CROSSCITYBENCH: A DIAGNOSTIC BENCHMARK FOR CROSS-CITY SPATIO-TEMPORAL GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-city spatio-temporal forecasting asks whether source-city information improves prediction when a target city provides limited observations and follows a different distribution. Existing evidence is difficult to compare because studies vary source-target pairs, target-data budgets, forecasting protocols, and reported metrics. CrossCityBench evaluates target-only learning, alignment, meta-learning, pre-training, knowledge distillation, and federated learning under matched target information and forecasting conditions. It first tests whether source access improves on the strongest target-only reference, then examines whether the recorded ordering persists across computational cost, missing-data sensitivity, city-pair sensitivity, Spatio-Temporal Pattern Bank (STPB) alignment, and federated collaboration cost. Under two core tasks and a matched three-day target-information protocol, D2MHyper and CrossST are the only evaluated transfer methods whose recorded average Mean Absolute Error (MAE) is below the strongest target-only value on both tasks. Accuracy and sensitivity produce different empirical orderings, so in-domain error alone is insufficient for selection. STPB is an optional prototype-alignment diagnostic, not an explanation of learned representations. The resulting guide ties each candidate to measured evidence and remains restricted to the evaluated traffic settings. CrossCityBench provides a falsifiable protocol for making transfer claims comparable and diagnostically traceable. Code, configurations, and data-processing resources are available at `https://anonymous.4open.science/r/CrossCityBench-87CB`.

1 INTRODUCTION

Spatio-temporal forecasting is often optimized for one city or sensor network, yet performance can decline in a target region with limited observations, different dynamics, or a non-aligned graph. Cross-city learning addresses this by transferring representations or model knowledge from data-rich sources. Existing approaches include distribution alignment (Tang et al., 2022; Ouyang et al., 2025), meta-learning (Finn et al., 2017; Lu et al., 2022), pre-training (Liu & Zhang, 2025; Liu et al., 2025c), knowledge distillation (Zheng et al., 2024), and federated learning (Wang et al., 2026; Yang et al., 2024). Because these approaches make different assumptions about transferable structure, their reported gains are not comparable unless the evaluation protocol is also controlled.

The central difficulty is evaluation inconsistency. Published results often use different source-target pairs, target-data budgets, prediction horizons, and training procedures (Bi et al., 2025; Liu & Zhang, 2025). A method can therefore appear effective because of its protocol rather than its transfer mechanism. Accuracy alone introduces a second ambiguity (Salaudeen et al., 2025). A model may achieve low average error while remaining sensitive to missing observations, cross-domain variation, or temporal concept drift (Zhao et al., 2026). Conversely, a method that is not the most accurate may provide a more economical operating point. These ambiguities prevent a reliable answer to a basic learning question: when does a source-enabled implementation outperform learning only from the target region? Figure 1 summarizes this setting and the coexistence of urban heterogeneity and transferable structure.

CrossCityBench uses a comparability contract that matches target information, horizons, and reporting rules. The benchmark records descriptive evidence of transfer utility only when a complete source-enabled implementation has lower target-test error than the strongest evaluated target-only reference

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.

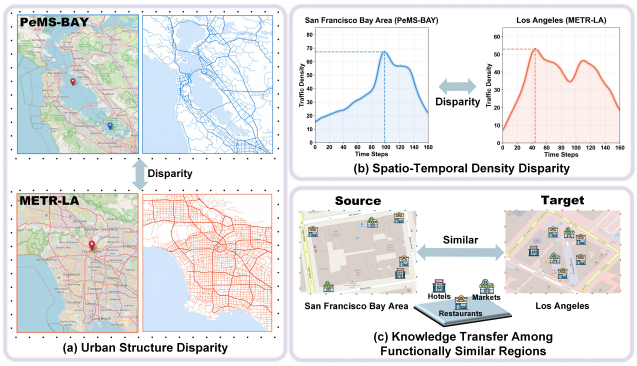


Figure 1: Cross-city forecasting under urban heterogeneity. Structural and temporal differences create source-target shifts, while functionally related regions may still share transferable patterns.

under the same target-information budget. Secondary diagnostics test whether that ordering persists under sensitivity and resource constraints. This fixed-seed criterion does not isolate the causal contribution of source information from method-specific operations.

The contributions are as follows:

- We define a matched comparability contract for five centralized learning paradigms and a separately declared data-local protocol for federated learning. Within each comparison family, target information, forecasting conditions, and reference models are controlled.
- We evaluate predictive utility, missingness and pair sensitivity, representation alignment, computational cost, and federated collaboration cost, with an explicit interpretation boundary for each coordinate.
- Under the recorded fixed-seed executions, transfer utility is selective. Only two evaluated transfer implementations, D2MHyper and CrossST, record lower average MAE than the strongest target-only value on both core tasks, while accuracy and sensitivity favor different choices.
- We provide an evidence-traceable selection guide rather than a global ranking, linking each candidate to a measured constraint and its extrapolation boundary.

The framework is extensible, but its empirical conclusions remain limited to the enumerated flow, speed, and multi-source settings.

2 Related Work

2.1 Cross-City Transfer for Spatio-Temporal Forecasting

Cross-city methods encode different assumptions about shared structure. Alignment reduces source-target discrepancy (Tang et al., 2022; Ouyang et al., 2023; Wang et al., 2022; Ouyang et al., 2025; 2024); meta-learning supports limited-data adaptation (Finn et al., 2017; Lu et al., 2022); and source selection limits negative transfer (Bi et al., 2025). Pre-training learns reusable representations (Liu & Zhang, 2025; Liu et al., 2025c; Han et al., 2025), distillation transfers teacher knowledge (Zheng et al., 2024), and federated methods collaborate without pooling raw observations (Wang et al., 2026; Zhang et al., 2024; Yang et al., 2024; Yuan et al., 2025). Different city pairs or target budgets still prevent direct preference claims.

2.2 Benchmarking Beyond Predictive Accuracy

Recent benchmark work questions accuracy-only domain-generalization evaluation (Salaudeen et al., 2025), while recent forecasting work identifies concept drift as a failure source (Zhao et al., 2026). Limited target data and urban shifts can also alter transfer behavior (Bi et al., 2025; Han et al., 2025). CrossCityBench therefore evaluates target-only and centralized-transfer methods under one matched protocol and federated methods under a separate data-local protocol, while reporting prediction,

sensitivity, and resource coordinates independently. Its contribution is controlled diagnostic evaluation rather than a new forecasting backbone.

2.3 Pattern-Based Diagnosis and Claim Scope

Pattern banks support urban continual forecasting (Liu & Zhang, 2026), while robust models address missingness and asynchronous observations (Jang et al., 2026; Wu et al., 2025). These studies motivate richer diagnostics but not a complete explanation from prototype similarity; we interpret STPB jointly with error and sensitivity rather than infer a mechanism from one score.

2.4 Positioning and Scope

Our evaluation differs from a conventional leaderboard in two respects. It asks whether an evaluated transfer implementation records lower error than the strongest target-only baseline, then tests whether that ordering persists under sensitivity and system constraints. Appendices B and C expand the literature boundary and scientific use while keeping the empirical claims proportional to the benchmark coverage.

3 CrossCityBench Design

CrossCityBench organizes cross-city forecasting into a common problem definition, a controlled data module, a method zoo, and a diagnostic evaluator. The components are introduced in this order so that each reported comparison follows from the same task and information budget.

3.1 Problem Formulation

We first formalize the prediction task and the sources of difficulty that the benchmark controls.

Definition 1 (Spatio-Temporal City Graph)**.** *We represent city $\mathcal{C}$ by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$:*

- *$\mathcal{V}$ contains N sensors or geographic regions.*
- *$\mathcal{E}$ records road connectivity or spatial proximity.*
- *$\mathbf{A} \in \mathbb{R}^{N \times N}$ encodes these topological relations.*

This representation supports the forecasting task under asymmetric data availability.

Definition 2 (Cross-City Spatio-Temporal Forecasting Task)**.** *For a data-rich source city $\mathcal{C}^s$ and a data-scarce target city $\mathcal{C}^t$, their temporal observations at time step τ are represented by feature matrices $\mathbf{X}^s_\tau \in \mathbb{R}^{N_s \times F}$ and $\mathbf{X}^t_\tau \in \mathbb{R}^{N_t \times F}$, respectively, where F denotes the feature dimension. With a past observation window spanning the length T_h, the input sequences are defined as:*

$$\mathcal{X}^s = (\mathbf{X}^s_{\tau - T_h + 1}, \mathbf{X}^s_{\tau - T_h + 2}, \ldots, \mathbf{X}^s_\tau) \in \mathbb{R}^{T_h \times N_s \times F}, \tag{1}$$

$$\mathcal{X}^t = (\mathbf{X}^t_{\tau - T_h + 1}, \mathbf{X}^t_{\tau - T_h + 2}, \ldots, \mathbf{X}^t_\tau) \in \mathbb{R}^{T_h \times N_t \times F}. \tag{2}$$

A data-driven mapping $f : \mathcal{X}^t \to \hat{\mathcal{Y}}^t$ is learned, producing the future estimates of length T_p:

$$\hat{\mathcal{Y}}^t = (\hat{\mathbf{X}}^t_{\tau+1}, \hat{\mathbf{X}}^t_{\tau+2}, \ldots, \hat{\mathbf{X}}^t_{\tau+T_p}) \in \mathbb{R}^{T_p \times N_t \times F}. \tag{3}$$

The mapping uses abundant source data $\mathcal{D}^s$ and limited target training data $\mathcal{D}^t_{train}$, and is evaluated on $\mathcal{D}^t_{test}$.

The task contains three coupled learning challenges:

(1) **Data scarcity**: The target city has extremely limited labeled data, i.e., $|\mathcal{D}^t_{\text{train}}| \ll |\mathcal{D}^s|$.

(2) **Distribution shifts**: The underlying data distributions differ between cities, i.e., $P_{\mathcal{C}^s}(\mathcal{X}, \mathcal{Y}) \neq P_{\mathcal{C}^t}(\mathcal{X}, \mathcal{Y})$.

(3) **Structural heterogeneity**: The graph structures are non-aligned, with different numbers of nodes ($N_s \neq N_t$) and potentially different topological characteristics between $\mathbf{A}^s$ and $\mathbf{A}^t$.

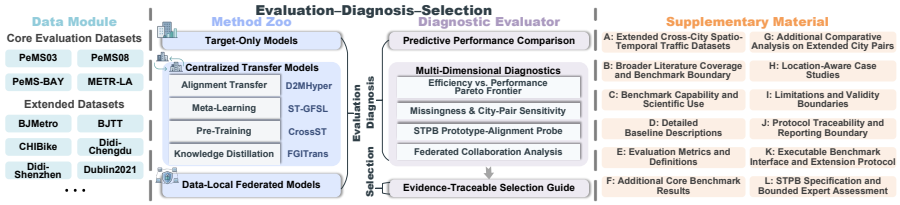


Figure 2: CrossCityBench pipeline. The data module fixes matched tasks, the method zoo instantiates transfer assumptions, and the evaluator tests predictive utility before sensitivity, representation alignment, and cost.

Table 1: The overview of core datasets

Datasets	Tasks	# Nodes	Interval	# Time Steps
PeMS03	Traffic Flow	358	5 min	26,208
PeMS08	Traffic Flow	170	5 min	17,856
PeMS-BAY	Traffic Speed	325	5 min	52,116
METR-LA	Traffic Speed	207	5 min	34,272

3.2 Benchmark Overall Design

CrossCityBench follows protocol consistency, diagnostic coverage, and transparent claim boundaries. Its data module fixes source-target tasks, its method zoo instantiates transfer assumptions, and its evaluator measures prediction, sensitivity, representation alignment, and cost (Figure 2).

We define the *comparability contract* as the target supervision, input and output horizons, chronological split, normalization, target-only reference, nominal resource budget, and reporting metrics held fixed within a comparison. Source interfaces remain paradigm-specific and are disclosed separately.

3.3 Data Module

The core protocol uses four traffic datasets (Table 1). We use *city domain* for a geographically and statistically distinct traffic-sensing domain. PeMS03 $\rightarrow$ PeMS08 is cross-region transfer within one statewide system, whereas PeMS-BAY $\rightarrow$ METR-LA is cross-metropolitan. This controlled scope does not represent full urban heterogeneity.

The controlled comparison uses two source-target scenarios:

- **Traffic flow:** PeMS03 $\rightarrow$ PeMS08. This is a cross-region transfer within the PeMS infrastructure and tests transfer under shared sensing semantics.
- **Traffic speed:** PeMS-BAY $\rightarrow$ METR-LA. This setting changes both the metropolitan region and graph topology, providing a more heterogeneous transfer test.

Appendix A documents the extended dataset collection, while Appendix G reports the evaluated multi-source city pairs. The catalogue indicates available coverage, whereas only the reported experiments support empirical claims.

3.4 Method Zoo

The method zoo contains 23 models organized by learning assumption. Included methods fit the common input-output task, expose source/target access, and provide a protocol-relevant comparison. Target-only models test source utility; alignment, meta-learning, pre-training, and distillation cover centralized transfer, while federated models test data-local collaboration. Table 2 states each assumption and limitation. Unequal family sizes mean results compare implementations, not average paradigm capability. Appendix D records full titles, adaptations, and coverage limits.

3.5 Diagnostic Evaluator

The diagnostic evaluator measures predictive error (MAE, Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE)), computational cost, missingness and pair sensitivity,

Table 2: Taxonomy of learning paradigms in the 23-model benchmark suite. The table states the transferable-structure assumption that each group tests; it is not a claim that all methods in a group perform equivalently.

Paradigm	Guiding principle	Evaluated methods	Potential strength	Limitation
Target-only learning	Learn city-specific dynamics without source transfer	GBRT, VAR, AGCRN, AllDeepSet, DCRNN, DyHSL, GRU, GWNet, STGCN, STG-NCDE	No source-induced bias	Limited target data can constrain estimation
Alignment transfer	Explicitly reduce source-target discrepancy	DASTNet, D2MHyper, DAGN, ST-DAAN	Can handle moderate domain shift	Pair mismatch and alignment cost
Meta-learning	Learn an initialization that adapts from few target observations	MAML, ST-GFSL	Fast adaptation when meta-tasks match	Sensitive to task mismatch and optimization
Pre-training	Learn reusable representations before target adaptation	CrossST, MTPB, STGCN-FT	Broad representation reuse	Requires a suitable pre-training corpus
Knowledge distillation	Transfer teacher knowledge to a compact student	FGITrans	Deployment-oriented compression	Teacher-student capability gap
Federated learning	Collaborate without pooling raw city data	FedCTPM, pFedCTP, FedGTP	Data-local collaboration	Communication and client heterogeneity

STPB alignment, and federated collaboration cost. The coordinates remain separate because aggregation would conceal trade-offs; Appendix E gives definitions and implementation details.

4 Experiments

Experiments first compare source-enabled implementations with target-only learning under matched target access, then test ordering persistence under efficiency, sensitivity, representation alignment, and data-local collaboration. Full federated results, extended city-pair analyses, and location-aware cases appear in Appendices F, G, and H.

4.1 Experimental Protocol

All models use 12 input steps to predict the next 12 after Z-score normalization. For centralized models, the three-day target-information budget contains all target observations available for fitting and checkpoint selection; the released split manifest identifies its chronological training and validation portions, while the held-out target test period is never used for training, tuning, or checkpoint selection. Z-score statistics are computed from the corresponding training partition and applied unchanged to validation and test observations, with source and target domains normalized separately. Missingness diagnostics retain the unmasked training statistics so that masking changes observation availability rather than preprocessing. For each core task, the target-only reference is the lowest value in the reported Average MAE column among the prespecified target-only candidates; this post-evaluation comparator is not used to select models or hyperparameters. Source information denotes any source-derived observation, parameter, prototype, teacher output, initialization, or federated update unavailable to that reference. The contract therefore matches target supervision and task definition rather than identical inputs across paradigms. We use one prespecified random seed and a shared nominal optimization budget where architectures permit. Centralized models use AdamW, learning rate 0.001, batch size 16, hidden dimension 64, early stopping on the declared validation portion, and at most 100 epochs. Within each federated client, observations follow the recorded chronological 6:2:2 training, validation, and test partition; optimization uses Adam with learning rate 0.003, batch size 64, 100 rounds, and one local epoch. Experiments use Linux, an Intel Xeon Gold Central Processing Unit (CPU), a Tesla V100 Graphics Processing Unit (GPU), and PyTorch 1.10. The protocol controls target information and nominal resources but neither individually optimizes every architecture nor supports statistical claims from small fixed-seed gaps. Appendix J specifies the information-access and reporting contract.

Table 3: Lowest recorded average MAE among evaluated implementations in each centralized paradigm. A slash denotes different models across tasks.

Paradigm	Representative model(s)	Flow	Speed
Target only	STG-NCDE/DyHSL	18.62	3.82
Alignment	D2MHyper	17.54	**2.74**
Meta-learning	ST-GFSL/MAML	23.75	4.90
Pre-training	CrossST	**14.67**	3.21
Distillation	FGITrans	14.76	3.97

Table 4: Predictive performance on cross-city traffic flow data (PeMS03 → PeMS08) under the three-day target-data budget.

Methods (Paradigms)	PeMS08 (Target Domain)											
	15 min			30 min			60 min			Average		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
Target-Only Learning (Paradigm 1)												
GBRT	27.11	44.28	16.41	29.35	46.93	17.79	34.12	52.76	21.05	29.68	47.37	18.09
VAR	30.04	44.90	18.38	32.37	48.58	19.85	37.83	56.31	23.81	32.86	49.16	20.33
AGCRN	26.48	46.25	13.59	26.65	46.92	13.61	32.49	52.35	17.22	28.03	48.09	14.50
AllDeepSet	19.90	29.33	13.66	25.37	34.25	37.15	36.84	52.75	25.88	26.72	38.65	19.82
DCRNN	17.06	26.46	12.71	20.11	31.26	14.29	26.61	40.51	19.51	20.53	31.66	14.82
DyHSL	16.87	25.46	12.70	18.95	29.44	14.09	23.45	35.98	17.05	19.16	29.64	14.71
GRU	23.79	33.02	24.61	33.34	42.95	38.50	37.22	51.33	45.44	30.02	40.80	34.02
GWNet	21.03	29.92	20.19	25.51	36.89	22.65	36.30	51.17	29.62	26.56	38.07	23.83
STGCN	18.85	29.14	12.87	23.65	36.55	15.60	33.31	51.41	20.49	24.61	38.00	15.95
STG-NCDE	16.94	25.59	13.34	18.02	28.71	14.32	22.31	35.11	17.02	18.62	29.41	14.72
Alignment-Based Transfer (Paradigm 2)												
DASTNet	17.79	26.65	13.03	20.64	31.09	14.56	27.14	40.02	18.80	21.15	31.98	15.47
D2MHyper	15.34	23.49	12.01	17.00	26.29	12.75	21.69	33.37	16.79	17.54	26.98	13.48
DAGN	16.83	24.53	12.57	17.73	26.86	14.03	21.78	33.88	17.20	18.36	27.95	14.56
ST-DAAN	18.33	27.37	12.27	21.98	32.99	15.19	29.33	43.47	19.78	22.44	33.63	15.23
Meta-Learning-Based Transfer (Paradigm 3)												
MAML	20.20	28.61	19.33	24.30	33.70	25.05	32.65	44.25	36.49	24.89	34.40	26.11
ST-GFSL	19.71	28.35	16.00	23.41	33.18	19.53	30.28	42.48	27.52	23.75	33.64	20.25
Pre-Training-Based Transfer (Paradigm 4)												
CrossST	**13.68**	**21.79**	**8.81**	**14.65**	**23.60**	**9.63**	16.25	**26.00**	**10.60**	**14.67**	**23.49**	**9.59**
MTPB	21.92	31.69	15.17	24.21	34.71	15.50	27.53	39.75	18.30	24.47	35.39	16.14
STGCN-FT	18.11	27.17	13.99	20.94	31.39	16.83	26.63	39.17	20.09	21.92	32.72	16.77
Knowledge-Distillation-Based Transfer (Paradigm 5)												
FGITrans	14.63	28.29	18.91	14.73	28.58	18.98	**14.93**	29.14	19.10	14.76	28.67	19.00

4.2 Predictive Performance Under Limited Target Data

Table 3 reports the lowest recorded average MAE among the implementations assigned to each centralized paradigm. Unequal group sizes make this a descriptive summary, not an estimate of paradigm-level superiority. Tables 4 and 5 provide the complete model-level results.

The results support conditional rather than automatic transfer utility. Under the recorded executions, only two evaluated transfer implementations, D2MHyper and CrossST, fall below the strongest target-only average MAE on both tasks. FGITrans falls below the target-only reference on flow but remains above it on speed, while the evaluated meta-learning implementations remain above target-only learning in both settings. D2MHyper records the lowest speed MAE, whereas CrossST records the lowest flow MAE and a speed MAE of 3.21, below the target-only reference of 3.82. These are implementation-level observations for the evaluated pairs. In particular, the single FGITrans result cannot support a paradigm-wide distillation claim, and small adjacent gaps are not statistically decisive. Tables 4 and 5 expose the target-only comparator and horizon-level variation behind the summary.

This section reports the full model-level results for the two core forecasting tasks and the federated comparison referenced by the main paper.

Federated learning addresses a different constraint because it replaces centralized raw-data access with data-local collaboration. Tables 9 and 10 in Appendix F show that FedCTPM records the lowest error and smallest MAE loss among the evaluated federated methods relative to the strongest centralized reference, while FedGTP has the lowest recorded mean uploaded tensor payload. These findings motivate different choices for utility-oriented and communication-oriented deployments.

4.3 Multi-Dimensional Diagnostic Evaluation

Predictive error provides the first view of transfer quality. We next examine efficiency, sensitivity, representation alignment, and federated cost to identify where the recorded ordering changes.

Table 5: Predictive performance on cross-city traffic speed data (PeMS-BAY → METR-LA) under the three-day target-data budget.

Methods (Paradigms)	METR-LA (Target Domain)											
	15 min			30 min			60 min			Average		
	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
Target-Only Learning (Paradigm 1)												
GBRT	8.73	16.05	16.57	9.81	18.53	18.59	11.28	20.77	20.93	9.73	18.16	18.38
VAR	8.38	15.32	15.14	9.48	16.93	17.04	11.04	19.28	20.28	9.45	16.84	17.12
AGCRN	5.47	10.86	9.56	6.43	12.94	11.91	8.03	15.56	15.51	6.45	13.05	12.02
AllDeepSet	3.39	6.58	9.14	4.10	8.03	11.21	5.14	10.31	14.87	4.09	8.02	11.33
DCRNN	3.44	6.89	8.94	4.07	8.25	11.67	5.21	9.94	15.54	4.08	8.15	11.72
DyHSL	3.23	6.31	8.32	3.77	7.88	10.74	4.75	9.62	14.03	3.82	7.83	10.76
GRU	3.59	7.11	9.37	4.28	8.75	12.27	5.69	10.46	16.69	4.36	8.58	12.41
GWNet	3.27	6.37	9.88	4.04	7.80	12.28	5.09	9.84	16.68	4.01	7.72	12.45
STGCN	3.40	6.51	9.22	3.99	8.15	11.57	5.14	10.18	14.96	4.04	8.12	11.67
STG-NCDE	3.71	6.87	7.47	4.90	10.25	10.47	6.89	14.06	14.90	4.76	10.10	10.20
Alignment-Based Transfer (Paradigm 2)												
DASTNet	3.72	7.50	9.30	4.57	9.52	12.07	6.01	11.58	16.25	4.59	9.07	12.07
D2MHyper	**2.31**	**4.36**	**6.09**	**2.66**	**5.00**	**6.99**	**3.55**	**7.17**	**10.45**	**2.74**	**5.28**	**7.47**
DAGN	3.04	5.57	7.24	3.29	6.32	8.49	3.93	7.41	10.78	3.33	6.35	8.63
ST-DAAN	3.16	5.98	8.02	3.74	7.56	10.68	4.95	9.47	15.24	3.81	7.53	10.89
Meta-Learning-Based Transfer (Paradigm 3)												
MAML	4.04	7.60	11.82	4.87	9.24	14.88	6.28	10.87	19.15	4.90	9.07	14.90
ST-GFSL	4.02	7.52	11.57	4.85	9.36	14.36	6.42	11.43	18.53	4.91	9.25	14.47
Pre-Training-Based Transfer (Paradigm 4)												
CrossST	2.85	5.55	7.58	3.26	6.58	9.21	3.72	7.58	10.93	3.21	6.40	9.02
MTPB	3.14	5.68	7.59	3.70	7.10	10.00	4.68	8.57	13.17	3.75	7.00	10.08
STGCN-FT	3.41	6.60	8.93	3.90	7.78	11.31	4.89	9.53	15.02	3.94	7.82	11.50
Knowledge-Distillation-Based Transfer (Paradigm 5)												
FGITrans	3.27	6.39	12.07	3.88	7.36	13.54	4.75	8.55	15.04	3.97	7.43	13.55

4.3.1 Efficiency versus Performance Pareto Frontier

Figure 3 compares MAE with inference latency, GPU memory, and model size to test whether accuracy gains remain attractive under resource constraints.

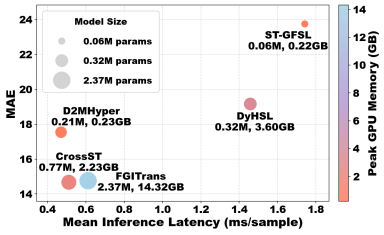


Figure 3: Efficiency versus predictive performance on the PeMS03 → PeMS08 flow task. Lower MAE, latency, memory, and model size are preferable; the panels report separate resource coordinates rather than one aggregate score.

Figure 3 places CrossST and D2MHyper in favorable regions of the observed accuracy-latency plane, without identifying which architectural component causes this position. Latency and peak memory are reported from the same V100 benchmark environment and are treated as descriptive system coordinates rather than hardware-independent estimates. No single paradigm dominates every coordinate: ST-GFSL combines weaker accuracy with high latency, while DyHSL and FGITrans require 3.6–14.3 Gigabytes (GB) of memory for average MAE differences below 1.5 relative to nearby alternatives. FGITrans therefore has favorable latency but high peak memory and is not uniformly efficient.

4.3.2 Sensitivity under Missingness and City-Pair Changes

We diagnose sensitivity to missing target data and city-pair choice. Figure 4(a) reports $\Delta\mathcal{M}_{\text{miss}}$ after randomly masking 10% to 50% of target training observations, while Figure 4(b) reports pair contrast $\Delta\mathcal{M}_{\text{shift}}$.

Figure 4(a) reports changes under 10% to 50% missing training data. CrossST changes by 0.34%–1.44%, D2MHyper by -8.42%–7.60%, ST-GFSL by 0.71%–9.33%, DyHSL from -0.06% to -2.35%, and FGITrans by 3.28%–10.84%. CrossST exhibits a narrow positive change range, whereas FGITrans shows the largest positive sensitivity. Negative values mean that a fixed execution obtained lower MAE after masking; they are treated as fluctuations rather than evidence that missingness improves a model. The preferred method therefore depends on whether nominal accuracy or missingness sensitivity is prioritized.

Here $\Delta\mathcal{M}_{\text{shift}}$ contrasts one designated homogeneous pair with one designated heterogeneous pair. It measures pair sensitivity under the reported construction, not the causal effect of an isolated distribution intervention; lower values indicate less observed degradation.

Figure 4(b) reports pair-contrast changes of 32.52% for D2MHyper, -5.24% for DyHSL, 16.75% for ST-GFSL, 0.28% for CrossST, and 0.55% for FGITrans. CrossST and FGITrans have the smallest positive changes in the evaluated pairs, whereas D2MHyper combines strong nominal speed accuracy with high pair sensitivity. Accuracy and sensitivity therefore support different choices.

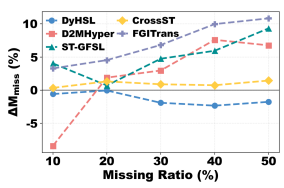
4.3.3 STPB Prototype-Alignment Probe

We diagnose representation alignment using STPB, a learned reference space of spatio-temporal prototypes. Figure 5 reports the average cosine similarity between model representation $\mathbf{r}_{\text{model}}$ and prototype vectors $\mathbf{p}_k$.

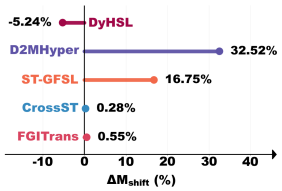
Figure 5 shows different similarities across the evaluated models. D2MHyper has the highest value (0.0743), followed by CrossST (0.0227), while DyHSL, ST-GFSL, and FGITrans are negative in the same reference space. Because the values are close to zero and no calibrated threshold is available, we interpret only their ordering within this probe, not the absolute magnitude of alignment. STPB cannot rank predictive quality, validate general interpretability, or override target-task results. Appendix L reports its construction and bounded assessment.

4.3.4 Federated Learning Scenario Analysis

For federated learning, Table 10 compares the recorded mean uploaded tensor payload $\mathcal{C}_{\text{comm}}$ with an auditable MAE utility loss $\mathcal{L}_{\text{util}}$. The loss is recomputed directly from Table 9 relative to the strongest centralized model on each dataset. FedCTPM preserves predictive utility most closely, whereas FedGTP uploads the fewest recorded megabytes per client and round. The same data-local collaboration setting therefore supports different candidates depending on whether predictive utility or uploaded payload is prioritized.



(a) Performance degradation ($\Delta\mathcal{M}_{\text{miss}}$) under increasing data missingness



(b) Performance degradation ($\Delta\mathcal{M}_{\text{shift}}$) under the designated city-pair contrast

Figure 4: Sensitivity diagnostics. Panel (a) reports relative MAE change when 10%–50% of target training observations are masked. Panel (b) reports change between the designated homogeneous and heterogeneous pairs, not an isolated distribution intervention. Lower values indicate less degradation; negative values are fixed-execution fluctuations.

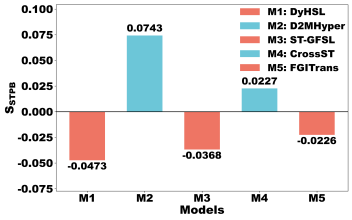


Figure 5: STPB prototype-alignment probe. Higher cosine similarity indicates closer alignment with the retained benchmark prototypes, not greater general interpretability.

5 DISCUSSION

5.1 WHAT THE BENCHMARK REVEALS

Three observations answer the benchmark questions. First, among the evaluated transfer implementations, only D2MHyper and CrossST record lower average MAE than the strongest target-only value on both core tasks. Second, accuracy, sensitivity, and cost produce different orderings: D2MHyper has the lowest speed MAE but larger change under the designated pair contrast, whereas CrossST has smaller change in the reported diagnostics but requires pre-training. Third, Appendix G shows that behavior changes with source composition. These fixed-seed observations support conditional guidance rather than a universal leaderboard.

The diagnostics motivate but do not establish mechanisms. CrossST's smaller recorded changes are compatible with lower dependence on the designated source-target alignment, while D2MHyper's nominal accuracy and larger pair contrast are compatible with greater dependence on the evaluated pair. The weaker meta-learning results are compatible with meta-task mismatch. Appendix I separates these hypotheses from measured evidence.

5.2 EVIDENCE-TRACEABLE SELECTION GUIDE

Table 6 maps each operating condition to an observed candidate, its evidence, and its principal limitation. It does not convert the reported settings into a deployment rule.

Table 6: Evidence-traceable selection guide. Candidates and limitations apply only to the evaluated settings.

Operating condition	Observed candidate and evidence	Scope limitation
Three-day budget; accuracy and sensitivity	CrossST: lowest flow MAE; speed MAE 3.21 $<$ target-only 3.82	Requires pre-training; not lowest speed MAE
Small positive change under the designated pair contrast	CrossST: $\Delta\mathcal{M}_{\mathrm{shift}} = 0.28\%$	Evidence is limited to evaluated pairs
Nominal speed accuracy	D2MHyper: lowest recorded speed MAE, 2.74	Pair sensitivity is 32.52%
Low latency with sufficient GPU memory	FGITrans: favorable latency	Peak memory is 14.32 GB; one distillation model
Federated communication constraint	FedGTP: lowest mean uploaded tensor payload, 0.1106 Megabytes (MB)	FedCTPM has smaller MAE utility loss; downloads and protocol overhead are excluded

The guide is closest to the evidence when a new setting resembles the evaluated tasks and uses the same target budget. Appendix G reports the limited multi-source check, while Appendix I records its interpretation boundaries.

For benchmark users, a candidate is first compared with target-only learning under the intended target-data budget. Candidates showing predictive utility are then compared on missingness, shift, latency, memory, or communication according to deployment constraints. This sequence prevents low nominal error, one favorable diagnostic, or a large source corpus from being sufficient evidence of transfer. A negative result remains informative because failure to beat target-only learning shows that the source interface does not justify its added information or cost in the evaluated setting.

Appendix K operationalizes this sequence as an auditable interface for adding future methods without changing the core comparison contract.

6 CONCLUSION

CrossCityBench evaluates whether source-enabled implementations record lower forecasting error under matched target access, horizons, and reporting conditions. Across two core tasks, the fixed-seed observations show selective transfer utility and disagreements among accuracy, sensitivity, efficiency, prototype alignment, and collaboration cost. The guide ties each recommendation to an explicit task, information budget, and diagnostic, so it remains limited to the evaluated settings. The comparison contract provides a reproducible basis for extending the evaluation without turning current observations into universal claims.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks with required disclosure, and the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to edit a research paper to improve readability. We have reviewed all AI-assisted work. All AI-assisted edits were manually checked by the authors against the original text to ensure that the technical terminology, numerical values, citations, claims, and intended scientific meaning remained unchanged. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This research uses publicly available traffic sensor datasets and does not involve human-subject recruitment or personally identifiable information. The datasets and accompanying resources are used in accordance with their documented research purposes and applicable access conditions. CrossCityBench is intended for research and diagnostic evaluation rather than direct deployment in safety-critical traffic management systems. Its results remain limited to the evaluated datasets, city pairs, target-data budgets, and model implementations. Differences in sensing coverage, infrastructure, and data availability across cities may affect model performance and should be considered before applying the benchmark conclusions to other regions or populations. The authors take responsibility for the integrity, documentation, and appropriate use of the released benchmark resources.

REPRODUCIBILITY STATEMENT

CrossCityBench is designed to make every reported comparison traceable to a declared task, information budget, model configuration, and evaluation procedure. Section 4.1 describes the experimental protocol, while Appendices D and E document the evaluated methods, benchmark adaptations, metrics, and diagnostic definitions. Appendix J connects the principal claims to their supporting evidence, and Appendix K specifies the run manifest, result schema, and admission procedure for future benchmark submissions. The dataset scope and processing requirements are described in Appendix A, and the STPB construction is detailed in Appendix L. The anonymous supplementary repository provides the released code, data-processing scripts, model configurations, environment specifications, and evaluation scripts associated with the reported benchmark results.

REFERENCES

Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17804–17815, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ce1aad92b939420fc17005e5461e6f48-Abstract.html.

Xiaojun Bi, Wudi Li, Yiwen Sun, Siyuan Qi, and Zheng Wang. Leveraging multiple source cities in selective transfer learning for traffic prediction with limited data. *IEEE Transactions on Intelligent Transportation Systems*, 26(10):15646–15659, 2025. doi: 10.1109/TITS.2025.3596309.

Yunyao Cheng, Peng Chen, Chenjuan Guo, Kai Zhao, Qingsong Wen, Bin Yang, and Christian S. Jensen. Weakly guided adaptation for robust time series forecasting. *Proceedings of the VLDB Endowment*, 17(4):766–779, 2023. doi: 10.14778/3636218.3636231.

Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are allset: A multiset function framework for hypergraph neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=hpBTIv2uy_E.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179.

Jeongwhan Choi, Hwangyong Choi, Jeehyun Hwang, and Noseong Park. Graph neural controlled differential equations for traffic forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6367–6374, 2022. doi: 10.1609/aaai.v36i6.20587.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135, 2017. URL https://proceedings.mlr.press/v70/finn17a.html.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

Shengnan Guo, Youfang Lin, Letian Gong, Chenyu Wang, Zeyu Zhou, Zekai Shen, Yiheng Huang, and Huaiyu Wan. Self-supervised spatial-temporal bottleneck attentive network for efficient long-term traffic forecasting. In *2023 IEEE 39th International Conference on Data Engineering*, pp. 1585–1596, 2023. doi: 10.1109/ICDE55515.2023.00125.

Jindong Han, Weijia Zhang, Hao Liu, Tao Tao, Naiqiang Tan, and Hui Xiong. BigST: Linear complexity spatio-temporal graph neural network for traffic forecasting on large-scale road networks. *Proceedings of the VLDB Endowment*, 17(5):1081–1090, 2024. doi: 10.14778/3641204.3641217.

Jindong Han, Hao Wang, Hui Xiong, and Hao Liu. Scalable pre-training of compact urban spatio-temporal predictive models on large-scale multi-domain data. *Proceedings of the VLDB Endowment*, 18(7):2149–2158, 2025. doi: 10.14778/3734839.3734851.

Junfeng Hu, Xu Liu, Zhencheng Fan, Yifang Yin, Shili Xiang, Savitha Ramasamy, and Roger Zimmermann. Prompt-based spatio-temporal graph transfer learning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 890–899, 2024. doi: 10.1145/3627673.3679554.

Jinkwan Jang, Hyungjin Park, Jinmyeong Choi, and Taesup Kim. Towards robust real-world multivariate time series forecasting: A unified framework for dependency, asynchrony, and missingness. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r4ZamwBE8P.

Songyu Ke, Zheyi Pan, Tianfu He, Yuxuan Liang, Junbo Zhang, and Yu Zheng. AutoSTG+: An automatic framework to discover the optimal network for spatio-temporal graph prediction. *Artificial Intelligence*, 318:103899, 2023. doi: 10.1016/j.artint.2023.103899.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJiHXGWAZ.

Aoyu Liu and Yaying Zhang. CrossST: An efficient pre-training framework for cross-district pattern generalization in urban spatio-temporal forecasting. In *2025 IEEE 41st International Conference on Data Engineering*, pp. 2935–2948, 2025. doi: 10.1109/ICDE65448.2025.00220.

Aoyu Liu and Yaying Zhang. A general spatio-temporal backbone with scalable contextual pattern bank for urban continual forecasting. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=LHSea6DI8U.

Chenxi Liu, Kethmi Hirushini Hettige, Qianxiong Xu, Cheng Long, Shili Xiang, Gao Cong, Ziyue Li, and Rui Zhao. ST-LLM+: Graph enhanced spatio-temporal large language models for traffic prediction. *IEEE Transactions on Knowledge and Data Engineering*, 37(8):4846–4859, 2025a. doi: 10.1109/TKDE.2025.3570705.

Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. TimeCMA: Towards LLM-empowered multivariate time series forecasting via cross-modality alignment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18): 18780–18788, 2025b. doi: 10.1609/aaai.v39i18.34067.

Zhanyu Liu, Guanjie Zheng, and Yanwei Yu. Multi-scale traffic pattern bank for cross-city few-shot traffic forecasting. *ACM Transactions on Knowledge Discovery from Data*, 19(4):1–24, 2025c. doi: 10.1145/3727622.

Bin Lu, Xiaoying Gan, Weinan Zhang, Huaxiu Yao, Luoyi Fu, and Xinbing Wang. Spatio-temporal graph few-shot learning with cross-city knowledge transfer. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1162–1172, 2022. doi: 10.1145/3534678.3539281.

Tinghui Luo, Ziquan Fang, Kaixuan Duan, Lu Chen, Panpan Feng, and Mingfan Lu. Towards online spatio-temporal prediction: A knowledge distillation driven continual learning approach. In *2025 IEEE 41st International Conference on Data Engineering*, pp. 2642–2655, 2025. doi: 10.1109/ICDE65448.2025.00199.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, Berlin, Heidelberg, 2005. doi: 10.1007/978-3-540-27752-1.

Jiaming Ma, Binwu Wang, Pengkun Wang, Zhengyang Zhou, Xu Wang, and Yang Wang. BiST: A lightweight and efficient bi-directional model for spatiotemporal prediction. *Proceedings of the VLDB Endowment*, 18(6):1663–1676, 2025. doi: 10.14778/3725688.3725697.

Jiqian Mo and Zhiguo Gong. Cross-city multi-granular adaptive transfer learning for traffic flow prediction. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11246–11258, 2023. doi: 10.1109/TKDE.2022.3232185.

Tong Nie, Wei Ma, Jian Sun, Yu Yang, and Jiannong Cao. Collaborative imputation of urban time series through cross-city meta-learning. *IEEE Transactions on Knowledge and Data Engineering*, 38(2):940–955, 2026. doi: 10.1109/TKDE.2025.3633492.

Xiaocao Ouyang, Yan Yang, Yiling Zhang, Wei Zhou, Jihong Wan, and Shengdong Du. Domain adversarial graph neural network with cross-city graph structure learning for traffic prediction. *Knowledge-Based Systems*, 278:110885, 2023. doi: 10.1016/j.knosys.2023.110885.

Xiaocao Ouyang, Yan Yang, Wei Zhou, Yiling Zhang, Hao Wang, and Wei Huang. CityTrans: Domain-adversarial training with knowledge transfer for spatio-temporal prediction across cities. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):62–76, 2024. doi: 10.1109/TKDE.2023.3283520.

Xiaocao Ouyang, Yanhua Li, Jie Zhang, Xin Yang, Yan Yang, Junbo Zhang, Wei Huang, Tianrui Li, and Zhiquan Liu. Enhancing cross-city spatio-temporal prediction via dynamic multi-scale hypergraph learning with domain adversarial training. *Knowledge-Based Systems*, 329:114352, 2025. doi: 10.1016/j.knosys.2025.114352.

Olawale Salaudeen, Nicole Chiou, Shiny Weng, and Sanmi Koyejo. Are domain generalization benchmarks with accuracy on the line misspecified? *Transactions on Machine Learning Research*, 2025. URL `https://openreview.net/forum?id=fNywRyqPQo`. Journal-to-Conference presentation at ICLR 2026.

Yihong Tang, Ao Qu, Andy H. F. Chow, William H. K. Lam, S. C. Wong, and Wei Ma. Domain adversarial spatial-temporal network: A transferable framework for short-term traffic forecasting across cities. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 1905–1915, 2022. doi: 10.1145/3511808.3557294.

Senzhang Wang, Hao Miao, Jiyue Li, and Jiannong Cao. Spatio-temporal knowledge transfer for urban crowd flow prediction via deep attentive adaptation networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(5):4695–4705, 2022. doi: 10.1109/TITS.2021.3055207.

Xing Wang, Jixiang Chen, Feifei Chen, Biao Jin, Fumin Zou, Lyuchao Liao, and Ruihao Zeng. FedCTPM: A federated cross-regional collaboration learning approach for traffic flow prediction. *Expert Systems with Applications*, 305:130779, 2026. doi: 10.1016/j.eswa.2025.130779.

Chenyu Wu, Zhipeng Ma, Junbo Zhang, Songyu Ke, and Yu Zheng. Spatio-temporal forecasting under open-world missingness with adaptive mixture-of-experts. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 3366–3375, 2025. doi: 10.1145/3746252.3761403.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph WaveNet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 1907–1913, 2019. doi: 10.24963/ijcai.2019/264.

Kaiwen Xia, Li Lin, Xinrui Zhang, Haotian Wang, Shuai Wang, and Tian He. A transferable spatio-temporal learning framework for cross-city logistics demand prediction. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5071–5082, 2025. doi: 10.1145/3711896.3737186.

Linghua Yang, Wantong Chen, Xiaoxi He, Shuyue Wei, Yi Xu, Zimu Zhou, and Yongxin Tong. FedGTP: Exploiting inter-client spatial dependency in federated graph-based traffic prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6105–6116, 2024. doi: 10.1145/3637528.3671613.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 3634–3640, 2018. doi: 10.24963/ijcai.2018/505.

Xie Yu, Jingyuan Wang, Yifan Yang, Qian Huang, and Ke Qu. BIGCity: A universal spatiotemporal model for unified trajectory and traffic state data analysis. In *2025 IEEE 41st International Conference on Data Engineering*, pp. 4455–4469, 2025. doi: 10.1109/ICDE65448.2025.00334.

Xiaoming Yuan, Zhenyu Luo, Ning Zhang, Ge Guo, Lin Wang, Changle Li, and Dusit Niyato. Federated transfer learning for privacy-preserved cross-city traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 26(4):4418–4431, 2025. doi: 10.1109/TITS.2025.3545445.

Yingqian Zhang, Chao Li, Shibo He, Xiangliang Zhang, and Jiming Chen. Toward generalized urban computing: Pretraining a spatial-temporal model for diverse urban tasks. *IEEE Transactions on Mobile Computing*, 24(10):10840–10852, 2025. doi: 10.1109/TMC.2025.3573373.

Yu Zhang, Hua Lu, Ning Liu, Yonghui Xu, Qingzhong Li, and Lizhen Cui. Personalized federated learning for cross-city traffic prediction. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pp. 5526–5534, 2024. doi: 10.24963/ijcai.2024/611. URL `https://www.ijcai.org/proceedings/2024/0611.pdf`.

Zhaoqi Zhang, Miao Xie, Pasquale Balsebre, Weiming Huang, Siqiang Luo, and Gao Cong. UrbanMFM: Spatial graph-based multiscale foundation models for learning generalized urban representation. *IEEE Transactions on Knowledge and Data Engineering*, 38(3):2064–2078, 2026. doi: 10.1109/TKDE.2026.3656202.

Yusheng Zhao, Xiao Luo, Wei Ju, Chong Chen, Xian-Sheng Hua, and Ming Zhang. Dynamic hypergraph structure learning for traffic flow forecasting. In *2023 IEEE 39th International Conference on Data Engineering*, pp. 2303–2316, 2023. doi: 10.1109/ICDE55515.2023.00178.

Zhiyuan Zhao, Haoxin Liu, and B. Aditya Prakash. Tackling time-series forecasting generalization via mitigating concept drift. In *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=emkvZ7NanK`.

Yuhao Zheng, Yishuo Cai, Zihao Cai, Changjun Fan, Senzhang Wang, and Jianxin Wang. FGITrans: Cross-city transformer for fine-grained urban flow inference. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 3415–3423, 2024. doi: 10.1145/3627673.3679855.

Jianping Zhou, Weida Wang, Bin Lu, Guanjie Zheng, Lei Bai, Xinbing Wang, and Chenghu Zhou. TRACK: Temporal decoupled kriging for inductive spatio-temporal graph. *IEEE Transactions on Knowledge and Data Engineering*, 38(6):3991–4005, 2026. doi: 10.1109/TKDE.2026.3674348.

SUPPLEMENTARY MATERIAL

Appendix roadmap. The supplement follows the logic of the main paper rather than repeating it. Appendix A documents the data space that motivates the benchmark. Appendix B separates broader context from tested baselines, and Appendix C explains what new comparisons the benchmark enables. Appendix D records baseline assumptions and adaptations, followed by metric definitions in Appendix E. Appendices F and G provide supplementary numerical evidence. Appendix H adds location-aware qualitative analysis. Appendix I consolidates claim boundaries. Appendices J, K, and L then close the chain with reproducibility, an executable extension protocol, and bounded STPB assessment details.

A EXTENDED CROSS-CITY SPATIO-TEMPORAL TRAFFIC DATASETS

To support evaluation under diverse urban conditions, we integrate a broad collection of public datasets encompassing various data types and environments, which facilitates research on cross-modal and cross-task data utilization.

A.1 DATASET STATISTICS AND DIVERSITY

Table 7 summarizes these resources using a common temporal-length convention. They exhibit substantial spatial and temporal diversity: node counts range from 44 (XMBRT) to over 1,000 (e.g., BJTT with 1,260), while recorded sequence lengths vary by more than an order of magnitude. This heterogeneity is intended to support future tests of data-processing pipelines across disparate scales; catalogue inclusion alone is not generalization evidence.

Table 7: Statistics of extended cross-city spatio-temporal traffic datasets

Datasets	#Nodes	Interval	# Time Steps	Datasets	#Nodes	Interval	# Time Steps
BJMetro	276	10min	4,752	BJTT	1260	5min	2,232
CHIBike	270	30min	4,416	Didi-Chengdu	524	10min	1,728
Didi-Shenzhen	627	10min	1,728	Dublin2021	33	5min	21,024
England	248	15min	17,376	HKData	617	5min	17,856
HZMetro	80	15min	2,400	Los-Loop	207	5min	8,928
NE-BJ	500	5min	8,928	NYCBike	250	30min	4,368
NYCTaxi	266	30min	4,368	PeMS04	307	5min	16,992
PeMS07	883	5min	28,224	PeMSD7(L)	1026	5min	12,672
PeMSD7(M)	228	5min	12,672	PeMS-Rainy	308	5min	21,024
Seattle-Loop	323	5min	105,120	SHMetro	288	15min	8,832
SZ-Taxi	156	5min	8,928	T-Drive	1024	60min	3,600
WHBT	251	5min	10,080	XMBRT	44	5min	7,488

A.2 POTENTIAL URBAN DATA EXTENSIONS

To clarify the scope beyond the evaluated traffic benchmarks, Figure 6 groups additional urban data sources that could support future cross-modal extensions. These sources are not part of the current benchmark release or empirical claims.

Transportation Sensor Data This category covers stationary sensors and imaging systems. Examples include Next Generation Simulation (NGSIM)[1] (vehicle trajectories), exiD[2] (highway interchanges with aerial imagery), and Cityscapes[3] (annotated street images from 50 cities).

Global Positioning System (GPS) Trajectory Data These datasets capture macro-level movement from GPS signals, including PeMS[4] (traffic flow), Tier and Bird e-scooter Application Programming Interfaces (APIs), DiDi Gaia[5] (ride-hailing trips), and conventional taxi trajectories.

[1] NGSIM Open Data portal
[2] https://www.exid-dataset.com
[3] https://www.cityscapes-dataset.com/
[4] http://pems.dot.ca.gov
[5] https://sts.didiglobal.com/

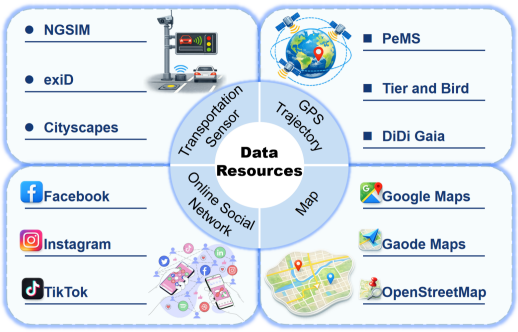


Figure 6: Potential urban data-source categories for future CrossCityBench extensions; these sources are outside the current evaluated benchmark.

Online Social Network Data Data from platforms such as Facebook, Instagram, and TikTok reflect human activity and are typically accessed via official APIs (e.g., Meta for Developers, TikTok Developer Portal) rather than direct downloads.

Map Data Geospatial data are provided by Google Maps and Gaode Maps (commercial APIs) and OpenStreetMap (OSM)[6], offering road infrastructure, points of interest, and live traffic conditions for urban analysis.

A.3 Scope of Core Evaluation Datasets and Extended Collections

The core experiments in Section 4 are conducted on established datasets (PeMS03, PeMS08, PeMS-BAY, METR-LA), a deliberate choice to ensure direct comparability and reproducibility with prior work. To further support evaluation across a wider spectrum of real-world urban heterogeneity, an extended collection of diverse datasets is integrated, as detailed in Table 7. This expanded resource is intended to enable the evaluation of analytical generalization under more challenging and varied conditions, facilitating research on knowledge transfer between cities with fundamentally different layouts, scales, or stages of development.

A.4 Supporting Complex Evaluations and Diagnostic Benchmarking

The diversity of the extended data allows exploration of complex, realistic forecasting problems beyond differences in scale. Integrating multi-modal data (Figure 6) can support future investigation of cross-modal migration and cross-task adaptation, transforming the benchmark from a conventional leaderboard into a diagnostic platform that probes data processing capabilities under more authentic urban conditions. This coverage defines directions for evaluating forecasting models across heterogeneous sensing and mobility settings.

The data catalogue therefore defines the intended expansion space, whereas the claims in the main paper remain tied to the city pairs for which results are reported. The next appendix applies the same distinction to methods by separating the evaluated 23-model suite from additional relevant literature.

A.5 Dataset Roles, Comparability, and Reuse

The catalogue is deliberately broader than the core scorecard. A dataset can play three distinct roles: a core-evaluation dataset supports a reported comparison under the fixed protocol; an extended-evaluation dataset supports a separately identified stress test or city-pair analysis; and a resource-catalogue dataset documents a plausible future domain without receiving an implied benchmark score. This separation prevents availability from being conflated with empirical validation.

For reuse, a new dataset should be mapped to the benchmark contract before model ranking is interpreted. The mapping records sensing modality, spatial unit, sampling interval, prediction target, source-target direction, chronological target budget, and missing-data rule. When one of these fields is

[6] https://www.openstreetmap.org

incompatible with the core setting, the result should be reported as an extension rather than combined with the central scorecard. This protocol-level distinction is especially important for metro, bike, taxi, and multimodal resources whose observation process differs from loop-detector traffic.

The catalogue also clarifies why larger coverage alone is not a generalization result. More cities can reduce a particular source-selection bias, but they do not remove differences in label semantics, sensor placement, or decision horizon. CrossCityBench therefore treats scale as an opportunity for subsequent matched evaluations rather than as evidence that the two core transfer results already hold everywhere.

B Broader Literature Coverage and Benchmark Boundary

The main paper emphasizes the models that can be compared under one matched protocol. A broader literature is retained here to preserve the intellectual context without implying that every cited method was reimplemented. This separation is important for an ICLR benchmark paper because breadth of citation and breadth of controlled evidence are different properties. It also addresses a practical reproducibility issue: a method can be relevant to the problem while remaining incomparable until its data interface, information budget, and evaluation horizon are matched.

B.1 Scalable and General Spatio-Temporal Modeling

Recent forecasting architectures improve long-horizon efficiency, automated design, or large-network scalability. Representative examples include the bottleneck-attention formulation of STBAN (Guo et al., 2023), automated architecture discovery in AutoSTG+ (Ke et al., 2023), linear-complexity graph forecasting in BigST (Han et al., 2024), and the lightweight bidirectional design of BiST (Ma et al., 2025). Broader urban representation models extend beyond one forecasting dataset, including BIGCity (Yu et al., 2025), compact multi-domain pre-training (Han et al., 2025), and UrbanMFM (Zhang et al., 2026). These works motivate scalability and representation-transfer axes, but they are not assigned benchmark scores unless they follow the same source-target protocol.

B.2 Transfer, Adaptation, and Source Selection

Cross-city transfer has expanded from adversarial alignment to prompt-based adaptation and source-aware transfer. CityTrans (Ouyang et al., 2024) and multi-granular adaptation (Mo & Gong, 2023) model city discrepancy explicitly. Prompt-based graph transfer (Hu et al., 2024) changes the adaptation interface, while selective transfer from multiple source cities (Bi et al., 2025) addresses negative transfer caused by an unsuitable source. The setting has also been extended to cross-city logistics demand (Xia et al., 2025) and inductive spatio-temporal kriging (Zhou et al., 2026). Together, these studies show why a single source-target pair cannot support a universal ranking.

B.3 Non-Stationarity, Missingness, and Continual Updating

Robust forecasting must distinguish static domain shift from changes that arise over time. Weakly guided adaptation studies robustness under evolving forecasting distributions (Cheng et al., 2023); continual distillation updates models for online spatio-temporal prediction (Luo et al., 2025); and adaptive mixture-of-experts addresses open-world missingness (Wu et al., 2025). Cross-city meta-learning has also been used for collaborative imputation (Nie et al., 2026). These directions motivate the missingness and shift diagnostics in CrossCityBench, but the present stress tests do not substitute for a complete continual-learning or imputation benchmark.

B.4 Foundation Models, Language Models, and Privacy

Urban pre-training increasingly targets multiple tasks rather than one forecasting output (Zhang et al., 2025). ST-LLM+ incorporates graph structure into Large Language Model (LLM) based traffic prediction (Liu et al., 2025a), while TimeCMA aligns time-series and language representations for multivariate forecasting (Liu et al., 2025b). In parallel, federated transfer learning addresses cross-city prediction without direct data pooling (Yuan et al., 2025). These approaches are retained in the

literature map because they define important future comparison groups. They are not inserted into the current result tables without matched runs, which keeps the experimental claims auditable.

Having established this broader boundary, Appendix D now specifies exactly which methods were evaluated and how their original assumptions were adapted to the common protocol. The subsequent metric appendix then defines the coordinates used to compare those matched implementations.

C Benchmark Capability and Scientific Use

CrossCityBench is not designed as a storage, indexing, or query-processing system. Its technical role is an evaluation and decision layer for machine-learning methods that consume spatio-temporal city data. This positioning matters because the benchmark contribution should be assessed by the comparisons it makes valid and the failure modes it reveals, rather than by database functionality that it does not claim to provide.

C.1 Controlled Cross-Paradigm Comparison

Existing cross-city studies commonly vary the source-target pair, target-data budget, prediction horizon, preprocessing rule, and evaluation metric at the same time. Under those conditions, a difference between published numbers cannot be attributed to a transfer mechanism. CrossCityBench fixes the task-facing variables and represents each paradigm as a distinct assumption about transferable structure. Target-only learning assumes that source information is unnecessary. Alignment assumes that source-target discrepancy can be reduced. Meta-learning assumes that prior tasks yield a useful adaptation trajectory. Pre-training assumes that reusable representations can be learned before target fitting. Distillation assumes that teacher knowledge can be compressed. Federated learning assumes that collaboration remains useful without pooling raw data.

The matched protocol turns these assumptions into comparable questions. The first asks whether a source-enabled implementation records lower error than a strong target-only reference. The second asks whether an accuracy advantage remains under missingness, city-pair changes, and resource constraints. The third asks whether a recommendation changes when centralized access is replaced by federated collaboration. No individual source paper provides all three comparisons under the same target budget.

C.2 Diagnostic Capability Beyond a Leaderboard

A single accuracy table identifies a winner only for one operating point. The diagnostic evaluator instead preserves separate coordinates for predictive error, latency, memory, model size, missing-data sensitivity, shift sensitivity, prototype alignment, communication cost, and utility loss. Keeping these coordinates separate is a methodological choice. An aggregate score would hide whether one method achieves a favorable average by trading away robustness or deployment feasibility.

The reported evidence illustrates why this separation is necessary. D2MHyper records the lowest speed MAE but has substantially larger change under the designated pair contrast than CrossST. FGITrans has favorable latency but high peak memory. FedCTPM has stronger predictive accuracy and smaller MAE loss relative to centralized learning, whereas FedGTP has a lower recorded uploaded payload. These measurements are not contradictory; they show that model selection is conditional on the deployment objective.

C.3 Traceable Decision Rather Than Universal Prescription

The empirical selection guide maps each recommendation to a measured coordinate and states the limitation that qualifies it. This traceability distinguishes the guide from a qualitative survey table. A reader can recover why a method was selected, which metric supports it, and which unmeasured condition could invalidate it. The guide is therefore an evidence summary for the evaluated traffic settings, not a learned policy or a universal decision rule.

The same principle governs future extensions. A new city or task should enter the guide only after its information budget, task semantics, and evaluation coordinates are aligned with the benchmark

contract. This rule preserves comparability as coverage grows and prevents a larger dataset catalogue from being mistaken for stronger empirical generalization.

D Detailed Baseline Descriptions

The main text introduces the taxonomy because it defines the unit of comparison. This appendix resolves it into model-level implementation choices, original assumptions, and benchmark adaptations. The details are necessary for reproducibility, but they are separated from the main-text taxonomy so that a reader does not mistake an adapted interface for a new algorithmic contribution.

D.1 Target-Only Data Modeling Pipelines

Ten pipelines learn spatio-temporal dynamics solely from target-city data. **GBRT** (Friedman, 2001) uses gradient boosting trees; **VAR** (Lütkepohl, 2005) applies multivariate autoregression; **AGCRN** (Bai et al., 2020) couples node-adaptive graph learning with GRUs; **AllDeepSet** (Chien et al., 2022) employs permutation-invariant hypergraph functions; **DCRNN** (Li et al., 2018) uses bidirectional graph diffusion in seq2seq; **DyHSL** (Zhao et al., 2023) captures higher-order correlations via hypergraph convolution; **GRU** (Cho et al., 2014) is a recurrent baseline; **GWNet** (Wu et al., 2019) combines self-adaptive graphs and dilated temporal convolutions; **STGCN** (Yu et al., 2018) couples graph and temporal convolutions; **STG-NCDE** (Choi et al., 2022) models continuous-time dynamics via neural differential equations.

D.2 Cross-City Data Transfer Models

These methods transfer knowledge from data-rich sources to data-scarce targets, categorized into four paradigms.

Alignment-Based Methods align distributions via adversarial learning or matching. **DASTNet** (Tang et al., 2022) uses adversarial domain adaptation; **D2MHyper** (Ouyang et al., 2025) combines multi-scale dynamic hypergraphs with adversarial training; **DAGN** (Ouyang et al., 2023) learns cross-city graph structure with domain discriminators; **ST-DAAN** (Wang et al., 2022) transfers crowd-flow knowledge through deep attentive adaptation.

Meta-Learning Methods enable fast few-shot adaptation. **MAML** (Finn et al., 2017) learns a reusable initialization; **ST-GFSL** (Lu et al., 2022) augments it with node-level meta-knowledge for spatio-temporal graphs.

Pre-Training Methods learn generalizable representations from multi-city data before fine-tuning. **CrossST** (Liu & Zhang, 2025) pre-trains a unified encoder-decoder; **MTPB** (Liu et al., 2025c) transfers traffic prototypes via clustering; **STGCN-FT** adapts the STGCN backbone (Yu et al., 2018) under the shared pre-training and fine-tuning protocol.

Knowledge Distillation Methods compress teacher models into lightweight students. **FGI-Trans** (Zheng et al., 2024) distills fine-grained spatial-temporal knowledge with multi-granularity objectives.

D.3 Applicability and Benchmark Adaptations

The taxonomy describes learning assumptions, while Table 8 records the conditions under which each group is compared. This distinction prevents an implementation adaptation from being mistaken for the original paper's protocol.

D.4 Data-Local Federation Frameworks

Federated frameworks enable collaborative learning without raw data sharing. **FedCTPM** (Wang et al., 2026) supports cross-city analytics with data locality; **pFedCTP** (Zhang et al., 2024) learns personalized models via privacy-preserving transfer; **FedGTP** (Yang et al., 2024) separates client-specific and global components in spatio-temporal graph networks.

Table 8: Applicability constraints and benchmark adaptations.

Method group	Required setting	Benchmark treatment
Target only	Target observations and a compatible graph or grid	Uses only the 3-day target-information budget for fitting and checkpoint selection; no source observations enter
Centralized transfer	Paired source and target domains with centralized access	Uses the same input horizon, output horizon, normalization, and target budget across alignment, meta-learning, pretraining, and distillation
Federated learning	Multiple client partitions with no raw-data pooling	Evaluated in a separate federated protocol; utility and communication are not merged with centralized accuracy
ST-DAAN	Originally defined for grid-based urban crowd flow	Adapted to the benchmark tensor interface; reported numbers describe this adaptation rather than the original paper's setting
STGCN-FT	A transferable backbone and a target fine-tuning stage	Benchmark variant built from STGCN; it is not presented as a separately published method

E Evaluation Metrics and Definitions

The main paper uses several diagnostic coordinates because no one metric captures the operating trade-off. This appendix defines each coordinate and its direction of preference so that the tables in the next appendix can be read as measurements of a common protocol rather than as incomparable reported numbers.

E.1 Accuracy and Data Efficiency

Accuracy under data scarcity is measured by MAE, RMSE, and MAPE: $\mathrm{MAE} = \frac{1}{M}\sum_{i=1}^{M}|y_i - \hat{y}_i|$, $\mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2}$, and $\mathrm{MAPE} = \frac{100\%}{M}\sum_{i=1}^{M}\left|\frac{y_i - \hat{y}_i}{y_i + \epsilon}\right|$, where y_i and $\hat{y}_i$ are true and estimated values, M is test samples, and ϵ ensures numerical stability. Lower values indicate better accuracy.

E.2 Efficiency and Cost

Practical resource consumption is reported as total training time, inference latency, peak GPU memory (GB), and trainable parameters. Because timing and memory are implementation- and hardware-dependent, comparisons are restricted to the common V100 environment and treated as descriptive system coordinates. Appendix K lists the measurement fields required for future comparable executions.

E.3 Sensitivity Diagnostics

We measure performance degradation under two diagnostic scenarios. For data missingness, a random fraction p_{miss} of observation-level entries in the target training partition is masked before fitting. Each model is retrained on the resulting masked training data, while its evaluation data remain unchanged. The relative MAE increase is $\Delta\mathcal{M}_{\text{miss}} = \frac{\mathrm{MAE}_{\text{masked}} - \mathrm{MAE}_{\text{full}}}{\mathrm{MAE}_{\text{full}}} \times 100\%$, where $\mathrm{MAE}_{\text{full}}$ and $\mathrm{MAE}_{\text{masked}}$ correspond to training on full and masked data; lower $\Delta\mathcal{M}_{\text{miss}}$ indicates smaller observed missingness sensitivity. For the city-pair contrast, we compare the designated heterogeneous and homogeneous pairs: $\Delta\mathcal{M}_{\text{shift}} = \frac{\mathrm{MAE}_{\text{hetero}} - \mathrm{MAE}_{\text{homo}}}{\mathrm{MAE}_{\text{homo}}} \times 100\%$, where a smaller $\Delta\mathcal{M}_{\text{shift}}$ indicates less observed degradation. We retain the subscript "shift" for consistency with the recorded benchmark output, but interpret the quantity only as a city-pair contrast because pair identity and distribution change together.

E.4 STPB Prototype-Alignment Probe

This metric quantifies how closely a model representation aligns with spatio-temporal prototypes encoded in the STPB, a learned reference space for bounded probing. Alignment is measured by the STPB similarity score, defined as $\mathcal{S}_{\text{STPB}} = \frac{1}{K}\sum_{k=1}^{K}\frac{\mathbf{r}_{\text{model}} \cdot \mathbf{p}_k}{\|\mathbf{r}_{\text{model}}\|\|\mathbf{p}_k\|}$, where K denotes the number of retained prototype patterns, such as periodicity and trend, and $\mathbf{r}_{\text{model}}$, $\mathbf{p}_k$ are the model representation and prototype vectors. Higher scores ($\mathcal{S}_{\text{STPB}} \in [-1, 1]$) indicate stronger alignment with these benchmark prototypes. They do not establish prototype universality, semantic completeness, or causal importance.

E.5 Collaboration Cost

This dimension analyzes trade-offs in data-local collaborative learning. The recorded communication footprint is the mean uploaded tensor payload per client and aggregation round, $\mathcal{C}_{\text{comm}} = (TC)^{-1} \sum_{t=1}^{T} \sum_{c=1}^{C} \text{size}(\boldsymbol{\theta}_c^{(t)})$, reported in MB under the common federated schedule. This definition keeps methods with the same number of rounds comparable without presenting a per-round payload as total training traffic. The statistic does not represent server downloads, serialization headers, protocol overhead, or wall-clock communication. To retain a value that can be checked against the accuracy table, we define MAE utility loss as $\mathcal{L}_{\text{util}} = \text{MAE}_{\text{fed}} - \min_{m \in \mathcal{M}_{\text{central}}} \text{MAE}_m$ on the same dataset. Smaller $\mathcal{L}_{\text{util}}$ means that federated accuracy is closer to the strongest evaluated centralized reference. It measures an empirical performance difference and is not a formal privacy guarantee.

All metrics follow the lower-is-better principle, except for $\mathcal{S}_{\text{STPB}}$ where a higher value is desirable.

F Additional Core Benchmark Results

The two complete centralized result tables are reported in the main paper so that its principal accuracy claims can be checked without consulting the supplement. This section retains the federated comparison because centralized transfer and data-local collaboration obey different information-access constraints. Keeping it separate prevents communication efficiency from being conflated with centralized prediction accuracy.

Table 9: Data analytics performance comparison: Centralized learning vs. Paradigm 6 (federated learning)

Categories	**Models**	**METR-LA**			**PeMS-BAY**		
		MAE	RMSE	MAPE (%)	MAE	RMSE	MAPE (%)
Centralized Learning	DCRNN	<u>4.04</u>	<u>9.38</u>	9.98	1.69	3.78	3.90
	GWNet	**3.92**	**9.25**	**9.68**	**1.58**	**3.52**	**3.58**
	STGCN	4.63	10.58	11.38	1.80	4.20	4.14
Federated Learning	FedCTPM	4.33	10.53	<u>9.94</u>	<u>1.68</u>	<u>3.68</u>	<u>3.77</u>
	pFedCTP	4.78	10.23	12.10	2.04	4.46	4.91
	FedGTP	4.44	10.86	10.41	1.75	3.83	3.85

Table 10: Federated communication and utility trade-off. Utility loss is recomputed from Table 9 relative to the lowest evaluated centralized MAE on the same dataset. Lower values are preferable in all three numeric columns.

Federated model	**Recorded communication (MB)**	**METR-LA MAE loss**	**PeMS-BAY MAE loss**
FedCTPM	1.0880	**0.41**	**0.10**
pFedCTP	0.2221	0.86	0.46
FedGTP	**0.1106**	0.52	0.17

The utility-loss entries follow directly from the MAE values in Table 9: the strongest centralized references are GWNet with 3.92 on METR-LA and 1.58 on PeMS-BAY. Using directly recomputable losses keeps the accuracy and collaboration analyses on one auditable scale. FedCTPM is therefore the observed candidate for predictive utility preservation, whereas FedGTP is the observed candidate for recorded communication cost. Neither comparison establishes a general privacy-utility frontier beyond the evaluated implementations.

G Additional Comparative Analysis on Extended City Pairs

The core tables establish conditional results for two controlled transfers. We next test whether that conditionality persists when the source set changes. The analysis extends evaluation to multi-source $\rightarrow$ single-target transfer using Chengdu (C), Shenzhen (S), METR-LA (M), and PeMS-BAY (P). Table 11 compares DCRNN as a target-only baseline, DASTNet as an alignment method, and ST-GFSL as a graph-specific meta-learning method. Pre-training and distillation are not included, so this is a limited sensitivity check on the decision logic rather than a new universal ranking.

G.1 Purpose of the Extended Comparison

The extended comparison addresses a specific threat to the core findings: a transfer mechanism may appear favorable because one source happens to resemble one target. Holding the prediction task fixed while changing the source composition provides a direct check on whether the ordering is pair dependent. The comparison does not claim coverage of every learning paradigm. Instead, it uses one target-only reference, one alignment method, and one graph meta-learning method to expose how source composition can reverse a conclusion drawn from the core pairs.

This distinction sharpens the benchmark contribution. A conventional leaderboard asks which method has the smallest error in one table. CrossCityBench asks whether the same recommendation persists after a change in target identity, source composition, sensitivity condition, or deployment constraint. ST-GFSL is weak in the two core transfers yet records the lowest MAE at every reported horizon in the limited three-method multi-source comparison, while DASTNet fails to improve over DCRNN for Shenzhen. Such reversals motivate a conditional selection guide.

G.2 Interpretation Safeguards

Three safeguards govern the interpretation of Table 11. First, comparisons are made within each target and horizon; absolute errors are not compared across different sensing scales. Second, a favorable result for one representative does not establish dominance of its entire paradigm. Third, the absence of pre-training and distillation methods from this table is treated as missing coverage rather than negative evidence. These rules preserve the distinction between what the reported measurements show and what a broader benchmark extension would still need to test.

The same safeguards explain why the extended catalogue in Appendix A is not presented as an evaluated result. Dataset availability increases the benchmark's reuse potential, whereas matched runs determine empirical scope. Separating those roles directly addresses the risk of treating a large resource list as proof of generalization.

The results in Table 11 provide a limited check on the selection logic in Table 6. DASTNet records lower MAE than DCRNN on three target groups but higher MAE at every Shenzhen horizon. ST-GFSL records the lowest MAE at every reported horizon in this three-method comparison despite its weaker core results. The ranking change shows that source composition and target identity affect the observed value of a transfer implementation. Consequently, the extended comparison supports conditional conclusions rather than a single hierarchy.

H Location-Aware Case Studies

The aggregate comparisons above establish that rankings depend on the evaluation axis and city pair. We now make this dependence concrete at the sensor level, where local traffic regimes reveal why a method that performs well on average can still miss a practically important pattern.

The preceding tables establish how rankings vary across city pairs but cannot show where a mean error hides a practically important failure. A visual case study across PeMS08 and METR-LA therefore evaluates five methods at central hubs, arterials, and local roads. It links aggregate diagnostics to sensor topology and is intended as qualitative evidence, not as a further quantitative comparison.

H.1 Visualization Methodology

For each city, four views are presented: a sensor map, and time-series data patterns for **central hubs** (complex, volatile patterns), **peripheral arterials** (regular commuter patterns), and **local roads** (data-scarce, low-volume). This consistent layout enables location-specific comparison.

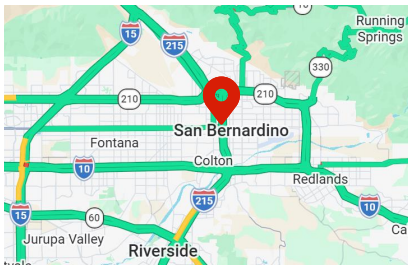
H.2 Case 1: Traffic Flow Data Analysis on PeMS08 Dataset

Figure 7 shows the PeMS08 sensor map and data analytics at three location types.

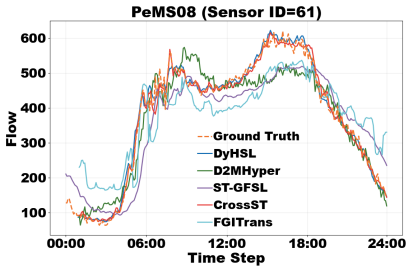
Key observations: at the central hub, several models deviate during peaks, while CrossST and DyHSL stay close to the ground truth. On arterials, D2MHyper and CrossST track peaks well, whereas

Table 11: Data analytics performance across multiple city pairs using DCRNN, DASTNet, and ST-GFSL.

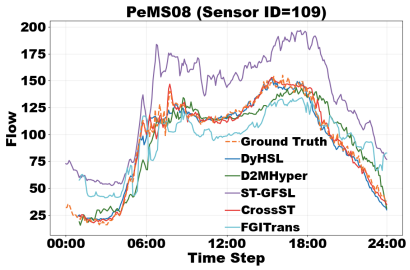
Transfer Tasks	Models	Horizon	MAE	RMSE	MAPE (%)
M,C,S → PeMS-BAY	DCRNN	5 min	1.46	2.56	3.02
		15 min	2.04	3.70	4.10
		30 min	2.65	5.17	5.87
	DASTNet	5 min	1.21	2.09	2.46
		15 min	1.75	3.30	3.64
		30 min	2.25	4.65	5.02
	ST-GFSL	5 min	1.10	1.86	2.24
		15 min	1.65	3.21	3.52
		30 min	2.09	4.43	4.65
P,C,S → METR-LA	DCRNN	5 min	2.68	4.81	6.97
		15 min	3.40	6.37	9.60
		30 min	4.36	7.85	12.57
	DASTNet	5 min	2.52	4.07	6.14
		15 min	2.96	5.55	8.37
		30 min	3.79	6.94	10.64
	ST-GFSL	5 min	2.36	4.15	6.18
		15 min	2.92	5.57	8.14
		30 min	3.66	6.99	10.57
M,P,S → Chengdu	DCRNN	10 min	2.32	3.57	9.99
		30 min	3.04	4.32	13.44
		60 min	3.65	5.00	16.36
	DASTNet	10 min	2.29	3.32	9.73
		30 min	2.94	4.04	12.85
		60 min	3.28	4.69	14.00
	ST-GFSL	10 min	2.21	3.26	9.40
		30 min	2.70	3.97	11.76
		60 min	3.15	4.58	13.55
M,P,C → Shenzhen	DCRNN	10 min	2.01	2.80	7.95
		30 min	2.43	3.53	10.38
		60 min	2.83	4.20	12.60
	DASTNet	10 min	2.05	3.03	8.48
		30 min	2.50	3.64	10.28
		60 min	2.86	4.15	11.73
	ST-GFSL	10 min	1.97	2.85	7.87
		30 min	2.33	3.57	10.08
		60 min	2.69	4.15	11.84



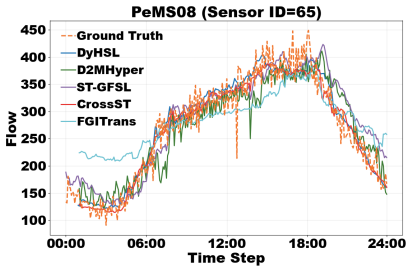
(a) Geographical map of the PeMS08 sensor network



(b) Data pattern at a central hub sensor



(c) Data pattern at a peripheral arterial sensor



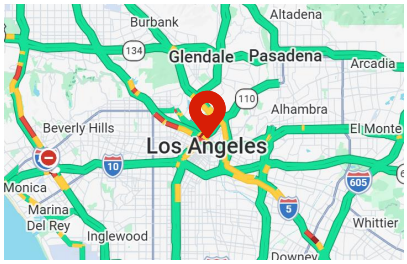
(d) Data pattern at a local road sensor

Figure 7: Case study on PeMS08 (flow data analysis).

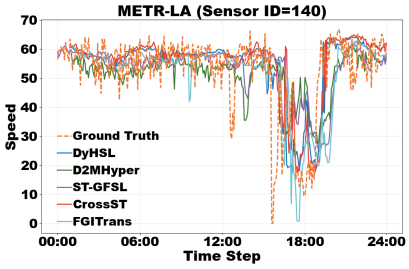
ST-GFSL overestimates flow and FGITrans oversmoothes peaks. On local roads, CrossST shows the most stable tracking, with DyHSL and D2MHyper also competitive. Overall, CrossST achieves more consistent temporal alignment, while distillation gives smooth but dampened estimates.

H.3 Case 2: Traffic Speed Data Analysis on METR-LA Dataset

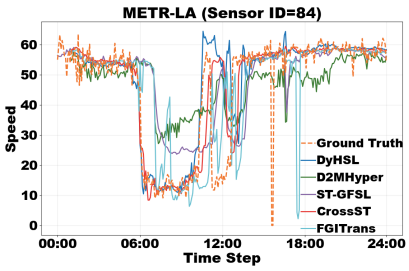
Figure 8 replicates the analysis for speed data on METR-LA.



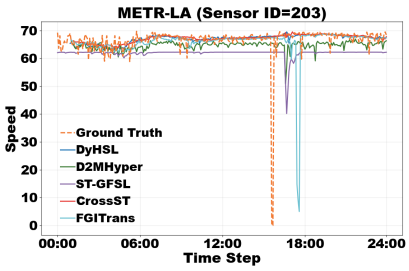
(a) Geographical map of the METR-LA sensor network



(b) Data pattern at a central hub sensor



(c) Data pattern at a peripheral arterial sensor



(d) Data pattern at a local road sensor

Figure 8: Case study on METR-LA (speed data analysis).

Key observations: central hubs show abrupt speed drops, with CrossST missing declines and FGITrans overreacting. On arterials, CrossST tracks ground truth steadily, while DyHSL has timing offsets and FGITrans recovers late. On local roads, CrossST and DyHSL are closest, whereas ST-GFSL occasionally drops. Location-dependent performance remains evident across tasks.

H.4 Guidelines for Location-Aware Paradigm Selection

The case studies yield three practical guidelines:

- **Central hubs** exhibit heterogeneous temporal regimes in the displayed examples. Pre-training or federated candidates may be considered when their additional source interface matches the intended constraint, while target-only and distillation results should be checked against local error patterns rather than rejected by paradigm label.
- **Peripheral arterials** show regular commuter patterns in the displayed cases. The observed traces motivate comparing alignment and pre-training candidates at these locations, but they do not establish a paradigm-level benefit or an efficiency trade-off beyond the evaluated models.
- **Local roads** show that the preferred candidate can differ from the aggregate ranking. In the displayed examples, the closer predictions from CrossST and DyHSL motivate a location-specific check rather than a general claim about one learning paradigm under Non-Independent and Identically Distributed (Non-IID) data.

These local examples provide qualitative context for the aggregate metrics. Before returning to the representation-level diagnostic, the following section consolidates the validity boundaries that govern both the quantitative tables and these case interpretations.

I LIMITATIONS AND VALIDITY BOUNDARIES

The main paper deliberately restricts its conclusions to the evidence produced by the matched protocol. This section separates limitations of empirical coverage from limitations of the evaluation design so that neither is hidden inside the selection guide.

Table 12 first maps each principal validity threat to its current control and residual boundary. The subsections that follow explain why those residual limitations constrain geographic, statistical, optimization, and mechanistic claims.

Table 12: Validity audit for the benchmark conclusions. A control reduces a threat but does not remove the stated residual limitation.

Validity threat	Current control	Residual limitation
Unequal task definitions	Shared input and output horizons, chronological target budget, preprocessing, and metrics	Method-specific structural operations remain different
Weak target-only reference	Ten target-only pipelines and comparison against the strongest observed baseline	Rankings remain conditional on the evaluated model set
Pair-specific transfer	Two core transfers plus multi-source comparisons	Core evidence covers traffic flow and speed rather than all urban tasks
Ranking uncertainty	Fixed seeds and explicit caution for small numerical gaps	Repeated runs and confidence intervals are not reported
Unequal tuning benefit	Shared optimization budget and disclosed adaptations	The protocol does not establish invariance to per-method tuning
Metric overreach	Separate diagnostic coordinates and bounded STPB interpretation	STPB does not explain internal causal mechanisms
Paradigm overreach	Model-level tables and qualified group statements	Some paradigms contain only one or two representatives

The audit provides the reading rule for the remainder of this section. The controls explain why the reported comparisons are informative, while the residual limitations explain why the ranking should not be extrapolated without a matched extension.

I.1 TASK AND GEOGRAPHIC SCOPE

The reported experiments concern traffic flow and traffic speed forecasting. They do not establish that the same paradigm ordering holds for air quality, human mobility, demand prediction, or other urban spatio-temporal tasks. The extended resource catalogue describes possible evaluation domains, but a listed dataset is not counted as empirical support until methods are executed under a matched task definition.

The two core transfers also have different meanings. PeMS03 to PeMS08 is a cross-region transfer within a shared California sensor infrastructure and mainly tests distribution shift under aligned measurement semantics. PeMS-BAY to METR-LA changes the metropolitan region and graph topology and is therefore the stronger cross-city setting. Treating both as equally heterogeneous would overstate the evidence. The multi-source comparisons broaden city composition, but only for the three methods that are actually reported.

I.2 RANKING UNCERTAINTY

The tables contain fixed-seed point estimates. Consequently, small gaps such as 0.09 MAE between CrossST and FGITrans on flow should not be read as statistically decisive. The directly supported observations concern cross-task and cross-axis contrasts, including that only two evaluated transfer implementations, D2MHyper and CrossST, fall below the strongest target-only reference on both core tasks. Repeated runs and confidence intervals remain necessary before fine-grained adjacent rankings can be treated as stable.

I.3 OPTIMIZATION AND ADAPTATION FAIRNESS

A shared optimization budget controls one source of variation, but it is not equivalent to providing every architecture with its individually optimal configuration. Conversely, using only published settings would confound the comparison with unequal tuning effort and different source-target protocols. CrossCityBench adopts the shared-budget interpretation and reports method-specific

structural operations separately. The resulting scores should therefore be read as controlled benchmark outcomes rather than method-specific upper bounds.

Several methods also require interface adaptations. ST-DAAN originates from a grid-based setting, STGCN-FT is a benchmark fine-tuning variant, and federated methods require client partitions that centralized methods do not use. Appendix D.3 records these differences. Results for an adapted implementation characterize that implementation under the benchmark contract, not every possible realization of the original method.

I.4 Diagnostic and Method-Zoo Coverage

STPB measures alignment with recurring prototypes. It can distinguish whether representations emphasize the selected pattern bank, but it cannot reveal a model's complete internal mechanism or establish general human interpretability. The five-expert comparison in Appendix L is therefore a bounded descriptive assessment rather than a universal validation.

The method zoo is intentionally broad but not exhaustive. Distillation is represented by FGITrans alone, and meta-learning by two methods, so statements about these groups remain model-level or explicitly qualified. Recent time-series foundation models and language-model adaptations are placed in the literature map rather than inserted into result tables without matched executions. These boundaries identify concrete extensions while keeping the present conclusions auditable.

I.5 Observed Relations and Mechanistic Hypotheses

The benchmark separates a measured relation from a proposed explanation. CrossST's small measured degradation is an observation. Reduced dependence on pair-specific alignment is a hypothesis compatible with that observation. D2MHyper's low nominal speed MAE and larger degradation under the designated pair contrast are observations. Greater reliance on the evaluated pair is a possible interpretation. Likewise, the weaker core results of MAML and ST-GFSL do not prove that meta-learning is unsuitable for cross-city forecasting; they are compatible with a mismatch between the available meta-tasks, optimization path, and target setting.

This distinction matters because coordinated diagnostics do not identify an internal cause. A causal statement would require an intervention that changes one candidate mechanism while holding other factors fixed. The current benchmark does not perform such interventions. It therefore reports where rankings change, associates those changes with observable conditions, and treats architectural explanations as hypotheses for future study.

J Protocol Traceability and Reporting Boundary

The preceding appendices provide the evidence behind each main-text claim. This final protocol note makes explicit what is held fixed, what remains method-specific, and how the reported scope should be read. It is included to address a recurring benchmark concern: numerical tables are only informative when a reader can distinguish a shared evaluation condition from a model-specific design choice.

J.1 Evidence Hierarchy

CrossCityBench separates four evidence levels so that breadth is not confused with validation. Core matched evidence uses the complete comparison contract and supports the principal findings in the main paper. Extended matched evidence changes source composition for the methods actually reported and tests whether the selection logic remains conditional. Diagnostic case studies localize aggregate behavior but do not establish a new ranking. Catalogue-only resources document future coverage and support reuse, but they provide no empirical evidence until methods are executed under the same task and information contract.

Table 13 makes these levels and their admissible interpretations explicit before the subsequent protocol clauses use them.

Table 13: Evidence hierarchy and admissible interpretation. Inclusion at a broader level does not inherit the claims supported by a more controlled level.

Evidence level	Protocol status	Admissible use	Excluded inference
Core matched evidence	Full target budget, task, reference, and diagnostic contract	Principal model-level comparisons and cross-axis ranking observations	Population-level claims over cities or tasks
Extended matched evidence	Changed source composition for the reported subset	Check whether conclusions remain conditional across listed settings	Complete paradigm ranking
Diagnostic case study	Location-aware visualization under selected models and locations	Qualitative context for aggregate error and pattern behavior	Statistical generalization or causal attribution
Catalogue-only resource	Dataset metadata without matched execution	Benchmark extension planning and dataset discovery	Evidence of method performance or generalization

This hierarchy also governs cross-references. A main-text conclusion may cite an extended result as a limited check, but the extended result cannot silently enlarge the population described by the conclusion. Similarly, a case visualization can illustrate where an error occurs, but it cannot establish why the model produced that error.

J.2 Shared Comparison Contract

For each centralized comparison, the source-target direction, 12-step input horizon, 12-step output horizon, chronological three-day target budget, normalization rule, target test split, and reported metrics are shared. Source information is defined broadly as any source-derived observation, parameter, prototype, teacher output, initialization, or federated update that is unavailable to a target-only model. Target-only models receive none of these objects. Transfer models receive source information only through the interface allowed by their paradigm. The contract therefore matches target supervision and task definition rather than claiming that heterogeneous paradigms consume identical inputs. It makes the central question falsifiable: whether a complete source-enabled implementation records lower error on the same target task than a strong target-only reference.

Table 14 records the source-derived and target-side information associated with each paradigm so that equal target access is not confused with identical internal inputs.

Table 14: Information-access interpretation used by the comparison contract. The table defines access categories and does not imply that every implementation within a category uses the same internal operation.

Paradigm	Source-derived access	Matched target access	Comparison meaning
Target only	None	Three chronological training days	Reference without source benefit or source bias
Alignment	Source observations or representations during alignment	Same target budget and forecasting task	Value of explicit discrepancy reduction
Meta-learning	Source tasks or source-derived initialization	Same target adaptation budget	Value of an adaptation-oriented prior
Pre-training	Source-pretrained parameters or prototypes	Same target fine-tuning budget	Value of reusable source representation
Distillation	Source-trained teacher outputs or parameters	Same target budget for the student interface	Value of the evaluated teacher-student transfer
Federated	Aggregated updates rather than pooled raw observations	Stated client partition and local budget	Utility and communication under a different access constraint

The federated rows are not merged into the centralized accuracy ranking because their information path and operational objective differ. They are evaluated with the same reporting discipline but interpreted through utility and communication coordinates appropriate to collaborative learning.

J.3 Method-Specific Operations

The contract does not erase operations that define a method. Alignment losses, meta-adaptation steps, prototype construction, teacher-student objectives, and federated aggregation remain method-specific. They are retained because removing them would turn the comparison into a backbone ablation rather than a comparison of transfer assumptions. The benchmark therefore standardizes the task and budget, not the internal mechanism. Where an original method requires a non-identical interface,

Appendix D.3 records the adaptation and the result is interpreted as a benchmark implementation result.

J.4 Scope of Statistical and Practical Claims

All reported values are fixed-seed point estimates under the stated protocol. The benchmark uses them to identify visible trade-offs across measured axes, not to assert significance for every small difference. Likewise, the extended city-pair results test whether the selection logic remains conditional when source composition changes; they do not establish a universal ranking over cities or tasks. The selection guide is therefore a traceable summary of reported evidence, while its limitations specify the experiments required before broader deployment claims can be made.

The same rule applies to negative degradation values in the missingness and shift diagnostics. A negative value means that the corresponding fixed execution produced a smaller MAE under the perturbed condition than under its reference condition. Without repeated executions, it is treated as an observed fluctuation and not as evidence that deleting information improves learning. Consequently, the diagnostic comparison emphasizes the magnitude and direction of visible changes while avoiding significance claims about values near zero.

J.5 Claim-to-Evidence Trace

The central claims can be reconstructed from specific evidence paths. The selective-transfer claim uses the complete core tables and the strongest target-only value within each task. The ranking-disagreement claim uses the predictive tables with the efficiency and sensitivity figures. The representation claim is limited to STPB similarity and the bounded expert comparison. The federated claim uses the federated accuracy table and the directly recomputed communication-utility table. Finally, the selection guide contains no independent result; every row restates one of these paths.

Table 15 summarizes these paths and states the boundary that accompanies every principal claim.

Table 15: Trace from principal claims to evidence and interpretation boundaries.

Claim	Primary evidence	Boundary
Transfer utility is selective	Core flow and speed model-level tables	Two core tasks, included implementations, fixed execution
Accuracy does not determine sensitivity	Core accuracy tables and sensitivity diagnostics	Reported missingness and city-pair constructions only
Accuracy does not determine efficiency	Accuracy-efficiency figure and recorded resources	Measured hardware and implementation environment
STPB adds an alignment coordinate	STPB figure, specification, and expert comparison	Prototype alignment rather than general interpretability
Federated choices depend on the objective	Federated accuracy and communication-utility tables	Evaluated clients, methods, and communication definition
Selection must remain conditional	All preceding evidence paths	No guarantee for an unseen city, task, or target budget

J.6 Reproducibility Checklist

A complete release for this benchmark should bind every table entry to four items: the dataset version and chronological split, the preprocessing and graph-construction configuration, the method-specific training configuration, and the script that computes prediction and diagnostic metrics. Random seeds and software dependencies belong to the same record because fixed code without a fixed execution environment does not reproduce a ranking. For adapted methods, the release should additionally identify the original assumption that was changed and the interface used in the benchmark.

This checklist separates two notions of fairness that can otherwise be confused. A shared optimization budget improves control because every method receives the same nominal resource allowance. Method-specific published settings may improve absolute performance because architectures do not share the same optimum. CrossCityBench uses the former interpretation for the reported comparison and states this choice explicitly; it does not claim that the resulting values are individually tuned upper bounds. Future releases can add a matched tuning budget without changing the benchmark's task definition.

J.7 Required Artifact Manifest

To make the protocol operational, an artifact release should expose the dependency environment, dataset preparation entry point, immutable split files, graph-construction files, training entry point for each evaluated method, evaluation entry point, STPB construction code, and scripts that regenerate every table and figure. Each result should record the method name, source and target identifiers, seed, target-data budget, prediction horizon, configuration file, and output metric file. This manifest is especially important for adapted or benchmark-defined variants such as STGCN-FT because a method name alone does not determine an implementation.

The manifest also provides a concrete completeness test. A benchmark claim is reproducible only if the corresponding method, configuration, and evaluation path are all present. A repository that contains datasets but omits a winning baseline or diagnostic implementation cannot reproduce the selection guide. The manuscript therefore treats artifact completeness as part of benchmark validity rather than as an optional software supplement.

J.8 Design-Choice Sensitivity for Future Executions

Three design axes should be varied when the benchmark is extended. The first is the target-data budget, which tests whether a recommendation persists as the target moves from severe scarcity toward ordinary supervision. The second is the source composition, which tests whether transfer depends on a particular source city. The third is the optimization allowance, which separates shared-budget comparison from method-specific tuning. Results from these axes should be reported separately because they answer different validity questions.

These analyses are specified as extension requirements and are not presented as completed evidence in the current version. This distinction avoids converting a desirable robustness check into an unsupported result. It also makes clear which additional evidence would be required to upgrade the present conditional conclusions into broader claims.

K Executable Benchmark Interface and Extension Protocol

The preceding traceability analysis identifies the evidence attached to the current tables. This section translates that analysis into an executable interface for future benchmark submissions. The interface does not add empirical claims. Instead, it specifies the records required to determine whether a new result is comparable with the reported evidence.

K.1 Run Manifest and Result Schema

Each execution should produce a machine-readable manifest that binds the prediction file to its data, method, and environment. At minimum, the manifest records the method identifier, code revision, source and target identifiers, chronological split identifier, target-data budget, input and output horizons, random seed, configuration identifier, checkpoint-selection rule, software environment, and hardware record. Efficiency records additionally specify batch size, warm-up, repetitions, synchronization, preprocessing inclusion, precision, and the GPU-memory counter used. A separate result file records MAE, RMSE, and MAPE for each horizon together with the average reported in the main tables. Diagnostic files use the same run identifier so robustness, efficiency, STPB alignment, and federated communication remain bound to the prediction that generated them. Table 16 summarizes the minimum record required to preserve comparability across future CrossCityBench submissions.

The split hash and run identifier are particularly important. A file named only by a dataset and model cannot reveal whether the three-day budget, target-test boundary, or checkpoint rule matches the benchmark. Conversely, a complete manifest permits an auditor to reject an incomparable execution without making a judgment about the model itself.

K.2 Admission and Audit Procedure

A new method is admitted in four steps. First, a protocol check verifies the source-target direction, target budget, chronological split, horizons, and normalization. Second, an access check identifies

Table 16: Minimum execution record for a comparable CrossCityBench submission.

Record block	Required fields	Comparability purpose
Data contract	Dataset version, source, target, split hash, normalization	Prevents changes in observations or target-test access from being attributed to a method
Learning contract	Method, seed, target budget, horizons, configuration, checkpoint rule	Distinguishes architectural gains from additional supervision or selection freedom
Source access	Source observations, parameters, prototypes, teacher outputs, or updates	Makes the transferred information explicit relative to the target-only reference
Metric output	Horizon-level and average MAE, RMSE, and MAPE	Supports direct regeneration of model-level and paradigm-level summaries
Diagnostic output	Mask, pair-shift, runtime, memory, STPB, and optional communication records	Links every secondary coordinate to the prediction run that produced it
Environment	Code revision, dependencies, device, precision, and runtime command	Separates benchmark protocol from implementation and hardware effects

all source-derived information and assigns the method to the appropriate learning paradigm. Third, an output check recomputes the reported metrics from saved target-test predictions. Fourth, a trace check verifies that every diagnostic and selection-guide statement points to a registered execution. Failure at one step does not imply that the method is ineffective, but it prevents direct comparison with the corresponding benchmark table.

The procedure also distinguishes required from optional coordinates. Predictive metrics and source-access disclosure are required for every submission. Missingness and pair-shift diagnostics are required before a method is described as robust. Runtime, memory, and parameter records are required before an efficiency claim is made. Communication records apply only to collaborative methods, while STPB remains an optional representation-alignment probe. This modular rule avoids penalizing an architecture for an inapplicable metric and prevents an absent diagnostic from being interpreted as favorable evidence.

K.3 Protocol-Preserving Extensions

Extensions should change one evaluation axis at a time. A target-budget extension retains the source-target pair, split boundary, horizons, and method configuration while changing the amount of target training data. A source-composition extension retains the target task and budget while changing the available source set. A tuning-budget extension retains the data interface while giving every method the same search allowance. Results from these extensions should be reported in separate blocks because they answer different questions and cannot be pooled into the core matched ranking.

New cities and modalities require an additional dataset card describing sensor semantics, spatial support, sampling interval, missing-value policy, licensing, and the mapping to the common prediction interface. Catalogue inclusion alone does not imply empirical coverage. A dataset contributes to a generalization claim only after the relevant methods are executed under a declared matched contract. This rule preserves the distinction between the breadth of the resource collection and the breadth of validated benchmark evidence.

K.4 Terminology and Abbreviation Policy

General metric and system abbreviations are expanded at first use in the main paper. Model identifiers such as DCRNN, STGCN, AGCRN, ST-GFSL, and FedGTP retain the official names used by their cited publications because several are treated as proper method names rather than manuscript-defined abbreviations. STPB is manuscript-defined and therefore always follows its first-use expansion as Spatio-Temporal Pattern Bank. This convention separates reproducible identifiers from newly introduced terminology and avoids assigning unsupported expansions to published model names.

Throughout the paper, *forecasting* denotes the task or research setting, *prediction* denotes a generated output or its evaluation, and *predict* is used as a verb. Official identifiers such as STG-NCDE and DyHSL are written as plain text without internal spaces or mathematical formatting.

L STPB Specification and Bounded Expert Assessment

Appendix H provides localized context, while the preceding executable interface links each diagnostic to a reproducible run record. We now specify the STPB quantity used across methods and compare its five-model ordering with a five-expert panel. This comparison is descriptive and does not constitute general interpretability validation.

Prototype Construction: Pattern segments are extracted from the PeMS03 and PeMS-BAY construction data, normalized, embedded, and grouped by K-means clustering. Target-test observations are excluded from prototype fitting. The resulting bank is a benchmark reference space rather than a universal catalogue of traffic semantics.

Prototype Selection: The number of clusters is determined by the elbow method, and only prototypes detected in both construction datasets are retained. This recurrence rule reduces dependence on a pattern unique to one construction dataset, but it does not establish universality outside the evaluated data.

Bounded Expert Comparison: Five domain experts rated the displayed model outputs on a 1-to-5 scale. We report only the mean rating and compare the resulting ordering with STPB. The available protocol record does not establish blinded assessment, randomized presentation, or an independently preregistered sampling rule. The comparison is therefore an exploratory sanity check rather than a controlled user study, and no inferential test is reported.

Table 17 records the construction choices, and Table 18 reports the corresponding bounded expert comparison. Reading the two tables together prevents the descriptive rank agreement from being detached from the prototype definition.

Table 17: Operational specification of STPB prototypes.

Item	Setting
Prototype source data	Pattern segments extracted from PeMS03 + PeMS-BAY
Candidate pattern pool	Trend, periodicity, peak-shift, burstiness, local fluctuation segments
Prototype construction	K-means clustering on normalized pattern embeddings
Prototype count and selection	$K = 8$, selected by the elbow method on within-cluster Sum of Squared Errors (SSE)
Prototype retention rule	Retain prototypes detected in both construction datasets
Similarity metric	Average cosine similarity between model representation and prototype vectors

Table 18: Descriptive STPB comparison with five experts; matching ranks apply only to the displayed models.

Model	STPB Similarity	Mean Expert Score	STPB Rank	Expert Rank
D2MHyper	0.0743	4.4	1	1
CrossST	0.0227	3.8	2	2
FGITrans	−0.0226	3.1	3	3
ST-GFSL	−0.0368	2.6	4	4
DyHSL	−0.0473	2.5	5	5

Tables 17 and 18 establish a bounded reading of STPB. The bank is operationally specified by $K = 8$ prototypes constructed from PeMS03 and PeMS-BAY, with recurrence required in both sources. The expert ordering matches the STPB ordering for the five displayed models, and D2MHyper and CrossST receive the two highest values under both views. This exact rank agreement is reported descriptively. The five models do not calibrate a universal STPB threshold, and the comparison is not treated as an independently replicated validation result. It supports STPB only as an exploratory

prototype-alignment proxy in the reported comparison and does not establish statistical significance, general interpretability validity, prototype universality, or an internal causal mechanism.

L.1 Representation and Assessment Reproduction Boundary

Reproducing the STPB coordinate requires more than the two summary tables. A complete artifact should retain the normalized prototype vectors, the model representations supplied to the similarity function, the registered extraction layer and any dimension-mapping rule, the mapping from each representation to its model and city pair, and the script that performs aggregation. These records make it possible to distinguish a change in representation extraction from a change in prototype construction. When a future architecture does not expose a representation compatible with the recorded interface, its STPB field should remain unreported rather than be replaced by an ad hoc projection that changes the diagnostic.

The expert comparison requires a parallel record containing the displayed cases, presentation order, rating prompt, model masking status, individual ratings, and aggregation rule. The current manuscript reports only the bounded descriptive comparison supported by the available protocol record. It therefore does not infer inter-rater reliability, population-level agreement, or causal interpretability. Future executions can strengthen this component through blinded presentation, randomized order, a prespecified sampling rule, and item-level ratings without changing the predictive benchmark itself.

This boundary preserves the intended role of STPB in the diagnostic sequence. Predictive error first establishes whether source access provides utility. Sensitivity and resource coordinates then test whether the observed preference persists under an operating constraint. STPB supplies an optional representation-level comparison only after those measurements and cannot override a negative predictive result. The resulting ordering is one evidence coordinate rather than a substitute for model selection.