bolton	flow, occupancy, speed
Madrid	flow, occupancy
Melbourne	flow
Torino	flow, occupancy, speed
Toronto	flow, occupancy
Toulouse	flow, occupancy
Bike-CHI	count
BjTT	congestion level, speed
METR-LA	speed
PEMS-BAY	speed
PEMS03	flow
PEMS04	flow
PEMS07	flow
PEMS08	flow
Taxi-BJ	inflow, outflow
Taxi-CHI	demand
Traffic-SH	speed

C.2 Training and baseline settings

MILU training. Experiments run on a cluster with four NVIDIA RTX PRO 6000 GPUs. We pretrain MILU using AdamW with BF16 computation and dropout disabled. Matrix parameters use weight decay 10^{-4}, other parameters use zero weight decay, and the gradient-norm clipping threshold is 5. Learning rates for multimodal encoders and the shared influence module range from 3×10^{-4} to 3×10^{-5}. Unfrozen backbone parameters, the decoder, and prediction head use rates from 10^{-5} to 10^{-6} during joint training, with cosine decay indexed by stage progress.

Test labels are excluded from gradient updates and normalization statistics.

Baseline implementations. Table 9 summarizes the baseline versions and implementations.

Table 9: Baseline versions and implementations.

Model	Version / weight	Implementation
MoST	Project reproduction	Pretrained MoST architecture with multimodal inputs
UrbanFM	STFormer reproduction	Pretrained on the public WorldST package
OpenCity	Official Plus checkpoint	Released pretrained weights
TimesFM	2.5-200M	Official point forecasts
Chronos	Chronos-2 Small	Univariate forecasts; predictive median
Moirai	2.0-R-small	Univariate forecasts; predictive median
HA	Historical average	Historical means estimated from target training data
E^2-CSTP / T3	Corresponding implementations	Model-specific training settings and input interfaces
STAEformer / STID / PatchSTG	Corresponding implementations	Model-specific training settings and spatial inputs

Pretraining and target adaptation. We use public checkpoints when available, or reproduce models with complete public pretraining data and code using the recommended settings. Models with

neither usable public weights nor complete public pretraining data are trained on the same data as MILU. Foundation models use the same 10% target-training data budget as MILU, while task-specific models use the full target training set. On the PeMS adaptation tasks, Chronos-2 Small, TimesFM 2.5-200M, and Moirai 2.0-R-small use LoRA: pretrained base weights remain frozen, and low-rank adapter parameters are updated on the target training subset. HA estimates historical means from the target training data to predict future observations.

Normalization. Input normalization follows the same observability constraint as MILU: statistics use permitted observed data and exclude future prediction targets and missing values. Models retain their native normalization forms, using training-set or visible-history statistics as appropriate.

Input and evaluation scope. Models receive the same length of visible spatio-temporal observations and predict the same target locations, retaining the history normalization and forecasting mapping associated with their training. Spatial and multimodal information follow each model's interface. All multimodal information used for forecasting is available before the forecast origin. Weather inputs use the hour containing the last historical observation. Comparisons with the same access to multimodal information and comparisons against spatio-temporal-only inputs are interpreted separately.

C.3 Task-level main results

Short and Long zero-shot evaluations use the same forecast origins, valid targets, and causal historical inputs within each horizon. MILU uses spatio-temporal observations for the single-modal group and available multimodal information for the multimodal group. Values are reported to three decimal places. Aggregation uses the supplied task-level values before rounding the aggregate.

Tables 10 and 11 expand the zero-shot results in Table 2; Tables 12 and 13 expand the few-shot and task-specific results in Table 3, using the target-data fractions specified in the main text. For each horizon, the zero-shot table includes all seven foundation models. For each horizon, one combined table compares task-specific and statistical baselines using 100% target data with foundation models adapted on 10% target data, retaining all 30 evaluation units.

MILU achieves the lowest zero-shot task-level MAE on 28 of 30 Short tasks and 27 of 30 Long tasks; the corresponding RMSE counts are 25 and 27.

Task labels use Flow (with in/out direction where applicable), Speed, Occupancy, Demand (including bike-trip counts), and Congestion. These labels do not change the prediction targets or evaluation units.

Cells contain MAE / RMSE. Each dataset-prediction-variable row is one equally weighted evaluation unit, also called a forecasting task in this paper. Overall scores aggregate all 30 tasks.

Table 10: Zero-shot forecasting performance by task (Short).

Dataset Task	MoST	UrbanFM	OpenCity	Chronos	TimesFM	Moirai	**MILU**
San Diego	28.133	24.874	56.458	29.771	31.074	34.811	**21.226**
Flow	42.000	38.082	78.609	44.875	46.804	50.549	**32.962**
California 2024	21.493	14.542	26.446	14.290	14.781	16.703	**10.496**
Flow	34.183	25.937	49.320	28.750	29.541	32.461	**21.146**
California 2024	0.952	0.767	1.276	0.781	0.764	0.830	**0.601**
Occupancy	2.345	2.214	3.143	2.451	2.329	2.364	**2.051**
California 2024	4.275	2.774	1.618	1.158	1.123	1.160	**1.059**
Speed	6.027	4.872	4.423	3.441	3.405	3.299	**2.771**
Bolton	82.350	71.212	112.089	84.147	82.206	85.075	**69.492**
Flow	132.842	115.373	174.769	134.600	132.660	134.317	**112.927**
Bolton	4.456	3.909	5.568	4.401	4.322	4.384	**3.687**
Occupancy	8.727	7.962	9.816	8.513	8.476	8.430	**7.596**
Bolton	9.722	9.085	10.411	9.266	9.121	8.999	**8.749**
Speed	17.570	16.626	17.525	18.052	17.958	17.386	**16.620**
Madrid	126.188	100.833	149.437	116.443	115.526	116.781	**99.900**
Flow	190.692	**155.711**	239.746	184.236	184.445	181.748	156.639
Madrid	3.249	2.693	3.405	2.884	2.767	2.811	**2.553**
Occupancy	6.529	**6.056**	6.897	6.544	6.377	6.292	6.095
Melbourne	43.231	42.946	62.762	40.588	41.250	40.964	**35.537**
Flow	64.618	63.096	89.521	60.926	61.527	60.579	**53.368**
Torino	108.489	96.764	147.267	109.876	109.620	112.923	**93.383**
Flow	162.863	147.905	220.769	168.536	169.642	171.889	**143.333**
Torino	3.971	3.543	5.090	3.943	3.890	4.061	**3.345**
Occupancy	7.527	7.086	8.797	7.532	7.526	7.636	**6.792**
Torino	8.348	7.918	7.964	7.661	7.560	7.694	**7.542**
Speed	13.163	12.622	12.505	12.260	12.214	12.384	**12.081**
Toronto	65.925	62.370	98.561	55.688	54.235	58.338	**50.979**
Flow	102.875	98.029	142.470	90.319	88.340	90.517	**81.332**
Toronto	3.900	3.532	4.715	3.209	**2.916**	3.110	2.965
Occupancy	9.472	8.855	10.417	9.146	**8.316**	8.423	8.410
Toulouse	124.723	106.108	174.161	128.295	126.028	126.036	**104.384**
Flow	176.867	153.336	260.746	191.347	187.293	182.672	**150.206**
Toulouse	1.792	1.568	2.227	1.799	1.680	1.722	**1.487**
Occupancy	3.429	3.203	4.026	3.605	3.405	3.413	**3.135**
Bike-CHI	0.108	0.121	0.300	0.066	0.066	0.064	**0.063**
Demand	**0.331**	0.339	0.581	0.337	0.333	0.335	0.332
BjTT	0.141	0.104	0.148	0.084	0.078	0.085	**0.073**
Congestion	0.350	0.325	0.417	0.352	0.339	0.331	**0.323**
BjTT	4.485	4.245	4.131	3.704	3.559	3.628	**3.439**
Speed	6.574	6.485	6.805	6.270	6.026	5.883	**5.609**
METR-LA	4.651	4.717	5.865	4.510	4.338	4.426	**3.957**
Speed	8.328	8.833	10.783	9.056	8.941	8.635	**7.844**
PEMS-BAY	2.642	2.764	3.595	2.498	2.380	2.413	**2.045**
Speed	5.528	5.996	7.813	6.020	5.808	5.587	**4.787**
PEMS03	23.953	19.959	41.198	24.090	24.690	26.892	**18.618**
Flow	36.378	31.761	61.655	38.081	39.255	41.741	**29.500**
PEMS04	31.014	26.211	53.742	30.759	31.925	35.103	**24.289**
Flow	45.510	40.313	75.730	46.910	48.522	51.658	**37.319**
PEMS07	36.619	31.015	64.609	35.094	36.507	40.598	**27.510**
Flow	52.535	46.730	91.793	54.221	56.502	60.565	**42.800**
PEMS08	26.131	21.982	45.207	25.063	25.949	28.825	**19.461**
Flow	38.019	33.768	63.870	39.051	40.085	42.931	**30.474**
Taxi-BJ	6.498	3.184	3.856	3.231	3.062	3.478	**2.936**
Flow (in)	10.346	5.999	7.029	5.941	5.589	6.369	**5.281**
Taxi-BJ	6.482	3.178	3.856	3.237	3.072	3.487	**2.948**
Flow (out)	10.318	5.991	7.053	5.968	5.623	6.405	**5.293**
Taxi-CHI	2.513	0.996	0.956	0.732	0.704	0.803	**0.658**
Demand	7.113	3.295	4.288	3.128	2.917	3.539	**2.608**
Traffic-SH	0.600	0.563	0.707	0.483	**0.398**	0.435	0.488
Speed	0.974	0.936	1.146	0.973	**0.796**	0.803	0.941

Table 11: Zero-shot forecasting performance by task (Long).

Dataset Task	MoST	UrbanFM	OpenCity	Chronos	TimesFM	Moirai	**MILU**
San Diego	42.383	36.651	72.353	45.874	48.812	55.827	**27.689**
Flow	64.681	57.165	102.225	70.075	74.386	82.644	**43.779**
California 2024	29.807	19.606	32.389	20.752	21.885	25.232	**13.279**
Flow	49.260	36.089	60.624	41.549	44.029	49.277	**27.595**
California 2024	1.282	0.997	1.514	1.048	1.064	1.201	**0.740**
Occupancy	3.002	2.664	3.532	2.896	2.908	3.054	**2.383**
California 2024	5.541	4.457	1.972	1.578	1.551	1.719	**1.364**
Speed	8.120	7.788	5.339	4.634	4.604	4.645	**3.642**
Bolton	111.348	88.266	137.388	111.386	110.979	119.965	**87.475**
Flow	174.682	142.789	211.492	175.979	177.289	187.907	**138.340**
Bolton	5.883	4.696	6.785	5.698	5.567	5.958	**4.599**
Occupancy	10.371	8.787	11.396	10.196	10.122	10.436	**8.702**
Bolton	11.863	11.048	11.467	10.609	10.420	10.755	**9.949**
Speed	20.970	19.208	18.754	19.701	19.595	19.340	**18.113**
Madrid	168.090	123.462	177.691	148.124	149.337	155.353	**123.335**
Flow	268.054	**205.054**	295.706	250.101	252.468	256.912	209.100
Madrid	4.036	3.167	3.851	3.347	3.266	3.383	**2.949**
Occupancy	7.934	6.920	7.744	7.340	7.310	7.316	**6.847**
Melbourne	59.589	61.368	72.905	54.037	56.476	58.549	**49.691**
Flow	92.484	91.902	104.243	83.061	84.501	86.850	**78.563**
Torino	139.396	113.889	169.718	137.775	138.346	150.058	**110.134**
Flow	218.449	185.994	261.917	221.043	222.996	236.284	**181.781**
Torino	4.910	4.014	5.601	4.651	4.613	4.991	**3.825**
Occupancy	8.818	7.864	9.668	8.630	8.665	8.997	**7.605**
Torino	9.040	9.267	8.091	7.867	**7.780**	8.157	7.955
Speed	14.103	14.907	12.646	12.553	12.571	12.883	**12.546**
Toronto	98.620	90.843	119.008	84.567	84.860	92.028	**80.202**
Flow	163.810	146.469	173.172	138.584	136.471	142.679	**129.401**
Toronto	5.396	4.830	5.489	4.596	4.122	4.353	**4.060**
Occupancy	12.240	11.112	11.835	12.192	10.605	10.606	**10.426**
Toulouse	164.615	**129.550**	191.997	164.629	160.778	165.586	133.978
Flow	242.919	**198.332**	287.792	256.056	249.336	251.308	205.962
Toulouse	2.222	1.792	2.402	2.082	1.979	2.087	**1.724**
Occupancy	4.126	3.613	4.314	4.023	3.941	3.959	**3.552**
Bike-CHI	0.117	0.126	0.315	0.066	0.067	0.065	**0.064**
Demand	0.345	0.350	0.611	0.342	0.339	0.341	**0.337**
BjTT	0.175	0.128	0.163	0.101	0.096	0.111	**0.087**
Congestion	0.414	0.370	0.455	0.399	0.391	0.391	**0.366**
BjTT	5.261	5.969	4.516	4.192	4.058	4.689	**3.956**
Speed	7.761	8.951	7.484	7.150	6.973	7.247	**6.492**
METR-LA	5.817	6.333	6.601	5.566	5.457	5.932	**4.974**
Speed	10.401	11.390	12.121	10.982	11.045	10.881	**9.771**
PEMS-BAY	3.467	3.958	4.187	3.308	3.257	3.521	**2.689**
Speed	7.257	8.180	9.008	7.773	7.767	7.639	**6.305**
PEMS03	33.769	27.424	51.829	35.497	36.947	41.360	**24.368**
Flow	52.570	45.063	77.790	57.123	59.455	64.638	**40.848**
PEMS04	42.322	34.826	65.164	44.000	46.231	51.891	**30.921**
Flow	62.444	54.075	92.110	67.660	70.420	76.671	**47.840**
PEMS07	50.433	42.241	77.616	50.876	53.430	60.538	**35.264**
Flow	73.834	64.296	110.767	79.465	83.150	90.702	**55.686**
PEMS08	35.979	29.542	55.954	36.378	38.548	43.320	**24.376**
Flow	52.613	45.595	79.275	56.374	59.300	64.795	**38.183**
Taxi-BJ	7.395	3.957	4.560	3.963	3.857	4.310	**3.675**
Flow (in)	11.534	7.576	8.397	7.379	7.181	7.994	**6.816**
Taxi-BJ	7.404	3.968	4.570	3.974	3.869	4.321	**3.688**
Flow (out)	11.547	7.592	8.412	7.397	7.198	8.010	**6.834**
Taxi-CHI	2.961	1.274	1.096	0.893	0.874	0.998	**0.801**
Demand	7.718	4.211	5.092	4.112	3.951	4.680	**3.491**
Traffic-SH	0.780	0.754	0.809	0.671	**0.585**	0.608	0.680
Speed	1.258	1.192	1.286	1.225	**1.044**	1.047	1.195

Table 12: Few-shot foundation models versus fully supervised baselines by task (Short).

	Task-specific (100% target data)							Foundation models (10% target data)					
Dataset Task	HA	E^2-CSTP	T3	STAE former	STID	Patch STG	MoST	Urban FM	Open City	Chronos	Times FM	Moirai	**MILU**
San Diego	36.795	19.941	18.441	18.135	19.629	18.310	21.264	18.752	21.034	28.379	28.815	28.316	**18.133**
Flow	58.062	30.371	28.862	**28.494**	30.802	28.775	31.295	29.298	31.359	44.052	44.697	46.384	28.831
California 2024	18.772	10.893	12.321	9.905	9.087	**8.527**	11.985	11.522	12.023	13.479	14.094	13.679	9.064
Flow	41.083	23.022	21.625	20.827	18.847	**18.167**	21.971	22.905	23.833	28.440	28.515	28.620	18.263
California 2024	1.074	0.723	0.787	0.678	0.637	0.635	0.673	0.651	0.692	0.706	0.752	0.616	**0.597**
Occupancy	2.982	2.256	2.390	2.163	2.076	2.074	2.595	2.123	2.429	2.441	2.306	2.339	**1.993**
California 2024	1.865	1.132	1.071	1.040	0.960	0.952	1.580	1.417	1.560	1.053	0.981	0.961	**0.945**
Speed	5.463	3.118	2.923	2.821	2.659	2.567	3.539	3.416	3.824	3.405	3.283	2.931	**2.488**
Bolton	83.526	74.474	74.953	78.622	**65.940**	71.167	77.807	70.050	72.741	82.091	80.438	80.788	67.814
Flow	134.295	114.494	117.317	104.520	**99.402**	103.808	117.662	112.885	119.354	135.865	132.216	126.793	102.683
Bolton	4.565	4.339	5.055	5.281	3.709	3.822	4.396	**3.349**	4.421	4.213	4.258	4.090	3.414
Occupancy	9.126	7.902	7.076	7.456	7.321	7.405	8.122	**6.890**	8.046	8.481	8.394	8.101	7.026
Bolton	9.606	9.785	11.212	**8.540**	9.004	9.438	9.552	8.943	10.115	9.222	9.072	8.850	8.687
Speed	19.930	16.248	**14.424**	16.054	16.224	16.765	17.009	16.358	17.073	17.882	17.886	17.222	16.066
Madrid	208.698	112.130	107.180	116.027	113.404	101.637	121.714	97.863	122.341	115.607	113.007	113.943	**95.490**
Flow	344.935	166.654	174.068	192.518	182.859	166.908	177.557	151.911	171.355	183.664	183.006	176.725	**146.334**
Madrid	5.011	2.619	2.745	2.658	2.723	2.862	3.105	2.615	3.178	2.823	2.745	2.706	**2.460**
Occupancy	10.106	**5.340**	6.207	6.036	6.061	6.321	5.865	5.861	6.144	6.498	6.202	6.115	6.095
Melbourne	41.898	28.996	25.480	26.242	**24.396**	25.372	41.316	40.743	44.659	39.140	39.719	33.141	27.002
Flow	69.778	42.404	38.615	41.217	**37.850**	38.537	60.566	59.819	67.796	59.485	60.571	56.766	40.956
Torino	90.752	82.313	75.996	73.937	**72.538**	74.707	104.558	87.348	92.346	103.460	104.197	104.152	76.838
Flow	145.125	125.409	118.039	116.806	**112.060**	114.523	155.069	137.370	161.004	163.880	166.135	159.685	118.400
Torino	3.660	3.003	2.899	2.993	2.861	2.904	3.827	3.155	3.409	3.809	3.757	3.641	**2.789**
Occupancy	7.626	**5.696**	6.255	6.321	6.297	6.200	6.788	6.206	6.288	7.456	7.421	7.147	6.042
Torino	8.877	7.278	6.966	6.950	6.924	6.983	7.622	7.329	7.699	7.373	7.210	6.865	**6.807**
Speed	14.255	11.376	11.358	11.443	11.261	11.372	11.290	11.818	12.381	12.391	11.652	11.131	**10.829**
Toronto	85.043	49.311	**38.046**	41.349	40.610	39.813	63.306	56.664	62.658	53.956	51.857	50.699	38.680
Flow	120.878	76.875	**60.410**	63.619	65.436	64.912	96.887	92.135	111.293	89.773	86.988	86.573	63.397
Toronto	5.330	2.716	2.361	2.557	**2.333**	2.454	3.869	3.440	4.300	3.165	2.893	2.989	2.525
Occupancy	11.293	6.828	6.926	7.359	**6.660**	6.924	8.930	8.272	8.906	9.109	8.097	8.139	7.248
Toulouse	83.005	84.144	79.042	75.589	**75.498**	77.480	118.660	104.028	132.232	122.975	122.981	119.699	80.281
Flow	138.168	117.894	118.742	111.532	**109.458**	116.159	166.793	150.061	189.586	186.655	181.120	171.806	114.473
Toulouse	1.721	1.306	1.305	1.286	**1.257**	1.337	1.713	1.534	1.630	1.739	1.603	1.686	1.324
Occupancy	3.479	**2.421**	2.942	2.884	2.841	2.933	3.089	3.143	3.238	3.599	3.328	3.148	2.938
Bike-CHI	0.086	0.067	0.057	0.059	0.059	**0.056**	0.083	0.061	0.069	0.061	0.060	0.060	0.057
Demand	0.331	0.254	0.318	0.343	0.341	0.310	**0.251**	0.302	0.278	0.328	0.313	0.309	0.310
BjTT	0.227	0.074	0.090	0.080	0.079	0.077	0.127	0.099	0.123	0.082	0.077	0.082	**0.069**
Congestion	0.480	0.308	0.340	0.306	0.323	0.311	**0.294**	0.318	0.318	0.339	0.326	0.314	**0.294**
BjTT	11.339	3.145	3.059	3.078	3.098	3.214	3.426	3.208	3.247	3.456	3.254	3.024	**3.019**
Speed	14.752	5.194	5.063	5.106	5.219	5.166	**4.984**	5.217	5.587	6.120	5.663	5.012	5.157
METR-LA	4.305	3.412	**3.134**	3.315	3.247	3.188	3.850	3.963	4.194	4.056	4.016	3.746	3.153
Speed	7.974	7.039	7.217	6.934	6.852	6.805	**6.688**	7.778	8.298	8.737	8.644	7.752	6.778
PEMS-BAY	2.745	1.762	1.566	1.628	1.552	1.534	2.070	2.235	2.223	2.340	2.170	1.949	**1.485**
Speed	5.417	3.959	3.726	3.767	3.627	3.633	4.137	5.126	5.306	5.674	5.615	4.801	**3.624**
PEMS03	26.334	17.599	**15.010**	15.358	15.333	15.479	18.856	18.481	20.478	22.670	23.365	22.554	15.536
Flow	48.294	27.513	**26.689**	27.558	27.414	27.417	28.590	29.838	33.929	36.344	37.767	38.180	27.918
PEMS04	25.817	19.823	**17.179**	18.223	18.193	18.359	24.288	24.186	26.734	29.476	30.365	29.782	18.619
Flow	41.368	31.938	**28.398**	30.182	29.765	29.809	34.539	38.183	42.511	46.631	47.319	47.675	30.143
PEMS07	30.083	20.890	**18.218**	18.651	19.611	19.007	27.265	27.377	30.266	32.827	34.530	31.990	18.351
Flow	56.276	36.478	31.294	32.067	32.796	32.045	40.318	42.464	47.602	52.573	54.308	52.521	**31.228**
PEMS08	23.154	15.599	**13.027**	13.100	14.219	13.894	19.912	20.149	21.287	23.828	23.942	23.235	13.485
Flow	40.432	26.022	22.254	22.629	23.285	22.732	28.118	31.442	33.884	38.480	38.108	37.809	**22.038**
Taxi-BJ	3.868	1.439	1.167	1.200	1.362	**1.142**	2.441	2.054	2.168	2.932	2.601	2.405	1.313
Flow (in)	7.958	2.880	**2.697**	2.769	2.875	2.740	4.207	4.192	4.447	5.860	5.313	5.052	2.737
Taxi-BJ	3.898	1.390	1.196	1.276	1.382	**1.192**	2.102	2.094	2.195	2.777	2.768	2.539	1.359
Flow (out)	8.085	2.763	**2.740**	2.850	2.867	2.804	3.855	4.154	4.468	5.919	5.375	5.012	2.825
Taxi-CHI	1.230	0.579	0.432	0.500	0.434	0.419	0.604	0.740	0.621	0.682	0.637	0.588	**0.411**
Demand	4.860	2.292	1.715	1.973	1.665	1.652	2.450	3.000	2.680	3.074	2.811	2.963	**1.645**
Traffic-SH	0.919	0.421	0.316	0.359	0.305	0.294	0.597	0.523	0.453	0.451	0.380	0.338	**0.293**
Speed	1.215	0.721	0.626	0.659	0.608	0.593	0.954	0.884	0.817	0.955	0.792	0.716	**0.592**

Table 13: Few-shot foundation models versus fully supervised baselines by task (Long).

	Task-specific (100% target data)						Foundation models (10% target data)					
Dataset Task	HA	E^2-CSTP	T3	STAE former	STID	Patch STG	MoST	Urban FM	Open City	Chronos	Times FM	Moirai	**MILU**
San Diego	38.367	22.983	21.868	20.739	21.819	20.999	26.884	23.378	25.071	39.174	38.018	36.779	**20.455**
Flow	60.604	35.220	34.799	32.958	34.816	33.340	39.589	37.096	37.527	65.934	63.649	61.706	**32.232**
California 2024	19.120	11.816	12.820	11.105	10.313	**9.428**	14.417	13.772	13.884	17.064	18.880	17.050	10.126
Flow	41.645	23.710	24.635	23.504	21.508	**20.484**	26.959	27.756	27.794	39.803	38.677	35.768	21.045
California 2024	1.066	0.838	0.846	0.739	0.775	0.802	0.830	0.800	0.822	0.882	1.007	0.784	**0.730**
Occupancy	3.024	2.512	2.583	2.381	2.406	2.336	2.932	2.435	2.709	2.868	2.846	2.909	**2.263**
California 2024	1.884	1.260	1.257	1.190	1.131	1.148	1.924	1.845	1.860	1.309	1.217	1.231	**1.121**
Speed	5.591	3.322	3.388	3.090	3.148	3.218	4.380	4.488	4.552	4.403	4.068	3.619	**2.950**
Bolton	108.783	86.160	103.776	84.129	**74.279**	74.485	93.485	82.554	83.679	101.297	103.792	101.306	74.688
Flow	172.309	131.363	151.240	118.841	109.204	**108.689**	140.137	133.563	135.897	176.729	171.143	155.533	112.352
Bolton	5.812	5.352	6.091	6.654	**3.980**	4.058	5.507	4.020	5.280	5.172	5.306	5.117	4.019
Occupancy	10.853	9.486	11.266	11.113	7.595	7.673	9.285	7.661	9.059	9.970	9.724	9.478	**7.594**
Bolton	10.930	9.964	12.015	**8.903**	9.506	9.583	11.144	10.352	10.799	10.408	10.061	10.127	9.163
Speed	21.578	**16.329**	19.508	16.846	17.144	17.582	18.884	18.006	18.132	19.135	19.231	18.443	16.939
Madrid	200.663	127.128	116.969	132.788	130.359	117.108	149.088	114.610	143.195	142.860	138.809	142.534	**113.211**
Flow	331.493	198.443	199.331	231.390	224.002	204.333	226.420	189.698	208.915	243.655	242.546	236.005	**184.342**
Madrid	5.205	2.942	2.971	2.925	3.032	3.136	3.593	2.967	3.542	3.193	3.203	3.080	**2.716**
Occupancy	10.546	**5.787**	6.671	6.535	6.621	6.837	6.585	6.482	6.751	7.138	6.918	6.833	6.307
Melbourne	41.981	32.910	28.174	29.419	**28.119**	28.966	50.607	51.173	50.967	47.374	49.665	40.629	30.234
Flow	70.069	48.495	**43.061**	47.072	43.831	43.970	75.837	76.538	77.543	75.235	78.575	70.546	46.403
Torino	92.592	86.078	78.155	75.813	**74.940**	77.700	122.044	95.002	98.674	114.901	117.063	118.199	77.884
Flow	161.404	134.427	125.796	123.651	**120.151**	121.760	183.363	155.074	177.577	198.270	208.025	187.009	122.203
Torino	3.654	3.199	2.964	3.043	**2.917**	2.974	4.319	3.359	3.596	4.119	4.113	3.956	2.935
Occupancy	7.693	**5.967**	6.405	6.493	6.428	6.355	7.318	6.588	6.641	8.251	8.155	7.717	6.295
Torino	8.736	7.524	7.016	6.990	6.965	7.035	7.814	7.819	7.818	7.481	7.375	7.074	**6.829**
Speed	13.970	11.671	11.459	11.516	11.360	11.472	11.655	12.578	12.515	12.594	11.751	11.384	**11.063**
Toronto	84.898	55.847	**45.655**	51.527	52.917	50.250	85.479	75.486	76.843	74.796	73.866	68.649	48.677
Flow	122.312	87.173	**72.108**	79.692	84.625	81.278	134.984	121.822	137.767	132.873	128.788	123.346	79.500
Toronto	5.369	4.098	**2.860**	3.037	3.031	3.049	5.263	4.478	5.156	4.347	4.033	3.965	3.271
Occupancy	11.239	9.341	**8.216**	8.656	8.328	8.446	11.269	10.353	10.442	11.758	10.134	10.064	9.125
Toulouse	84.005	88.070	81.696	77.296	**77.041**	79.619	140.560	120.924	140.075	142.478	149.312	137.628	83.577
Flow	138.962	126.957	126.431	117.373	**114.159**	122.840	198.562	185.109	206.141	233.771	220.957	207.817	128.594
Toulouse	1.644	1.394	1.337	1.319	**1.293**	1.357	1.906	1.675	1.701	1.884	1.718	1.910	1.347
Occupancy	3.384	**2.541**	3.001	2.962	2.912	3.006	3.368	3.384	3.387	4.006	3.615	3.415	3.030
Bike-CHI	0.086	0.072	0.061	0.062	0.063	**0.060**	0.089	0.066	0.074	0.063	0.062	0.063	0.061
Demand	0.330	0.341	0.337	0.359	0.363	0.329	0.363	**0.318**	0.384	0.342	0.327	0.320	0.329
BjTT	0.226	0.082	0.099	0.088	0.093	0.084	0.146	0.115	0.136	0.094	0.090	0.099	**0.079**
Congestion	0.477	0.326	0.358	0.323	0.348	0.327	0.324	0.349	0.339	0.368	0.357	0.343	**0.314**
BjTT	11.345	3.316	3.219	3.224	3.262	3.255	3.740	3.704	3.497	3.734	3.502	3.410	**3.199**
Speed	14.760	5.367	5.216	5.273	5.294	5.284	5.404	5.978	5.868	6.618	6.021	5.514	**5.173**
METR-LA	4.266	4.370	3.473	3.517	3.502	3.449	4.408	4.654	4.593	4.574	4.549	4.344	**3.278**
Speed	7.973	7.734	7.311	7.286	7.303	7.158	7.491	8.823	8.968	9.689	9.922	8.682	**7.064**
PEMS-BAY	2.706	2.436	1.862	1.900	1.902	1.909	2.578	2.880	2.656	2.937	2.758	2.552	**1.832**
Speed	5.331	5.132	**4.299**	4.319	4.307	4.361	5.124	6.383	6.251	7.009	6.962	6.074	4.351
PEMS03	26.221	19.390	**17.181**	17.327	17.494	17.448	23.044	22.495	23.907	29.058	30.920	28.492	17.842
Flow	47.627	30.024	29.443	31.139	29.794	**29.203**	34.543	36.819	38.779	48.930	51.416	47.725	30.038
PEMS04	26.192	21.555	**19.379**	20.710	20.197	20.305	29.173	28.821	30.650	38.001	38.926	36.755	20.901
Flow	41.857	33.611	**31.533**	32.763	32.282	32.632	41.006	44.772	47.865	66.872	64.387	58.745	33.128
PEMS07	30.279	23.265	21.154	**21.053**	22.544	22.302	33.589	33.783	35.593	41.159	44.139	40.157	22.362
Flow	56.649	38.869	**35.962**	36.144	37.153	37.324	48.948	51.494	55.967	71.369	73.157	66.045	36.734
PEMS08	23.150	17.167	14.965	**14.788**	16.265	15.789	24.327	24.537	25.212	30.817	30.325	29.045	16.177
Flow	40.331	27.869	25.430	**25.277**	26.101	26.329	33.977	37.906	40.170	53.027	49.975	47.033	26.081
Taxi-BJ	3.898	1.683	1.362	1.394	1.785	**1.334**	2.845	2.466	2.555	3.479	3.117	2.872	1.376
Flow (in)	8.079	3.164	2.963	3.028	3.496	**2.878**	4.581	4.785	5.052	7.036	6.214	5.877	2.915
Taxi-BJ	3.907	1.606	**1.369**	1.466	1.799	1.376	2.397	2.476	2.564	3.271	3.291	3.007	1.427
Flow (out)	8.089	3.033	2.985	3.116	3.519	**2.956**	4.220	4.801	5.068	7.122	6.212	5.715	3.031
Taxi-CHI	1.215	0.592	0.460	0.526	0.507	0.438	0.658	0.830	0.674	0.756	0.713	0.654	**0.425**
Demand	4.835	2.381	1.919	2.171	2.112	1.837	2.686	3.482	3.025	3.873	3.512	3.487	**1.805**
Traffic-SH	0.919	0.493	0.377	0.383	0.406	0.355	0.758	0.650	0.534	0.571	0.496	0.429	**0.354**
Speed	1.215	0.836	0.710	0.734	0.733	0.675	1.164	1.054	0.926	1.147	1.018	0.858	**0.659**

C.4 SCALING RESULTS AND MODEL CONFIGURATIONS

Data and model scaling. For data scaling, we independently train the M model on 12.5%, 25%, 50%, and 100% of valid pretraining windows at a fixed update count. For model scaling, we train S, M, and L on all pretraining data with fixed update counts and depth; their widths, influence ranks, and parameter counts are listed in Table 17. Both experiments use zero-shot evaluation with the same test pools, aggregation, and paired on/off protocol. Table 15 reports data-scaling results, and Table 16 reports model-scaling results, with MAE above RMSE in each cell. Single-modal on and off predictions coincide, and the full-data and M rows refer to the same model.

Modality scaling. We fix the pretrained main model without target-task adaptation and progressively enable Event, Weather, Remote, and Road in this order. Historical spatio-temporal observations remain available throughout and are excluded from the modality count. All four conditions use the same 19 multimodal tasks, test samples, target masks, and model weights. Original availability masks determine which enabled modalities are available for each sample. Table 14 reports aggregate MAE for each step.

Table 14: Effect of modality count on zero-shot forecasting performance.

Count	Enabled modalities	Short	Long
1	Event	7.658	9.430
2	Event + Weather	7.571	9.333
3	Event + Weather + Remote	7.528	9.298
4	Event + Weather + Remote + Road	7.478	9.262

Table 15: Effect of pretraining data size on zero-shot forecasting performance.

Data / horizon	Overall on	Overall off	Multimodal on	Multimodal off	Single-modal
12.5% / Short	6.131 11.924	6.203 11.938	8.530 16.142	8.690 16.171	3.466 7.066
12.5% / Long	7.697 14.964	7.796 14.986	10.689 20.044	10.907 20.090	4.365 9.032
25% / Short	5.963 11.716	6.068 11.739	8.239 15.828	8.469 15.877	3.412 6.969
25% / Long	7.455 14.650	7.602 14.680	10.267 19.580	10.589 19.643	4.289 8.877
50% / Short	5.620 11.431	5.715 11.459	7.692 15.377	7.899 15.436	3.268 6.849
50% / Long	7.011 14.228	7.164 14.268	9.554 18.920	9.885 19.005	4.108 8.696
100% / Short	5.469 11.121	5.589 11.152	7.478 14.909	7.738 14.974	3.186 6.703
100% / Long	6.802 13.850	6.980 13.898	9.262 18.373	9.648 18.472	3.991 8.501

Parameter totals include the backbone and multimodal prediction head, excluding external frozen image and text encoders. M is the default model used in the main experiments.

C.5 FUSION, PRETRAINING, AND COMPONENT ABLATION DETAILS

Table 18 reports zero-shot ablations on the 19 multimodal tasks, using MAE and RMSE. Each variant is trained once with common initialization for unchanged components, data order, and evaluation samples. Architecture and training settings follow full MILU except for the factor under study.

Fusion variants. Concat, Cross-attn, and FiLM replace the mapping from modality keys to forecasting representations while retaining the backbone, modality encoders, and input-validity

Table 16: Effect of model size on zero-shot forecasting performance.

Scale / horizon	Overall on	Overall off	Multimodal on	Multimodal off	Single-modal
S / Short	5.904 11.588	6.005 11.620	8.190 15.638	8.412 15.706	3.354 6.905
S / Long	7.403 14.562	7.558 14.605	10.254 19.469	10.594 19.559	4.217 8.818
M / Short	5.469 11.121	5.589 11.152	7.478 14.909	7.738 14.974	3.186 6.703
M / Long	6.802 13.850	6.980 13.898	9.262 18.373	9.648 18.472	3.991 8.501
L / Short	5.429 11.089	5.556 11.121	7.408 14.856	7.682 14.923	3.174 6.691
L / Long	6.772 13.817	6.946 13.864	9.214 18.319	9.593 18.418	3.978 8.489

Table 17: Model configurations and parameter counts.

Configuration	S	M	L
Hidden dimension	64	128	256
Attention heads	4	8	16
Dimension per head	16	16	16
Encoder layers	6	6	6
Decoder layers	1	1	1
FFN dimension	256	512	1024
Influence rank	16	32	64
Backbone parameters	472,309	1,870,313	7,443,409
Multimodal head parameters	168,216	440,856	1,308,696
Total parameters	640,525	2,311,169	8,752,105

Table 18: Ablation results for fusion, pretraining, and model components.

Variant	Short MAE	Short RMSE	Long MAE	Long RMSE
MILU	7.478	14.909	9.262	18.373
A. Fusion				
Concat	7.782	15.298	9.782	18.978
Cross-attn	7.606	15.131	9.500	18.663
FiLM	7.615	15.040	9.376	18.656
B. Pretraining				
Joint	7.658	15.103	9.491	18.642
No gain	7.540	15.038	9.470	18.615
No subset	7.530	15.019	9.384	18.588
C. Components				
No query	7.538	14.971	9.382	18.516
Private influence	7.641	15.206	9.513	18.681
No content	7.582	15.079	9.525	18.768

rules. For alternatives requiring a single Event representation, candidate keys are pooled using the weights in Appendix A.2; MILU instead pools influences after candidate-specific historical reading. Let c_i collect modality keys and validity indicators, and let $g_i = \mathbf{1}[n_i > 0]$. Concat applies $h_i^\star = h_i + g_i F_{\mathrm{cat}}([h_i; c_i])$ with a linear map F_{cat}. Cross-attn uses h_i to query valid modality keys and adds the retrieved content to h_i. FiLM replaces MILU's historical retrieval, shared influence aggregation, and bounded modulation with content-conditioned affine modulation. It directly generates channel-wise parameters a_i, b_i from c_i and applies $h_i^\star = h_i \odot (1 + g_i a_i) + g_i b_i$. All three retain the original spatio-temporal prediction path when no valid multimodal information is available.

Pretraining variants. Joint updates the parameters permitted in the joint phase of MILU from the start, removing the initial frozen-backbone phase. No gain removes both the on/off and on/shuf gain terms while retaining prediction supervision. No subset removes $\lambda_1 \ell_{\mathrm{sub}}$ while retaining the main training samples and their modality combinations. These variants isolate staged optimization, relative gain supervision, and modality-subset supervision, respectively.

Component variants. No query sets $q_i = 0$ only in historical reading weights, retaining modality keys, historical keys and values, and query-conditioned Event candidate pooling. It therefore tests query dependence in historical reading specifically.

Private influence replaces both shared W_V, P with modality-specific, bias-free $W_V^m \in \mathbb{R}^{r \times d}$ and $P_{\mathrm{out}}^m \in \mathbb{R}^{d \times r}$. Using the modality-conditioned historical summary ρ_i^m from Appendix D.1, it forms and combines predictive influences as

$$e_i^m = W_V^m \rho_i^m, \qquad a_i = \frac{\sum_{m \in \mathcal{A}_i} P_{\mathrm{out}}^m e_i^m}{\sqrt{\max(|\mathcal{A}_i|, 1)}}, \qquad h_i^\star = h_i \odot \exp(\beta \tanh(a_i)). \tag{14}$$

Each modality learns in its own r-dimensional influence space and projects into the forecasting channels before aggregation and bounded modulation. The output projection P_{out}^m differs from the modality-key projection P_m in Section 4.2. Branches start from the shared projections and then learn independently. The variant adds $2(M-1)dr$ parameters relative to MILU, or 24,576 for the default four-modality M model with $M = 4, d = 128, r = 32$.

No content zeros modality keys before historical reading and Event candidate pooling, making the latter uniform over valid candidates. It retains modality availability, forecasting queries, historical keys and values, and the aggregation and modulation paths. This removes content information while preserving availability information. The on/shuf term is zero because no content association changes; the remaining loss terms and coefficients are unchanged.

C.6 Modality contributions and held-out combination results

Modality contributions. We compare off, individual modalities, and all available modalities under zero-shot and 10% target-data adaptation. Each setting uses fixed weights and the same 19 multimodal tasks and test samples. Unavailable modalities contribute the same-checkpoint off error. For each metric, we first compute the geometric mean of task errors with equal task weights, then calculate the gain as $100(1 - E_{\mathrm{on}}/E_{\mathrm{off}})$. Each setting uses its own same-weight off reference. For zero-shot Long, the off errors are 9.648/18.472 and the all-modality errors are 9.262/18.373; Table 19 reports gains and adaptation errors.

Figure 7 shows Long MAE gains for individual and joint modality use. Each modality improves forecasting, while joint use yields the largest aggregate MAE gain under both zero-shot evaluation and target adaptation. Short adaptation results also support this benefit from joint use.

Combination holdout. We retain the main experiment's training, validation, and test splits and the same 19 multimodal evaluation tasks. Starting from the pretrained spatio-temporal backbone and prediction head, we initialize fresh trainable multimodal modules and use the full-MILU pretraining procedure. Among Road (A), Remote (B), and Weather (C), training uses A+C and B+C with individual-modality subset supervision. Event information is added independently when available, without changing the restriction that A and B are never jointly enabled. Validation follows the same combination restrictions.

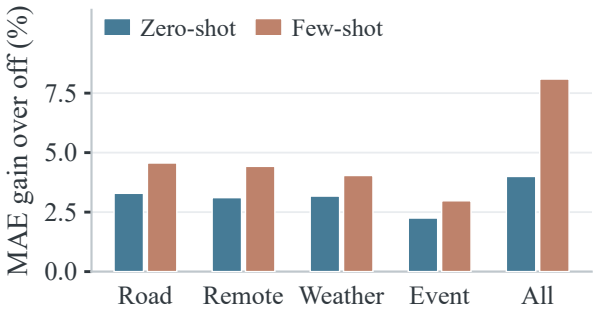

Figure 7: Individual and joint modality contributions.

Table 19: Individual and joint modality contributions. Zero-shot results report relative gains; few-shot results report forecasting errors.

Input / horizon	**MAE gain (%)**	**RMSE gain (%)**
Zero-shot		
Zero-shot: off / Long	0.000	0.000
Zero-shot: Road / Long	3.291	0.401
Zero-shot: Remote / Long	3.108	0.454
Zero-shot: Weather / Long	3.174	0.485
Zero-shot: Event / Long	2.251	0.300
Zero-shot: All available / Long	3.999	0.535
Few-shot		
Input / horizon	MAE	RMSE
off / Short	7.130	14.157
off / Long	8.023	15.665
Road / Short	6.727	13.878
Road / Long	7.657	15.240
Remote / Short	6.860	13.901
Remote / Long	7.668	15.191
Weather / Short	6.835	13.948
Weather / Long	7.699	15.232
Event / Short	6.891	13.899
Event / Long	7.784	15.332
All available / Short	6.600	13.218
All available / Long	7.374	14.630

Testing fixes the weights and disables Event, comparing off, A, B, C, A+C, B+C, A+B, and A+B+C on the same complete set of test samples. Specified modalities are enabled through the original availability masks; tasks and samples are retained even when some or all specified modalities are unavailable. Thus, tasks with only Event information, such as BjTT, use spatio-temporal inputs under every condition. All conditions, including off, use identical task weights and the checkpoint trained for this combination-holdout experiment. Comparing A+B with off and its individual modalities tests new pairings; comparing A+B+C with all pairs tests extension to three modalities. Figure 4 and Table 20 report the results.

Table 20: Forecasting performance with seen and unseen modality combinations.

Input	Role	Short: MAE / RMSE	Long: MAE / RMSE
off	Reference	7.796 15.009	9.650 18.470
Road	Reference	7.570 14.946	9.402 18.393
Remote	Reference	7.642 14.991	9.441 18.456
Weather	Reference	7.688 14.988	9.458 18.465
Road + Weather	Seen pair	7.511 14.905	9.364 18.393
Remote + Weather	Seen pair	7.623 14.964	9.429 18.458
Road + Remote	Unseen pair	7.534 14.901	9.387 18.370
Road + Remote + Weather	Unseen triple	7.485 14.903	9.353 18.369

C.7 Multimodal transfer results and reporting

Paired evaluation and gains. We compare predictions with all available multimodal information (on) and ST-only predictions (off) using pretrained weights for zero-shot evaluation and weights adapted with 10% of target training data for few-shot evaluation. Within each pair, model weights, target locations, history, test origins, and valid-target masks remain fixed. Both conditions retain the same test samples, including those with incomplete multimodal information. For a fixed task set, let E_{on} and E_{off} denote the aggregate MAE or RMSE under the two conditions. The relative gain is $G_{\mathrm{on/off}} = 100(1 - E_{\mathrm{on}}/E_{\mathrm{off}})$; positive values indicate lower error with multimodal information.

Following Figure 3, Flow, Occupancy, Speed, and Congestion contain 8, 6, 4, and 1 tasks, respectively. Overall aggregates all 19 tasks with equal task weights before computing the gain. Gains are computed from paired geometric means rather than averaging group percentages.

Interpreting comparisons across tasks. For positive task errors E_k (MAE or RMSE), aggregate errors for models A and B satisfy

$$\frac{\mathrm{GeoE}_A}{\mathrm{GeoE}_B} = \exp\Big(\frac{1}{K}\sum_{k=1}^{K}\log\frac{E_{A,k}}{E_{B,k}}\Big). \tag{15}$$

The same proportional error reduction on one task therefore changes the log aggregate independently of sample count or the original error magnitude. This represents relative performance with equal task weights (Fleming & Wallace, 1986). A positive unit-conversion factor applied to all models on a task cancels in the score ratio; the absolute Geo score itself is not dimensionless. Scores from different test pools are not directly comparable.

Figure 3 groups Long/MAE gains by prediction variable; Table 21 complements it with Short and RMSE results using the same prediction-variable groups.

For Short zero-shot occupancy, MAE improves by 1.351% while RMSE increases by 3.195%.

C.8 Hyperparameter analysis

We compare three values for each of six hyperparameters using validation MAE during early training. Table 22 reports relative validation MAE changes with respect to the selected value for each parameter, $\Delta E = 100(E_{\mathrm{candidate}}/E_{\mathrm{reference}} - 1)$; positive values indicate increased error. The selected configuration is used for full pretraining.

Table 21: Forecasting errors and multimodal gains by prediction variable.

Condition / horizon	Target pool	off: MAE / RMSE	on: MAE / RMSE	Gain (%): MAE / RMSE
Zero-shot / Short	Flow (8)	48.765 78.068	47.813 76.329	1.952 2.228
Zero-shot / Short	Occupancy (6)	2.119 4.914	2.090 5.071	1.351 -3.195
Zero-shot / Short	Speed (4)	4.319 7.623	3.937 7.474	8.839 1.952
Zero-shot / Short	Congestion (1)	0.076 0.327	0.073 0.323	3.947 1.223
Zero-shot / Short	Overall (19)	7.738 14.974	7.478 14.909	3.362 0.434
Zero-shot / Long	Flow (8)	63.955 104.252	62.515 103.630	2.251 0.596
Zero-shot / Long	Occupancy (6)	2.581 5.850	2.540 5.848	1.584 0.031
Zero-shot / Long	Speed (4)	5.105 8.674	4.546 8.562	10.944 1.299
Zero-shot / Long	Congestion (1)	0.090 0.366	0.087 0.366	3.333 0.000
Zero-shot / Long	Overall (19)	9.648 18.472	9.262 18.373	3.999 0.535
Few-shot / Short	Flow (8)	44.157 68.477	40.227 63.913	8.899 6.665
Few-shot / Short	Occupancy (6)	2.008 5.026	1.898 4.715	5.484 6.180
Few-shot / Short	Speed (4)	3.822 7.220	3.604 6.874	5.697 4.793
Few-shot / Short	Congestion (1)	0.080 0.349	0.069 0.294	13.729 15.784
Few-shot / Short	Overall (19)	7.130 14.157	6.600 13.218	7.427 6.627
Few-shot / Long	Flow (8)	49.127 78.490	44.987 73.090	8.426 6.880
Few-shot / Long	Occupancy (6)	2.297 5.379	2.165 5.160	5.729 4.083
Few-shot / Long	Speed (4)	4.317 8.026	3.870 7.313	10.352 8.892
Few-shot / Long	Congestion (1)	0.088 0.349	0.079 0.314	10.227 10.029
Few-shot / Long	Overall (19)	8.023 15.665	7.374 14.630	8.092 6.606

The selected loss weights, gain margin, and modulation bound yield the lowest validation MAE during early training among their tested values. Increasing the influence dimension from 32 to 64 reduces error by only 0.035%, whereas reducing it to 16 increases error by 1.972%. We therefore retain $r = 32$ as a compact choice.

C.9 Computing efficiency

Figure 8 compares forecasting accuracy, parameter count, and inference latency.

Table 22: Hyperparameter sensitivity during early training. Selected values are bold; Δ denotes relative validation MAE change (%).

Parameter	Value	Δ	Selected	Δ	Value	Δ
λ_1	0.05	+1.824	**0.2**	0.000	1	+0.953
λ_2	0.1	+1.725	**1**	0.000	10	+4.360
δ	0.005	+1.867	**0.01**	0.000	0.05	+4.916
β	0.1	+1.140	**0.2**	0.000	0.4	+3.280
MSE weight	0.05	+0.853	**0.2**	0.000	1	+4.619
r	16	+1.972	**32**	0.000	64	-0.035

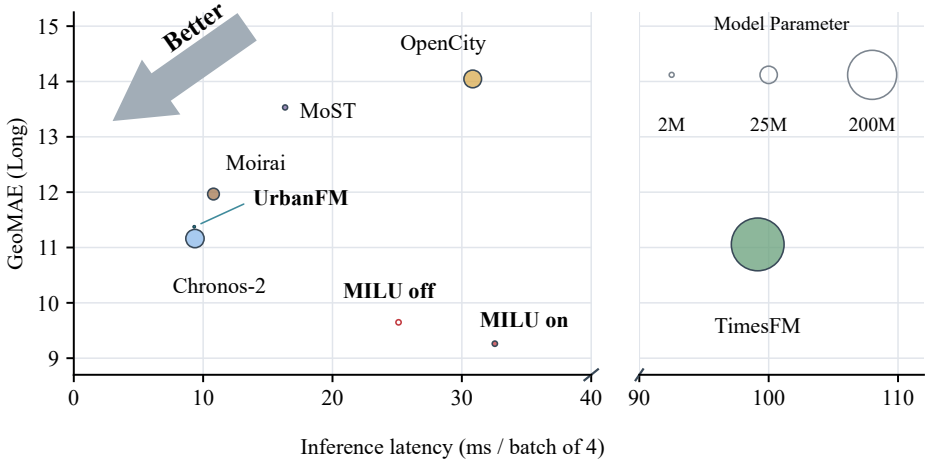

Figure 8: Forecasting accuracy and efficiency. Bubble area represents forecasting parameters; MILU off uses ST-only inputs. Latency uses cached multimodal features. The horizontal axis omits 40–90 ms.

Measurement protocol. We measure inference on one NVIDIA A100-SXM4-80GB using batches of four requests, each with 24 historical steps, 24 forecast steps, 16 target spatial units, and one target variable. Each call returns 1,536 forecasts. We report the median and 10th–90th percentiles over 30 calls after 10 warm-up calls, synchronizing CUDA before and after each call. Timing includes transfer of prepared CPU inputs to the GPU, model inference, inverse normalization, target-unit selection, and return of predictions to the CPU. Data loading, batch preparation, and frozen image/text encoding are performed before timing; multimodal features are cached.

Model configurations. Measurements use the forecasting checkpoints and native adapters corresponding to the reported results. MILU on/off use the same checkpoint, with multimodal information enabled or disabled. MoST and UrbanFM use the evaluated reproductions; the released models are OpenCity Plus, Chronos-2 Small, TimesFM 2.5 (200M), and Moirai 2.0-R Small. MILU, MoST, and Chronos-2 use BF16 autocast with FP32 weights; UrbanFM uses its native BF16 inference path with stability fallback. OpenCity, TimesFM, and Moirai use FP32. TF32 is disabled, and all models run in evaluation and inference mode with two CPU threads and without compilation. Native spatial support and internal padding are retained. TimesFM uses its point-forecast path with internal batch/context/output sizes of 512/32/128, clearing prediction memoization before each call.

Parameter and memory accounting. Forecasting parameters include all loaded prediction modules. MILU on/off therefore have identical parameter counts. The frozen image and text encoders used by MILU and MoST contain 223.357M additional parameters and produce cached features offline. Peak memory is PyTorch peak allocated GPU memory in an isolated, warmed inference process. Table 23 reports parameter counts, inference latency, peak memory, and Long GeoMAE on the 19 multimodal tasks, consistent with the main results.

Table 23: Forecasting accuracy and computational efficiency. Parameter counts are in millions; latency is median [10th, 90th percentile] in milliseconds per batch.

Model	Parameters (M)	GeoMAE	Latency (ms)	Peak memory (MiB)
MoST	2.086	13.531	16.34 [16.28, 16.46]	26.47
UrbanFM	0.400	11.375	9.31 [9.26, 9.37]	19.51
OpenCity	25.975	14.046	30.84 [30.82, 30.88]	213.87
Chronos-2	27.935	11.162	9.37 [9.32, 9.44]	122.07
TimesFM	231.289	11.056	99.15 [98.98, 99.41]	1043.23
Moirai	11.387	11.966	10.80 [10.72, 10.85]	56.92
MILU off	2.311	9.648	25.11 [25.01, 25.57]	67.15
MILU on	2.311	**9.262**	32.55 [32.46, 33.12]	73.88

D Theoretical Analysis of Shared Influence Modeling

D.1 Structural factorization and the response model

Our analysis isolates the shared-response structure of MILU through a linear statistical model. In MILU, modality-conditioned retrieval forms input-dependent summaries, which share the mapping PW_V before nonlinear modulation and prediction. The model below captures this response sharing to study estimation and aggregation errors.

Pooling a predictive response connects to the statistical benefit of shared structure in multitask representation learning (Maurer et al., 2016).

In the shared influence module, write $\bar{z}_{i,t} = \mathrm{LN}_s(z_{t,n(i)})$. Linearity of the value projection gives

$$e_i^m = \sum_t \alpha_{i,t}^m W_V \bar{z}_{i,t} = W_V \rho_i^m, \qquad \rho_i^m = \sum_t \alpha_{i,t}^m \bar{z}_{i,t}. \tag{16}$$

For Event, let ω_{ij} be the query-conditioned pooling weight of a valid candidate and $\alpha_{ij,t}$ its historical reading weight. Define $\rho_i^{\mathrm{Event}} = \sum_j \omega_{ij} \sum_t \alpha_{ij,t} \bar{z}_{i,t}$; the same equality holds. Hence

$$\bar{\rho}_i = \frac{\sum_{m \in \mathcal{A}_i} \rho_i^m}{\sqrt{\max(|\mathcal{A}_i|, 1)}}, \qquad a_i = PW_V \bar{\rho}_i. \tag{17}$$

This yields the shared response $W = PW_V$ with $\mathrm{rank}(W) \le r$, which generates the input to bounded modulation. An empty set produces a zero summary and prediction from spatio-temporal observations alone.

Consider $M \ge 2$ modalities with content $c_m = A_m \xi_m$, where latent representations $\xi_m \sim \mathcal{N}(0, I_p)$ are independent and orthogonal A_m describe modality-specific transforms. The centered forecasting residual is $\zeta = \mu + \varepsilon$, with signal $\mu = \vartheta^\top \sum_m \pi_m \xi_m$, shared response $\vartheta \in \mathbb{R}^p$, and independent noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$. Fixed aggregation weights satisfy $\pi_m > 0$ and $\sum_m \pi_m^2 = 1$.

The predictors use concatenated content (C), fixed-weight content aggregation $v_A = \sum_m \pi_m c_m$ (A), or shared influences $v_U = \sum_m \pi_m L_m c_m$ (U), where L_m maps content into the shared influence space. All fit least squares without an intercept on the same $n > Mp + 1$ independent samples. Risk $\mathcal{E}_F = \mathbb{E}[(\widehat{f}_F - \mu)^2]$, $F \in \{C, A, U\}$, averages error relative to signal μ over training and independent test samples.

This appendix uses full method names as subscripts: $U = \mathrm{Influence}$, $A = \mathrm{Attn}$, and $C = \mathrm{Concat}$; $n_{\mathrm{tr}} = n$, $B_{\mathrm{Influence}} = B_U$, and $B_{\mathrm{Attn}} = B_A$.

The variables c_m, A_m, L_m defined above form a statistical abstraction at the response level: private content coordinates undergo modality-conditioned influence transformations and invoke the same

response in common coordinates. Orthogonal A_m give modalities equal scales and condition numbers, focusing the comparison on task directions in their coordinates. The common ϑ represents a reusable predictive rule. For urban forecasting, this corresponds to different information selecting distinct historical dynamics while reusing the relationship between those dynamics and forecasting adjustments. Here, p is the statistical influence dimension, and $\mathrm{rank}(PW_V)$ is the rank of an actual linear operator.

Concat retains all content blocks and estimates their responses separately. Attn aggregates content values with fixed attention weights; the product of a single shared linear value projection and a linear predictor can be absorbed into w. The common weight scale can be absorbed into the predictor to obtain $\sum_m \pi_m^2 = 1$. Fixed aggregation weights focus the analysis on representation coordinates and predictive responses. Corollary 1 conditions on representations learned from independent pilot data; D.4 numerically verifies the attainability of its gain condition.

D.2 Proof of Theorem 1

Theorem 1 (Prediction gains from shared influences). If $L_m = A_m^\top$, let $\bar{A} = \sum_m \pi_m^2 A_m$ and define the content-aggregation approximation error $B_A = \|\vartheta\|^2 - \|\bar{A}\vartheta\|^2$. Then

$$\mathcal{E}_U = \frac{\sigma^2 p}{n-p-1}, \qquad \mathcal{E}_C = \frac{\sigma^2 Mp}{n-Mp-1}, \qquad \mathcal{E}_A = B_A + \frac{(\sigma^2 + B_A)p}{n-p-1}. \tag{18}$$

Response sharing reduces the estimation dimension from Mp to p, lowering variance; when $B_A > 0$, shared influences also eliminate the approximation error from direct content aggregation.

Gaussian regression risk lemma. Let a feature $v \in \mathbb{R}^k$ and the target conditional mean μ have a joint zero-mean Gaussian distribution, with $\Sigma = \mathbb{E}[vv^\top] \succ 0$ and $g = \mathbb{E}[v\mu]$. For $\zeta = \mu + \varepsilon$ with independent $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, fit least squares without an intercept using $n_{\mathrm{tr}} > k+1$ independent samples. Define

$$w^* = \Sigma^{-1} g, \qquad B = \mathbb{E}[\mu^2] - g^\top \Sigma^{-1} g. \tag{19}$$

Then an independent test input from the same distribution satisfies

$$\mathbb{E}[(\widehat{w}^\top v - \mu)^2] = B + (\sigma^2 + B)\frac{k}{n_{\mathrm{tr}} - k - 1}. \tag{20}$$

Proof. Gaussian conditioning gives $\mu = w^{*\top} v + \eta$, where η is independent of v, has zero mean, and has variance B. The regression residual $\eta + \varepsilon$ therefore has variance $B + \sigma^2$. Writing the training design as X, we have $\mathbb{E}[\widehat{w} \mid X] = w^*$ and

$$\mathrm{Cov}(\widehat{w} \mid X) = (\sigma^2 + B)(X^\top X)^{-1}. \tag{21}$$

Whiten the design as $X = G\Sigma^{1/2}$, where entries of G are independent standard normal variables. Let g_j be its column j and H_{-j} the projection onto the span of the remaining columns. The Schur complement gives

$$[(G^\top G)^{-1}]_{jj} = \frac{1}{g_j^\top (I - H_{-j}) g_j}. \tag{22}$$

Conditional on the remaining columns, the denominator has a chi-squared distribution with $n_{\mathrm{tr}} - k + 1$ degrees of freedom. Integrating its density gives $\mathbb{E}[U^{-1}] = 1/(\nu - 2)$ for $U \sim \chi^2_\nu$ and $\nu > 2$. Off-diagonal expectations are zero by column-sign symmetry, so

$$\mathbb{E}[(X^\top X)^{-1}] = \frac{\Sigma^{-1}}{n_{\mathrm{tr}} - k - 1}. \tag{23}$$

The independent test residual has zero expected cross term. The risk is thus $B + (\sigma^2 + B)\,\mathrm{tr}(\Sigma\mathbb{E}[(X^\top X)^{-1}])$, proving the lemma.

Applying the lemma to the three representations. Concat uses the Mp-dimensional concatenation of all c_m, whose covariance is I_{Mp}. Its true coefficients are $b_m^* = \pi_m A_m \vartheta$, giving $B = 0$ and the stated $\mathcal{E}_{\mathrm{Concat}}$. Ideal common influence coordinates yield $v_{\mathrm{Influence}} = \sum_m \pi_m \xi_m$, with covariance I_p and true response ϑ. Hence $B = 0$, giving $\mathcal{E}_{\mathrm{Influence}}$.

Content attention uses $v_{\mathrm{Attn}} = \sum_m \pi_m A_m \xi_m$. Modality independence, orthogonality, and weight normalization imply

$$\mathbb{E}[v_{\mathrm{Attn}} v_{\mathrm{Attn}}^\top] = I_p, \qquad \mathbb{E}[v_{\mathrm{Attn}}\mu] = \bar{A}\vartheta, \qquad \mathbb{E}[\mu^2] = \|\vartheta\|^2. \tag{24}$$

Thus $B = \|\vartheta\|^2 - \|\bar{A}\vartheta\|^2$. Expanding the weighted squared deviations and using $\sum_m \pi_m^2 \|A_m\vartheta\|^2 = \|\vartheta\|^2$ yields B_{Attn} in the theorem. The lemma gives the third risk formula. Since $k/(n_{\mathrm{tr}} - k - 1)$ strictly increases with k, $M \geq 2$, and $\sigma^2 > 0$, shared influence has lower risk than Concat. Its risk reduction over Attn is $B_{\mathrm{Attn}}(n_{\mathrm{tr}} - 1)/(n_{\mathrm{tr}} - p - 1)$, completing the proof. Evaluating mean squared error against noisy test labels ζ adds the same σ^2 to all three risks.

D.3 Coordinate-learning error and proof of Corollary 1

Corollary 1 (Prediction gains with learned influences). Under this model, condition on linear maps L_m learned on a common independent pilot set and then fixed, with positive-definite covariance for v_U. Let $B_U = \min_w \mathbb{E}[(\mu - w^\top v_U)^2]$ be their predictive approximation error. Shared influences have strictly lower risk than both alternatives if and only if

$$B_U < T(n) := \min\left\{ B_A, \frac{\sigma^2(M-1)p}{n - Mp - 1} \right\}. \tag{25}$$

This threshold specifies the representation error tolerated while retaining both gains. The proof follows below, and numerical verification is provided in Appendix D.4.

Condition on maps L_m learned from independent pilot data and write $D_m = L_m A_m$. The aggregated influence $v_{\mathrm{Influence}} = \sum_m \pi_m D_m \xi_m$ satisfies

$$\Sigma_{\mathrm{Influence}} = \sum_m \pi_m^2 D_m D_m^\top \succ 0, \qquad g_{\mathrm{Influence}} = \sum_m \pi_m^2 D_m \vartheta. \tag{26}$$

The lemma gives its optimal linear approximation error and response-estimation risk:

$$\begin{aligned} B_{\mathrm{Influence}} &= \|\vartheta\|^2 - g_{\mathrm{Influence}}^\top \Sigma_{\mathrm{Influence}}^{-1} g_{\mathrm{Influence}}, \\ \mathcal{E}_{\mathrm{Influence}} &= B_{\mathrm{Influence}} + \frac{(\sigma^2 + B_{\mathrm{Influence}})p}{n_{\mathrm{tr}} - p - 1}. \end{aligned} \tag{27}$$

The best predictor incurs no more error than the candidate predictor ϑ. Independence and zero means give

$$\begin{aligned} B_{\mathrm{Influence}} &\leq \mathbb{E}[(\mu - \vartheta^\top v_{\mathrm{Influence}})^2] \\ &= \sum_m \pi_m^2 \|(D_m - I_p)^\top \vartheta\|^2 = \delta. \end{aligned} \tag{28}$$

Direct subtraction yields

$$\begin{aligned} \mathcal{E}_{\mathrm{Attn}} - \mathcal{E}_{\mathrm{Influence}} &= \frac{n_{\mathrm{tr}} - 1}{n_{\mathrm{tr}} - p - 1}(B_{\mathrm{Attn}} - B_{\mathrm{Influence}}), \\ \mathcal{E}_{\mathrm{Concat}} - \mathcal{E}_{\mathrm{Influence}} &= \frac{n_{\mathrm{tr}} - 1}{n_{\mathrm{tr}} - p - 1}\left[\frac{\sigma^2(M-1)p}{n_{\mathrm{tr}} - Mp - 1} - B_{\mathrm{Influence}}\right]. \end{aligned} \tag{29}$$

The common prefactor of both risk differences is positive, so $B_U < T(n)$ in Corollary 1 is necessary and sufficient for both strict gains. Bounding B_U by δ further gives the stronger sufficient coordinate-error condition $\delta < T(n)$. The quantity B_U measures optimal linear approximation error in the learned representation, includes learning or modeling error in the influence transformation, and is invariant to invertible common changes of coordinates. Ideal common coordinates may also use any common invertible transformation: if every $D_m = R$ with invertible R, the predictor $R^{-\top}\vartheta$ gives $B_{\text{Influence}} = 0$. Common coordinates allow modalities to produce different influences within the same sample.

Size of the effective range. Write $D = Mp$ and $\Delta d = (M-1)p$. Normalizing by noise variance gives

$$\frac{B_U}{\sigma^2} < \min\{\frac{B_A}{\sigma^2}, \frac{\Delta d}{n - D - 1}\}. \tag{30}$$

For signal variance $S = \|\vartheta\|^2 > 0$, define $\rho_U = B_U/S$, $\rho_A = B_A/S$, and $\text{SNR} = S/\sigma^2$. Equivalently, $\rho_U < \min\{\rho_A, \Delta d/[\text{SNR}(n - D - 1)]\}$. This places retained predictive influences, content-aggregation error, and response-estimation signal-to-noise ratio on a common scale. For fixed learned representations and $B_U > 0$, simultaneous gains hold exactly when

$$B_U < B_A, \qquad D + 1 < n < D + 1 + \frac{\sigma^2 \Delta d}{B_U}. \tag{31}$$

If $B_U = 0$ and $B_A > 0$, every $n > D + 1$ satisfies the condition. Increasing the response-training sample size reduces Concat's repeated-estimation cost, so the tolerance for a fixed nonzero B_U shrinks with n. More accurate influence transformations reduce B_U and expand the favorable range. The other tolerance, content-aggregation error B_A, is independent of response-training sample size.

D.4 Numerical verification of the corollary condition

Under the Gaussian response model in Appendix D.1, we examine whether finite noisy supervision can learn shared influence representations satisfying the corollary condition. We first learn the mappings into the shared influence space from independent samples, then fix the representations and fit shared-influence, concatenated-content, and fixed-weight content-aggregation predictors on the same training data. We compare their prediction risks relative to the conditional mean on an independent test distribution.

Across the generated scenarios, the approximation error of the learned representations satisfies the corollary condition. Mean prediction risks are 0.132 for shared influences, 0.231 for concatenated content, and 0.800 for fixed-weight content aggregation; shared influences attain lower mean risk in every scenario. These results show that finite noisy supervision can learn representations satisfying the condition and realize both gains predicted by the theory.