How Much Structural Abstraction Does Traffic Forecasting Need? A Benchmark for Controllable Multimodal Representations

Anonymous authors
Paper under double-blind review

Abstract

Traffic forecasting combines temporal, visual, and graph representations, yet it remains unclear how much structural abstraction should be provided to a model rather than inferred from raw observations. Benchmarks make this question difficult to isolate because representation type, graph construction, and model architecture are usually varied together. We introduce GB-Traffic, a representation-centric benchmark that aligns temporal signals, multiscale visual encodings, and functional temporal graphs while controlling structural abstraction from coarse to fine. All representations are constructed from historical observations, allowing abstraction level to be studied without changing the prediction target. The benchmark covers six datasets spanning highway, urban-road, bus rapid transit, and metro systems under a unified forecasting protocol and evaluates granularity controllability, functional-edge verification, multimodal complementarity, structural utility across backbones, cross-granularity transfer, and zero-shot cross-domain transfer. Results show that increasing structural resolution is not uniformly beneficial. Intermediate abstraction provides a consistent balance between local detail and structural coherence, while functional temporal graphs provide complementary information beyond physical connectivity across settings. Cross-domain results show that transferability depends on the representation regime. These observations establish structural abstraction as an explicit experimental variable in spatio-temporal forecasting rather than a fixed preprocessing decision. Code, representations, and evaluation resources are available at the anonymous repository: `https://anonymous.4open.science/r/GB-Traffic-5FCB`.

1 Introduction

Traffic forecasting models are commonly evaluated on numerical sensor streams together with one physical adjacency graph. This protocol has accelerated model development, but it also fixes the representation before learning begins. Two sensors that are geographically distant can exhibit synchronized morning peaks, while adjacent locations can serve different functional roles because of road hierarchy, commuting direction, or local demand. A benchmark that exposes only the raw sequence and a single graph therefore makes it difficult to study whether forecasting performance depends on the *level and type of structural abstraction* provided to a model.

This limitation is becoming more consequential as time-series research moves toward representation-centric evaluation. Recent work has examined forecasting components at module level rather than only comparing end-to-end models (Zhao et al., 2026), benchmarked heterogeneous time-series encodings (Wu et al., 2026), and studied how paired textual or multiscale semantic representations interact with temporal signals (Li et al., 2026b; Shang et al., 2026). Traffic forecasting, however, still lacks a benchmark in which the graph representation of the *same historical window* can be changed systematically while the prediction target and preprocessing protocol remain fixed.

We identify three representation gaps. First, sensor-wise numerical sequences leave recurring temporal motifs implicit. Second, a physical graph equates proximity with predictive dependence even when functional similarity is long-range. Third, existing resources expose little control over representation granularity, so a researcher cannot cleanly ask whether a model benefits from broad

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.

structural summaries, fine local detail, or an intermediate level. These gaps concern the benchmark representation itself rather than a particular forecasting architecture.

Granular-ball computing offers a natural route to controllable abstraction by grouping locally homogeneous regions at multiple resolutions (Xia et al., 2019; 2024; Xie et al., 2024b). Building on the GRanular-ball Image Graph (GRIG) formulation (Dai et al., 2025), we develop the Traffic-aware GRanular-ball Image Graph (T-GRIG) for traffic time-series images. T-GRIG emphasizes the temporal axis, aggregates multichannel merge purity, adapts its threshold using training-set sensor variability, and constructs functional edges between regions with similar traffic dynamics. Its benchmark role is a matched representation intervention rather than a new forecaster: the target and forecasting protocol remain fixed while supplied abstraction changes.

GB-Traffic packages these graph views with the raw temporal sequence and a multiscale image. The benchmark spans six datasets from highway, urban-road, bus rapid transit, and metro systems. All modalities are aligned to the same 12-step historical window and 12-step prediction target, so representation changes can be compared without changing the temporal task. Within this unified five-minute protocol, the 60-minute endpoint is the longest evaluated forecasting horizon. This makes the benchmark useful for questions that are difficult to isolate in ordinary traffic datasets: whether functional edges provide information beyond a physical graph, whether granularity yields a predictable accuracy-complexity tradeoff, whether graph benefits persist across forecasting backbones, and whether a representation trained at one resolution transfers to another.

Our contributions are fourfold. First, we introduce a structured multimodal benchmark that aligns raw sequences, multiscale images, and three controllable granular-ball graph resolutions under one data protocol. Second, we develop T-GRIG as a traffic-specific representation pipeline and make its purity, thresholding, and edge-construction rules explicit. Third, we organize evaluation around representation questions rather than a single aggregate score, covering granularity controllability, functional-edge verification, multimodal complementarity, structural-interface comparison across backbones, component attribution, cross-granularity transfer, and zero-shot cross-domain stress tests. Fourth, we release the benchmark construction pipeline and unified loader to support reproducible representation-centric traffic forecasting.

2 Related Work

Traffic forecasting benchmarks and evaluation. Public traffic benchmarks such as PeMS and METR-LA have supported rapid progress in spatio-temporal prediction, but the dominant protocol still exposes numerical streams together with one sensor-level graph. Recent forecasting research spans self-supervised spatio-temporal models, long-range graph Transformers, dynamic graph learning, frequency-aware models, and foundation-model perspectives (Li et al., 2025d; Jin et al., 2025; Zhang et al., 2025b; 2024b; Mao et al., 2025). Broader reviews also emphasize robustness, graph learning, and the growing role of generative and representation-centric methods in transportation forecasting (Liu et al., 2025; Li et al., 2025c; Yan & Li, 2025; Ju et al., 2026; Li et al., 2025b). Recent ICLR work increasingly evaluates representations and modules rather than only complete end-to-end systems: TimeRecipe studies module-level forecasting choices (Zhao et al., 2026), while SciTS evaluates heterogeneous scientific time-series representations (Wu et al., 2026). GB-Traffic complements these directions by making the structural abstraction supplied to a forecaster an explicit benchmark variable.

Multimodal traffic and time-series representations. Traffic prediction has incorporated events, geo-semantic context, images, graph structure, and multiple temporal views (Han et al., 2024; Masuda et al., 2025; Lv et al., 2025; Guo, 2026; Zhao et al., 2023; Zhou et al., 2024; Jiang et al., 2024). Related multimodal mobility work studies temporal multimodal learning, vehicle trajectories, driving-scene representations, and topology-aware scene understanding (Park et al., 2022; Azadani & Boukerche, 2022; Wu et al., 2025; Rong et al., 2024). MM-TSI converts traffic sequences into multiscale time-series images (Lv et al., 2025), whereas BjTT aligns traffic observations with contextual information (Zhang et al., 2024a). Work on multimodal large models and multisource fusion further highlights the importance of aligned heterogeneous representations (Yin et al., 2024; Li et al., 2024a; Yang et al., 2025b; Li et al., 2025a; Wu et al., 2023; Li et al., 2024b). In ICLR 2026, TaTS studies paired text and time series (Li et al., 2026b), while MSH-LLM aligns time-series and

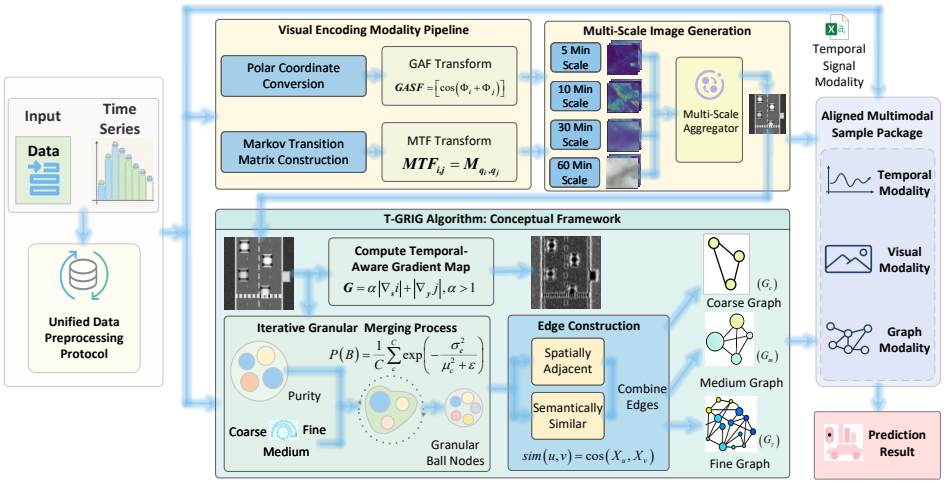


Figure 1: GB-Traffic construction pipeline. Each historical window is aligned across the raw temporal signal, multiscale image, and coarse, medium, and fine functional temporal graphs generated by T-GRIG.

language-model representations through multiscale hypergraphs (Shang et al., 2026). These studies motivate richer inputs, but they do not provide a benchmark in which the graph abstraction of the same traffic window is explicitly varied while the target remains fixed.

Graph structure for traffic forecasting. Graph neural networks (GNNs) remain central to traffic forecasting. Spatio-Temporal Graph Convolutional Network (STGCN) uses a fixed graph convolutional structure (Yu et al., 2018), Graph WaveNet (GWNet) learns an adaptive dependency matrix (Wu et al., 2019), Propagation Delay-Aware Dynamic Long-Range Transformer (PDFormer) models dynamic and long-range dependencies (Jiang et al., 2023), and Graph Attention Network (GAT) provides a general attention mechanism for graph-structured data (Veličković et al., 2018). Additional traffic models explore hierarchical, adaptive, frequency-decomposed, and dynamically evolving graph structures (Guo et al., 2021; Duan et al., 2024; Xu et al., 2026; Zhang et al., 2025b). These methods improve how a model learns from a graph, but the node system is still typically fixed at the sensor level. GB-Traffic instead asks what happens when the representation itself is changed from a physical sensor graph to a controllable hierarchy of functional temporal regions.

Granular-ball and multi-granularity representation. Granular-ball computing constructs adaptive coarse-to-fine abstractions and has been studied for scalable classification, clustering, rough-set learning, and image representation (Xia et al., 2019; 2024; Xie et al., 2024b;a; Su et al., 2025a; Zhang et al., 2023). GRIG extends this idea to image graphs by grouping locally homogeneous regions (Dai et al., 2025), while later work investigates granular-ball density and multi-granularity feature fusion (Su et al., 2025b; Chen et al., 2025). T-GRIG adapts, rather than replaces, the granular-ball primitive: temporal-axis weighting, multichannel purity, training-derived thresholding, and functional edges make abstraction a controlled traffic-specific benchmark intervention.

3 GB-Traffic Benchmark Construction

Figure 1 summarizes the benchmark pipeline. A historical traffic window is preserved as a raw sequence, transformed into a multiscale visual representation, and converted by T-GRIG into coarse, medium, and fine functional temporal graphs. All views share the same historical window and future target.

3.1 Unified Data Protocol

GB-Traffic contains six representative datasets covering four transportation domains. Table 1 lists the benchmark statistics. All series are resampled to 5-minute intervals, samples are split chronologically into training, validation, and test sets with ratio 6:2:2, and normalization or sensor-level

Table 1: Datasets in GB-Traffic under the unified 12-step-to-12-step protocol.

Dataset	Domain	Nodes	Samples	Time span
PeMS04	Highway	307	16,992	Jan. to Feb. 2018
PeMS08	Highway	170	17,856	Jul. to Aug. 2016
METR-LA	Urban road	207	34,272	Mar. to Jun. 2012
XMBRT	Bus rapid transit	44	4,297	2019
HZMetro	Metro	80	5,593	Jan. 2019
BJMetro	Metro	276	2,677	Feb. to Mar. 2018

statistics used during representation construction are computed from the training split only. The standard task observes 12 historical steps and predicts the next 12.

The unified protocol is important for representation comparison. High-frequency observations are aggregated to 5-minute intervals. Lower-frequency series use linear interpolation only within the causal history available at each forecast origin: every interpolated history value must be supported by native observations available no later than that origin. Chronological splitting precedes normalization and adaptive-threshold statistics, preventing target or test-period information from entering representation construction.

3.2 Aligned Representation Modalities

Temporal signal. The temporal modality is the original 12-step numerical sequence and provides the common input representation for conventional forecasting models.

Multiscale image. Following the multiscale imaging idea in MM-TSI (Lv et al., 2025), each historical window is transformed at 5-, 10-, 30-, and 60-minute scales. At each scale we retain the raw sequence, Gramian Angular Field (GAF), and Markov Transition Field (MTF) representation, yielding 12 source channels. These channels are resized to 64×64 and grouped into a three-channel image for convolutional neural networks (CNNs). T-GRIG operates on the original multichannel representation before this three-channel compression.

Functional temporal graph. A T-GRIG node corresponds to a coherent region of the transformed temporal representation rather than a physical sensor. Edges within this per-window graph encode region adjacency or feature-level similarity. For cross-sensor forecasting, region-level features are matched across sensors and converted to the sensor-aligned structural interface consumed by the forecaster. The per-window graph and forecasting interface are therefore distinct objects; Appendix C.1 states their dimensional contract and the adapter information required for reproducibility.

Figure 2 illustrates the alignment. The raw series and image preserve the temporal observation, the physical graph expresses road-network structure, and T-GRIG exposes three progressively finer functional abstractions for the same sample.

3.3 T-GRIG: Controllable Functional Graph Construction

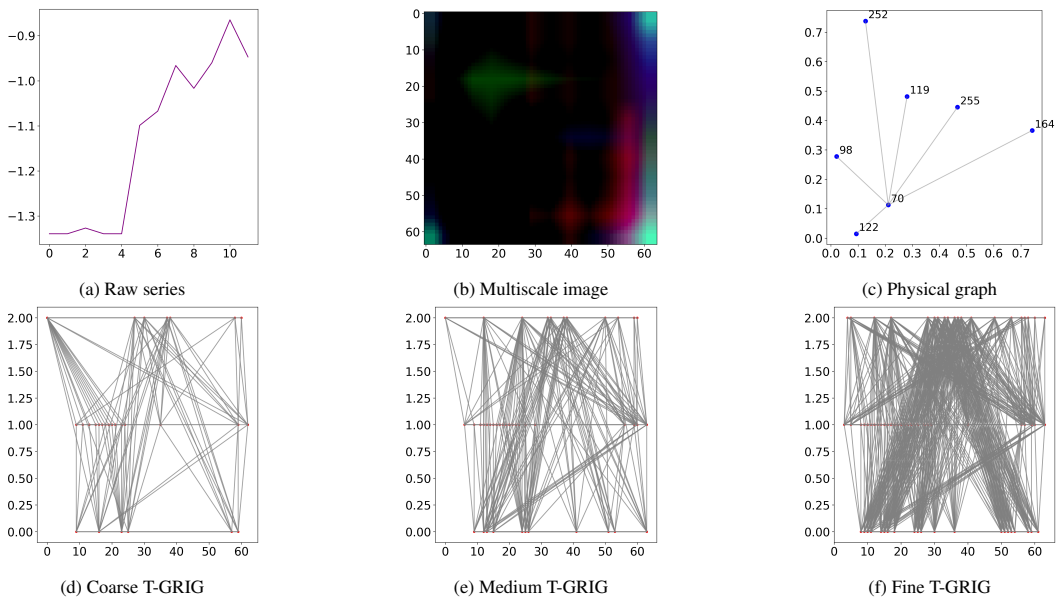
T-GRIG uses four traffic-specific mechanisms. First, a temporally weighted gradient supplies the boundary signal used during region formation. Second, merge purity aggregates coherence across all representation channels. Third, the merge threshold is adjusted by training-set sensor variability, discouraging over-merging at volatile locations. Fourth, functional edges are added when normalized node-feature cosine similarity exceeds $\theta_{\mathrm{sim}} = 0.8$.

Let $\nabla_t I$ denote the image gradient along the designated temporal axis and $\nabla_s I$ the complementary image axis. The weighted gradient is

$$G = \alpha|\nabla_t I| + |\nabla_s I|, \qquad \alpha = 1.5.$$

For a region union $b \cup b'$, joint purity is

$$P(b \cup b') = \frac{1}{C} \sum_{c=1}^{C} \exp\left(-\frac{\sigma_c^2}{\mu_c^2 + \epsilon}\right),$$



(a) Raw series (b) Multiscale image (c) Physical graph

(d) Coarse T-GRIG (e) Medium T-GRIG (f) Fine T-GRIG

Figure 2: Aligned views for one PeMS04 sample under the same historical window and target.

Table 2: Average T-GRIG statistics across the six datasets.

Granularity	Nodes	Degree	Clustering	Density
Coarse	18.30	4.20	0.31	0.12
Medium	37.60	6.80	0.45	0.09
Fine	82.10	11.30	0.62	0.06

where μ_c and σ_c are channel-wise statistics. The adaptive factor is

$$\tau_{\text{adapt}} = 1 + \beta \log(1 + \sigma_s^2),$$

with $\beta = 0.2$ and sensor variance σ_s^2 estimated from training data only. The operational merge condition is

$$T_s = P_{\text{thr}} \tau_{\text{adapt}}, \qquad P(b \cup b') \geq T_s.$$

Because $P \in (0, 1]$, an effective threshold above one simply rejects that candidate union; no additional clipping rule is assumed here. We use $P_{\text{thr}} \in \{0.85, 0.90, 0.95\}$ for coarse, medium, and fine graphs. Unless otherwise specified, the reported construction uses $\alpha = 1.5$, $\beta = 0.2$, and $\theta_{\text{sim}} = 0.8$; parameter selection is confined to validation data and does not use the test split. Each node uses a label-free 19-dimensional schema covering geometry, channel statistics, and traffic-summary descriptors. Appendix A gives the construction contract.

4 Benchmark Evaluation

We organize evaluation around six representation questions. **Q1** asks whether purity thresholds provide predictable control over graph abstraction. **Q2** asks whether T-GRIG edges capture functional temporal relationships not explained by proximity alone. **Q3** asks whether the graph modality adds information beyond time and image inputs. **Q4** asks whether a T-GRIG-derived structural interface improves forecasting across backbones and whether individual T-GRIG components matter. **Q5** asks whether a representation trained at one granularity remains useful when graph resolution changes. **Q6** treats zero-shot cross-domain transfer as a stress test of representation reuse. Mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) are used where applicable; smaller values are better. MAPE excludes zero-valued targets from percentage averaging, while MAE and RMSE retain all valid targets.

Two quantitative protocols are kept separate. The *multimodal configuration protocol* changes the available modalities from time only to time plus physical graph, time plus image, and time plus

image plus T-GRIG. The *structural-interface protocol* holds the forecasting backbone fixed and replaces its graph input. Results from the two protocols are not numerically pooled because the input information differs.

4.1 Q1: Granularity Controllability

The purity threshold produces a controlled accuracy-complexity tradeoff. On PeMS04, Table 3 shows a non-monotonic MAE profile under the same forecasting setup: medium granularity achieves the lowest 60-minute MAE, while coarse and fine graphs remain within a stable error range. We interpret this sweep as evidence that the three graph levels form meaningful abstraction settings, not as a universal optimum for medium granularity.

Table 3: Purity-threshold sensitivity on PeMS04.

P_{thr}	Average nodes	60-min MAE
0.82	12.4	20.52
0.85 (coarse)	18.3	19.88
0.88	28.7	19.62
0.90 (medium)	**37.6**	**19.47**
0.92	48.2	19.55
0.95 (fine)	82.1	19.78
0.97	105.6	20.10

The structural statistics and prediction error move differently with the threshold. Node count rises monotonically from coarse to fine graphs, whereas forecasting error reaches its minimum at the intermediate setting. This distinction matters because controllable granularity is useful only if the benchmark exposes genuinely different representations without making resolution itself a proxy for performance.

4.2 Q2: Functional Temporal Edge Verification

T-GRIG is designed to connect temporally compatible regions even when physical proximity is weak. Across all six datasets, Table 4 reports mean Pearson correlation 0.72 for T-GRIG functional edges, compared with 0.54 for distance-based edges and 0.18 for random pairs. Representative long-range synchronization and complementary lead-lag cases are moved to Appendix D, keeping the main paper focused on quantitative verification.

Table 4: Mean Pearson correlation for T-GRIG-derived functional pairs, distance-based pairs, and random pairs.

Dataset	T-GRIG functional	Distance-based	Random
PeMS04	0.71	0.53	0.17
PeMS08	0.73	0.56	0.19
METR-LA	0.70	0.52	0.18
XMBRT	0.74	0.58	0.20
HZMetro	0.69	0.51	0.16
BJMetro	0.72	0.55	0.17
Overall	**0.72**	0.54	0.18

The edge analysis provides two complementary checks. Functional edges are more temporally coherent on average than distance-based or random connections; among long-distance functional edges above 5 km on PeMS04, 82% align morning peaks within 10 minutes, with mean correlation 0.79. Because feature similarity helps define these edges, correlation is a construction sanity check rather than independent validation; Q3–Q4 test downstream predictive utility.

4.3 Q3: Multimodal Complementarity and Horizon Robustness

The multimodal configuration protocol measures the incremental effect of exposing additional representations to a common forecasting task. Table 5 reports mean absolute error (MAE) at 15, 30,

and 60 minutes for PeMS04 and METR-LA. The 60-minute endpoint is the longest horizon in the unified 12-step benchmark and is used here as the long-horizon stress point. Adding the image improves over time and physical-graph inputs, while adding the medium T-GRIG representation further reduces error throughout the horizon.

Table 5: Multimodal configuration protocol across prediction horizons. Values are MAE.

Dataset	Configuration	15 min	30 min	60 min
PeMS04	Time only	18.52	21.34	25.10
	Time + physical graph	17.20	19.85	23.40
	Time + image	16.00	18.30	21.12
	Time + image + T-GRIG	**15.20**	**17.10**	**19.47**
METR-LA	Time only	4.28	5.12	6.35
	Time + physical graph	3.85	4.55	5.60
	Time + image	3.62	4.28	5.22
	Time + image + T-GRIG	**3.45**	**4.05**	**4.95**

Relative to the image-only configuration, T-GRIG reduces MAE on PeMS04 by 5.0%, 6.6%, and 7.8% at 15, 30, and 60 minutes, respectively; the corresponding reductions on METR-LA are 4.7%, 5.4%, and 5.2%. The gain therefore does not collapse at the longest evaluated horizon. Complete MAE/RMSE/MAPE results remain in Appendix E. These comparisons establish modality complementarity rather than implying that the graph replaces either the image or raw temporal signal.

4.4 Q4: Structural Utility Across Backbones and Components

We next compare structural inputs while holding the forecasting backbone fixed. Table 6 compares each model with its physical graph and the T-GRIG-derived medium structural interface on PeMS04. The T-GRIG-derived pipeline lowers 60-minute MAE by 6.8% to 9.2% across STGCN (Yu et al., 2018), GWNet (Wu et al., 2019), GAT (Veličković et al., 2018), and PDFormer (Jiang et al., 2023). Because the interface includes the region-to-sensor transformation, Q4 evaluates the complete representation pipeline rather than the region partition alone.

Table 6: Structural-interface comparison on PeMS04: 60-minute MAE.

Backbone	Physical graph	T-GRIG-derived	Improvement
STGCN	22.05	20.12	8.7%
GWNet	21.35	19.47	8.8%
GAT	21.90	19.88	9.2%
PDFormer	18.92	17.64	6.8%

Table 7 compares physical, adaptive, single-granularity, and multi-granularity graph sources. Relative to the T-GRIG-derived medium interface, the multi-granularity interface further reduces MAE by 3.2% on PeMS04 and 3.0% on METR-LA.

Table 7: Structural-source comparison at 60 minutes; parentheses denote gains over the physical graph.

Graph source	PeMS04	METR-LA
Physical	21.35	5.30
Adaptive (GWNet)	20.78 (2.7%)	5.10 (3.8%)
T-GRIG-derived medium	19.47 (8.8%)	4.95 (6.6%)
T-GRIG-derived multi-granularity	**18.85** (11.7%)	**4.80** (9.4%)

The component ablation in Table 8 tests whether the gains depend on one isolated design choice. Relative to the full model, proximity-only edges and single-channel purity increase MAE by 3.4% and 2.8%, while removing temporal-axis weighting and adaptive thresholding increases it by 2.1% and 1.5%, respectively.

Table 8: T-GRIG component ablation on PeMS04 using 60-minute MAE.

Variant	MAE
Full T-GRIG	**19.47**
Without temporal-axis weighting ($\alpha = 1$)	19.88
Without adaptive thresholding ($\beta = 0$)	19.76
Single-channel purity	20.02
Proximity-only edges	20.13

The backbone comparison asks whether the T-GRIG-derived pipeline remains useful as forecasting architecture changes; the component ablation asks whether its construction is reducible to one heuristic. The evidence supports pipeline-level utility across these backbones, without isolating the adapter from the region graph or implying universal optimality.

4.5 Q5: Cross-Granularity Generalization

Cross-granularity evaluation asks whether a model trained at one graph resolution remains useful when evaluated at another without retraining for the test resolution. Figure 3 summarizes same-granularity and cross-granularity results over all six datasets. Medium granularity provides the strongest average tradeoff: in the same-granularity setting it reaches average MAE/RMSE/MAPE of 12.84/22.41/21.50%, while its cross-granularity averages are 13.33/23.09/22.51%.

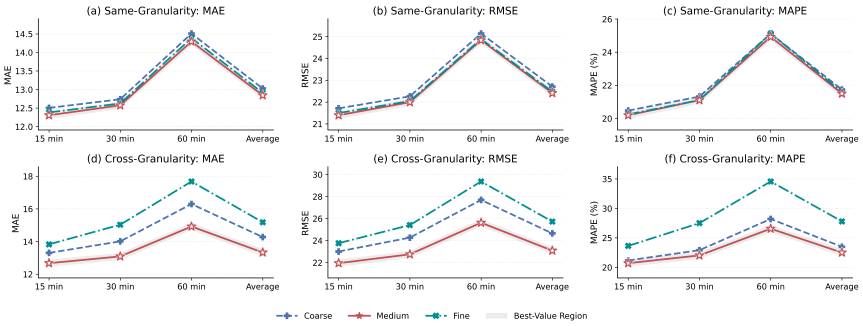


Figure 3: Average same-granularity and cross-granularity prediction results across six datasets.

Relative to coarse granularity, medium reduces the cross-granularity averages by 6.59%, 6.32%, and 4.15% for MAE, RMSE, and MAPE; relative to fine granularity, the reductions are 12.18%, 10.23%, and 19.01%. Fine graphs degrade more strongly under transfer, suggesting sensitivity to resolution-specific local detail. At dataset level, medium's maximum same-granularity excess error is 1.17%, while coarse reaches 4.78% on HZMetro and 3.22% on PeMS08. Appendix F gives the complete excess-error analysis and shows METR-LA as the main exception, where coarse granularity is slightly better.

4.6 Q6: Zero-Shot Cross-Domain Stress Test

The medium-granularity representation is evaluated in four directed zero-shot transfers forming two bidirectional dataset pairs, with no target labels used for adaptation. Figure 4 reports degradation relative to the corresponding target-supervised oracle. Transfer is asymmetric: PeMS08→PeMS04 has the lowest reported ΔRMSE of 17.29%, while HZMetro→XMBRT has the lowest ΔMAPE of 8.34%. Because the strongest direction depends on both metric and source-target family, we treat this experiment as a stress test of representation reuse rather than evidence of universal domain invariance.

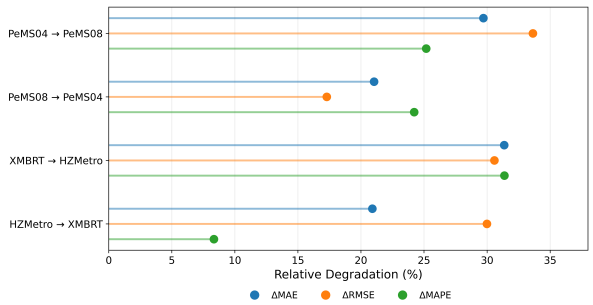


Figure 4: Relative degradation across four directed zero-shot transfers. Lower is better.

Zero-shot interpretation. The target-supervised oracle is used only to express post-evaluation degradation, not for adaptation, and the two directions within each dataset pair are reported separately.

Protocol integrity. Modality addition, structural-interface substitution, cross-granularity transfer, and cross-domain transfer use separate information regimes; Appendix B gives the full control matrix and leakage checks.

Table 9: Cross-axis quantitative synthesis and evidence scope.

Q	Dataset coverage	Headline quantitative evidence	Interpretation
Q1	PeMS04 + six-dataset statistics	Medium: 37.6 nodes and 19.47 MAE; coarse/fine: 18.3/82.1 nodes	non-monotonic accuracy vs. resolution
Q2	six datasets	Mean correlation 0.72 vs. 0.54 distance-based and 0.18 random	functional coherence
Q3	PeMS04, METR-LA	60-min MAE gain over image-only: 7.8% and 5.2%	modality complementarity
Q4	PeMS04 (backbones); PeMS04/METR-LA (sources)	6.8 to 9.2% across backbones; multi-granularity improves 11.7%/9.4% over physical	structural utility
Q5	six datasets	Medium reduces cross-granularity MAE by 6.59% vs. coarse and 12.18% vs. fine	resolution robustness
Q6	2 pairs / 4 directions	Best reported ΔRMSE 17.29%; best reported ΔMAPE 8.34%	exploratory domain reuse

The scope is claim-specific: six-dataset conclusions are used only where all six datasets are evaluated, while supervised multimodal and structural-interface claims remain restricted to the reported configurations. Table 9 is an evidence summary rather than a pooled leaderboard.

5 Conclusion

GB-Traffic reframes traffic forecasting around a controlled representation question: how much structural abstraction should be supplied to a model? The benchmark aligns raw temporal signals, multi-scale visual encodings, and coarse-to-fine functional temporal graphs under the same historical observations and prediction target. The quantitative evidence in Table 9 gives a conditional rather than maximal answer: increasing structural resolution is not uniformly beneficial, and intermediate abstraction provides the most consistent balance between local detail and structural coherence in the reported benchmark. Functional temporal structure remains complementary to physical connectivity across the reported horizon and backbone evaluations, while cross-granularity and zero-shot experiments show that transferability depends on the representation regime. These results support treating structural abstraction as an explicit experimental variable rather than a fixed preprocessing decision, with cross-domain reuse kept as an exploratory claim rather than a guarantee beyond the controlled settings evaluated here. This protocol also enables cleaner comparisons across alternative representation choices.

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

ETHICS STATEMENT

GB-Traffic is constructed from public, anonymized traffic datasets and does not involve human-subject recruitment or personally identifiable information. The benchmark is intended for research on traffic forecasting and representation learning rather than direct safety-critical traffic control. Author-generated benchmark artifacts are released under CC BY 4.0 and code under the MIT license, while the original constituent datasets remain subject to their own access and redistribution terms.

REPRODUCIBILITY STATEMENT

Section 3 specifies the unified data protocol, aligned modalities, and T-GRIG construction, while Section 4 states the evaluation questions and separates the multimodal and structural-interface protocols. The appendix provides the complete T-GRIG algorithm, alignment and storage details, full modality-level prediction tables, cross-granularity diagnostics, zero-shot cross-domain results, representation visualizations, and reproducibility details at the representation, interface, and forecasting levels. The released resources include data-generation scripts, a unified Python loader, experiment configuration files, and environment information; optimizer-level settings are treated as configuration-level records rather than reconstructed from prose.

REFERENCES

Tomoki Arita and Keisuke Okumura. Local guidance for configuration-based multi-agent pathfinding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29296–29304, 2026.

Mozhgan Nasr Azadani and Azzedine Boukerche. A novel multimodal vehicle path prediction method based on temporal convolutional networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):25384–25395, 2022.

Lianhua Chen, Hao Zhai, Yuncan Ouyang, Guochao Zhang, and You Yang. Gbfusion: Adaptive granular-ball computing and multi-granularity feature perception for multi-focus image fusion. *Knowledge-Based Systems*, 330:114789, 2025.

Dongdong Cheng, Xiaocui Jiang, Shuyin Xia, Guoyin Wang, Jinlong Huang, Sulan Zhang, and Yi Wang. Fast spectral clustering via pseudo-label-based granular-ball division for large-scale data. *IEEE Transactions on Knowledge and Data Engineering*, 38(5):2807–2817, 2026.

Dawei Dai, Fan Chen, Shuyin Xia, Long Yang, Guan Wang, Guoyin Wang, and Xinbo Gao. An adaptive multi-granularity graph representation of image via granular-ball computing. *IEEE Transactions on Image Processing*, 34:2986–2999, 2025.

Hongyang Du, Zongxia Li, Dawei Liu, Runhao Li, Haoyuan Song, Qingyu Zhang, Yubo Wang, Jingcheng Ni, Shihang Gui, Congchao Dong, and Tao Hu. A cookbook of 3d vision: Data, learning paradigms, and application. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 8376–8390, 2026.

Wenying Duan, Tianxiang Fang, Hong Rao, and Xiaoxi He. Pre-training identification of graph winning tickets in adaptive spatial-temporal graph neural networks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 701–712, 2024.

Kan Guo, Yongli Hu, Yanfeng Sun, Sean Qian, Junbin Gao, and Baocai Yin. Hierarchical graph convolution network for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 151–159, 2021.

Yunfei Guo. Multimodal spatio-temporal fusion: A generalizable gcn-lstm with attention framework for urban application. *Information Fusion*, 131:104164, 2026.

Haoyu Han, Mengdi Zhang, Min Hou, Fuzheng Zhang, Zhongyuan Wang, Enhong Chen, Hongwei Wang, Jianhui Ma, and Qi Liu. Stgcn: A spatial-temporal aware graph learning method for poi recommendation. In *Proceedings of the IEEE International Conference on Data Mining*, pp. 1052–1057. IEEE, 2020.

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

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pp. 22118–22133, 2020.

Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4365–4373, 2023. doi: 10.1609/aaai.v37i4.25556.

Weiwei Jiang, Yang Zhang, Haoyu Han, Zhaolong Huang, Qiting Li, and Jianbin Mu. Mobile traffic prediction in consumer applications: A multimodal deep learning approach. *IEEE Transactions on Consumer Electronics*, 70(1):3425–3435, 2024.

Di Jin, Cuiying Huo, Jiayi Shi, Dongxiao He, Jianguo Wei, and Philip S Yu. Llgformer: Learnable long-range graph transformer for traffic flow prediction. In *Proceedings of the ACM on Web Conference*, pp. 2860–2871, 2025.

Wei Ju, Siyu Yi, Yifan Wang, Zhiping Xiao, Zhengyang Mao, Hourun Li, Yiyang Gu, Yifang Qin, Nan Yin, Senzhang Wang, et al. A survey of graph neural networks in real world: Imbalance, noise, privacy and ood challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3036–3055, 2026.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024a.

Jipu Li, Yi He, Ye Li, Helai Huang, Dan Wu, and Jieling Jin. Vehicle real-time collision risk prediction: A multi-modal learning approach for diverse urban road scenarios based on a large-scale near-crash event dataset. *Engineering Applications of Artificial Intelligence*, 157:111299, 2025a.

Xinde Li, Fir Dunkin, and Jean Dezert. Multi-source information fusion: Progress and future. *Chinese Journal of Aeronautics*, 37(7):24–58, 2024b. doi: 10.1016/j.cja.2023.12.009.

Xuyan Li, Jie Wang, and Zheng Yan. Can graph neural networks be adequately explained? a survey. *ACM Computing Surveys*, 57(5):1–36, 2025b.

Xuzhao Li, Xuchen Li, Shiyu Hu, Yongzhen Guo, and Wentao Zhang. Verifybench: A systematic benchmark for evaluating reasoning verifiers across domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31796–31804, 2026a.

Ying Li, Fan Bai, Cheng Lyu, Xiaobo Qu, and Yang Liu. A systematic review of generative adversarial networks for traffic state prediction: Overview, taxonomy, and future prospects. *Information Fusion*, 117:102915, 2025c.

Zetao Li, Zheng Hu, Peng Han, Yu Gu, and Shimin Cai. Ssl-stmformer self-supervised learning spatio-temporal entanglement transformer for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12130–12138, 2025d.

Zihao Li, Xiao Lin, Zhining Liu, Jiaru Zou, Ziwei Wu, Lecheng Zheng, Dongqi Fu, Yada Zhu, Hendrik Hamann, Hanghang Tong, and Jingrui He. Language in the flow of time: Time-series-paired texts weaved into a unified temporal narrative. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/29de5722d7d2a6946c08d0c0162a1c71-Abstract-Conference.html.

Lihao Liu, Yanqi Cheng, Zhongying Deng, Shujun Wang, Dongdong Chen, Xiaowei Hu, Pietro Liò, Carola-Bibiane Schönlieb, and Angelica Aviles-Rivero. Trafficmot: A challenging dataset for multi-object tracking in complex traffic scenarios. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1265–1273, 2024.

Shuaiyu Liu, Song Wu, Jie Xu, Yazhou Ren, Yang Yang, Xiaorong Pu, and Guoying Wang. Views attention fusion of granular-ball fuzzy representations split for improved multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23810–23818, 2026a.

Yang Liu, Jiyao Yang, Madhawa Perera, Pan Ji, Dongwoo Kim, Min Xu, Tianyang Wang, Saeed Anwar, Tom Gedeon, Lei Wang, et al. Representation-centric survey of supervised skeletal action recognition and the new benchmark. *Pattern Recognition*, 180:114140, 2026b.

Zewen Liu, Xiaoda Wang, Bohan Wang, Zijie Huang, Carl Yang, and Wei Jin. Graph odes and beyond: A comprehensive survey on integrating differential equations with graph neural networks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6118–6128, 2025.

Yuhuan Lu, Pengpeng Xu, Wei Wang, Zhen Zhang, Han Liu, and Xiping Hu. Knowlcp: Knowledge augmented lane change prediction for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38979–38987, 2026.

Qinzhi Lv, Lijuan Liu, Ruotong Yang, and Yan Wang. Multimodal urban traffic flow prediction based on multi-scale time series imaging. *Pattern Recognition*, 164:111499, 2025.

Yingchi Mao, Hongliang Zhou, Ling Chen, Rongzhi Qi, Zhende Sun, Yi Rong, Xiaoming He, Mingkai Chen, Shahid Mumtaz, Valerio Frascolla, et al. A survey on spatio-temporal prediction: From transformers to foundation models. *ACM Computing Surveys*, 58(4):1–36, 2025.

Shuntaro Masuda, Fumiya Matsuno, Itsuki Hirai, and Toshihiko Yamasaki. Integrating multimodal geo-semantic contexts for traffic prediction. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia Workshops*, pp. 14:1–14:6, 2025. doi: 10.1145/3769748.3773353.

Hyoshin Park, Justice Darko, Niharika Deshpande, Venktesh Pandey, Hui Su, Masahiro Ono, Dedrick Barkley, Larkin Folsom, Derek Posselt, and Steve Chien. Temporal multimodal multivariate learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3722–3732, 2022.

Aleksandr Poslavsky, Alexander D'yakonov, Yuriy Dorn, and Andrey Zimovnov. Vk-lsvd: A large-scale industrial dataset for short-video recommendation. In *Proceedings of the ACM Web Conference*, pp. 8657–8660, 2026.

Fu Rong, Wenjin Peng, Meng Lan, Qian Zhang, and Lefei Zhang. Driving scene understanding with traffic scene-assisted topology graph transformer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10075–10084, 2024.

Zongjiang Shang, Dongliang Cui, Binqing Wu, and Ling Chen. Multi-scale hypergraph meets llms: Aligning large language models for time series analysis. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fa0126bb7ebad258bf4ffdbbac2dd787-Abstract-Conference.html.

Lifeng Shen, Liang Peng, Ruiwen Liu, Shuyin Xia, and Yi Liu. Finding time series anomalies using granular-ball vector data description. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25295–25303, 2026.

Peng Su, Shudong Huang, Weihong Ma, Deng Xiong, and Jiancheng Lv. Multi-view granular-ball contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20637–20645, 2025a.

Xinyu Su, Xiwen Wang, Dezhong Peng, Xiaomin Song, Huiming Zheng, and Zhong Yuan. Identifying outliers via local granular-ball density. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):18956–18967, 2025b.

Weiwei Sun, Shengyu Feng, Shanda Li, and Yiming Yang. Co-bench: Benchmarking language model agents in algorithm search for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33126–33134, 2026.

Zhaomin Tong, Ziyi Zhang, Rui An, Yaolin Liu, Huiting Chen, Jiwei Xu, and Shihang Fu. Detecting anomalous commuting patterns: Mismatch between urban land attractiveness and commuting activities. *Journal of Transport Geography*, 116:103867, 2024.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Sihan Wang, Zhong Yuan, Shitong Cheng, Hongmei Chen, and Dezhong Peng. Granular-ball computing-based random walk for anomaly detection. *Pattern Recognition*, 165:111588, 2025.

Ladyna Wittscher. A survey on design choices for self-supervised learning in computer vision. *Artificial Intelligence Review*, 59(4):112, 2026.

Song Wu, Xiaoyong Li, Wei Dong, Senzhang Wang, Xiaojiang Zhang, and Zichen Xu. Multi-source and heterogeneous marine hydrometeorology spatio-temporal data analysis with machine learning: A survey. *World Wide Web*, 26(3):1115–1156, 2023.

Wen Wu, Ziyang Zhang, Liwei Liu, Xuenan Xu, Jimin Zhuang, Ke Fan, Qitan Lv, Junlin Liu, Chen Zhang, Zheqi Yuan, Siyuan Hou, Tianyi Lin, Kai Chen, Bowen Zhou, and Chao Zhang. Scits: Scientific time series understanding and generation with llms. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ba1d33849b963efc6b5d3082ad68f480-Abstract-Conference.html.

Yanhao Wu, Haoyang Zhang, Tianwei Lin, Lichao Huang, Shujie Luo, Rui Wu, Congpei Qiu, Wei Ke, and Tong Zhang. Generating multimodal driving scenes via next-scene prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6844–6853, 2025.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 1907–1913, 2019. doi: 10.24963/ijcai.2019/264. URL https://www.ijcai.org/proceedings/2019/264.

Shuyin Xia, Yunsheng Liu, Xin Ding, Guoyin Wang, Hong Yu, and Yuoguo Luo. Granular ball computing classifiers for efficient, scalable and robust learning. *Information Sciences*, 483:136–152, 2019.

Shuyin Xia, Xiaochuan Dai, Guoyin Wang, Xinbo Gao, and Elisabeth Giem. An efficient and adaptive granular-ball generation method in classification problem. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5319–5331, 2024.

Shuyin Xia, Xinjun Ma, Zhiyuan Liu, Cheng Liu, Sen Zhao, and Guoyin Wang. Graph coarsening via supervised granular-ball for scalable graph neural network training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12872–12880, 2025.

Haolong Xiang, Peisi Wang, Xiaolong Xu, Kun Yi, Xuyun Zhang, Quan Z Sheng, Amin Beheshti, and Wei Fan. Mtp: Exploring multimodal urban traffic profiling with modality augmentation and spectrum fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 27037–27045, 2026.

Jiang Xie, Chunfeng Hua, Shuyin Xia, Yuxin Cheng, Guoyin Wang, and Xinbo Gao. W-gbc: An adaptive weighted clustering method based on granular-ball structure. In *Proceedings of the IEEE 40th International Conference on Data Engineering*, pp. 914–925. IEEE, 2024a.

Qin Xie, Qinghua Zhang, Shuyin Xia, Fan Zhao, Chengying Wu, Guoyin Wang, and Weiping Ding. Gbg++: A fast and stable granular ball generation method for classification. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(2):2022–2036, 2024b.

Siyi Xie, Hanxin Zhu, Xinyi Chen, Tianyu He, Xin Li, and Zhibo Chen. Sonic4d: Spatial audio generation for immersive 4d scene exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11087–11095, 2026.

Jinpeng Xu, Jing Yang, Yaqun Huang, Lip Yee Por, Xiaotong Chen, and Chunna Zhao. Dfgnet: A dual-pathway graph neural network via frequency decomposition for spatiotemporal forecasting. *Expert Systems with Applications*, 297:129518, 2026.

Huan Yan and Yong Li. Generative ai for intelligent transportation systems: Road transportation perspective. *ACM Computing Surveys*, 57(12):1–45, 2025.

Jie Yang, Zhuangzhuang Liu, Guoyin Wang, Qinghua Zhang, Shuyin Xia, Di Wu, and Yanmin Liu. Constructing three-way decision with fuzzy granular-ball rough sets based on uncertainty invariance. *IEEE Transactions on Fuzzy Systems*, 33(6):1781–1792, 2025a.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025b.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

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

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

Qianru Zhang, Xinyi Gao, Haixin Wang, Siu Ming Yiu, and Hongzhi Yin. Efficient traffic prediction through spatio-temporal distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1093–1101, 2025a.

Qinghua Zhang, Chengying Wu, Shuyin Xia, Fan Zhao, Man Gao, Yunlong Cheng, and Guoyin Wang. Incremental learning based on granular ball rough sets for classification in dynamic mixed-type decision system. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):9319–9332, 2023.

Xiaomei Zhang, Ziqin Jiang, and Ping Lou. Triple dynamic graph convolutional recurrent network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 26(8):12647–12660, 2025b.

Xingyu Zhang, Siyu Zhao, Zeen Song, Huijie Guo, Jianqi Zhang, Changwen Zheng, and Wenwen Qiang. Not all frequencies are created equal: Towards a dynamic fusion of frequencies in time-series forecasting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 4729–4737, 2024b.

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, 2023.

Zhiyuan Zhao, Juntong Ni, Shangqing Xu, Haoxin Liu, Wei Jin, and B. Aditya Prakash. Timerecipe: A time-series forecasting recipe via benchmarking module level effectiveness. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/f46ddea413df86832418c5e04e59644f-Abstract-Conference.html.

Bodong Zhou, Jiahui Liu, Songyi Cui, and Yaping Zhao. A large-scale spatio-temporal multimodal fusion framework for traffic prediction. *Big Data Mining and Analytics*, 7(3):621–636, 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 complete T-GRIG construction, while Appendices B and C define the controlled evaluation protocols, structural interface, alignment, and storage contract. Appendices D–H provide complementary qualitative and quantitative evidence for representation alignment, complete forecasting performance, cross-granularity generalization, zero-shot cross-domain transfer, and representation interpretability. Appendices I and J discuss extended benchmark use cases and representation failure modes. Appendices K–M document the artifact manifest, broader benchmark and granular-computing context, and the protocol-by-dataset evidence map. Appendices N–P close the supplement with the experimental reporting record, reproducibility and statistical-reporting boundaries, and the benchmark release scope.

A COMPLETE T-GRIG CONSTRUCTION

T-GRIG follows three principles. **Temporal dominance** emphasizes changes along the temporal axis. **Multi-channel consistency** aggregates coherence evidence across the original multiscale channels. **Functional connectivity** allows edges to reflect temporal compatibility in addition to local adjacency. These principles make the graph a representation transformation rather than a task-specific change to the forecasting target.

Let $I \in \mathbb{R}^{H\times W\times C}$ denote the original multichannel traffic representation. While the three-channel image is used for CNN compatibility, T-GRIG retains the full channel set during graph construction. The weighted gradient is

$$G = \alpha|\nabla_t I| + |\nabla_s I|, \qquad \alpha = 1.5,$$

where ∇_t is the designated temporal axis and ∇_s the complementary image axis. The segmentation stage uses the resulting map to prioritize lower-contrast shared boundaries before stronger temporal boundaries; a recorded merge-order seed resolves ties. Using axis names rather than fixed image coordinates avoids tying the definition to a rendering convention.

For a candidate union $b \cup b'$, purity is

$$P(b \cup b') = \frac{1}{C}\sum_{c=1}^{C} \exp\left(-\frac{\sigma_c^2}{\mu_c^2 + \epsilon}\right).$$

The adaptive factor

$$\tau_{\text{adapt}} = 1 + \beta \log(1 + \sigma_s^2)$$

uses historical sensor variance estimated from the training split. The operational threshold is $T_s = P_{\text{thr}}\tau_{\text{adapt}}$, and a merge is accepted if $P(b\cup b') \geq T_s$; because purity is bounded by one, a threshold above one rejects the candidate union rather than invoking an unstated clipping step. After merging, each ball is represented by a 19-dimensional, label-free vector spanning geometry, image statistics, and traffic-summary descriptors. Edges are added for region adjacency or normalized-feature cosine similarity above $\theta_{\text{sim}} = 0.8$.

Complexity. Gradient computation and pixel initialization require $\mathcal{O}(HW)$ time. With maintained spatial adjacency, the merging stage performs at most $\mathcal{O}(HW)$ successful merges plus local neighborhood checks. The functional-edge step is the potential bottleneck: naive all-pairs similarity costs $\mathcal{O}(K^2 d)$ for K final balls and d-dimensional features. Candidate pruning or indexed similarity search reduces practical cost. Space is dominated by the multichannel image and the final graph.

Determinism and parameter policy. The graph is determined once the historical window, training-derived sensor statistics, purity threshold, gradient-priority merge order, tie-breaking seed, and edge rule are fixed. The reported construction uses $\alpha = 1.5$, $\beta = 0.2$, and $\theta_{\text{sim}} = 0.8$ unless otherwise specified; P_{thr} is varied as the granularity intervention. Reproducibility depends on the merge-order seed, feature normalization before cosine similarity, and candidate-pruning setting because these choices can affect ties and runtime.

Algorithm 1 T-GRIG: Traffic-aware Granular-Ball Image Graph Construction

Require: Multichannel traffic representation $I \in \mathbb{R}^{H\times W\times C}$, purity threshold P_{thr}, adaptive factor β, similarity threshold θ_{sim}, fixed merge-order seed
Ensure: Granular-ball graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$
Compute $G = \alpha|\nabla_t I| + |\nabla_s I|$, with $\alpha = 1.5$.
Compute $\tau_{\mathrm{adapt}} = 1 + \beta \log(1 + \sigma_s^2)$ and $T_s = P_{\mathrm{thr}}\tau_{\mathrm{adapt}}$.
Initialize each pixel as one ball in $\mathcal{B}$; order candidate pairs by increasing mean boundary value in G (fixed-seed tie breaking).
repeat
  Set `merged` $\leftarrow$ false.
  for each ball $b \in \mathcal{B}$ in the current gradient-priority order **do**
    for each spatial neighbor $b' \in \mathcal{N}(b)$ **do**
      Compute $P(b \cup b') = \frac{1}{C}\sum_c \exp[-\sigma_c^2/(\mu_c^2 + \epsilon)]$.
      if $P(b \cup b') \geq T_s$ **then**
        Merge b and b', update $\mathcal{B}$, set `merged` $\leftarrow$ true, and break.
      end if
    end for
  end for
until `merged` is false
Compute the fixed 19-dimensional feature X_v for each final ball v.
Normalize X_v and add (u, v) if the balls share a boundary or $\cos(X_u, X_v) > \theta_{\mathrm{sim}}$.
return $\mathcal{G} = (\mathcal{V}, \mathcal{E}, X)$.

B Controlled Evaluation Protocols

GB-Traffic uses multiple evaluation protocols because each isolates a different representation question. They are intentionally not collapsed into a single leaderboard.

Table 10: Controlled factors in the principal evaluation protocols.

Protocol	Changed factor	Held fixed
Granularity sweep	P_{thr} and graph resolution	Dataset, historical window, target, backbone, split
Functional-edge diagnostic	Edge-selection rule / pair type	Historical observations and diagnostic statistic
Multimodal configuration	Available modality set	Dataset, target horizon, source training protocol
Structural interface	Physical/adaptive/T-GRIG-derived source	Backbone, target, horizon, training protocol
Cross-granularity	Train/test graph resolution	Dataset, sample identity, target definition
Zero-shot transfer	Source-target domain pair	Medium representation rule and no target-label adaptation

Within-protocol interpretation. The multimodal configuration and structural-interface experiments answer different questions. The former changes the information made available to the model, whereas the latter changes the structural source while holding the forecasting backbone fixed. Similarly, cross-granularity transfer changes representation resolution within a dataset, whereas cross-domain transfer changes the traffic domain. Reporting them separately prevents a gain caused by an additional modality from being attributed solely to graph topology.

Leakage boundary. Chronological splitting is performed before normalization, sensor-variance estimation, or any representation statistics are fitted. Every modality for a sample is generated only from the same historical input window, and the future target is attached after representation construction. This ordering is part of the benchmark definition because otherwise a data-dependent abstraction rule could inadvertently use target-period information.

Long-horizon interpretation. The unified benchmark predicts 12 future five-minute steps. We therefore use the 60-minute endpoint as the longest forecasting horizon supported by the unified protocol rather than introducing a new target definition. The main paper reports horizon-resolved MAE; Appendix E retains RMSE and MAPE for the same 15-, 30-, and 60-minute checkpoints.

C Protocol, Alignment, and Storage Details

C.1 Per-Window Graph and Cross-Sensor Interface

GB-Traffic distinguishes two related graph objects. The first is the per-window granular-ball graph generated from one sensor's transformed historical sequence; its nodes are coherent image regions and do not correspond one-to-one with physical traffic sensors. The second is the sensor-aligned structural interface used in cross-sensor forecasting comparisons, for which region-level features are matched across sensors before input to a forecasting backbone.

This distinction is necessary for dimensional consistency: a per-window graph with tens of granular-ball nodes cannot be substituted node-for-node for a sensor graph containing hundreds of traffic units. Cross-sensor alignment is obtained by region-feature matching followed by a region-to-sensor adapter. Structural-interface comparisons are defined by the adapter, its normalization, and its configuration; these quantities are held fixed when physical, adaptive, and T-GRIG-derived sources are compared.

The structural-interface results therefore apply to the complete specified T-GRIG-derived pipeline rather than to arbitrary region graphs inserted unchanged into sensor-level GNNs. They should be interpreted as pipeline-level evidence unless the region-to-sensor adapter is separately ablated.

C.2 Structural-Interface Definition

The benchmark distinguishes the per-window region graph from the sensor-aligned structural interface consumed by forecasting backbones. At granularity g, T-GRIG produces region features $\mathbf{X}^{(g)} \in \mathbb{R}^{K_g \times 19}$ and region edges $\mathbf{E}^{(g)}$, whereas a sensor-level forecaster operates on N traffic units. We denote the archived region-to-sensor transformation abstractly as

$$\Phi_g\Big(\mathbf{X}^{(g)}, \mathbf{E}^{(g)}; \omega\Big) = \mathbf{S}^{(g)}, \qquad \mathbf{S}^{(g)} \in \mathbb{R}^{N \times d_s},$$

where ω denotes the adapter configuration. The equation specifies the dimensional contract without equating region nodes with sensors. Reproducing a structural-interface result therefore requires the corresponding adapter implementation, normalization, configuration, and output dimensionality.

Table 11: Dimensional contract for the T-GRIG-derived forecasting interface.

Stage	Object	Reproducibility requirement
T-GRIG construction	$K_g \times 19$ region features + region edges	report granularity, normalization, merge order, and edge rule
Region-to-sensor adapter	Φ_g : $\{K_g$-region graph$\} \rightarrow N$-aligned interface	expose matching/projection rule, configuration, and whether it is deterministic or learned
Forecasting backbone	sensor-level model input	keep backbone and training protocol fixed in structural-interface comparisons
Evaluation	common future target	use identical sample identity, horizon, and metric aggregation

The adapter implementation and configuration define Φ_g operationally for each experiment record. Structural-interface results refer to the joint region-graph-plus-adapter representation and should not be interpreted as direct node-for-node substitution of region graphs into sensor-level GNNs or as an adapter-free estimate of region-graph utility.

C.3 Temporal Alignment and Leakage Control

All raw sequences are represented at 5-minute intervals. High-frequency observations are aggregated, while lower-frequency series are linearly interpolated subject to the causal eligibility rule above: an interpolated history point is used only when its supporting native observations are available by the forecast origin. Each sample uses a 12-step input and 12-step target, with chronological train/validation/test splits and training-only normalization and variance statistics.

Every modality for a sample is derived exclusively from the historical input window. The raw sequence, image channels, and T-GRIG nodes therefore share the same information boundary. The future 12-step target is attached only after representation construction. This separation is necessary because an adaptive representation benchmark would otherwise risk leaking target-period statistics into the graph itself.

C.4 Multiscale Visual Representation

At four temporal scales, 5, 10, 30, and 60 minutes, the pipeline retains the raw sequence, a Gramian Angular Field (GAF), and a Markov Transition Field (MTF) representation. The resulting 12 channels are resized to a common 64×64 spatial support. For the visual modality, channels are grouped into three aggregates compatible with standard CNNs; T-GRIG instead consumes the original multichannel representation before compression.

The distinction between the visual and graph pipelines is important. The image modality intentionally compresses the channel set to a conventional visual interface, whereas T-GRIG keeps channel-wise information for purity estimation. Consequently, an improvement from the graph modality cannot be reduced to merely feeding the same three-channel image through another encoder.

C.5 Sample Storage and Loader

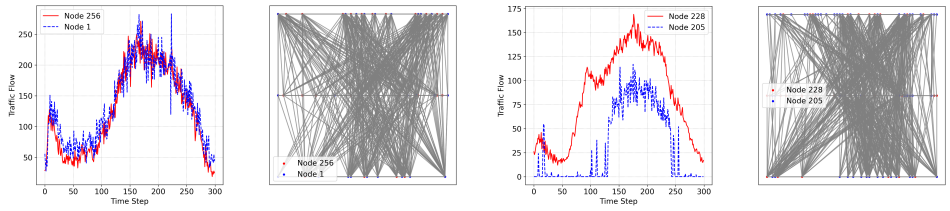
Each sample is stored under a dataset/sensor/timestamp hierarchy and contains the 12-step raw sequence, the three-channel image, coarse/medium/fine region-graph tensors, and the future label tensor. For granularity g, the raw T-GRIG representation stores region features $\mathbf{X}^{(g)} \in \mathbb{R}^{K_g \times 19}$ and an edge index $\mathbf{E}^{(g)} \in \mathbb{N}^{2 \times E_g}$, where K_g is the number of granular-ball regions at that resolution. This notation is distinct from the N physical traffic units used by sensor-level forecasting backbones. When a forecasting experiment uses the T-GRIG-derived structural interface, the released artifact must additionally store or regenerate the region-to-sensor adapter output defined in Appendix C.2. A unified `GBTrafficDataset` loader supports modality selection, batching, and cross-dataset evaluation without changing the downstream data interface.

The benchmark loader returns a common sample identifier across modalities, making it possible to verify that a time-series window, image, region graph, sensor-aligned structural interface, and future target originate from the same timestamp. This alignment check guards against silent sample reordering across otherwise valid modalities.

D Representation Alignment and Qualitative Examples

Figure 2 in the main paper already shows the aligned PeMS04 temporal, visual, physical-graph, and coarse-to-fine T-GRIG views for one historical window and target. This appendix therefore focuses on qualitative functional-edge cases that are not shown in the main text.

D.1 Qualitative Functional-Edge Cases



(a) Long-range synchronized pattern (b) Adjacent complementary pattern

Figure 5: Representative functional temporal edges on PeMS04 at medium granularity.

Functional-edge diagnostic. The two examples represent different notions of functional relation. A synchronized pair has similar temporal shape despite spatial separation, whereas a complementary pair can exhibit a stable lead-lag relation. The mean-correlation table in the main paper therefore summarizes one form of coherence, not an exhaustive definition of functional temporal utility. The additional PeMS04 diagnostic shows that 82% of long-distance connected pairs above 5 km align their morning peaks within 10 minutes and have mean correlation 0.79.

E Complete Forecasting Results

E.1 Multimodal Configuration Protocol

Table 12 reports the complete 15-, 30-, and 60-minute results for the multimodal configuration protocol. Because this protocol changes the available modalities, its values should not be numerically conflated with structural-interface comparison.

Table 12: Multimodal configuration protocol under the 12-step-to-12-step setting. Each cell is MAE/RMSE/MAPE(%).

Configuration	15 min	30 min	60 min	Average
PeMS04				
Time only	18.52/28.33/12.45	21.34/33.56/14.20	25.10/38.90/16.80	21.65/33.60/14.48
Time + physical graph	17.20/26.10/11.80	19.85/30.20/13.10	23.40/35.50/15.20	20.15/30.60/13.37
Time + image	16.00/24.50/10.90	18.30/28.10/12.00	21.12/32.00/13.50	18.47/28.20/12.13
Time + image + T-GRIG	**15.20/23.00/10.20**	**17.10/26.00/11.00**	**19.47/29.50/12.10**	**17.26/26.17/11.10**
METR-LA				
Time only	4.28/8.15/10.22	5.12/9.88/12.45	6.35/11.90/15.10	5.25/9.98/12.59
Time + physical graph	3.85/7.20/9.15	4.55/8.70/10.80	5.60/10.45/13.25	4.67/8.78/11.07
Time + image	3.62/6.88/8.55	4.28/8.10/10.00	5.22/9.70/12.40	4.37/8.23/10.32
Time + image + T-GRIG	**3.45/6.50/8.10**	**4.05/7.65/9.40**	**4.95/9.20/11.70**	**4.15/7.78/9.73**

The improvement is consistent across the three reported horizons. On both datasets, the ranking of the four configurations remains unchanged, and the difference between the image-only and T-GRIG-enhanced configurations is maintained at the 60-minute horizon. The table is not used to claim that one fusion architecture is universally optimal; its role is to verify graph-modality complementarity under this controlled protocol.

E.2 Graph-Source and Component Diagnostics

The structural-interface results in the main paper isolate the source of structural information while keeping the backbone fixed. Component ablations isolate the construction rules within T-GRIG. These tests answer different questions and should not be merged into a single effect size.

The complete graph-source values are already reported in Table 7, while the component variants are reported separately in Table 8. Keeping those two protocols separate avoids duplicating their values in the supplement and preserves the distinction between changing the supplied structural source and ablating T-GRIG construction rules.

Complete supervised multimodal T-GRIG results under this configuration protocol are reported for PeMS04 and METR-LA. We therefore restrict the supervised multimodal claim to these two datasets. XMBRT, HZMetro, and BJMetro remain part of the benchmark and enter the six-dataset structural and transfer diagnostics reported elsewhere.

E.3 Horizon-Specific Complementarity

The horizon profile can be summarized by the relative MAE reduction obtained when T-GRIG is added to the time-plus-image configuration. On PeMS04, the reduction grows from 5.0% at 15 minutes to 6.6% at 30 minutes and 7.8% at 60 minutes. On METR-LA, the corresponding reductions are 4.7%, 5.4%, and 5.2%. The two datasets therefore do not exhibit the same monotonic trend, but both retain a positive contribution at the 60-minute endpoint. This is the basis for the manuscript's long-horizon claim: the structural representation remains useful at the longest horizon defined by the common 12-step target, not at arbitrary future horizons.

F Cross-Granularity Diagnostics

F.1 Overall Performance

The same-granularity protocol trains and tests at the same graph resolution, whereas the cross-granularity protocol transfers a trained model to the other two resolutions. Medium granularity has average MAE/RMSE/MAPE of 12.84/22.41/21.50% in the same-granularity setting and 13.33/23.09/22.51% under cross-granularity transfer.

Relative to coarse granularity, the cross-granularity averages improve by 6.59%, 6.32%, and 4.15%; relative to fine granularity, the improvements are 12.18%, 10.23%, and 19.01%. This pattern suggests that an intermediate partition can preserve enough local structure for prediction without binding the model too tightly to one high-resolution graph.

F.2 Dataset-Level Stability

Figure 6 reports relative excess error, defined as the performance gap between a granularity and the best granularity under the same dataset and evaluation condition. Medium granularity is not the best on every dataset, but its excess error remains small across the six datasets. The maximum same-granularity excess error is 1.17% for medium, whereas coarse reaches 4.78% on HZMetro and 3.22% on PeMS08.

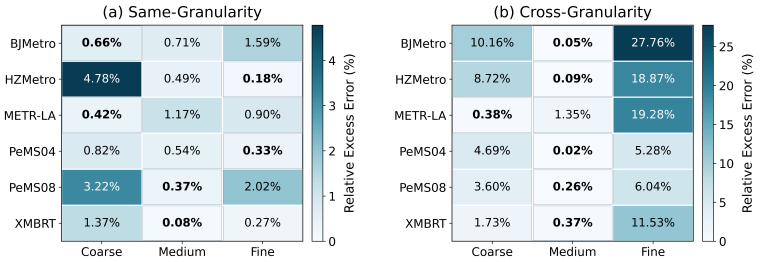


Figure 6: Dataset-level relative excess error across graph granularities. Lower is better.

Under cross-granularity evaluation, medium is the best choice on five of the six datasets; METR-LA is the exception, where coarse granularity is slightly better. The result motivates treating medium as a robust operating point rather than a universal optimum.

G Zero-Shot Cross-Domain Transfer

The zero-shot protocol evaluates the medium-granularity graph representation in four directed cross-domain transfers forming two bidirectional dataset pairs, without target-label adaptation. Figure 4 in the main paper reports degradation from the corresponding target-supervised oracle. Lower values indicate stronger transfer.

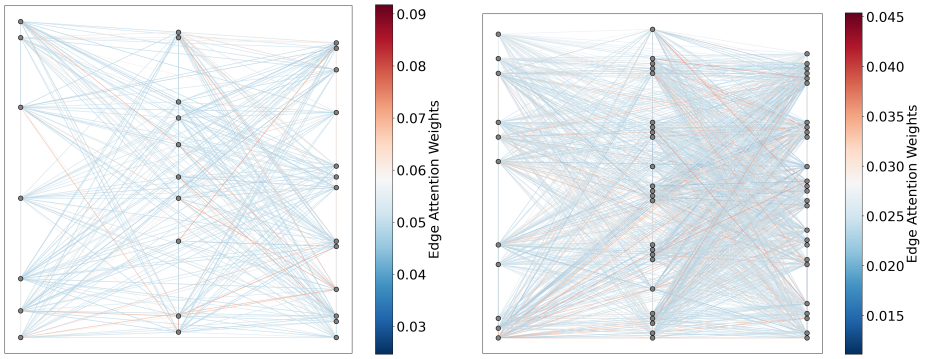
Transfer is asymmetric: PeMS08→PeMS04 gives the lowest reported ΔRMSE of 17.29%, while HZMetro→XMBRT gives the lowest ΔMAPE of 8.34%. Differences in network scale, transportation mode, and temporal distribution can all contribute to this asymmetry. The experiment therefore supports feasibility of representation reuse across some domain pairs, not a domain-invariance guarantee.

Interpretation boundary. The target-supervised oracle is used only to express relative degradation after evaluation. No target labels are used to adapt the zero-shot representation. Because only two bidirectional dataset pairs, comprising four directed transfers, are tested, the analysis should be read as a stress test rather than a comprehensive domain-generalization benchmark.

H Representation Interpretability and Benchmark Use Cases

H.1 Granularity-Dependent Attention

Figure 7 projects attention weights from the same forecasting model onto coarse and fine T-GRIG graphs. Coarse graphs spread attention across broad traffic phases, whereas fine graphs emphasize shorter high-variation intervals. The observation motivates adaptive granularity selection but is not used as a quantitative performance claim.



(a) Coarse granularity (b) Fine granularity

Figure 7: Granularity-dependent attention for the same representative sample.

The coarse view emphasizes broad phases that span substantial parts of the historical window, while the fine view concentrates on shorter changes. This qualitative behavior is consistent with the intended abstraction hierarchy. It should not be interpreted as proof that one resolution is more interpretable in general, because attention depends on the forecasting model and selected sample.

I Extended Benchmark Use Cases

The aligned temporal, visual, and graph views support several prospective benchmark extensions. These directions are not empirical claims of the present paper.

Vision-language grounding. Because image and graph regions originate from the same historical window, future work can test whether textual descriptions remain consistent across coarse, medium, and fine abstractions. Such evaluations should preserve graph-to-image correspondence and measure grounding consistency directly.

Parameter-efficient adaptation. The granularity family provides a controlled representation shift for adapters, prompts, or low-rank updates. A future protocol can train at one resolution, adapt to another under a fixed parameter budget, and report forecasting error together with trainable parameter count.

Test-time adaptation. The six datasets also permit source-target experiments with unlabeled target samples under a fixed representation-construction rule. Zero-shot transfer, unsupervised adaptation, and target-supervised fine-tuning should remain separate information regimes.

J Representation Failure Modes and Diagnostics

Controllable abstraction introduces failure modes that are easy to hide if only final forecasting error is reported. Table 13 summarizes the principal risks implied by the construction pipeline together with diagnostics that expose each failure mode.

Coarsening versus fragmentation. A lower purity threshold reduces node count and can suppress short-lived local structure, whereas a very high threshold creates many small regions that can bind a

Table 13: Representation failure modes and corresponding diagnostics.

Failure mode	Observable symptom	Diagnostic
Over-coarsening	Distinct temporal phases merge into a small number of regions	Node-count trend, purity sweep, and horizon-wise error
Over-fragmentation	Fine graph becomes resolution-specific and transfer degrades	Cross-granularity excess error and graph density
Spurious functional edges	High-similarity edges do not correspond to stable temporal relations	Correlation, peak alignment, and lead-lag case inspection
Training-statistic leakage	Graph threshold reflects validation or test behavior	Recompute using training-only variance and verify split boundaries
Modality conflation	Gain from extra information is reported as graph-only gain	Separate multimodal configuration from structural-interface comparison
Domain overclaim	A few successful pairs are treated as universal transferability	Report every source-target direction and oracle-relative degradation

model to a particular resolution. The U-shaped PeMS04 sensitivity curve and the cross-granularity results provide evidence for both risks. They motivate reporting structural statistics together with error rather than selecting a granularity from forecasting accuracy alone.

Semantic-edge validation. Cosine similarity above θ_{sim} is an operational edge rule, not a proof of functional equivalence. The benchmark therefore combines mean connected-pair correlation with long-distance synchronization and lead-lag examples. Future extensions could add lagged correlation, mutual information, or task-conditional edge relevance, but such alternatives should be compared under the same historical windows and train-only preprocessing.

Protocol leakage and comparability. Adaptive thresholding depends on historical sensor variance. If that variance were estimated from the test period, the graph itself would carry information from the evaluation distribution. The released protocol avoids this by fitting normalization and sensor statistics on the training split before constructing validation or test representations. The same principle applies to any future learned graph-construction module.

K Artifact Manifest and Data Card

A GB-Traffic sample is defined by one dataset, sensor or local unit, and timestamp. All modalities derived from that identifier correspond to the same historical window and future target. Table 14 records the released artifacts and their benchmark role.

Table 14: Per-sample artifact manifest.

Artifact	Stored form	Role
Raw temporal signal	12-step numerical sequence	Highest-fidelity history and conventional forecasting input
Multiscale visual view	64×64 three-channel image	CNN-compatible summary of multiscale transformed channels
Coarse T-GRIG	Graph tensor with node features and edge index	Broad functional temporal abstraction
Medium T-GRIG	Graph tensor with node features and edge index	Intermediate abstraction used by the primary comparisons
Fine T-GRIG	Graph tensor with node features and edge index	High-resolution functional temporal abstraction
Future label	12-step target tensor	Shared prediction target, attached after representation construction
Sample identifier	Dataset / unit / timestamp key	Cross-modality alignment and provenance

For granularity g, the raw T-GRIG graph tensor contains a $K_g \times 19$ region-feature matrix and a $2 \times E_g$ edge-index representation. The feature schema is fixed across datasets and contains geometry descriptors (including region size and centroid), channel-level statistical summaries, and traffic summaries including peak and periodicity descriptors; no target values or future observations enter these features. The artifact specification records the exact feature ordering and normalization

together with the region-to-sensor adapter, linking raw region graphs and downstream structural interfaces to the same source window.

Release scope. The benchmark data originate from public, anonymized traffic datasets. GB-Traffic adds derived representations and aligned sample packaging rather than new personally identifiable information. Redistribution remains subject to the access and licensing conditions of the constituent datasets; the benchmark transformation version and raw-to-derived sample mapping provide provenance for the derived artifacts.

L Broader Benchmark and Granular-Computing Context

The representation-centric design is consistent with benchmarks that make a controlled factor explicit rather than reporting only a final task score. Graph-learning benchmarks standardize datasets and protocols (Hu et al., 2020), while recent benchmark efforts isolate algorithm search, verifier reliability, industrial recommendation, immersive multimodal generation, representation-centric action recognition, 3D vision design choices, and self-supervised learning configurations (Sun et al., 2026; Li et al., 2026a; Poslavsky et al., 2026; Xie et al., 2026; Liu et al., 2026b; Du et al., 2026; Wittscher, 2026). Adjacent transportation work further illustrates heterogeneous structural representations in tracking, lane-change prediction, multimodal traffic profiling, spatio-temporal distillation, commuting analysis, path finding, and recommendation (Liu et al., 2024; Lu et al., 2026; Xiang et al., 2026; Zhang et al., 2025a; Tong et al., 2024; Arita & Okumura, 2026; Han et al., 2020).

Granular-ball methodology has also expanded to time-series anomaly detection, multi-view fuzzy representation, scalable spectral clustering, supervised graph coarsening, random-walk anomaly detection, and uncertainty-aware three-way decision models (Shen et al., 2026; Liu et al., 2026a; Cheng et al., 2026; Xia et al., 2025; Wang et al., 2025; Yang et al., 2025a). GB-Traffic differs from these directions by treating granularity itself as a benchmark-controlled forecasting variable rather than only an internal optimization device.

M Protocol-by-Dataset Evidence Map

The benchmark contains six datasets, but not every experimental question requires the same supervised training protocol. Table 15 makes the evidence population explicit so that dataset breadth is not confused with universal coverage of every experiment.

Table 15: Dataset coverage by evaluation protocol.

Protocol	Dataset population	Interpretation boundary
Graph statistics	all six datasets	describes controllable abstraction, not predictive superiority
Functional-edge correlation	all six datasets	tests temporal coherence of selected pairs
Multimodal forecasting	PeMS04 and METR-LA	supports modality complementarity on the core supervised protocol
Structural-interface comparison	reported core configurations	tests structural source under fixed backbones
Cross-granularity transfer	all six datasets	supports robustness to representation-resolution change
Zero-shot cross-domain	2 bidirectional pairs / 4 directions	exploratory cross-domain reuse only

This separation also defines how benchmark extensions enter the evidence map. Additional supervised results on XMBRT, HZMetro, or BJMetro belong to the same multimodal protocol rather than to the six-dataset structural diagnostics; additional domain pairs extend the zero-shot matrix while retaining the target-label restriction.

Evidence density. For each claim, the benchmark prefers a direct diagnostic over a larger but less controlled collection of scores. Granularity controllability is checked with node-count and error sweeps; functional connectivity is checked with pair-level temporal statistics; modality complementarity is checked with horizon-resolved forecasting; structural utility is checked across backbones and graph sources; transfer robustness is checked under resolution and domain shifts.

N EXPERIMENTAL RECORD FOR GB-TRAFFIC STUDIES

A GB-Traffic experiment is characterized by the dataset version, source and target periods, historical and prediction horizons, modality subset, graph granularity, T-GRIG construction parameters, forecasting backbone, training seed, checkpoint rule, and the representation axis being changed (modality availability, structural source, granularity, or domain). A final MAE without this context does not identify which abstraction variable produced the observed difference.

Table 16: Minimum reporting record for a GB-Traffic experiment.

Field	Recorded information
Data protocol	Dataset version, timestamps, 5-minute resampling rule, chronological split boundaries
Temporal task	Input length, target length, reported horizons, metric aggregation
Representation	Raw/image/graph modalities, coarse/medium/fine or multi-granularity setting
T-GRIG construction	P_{thr}, α, β, θ_{sim}, merge ordering, feature normalization
Structural interface	Physical/adaptive/T-GRIG-derived source and exact region-to-sensor adapter when applicable
Optimization	Backbone, optimizer, seed, run count, dispersion statistic, early stopping, checkpoint-selection rule
Transfer protocol	Same-granularity, cross-granularity, or zero-shot domain direction and target-label status
Artifact trace	Code commit, processed-representation version, configuration file, output checksum when released

O REPRODUCIBILITY AND STATISTICAL REPORTING

The paper specifies the representation-construction parameters used by T-GRIG, including P_{thr}, α, β, θ_{sim}, the 12-step input/output protocol, merge ordering, and training-only estimation of adaptive statistics. Optimizer, learning rate, batch size, epoch budget, early stopping, and checkpoint-selection settings belong to the experiment configurations associated with each reported result rather than to the representation definition itself.

Statistical reporting boundary. The reported backbone, graph-source, and component-ablation tables contain point estimates and do not tabulate repeated-seed dispersion statistics. We therefore make no uncertainty claim beyond the reported runs. When repeated stochastic runs are available, the run count and dispersion statistics provide the corresponding uncertainty information.

Reproducibility has three distinct levels. *Representation reproducibility* concerns whether the same historical window regenerates identical image channels, graph nodes, and graph edges under a fixed merge order. *Interface reproducibility* concerns whether the same T-GRIG region graph yields the same sensor-aligned structural input. *Forecasting reproducibility* additionally fixes the backbone configuration, optimization settings, seed handling, checkpoint selection, and metric aggregation.

Table 17: Three levels of reproducibility in GB-Traffic.

Level	Fixed quantities	Consistency check
Representation	input window, transforms, thresholds, merge order	image hash, node/edge counts
Interface	region-to-sensor adapter and normalization	aligned structural tensor shape and checksum
Forecasting	backbone, optimizer, seed, checkpoint rule	configuration file, checkpoint id, metric log

P RELEASE SCOPE AND REPRODUCIBILITY

The benchmark uses public, anonymized traffic data. The current scope is limited to six transportation datasets, 5-minute resampling, a fixed 12-step history and prediction horizon, and graph representations derived from the chosen multiscale image construction. T-GRIG uses one family of purity thresholds and a cosine-similarity edge rule; other construction operators may produce different abstraction behavior. Cross-domain analysis covers a limited set of source-target pairs and should not be interpreted as universal transferability.

A reproducible release or reimplementation is specified by:

1. the original dataset version, resampling rule, chronological split boundaries, and training-only normalization statistics;

2. the four temporal scales, image-channel construction, resize operation, and channel grouping used for the visual modality;
3. P_{thr}, α, β, θ_{sim}, merge-order seed, node-feature normalization, and candidate-pruning rule for T-GRIG;
4. the exact modality set used by each forecasting experiment so that modality-fusion and structural-interface results are not conflated;
5. the exact cross-sensor matching or projection adapter used to convert per-window T-GRIG region features into the structural interface consumed by sensor-level forecasting backbones;
6. the forecasting backbone, training seeds, number of independent runs, dispersion statistic when available, checkpoint-selection rule, prediction horizons, and metric aggregation;
7. the source-target direction and target-supervision status for every cross-domain experiment.

Claim-to-evidence map. Granularity controllability is supported by monotonic graph statistics and the purity-threshold sweep. Functional connectivity is supported by connected-pair correlation and long-distance peak-alignment diagnostics. Predictive utility is supported separately by multimodal complementarity, structural-interface comparison across four backbones, graph-source comparison, and component ablation. Cross-resolution robustness is supported by same- and cross-granularity evaluation. Cross-domain reuse is exploratory and limited to the reported zero-shot pairs.

Artifact Consistency

Several low-cost consistency checks detect silent representation mismatches before forecasting begins (Table 18). They are useful because GB-Traffic aligns several derived views to one sample identifier, and a tensor can be syntactically valid while still corresponding to the wrong historical window, granularity, or sensor interface.

Table 18: Artifact-level consistency checks for released benchmark instances.

Check	Expected invariant	Failure indication
Sample alignment	raw, image, graph, and label share dataset/unit/timestamp key	cross-modality sample mismatch
Temporal boundary	every derived view uses only the 12-step historical window	target leakage into representation construction
Granularity ordering	node count increases from coarse to medium to fine for the same sample family	inconsistent threshold or merge configuration
Interface dimension	sensor-aligned structural object matches the forecaster's N traffic units	region graph used without the required adapter
Configuration trace	reported run resolves to one immutable preprocessing and model configuration	result cannot be reproduced from the released artifact

Release validation. These invariants can be verified automatically on a deterministic sample set and stored with the benchmark version as identifiers, tensor shapes, node/edge counts, adapter version, and configuration hashes. They do not replace forecasting replication, but they separate data-generation failures from model-training variation.

Artifact Provenance

Result provenance in a representation-centric benchmark depends on the exact artifacts used to produce each measurement. For GB-Traffic, changes in preprocessing, graph granularity, the region-to-sensor adapter, or the forecasting configuration can alter the information supplied to the model even when the dataset name is unchanged. A machine-readable run manifest provides this binding for each reported result.

Result-level traceability. The run manifest associates each reported table row or plotted point with one immutable configuration record rather than only with a model name. This is particularly important for Tables 5, 6, 7, and 8, Figure 3, and Figure 4, where different representation or transfer protocols intentionally expose different information to the forecasting model. The corresponding configuration file records implementation settings that are not repeated in prose.

Table 19: Provenance fields linking reported results to released artifacts.

Artifact layer	Minimum identifier	Purpose
Dataset snapshot	source version, split boundaries, checksum	fixes the raw observation population
Representation generator	code commit, preprocessing config	fixes resampling, imaging, and graph construction
T-GRIG graph	granularity parameters, merge-order seed, graph checksum	identifies the exact region representation
Structural adapter	implementation version, adapter config, output shape	fixes the K_g-region to N-unit interface
Forecasting run	backbone config, seed, checkpoint id	binds optimization and model state to the result
Metric record	horizon, aggregation rule, output checksum	links table and figure values to the run record

Versioning rule. Versioning separates representation changes from model-only changes: modifications to dataset preprocessing, T-GRIG construction, the region-to-sensor adapter, or metric aggregation define a new benchmark artifact version, whereas model-only changes can reuse an existing representation version when the input tensors and interface are unchanged.

Future work should evaluate dynamic graph updates, adaptive granularity selection, missing-modality settings, broader geographic coverage, and additional representations not derived from the current image pipeline.