WHAT TRANSFERS ACROSS CITIES? INVARIANT REPRESENTATION LEARNING FOR ROBUST SPATIO-TEMPORAL FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-city forecasting requires transferring knowledge across urban environments that share predictive structure yet differ in traffic dynamics, observation density, and latent environmental factors. Many existing approaches align source and target distributions, but alignment can also retain source-specific correlations and induce negative transfer when domain mismatch becomes substantial. We introduce CauCrossNet, an invariant representation transfer framework that separates relatively stable predictive structure from transient environment-specific variation and promotes prediction consistency across alternative stable-space summaries. The learned invariant space provides a basis for target adaptation. Target-side augmentation is conditioned on stable representations, while source knowledge is integrated according to representation-level transfer reliability rather than unconditional similarity. This design is causality-inspired but does not assume an identifiable structural causal model or require interventional observations. Across multiple city pairs, we evaluate standard transfer, limited-target-data adaptation, temporal shifts, and longer horizons. Results show that selective invariant transfer improves forecasting across source-target relations, remains effective under mismatch and limited target data, and reduces degradation from poorly matched source information through reliability-aware integration. These findings suggest that robust cross-city forecasting depends not only on reducing domain discrepancy, but also on identifying predictive structure that remains stable, useful, and transferable across environments. Code and experimental configurations are available at the anonymous repository: `https://anonymous.4open.science/r/CauCrossNet-040B`.

1 INTRODUCTION

Cross-city spatio-temporal forecasting asks a model trained with data from one urban environment to remain useful in another. Data-rich cities can provide valuable inductive structure for cities with short histories or sparse sensing, which motivates few-shot and cross-city transfer protocols (Lu et al., 2022; Liu et al., 2025; Cui et al., 2026). The difficulty is that two cities rarely differ only by a simple covariate shift. Network geometry, sensor placement, traffic demand, weather, calendar effects, and latent environmental factors can change together, so a source representation can mix reusable predictive structure with city-specific correlations (Ouyang et al., 2025; Shao et al., 2025).

Most cross-city methods address this mismatch by aligning source and target distributions, adapting model parameters, or learning cross-city correspondences (Tang et al., 2022; Ouyang et al., 2023; Deng et al., 2026). These strategies are effective when the aligned features are genuinely transferable. They are less reliable when transient source patterns happen to correlate with the target. Strong alignment can then preserve the wrong information, and increasing source-target mismatch can turn positive transfer into negative transfer. The central question is therefore not only *how much* source information should be aligned, but *what* should remain transferable across environments.

Figure 1 summarizes this perspective. A source city contains predictive factors with different degrees of environmental stability. Some patterns, such as recurrent demand transitions around similar functional regions, can remain useful after transfer. Other patterns depend strongly on local topology,

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.

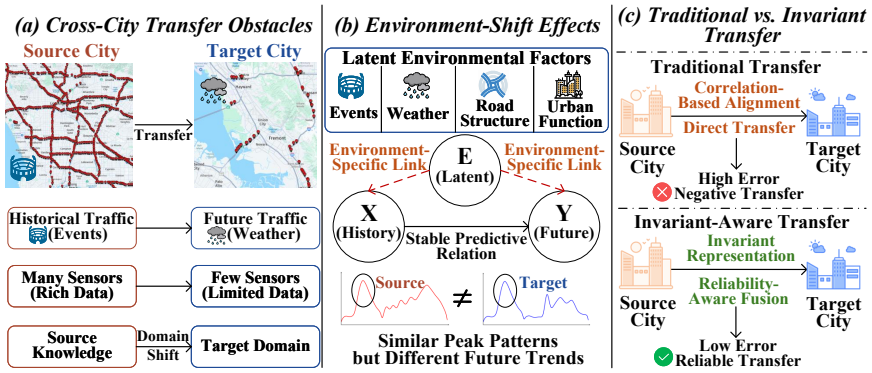


Figure 1: Motivation for invariant cross-city transfer. Observational city shifts mix reusable predictive structure with environment-specific correlations; stable representations and reliability-aware fusion are used to reduce negative transfer.

sensing density, or temporary context. We consequently view cross-city forecasting as an invariant representation transfer problem: learn a representation that preserves relatively stable predictive information, separate it from transient environment-specific variation, and use the resulting representation to control both target adaptation and source integration.

This viewpoint is causality-inspired but deliberately weaker than causal identification. Hidden environmental factors motivate separating stable from transient predictive representations, yet standard traffic benchmarks do not provide interventions that would identify a structural causal model. We therefore do not interpret a learned invariant representation as a causal variable, and we do not claim that representation alignment blocks all confounding. Instead, alternative stable-space summaries provide a practical consistency signal: if predictions remain stable across these summaries, the model is encouraged to rely less on brittle source-specific shortcuts. This distinction follows the broader lesson that causal interpretations require assumptions beyond observational predictive success (Pearl, 2009; Zhao et al., 2023a; Zhu et al., 2024; Abbavaram et al., 2026).

We instantiate this principle in CauCrossNet and interpret its role through invariant representation transfer. To keep notation consistent with the implementation, we retain three module identifiers and introduce them by function: the stable-representation distillation module (**CauKD**), the invariant-conditioned curriculum module (**CauCL**), and reliability-aware fusion (**RAFusion**). These functional descriptions do not imply causal identification. CauKD learns stable and transient representations and regularizes prediction consistency across alternative summaries. CauCL conditions target-side augmentation and curriculum scheduling on the stable representation. RAFusion estimates transfer reliability from invariant-space mismatch and gates source and target predictions. The modules form a directional pipeline in which representation learning defines the space used by later adaptation and fusion.

The resulting contribution is not a new claim that cross-city forecasting requires an identifiable causal graph. It is a transfer principle with three concrete consequences. First, domain alignment is applied to a separated invariant branch rather than to all source features. Second, target augmentation is conditioned on the same stable branch instead of unrestricted source or target embeddings. Third, source information is weighted by representation-level reliability rather than by unconditional similarity. Across standard transfer, limited-target-data adaptation, natural temporal shifts, and longer horizons, the experiments test when this selective transfer is preferable to ordinary alignment.

Our contributions are summarized as follows.

- We formulate cross-city forecasting as **invariant representation transfer**, separating stable predictive structure from transient environment-specific variation without assuming that either component is causally identifiable from observational traffic data.
- We develop a **coupled invariant-transfer representation** that combines stable/transient separation, prototype-mediated prediction consistency, and source-target invariant alignment. The same

stable space governs target augmentation and later source-admission decisions rather than acting as an isolated auxiliary representation.

- We introduce **reliability-aware source integration**, where transfer weights depend on mismatch in the invariant space together with predictive uncertainty. This explicitly reduces source contribution when the learned source structure is poorly matched to the target city.
- We evaluate the framework on traffic flow and speed transfer across multiple city pairs, including standard, limited-data, natural-shift, and longer-horizon settings, and provide a detailed appendix with additional pairs, baseline descriptions, sensitivity, efficiency, and statistical analyses.

2 Related Work

Cross-city spatio-temporal forecasting. Cross-city traffic forecasting has been studied through adversarial alignment, graph transfer, few-shot adaptation, and cross-city correspondence learning. Domain Adversarial Spatial-Temporal Network (DASTNet) learns domain-shared embeddings (Tang et al., 2022); Domain Adversarial Graph Network (DAGN) learns cross-city graph structure (Ouyang et al., 2023); Cross-City Correlation Learning (CCL) discovers transferable node and region correspondences (Deng et al., 2026); and Dynamic Multi-Scale Hypergraph learning (D2MHyper) combines higher-order structure with domain adversarial training (Ouyang et al., 2025). Few-shot alternatives transfer meta-knowledge or traffic-pattern banks to data-scarce targets (Lu et al., 2022; Liu et al., 2025). PLM-CUP recently uses a pre-trained language model as an auxiliary source domain for cross-city urban-flow transfer (Zhou et al., 2026); its published protocol uses rasterized city-flow grids rather than road-sensor graphs, so we treat it as complementary rather than mixing cross-protocol numbers. Our focus is selective reuse in graph-based sensor forecasting, where representation roles are separated before source information is gated.

Invariant representation learning under shift. Domain generalization asks models to preserve predictive structure when test distributions differ from training (Gulrajani & Lopez-Paz, 2021). Traffic-specific work includes CauSTG, which captures invariant spatio-temporal relations for out-of-distribution forecasting (Zhou et al., 2023), STEVE, which learns basis confounder representations for robust traffic prediction (Ji et al., 2025), and CCHMM, a deep latent causal generative model for multimodal traffic prediction (Zhao et al., 2023a). These studies motivate robust structure learning but do not make observational benchmarks sufficient for identifying a structural causal model. Recent representation analysis further shows that hidden-confounder shifts can make strict invariance insufficient because environment-specific information may remain useful (Abbavaram et al., 2026). We therefore retain a transient branch and treat prediction consistency across alternative stable-space summaries as a robustness regularizer, not a front-door estimator.

Transfer reliability and negative transfer. Negative transfer occurs when source features are poorly matched to the target task. Recent work studies both explicit safeguards against negative transfer and representation-level estimates of source-target compatibility (Sun et al., 2026; Xu et al., 2026; Diniz et al., 2026). These results support a broader shift from unconditional reuse toward selective integration. Our setting differs because compatibility is estimated inside a spatio-temporal forecasting model: invariant-space mismatch and predictive uncertainty jointly modulate source contribution, while a target-specific branch preserves information that need not transfer. Reliability is therefore tied to the representation the framework explicitly intends to share across cities.

3 Invariant Representation Transfer

3.1 Problem Formulation

Let $e \in \{S, T\}$ denote source and target cities. A city provides a multivariate traffic history $\mathbf{X}^e_{1:L} \in \mathbb{R}^{L \times N_e \times D}$ and future states $\mathbf{Y}^e_{L+1:L+H}$. N_e is the number of sensors, D the input dimension, L the look-back length, and H the prediction horizon. The source city has substantially more usable observations, while the target city may provide only a short adaptation interval. The goal is a target predictor for $\mathbf{Y}^T$ that benefits from source data without assuming that all source features transfer.

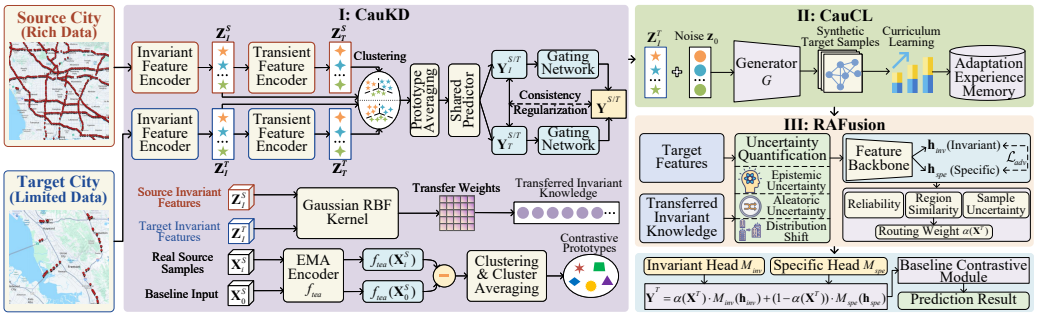


Figure 2: Framework overview. CauKD learns stable and transient representations, CauCL performs invariant-conditioned target augmentation, and RAFusion integrates source information according to representation-level transfer reliability. The module identifiers are retained for implementation consistency; the interpretation is predictive and does not assume structural causal identification.

We describe each city and temporal regime by an environment variable E. The observational distribution $P_E(\mathbf{X}, \mathbf{Y})$ can change across cities and time. Rather than imposing equality of complete feature distributions, we seek a representation decomposition

$$(\mathbf{Z}_I^e, \mathbf{Z}_E^e) = F(\mathbf{X}^e), \tag{1}$$

where $\mathbf{Z}_I$ is encouraged to retain relatively stable predictive structure and $\mathbf{Z}_E$ captures residual environment-specific variation. The subscript I denotes the intended invariant role and does not imply causal identifiability.

The transfer problem then has three parts. We first learn $\mathbf{Z}_I$ so that it remains predictive and comparatively stable across source and target environments. We then condition target augmentation on $\mathbf{Z}_I^T$. Finally, source predictions are fused with target predictions according to a reliability score computed from invariant-space mismatch. This ordering prevents later modules from defining transferability independently of the representation learner.

Definition 1 (Operational transferability)**.** *For the observed family of source, target, and temporal environments, a representation is* operationally transferable *when it satisfies three empirical requirements: it remains prediction-relevant after target adaptation, its source-target discrepancy is controlled in the representation space intended for transfer, and admitting source information through the learned gate does not reduce validation utility relative to the target-local path. This definition is deliberately operational. It does not assert population-level invariance, causal sufficiency, or uniqueness of the learned representation.*

This definition connects the title question to the architecture. During training, observed source and target histories are encoded into stable and transient branches; CauCL uses the stable target representation only for augmentation and curriculum construction, while supervised forecasting is computed on observed target history-future pairs. RAFusion then combines source-transfer and target-local predictions using mismatch and uncertainty derived from observed histories and model outputs. At inference, target generation is inactive and no ground-truth future enters the gate.

Why selective alignment is necessary. Let $\mathbf{Z} = (\mathbf{Z}_I, \mathbf{Z}_E)$. Aligning the complete representation does not reveal which coordinates are reusable: it can suppress target-specific structure or preserve a source shortcut. We therefore apply the cross-city discrepancy penalty only to $\mathbf{Z}_I$ and leave $\mathbf{Z}_E$ free to encode residual variation. This is an architectural restriction, not a proof of invariance. Its empirical test is predictive: transfer should remain useful across environment changes, and reliability-aware source weighting should matter more as mismatch increases.

3.2 Stable and Transient Representation Learning

The invariant learner uses two parallel spatio-temporal encoders with separate attention projections. Here, $\mathrm{Attn}(\cdot)$ denotes the shared attention operator with branch-specific query, key, and value pro-

jections. For a source sequence, the two branches are

$$\begin{aligned}\mathbf{H}_I^S &= \mathrm{Attn}(\mathbf{X}^S\mathbf{W}_Q^I, \mathbf{X}^S\mathbf{W}_K^I, \mathbf{X}^S\mathbf{W}_V^I),\\ \mathbf{H}_E^S &= \mathrm{Attn}(\mathbf{X}^S\mathbf{W}_Q^E, \mathbf{X}^S\mathbf{W}_K^E, \mathbf{X}^S\mathbf{W}_V^E),\end{aligned} \tag{2}$$

with the same construction applied to the target city. Temporal stacking and projection produce $\mathbf{Z}_I^e$ and $\mathbf{Z}_E^e$. The branches are not labeled by observed causal variables. Their roles are learned through consistency, alignment, and the downstream transfer objective.

Prediction consistency across stable-space summaries. A single representation can satisfy a domain-alignment loss while still containing source-specific shortcuts. We therefore construct K prototype summaries $\{\mathbf{p}_{I,k}^e\}_{k=1}^K$ from the stable branch and evaluate a shared prediction head g_θ through alternative summaries. Let $\hat{\mathbf{Y}}^e$ denote the ordinary prediction and $\hat{\mathbf{Y}}_{I,k}^e = g_\theta(\mathbf{p}_{I,k}^e, \mathbf{Z}_E^e)$ the prediction routed through prototype k. We use

$$\mathcal{L}_{\mathrm{cons}} = \sum_{e\in\{S,T\}} \frac{1}{K}\sum_{k=1}^{K} \left\| \hat{\mathbf{Y}}_{I,k}^e - \hat{\mathbf{Y}}^e \right\|_2^2. \tag{3}$$

This objective encourages the prediction to remain stable across multiple summaries of the intended transferable space. It is not optimized in isolation: supervised forecasting preserves prediction relevance, while invariant alignment and branch separation constrain the roles assigned to $\mathbf{Z}_I$ and $\mathbf{Z}_E$. A solution that simply ignores the stable branch is therefore not the intended optimum of the coupled objective, although the formulation does not provide a formal non-collapse guarantee. The consistency term is motivated by alternative predictive pathways but is not a front-door estimator and does not identify an intervention effect.

Invariant-space alignment. We compare temporally pooled source and target stable representations using maximum mean discrepancy (MMD):

$$\mathcal{L}_{\mathrm{inv}} = \mathrm{MMD}^2(P(\mathbf{Z}_I^S), P(\mathbf{Z}_I^T)). \tag{4}$$

For node-level correspondence, a Gaussian radial basis function (RBF) kernel defines

$$\phi_I(\mathbf{z}_{I,i}^S, \mathbf{z}_{I,j}^T) = \exp\!\left(-\lambda\|\mathbf{z}_{I,i}^S - \mathbf{z}_{I,j}^T\|_2^2\right), \tag{5}$$

followed by a temperature-normalized source weight. Unlike unconditional similarity, this correspondence is computed after stable/transient separation.

Baseline-centered prototypes. To reduce dependence on city-wide offsets, an exponential moving average (EMA) teacher forms centered prototypes

$$\mathbf{p}_k = \frac{1}{|\mathcal{C}_k|}\sum_{\mathbf{X}_i\in\mathcal{C}_k} \left[f_{\mathrm{tea}}(\mathbf{X}_i) - f_{\mathrm{tea}}(\mathbf{X}_0)\right], \tag{6}$$

where $\mathbf{X}_0$ is a fixed baseline pattern and $\mathcal{C}_k$ a representation cluster. Centering is a robustness device, not a counterfactual intervention.

3.3 Invariant-Conditioned Target Augmentation

Limited target data makes it difficult to estimate cross-city differences directly. CauCL therefore augments target observations only after the stable representation has been learned. A conditional generator receives noise $\mathbf{z}$ together with $\tilde{\mathbf{Z}}_I^T$ and produces $\tilde{\mathbf{X}}^T = G(\mathbf{z} \mid \tilde{\mathbf{Z}}_I^T)$. Its objective combines the ordinary generative adversarial network (GAN) term with target-distribution matching and baseline consistency:

$$\begin{aligned}\mathcal{L}_{\mathrm{aug}} = \mathcal{L}_{\mathrm{GAN}} &+ \lambda_m \,\mathrm{MMD}^2(P(\tilde{\mathbf{X}}^T), P(\mathbf{X}^T))\\ &+ \lambda_b\|G(\mathbf{z} \mid \tilde{\mathbf{Z}}_I^T) - G(\mathbf{z}_0 \mid \tilde{\mathbf{Z}}_I^T)\|_2^2.\end{aligned} \tag{7}$$

The conditioning variable is the key restriction: augmentation cannot freely exploit the source-specific branch. Generated target observations are used by the augmentation objective and curriculum construction; supervised $\mathcal{L}_{\mathrm{pred}}$ is always evaluated on observed target history-future pairs. Thus

CauCL affects forecasting through target-side representation regularization and sample scheduling and does not require synthetic future labels or pseudo-labels.

Training examples are introduced through a curriculum based on two observable quantities: predictive dispersion under stochastic forward passes and representation mismatch to target invariant prototypes. For target sample i,

$$d_i = \alpha \operatorname{Var}_b(\hat{\mathbf{Y}}_i^{(b)}) + \beta \operatorname{MMD}^2(\mathbf{Z}_{I,i}^T, \mathcal{P}_I^T). \tag{8}$$

A threshold γ_t expands as training stabilizes, so examples close to the learned target invariant structure are introduced before higher-uncertainty examples. Memory-assisted adaptation is described in Appendix A.

3.4 Reliability-Aware Source Integration

A representation can be approximately invariant on average while still transferring poorly to a particular target city. RAFusion therefore estimates reliability from stable-space mismatch. The global shift score is

$$u_{\text{shift}} = \operatorname{MMD}^2(P(\mathbf{Z}_I^S), P(\mathbf{Z}_I^T)), \qquad R_g = \exp(-\gamma u_{\text{shift}}). \tag{9}$$

For target node j, R_j denotes node-level transfer reliability after combining the global shift discount R_g with compatibility to source invariant prototypes. We keep the regional term s_j separate because it represents local structural compatibility rather than the global-to-node reliability score. Predictive uncertainty is estimated through stochastic forward passes, yielding epistemic uncertainty u_{ep}; an auxiliary variance head provides aleatoric uncertainty u_{al}. The fusion gate is

$$\alpha_j = \sigma(w_1 R_j + w_2 s_j - w_3 u_{ep,j} - w_4 u_{al,j}), \tag{10}$$

where s_j is invariant-space regional similarity. All gate inputs are computed from observed histories, learned representations, source prototypes, or stochastic model outputs; no ground-truth future $\mathbf{Y}^T$ is used at inference. Thus R_j discounts source reuse by transfer reliability, whereas s_j captures local structural match. The final target prediction is

$$\hat{\mathbf{Y}}_j^T = \alpha_j \hat{\mathbf{Y}}_j^{S\to T} + (1-\alpha_j)\hat{\mathbf{Y}}_j^T. \tag{11}$$

Thus source information receives less weight when representation mismatch or uncertainty is high. This implements the abstract principle that transfer reliability is representation-level rather than an unconditional source-target similarity score.

3.5 Optimization and Interpretation

The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_1 \mathcal{L}_{\text{cons}} + \lambda_2 \mathcal{L}_{\text{inv}} + \lambda_3 \mathcal{L}_{\text{aug}} + \lambda_4 \mathcal{L}_{\text{dis}}, \tag{12}$$

where $\mathcal{L}_{\text{pred}}$ is the supervised forecasting loss and $\mathcal{L}_{\text{dis}}$ separates the intended stable and environment-specific representation roles. Optimization proceeds in three stages: representation learning, target augmentation, and joint reliability-aware refinement. Appendix A specifies the separation objective, information flow, and optimization sequence.

The framework makes a predictive invariance claim rather than a causal-identification claim. A low $\mathcal{L}_{\text{inv}}$ means the learned stable representations are close under the selected discrepancy, not that all confounders have been removed. Likewise, the consistency loss encourages robustness across alternative representation summaries but does not identify a mediator in the causal-inference sense. These boundaries matter because the evaluation contains observational city shifts rather than randomized interventions.

4 Experiments

4.1 Experimental Protocol and Evaluation Axes

The experimental design is organized around the title question rather than around one aggregate accuracy score. We evaluate five complementary axes. **E1, standard transfer**: does selective invariant transfer improve forecasting when a fixed target adaptation budget is available? **E2, representation role**: which components support the selective-transfer pipeline? **E3, temporal stability**:

does the learned representation remain predictive across workday and non-workday regimes? **E4, mismatch and scarcity**: how does the model behave under limited target data and a cross-region out-of-distribution (OOD) transfer? **E5, stress and cost**: do gains persist at longer horizons and under realistic efficiency constraints? This organization separates ordinary predictive accuracy from evidence about when source information should be reused.

Datasets and protocol. We use three traffic-flow datasets from the California Performance Measurement System (PeMS), namely PeMS03, PeMS04, and PeMS08, and two traffic-speed datasets, METR-LA and PeMS-BAY. All use 5-minute intervals. The model observes 12 steps and predicts the next 12. The chronological split is 7:1:2, and standard target adaptation uses seven days of target training data. Flow transfer uses PeMS03 as the primary source with PeMS08 and PeMS04 as targets. Speed transfer evaluates both PeMS-BAY→METR-LA and METR-LA→PeMS-BAY. All data are z-score normalized using training statistics. Dataset periods, transfer roles, and leakage safeguards are in Appendix B.

Training uses AdamW with learning rate 10^{-3}, batch size 16, hidden dimension 64, up to 100 epochs, and early stopping. Experiments run on an NVIDIA H100 graphics processing unit (GPU). We report mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). Tables report means over five runs. The principal tables contain aggregated means rather than per-run dispersion, so we do not attach unverified $\pm$std values; significance claims are restricted to the post-hoc comparisons for which test statistics are reported.

Baselines. Nineteen baselines cover traditional, single-city, and cross-city methods. Traditional methods are gradient boosted regression trees (GBRT) and vector autoregression (VAR). Single-city baselines include Adaptive Graph Convolutional Recurrent Network (AGCRN), AllDeepSet, Diffusion Convolutional Recurrent Neural Network (DCRNN), Dynamic Hypergraph Structure Learning (DyHSL), Gated Recurrent Unit (GRU), Graph WaveNet (GWNet), Spatio-Temporal Graph Convolutional Network (STGCN), and Spatio-Temporal Graph Neural Controlled Differential Equation (STG-NCDE). Cross-city baselines include CCL, DASTNet, DAGN, D2MHyper, Model-Agnostic Meta-Learning (MAML), Multi-Scale Traffic Pattern Bank (MTPB), Spatio-Temporal Deep Attentive Adaptation Network (ST-DAAN), Spatio-Temporal Graph Few-Shot Learning (ST-GFSL), and STGCN fine-tuning (STGCN-FT) (Bai et al., 2020; Chien et al., 2022; Li et al., 2018; Zhao et al., 2023b; Ouyang et al., 2025; Deng et al., 2026; Tang et al., 2022; Lu et al., 2022). Appendix C gives canonical references and comparison roles.

4.2 Cross-City Transfer Across Primary Pairs (E1)

Table 1 reports the average across all 12 forecast steps for the two primary directions; the displayed 15, 30, and 60 minute checkpoints are diagnostic horizons rather than the values averaged to form the "Average" column. Complete horizon-level results for all four directions are retained in Appendix D.

On PeMS03→PeMS08, D2MHyper is the strongest baseline by average MAE at 17.55, while the proposed framework reaches 14.64, a 16.6% reduction. On PeMS-BAY→METR-LA, the proposed model obtains MAE 2.03, RMSE 4.71, and MAPE 5.40%, compared with CCL at MAE 2.41. The additional flow target and reverse speed direction preserve the same average-MAE ordering, but the gain varies by pair. This heterogeneity is important: the empirical claim is not that one fixed transfer rule yields a constant gain, but that explicitly deciding what to transfer remains useful across heterogeneous source-target relations.

4.3 What Is Being Transferred? Components and Temporal Stability (E2, E3)

Figure 3(a) reports coarse and fine ablations on PeMS03→PeMS08. Removing CauCL raises average MAE from 14.64 to 19.55; removing invariant consistency regularization (ICR), Dynamic Graph Augmentation (DGA), or reliability routing (RR) also degrades performance. The fine variants keep the forecasting backbone, data split, and optimization protocol fixed while disabling one internal mechanism, so they are matched within-model controls for component necessity. They do not replace separately trained target-only, generic-MMD, or DANN controls, and we therefore do not attribute the full cross-city gain to one mechanism in isolation; Appendix G reports the exact deltas.

Table 1: Average standard-transfer performance. MAPE is reported in percent. Best results are bold and the strongest baseline is underlined when unambiguous.

Method	PeMS03→PeMS08			PeMS-BAY→METR-LA		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE
GBRT	29.69	47.36	18.10	9.74	18.11	18.33
VAR	32.88	49.15	20.32	9.46	16.86	17.14
AGCRN	28.06	48.04	14.53	6.46	13.00	11.97
AllDeepSet	26.75	38.68	19.77	4.10	8.04	11.36
DCRNN	20.56	31.68	14.85	4.10	8.10	11.67
DyHSL	19.19	29.59	14.67	3.81	7.78	10.70
GRU	30.05	40.83	34.05	4.37	8.53	12.35
GWNet	26.59	38.02	23.78	4.02	7.74	12.47
STGCN	24.64	38.03	15.98	4.06	8.07	11.60
STG-NCDE	18.65	29.36	14.66	4.77	10.05	10.13
CCL	18.03	27.98	<u>11.86</u>	<u>2.41</u>	5.61	<u>6.47</u>
DASTNet	21.17	32.00	15.50	4.60	9.00	12.00
DAGN	18.39	27.90	14.51	3.33	6.30	8.57
D2MHyper	<u>17.55</u>	<u>27.00</u>	13.44	2.75	<u>5.30</u>	7.50
MAML	24.91	34.43	26.14	4.91	9.01	14.85
MTPB	24.50	35.42	16.17	3.76	6.95	10.02
ST-DAAN	22.47	33.58	15.18	3.82	7.48	10.90
ST-GFSL	23.78	33.69	20.29	4.92	9.18	14.40
STGCN-FT	21.95	32.67	16.71	3.95	7.76	11.44
Ours	**14.64**	**25.55**	**9.91**	**2.03**	**4.71**	**5.40**

Table 2: Predictive consistency across workday and non-workday environments.

Method	Avg. MAE	$\Delta\mathrm{MAE}_{env}$	Worst MAE
VAR	32.96	6.78	35.15
DyHSL	20.07	4.13	21.41
CCL	18.59	4.06	19.91
Ours	**14.60**	**1.34**	**15.03**

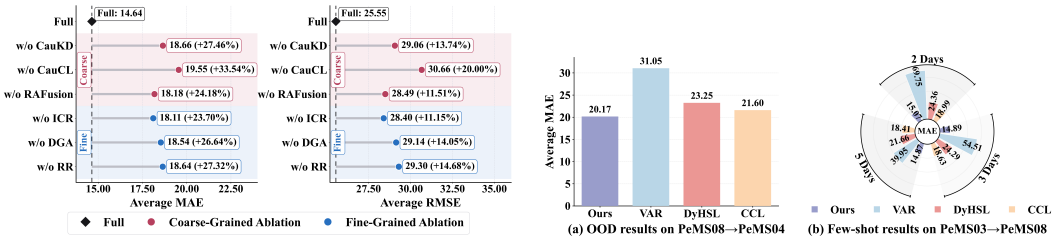
To evaluate predictive stability without asserting causal recovery, we partition PeMS03→PeMS08 by workday and non-workday environments. Table 2 reports average MAE, the environment gap $\Delta\mathrm{MAE}_{env}$, and worst-environment MAE. The proposed model has the lowest average error, the smallest environment gap, and the lowest worst-environment error among the listed methods.

The environment gap is deliberately predictive and is not a hidden-confounder intervention. It tests whether the transferred representation remains useful across a natural temporal regime change; it does not demonstrate removal of an unobserved confounder, identify a causal mechanism, or establish invariance outside the observed environments.

4.4 Target Scarcity and Cross-Region Mismatch (E4)

Figure 3(b) evaluates two difficult regimes. The cross-region OOD setting transfers PeMS08 to PeMS04, while the few-shot setting transfers PeMS03 to PeMS08 using 2, 3, or 5 days of target data. Under OOD transfer, the proposed framework attains MAE 20.17, compared with 21.60 for CCL, 23.25 for DyHSL, and 31.05 for VAR. Under few-shot transfer, its MAE remains near 14.9 across the three reported target-data budgets. These settings probe different failure pressures: the OOD transfer changes the city pair, whereas few-shot adaptation reduces the evidence available to estimate target structure.

The panels are complementary: the ablation tests which transfer components are necessary, whereas the OOD and few-shot analysis tests whether their value persists under stronger mismatch or reduced target evidence.



(a) Ablation on PeMS03→PeMS08. (b) OOD and few-shot MAE.

Figure 3: Component and robustness analyses. The ablation covers modules and internal components; OOD and few-shot tests stress cross-region mismatch and target-data scarcity.

4.5 Cross-Regime Quantitative Evidence for Transferability (E5)

Table 3 consolidates the principal quantitative stress tests using the lowest-error comparator reported under each corresponding protocol. All entries use smaller-is-better metrics, so the final column reports the relative reduction of the proposed framework. The table is a synthesis of already reported experiments rather than a new pooled metric.

Table 3: Cross-regime quantitative evidence. Each comparator is the lowest-error baseline reported under the same protocol.

Evaluation axis	Metric	Ours	Best comparator	Reduction
PeMS03→PeMS08	Avg. MAE	14.64	17.55 (D2MHyper)	16.6%
PeMS-BAY→METR-LA	Avg. MAE	2.03	2.41 (CCL)	15.8%
PeMS03→PeMS04	Avg. MAE	20.33	21.62 (CCL)	6.0%
METR-LA→PeMS-BAY	Avg. MAE	1.26	1.53 (D2MHyper)	17.6%
Workday/non-workday	$\Delta\mathrm{MAE}_{env}$	1.34	4.06 (CCL)	67.0%
Cross-region OOD	MAE	20.17	21.60 (CCL)	6.6%
150-min flow forecast	MAE	20.40	34.47 (CCL)	40.8%

Gains vary across city pairs, while the workday/non-workday gap contracts from 4.06 for CCL to 1.34 and the 150-minute flow MAE improves by 40.8% over CCL. These results support source-target-dependent selective transfer rather than a fixed transfer gain. The model has 0.65M parameters and requires 2.04 GPU-hours in the reported training configuration. The 7.22 ms value is retained only as implementation timing and is not used for cross-method efficiency claims. The post-hoc analysis does not support a pairwise superiority claim over CCL. We also do not claim standalone uncertainty-calibration superiority or robustness to the four loss weights λ_1–λ_4, because neither is directly benchmarked. Additional details are in Appendices E–G.

Attribution boundary. The evidence supports the complete pipeline, but it does not directly identify the roles of $\mathbf{Z}_I$ and $\mathbf{Z}_E$, benchmark RAFusion as a calibrated uncertainty estimator, sweep λ_1–λ_4, or replace matched target-only, generic-MMD, and DANN controls. We therefore avoid attributing the full cross-city gain to any single mechanism.

5 Conclusion

We study cross-city spatio-temporal forecasting by asking what should transfer rather than aligning complete domains. CauCrossNet uses separate representation branches for intended stable and environment-specific information, adapts the target in the stable space, and gates source reuse by transfer reliability. Across four transfer directions and multiple stress regimes, the results support selective transfer when unconditional reuse becomes unreliable. The evidence concerns predictive robustness over the evaluated environments, not identification of a structural causal mechanism. This framing separates practical transfer gains from stronger causal interpretations across the evaluated transfer regimes.

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 study uses public traffic sensor benchmarks and does not involve human-subject recruitment or personally identifiable information. The framework is intended for research on cross-city forecasting and should not be deployed by itself for safety-critical traffic control. Performance may change with sensing coverage, network topology, infrastructure, demographic travel patterns, and data availability across cities. The learned invariant representation is a predictive object and should not be interpreted as an identified causal structure. Operational use would require local validation, monitoring under distribution shift, and assessment of whether transferred source information is appropriate for the target environment.

REPRODUCIBILITY STATEMENT

Section 4 reports the data split, target-data budget, prediction horizon, optimization settings, evaluation metrics, and primary baselines. Appendices A and B document notation, module dependencies, progressive optimization, the branch-separation implementation boundary, operational transferability criteria, terminology boundaries, alternative design assumptions, dataset periods, transfer directions, and leakage controls. Appendix D reports horizon-level results for all four city-pair directions, while Appendices F, G, H, and I provide sensitivity, efficiency, representation diagnostics, statistical details, failure modes, threats to validity, and a cross-regime analysis. Code and experimental configurations are available through the anonymous repository linked in the abstract.

REFERENCES

Gowtham Reddy Abbavaram, Celia Rubio-Madrigal, Rebekka Burkholz, and Krikamol Muandet. When shift happens - confounding is to blame. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d285ab51e167a923f07e6137d802305a-Abstract-Conference.html.

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.

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.

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.

Jeongwhan Choi, Hwangyong Choi, Jeehyun Hwang, and Noseong Park. Graph neural controlled differential equations for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6367–6374, 2022.

Zhe Cui, Di Zang, Junqi Zhang, and Keshuang Tang. Cross-city and predictive resilience assessment for urban traffic networks via transfer learning. *Engineering Applications of Artificial Intelligence*, 164:113286, 2026.

Zhiyuan Deng, Zhe Wu, Li Su, Xinfeng Zhang, Yaowei Wang, and Qingming Huang. Cross-city correlation learning for traffic forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 27(2):2252–2267, 2026. doi: 10.1109/TITS.2025.3647158.

Raphaella Diniz, Jackson de Faria, and Martin Ester. PAS: Estimating the target accuracy before domain adaptation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d9b1e3397492bed99c1bf7c355726d8b-Abstract-Conference.html.

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*, pp. 1126–1135, 2017.

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

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Jiahao Ji, Wentao Zhang, Jingyuan Wang, and Chao Huang. Seeing the unseen: Learning basis confounder representations for robust traffic prediction. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 577–588, 2025.

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.

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):94:1–94:24, 2025. 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.

Helmut Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, Berlin, Heidelberg, 2005.

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.

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.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

Junqi Shao, Shen Li, Ke Zhang, Anyou Wang, and Meng Li. Cross-city traffic prediction based on deep domain adaptive transfer learning. *Transportation Research Part C: Emerging Technologies*, 176:105152, 2025.

Hui Sun, Zheng Xie, Hao-Yuan He, and Ming Li. Mitigating negative transfer via reducing environmental disagreement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7): 8485–8502, 2026.

Yihong Tang, Ao Qu, Andy H. F. Chow, William H. K. Lam, Sze Chun 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.

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.

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.

Dongwei Xu, Yufu Tang, Jingfei Ju, Zefeng Yu, Jiaying Zheng, Tongcheng Gu, and Haifeng Guo. Cross-city traffic state prediction based on knowledge transfer framework. *Expert Systems with Applications*, 272:126747, 2025.

Yichen Xu, Ryumei Nakada, Linjun Zhang, and Lexin Li. Residual feature integration is sufficient to prevent negative transfer. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/7d998ff3e6ecce821986c9dd876be9bc-Abstract-Conference.html.

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.

Yu Zhao, Pan Deng, Junting Liu, Xiaofeng Jia, and Mulan Wang. Causal conditional hidden markov model for multimodal traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4929–4936, 2023a.

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

Qiang Zhou, Xudong Tong, Yuting Liu, Chuanxing Liu, and Jingjing Gu. Exploiting pre-trained language model for cross-city urban flow prediction guided by information-theoretic analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16513–16521, 2026. doi: 10.1609/aaai.v40i19.38691.

Zhengyang Zhou, Qihe Huang, Kuo Yang, Kun Wang, Xu Wang, Yudong Zhang, Yuxuan Liang, and Yang Wang. Maintaining the status quo: Capturing invariant relations for OOD spatiotemporal learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3603–3614, 2023.

Qihui Zhu, Shenwen Chen, Tong Guo, Yisheng Lv, and Wenbo Du. A spatio-temporal approach with self-corrective causal inference for flight delay prediction. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):20820–20831, 2024.

SUPPLEMENTARY MATERIAL

Appendix roadmap. The supplement follows the representation-to-evidence-to-reproducibility logic of the main paper rather than repeating it. Appendix A specifies the method construction, optimization sequence, representation roles, inference-time information flow, and operational transferability criteria. Appendix B records the datasets, transfer directions, metric aggregation, and partition safeguards, while Appendix C separates broader baseline context from the comparison roles used in the experiments. Appendices D–G provide complementary evidence through complete horizon-level transfer results, longer-horizon analysis, sensitivity and efficiency, and representation visualization with statistical analysis. Appendices H and I consolidate discussion, limitations, failure modes, threats to validity, and transferability across evaluation regimes. Appendix J closes the supplement with the evidence boundaries and reproducibility requirements associated with the main claims.

A ADDITIONAL METHOD DETAILS

A.1 NOTATION AND MODULE DEPENDENCIES

Table 4 summarizes the symbols that connect representation learning, augmentation, and reliability-aware fusion. The same stable representation $\mathbf{Z}_I$ is reused by all three modules, which is important for interpreting the method as a single transfer pipeline rather than three independent regularizers.

Table 4: Notation used by the invariant representation transfer pipeline.

Symbol	Meaning
$\mathbf{Z}_I^e$	Representation intended to carry stable predictive structure in city e
$\mathbf{Z}_E^e$	Environment-specific residual representation in city e
$\mathcal{L}_{\text{cons}}$	Prediction consistency across stable-space prototype summaries
$\mathcal{L}_{\text{inv}}$	Source-target MMD penalty applied to $\mathbf{Z}_I$
$\mathcal{L}_{\text{aug}}$	Invariant-conditioned target augmentation objective
R_j	Target-node transfer reliability derived from invariant-space mismatch
α_j	Reliability gate between transferred-source and target-local predictions

The dependency order is deliberate. CauKD determines the representation in which transferability is judged. CauCL can then synthesize target observations only through $\mathbf{Z}_I^T$, and RAFusion measures source-target compatibility in the same space. Consequently, later modules cannot independently redefine which source features are considered transferable. This coupling is central to the method's inductive bias.

A.2 PROGRESSIVE OPTIMIZATION

Training uses three stages so that target augmentation and source reliability are defined only after an invariant representation has been initialized. Algorithm 1 summarizes the procedure.

Algorithm 1 Progressive invariant representation transfer

Require: Source data $\mathcal{D}^S$, target data $\mathcal{D}^T$, source/target graphs, hyperparameters
Ensure: Forecasting parameters Θ

Stage 1: representation separation
Train stable and transient encoders with $\mathcal{L}_{\text{pred}}$, $\mathcal{L}_{\text{cons}}$, $\mathcal{L}_{\text{inv}}$, and the branch-separation term $\mathcal{L}_{\text{dis}}$.
Build baseline-centered invariant prototypes with the EMA teacher.
Stage 2: invariant-conditioned target adaptation
Fit the conditional generator using $\mathbf{Z}_I^T$ and construct curriculum scores d_i.
Expand the target curriculum as validation performance stabilizes.
Stage 3: reliability-aware refinement
Estimate invariant-space mismatch and predictive uncertainties.
Jointly refine source-transfer and target-local predictors with the reliability gate in Eq. (10).

The memory-assisted adaptation component stores representative target input/future pairs and penalizes forgetting during target updates. For target task $\mathcal{D}_{task}^{T}$ and memory $\mathcal{M}$, the update is written as

$$\theta' = \theta - \alpha \nabla_\theta \left[\mathcal{L}(\theta, \mathcal{D}_{task}^{T}) + \lambda_{\text{mem}} \sum_{(\mathbf{X}_m, \mathbf{Y}_m) \in \mathcal{M}} \|\hat{\mathbf{Y}}_m - \mathbf{Y}_m\|_2^2 \right]. \tag{13}$$

Thus the memory penalty is part of the differentiated adaptation objective; it is auxiliary to invariant transfer and is not interpreted causally.

A.3 Disentanglement and Complexity

The stable and environment-specific representations are further separated by the binary branch-role objective

$$\mathcal{L}_{\text{dis}} = -\mathbb{E}[\log D(\mathbf{h}_I)] - \mathbb{E}[\log(1 - D(\mathbf{h}_E))]. \tag{14}$$

This term trains the two representation roles to remain distinguishable; it is not used as evidence that $\mathbf{h}_I$ is source-target invariant. Cross-city invariance pressure is supplied separately by $\mathcal{L}_{\text{inv}}$ in Eq. (4), avoiding the stronger and incorrect interpretation of $\mathcal{L}_{\text{dis}}$ as an adversarial domain-invariance objective. For sequence length L, N nodes, and hidden dimension D_h, the dominant CauKD attention cost is $\mathcal{O}(LN^2 D_h)$. Conditional augmentation costs approximately $\mathcal{O}(BND_h)$ per batch and reliability routing is linear in the number of retained node representations after prototype construction. Discovery and augmentation are training-time costs; the online predictor uses the trained representation encoders and fusion gate.

A.4 Why the Transient Branch Is Retained

The stable branch is not intended to replace the complete city representation. A target city can contain locally predictive factors that are not shared with the source, so removing the residual branch would turn invariance into an information bottleneck. This design choice is also consistent with recent ICLR evidence that hidden-confounder shifts can invalidate strict invariance assumptions and that residual target features can reduce negative transfer (Abbavaram et al., 2026; Xu et al., 2026). The method therefore transfers through $\mathbf{Z}_I$ while allowing $\mathbf{Z}_E$ to remain target-specific. The reliability gate combines these roles at prediction time instead of forcing complete source-target alignment.

A.5 Reliability-Gate Behavior and Edge Cases

The gate in Eq. (10) should be interpreted as a transfer controller rather than a calibrated probability that the source city is "correct." When invariant-space mismatch is small and predictive uncertainty is low, α_j can increase and the transferred-source path contributes more strongly. When mismatch or uncertainty increases, the target-local path receives more weight. The two limits are intentionally asymmetric: $\alpha_j \to 1$ corresponds to strong source reuse, whereas $\alpha_j \to 0$ recovers a predominantly target-local prediction rather than forcing a harmful transfer.

This design also exposes two edge cases. If both source and target representations are poorly estimated, a small measured discrepancy can be falsely reassuring; validation-based calibration is therefore required. Conversely, a genuinely useful source pattern can appear mismatched because of sensor sparsity or short target history, leading the gate to underuse transferable information. The framework addresses neither case by assuming the reliability score is perfect. Instead, the transient branch and target-local head provide a fallback, while the sensitivity and OOD analyses test whether the full system degrades gracefully when the transfer signal becomes harder to estimate.

Reliability-score traceability. R_j is the node-level reliability produced by combining the global discount R_g in Eq. (9) with target-node compatibility to source invariant prototypes; s_j remains a separate regional similarity term in Eq. (10). This separation prevents the same local similarity from being counted twice. Importantly, neither quantity depends on the held-out future target value, so the routing rule is deployable at inference time.

A.6 Loss Roles and Module Coupling

The objective in Eq. (12) combines terms with different purposes. Their roles are separated because a low value of one loss does not imply success of the others. In particular, distribution alignment alone can reduce source-target discrepancy without preserving forecasting utility, while supervised prediction alone can fit a target interval without learning a representation that remains useful under transfer.

Table 5: Functional decomposition of the training objective. The terms are complementary rather than interchangeable certificates of invariance.

Objective	Primary component	Intended role in the pipeline
$\mathcal{L}_{\text{pred}}$	Shared predictor	Preserve target-relevant forecasting information.
$\mathcal{L}_{\text{cons}}$	CauKD	Discourage prediction changes across alternative summaries of the intended transferable representation.
$\mathcal{L}_{\text{inv}}$	CauKD	Reduce source-target discrepancy only in the branch intended for transfer.
$\mathcal{L}_{\text{aug}}$	CauCL	Constrain target augmentation to information routed through the stable representation.
$\mathcal{L}_{\text{dis}}$	Representation split	Encourage stable and environment-specific branches to play different roles.

This decomposition clarifies the ablation logic. Removing $\mathcal{L}_{\text{inv}}$ tests whether selective alignment is useful once the representation has been split. Removing the consistency term tests whether predictive behavior remains stable across alternative summaries of the same branch. Removing CauCL tests whether a learned transferable representation still helps when target evidence is scarce. RAFusion is evaluated separately because it acts after representation learning and controls the amount of source information admitted at prediction time. The central claim therefore concerns the coupled pipeline rather than any single regularizer in isolation.

A.7 Training-Time and Inference-Time Information Flow

The three modules do not impose the same deployment cost. Prototype construction, conditional target generation, curriculum scheduling, and memory-assisted adaptation are training-time operations. At inference, the target sequence is encoded into stable and transient representations, the source-transfer and target-local prediction paths are evaluated, and RAFusion combines them through the learned gate. This separation matters for two reasons. First, it explains why augmentation can improve low-data adaptation without requiring synthetic sample generation for every deployed prediction. Second, it prevents training-only information from being treated as an additional test-time signal when reporting inference latency.

The representation path also preserves a target-local fallback. Even when the transfer gate suppresses source information, the local branch remains available, so failure of the source match does not force extrapolation from an incompatible source city. Conversely, when the learned logit in Eq. (10) strongly favors source reuse, the transferred path receives most of the weight. The gate therefore implements a continuum between local prediction and source-assisted prediction rather than a binary transfer decision.

A.8 Operational Transferability as an Evidence Chain

Definition 1 is intentionally empirical, so each requirement should be connected to a different observable rather than collapsed into one alignment score. Table 6 makes this connection explicit. The separation is important for interpretation: prediction relevance asks whether the representation helps the target task, controlled discrepancy asks whether the intended shared branch is comparatively stable across observed environments, and gated utility asks whether admitting source information is beneficial once mismatch and uncertainty are considered.

This view also clarifies why a single scalar "domain distance" is not used as the paper's final transferability criterion. A source-target representation can be close but uninformative, predictive but unstable, or stable on average while unreliable for a particular target node. The three-stage frame-

Table 6: Evidence chain for the operational transferability definition. No single row is treated as a complete certificate that a representation should transfer.

Requirement	Mechanism	Observable evidence	Main caveat
Prediction relevance	Supervised prediction with stable/transient representation split	Direction-specific MAE/RMSE/MAPE and component ablations	Accuracy alone does not establish invariance
Controlled discrepancy	$\mathcal{L}_{\text{inv}}$ on $\mathbf{Z}_I$ plus consistency regularization	Stable-space alignment objective and workday/non-workday environment gap	Low discrepancy can result from an information bottleneck
Positive gated utility	RAFusion with a target-local fallback	OOD/few-shot behavior and reliability-routing ablation	Gate quality depends on mismatch and uncertainty calibration

work addresses these cases with different mechanisms rather than assuming that one discrepancy measure is sufficient.

A.9 Terminology and Claim Discipline

The identifiers CauKD, CauCL, and RAFusion are retained to preserve correspondence with the implementation and reported experiment configuration. Their functional descriptions were defined at first use in the main paper; subsequent occurrences use the identifiers only. They are treated as representation-learning, target-adaptation, and reliability-fusion modules, not as estimators of causal effects. Likewise, "invariant", "stable", and "environment-specific" denote roles induced by the optimization objectives and evaluated environments. They do not establish that a unique latent causal decomposition has been recovered. The evaluation therefore treats forecasting utility, environment sensitivity, and gated source utility as predictive evidence, while causal identification remains outside the claim.

A.10 Alternative Transfer Designs and Non-Equivalence

Several simpler designs are plausible but encode different assumptions. Aligning the complete representation assumes that target-specific features should also become source-like. Removing the transient branch assumes that all predictive information ought to be shared. Unconditioned target augmentation allows synthetic samples to depend on features that the representation learner has not judged transferable. Unconditional source fusion assumes that source utility is independent of the target mismatch. Table 7 summarizes why these alternatives are not equivalent to the proposed selective-transfer pipeline.

Table 7: Conceptual alternatives to selective invariant transfer. The table describes modeling assumptions rather than additional experimental baselines.

Design	Implicit assumption	Main risk under city shift
Full-representation alignment	All predictive dimensions should be shared	Target-specific structure can be suppressed or source shortcuts can be retained
No transient branch	Invariance should dominate local specialization	The shared representation can become an information bottleneck
Unconditioned augmentation	Any learned target feature is safe for synthesis	Generated samples can amplify transient or city-specific structure
Unconditional source fusion	A useful source remains useful for every target state	Poorly matched source information can cause negative transfer
Selective pipeline	Shared, local, and unreliable information play different roles	Requires accurate representation separation and reliability calibration

B Datasets and Experimental Configuration

All datasets are divided chronologically into training, validation, and test splits with ratio 7:1:2. The standard cross-city protocol uses seven days of target training observations, except in the few-shot study where 2, 3, and 5 days are evaluated. Input and prediction lengths are 12 steps. "Average" metrics aggregate all 12 forecast steps, whereas 15, 30, and 60 minutes are selected checkpoints. Training uses AdamW with learning rate 10^{-3}, batch size 16, hidden dimension 64, and at most 100

Table 8: Traffic datasets used in the cross-city evaluation.

Dataset	Type	Sensors	Period	Interval
PeMS03	Flow	358	09/2018 to 11/2018	5 min
PeMS04	Flow	307	01/2018 to 02/2018	5 min
PeMS08	Flow	170	07/2016 to 08/2016	5 min
METR-LA	Speed	207	03/2012 to 06/2012	5 min
PeMS-BAY	Speed	325	01/2017 to 05/2017	5 min

epochs with early stopping. Z-score statistics are computed from the training partition. Experiments were run on an NVIDIA H100 graphics processing unit.

Table 9: Source-target pairs and their roles in the evaluation.

Task	Source	Target	Role in evaluation
Flow	PeMS03	PeMS08	Primary standard transfer and few-shot analysis
Flow	PeMS03	PeMS04	Additional transfer direction
Flow	PeMS08	PeMS04	Cross-region OOD stress test
Speed	PeMS-BAY	METR-LA	Primary standard transfer and long-horizon analysis
Speed	METR-LA	PeMS-BAY	Reverse-direction transfer

Partition and leakage safeguards. Chronological split boundaries are fixed before normalization or target adaptation. Sliding input-output windows are constructed within each partition, so no window crosses a train/validation or validation/test boundary. Training statistics are used for z-score normalization, validation data for model selection and early stopping, and the held-out test interval only for final reporting. The target-data budget refers only to the target training interval; standard, few-shot, and OOD protocols therefore preserve validation/test separation.

B.1 Protocol Semantics and Comparability

The reported regimes answer different transfer questions and therefore should not be pooled as if they used the same information budget. In the standard setting, seven days of target training data are available for adaptation. The few-shot study reduces this target evidence to two, three, or five days while keeping the evaluation target fixed. The cross-region OOD study changes the source-target pairing and is used as a stronger mismatch stress test. Longer-horizon experiments retain the transfer pair but increase the prediction range. These distinctions are important because a method can be robust to one form of shift while remaining sensitive to another.

The relatively flat few-shot curve between two and five target days is interpreted only over this narrow adaptation range. It indicates early saturation for this particular transfer pair, not that target observations are generally unnecessary or that the same behavior should hold for more severe data scarcity.

For fair comparison, model selection is performed only with the corresponding training and validation partitions. The test interval is not used to tune representation alignment, curriculum thresholds, reliability weights, or early stopping. Source and target normalization uses training statistics from the relevant protocol. The appendix reports complete pair identities so that gains are not averaged across incompatible directions without preserving the direction of transfer.

B.2 Metric Aggregation and Reporting Semantics

The main paper reports mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) using the evaluation definitions used for the reported experiments. For each city-pair direction, the "Average" columns aggregate all 12 forecast steps; the 15-, 30-, and 60-minute values are diagnostic checkpoints and are not averaged to form the summary. Flow and speed tasks are never pooled because their units and error scales differ.

Relative improvements are computed only within the same transfer direction and metric. For a baseline error E_b and the proposed error E_o, the reported relative reduction is $(E_b - E_o)/E_b \times 100\%$. This convention is used for statements such as the average-MAE reduction on PeMS03→PeMS08 and the longer-horizon reductions over CCL. The purpose of the percentage is descriptive comparison within one protocol, not normalization across unrelated datasets.

The manuscript reports the conventional MAPE expression $\mathrm{MAPE} = \frac{1}{\tau N}\sum_{t,n}|(x_{n,t}-\hat{x}_{n,t})/x_{n,t}|$. The reported experiment record does not separately specify a zero-target mask or denominator-clipping constant, so we do not reconstruct one from aggregate values. MAPE is therefore treated as a secondary metric and no conclusion relies on MAPE alone, especially in low-flow regimes.

C Detailed Baseline Descriptions

The evaluation contains 19 methods in three categories. We use canonical method references where a baseline is a standard architecture, and the cross-city paper itself where the baseline is defined by a transfer protocol.

Traditional and single-city baselines. GBRT follows gradient boosting (Friedman, 2001), while VAR provides a classical multivariate linear reference (Lütkepohl, 2005). AGCRN learns node-adaptive graph structure with recurrence (Bai et al., 2020). AllDeepSet uses multiset functions for hypergraph neural networks (Chien et al., 2022). DCRNN models directed diffusion dynamics (Li et al., 2018). DyHSL learns time-varying hypergraph relations (Zhao et al., 2023b). GRU is the recurrent reference (Cho et al., 2014). GWNet combines adaptive adjacency with dilated temporal convolution (Wu et al., 2019). STGCN combines spatial graph convolution with temporal convolution (Yu et al., 2018). STG-NCDE models continuous-time spatio-temporal dynamics (Choi et al., 2022).

Cross-city transfer baselines. CCL learns node and region correspondences (Deng et al., 2026). DASTNet learns domain-shared embeddings adversarially (Tang et al., 2022). DAGN jointly learns cross-city graph structure and transferred representations (Ouyang et al., 2023). D2MHyper combines dynamic multi-scale hypergraphs with domain adversarial learning (Ouyang et al., 2025). MAML is the generic fast-adaptation baseline (Finn et al., 2017). MTPB transfers clustered traffic patterns to data-scarce cities (Liu et al., 2025). ST-DAAN performs attention-based spatio-temporal knowledge transfer (Wang et al., 2022). ST-GFSL transfers node-level meta-knowledge while reconstructing target graphs (Lu et al., 2022). STGCN-FT denotes the STGCN fine-tuning reference used in the evaluation protocol; the associated cross-city knowledge-transfer study is cited in Xu et al. (2025).

Why these baseline families are complementary

The three baseline groups control for different explanations of the observed gains. Traditional methods test whether transfer is needed at all relative to simple temporal structure. Single-city deep models test whether a strong spatio-temporal backbone can compensate for limited target data without an explicit cross-city mechanism. Cross-city methods are the closest comparison because they already use alignment, graph correspondence, meta-learning, or pattern transfer. D2MHyper is particularly relevant because it combines higher-order graph structure with adversarial domain alignment, while CCL directly learns cross-city correspondences. The proposed method is therefore not compared only against non-transfer baselines; it is evaluated against alternatives that already address the source-target mismatch but differ in how they decide what information should be shared.

The comparison remains protocol-bound. Different papers may use different preprocessing, graph construction, or target-data budgets, so the tables use one common evaluation pipeline rather than mixing headline numbers from unrelated settings. This is why the paper reports complete direction-specific values instead of selecting the best number from each publication.

D Complete Horizon-Level Transfer Results

Tables 10 and 11 provide the two primary directions. Tables 12 and 13 report the additional flow and reverse speed directions. Bold and underline denote the best and second-best values within each table.

Table 10: PeMS03→PeMS08 horizon-level results. MAPE is in percent.

Method	15 min			30 min			60 min			Average		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
GBRT	27.13	44.25	16.36	29.32	46.95	17.81	34.15	52.71	21.07	29.69	47.36	18.10
VAR	29.99	44.93	18.41	32.40	48.53	19.81	37.79	56.34	23.84	32.88	49.15	20.32
AGCRN	26.43	46.27	13.62	26.68	46.87	13.56	32.44	52.38	17.25	28.06	48.04	14.53
AllDeepSet	19.85	29.36	13.61	25.40	34.20	37.18	36.87	52.70	25.83	26.75	38.68	19.77
DCRNN	17.08	26.41	12.66	20.06	31.29	14.32	26.64	40.46	19.46	20.56	31.68	14.85
DyHSL	16.82	25.49	12.73	18.98	29.39	14.04	23.40	36.01	17.08	19.19	29.59	14.67
GRU	23.82	32.95	24.55	33.29	42.98	38.53	37.25	51.28	45.40	30.05	40.83	34.05
GWNet	20.98	29.95	20.22	25.54	36.84	22.60	36.25	51.20	29.65	26.59	38.02	23.78
STGCN	18.87	29.09	12.82	23.60	36.58	15.63	33.34	51.36	20.44	24.64	38.03	15.98
STG-NCDE	16.89	25.62	13.37	18.05	28.66	14.27	22.26	35.16	17.07	18.65	29.36	14.66
CCL	15.73	24.44	10.20	17.82	27.63	11.63	21.80	33.01	14.62	18.03	27.98	11.86
DASTNet	17.81	26.60	12.98	20.60	31.12	14.59	27.17	39.97	18.75	21.17	32.00	15.50
DAGN	16.78	24.56	12.59	17.76	26.81	13.98	21.73	33.91	17.23	18.39	27.90	14.51
D2MHyper	15.36	23.45	11.96	16.95	26.32	12.78	21.72	33.32	16.74	17.55	27.00	13.44
MAML	20.23	28.56	19.28	24.25	33.73	25.08	32.68	44.20	36.44	24.91	34.43	26.14
MTPB	21.95	31.64	15.12	24.16	34.74	15.53	27.56	39.70	18.25	24.50	35.42	16.17
ST-DAAN	18.29	27.40	12.30	22.01	32.95	15.14	29.28	43.50	19.81	22.47	33.58	15.18
ST-GFSL	19.74	28.30	15.95	23.36	33.21	19.56	30.31	42.43	27.47	23.78	33.69	20.29
STGCN-FT	18.06	27.20	14.02	20.97	31.34	16.78	26.58	39.20	20.12	21.95	32.67	16.71
Ours	**13.38**	**23.22**	**9.00**	**14.61**	**25.59**	**9.83**	**16.59**	**28.62**	**11.39**	**14.64**	**25.55**	**9.91**

Table 11: PeMS-BAY→METR-LA horizon-level results. MAPE is in percent.

Method	15 min			30 min			60 min			Average		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
GBRT	8.69	16.07	16.60	9.84	18.48	18.54	11.23	20.80	20.95	9.74	18.11	18.33
VAR	8.40	15.27	15.09	9.43	16.96	17.07	11.07	19.23	20.23	9.46	16.86	17.14
AGCRN	5.43	10.88	9.59	6.45	12.90	11.86	7.98	15.59	15.54	6.46	13.00	11.97
AllDeepSet	3.40	6.53	9.09	4.05	8.06	11.24	5.17	10.26	14.82	4.10	8.04	11.36
DCRNN	3.39	6.92	8.97	4.10	8.20	11.61	5.16	9.97	15.57	4.10	8.10	11.67
DyHSL	3.18	6.34	8.35	3.80	7.83	10.69	4.70	9.65	14.05	3.81	7.78	10.70
GRU	3.54	7.14	9.40	4.31	8.70	12.22	5.64	10.49	16.72	4.37	8.53	12.35
GWNet	3.30	6.32	9.83	3.99	7.83	12.31	5.12	9.79	16.63	4.02	7.74	12.47
STGCN	3.34	6.54	9.25	4.02	8.10	11.51	5.09	10.21	14.98	4.06	8.07	11.60
STG-NCDE	3.66	6.90	7.50	4.93	10.20	10.42	6.84	14.09	14.93	4.77	10.05	10.13
CCL	1.67	3.61	4.04	2.49	5.51	6.52	3.49	7.62	10.07	2.41	5.61	6.47
DASTNet	3.67	7.53	9.32	4.59	9.47	12.00	5.94	11.61	16.28	4.60	9.00	12.00
DAGN	2.99	5.59	7.27	3.31	6.27	8.43	3.88	7.44	10.81	3.33	6.30	8.57
D2MHyper	2.33	4.32	6.04	2.61	5.03	7.03	3.58	7.12	10.40	2.75	5.30	7.50
MAML	3.99	7.63	11.85	4.90	9.19	14.83	6.23	10.90	19.18	4.91	9.01	14.85
MTPB	3.09	5.71	7.62	3.73	7.05	9.94	4.63	8.60	13.20	3.76	6.95	10.02
ST-DAAN	3.11	6.01	8.05	3.77	7.51	10.63	4.90	9.50	15.27	3.82	7.48	10.90
ST-GFSL	3.96	7.55	11.60	4.88	9.31	14.30	6.37	11.46	18.56	4.92	9.18	14.40
STGCN-FT	3.35	6.62	8.96	3.91	7.72	11.25	4.84	9.56	15.05	3.95	7.76	11.44
Ours	**1.38**	**2.76**	**3.72**	**2.08**	**4.50**	**5.35**	**3.05**	**6.56**	**8.69**	**2.03**	**4.71**	**5.40**

CROSS-PAIR INTERPRETATION

The four transfer directions expose two distinct sources of difficulty. Flow transfer changes both topology and traffic-demand distribution across PeMS regions, while the two speed directions reverse the source and target roles between PeMS-BAY and METR-LA. The latter is useful because a method that benefits only from one favorable source-target ordering would not be expected to retain the same qualitative advantage after reversal. Across the reported tables, the proposed model remains strongest on average in all four directions, while the magnitude of improvement varies substantially. We therefore interpret the evidence as support for selective transfer rather than a city-independent constant gain.

The horizon-level results also clarify the role of reliability-aware fusion. The strongest baseline changes across metrics and city pairs, which argues against treating any single source alignment mechanism as uniformly sufficient. The proposed framework instead changes the amount of source contribution through the learned reliability gate, so the empirical claim concerns robustness of the transfer rule across heterogeneous pairs rather than universal superiority of a fixed source representation.

Table 12: PeMS03→PeMS04 horizon-level results. MAPE is in percent.

Method	15 min			30 min			60 min			Average		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
GBRT	29.99	44.93	18.41	32.40	48.53	19.81	37.79	56.34	23.84	32.88	49.15	20.32
VAR	30.27	45.17	19.48	33.60	50.12	21.71	41.74	61.20	29.42	34.50	51.19	22.88
AGCRN	22.89	36.77	16.24	23.95	38.31	18.08	28.83	44.26	22.37	24.93	39.40	18.73
AllDeepSet	34.85	42.86	21.77	40.78	55.28	33.85	50.84	68.11	44.94	40.55	55.03	33.74
DCRNN	21.42	33.29	16.83	25.10	38.54	19.36	33.51	50.04	25.63	25.84	39.47	19.96
DyHSL	20.64	32.33	15.93	23.25	36.14	17.60	28.44	43.42	22.28	23.61	36.58	18.27
GRU	24.01	36.50	17.30	39.14	54.41	42.20	51.95	73.34	48.31	37.64	53.75	34.81
GWNet	24.40	36.24	22.65	29.94	43.66	26.80	42.51	60.53	35.64	31.28	45.42	27.79
STGCN	23.12	35.60	17.73	27.09	41.65	19.94	35.81	55.26	24.66	28.06	43.21	20.47
STG-NCDE	21.29	32.99	16.58	23.29	35.76	18.25	28.60	43.92	22.95	24.05	36.97	18.88
CCL	19.41	30.84	13.31	21.39	33.52	14.89	25.19	38.21	18.33	21.62	33.83	15.18
DASTNet	21.51	33.16	16.51	24.45	37.56	18.13	31.31	47.01	23.08	25.11	38.00	18.21
DAGN	20.79	31.21	15.86	22.58	34.58	17.99	28.09	43.55	22.17	23.51	36.01	18.15
D2MHyper	19.91	30.65	14.63	22.24	34.47	16.88	28.51	43.47	20.42	23.06	35.49	17.23
MAML	25.42	37.93	16.48	29.36	43.28	21.84	38.08	54.58	32.33	30.16	44.16	22.50
MTPB	23.53	35.46	19.90	24.47	37.89	20.86	29.94	44.91	24.68	26.06	39.06	21.99
ST-DAAN	22.25	33.96	17.67	26.52	39.86	19.54	35.27	51.82	27.07	27.21	40.79	20.57
ST-GFSL	24.76	35.12	15.93	28.22	41.92	18.97	36.08	52.56	26.28	28.95	42.80	19.98
STGCN-FT	21.98	33.85	16.62	25.17	37.96	18.87	30.96	46.33	22.37	26.08	39.32	19.20
Ours	**18.75**	**30.59**	**12.91**	**20.32**	**32.95**	**14.18**	**22.73**	**36.32**	**16.13**	**20.33**	**32.98**	**14.18**

Table 13: METR-LA→PeMS-BAY horizon-level results. MAPE is in percent.

Method	15 min			30 min			60 min			Average		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
GBRT	2.47	5.06	6.19	2.86	6.02	7.25	3.39	7.22	8.74	2.85	5.97	7.24
VAR	2.48	4.87	6.04	2.92	5.85	7.26	3.61	7.13	9.18	2.94	5.83	7.32
AGCRN	1.60	3.38	3.46	1.93	4.30	4.26	2.57	5.89	5.92	1.98	4.51	4.41
AllDeepSet	1.58	3.45	3.37	2.10	4.77	4.86	2.89	6.62	6.88	2.10	4.70	4.82
DCRNN	1.45	3.18	3.13	1.99	4.57	4.48	2.71	6.20	6.65	1.98	4.46	4.54
DyHSL	1.48	3.13	3.15	1.92	4.35	4.43	2.43	5.96	5.55	1.87	4.17	4.34
GRU	1.60	3.48	3.32	2.16	4.82	5.02	3.02	6.52	7.97	2.18	4.73	5.23
GWNet	1.73	3.55	4.33	2.28	4.88	5.37	3.07	6.53	8.06	2.28	4.80	5.86
STGCN	1.60	3.41	3.60	2.10	4.81	4.96	2.76	6.32	6.70	2.10	4.71	4.95
STG-NCDE	1.99	3.41	4.46	2.65	5.42	6.12	3.17	7.02	7.66	2.42	5.13	5.63
CCL	1.37	2.43	2.83	1.61	3.65	3.47	2.55	5.19	6.02	1.75	3.85	3.89
DASTNet	1.62	3.51	3.40	2.17	4.98	4.80	2.98	6.73	7.04	2.17	5.00	5.00
DAGN	1.35	2.74	2.95	1.61	3.46	3.56	2.09	4.58	4.78	1.65	3.47	3.66
D2MHyper	1.28	2.60	2.92	1.49	3.25	3.50	2.00	4.52	4.93	1.53	3.30	3.62
MAML	1.72	3.46	4.15	2.21	4.75	5.55	2.96	6.35	7.71	2.20	4.69	5.60
MTPB	1.55	2.93	3.20	1.97	3.95	4.37	2.48	5.03	5.86	1.94	3.83	4.33
ST-DAAN	1.51	3.24	3.24	2.06	4.61	4.87	2.82	6.27	7.46	2.05	4.52	4.99
ST-GFSL	1.76	3.49	4.23	2.24	4.77	5.74	2.96	6.32	7.92	2.25	4.69	5.76
STGCN-FT	1.62	3.47	3.71	2.10	4.74	5.06	2.79	6.30	7.07	2.12	4.71	5.15
Ours	**0.99**	**2.27**	**2.59**	**1.28**	**2.94**	**2.72**	**1.90**	**4.41**	**4.49**	**1.26**	**3.12**	**2.82**

Cross-Pair Asymmetry and Metric Consistency

The complete tables show that transfer gains are not symmetric or constant. PeMS03→PeMS04 has the smallest average-MAE reduction among the four directions, whereas the reverse speed transfer remains substantially stronger. This asymmetry is informative because it rules out an interpretation in which the method contributes a fixed regularization benefit independent of the source-target relation. The main-paper comparison therefore reports direction-specific reductions rather than one pooled percentage.

Metric-wise comparisons provide a second check. The strongest reference can differ across MAE, RMSE, and MAPE, especially in the speed tasks, yet the proposed framework remains strongest on the averaged metrics reported in the main paper. We consequently treat average MAE as the compact direction-level summary while retaining RMSE and MAPE here to expose cases in which ranking depends on the error measure. No cross-dataset aggregate is formed because flow and speed datasets have different physical scales and metric distributions.

E LONGER-HORIZON ANALYSIS

The prediction horizon is extended to 90, 120, and 150 minutes for PeMS03→PeMS08 and PeMS-BAY→METR-LA. This setting is more demanding because both local forecasting error and transferred mismatch can accumulate with horizon.

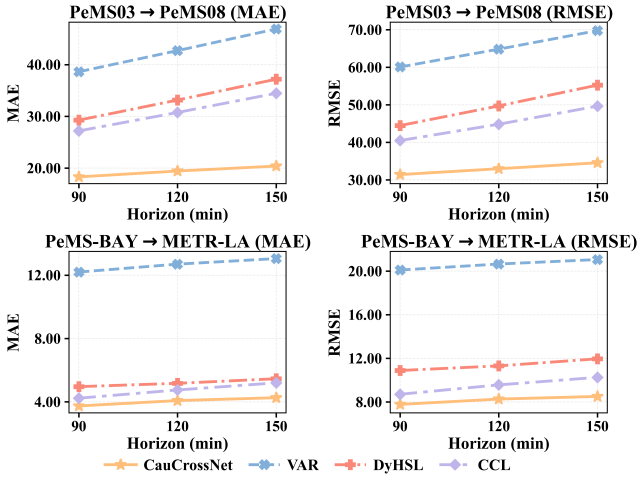


Figure 4: Longer-horizon MAE and RMSE comparisons on the two primary transfer directions.

On PeMS03→PeMS08, MAE increases from 18.31 to 20.40 for the proposed framework, whereas CCL increases from 27.20 to 34.47. The corresponding MAE reductions over CCL are 32.7%, 36.8%, and 40.8% at 90, 120, and 150 minutes. On PeMS-BAY→METR-LA, the reported reductions over CCL also increase with horizon. We interpret this trend as evidence that selective invariant transfer becomes more useful as long-horizon prediction magnifies environment-specific variation, while avoiding a causal claim about the underlying mechanism.

F SENSITIVITY AND EFFICIENCY

Sensitivity analysis should be read together with the role of each parameter. Table 14 separates representation, augmentation, reliability, and optimization controls so that similar numerical ranges are not interpreted as equivalent mechanisms.

Table 14: Functional roles of the main invariant-transfer hyperparameters.

Parameter	Component	Functional role
τ	CauKD	Temperature controlling source-target invariant correspondence
γ	RAFusion	Converts invariant-space shift into a transfer-reliability score
η	CauCL	Controls curriculum expansion during target adaptation
λ_1	CauKD	Weights prediction-consistency regularization
λ_2	CauKD	Weights source-target invariant-space alignment
λ_3	CauCL	Weights invariant-conditioned augmentation
λ_4	Disentanglement	Weights stable/transient branch separation

Table 15: Parameter sensitivity and efficiency under the reported configuration.

Category	Metric	Value
MAE variation	Invariant parameters (τ, γ, η)	$< 4.0\%$
	Other swept settings	$< 2.0\%$
Efficiency	Parameters	0.65M
	GPU-hours (training)	2.04 h
	Inference latency	7.22 ms
Complexity	CauKD	$\mathcal{O}(LN^2D_h)$
	CauCL	$\mathcal{O}(BND_h)$
	RAFusion	$\mathcal{O}(ND_h)$

The reported variation summarizes forecasting sensitivity over the hyperparameters included in the reported sweep rather than stability of any uniquely identified latent mechanism. The sweep does not separately vary λ_1–λ_4, so their sensitivity remains outside the reported evidence. Low sensitivity to τ or γ therefore does not prove invariance under arbitrary city shifts. The 7.22 ms value is retained as an implementation timing measurement; because the manuscript does not separately specify batch-size, warm-up, or device-synchronization details for that measurement, it is not used for cross-method efficiency claims. Target generation and curriculum construction are training-time operations and are inactive during forecasting inference.

G Representation Visualization and Statistical Analysis

Figure 5 visualizes PeMS03→PeMS08 over a 24-hour period. Before invariant representation learning, source and target samples are more separated. After CauKD, samples assigned to the same traffic-pattern group are more compact and better mixed across domains. The visualization supports reduced representation discrepancy, while the forecasting trace shows that this alignment remains prediction-relevant.

Direct-diagnostic boundary. The reported results do not include per-run scalar discrepancy summaries for $\mathbf{Z}_I$ versus $\mathbf{Z}_E$ or node-level RAFusion gate distributions. We therefore do not derive a quantitative stable-versus-transient or gate-calibration table from aggregate forecasting values. Figure 5 is used only as qualitative representation evidence; the quantitative transfer claim rests on the environment-gap, ablation, OOD, few-shot, and horizon analyses reported separately.

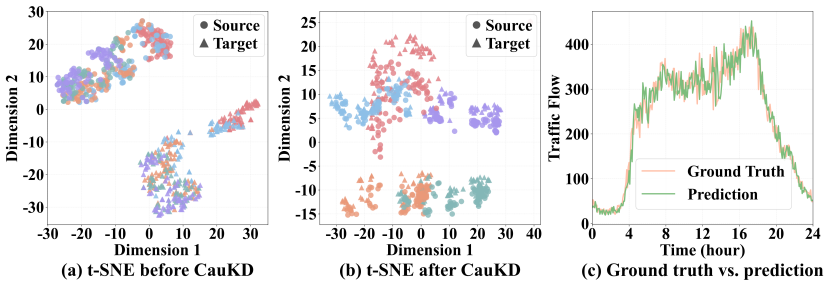


Figure 5: Representation and prediction case study: source-target embeddings before and after CauKD, followed by ground truth and prediction.

Matched within-model controls. The fine-grained ablations disable one internal mechanism under the same PeMS03→PeMS08 backbone and data protocol. Relative to the full model, removing ICR raises MAE by 23.7%, removing DGA raises it by 26.6%, and removing RR raises it by 27.3%. The corresponding RMSE increases are 11.2%, 14.1%, and 14.7%, respectively. These are closer controls for mechanism attribution than cross-paper baselines, although they do not replace a separately trained no-transfer, DANN, or generic-MMD matched baseline.

Table 16: Matched within-model controls from the reported fine-grained ablation.

Variant	MAE	ΔMAE	RMSE	ΔRMSE
Full	14.64	0.0%	25.55	0.0%
w/o ICR	18.11	+23.7%	28.40	+11.2%
w/o DGA	18.54	+26.6%	29.14	+14.1%
w/o RR	18.64	+27.3%	29.30	+14.7%

All-method rank audit. Using the four standard transfer directions as independent blocks and their 12-step averages, a 20-method Friedman audit rejects equal ranks for MAE ($\chi^2 = 61.67$, $p = 2.10 \times 10^{-6}$, $W = 0.811$), RMSE ($\chi^2 = 58.44$, $p = 6.80 \times 10^{-6}$, $W = 0.769$), and MAPE ($\chi^2 = 62.50$, $p = 1.55 \times 10^{-6}$, $W = 0.822$). Horizons are not treated as independent blocks; all three omnibus tests also pass the Bonferroni threshold $0.05/3$.

The all-method audit establishes a global ranking difference while avoiding a selective four-method subset. With only four independent city-pair blocks, however, post-hoc pairwise power remains limited; we therefore make no pairwise superiority claim over CCL or D2MHyper. This is consistent

Table 17: Average ranks over all 20 methods across the four standard transfer directions. Lower is better.

Method	MAE	RMSE	MAPE	Method	MAE	RMSE	MAPE
Ours	1.00	1.00	1.00	STG-NCDE	11.00	11.50	9.00
D2MHyper	2.50	2.25	2.75	STGCN	11.62	13.38	11.25
CCL	2.75	3.50	2.50	AGCRN	12.38	13.75	8.00
DAGN	3.75	3.50	3.75	GWNet	14.25	13.25	17.50
DyHSL	5.50	6.50	6.50	ST-GFSL	14.50	12.62	15.25
MTPB	8.25	7.50	9.25	AllDeepSet	14.50	14.50	13.00
DCRNN	8.75	9.25	9.50	MAML	15.00	13.12	17.00
ST-DAAN	9.50	9.25	10.50	GRU	16.25	16.00	17.25
STGCN-FT	10.00	9.88	11.25	GBRT	18.50	18.75	16.25
DASTNet	10.75	11.50	10.25	VAR	19.25	19.00	18.25

with the earlier four-method Nemenyi result, where the CauCrossNet–CCL rank difference was 1.00, below the reported critical difference of 1.354.

The visualization and statistical analysis play different roles. Representation plots provide qualitative evidence that source and target embeddings become better mixed after invariant learning, whereas rank-based tests summarize forecasting comparisons across evaluated settings. Neither is used as a substitute for the regime-specific robustness analyses that follow. The next section therefore turns from aggregate performance to explicit failure modes and claim boundaries.

H Discussion, Limitations, and Scope

H.1 What Should Transfer, What Should Remain Local

The experiments support a three-way interpretation of cross-city representation reuse. First, some predictive components are sufficiently stable across the evaluated environments to justify alignment and reuse. Second, some components remain target-specific and should stay available through the transient branch rather than being forced into a shared representation. Third, some source information is too unreliable for the current target and should be down-weighted even when it was useful in the source city. This distinction is the operational answer to the title question.

The three modules instantiate these roles in sequence. CauKD proposes a representation intended to carry stable predictive structure. CauCL asks whether that representation remains useful when target evidence is scarce by conditioning augmentation on it. RAFusion then treats transferability as a graded decision rather than a binary property by reducing source contribution when representation mismatch or uncertainty grows. The architecture therefore does not assume that all information surviving domain alignment should be transferred.

This interpretation is deliberately weaker than causal invariance. A low MMD value or a small workday/non-workday gap does not show that latent environmental factors have been removed, and the prototype-consistency objective does not identify a mediator in the causal-inference sense. The evidence establishes predictive robustness over the evaluated environments. It does not establish uniqueness, causal sufficiency, or invariance under arbitrary future urban shifts.

H.2 Limitations of the Current Evidence

The evaluation covers multiple flow and speed pairs, but the number of cities and sensing modalities remains limited. All reported benchmarks are traffic sensor networks with five-minute sampling, so the evidence does not directly establish transfer to heterogeneous modalities such as camera streams, event logs, weather-only sensing, or irregular mobile traces. The four principal transfer directions also provide stronger evidence about pair-dependent robustness than about universal unseen-city generalization.

The stable/transient decomposition is operationally induced by the loss design. It can fail if target-specific factors enter the stable branch or reusable patterns remain transient. The current natural-shift tests diagnose predictive stability but do not manipulate hidden confounders or identify a unique invariant representation. Likewise, target augmentation inherits representation error and can reduce

diversity. These limitations are why lower source-target discrepancy is not treated as sufficient evidence of deconfounding or causal recovery.

Reliability-aware fusion introduces its own calibration requirement. The gate depends on invariant-space similarity and predictive uncertainty, both of which can be poorly estimated when the target interval is very short. The present study does not benchmark these uncertainty estimates as standalone probabilistic predictors against Bayesian uncertainty baselines; they are evaluated only through their routing role. A miscalibrated gate can therefore underuse useful source information or retain source information that should be suppressed.

H.3 Failure Modes and Diagnostic Symptoms

Three primary failure modes follow directly from the framework assumptions. If cities share very little predictive structure, the stable branch can collapse toward a weak common denominator and the target-local branch must carry most of the prediction. If the target training interval is too short to estimate stable-space mismatch reliably, RAFusion can miscalibrate source reliability. If an important target-specific factor is incorrectly absorbed into the stable branch, CauCL can amplify that modeling error through augmentation.

These failure modes imply different observable symptoms and should not be collapsed into one undifferentiated "domain gap." Table 18 summarizes the corresponding diagnostics.

Table 18: Diagnostic view of failure modes. The symptoms distinguish representation, augmentation, and reliability failures rather than attributing every error to a single domain shift.

Failure mode	Observable symptom	Appropriate interpretation or response
Weak shared structure	Small or unstable benefit from source reuse; target-local contribution dominates	Reduce alignment pressure and rely more strongly on the transient or target-local path
Reliability miscalibration	Source weight remains high under large mismatch or high uncertainty, or remains low when validation utility is positive	Recalibrate RAFusion without assuming that the representation itself is invalid
Stable/transient misassignment	Lower representation discrepancy but loss of target-local predictive detail	Revisit the separation and consistency weights rather than increasing alignment strength
Augmentation amplification	Synthetic target samples improve training fit but worsen validation forecasting	Restrict curriculum admission or reduce the contribution of generated target samples
Shortcut persistence	Low average discrepancy but large workday/non-workday or OOD degradation	Treat the learned stable branch as insufficiently robust for that environment

These failure modes also explain why both stable and transient representations are retained. Recent representation-learning results similarly show that strict invariance can fail under hidden-confounder shifts and that residual target features can protect against negative transfer (Abbavaram et al., 2026; Xu et al., 2026). The intended contribution is therefore selective reuse rather than maximal invariance.

H.4 Scope of the Causality-Inspired Interpretation

The causal motivation is used to question indiscriminate correlation transfer, not to assert structural identification. No randomized intervention is observed, no do-effect is estimated, and no learned coordinate is labeled as a causal variable. The role of the mediator-inspired consistency loss is to discourage predictions from depending on one brittle representation route. Its empirical support comes from forecasting performance and natural-shift diagnostics, not from a causal recovery benchmark.

This boundary matters when interpreting negative transfer. If the model down-weights a source city under OOD transfer, the result supports a representation-level reliability decision. It does not identify which latent city factor caused the mismatch. Similarly, improved robustness at long horizons is compatible with the intended stable/transient decomposition, but does not prove that the stable branch corresponds to a true causal mechanism.

The current evidence leaves three specific gaps. First, the stable/transient roles are supported indirectly by matched internal ablations, temporal stability, and visualization rather than by a direct quantitative $\mathbf{Z}_I$ versus $\mathbf{Z}_E$ role test. Second, RAFusion is evaluated through forecasting and routing ablations rather than against standalone Bayesian uncertainty-calibration baselines. Third, the main loss weights λ_1–λ_4 are not swept. Broader city families and controlled environment changes would further separate representation stability, routing quality, and augmentation effects. PLM-CUP is discussed as the closest language-model-based cross-city alternative, but its raster-grid, hourly, 6-to-1 protocol is not numerically mixed with our road-sensor, five-minute, 12-to-12 results.

H.5 Threats to Validity

Construct validity. The paper uses predictive consistency, representation discrepancy, and reliability-gated utility as observable proxies for transferability. These quantities operationalize the intended concept but do not prove that the learned stable branch is the unique invariant representation. In particular, lower MMD can arise from compression or oversmoothing, which is why the paper requires forecasting utility and environment-gap evidence in addition to discrepancy reduction.

Internal validity. CauKD, CauCL, and RAFusion are coupled sequentially, so removing an upstream module changes the input distribution seen by downstream modules. Coarse module ablations therefore estimate the effect of removing a pipeline component, not the isolated causal effect of one scalar loss. Fine-grained ablations are reported separately to reduce this ambiguity, but complete factorization of all interactions would require a larger factorial experiment.

External validity. The evaluated datasets cover traffic flow and speed with five-minute sampling and a limited family of sensor networks. The four standard transfer directions, one stronger OOD pair, temporal regime split, and few-shot budgets provide heterogeneous evidence within this domain, but they do not establish universality across other cities, modalities, sampling rates, or transportation systems.

Statistical validity. The all-method Friedman audit uses all 20 methods over the four standard transfer directions and rejects equal global rankings for MAE, RMSE, and MAPE. Because only four independent city-pair blocks are available, we do not convert this global result into a pairwise significance claim against CCL or D2MHyper. Direction-specific effect sizes and robustness diagnostics remain complementary descriptive evidence rather than substitutes for run-level tests in every setting.

I Transferability Across Evaluation Regimes

The experiments deliberately use several forms of shift because no single benchmark regime fully answers the title question. Table 19 separates what each regime probes from what it cannot establish. This distinction prevents standard transfer accuracy from being used as a substitute for evidence about temporal stability, data scarcity, or negative transfer.

The main conclusion is supported by triangulation rather than by a single scalar domain-distance measure. Standard transfer establishes that source knowledge is useful under the base protocol. The temporal split checks whether the learned representation remains predictive under a within-pair regime change. Few-shot adaptation stresses the target-data budget, cross-region OOD transfer stresses city mismatch, and longer horizons stress cumulative forecasting error. Agreement across these axes supports the narrower claim that selective representation transfer is robust across the evaluated forms of shift. It does not establish that one latent representation is universally invariant.

The same cross-regime comparison also clarifies the role of RAFusion. A favorable standard-transfer result alone cannot show that reliability-aware fusion is necessary, because closely matched cities may tolerate unconditional reuse. Its role is most directly probed when mismatch or uncertainty increases. The OOD, few-shot, and longer-horizon results are therefore interpreted jointly with the RAFusion ablation rather than as independent evidence for a causal mechanism.

Table 19: Interpretive role of the evaluation regimes. The regimes provide complementary evidence and are not treated as interchangeable tests of invariance.

Regime	Primary question	Supported interpretation	Not established
Standard transfer	Does source knowledge improve a target with a fixed adaptation budget?	Overall utility of selective invariant transfer	Causal invariance or worst-case robustness
Workday/non-workday	Does performance remain stable across a natural temporal regime change?	Predictive consistency under temporal shift	Structural causal recovery
Few-shot	Is the transfer bias useful when target evidence is limited?	Data-efficiency of invariant-conditioned adaptation	Zero-shot universality
Cross-region OOD	Does the framework remain useful under a more mismatched target?	Robustness to stronger source-target mismatch	Arbitrary unseen-city guarantees
Long horizon	Does transfer remain useful as forecasting difficulty increases with horizon?	Horizon robustness of selective source reuse	Stability under every future traffic regime

J Evidence Boundaries and Reproducibility

The invariant-transfer interpretation depends on a set of fixed choices that should be reported when the framework is moved to a new city pair.

1. Report the chronological source and target splits, target adaptation budget, normalization statistics, graph construction, and sensor coverage.
2. Report the stable/transient encoder architecture, prototype construction, MMD kernel settings, weights in Eq. (12), and the optimization rule used for $\mathcal{L}_{\mathrm{dis}}$.
3. Report the target augmentation conditioning variables, generator configuration, curriculum schedule, and whether memory-assisted adaptation is enabled.
4. Report how epistemic and aleatoric uncertainty are estimated, the exact construction of R_j, and how the reliability gate is calibrated on validation data.
5. Reserve the held-out test split for final evaluation. Reliability weights and curriculum thresholds should not be chosen from test errors.
6. Interpret cross-environment stability as predictive robustness. Do not infer an identified causal mechanism without additional assumptions or intervention data.

Minimal reporting template. For a new city pair, a reproducible result should identify the source and target periods, sensor counts, graph construction, adaptation budget, normalization statistics, prediction horizons, random-run aggregation, and the exact checkpoint-selection rule. For the transfer mechanism, it should additionally record the dimensionality of $\mathbf{Z}_I$ and $\mathbf{Z}_E$, the MMD kernel and bandwidth, prototype count, branch-separation optimization rule, curriculum thresholds, uncertainty estimator, node-level reliability construction, and the distribution of RAFusion weights on validation data. These items are sufficient to distinguish changes in the data protocol from changes in the transfer mechanism itself.

When reporting ablations, only one component should be disabled at a time unless the purpose is explicitly to test a reduced pipeline. In particular, removing CauKD changes the representation consumed by CauCL and RAFusion, so that ablation should be interpreted as a pipeline-level intervention rather than the isolated removal of one scalar loss. This dependency is why the main paper reports both coarse module ablations and finer internal ablations.

Table 20: Evidence boundaries for the principal claims. The table separates what the current experiments support from stronger interpretations that are not made.

Topic	Evidence reported	Boundary of the claim
Hidden confounding	Workday/non-workday stability, OOD transfer, and mismatch-aware routing	Natural shifts test predictive robustness; they do not manipulate or remove a hidden confounder
Causality-inspired mechanism	Selective stable-branch alignment, consistency regularization, centered prototypes, and reliability gating	MMD, prototype centering, and replay/memory mechanisms are predictive regularizers, not do-operator interventions or identifiable causal estimators
Stable/transient roles	Fine-grained ablations, representation visualization, and environment-gap results	No direct quantitative $\mathbf{Z}_I$ versus $\mathbf{Z}_E$ role-identification test is reported
Reliability/uncertainty	RAFusion ablation and stronger-mismatch evaluations	Uncertainty signals are tested as routing inputs, not as standalone calibrated probabilistic forecasts against Bayesian baselines
Statistical comparison	All-20-method Friedman audit across four independent transfer directions	Omnibus Friedman tests reject equal rankings; pairwise superiority over CCL or D2MHyper is not claimed
Loss sensitivity	Functional roles of λ_1–λ_4 are specified	The four main loss weights are not individually swept, so loss-weight robustness is not claimed
LLM/PLM comparison	PLM-CUP is included in related work as a language-model-based cross-city approach	Its raster-grid/hourly/6-to-1 protocol is not directly comparable to the road-sensor/five-minute/12-to-12 benchmark

Core configuration snapshot. Table 21 collects the settings explicitly reported in the manuscript and experiment description. It provides a compact reproduction cross-check rather than replacing the configuration files.

Table 21: Compact configuration snapshot for the reported experiments.

Item	Reported setting
Input history/standard horizon	12/12 steps
Sampling interval	5 minutes
Chronological split	7:1:2
Standard target adaptation budget	7 days
Few-shot target budgets	2, 3, and 5 days
Optimizer/learning rate	AdamW/10^{-3}
Batch size/hidden dimension	16/64
Maximum epochs	100 with early stopping
Normalization	Training-partition z-score statistics
Compute device	NVIDIA H100 GPU
Reported aggregation	Five-run averages