CAUSALGAP: BENCHMARKING REPRESENTATION RESPONSES IN TRAFFIC PREDICTION UNDER STRUCTURAL INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Traffic prediction is a cornerstone of intelligent transportation, yet model reliability under sudden structural changes remains poorly understood. Existing evaluations quantify predictive degradation under structural interventions, but provide limited evidence on whether model representations register those changes or remain rigid when performance deteriorates. We introduce CausalGap, a diagnostic framework centered on the Representation Response Landscape, a two-dimensional state space defined by Dual-Metric Intervention Sensitivity (DMIS) and Relative Degradation (RD), which measure representation responsiveness and normalized post-intervention error increase, respectively. A constructive error-only impossibility example motivates joint analysis of response and degradation within a four-regime landscape, including pathological rigidity, characterized by high degradation and low responsiveness at a declared representation interface. To reduce structural-access confounding, CausalGap introduces a Dual-Track Structural Access Interface separating native and controlled access. The primary falsifiable test asks whether weak-intervention response provides incremental information about stronger unseen degradation beyond prespecified weak-error trajectory summaries and controls, complemented by an information-matched post-hoc steering test of explicit DMIS parameterization. The benchmark protocol supports topology-input stress testing and intervention-consistent response-failure analysis across heterogeneous traffic predictors and multiple observational, controlled, and simulation-based settings, while treating real-world disruption evidence separately when representation response is unavailable. Matched-error comparisons, invariance checks, and alternative response measures assess measurement stability and diagnostic consistency across model families and settings. CausalGap provides a diagnostic benchmark for representation-response analysis under structural interventions. Resources and evaluation instructions are available at `https://anonymous.4open.science/r/CausalGap-80CB`.

1 INTRODUCTION

Traffic prediction models are typically trained under fixed road-network topologies (Yu et al., 2018; Li et al., 2018; Wu et al., 2019; Bai et al., 2020). Real-world disruptions, such as accidents, closures, and emergency controls, can abruptly violate this assumption, causing prediction errors that propagate through the network. As deployment scales, the gap between static model assumptions and dynamic network conditions widens, exposing fundamental vulnerabilities in real-world systems.

This widening gap manifests as three critical challenges, as illustrated in Figure 1. First, existing evaluations primarily report accuracy drops, leaving representation changes unexamined and making it difficult to determine whether a model has registered the altered topology. Second, conventional error metrics cannot separate beneficial invariance from harmful rigidity, conflating robust insensitivity with a frozen state disconnected from the environment. Third, heterogeneous architectures receive unequal structural information, complicating cross-model comparisons and creating diagnostic blind spots. These challenges motivate a falsifiable question: after accounting for weak-intervention error trajectories, does representation response predict failure under a stronger unseen intervention? Error alone cannot answer it because models with nearly identical current errors can

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.

exhibit sharply different representation responses and subsequently diverge. This formulation does not require a strong global association between response and contemporaneous degradation.

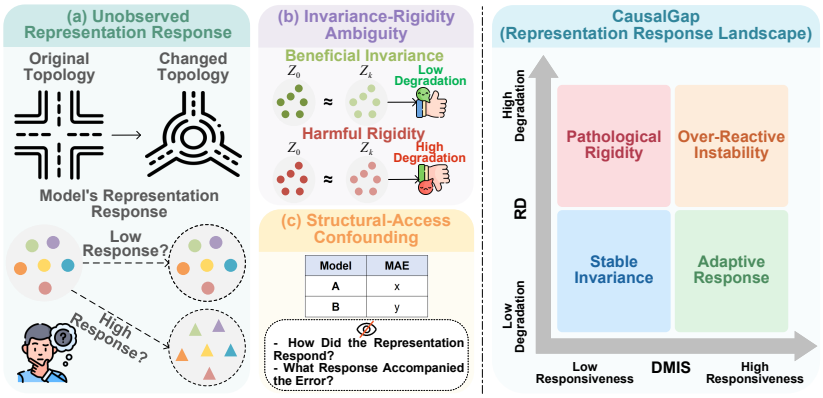

Figure 1: **Motivation for CausalGap.** Error-only evaluation leaves representation responses unobserved, cannot distinguish beneficial invariance from harmful rigidity, and confounds representation-response comparisons across architectures with unequal access to structural information.

The three evaluation gaps in Figure 1 motivate the joint response-degradation view used throughout the paper. The name CausalGap refers to the gap between declared structural interventions and observed representation responses, rather than identification of latent causal mechanisms. CausalGap makes three contributions centered on the Representation Response Landscape. (1) It jointly measures representation responsiveness through DMIS and relative degradation through RD. A constructive impossibility example shows that prediction error alone does not determine selected-interface response, while four operational regimes distinguish stable invariance, adaptive response, over-reactive instability, and pathological rigidity. (2) A Controlled Structural-Access Principle, implemented through a Dual-Track Structural Access Interface (DSAI), exposes native-access differences and reduces structural-information confounding in controlled comparisons. (3) A held-out-model-family weak-to-strong test evaluates response value beyond prespecified weak-error trajectory summaries and controls; information-matched post-hoc steering is a secondary test of explicit DMIS parameterization rather than deployable adaptation.

2 Related Work

We organize related work by the role it plays in CausalGap: structural robustness in traffic forecasting, representation comparison, and failure-mode diagnosis. This separation clarifies why output degradation alone does not subsume the proposed response coordinate.

2.1 Limitations of Traffic Prediction under Structural Interventions

Many traffic predictors use fixed-topology protocols (Yu et al., 2018; Li et al., 2018; Wu et al., 2019; Bai et al., 2020), while conventional road-traffic studies mainly focus on predictive accuracy (Cheng et al., 2022). Time-series foundation models have also emerged (Li et al., 2025; Gruver et al., 2023; Jin et al., 2024; Ansari et al., 2024), but their structural responses remain unstudied. Related robust, causal, and structure-aware methods include STEVE for confounder representations (Ji et al., 2025), CausalGRIT for causal graph reasoning (Liu et al., 2025), CCHMM for structured causal transitions (Zhao et al., 2023), and TF-MVGNN for multiple graph views (Cheng et al., 2025). DCGCN models dynamic causal structure, whereas ST-Norm refines temporal and spatial components (Lin et al., 2023; Deng et al., 2021). Prior robustness work primarily quantifies output degradation or transfer performance; CausalGap instead evaluates the joint state of selected-interface response and predictive degradation and tests its weak-to-strong predictive value.

2.2 Diagnostic Benchmarks and Representation Analysis Methods

DomainBed (Gulrajani & Lopez-Paz, 2021), WILDS (Koh et al., 2021), and ImageNet-C (Hendrycks & Dietterich, 2019) standardize evaluation under domain, real-world distribution, and corruption shifts. CausalGap extends these traditions by testing whether intervention-induced response adds held-out value beyond prespecified weak-error trajectory summaries. It draws on Centered Kernel Alignment (CKA), Singular Vector Canonical Correlation Analysis (SVCCA), Representational Similarity Analysis (RSA), and Projection-Weighted Canonical Correlation Analysis (PWCCA) (Kornblith et al., 2019; Raghu et al., 2017; Kriegeskorte et al., 2008; Morcos et al., 2018). These tools usually compare static or within-distribution representations; CausalGap instead couples response with degradation and tests whether weak-intervention response adds held-out predictive value beyond the same weak-error summaries, including against these established alternatives.

2.3 Positioning Pathological Rigidity within the Failure Mode Taxonomy

Pathological rigidity differs from known failure modes. Neural collapse describes convergent class-level geometry near the terminal phase of training (Papyan et al., 2020), shortcut learning exploits spurious correlations (Geirhos et al., 2020), out-of-distribution degradation refers broadly to performance loss under distribution shift (Koh et al., 2021), and overconfidence reflects a mismatch between predictive confidence and empirical correctness (Guo et al., 2017). In contrast, pathological rigidity denotes high predictive degradation accompanied by limited change at the selected representation interface. Within the proposed taxonomy, this combination requires jointly analyzing representation response and prediction error, precisely the capability of the response-error landscape.

3 CausalGap: Representation-Response Diagnostic Framework

The framework asks whether a declared structural intervention is registered at a selected representation interface and how that response co-occurs with degradation. We first define the coordinates, then the measurement requirements, controlled access, and weak-to-strong tests.

3.1 Problem Formulation and Diagnostic Interpretation

Let $E_0 = (X_0, A_0, Y_0)$ be the baseline environment and T_k produce $E_k = (X_k, A_k, Y_k)$. Without intervention-consistent outcomes, evaluation is an input-level stress test. For $f = g \circ z$, define $Z_k^{(0)} = z(X_k, A_0)$ and $Z_k^{(T)} = z(X_k, A_k)$. Intervention Prediction Error (IPE) is mean absolute error over the common evaluated samples, nodes, and horizons in E_k.

Error-Only Measurement Limitation. For a nontrivial intervention, two predictors can have identical outputs and error while their selected representations undergo unequal changes; Appendix A.6 constructs such a pair. Hence predictions alone cannot identify representation response.

Low response alone may reflect beneficial invariance. CausalGap therefore interprets response jointly with predictive degradation and treats the regions as operational diagnostics, not latent mechanisms.

Definition 1 (Operational Pathological-Rigidity Region)**.** *A model-intervention instance is assigned to the operational pathological-rigidity region if* $(\mathrm{DMIS}_k, \mathrm{RD}_k)$ *falls in a low-response, high-degradation region with calibration-estimated thresholds. The label denotes a diagnostic region, not a latent mechanism. We set* $\mathrm{DMIS}_{\mathrm{norm}} = \mathrm{clip}_{[0,1]}(\mathrm{DMIS}_k/\delta_Z)$, $\mathrm{RD}_{\mathrm{norm}} = \mathrm{clip}_{[0,1]}(\mathrm{RD}_k/\delta_R)$, *and define the Rigidity Index (RI) as* $\mathrm{RI} = (1 - \mathrm{DMIS}_{\mathrm{norm}})\mathrm{RD}_{\mathrm{norm}}$. *Clipping makes RI bounded and threshold-relative, not globally rank-preserving.*

3.2 State Space Coordinate: DMIS

DMIS measures standardized representation change per unit structural distance. "Dual-Metric" denotes angular and normalized-magnitude displacement. For the centered topology-controlled pair

$\bar{Z}_k^{(0)}, \bar{Z}_k^{(T)}$, define

$$D_{Z,k} = \alpha\big(1 - \cos_F(\bar{Z}_k^{(T)}, \bar{Z}_k^{(0)})\big) + (1-\alpha)\frac{\|\bar{Z}_k^{(T)} - \bar{Z}_k^{(0)}\|_F}{\max(\|\bar{Z}_k^{(0)}\|_F, \epsilon_Z)},$$

where $\alpha \in (0,1)$ defaults to 0.5 and $\epsilon_Z = 10^{-6}$. Degenerate pairs, for which either centered norm is below ϵ_Z, use a prespecified finite cosine convention and are excluded from primary-landscape claims; Appendix A.6 gives details. The Dual-Metric Intervention Sensitivity is

$$\mathrm{DMIS}_\alpha(f, T_k) = \frac{D_{Z,k}}{\max(d_{\mathcal{T},k}, \epsilon_D)},$$

with $\epsilon_D = 10^{-6}$. Holding X_k fixed isolates response to replacing A_0 by A_k. This is the native-distance form $\mathrm{DMIS}^{\mathrm{raw}}$; pooled cross-family analyses use the calibration-normalized $\mathrm{DMIS}^{\mathrm{cal}}$ defined in Appendix A.2. DMIS is invariant under shared orthogonal transformations and, when the numerical stabilizer remains inactive before and after transformation, shared positive scaling; centering removes shared translations. It diagnoses the selected interface, not the complete internal mechanism. This paired contrast measures interface sensitivity rather than a natural direct causal effect because X_k may itself be generated under A_k.

3.3 Diagnostic Requirements and Error-Only Measurement Limitation

Representation Equivalence Principle. As a measurement convention, nondegenerate representation pairs related by a shared orthogonal transformation, or by shared positive scaling for which the stabilizer remains inactive, are equivalent for response measurement. This convention does not cover every functionally equivalent representation; the prediction head affects error but not the response coordinate.

A response coordinate is suitable as one axis of the taxonomy if it is invariant under the adopted equivalence and sensitive to topology response; jointly with RD, the two coordinates distinguish the four operational combinations.

Proposition 1 (Error-Response Measurement Impossibility)**.** *For any nontrivial intervention, equal predictive error need not imply equal selected-interface response: even with fixed dimension and a common readout, there exist f_1, f_2 with $\mathrm{IPE}_k(f_1) = \mathrm{IPE}_k(f_2)$ but $\mathrm{DMIS}_k(f_1) \neq \mathrm{DMIS}_k(f_2)$. Thus output error alone cannot identify the response-error regime.*

The primary taxonomy uses calibration-estimated axis-aligned thresholds; Gaussian mixture model (GMM) partitioning is used only as a robustness check and does not establish discrete latent mechanisms. Low DMIS alone is not labeled rigid: pathological rigidity requires both low response and high RD under the calibrated thresholds.

3.4 Projection Statistics and Evaluation Principle

The absolute post-intervention error is IPE_k. Relative degradation is

$$\mathrm{RD}_k = \frac{\mathrm{IPE}_k - \mathrm{IPE}_0}{\mathrm{IPE}_0 + \tau}, \quad \tau = 0.05\,\mathrm{std}(Y_0).$$

Across K interventions, Average Relative Degradation (ARD) is $\mathrm{ARD} = \frac{1}{K}\sum_{k=1}^{K}\mathrm{RD}_k$. The landscape uses $(\mathrm{DMIS}_k, \mathrm{RD}_k)$.

Controlled Structural-Access Principle. The Dual-Track Structural Access Interface (DSAI) uses Track A for native input and Track B for a shared low-capacity topology encoder. Track B reduces, but does not eliminate, structural-information differences; the two access conditions are kept conceptually separate and are never pooled into a single ranking.

3.5 Incremental Diagnostic Value Hypothesis

We test two claims. Let $\mathcal{C}$ contain prespecified magnitude/baseline-error and generalizable architecture controls; identity indicators are excluded. Let $\mathcal{H}_{RD}$ contain weak-RD mean, normalized-intensity slope, and maximum. **H1 (response value)** compares standardized ridge predictors

$\mathcal{M}_{\mathcal{C}+\mathcal{H}_{RD}}$ and $\mathcal{M}_{\mathcal{C}+\mathcal{H}_{RD}+\overline{\mathrm{DMIS}}^{\mathrm{cal,weak}}}$, pairing each intervention family once with its strong outcome. Its confirmatory endpoint is leave-one-model-family-out macro mean absolute error (macro-MAE) on strong-intervention RD; Appendix A.7 specifies risk and secondary analyses. H1 does not require strong marginal DMIS-RD correlation; it tests conditional weak-to-strong value. **H2 (explicit-summary utility)** asks whether explicitly providing DMIS improves post-hoc steering beyond a fixed-capacity baseline receiving the same raw information:

$$\Delta\mathrm{IPE}_{\mathcal{M}_{\mathrm{matched}}+\mathrm{DMIS}} > \Delta\mathrm{IPE}_{\mathcal{M}_{\mathrm{matched}}},$$

where $\mathcal{M}_{\mathrm{matched}}$ exposes $Z_k^{(T)}$, RD, $D_{Z,k}$, $d_{\mathcal{T},k}$, and their source variables at matched capacity. H2 uses within-family $\mathrm{DMIS}^{\mathrm{raw}}$ with no calibrated-distance transformation; because it is deterministic given these constituents, H2 tests parameterization utility, not additional information. Here $\Delta\mathrm{IPE} = \mathrm{IPE}_{\mathrm{before}} - \mathrm{IPE}_{\mathrm{after}}$. RD uses labels, so steering is post hoc, not deployable adaptation or causal identification.

4 Experimental Setup

Accordingly, the protocol separates evidence roles: controlled synthetic dynamics support the confirmatory H1 analysis and the prespecified post-hoc H2 steering test, while abstract SUMO supplies intervention-consistent rank/regime replication. Observational datasets and the real-world event are documented as non-counterfactual contexts and are excluded from the reported response-failure analyses. This separation prevents simulator alignment or missing counterfactual outcomes from being interpreted as confirmatory evidence.

4.1 Datasets and Intervention Protocol

The benchmark documents 8 traffic datasets: NYCTaxi, NYCBike1, NYCBike2, BJTaxi, METR-LA, PeMS-BAY, PeMS04, and PeMS08; details are in Appendix A.3. Controlled nonlinear graph dynamics provide the confirmatory weak-to-strong setting. SUMO provides a separate simulation replication built from a connected METR-LA subgraph: it preserves the selected topology and is aligned only to coarse traffic statistics, rather than reconstructing the real METR-LA road network or trajectories.

Structural interventions include edge removal, edge-weight modification, and node isolation at multiple intensities. Time shifts are observational-shift controls. In the SUMO replication, the simulator generates intervention-consistent Y_k for DMIS, RD, model-rank correspondence, and four-regime replication; it is not used for held-out-setting H1 or steering. The observational datasets and the I-405 closure remain documented contexts, but do not enter the reported response-failure analyses. A calibration split disjoint from test fixes intervention normalizers and regime thresholds; steering parameters are calibrated only in the controlled synthetic setting. Sibling perturbations sharing graph/checkpoint/seed remain in one outer fold, and H1 forms one prespecified weak-history summary before one strong pairing.

4.2 Models and Baselines

The benchmark evaluates 21 models in five prespecified families: Forecasting spatio-temporal graph neural networks (STGNNs), Robust and Structure-Aware Models, Structural-Dynamics Models, Generic Time-Series Models, and Foundation Models. Core analyses use fixed benchmark configurations; additional domain-generalization retraining is outside the primary evidence scope. Appendix A.4 gives full model names, citations, assignment criteria, and implementations.

The five model families and all assignments are fixed from architectural criteria before H1 outcomes are inspected. The primary analysis uses the representation immediately preceding the prediction head as a common extraction interface. For topology-agnostic models, Track B evaluates the backbone plus its declared topology adapter; the adapter controls are specified in Appendix A.5, while extraction-layer sensitivity is outside the submitted analysis.

4.3 Foundation-Model Configuration

Family V models use one prespecified benchmark configuration in all core analyses. Zero-shot, linear-probe, Low-Rank Adaptation (LoRA), and full fine-tuning sweeps are treated as secondary adaptation sensitivities rather than components of the primary H1, H2, or response-landscape protocol.

4.4 Representation Steering

The steering experiment compares multiple directions for updating the representation of a pathologically rigid model. All directions are applied at equal norm, and error reduction is measured. Figure 2 illustrates the setup.

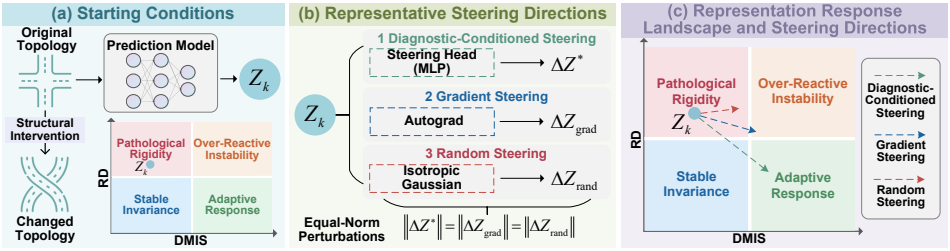

Figure 2: Overview of representation steering experiment. Schematic of diagnostic-conditioned steering comparison and diagnostic-state steering trajectories.

As Figure 2 makes explicit, a low-capacity multilayer perceptron (MLP) steering head h_ϕ produces the diagnostic-conditioned direction from $Z_k^{(T)}$, $\mathrm{DMIS}_k^{\mathrm{raw}}$, and RD_k. Capacity-matched controls expose $D_{Z,k}$, $d_{\mathcal{T},k}$, and their source variables without DMIS; a constant-zero coordinate preserves input dimension. Shuffled, gradient, and random controls test alignment and direction specificity. All heads use calibration interventions with z and g frozen; hyperparameter selection never accesses held-out strong interventions. The prespecified H2 contrast compares explicit DMIS against the strongest information-matched baseline. Because RD uses labels, this is an oracle, post-hoc parameterization test, not label-free adaptation.

4.5 Counterexample and Invariance Checks

Weak-error comparability analysis forms all candidate model pairs under weak edge-weight interventions using the prespecified RD tolerance $\eta_m = 0.05$, then compares high- versus low-DMIS members on stronger held-out interventions from the same family. The GWNet-PatchTST pair is retained only as an illustrative case; Appendix A.7.4 fixes the complete-cohort protocol. Cross-family transfer to edge removal is exploratory. For each supported transformation, the invariance audit evaluates 100 nondegenerate paired representation instances under shared orthogonal rotation, common feature permutation, and admissible positive scaling.

4.6 Robustness Analyses

The robustness analyses reuse saved predictions and representations where applicable. Threshold/GMM stability, weak-error comparability, alternative response measures, invariance, the DSAI access-sensitivity comparison (Appendix A.7.7), and the primary DMIS mixing-weight sensitivity are post-processing analyses. H2 trains separate steering heads and is restricted to controlled synthetic dynamics.

5 Empirical Results

Following this evidence hierarchy, Sections 5.1-5.4 report association, incremental value, simulator replication, and regime sensitivity, while Sections 5.5-5.9 address weak-error comparability, alternative measures, invariance, steering, and DMIS mixing-weight sensitivity.

5.1 Association Between Representation Response and Predictive Degradation

Table 1 reports the pooled cross-family DMIS-RD association in intervention-consistent settings; downstream H1 and SUMO audits are collected in Appendix A.7.

Table 1: Cross-family association between calibration-normalized DMIS and RD in intervention-consistent settings.

Metric	Value	95% CI lower	95% CI upper
Pearson r	-0.006	-0.025	0.040
Spearman ρ	0.251	0.141	0.361
Kendall τ	0.179	0.106	0.251

As summarized in Table 1, pooled DMIS and RD exhibit essentially no global linear association (Pearson $r = -0.006$), while rank statistics indicate a modest positive monotonic association (Spearman $\rho = 0.251$, Kendall $\tau = 0.179$). Figure 3 visualizes the same contrast over the full range and a central-range zoom. DMIS measures response magnitude, not response quality, so large responses may be adaptive or over-reactive. Thus DMIS is not a linear proxy for RD; H1 remains the confirmatory incremental test.

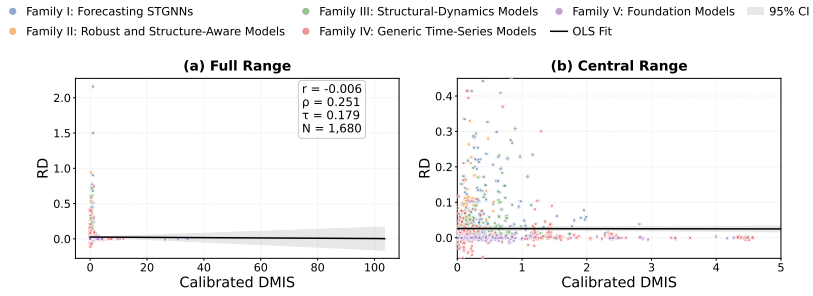

Figure 3: **Cross-family DMIS-RD association.** (a) Full range; (b) $DMIS \leq 5$ zoom for visualization only. All $N = 1{,}680$ observations enter the statistics.

The full-range and zoomed views in Figure 3 show that the modest rank association is not driven by a single narrow region of the response scale, motivating the conditional H1 test rather than a marginal-correlation claim.

5.2 Incremental Diagnostic Value Verification

Building on the association analysis, H1 tests whether weak-intervention DMIS adds information beyond the prespecified weak-RD trajectory summary in the controlled synthetic setting. Leave-one-model-family-out macro-MAE is the confirmatory endpoint. Figure 4 summarizes the primary nested-risk comparison and its weak-history/intensity specification checks; exact risks are tabulated in Appendix A.7.1, Table 9. H2 steering is separate from H1; Appendix A.7.8 specifies that test.

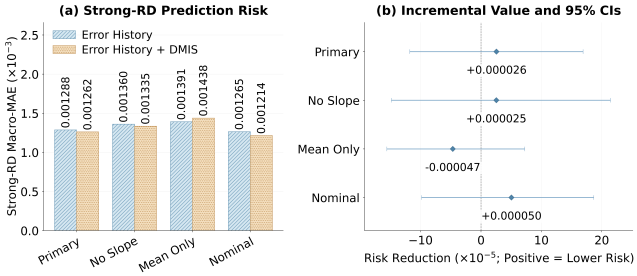

Figure 4: Incremental diagnostic value of DMIS under weak-history and intensity specifications.

The primary point estimate decreases macro-MAE from 0.001288 to 0.001262, a reduction of 2.6×10^{-5}. The no-slope and nominal-intensity checks retain a positive point estimate, whereas the mean-only history reverses direction. Relative to the uncertainty shown in Figure 4, H1 is therefore interpreted as a small, specification-sensitive incremental point estimate rather than a uniform improvement across reduced baselines.

5.3 Operational Regimes and SUMO Replication

SUMO is an intervention-consistent replication environment, not a reconstruction of real METR-LA dynamics or an H1 setting. Figure 5 compares cross-environment model-rank correspondence for DMIS and RD.

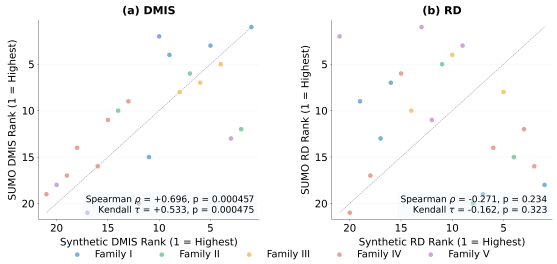

Figure 5: Model-rank replication between controlled synthetic dynamics and the abstract SUMO scenario.

As Figure 5 shows, DMIS rankings have stronger cross-environment correspondence (Spearman $\rho = 0.696$, $p < 0.001$) than RD rankings, so response and degradation ordering are not interchangeable. Figure 6 then reports the four SUMO regimes by model family: Families I-III show larger over-reactive shares, whereas Families IV-V show larger pathological-rigidity shares. These are descriptive abstract-simulator results, not claims of real METR-LA reconstruction; exact proportions are in Appendix A.7.2, Table 10.

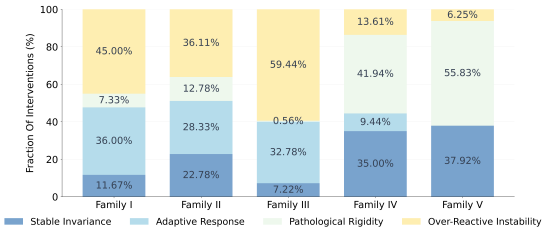

Figure 6: Operational regime distributions across model families in the abstract SUMO replication.

5.4 Sensitivity of Operational Regime Assignments

Building on the family-level patterns in Figure 6, we perturb the primary $q = 0.50$ threshold to $q \in \{0.40, 0.60\}$ and compare the calibrated partition with a four-component calibration-fit GMM. Table 2 reports held-out agreement; Appendix A.7.3 gives the full breakdown.

Table 2: Operational-regime sensitivity to local threshold perturbations and GMM partitioning. GMM agreement is measured against the primary $q = 0.50$ partition.

Setting	GMM ARI	GMM NMI	Min. agreement over $q \in \{0.40, 0.60\}$
Controlled synthetic	0.135	0.253	0.810
Abstract SUMO replication	0.346	0.540	0.810

Table 2 shows at least 81.0% local-threshold agreement in both environments, but weaker GMM recovery (ARI/NMI $0.135/0.253$ synthetic; $0.346/0.540$ SUMO). The quadrants are therefore locally stable operational diagnostics, not empirically recovered latent modes.

5.5 Weak-Error Comparability Audit

We next ask whether comparable weak-error behavior can mask different strong-intervention degradation. Under the prespecified $\eta_m = 0.05$ tolerance, all 210 candidate pairs are eligible, so the audit is non-selective; Appendix A.7.4, Table 12 gives the full construction. The aggregate weak-RD gap is 0.000155; strong RD is 0.000346 for higher-DMIS and 0.000561 for lower-DMIS members (difference -0.000215). This descriptive separation is compatible with weak-to-strong diagnostic value, but is neither selective-matching evidence nor a standalone inferential result.

5.6 Robustness Across Response Measures

To test coordinate dependence, we repeat the frozen H1 protocol with CKA, SVCCA, RSA, and PWCCA. Rank agreement with DMIS remains high (Spearman 0.923-0.976); CKA, SVCCA, and PWCCA preserve the positive H1 direction, whereas RSA is slightly negative. Full results are in Appendix A.7.5.

5.7 Invariance Sanity Checks

Figure 7 summarizes 100 valid instances per transformation. Shared rotation, feature permutation, and admissible positive scaling yield maximum absolute DMIS deviations of 4.90×10^{-15}, 5.83×10^{-15}, and 4.26×10^{-14}, respectively.

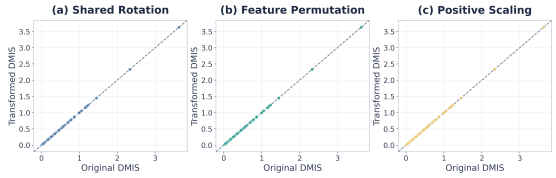

Figure 7: DMIS invariance under supported shared representation transformations.

All three checks in Figure 7 give Spearman rank agreement 1.000 and operational-regime agreement 100%, confirming numerical implementation stability; Appendix A.7.6 gives the complete audit.

5.8 H2 Steering Check

H2 compares explicit DMIS with matched controls. Matched+DMIS yields 0.1007% error reduction versus 0.0582% for the information-matched baseline and 0.0904% for shuffled DMIS. The $+0.0425$-point baseline margin but $+0.0103$-point placebo margin indicates a modest, configuration-specific parameterization effect; gradient and random directions remain auxiliary (Appendix A.7.8).

5.9 Sensitivity to the DMIS Mixing Weight

Across $\alpha \in \{0.25, 0.50, 0.75\}$, model-rank agreement remains at least 0.995, the Synthetic H1 risk reduction stays positive, and SUMO regime agreement is 1.000; Appendix A.7.9 reports the full audit.

6 Conclusion

CausalGap pairs DMIS with RD to diagnose representation response under structural interventions, with DSAI separating native and controlled structural access. H1 is small and specification-sensitive; alternative-measure, invariance, DSAI access-sensitivity, and mixing-weight audits support measurement robustness. Abstract SUMO replicates rank/regime patterns, while H2 remains a modest configuration-specific effect. The benchmark complements error-based evaluation without identifying latent causal mechanisms; limitations are interface choice, architectural asymmetry, and intervention scope.

AI USE STATEMENT

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

ETHICS STATEMENT

This research uses publicly available traffic sensor datasets, simulated traffic dynamics, and a documented freeway-closure case study; it does not involve human-subject recruitment or personally identifiable information. The datasets and accompanying resources are used in accordance with their documented research purposes and applicable access conditions. CausalGap is intended for research and diagnostic evaluation, not for direct deployment in safety-critical traffic-management systems. Its conclusions are limited to the evaluated datasets, interventions, outcome-generation settings, architectures, and representation interfaces. Differences in sensing coverage, network topology, infrastructure, and data availability across cities may affect the diagnostic results and should be considered before applying them to other regions or operational settings. The authors take responsibility for the integrity, documentation, and appropriate use of the released benchmark resources.

REPRODUCIBILITY STATEMENT

CausalGap is designed to make each reported comparison traceable to a declared dataset, structural intervention, outcome-generation setting, model configuration, representation interface, and evaluation procedure. Section 4 describes the primary experimental protocol. Appendix A.2 specifies intervention construction, calibration, simulator-based outcome generation, and real-world event metadata; Appendices A.3 and A.4 document the datasets, topology construction, evaluated models, and benchmark adaptations. Appendix A.7 provides extended results, sensitivity analyses, statistical specifications, and robustness checks. The anonymous supplementary repository provides the released code, data-processing scripts, intervention generators, model configurations, environment specifications, and evaluation scripts associated with the reported benchmark results.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

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

Haoyuan Cheng, Yufeng Wang, Jianhua Ma, and Qun Jin. Tf-mvgnn: An accurate traffic forecasting framework based on spatial-temporal graph neural network through exploiting multiple-view graph construction and learning. *Neural Computing and Applications*, 37(20):14657–14671, 2025. doi: 10.1007/s00521-024-10508-4.

Zeyang Cheng, Jian Lu, Huajian Zhou, Yibin Zhang, and Lin Zhang. Short-term traffic flow prediction: An integrated method of econometrics and hybrid deep learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5231–5244, 2022. doi: 10.1109/TITS.2021.3052796.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine*

Learning, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Jinliang Deng, Xiusi Chen, Renhe Jiang, Xuan Song, and Ivor W Tsang. St-norm: Spatial and temporal normalization for multi-variate time series forecasting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 269–278, 2021. doi: 10.1145/3447548.3467330.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19622–19635, 2023. URL https://arxiv.org/abs/2310.07820.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Shengnan Guo, Youfang Lin, Huaiyu Wan, Xiucheng Li, and Gao Cong. Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5415–5428, 2022. doi: 10.1109/TKDE.2021.3056502.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of the International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

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. doi: 10.1145/3690624.3709201.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pinyu Chen, Yuxuan Liang, Yuanfang Li, Shirui Pan, and Qingsong Wen. Time-llm: Time series forecasting by reprogramming large language models. In *Proceedings of the International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Unb5CVPtae.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the International Conference on Machine Learning*, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the International Conference on Machine Learning*, pp. 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

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

Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, and Bin Yang. Tsfm-bench: A comprehensive and unified benchmark of foundation models for time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5595–5606, 2025. doi: 10.1145/3711896.3737442.

Junpeng Lin, Ziyue Li, Zhishuai Li, Lei Bai, Rui Zhao, and Chen Zhang. Dynamic causal graph convolutional network for traffic prediction. In *Proceedings of the IEEE 19th International Conference on Automation Science and Engineering (CASE)*, pp. 1–8, 2023. doi: 10.1109/CASE56687.2023.10260564.

Sheng Liu, Jun Ying, and Fan Bai. Causalgrit: Causal graph reasoning for traffic congestion prediction–from statistical association to casual intervention. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, pp. 173–185, 2025. doi: 10.1145/3748636.3762721.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *Proceedings of the International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=JePfAI8fah.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In *Proceedings of the Advances in neural information processing systems*, volume 31, 2018.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *Proceedings of the International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Vardan Papyan, Xiao Y Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Proceedings of the Advances in neural information processing systems*, volume 30, 2017.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, et al. Lag-llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023. URL https://arxiv.org/abs/2310.08278.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024. URL https://arxiv.org/abs/2402.02592.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *Proceedings of the International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

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

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

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

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. doi: 10.1609/aaai.v37i4.25619.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the International Conference on Machine Learning*, pp. 27268–27286, 2022. URL `https://proceedings.mlr.press/v162/zhou22g.html`.

SUPPLEMENTARY MATERIAL

Appendix roadmap. The supplement follows the measurement-to-evidence-to-reproducibility logic of the main paper rather than repeating it. Appendix A.2 specifies structural interventions, calibration, intervention-consistent outcome generation, and the scope of the abstract SUMO replication. Appendix A.3 records the benchmark datasets and topology construction, while Appendices A.4 and A.5 document model-family assignments, benchmark adaptations, training details, representation interfaces, and the Dual-Track Structural Access Interface. Appendix A.6 provides the error-response measurement construction and the supported invariances of DMIS. Appendix A.7 then collects the numerical evidence underlying Sections 5.1-5.9, including H1, SUMO replication, regime sensitivity, weak-error comparability, alternative response measures, invariance, DSAI access sensitivity, H2 steering, and DMIS mixing-weight sensitivity. Appendix A.8 closes the supplement with artifact and reproducibility checks.

A ADDITIONAL EXPERIMENTAL DETAILS AND RESULTS

A.1 NOTATION AND ABBREVIATIONS

Table 3 standardizes methodological abbreviations. Official model names are introduced with their expansions in Appendix A.4; thereafter the short names are used throughout. Stylized model and dataset names whose source publications do not define stable expansions, including STEVE, METR-LA, and PeMS-BAY, are treated as official proper names rather than expanded as abbreviations.

Table 3: Notation and abbreviations used in the paper.

Short form	Meaning	Short form	Meaning
ARD	Average Relative Degradation	DSAI	Dual-Track Structural Access Interface
DMIS	Dual-Metric Intervention Sensitivity	IPE	Intervention Prediction Error
RD	Relative Degradation	RI	Rigidity Index
STGNN	spatio-temporal graph neural network	DG	domain generalization
SUMO	Simulation of Urban Mobility	GMM	Gaussian mixture model
MLP	multilayer perceptron	LoRA	Low-Rank Adaptation
CKA	Centered Kernel Alignment	SVCCA	Singular Vector Canonical Correlation Analysis
RSA	Representational Similarity Analysis	PWCCA	Projection-Weighted Canonical Correlation Analysis
ARI	adjusted Rand index	NMI	normalized mutual information
CI	confidence interval	CSV	comma-separated values
GPU	graphics processing unit	CPU	central processing unit

The notation in Table 3 is used consistently in the protocol and result appendices below.

A.2 INTERVENTION CALIBRATION AND REPRODUCIBILITY

This section describes the structural intervention protocols, generation of intervention-consistent outcomes, construction of the abstract SUMO replication, and real-world event metadata. The SUMO component is designed for controlled replication of structural-response patterns; it is not intended to reconstruct the real METR-LA network, origin-destination demand, or pointwise trajectories.

A.2.1 STRUCTURAL INTERVENTION PROTOCOL

Structural interventions are generated from the observed road-network topologies. Three structural intervention types are used:

- **Edge removal**: an s fraction of all original edges is removed, where $s \in \{0.05, 0.10, 0.15, 0.20\}$. Removed edges are sampled without replacement from a candidate pool consisting of the top 40% of edges ranked by edge betweenness centrality. Edge

betweenness is computed once on A_0. For directed graphs, directed betweenness is used; when A_{ij} represents affinity, centrality distances are defined as $c_{ij} = 1/(A_{ij} + \epsilon_w)$, with $\epsilon_w = 10^{-6}$. When physical road length is available, the recorded length is used directly.

- **Edge-weight modification**: 15% of edges are sampled, and their weights are multiplied by a factor $\gamma \in \{0.2, 0.4, 0.6, 0.8\}$. The corresponding intervention strength is $1 - \gamma$.
- **Node isolation**: a fraction $s \in \{0.1, 0.2, 0.4, 0.6\}$ of nodes is structurally isolated. The intervened adjacency is $A_k = M_k A_0 M_k$, where M_k is a diagonal mask matrix, while node observations remain unchanged ($X_k = X_0$). Prediction error is evaluated on unmasked nodes only.

Time shifts are treated as observational-shift controls, not structural interventions. Intervention sampling uses seeds 0-19, while model training uses seeds 0-9. Each intervention is repeated with 20 independent seeds per type and intensity.

Separation of temporal and intervention splits. The temporal calibration split $\mathcal{D}_{\text{cal}}$ is distinguished from the weaker intervention set $\mathcal{I}_{\text{weak}}$. Thresholds, intervention-magnitude normalizers, and steering heads are fitted exclusively on $\mathcal{D}_{\text{cal}} \times \mathcal{I}_{\text{weak}}$. All reported weak-to-strong predictions are evaluated on $\mathcal{D}_{\text{test}} \times \mathcal{I}_{\text{strong}}$, with no test labels used for model selection or threshold calibration. The specific intensity assignments are summarized in Table 4.

Table 4: Temporal and intervention split matrix.

Purpose	Temporal split	Structural exposure
Standard model training	$\mathcal{D}_{\text{train}}$	Baseline topology only
Early stopping	$\mathcal{D}_{\text{val}}$	Baseline topology only
Threshold/distribution/GMM calibration	$\mathcal{D}_{\text{cal}}$	$\mathcal{I}_{\text{weak}}$
Steering head training	$\mathcal{D}_{\text{cal}}$	$\mathcal{I}_{\text{weak}}$
Weak-to-strong evaluation	$\mathcal{D}_{\text{test}}$	$\mathcal{I}_{\text{strong}}$

Using the split structure in Table 4, for edge removal, $\mathcal{I}_{\text{weak}} = \{0.05, 0.10, 0.15\}$ and $\mathcal{I}_{\text{strong}} = \{0.20\}$; for weight modification, $\mathcal{I}_{\text{weak}} = \{\gamma = 0.8, 0.6, 0.4\}$ and $\mathcal{I}_{\text{strong}} = \{\gamma = 0.2\}$; for node isolation, $\mathcal{I}_{\text{weak}} = \{0.1, 0.2, 0.4\}$ and $\mathcal{I}_{\text{strong}} = \{0.6\}$. Thus every confirmatory intervention family has three prespecified weak levels and one held-out strong level. Test labels are used only for final evaluation and the explicitly label-dependent post-hoc steering analysis; they are never used for model selection, threshold estimation, magnitude normalization, or steering-head training.

Scope of diagnostic claims. The controlled synthetic dynamics support the confirmatory weak-to-strong H1 analysis and the completed H2 steering test. The abstract SUMO scenario supports intervention-consistent DMIS, RD, model-rank correspondence, four-regime distributions, and regime-assignment replication only. Observational datasets without counterfactual outcomes are not used in the reported response-failure analyses.

Generation of intervention-consistent outcomes. The controlled scenario generator modifies the topology, while the corresponding traffic outcomes are generated by an intervention-aware dynamics module. For every baseline scenario, exogenous inputs and simulation randomness are paired, and baseline and intervened trajectories are generated as

$$(X_0, Y_0) \sim P(\cdot \mid A_0, \omega), \qquad (X_k, Y_k) \sim P(\cdot \mid A_k, \omega),$$

using the same exogenous realization ω. In the synthetic benchmark this isolates the declared structural change. In SUMO, the same abstract demand template and random seed are reused across each baseline/intervention pair so that the simulator supplies intervention-consistent Y_k; this pairing is used for replication metrics, not for H1 or steering.

For NYCTaxi, NYCBike1/2, BJTaxi, PeMS-BAY, PeMS04, and PeMS08, counterfactual outcomes are unavailable; they are therefore excluded from the reported response-failure analyses. METR-LA additionally supplies the connected subgraph and coarse statistics used to construct the abstract SUMO replication. The I-405 event is retained only as event metadata. Table 5 summarizes the admitted evidence roles before the detailed metric definitions.

Table 5: Evidence scope fixed before evaluation. A checkmark denotes an admitted analysis used in the reported evidence.

Setting	Counterfactual Y_k	DMIS	RD	Regime	H1/H2	Evidence role
Controlled synthetic	✓	✓	✓	✓	✓	H1 confirmatory; H2 post-hoc
Abstract SUMO replication	✓	✓	✓	✓	-	Simulation replication

As Table 5 makes explicit, only controlled synthetic dynamics support the confirmatory H1 analysis and the prespecified post-hoc H2 test; SUMO has the narrower role of intervention-consistent replication.

Paired representation and prediction error evaluation. The paired representation comparison and the predictive-error comparison serve different purposes. DMIS holds X_k fixed and varies only the topology input:

$$Z_k^{(0)} = z(X_k, A_0), \qquad Z_k^{(T)} = z(X_k, A_k).$$

RD compares prediction errors in the baseline and intervened environments using their corresponding targets:

$$\mathrm{IPE}_k = \frac{1}{|\mathcal{S}_k||\mathcal{U}_k|H} \sum_{s\in\mathcal{S}_k} \sum_{v\in\mathcal{U}_k} \sum_{h=1}^{H} \left| \widehat{Y}_{k,s,v,h} - Y_{k,s,v,h} \right|,$$

where $\mathcal{S}_k$, $\mathcal{U}_k$, and H are the common sample, evaluated-node, and forecast-horizon supports used by the paired baseline and intervention. Then

$$\mathrm{RD}_k = \frac{\mathrm{IPE}_k - \mathrm{IPE}_0}{\mathrm{IPE}_0 + \tau},$$

where IPE_k uses (X_k, A_k, Y_k), and IPE_0 uses (X_0, A_0, Y_0). For node isolation, $\mathcal{U}_k$ is the same unmasked node set in both errors.

Intervention magnitude. For edge removal and edge-weight modification, intervention magnitude is defined as

$$d_{\mathcal{T},k} = \frac{\|A_k - A_0\|_F}{\|A_0\|_F + \epsilon_G}, \qquad \epsilon_G = 10^{-6}.$$

For node isolation,

$$d_{\mathcal{T},k} = \frac{|\mathcal{V}_{\mathrm{iso},k}|}{N}.$$

Only interventions satisfying $d_{\mathcal{T},k} \geq \delta_A = 0.01$ are evaluated. Thus all analyzed interventions are nontrivial, ensuring $d_{\mathcal{T},k} > 0$.

Within- and cross-family DMIS. Within each intervention family, raw DMIS is computed using $d_{\mathcal{T},k}$. For cross-family analysis, magnitudes are mapped relative to the fixed calibration reference set $\mathcal{C}_{d,t}$:

$$\widetilde{d}_{\mathcal{T},k} = \min\left\{ \frac{1 + \sum_{j\in\mathcal{C}_{d,t}} \mathbb{I}(d_{\mathcal{T},j} < d_{\mathcal{T},k}) + \frac{1}{2}\sum_{j\in\mathcal{C}_{d,t}} \mathbb{I}(d_{\mathcal{T},j} = d_{\mathcal{T},k})}{|\mathcal{C}_{d,t}| + 1}, 1 \right\}.$$

The calibration mapping is fitted separately for each dataset and intervention family using $\mathcal{D}_{\mathrm{cal}} \times \mathcal{I}_{\mathrm{weak}}$, and is then held fixed. We define

$$\mathrm{DMIS}^{\mathrm{raw}}_{k,\alpha} = \frac{D_{Z,k}}{\max(d_{\mathcal{T},k}, \epsilon_D)}, \qquad \mathrm{DMIS}^{\mathrm{cal}}_{k,\alpha} = \frac{D_{Z,k}}{\max(\widetilde{d}_{\mathcal{T},k}, \epsilon_D)}.$$

Primary H1 and cross-family comparisons use $\mathrm{DMIS}^{\mathrm{cal}}$, because they pool intervention families with different native distance scales. Within-family analyses repeat $\mathrm{DMIS}^{\mathrm{raw}}$ as a prespecified sensitivity check. The calibration mapping uses only $\mathcal{D}_{\mathrm{cal}} \times \mathcal{I}_{\mathrm{weak}}$ and is frozen before any strong outcome is inspected.

Aggregation. For a metric M, seed-level results are first averaged within each model-dataset-intervention-family-intensity configuration. The family-balanced summary is

$$M_{\text{macro}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{1}{|\mathcal{I}_t|} \sum_{k \in \mathcal{I}_t} M_{t,k}.$$

Thus each intervention family receives equal weight irrespective of its number of intensity levels. Confidence intervals are computed separately for each analysis and are described in the corresponding table captions.

Synthetic dynamics. The synthetic graph dynamics benchmark is defined by the discrete-time nonlinear graph dynamical system

$$x_{t+1} = \sigma\left(\beta_1 x_t + \beta_2 \widetilde{A}_k x_t + u_t\right) + \epsilon_t,$$

where $x_t \in \mathbb{R}^N$, $\widetilde{A}_k$ is the row-normalized adjacency matrix, $\beta_1 = 0.7$, $\beta_2 = 0.2$, σ is the element-wise ReLU activation, $u_t \sim \mathcal{N}(0, 0.1^2 I)$ is an exogenous input, and $\epsilon_t \sim \mathcal{N}(0, 0.05^2 I)$ is process noise. The system is simulated for 200 burn-in steps, after which 5,000 consecutive snapshots are generated. Baselines and interventions use the same random seed for ϵ_t and u_t. The input horizon is $H = 12$, and the prediction target is the next 12 steps: $Y_t = [x_{t+1}, x_{t+2}, \ldots, x_{t+12}]$. This matches the real-data forecasting setting.

SUMO intervention timeline. The SUMO replication uses an abstract road network constructed from a connected METR-LA subgraph. The network preserves the selected graph connectivity but does not reproduce the real road geometry or real origin-destination matrix. A simple demand template is tuned only so that baseline simulations match a small set of coarse METR-LA statistics. The last pre-intervention observation is recorded at t_0; the intervention is then applied and the target covers $[t_0 + 1, t_0 + F]$ with $F = 12$. The input window $[t_0 - H + 1, t_0]$ is entirely pre-intervention. Demand-template parameters and random seeds are held fixed across each paired run, and routes are recomputed after the structural change. Edge removal closes selected abstract road segments, edge-weight modification changes their maximum speeds by γ, and node isolation disables incident segments. Failed insertion, teleport, and disconnected-pair rates are logged as simulator diagnostics.

A.2.2 Controlled Scenario Generator

Algorithm 1 summarizes the generation process. The algorithm receives a type-specific parameter η: for edge removal and node isolation, η is the fraction s; for weight modification, η is the weight factor γ, and the strength is $1 - \gamma$.

A.2.3 Abstract SUMO Scenario Construction and Coarse Alignment

The paired-output contract returned by Algorithm 1 is retained in the abstract SUMO replication, while the simulator itself is aligned only at a coarse statistical level.

A connected subgraph is selected from the METR-LA sensor graph and converted into an abstract SUMO road network whose purpose is to preserve the selected connectivity pattern under structural interventions. Node positions, road lengths, lane attributes, and demand are simulation parameters rather than claims about the corresponding physical roads. We do not fit a real OD matrix, reconstruct road geometry, align individual SUMO edges to physical segments, or minimize pointwise trajectory error.

Alignment uses only a small prespecified set of coarse baseline statistics computed on the selected METR-LA subgraph: mean speed, speed variability, congestion proportion, and peak-to-valley ratio. The abstract demand template and global simulator parameters are adjusted on the training period to bring these summaries into comparable ranges; no test-period statistic is used. The alignment is therefore a plausibility check for the simulation baseline, not a digital-twin calibration or evidence of physical equivalence. Consistent with this scope, the resulting SUMO outputs are used only to generate intervention-consistent Y_k for DMIS, RD, rank correspondence, and regime replication. They are excluded from held-out-setting H1 and from all steering-head training or evaluation.

Algorithm 1: Controlled Scenario Generator (CSG)

Require: Graph $\mathcal{G} = (V, E)$, intervention type t, type-specific parameter η, seed r
Initialize $A \leftarrow A_0$, $X \leftarrow X_0$, $Y \leftarrow Y_0$
if $t =$ edge removal **then**
  Compute betweenness centrality on A
  Form candidate pool from top 40% edges
  Sample η fraction of all edges from the pool with seed r, remove from A
else if $t =$ weight mod **then**
  Sample 15% edges with seed r
  Multiply weights of sampled edges by factor η
else if $t =$ node isolation **then**
  Sample η fraction of nodes with seed r
  Set $A \leftarrow MAM$, where M is the mask
  Leave X unchanged, evaluate on unmasked nodes
end if
if simulator available **then**
  Generate (X_k, Y_k) under A_k with common random numbers ω
else
  Keep (X_0, Y_0); this is an input perturbation only
end if
return intervened adjacency A_k, modified input X_k, target Y_k, isolated node set if applicable

A.2.4 I-405 Closure Case Metadata

The closure occurred from 22:00 on 2023-11-15 to 05:00 on 2023-11-17, affecting northbound LA-405 between Skirball Center Dr and Mulholland Dr. Data from Caltrans PeMS. The closure reduced capacity on a segment with typical volume 4,200 veh/h, affecting eight upstream detectors. Station IDs and exclusion rules are in the repository.

The I-405 closure is retained as contextual event metadata rather than an intervention-identifiable benchmark. The sensor/event metadata do not by themselves define a unique intervened adjacency A_k, so the primary manuscript does not assign DMIS, RD, regime labels, or rigidity prevalence to this event. It is not used to fit thresholds, GMM components, regression coefficients, steering heads, or any H1/H2 statistic. Any optional topology reconstruction released with the repository is descriptive and excluded from the reported analyses.

A.3 Benchmark Datasets

Table 6 summarizes the eight datasets. All splits are chronological 7:1:2. Z-score statistics are computed on train only; interpolation may use within-split bracketing but not cross-split observations; forward-filling is reported as a sensitivity.

Table 6: Statistics of the eight benchmark datasets.

Dataset	#Nodes	#Timesteps	Time Span	Interval	Type
NYCTaxi	200	2,880	Jan-Mar 2015	30 min	Taxi flow
NYCBike1	128	4,392	Apr-Sep 2014	1 hour	Bike flow
NYCBike2	200	2,976	Jul-Aug 2016	30 min	Bike flow
BJTaxi	1024	5,596	Mar-Jun 2015	30 min	Taxi flow
METR-LA	207	34,272	Mar-Jun 2012	5 min	Traffic speed
PeMS-BAY	325	52,116	Jan-May 2017	5 min	Traffic speed
PeMS04	307	16,992	Jan-Feb 2018	5 min	Traffic flow
PeMS08	170	17,856	Jul-Aug 2016	5 min	Traffic flow

The dataset scales and sampling intervals in Table 6 motivate reporting topology construction separately. Table 7 specifies how the graph topology is constructed for each dataset. All data-derived topologies are constructed using the training split only.

Table 7: Dataset-specific topology construction.

Dataset	Node definition	Edge construction	Edge weight
NYCTaxi	Spatial zones	Geographic adjacency	Binary / distance
NYCBike1	Stations/zones	Geographic adjacency	Binary / distance
NYCBike2	Stations/zones	Geographic adjacency	Binary / distance
BJTaxi	Spatial zones	Geographic adjacency	Binary / distance
METR-LA	Sensors	Road-network adjacency	Distance kernel
PeMS-BAY	Sensors	Road-network adjacency	Distance kernel
PeMS04	Sensors	Road-network adjacency	Distance kernel
PeMS08	Sensors	Road-network adjacency	Distance kernel

Following the construction rules summarized in Table 7, geographic adjacency uses dataset-native zone adjacency when available; otherwise the distance threshold is selected using the training split only and then frozen, with the selected values released in the benchmark configuration. Edge weights are set to $\exp(-d_{ij}/\sigma^2)$, where d_{ij} is geographic distance and σ is the median distance among connected training edges. No topology-construction parameter is updated on calibration or test data.

A.4 Detailed Description of Baseline Models

The benchmark comprises 21 models spanning five prespecified architectural families, together with five DG methods adapted as robust baselines. All models are configured according to their original publications and official implementations, using the hyperparameter settings described in Section 4.2. A concise description of each method is provided below, highlighting its core architectural principle and its relevance to representation-response evaluation. Model names, original references, and implementation details are cross-checked with Section 4.2 to maintain consistency.

Family assignment follows the model's published architecture and inference interface, not its observed CausalGap score. The criteria and assignments in Table 8 were fixed before H1 outcomes were inspected. A hybrid model is assigned by the component that determines how topology enters the representation used for prediction; no model changes family across datasets, adaptation levels, or DSAI tracks.

Primary-evidence scope. The confirmatory H1 matrix targets the 21 benchmark architectures in the controlled synthetic setting. SUMO is audited separately as a replication matrix for DMIS, RD, model ranking, and operational regimes. Failed or inadmissible runs are retained in the released manifest rather than silently removed. Table 8 records the architecture-based family definitions used by the outer-fold analysis.

Table 8: Prespecified model-family criteria and fixed assignments.

Family	Assignment criterion	Members
I	Forecasting backbone contains a graph operator for spatial dependence.	STGCN, DCRNN, ASTGNN, Graph WaveNet, AGCRN
II	Traffic-specific objective or module explicitly targets robustness, invariance, or transfer.	STEVE, DCGCN, ST-Norm
III	Published model explicitly parameterizes causal or multi-view structural dynamics.	CausalGRIT, CCHMM, TF-MVGNN
IV	General time-series forecaster trained for the evaluated dataset, without traffic-specific structural design.	PatchTST, iTransformer, TimesNet, FEDformer, DLinear, Autoformer
V	Pretrained general-purpose time-series foundation model evaluated under one prespecified benchmark configuration.	Chronos, TimesFM, Moirai, Lag-Llama

The assignments in Table 8 therefore encode architectural criteria rather than observed CausalGap behavior. The boundary between Families II and III is operational: Family II contains methods whose

primary published contribution is robust, invariant, or transferable traffic prediction, even when they estimate a dynamic or causal graph; Family III requires an explicit structural-dynamics state or multi-view structural mechanism to be part of the forecasting formulation. Thus DCGCN remains in Family II because its dynamic causal graph supports robust traffic prediction, whereas CausalGRIT and CCHMM expose structural-dynamics states as central modeling objects. These rules are applied without reference to DMIS, RD, or H1 outcomes.

FAMILY I - FORECASTING STGNNS

- **STGCN** (Yu et al., 2018): A spatio-temporal graph convolutional network that integrates graph convolution and temporal gated convolution to capture spatial and temporal dependencies jointly, employs Chebyshev polynomial approximation for efficient spectral graph convolution, and serves as a canonical spatio-temporal baseline.
- **DCRNN** (Li et al., 2018): A diffusion convolutional recurrent neural network that models traffic flow as a diffusion process on directed graphs. The encoder-decoder architecture with scheduled sampling makes this model a strong representative of recurrent-based spatio-temporal models.
- **ASTGNN** (Guo et al., 2022): An attention-enhanced spatio-temporal graph network that decouples spatial and temporal learning, dynamically adjusting spatial attention weights and temporal dependencies to provide more flexible modeling capacity under varying traffic conditions.
- **GWNet** (Wu et al., 2019): Graph WaveNet, which combines a self-adaptive adjacency matrix with stacked temporal convolutions, learning hidden spatial dependencies directly from data without requiring a predefined graph. This data-driven structure learning makes it particularly interesting for testing sensitivity to topological interventions.
- **AGCRN** (Bai et al., 2020): An adaptive graph convolutional recurrent network where node-specific parameters and a learnable adjacency matrix are inferred automatically, capturing fine-grained spatial correlations that provide a complementary perspective to fixed-graph methods.

FAMILY II - ROBUST AND STRUCTURE-AWARE MODELS

- **STEVE** (Ji et al., 2025): A framework that learns basis confounder representations to disentangle environment-specific and invariant features for robust traffic prediction. By isolating confounders, the framework aims to maintain prediction accuracy when distribution shifts occur.
- **DCGCN** (Lin et al., 2023): A dynamic causal graph convolutional network that infers time-varying causal relations between traffic nodes. The evolving causal structure provides a testbed for evaluating how well a model adapts when the underlying causal graph changes.
- **ST-Norm** (Deng et al., 2021): A normalization method that introduces temporal and spatial normalization modules to refine high-frequency and local components of multivariate time series. These modules can be integrated into canonical forecasting backbones to improve predictive modeling.

FAMILY III - STRUCTURAL-DYNAMICS MODELS

- **CausalGRIT** (Liu et al., 2025): A causal graph reasoning model that employs causal graph reasoning to move from statistical association to causal intervention, inferring the causal structure of traffic congestion. The built-in causal discovery module serves as a strong reference for measuring adaptive representation changes under interventions.
- **CCHMM** (Zhao et al., 2023): A causal conditional hidden Markov model that captures multimodal traffic dynamics via structured causal transitions. The explicit causal hidden states allow direct inspection of how the model's internal belief state responds to topological perturbations.
- **TF-MVGNN** (Cheng et al., 2025): A multi-view graph neural network that constructs multiple views of the traffic graph and learns a unified representation. By integrating complementary structural perspectives, this model tests whether multi-view fusion enhances structural-response robustness.

FAMILY IV - GENERIC TIME-SERIES MODELS

- **PatchTST** (Nie et al., 2023): A Transformer-based model that segments time series into patches and applies a standard Transformer encoder, treating each patch as a token. The channel-independent design makes this model an important baseline for examining how the absence of explicit spatial modeling affects structural responsiveness.
- **iTransformer** (Liu et al., 2024): An inverted Transformer that inverts the standard architecture by treating variates as tokens and temporal steps as features, capturing cross-variate dependencies. This alternative tokenization scheme provides a contrasting view of how spatial information is processed without a predefined graph.
- **TimesNet** (Wu et al., 2023): A temporal 2D-variation model that transforms 1D time series into 2D tensors to capture both intra-period and inter-period variations using convolutional kernels. Multi-scale temporal modeling helps diagnose whether rigidity originates from a lack of multi-scale temporal processing.
- **FEDformer** (Zhou et al., 2022): A frequency-enhanced Transformer that utilizes attention with Fourier and wavelet transforms to capture long-range dependencies. The frequency-domain perspective offers a distinct representation space for testing sensitivity to structural interventions.
- **DLinear** (Zeng et al., 2023): A simple yet effective linear model that decomposes time series into trend and seasonal components for prediction. As a non-attention, non-graph baseline, this model sets a lower bound on adaptability and reveals whether complex architectures introduce additional rigidity.
- **Autoformer** (Wu et al., 2021): An auto-correlation Transformer that leverages an auto-correlation mechanism instead of standard self-attention to discover period-based dependencies. The decomposition architecture progressively separates trend and seasonal effects, providing yet another lens on representation dynamics.

FAMILY V - FOUNDATION MODELS

- **Chronos** (Ansari et al., 2024): A time series foundation model that tokenizes time series into a language-like representation and trains a large language model backbone for probabilistic prediction. Pretraining on massive heterogeneous time series allows evaluation of whether scale alone can compensate for the lack of structural priors.
- **TimesFM** (Das et al., 2024): A decoder-only foundation model pretrained on a vast corpus of real and synthetic time series. Its pretrained forecasting capability is evaluated under the prespecified benchmark interface to measure representation response under topological interventions.
- **Moirai** (Woo et al., 2024): A unified transformer model designed to handle arbitrary time series forecasting tasks with varying frequencies and dimensions. The architecture permits investigation of whether multi-task pretraining induces more adaptive or more rigid representations.
- **Lag-Llama** (Rasul et al., 2023): A decoder-only probabilistic foundation model pretrained on univariate time series using lagged covariates. Examining its representation response to structural changes reveals whether language model priors hinder or help representation-response generalization.

The five required model families span graph-based spatio-temporal predictors, robust/structure-aware architectures, structural-dynamics models, generic time-series forecasters, and pretrained foundation models. Their fixed benchmark configurations define the evidence set used by H1 and the response-landscape analyses.

A.5 TRAINING DETAILS

Training protocol.

- Optimizer: Adam.
- Learning rate: selected per model from $\{10^{-3}, 10^{-4}\}$.

- Batch size: 32.
- Maximum epochs: 200.
- Early stopping patience: 10 epochs on validation loss.
- Input horizon: 12 steps; output horizon: 12 steps.
- Evaluation metric: MAE.
- Models trained from scratch use MAE as the optimization objective unless their original implementation requires a probabilistic likelihood objective.
- For probabilistic foundation models, point forecasts are obtained from the predictive median using each model's native inference interface. When sampling is required, 100 samples are used; deterministic or quantile-output models use their native median or point forecast without artificial Monte Carlo sampling.
- Random seeds: 0-9 for model training.
- Each model is first configured according to its official implementation. Learning rate is then selected from the common grid when both values are compatible with the model. Foundation models use the frozen benchmark configuration; no additional adaptation sweep enters the reported analysis.

Representation extraction: the layer immediately preceding the final prediction head is used. Rows are aligned across paired representations. No learned pooling is introduced; a fixed flattening or average pooling is applied if needed. The pre-head interface is primary; layer sensitivity is optional and is reported only when cached intermediate representations are available.

A.5.1 DSAI Track B Encoding and Interpretation

The Dual-Track Structural Access Interface (DSAI) standardizes the source and intervention status of structural information. Graph-native models receive A_k through native operators; topology-agnostic models receive $e(A_k) \in \mathbb{R}^{N\times 64}$ from a shared low-capacity encoder. One encoder is trained per dataset on training-only baseline and $\mathcal{I}_{\text{aug}}$ topologies with an adjacency-autoencoding objective, then frozen and shared across architectures. A model-specific linear projection maps $e(A_k)$ to the model dimension and is trained jointly with the forecasting model on $\mathcal{D}_{\text{train}}$, then frozen. Encoder params: 50K; projection params: $\leq 5K$.

Interpretation of the two tracks. For models that do not natively consume topology, DMIS in Track A is zero by construction. Track A measures system-level structural responsiveness; Track B is the primary track for cross-architecture representation-response diagnosis. In Track B, DMIS is computed from the post-fusion representation immediately preceding the prediction head. The adapter-only quantity DMIS_e is an interface diagnostic and is not used as a reported endpoint.

Capacity-matched control. For every topology-agnostic architecture, a control with the same projection layer and parameter count but a constant zero structural embedding is included.

The two tracks expose, rather than fully eliminate, confounding due to unequal structural-information access: Track A measures native-system behavior, whereas Track B evaluates modified systems under a shared topology interface. Track B is therefore a controlled comparison, not an unbiased correction of the original architectures, and the two tracks are never pooled into one ranking.

A.6 Proofs of Supported Properties

A.6.1 Measurement Impossibility of Representation Response from Prediction Error Alone

Proof of Proposition 1. Let $H(X) \in \mathbb{R}^{n\times d}$, $d \geq 2$, be a nonzero centered representation and let the common scalar prediction head be $g(Z) = Zw$ for some nonzero $w \in \mathbb{R}^d$. Choose a unit vector $v \in \mathbb{R}^d$ with $v^\top w = 0$, and a centered topology-dependent vector $u(A) \in \mathbb{R}^n$ satisfying $u(A_k) \neq u(A_0)$. For any nontrivial T_k, define the same-dimensional representations

$$z_1(X, A) = H(X), \qquad z_2(X, A) = H(X) + \epsilon\, u(A)v^\top.$$

Because $v^\top w = 0$, both models use the same head and have identical predictions, hence identical IPE_k. Yet z_1 is topology-invariant while the centered response of z_2 changes by $\epsilon[u(A_k) - u(A_0)]v^\top$, which is nonzero. Therefore

$$\mathrm{DMIS}_k(f_1) = 0, \qquad \mathrm{DMIS}_k(f_2) > 0.$$

The example uses neither an appended unused channel nor a different readout: it exploits a null direction already present in a fixed-dimensional interface. Thus output error alone does not determine response at that declared interface. □

Scope of the construction. The construction is a measurement-impossibility example, not a general theorem of causal-mechanism non-identifiability. It shows that selected-interface response can vary inside a prediction-head null direction without changing outputs. Consequently, neither high nor low DMIS alone establishes causal influence. The result only motivates observing response in addition to error; it does not claim that DMIS and RD uniquely or completely identify an underlying mechanism.

A.6.2 Supported Invariances of DMIS

For numerical completeness, $\cos_F$ is the standard Frobenius cosine when both centered norms exceed ϵ_Z, is set to 1 when both are below it, and to 0 when only one is. These conventions only return finite values; all such pairs are flagged as degenerate and excluded from primary-landscape claims.

Proof. Let $\bar{Z}_k^{(0)}, \bar{Z}_k^{(T)}$ be a nondegenerate centered pair. For a shared orthogonal transformation Q, $Q^\top Q = I_d$, both Frobenius inner products and norms are preserved; hence the cosine and magnitude terms, and therefore DMIS, are unchanged. For a shared positive scaling $c > 0$,

$$\bar{Z}_k^{(0)\prime} = c\bar{Z}_k^{(0)}, \qquad \bar{Z}_k^{(T)\prime} = c\bar{Z}_k^{(T)}.$$

The cosine term is unchanged. If the magnitude stabilizer is inactive before and after scaling,

$$\|\bar{Z}_k^{(0)}\|_F > \epsilon_Z, \qquad c\|\bar{Z}_k^{(0)}\|_F > \epsilon_Z,$$

then its numerator and denominator both scale by c, so the magnitude term is also unchanged. Centering removes shared translations. Thus DMIS is invariant under shared orthogonal transformations and is invariant under shared positive scaling whenever the numerical stabilizer remains inactive. Exact scale invariance is not claimed for near-zero pairs; these are flagged as degenerate and analyzed separately. □

A.6.3 Necessity of Observing Both Response and Error for the Proposed Taxonomy

The construction proves that error does not determine selected-interface response. Conversely, response alone cannot distinguish low- from high-degradation cases. Thus observing both coordinates is necessary for the predefined taxonomy, although they are neither claimed sufficient nor uniquely minimal for arbitrary mechanism descriptions.

A.7 Additional Experimental Results

This appendix follows the evidence roles fixed in Section 4. H1 and H2 use the controlled synthetic setting. The abstract SUMO scenario is used only for DMIS/RD, model-rank, and operational-regime replication. Observational datasets and I-405 event metadata are excluded from the reported response-failure analyses.

A.7.1 H1 Protocol and Specification Audit

For prediction unit i, let $r_{ij} = \mathrm{RD}_{ij}^{\mathrm{weak}}$ at normalized weak intensity s_{ij}, $j = 1, \ldots, J_i$. The prespecified error-history vector is

$$\mathcal{H}_{RD,i} = \left(\bar{r}_i, \widehat{\beta}_i, r_i^{\max}\right), \quad \bar{r}_i = \frac{1}{J_i}\sum_j r_{ij}, \quad \widehat{\beta}_i = \frac{\sum_j (s_{ij} - \bar{s}_i)(r_{ij} - \bar{r}_i)}{\sum_j (s_{ij} - \bar{s}_i)^2}, \quad r_i^{\max} = \max_j r_{ij}.$$

H1 compares $\mathcal{M}_{\mathcal{C}+\mathcal{H}_{RD}}$ with $\mathcal{M}_{\mathcal{C}+\mathcal{H}_{RD}+\overline{\mathrm{DMIS}}^{\mathrm{cal,weak}}}$. Each model-intervention-family contributes one unit before pairing with its single held-out strong outcome. The outer split leaves one model family out; ridge regularization and all standardization are selected inside the outer training folds. Model and family identity indicators are excluded.

Table 9 gives the exact numerical values underlying Figure 4.

Table 9: Completed H1 weak-history/intensity specification audit. All rows use the same leave-one-model-family-out outer split in controlled synthetic dynamics.

Specification	Error history	+DMIS	Risk reduction
Primary	0.001288	0.001262	+0.000026
No slope	0.001360	0.001335	+0.000025
Mean only	0.001391	0.001438	-0.000047
Nominal intensity	0.001265	0.001214	+0.000050

As Table 9 shows, the primary point estimate is positive but small; removing slope preserves its direction, the weak-error mean alone reverses it, and nominal rather than normalized intensity restores a positive estimate. The reported aggregate is the prespecified leave-one-model-family-out macro endpoint; it should not be interpreted as evidence that the direction is uniform within every held-out family.

A.7.2 Completed SUMO Rank and Regime Replication

The SUMO component uses the abstract connected-subgraph construction in Appendix A.2 and supports the replication results in Section 5.3. It does not enter H1 or H2. Model-rank correspondence is reported in Figure 5: DMIS gives Spearman $\rho = 0.696$ with $p < 0.001$, whereas RD gives weaker correspondence. The family-wise regime proportions underlying Figure 6 are listed in Table 10 for auditability.

Table 10: Family-wise operational-regime proportions in the abstract SUMO replication (%).

Model family	Stable	Adaptive	Pathological rigidity	Over-reactive
Family I	11.67	36.00	7.33	45.00
Family II	22.78	28.33	12.78	36.11
Family III	7.22	32.78	0.56	59.44
Family IV	35.00	9.44	41.94	13.61
Family V	37.92	0.00	55.83	6.25

The exact percentages in Table 10 are the numerical basis for the family-level pattern summarized by Figure 6; they remain descriptive simulator results.

A.7.3 Regime-Assignment Sensitivity: Section 5.4

Section 5.4 reports the held-out summary; Table 11 gives the full local-threshold breakdown.

This analysis evaluates stability of the four operational labels without retraining forecasting models. The primary thresholds are calibration medians, $q = 0.50$, estimated separately by setting and intervention family. Local sensitivity uses $q \in \{0.40, 0.60\}$. A four-component GMM is fitted only to calibration $\mathrm{DMIS}, \mathrm{RD}$ pairs and then used to assign held-out units. Adjusted Rand index (ARI) and normalized mutual information (NMI), both permutation-invariant, quantify agreement with the primary threshold partition; no component-to-regime label matching or silhouette analysis is required.

As Table 11 shows, local quantile perturbations preserve 81.0%-81.5% of labels, whereas the GMM recovers the threshold partition only weakly in controlled synthetic dynamics and moderately in SUMO. H1 is not refit in this experiment because H1 uses continuous DMIS rather than regime labels. The two diagnostics answer different questions: the former measures local robustness of the declared thresholds, while the latter tests whether the same geometry emerges from an unconstrained four-component partition. Their divergence therefore supports an operational, threshold-calibrated interpretation of the four regimes rather than a latent-cluster claim.

Table 11: Regime-assignment sensitivity for Section 5.4. Thresholds/GMM are fitted on calibration data and evaluated on held-out units.

Setting	GMM ARI	GMM NMI	Agreement at $q = 0.40$	Agreement at $q = 0.60$
Controlled synthetic	0.135	0.253	0.810	0.815
Abstract SUMO replication	0.346	0.540	0.810	0.810

A.7.4 Weak-Error Comparability Audit: Section 5.5

Pairwise comparison of the 21 benchmark models yields $\binom{21}{2} = 210$ unordered candidate pairs. Eligibility is defined by the prespecified weak-intervention relative-degradation criterion $|\mathrm{RD}_i^{\mathrm{weak}} - \mathrm{RD}_j^{\mathrm{weak}}| \leq \eta_m$, with $\eta_m = 0.05$. Table 12 records both cohort construction and the outcome summary used in Section 5.5.

Table 12: Weak-error comparability audit and outcome summary.

Item	Value
Number of models	21
Candidate model pairs	210
Prespecified tolerance	$\eta_m = 0.05$
Eligible pairs	210
Aggregate weak-RD gap	0.000155
Strong RD, higher DMIS	0.000346
Strong RD, lower DMIS	0.000561
Higher-minus-lower difference	-0.000215

Table 12 shows that all 210 candidate pairs satisfy the criterion, so the eligibility rate is 100% and $\eta_m = 0.05$ is non-selective for this cohort. The analysis is therefore interpreted as a descriptive weak-error comparability audit, not as evidence created by selective matching. Because a model can participate in multiple unordered pairs, the 210 pairwise comparisons are not treated as 210 independent inferential units.

A.7.5 Alternative Representation-Response Measures: Section 5.6

Table 13 reports the alternative-coordinate audit under the frozen H1 protocol. CKA, SVCCA, RSA, and PWCCA are computed from the same row-aligned representation pairs and evaluated using the same H1 outer folds, weak-to-strong target, controls, and ridge-selection procedure. The purpose is directional robustness of incremental value and response ordering, not selection of the most favorable metric.

Table 13: Alternative response measures under the frozen H1 folds.

Response measure	Rank agreement with DMIS	H1 risk reduction	Direction vs. primary
$1-$CKA	0.923	0.000041	Same
$1-$SVCCA	0.964	0.000143	Same
RSA distance	0.976	-0.000019	Different
$1-$PWCCA	0.970	0.000121	Same

As shown in Table 13, all four alternative measures remain strongly aligned with DMIS in response ordering, with Spearman correlations between 0.923 and 0.976. Among them, CKA, SVCCA, and PWCCA retain the positive H1 risk-reduction direction, while RSA shows similarly strong ordering agreement but a slightly negative incremental effect. These results indicate that the response ordering identified by DMIS is stable across substantially different representation-similarity measures, whereas the incremental weak-to-strong predictive contribution shows some dependence on the chosen response coordinate.

A.7.6 Invariance Verification: Section 5.7

Table 14 reports the invariance audit corresponding to Section 5.7. We evaluate 100 nondegenerate representation pairs for each transformation, using the frozen H1 representations and the operational-

regime thresholds fixed in the preceding analysis. The evaluated pairs cover all 21 model architectures and the three intervention families. Shared rotation and feature permutation are applied jointly to the baseline and intervened representations, while positive scaling is evaluated only for admissible cases in which the numerical stabilizer remains inactive.

Table 14: DMIS invariance under supported shared representation transformations.

Check	Valid Instances	Mean DMIS Deviation	Maximum DMIS Deviation	Spearman ρ	Regime Agreement
Shared Rotation	100	7.54×10^{-16}	4.897×10^{-15}	1.000	100.0%
Feature Permutation	100	8.67×10^{-16}	5.825×10^{-15}	1.000	100.0%
Admissible Positive Scaling	100	1.93×10^{-15}	4.258×10^{-14}	1.000	100.0%

As shown in Table 14, all three transformations induce only negligible numerical changes in DMIS, with maximum deviations on the order of 10^{-14} or smaller. Spearman rank agreement is 1.000 across all evaluated instances, and no operational regime label changes are observed. These results numerically verify that the implementation of DMIS is stable under the supported shared representation transformations.

A.7.7 DSAI Access-Sensitivity Audit

We evaluate the sensitivity of the diagnostic landscape to the declared structural-access protocol in controlled synthetic dynamics. Track A uses native model-specific structural access, whereas Track B uses the controlled benchmark interface. The two tracks are analyzed separately and are never pooled into a single ranking.

Table 15: DSAI Track A/B audit in controlled synthetic dynamics. Track A and Track B are reported separately and are never pooled.

Comparison	Track A	Track B	Cross-Track Summary
Model DMIS Ranking	Mean DMIS $= 0.225592$	Mean DMIS $= 0.876778$	Kendall $\tau_b = 0.391$
Operational-Regime Distribution	Stable 34.3%; Adaptive 16.8%; Over-Reactive 17.9%; Rigid 31.0%	Stable 17.2%; Adaptive 31.3%; Over-Reactive 25.6%; Rigid 26.0%	Agreement $= 66.1\%$

Table 15 shows substantial descriptive differences between the two structural-access settings. Mean DMIS increases from 0.225592 under Track A to 0.876778 under Track B, while the model-level rank correspondence remains moderate ($\tau_b = 0.391$). The regime composition also changes across tracks, with 66.1% of assignments preserved. These results motivate reporting the two access protocols separately rather than interpreting them as a single pooled diagnostic ranking.

A.7.8 H2 Information-Matched Steering: Section 5.8

Table 16 reports the H2 point-estimate audit. H2 is restricted to controlled synthetic dynamics and is intentionally separated from H1 and SUMO replication. The frozen comparison is Matched+DMIS versus Matched without DMIS. Both heads receive $Z_k^{(T)}$, RD, $D_{Z,k}$, $d_{\mathcal{T},k}$, and the source variables used to compute them; only the former receives the explicit DMIS coordinate. The baseline receives a constant-zero coordinate to preserve input dimensionality. All heads use identical depth, width, optimization, early stopping, and equal-norm perturbations. Shuffled-DMIS, gradient, and random directions are auxiliary controls rather than additional H2 contrasts.

Table 16: H2 steering results in controlled synthetic dynamics.

Direction	Error reduction (%)	Difference vs. matched baseline	Role
Matched without DMIS	0.058224	-	Prespecified H2 baseline
Matched+DMIS	0.100722	+0.042498	Prespecified H2 condition
Shuffled DMIS	0.090435	+0.032211	Placebo
Gradient	3.147252	+3.089028	Auxiliary
Random	-0.006971	-0.065195	Auxiliary

As shown in Table 16, Matched+DMIS increases the point-estimated error reduction from 0.0582% to 0.1007%, a 0.0425 percentage-point advantage over the information-matched baseline. It exceeds the shuffled-DMIS placebo by only 0.0103 percentage points, so the completed H2 point-estimate audit is interpreted as a modest, configuration-specific parameterization effect rather than a large or

uncertainty-resolved steering gain. The gradient direction produces a substantially larger reduction but uses post-intervention label information and is therefore only a label-oracle auxiliary reference, while the random direction slightly degrades performance.

A.7.9 Mixing-Weight Sensitivity: Section 5.9

We evaluate the sensitivity of DMIS to the angular–magnitude mixing weight using $\alpha \in \{0.25, 0.50, 0.75\}$, with $\alpha = 0.50$ retained as the primary setting. DMIS is recomputed from the saved representation responses without retraining or rerunning the forecasting models. For Controlled Synthetic, we examine model-level rank agreement and the direction of the frozen H1 point estimate. For Abstract SUMO, we evaluate model-rank agreement and operational-regime stability.

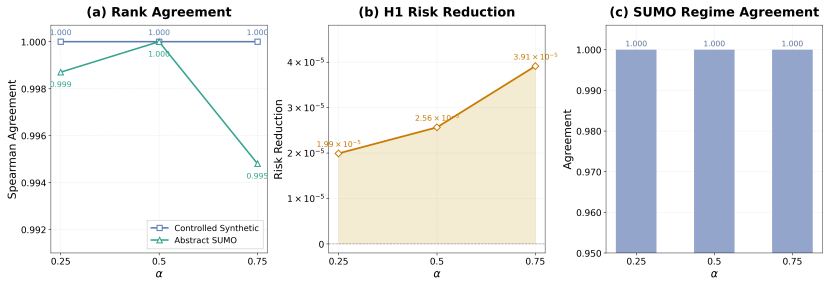

Figure 8: The sensitivity of DMIS mixing weight α. (a) Spearman agreement of model-level DMIS rankings relative to the primary $\alpha = 0.50$ setting in Controlled Synthetic and Abstract SUMO. (b) H1 risk reduction in Controlled Synthetic across the tested mixing weights. (c) Agreement of Abstract SUMO operational-regime assignments with the primary $\alpha = 0.50$ configuration.

As shown in Figure 8(a), the Controlled Synthetic ranking is unchanged across all tested values of α, while Abstract SUMO remains highly consistent with the reference ranking, with Spearman agreement above 0.994. Figure 8(b) shows that the H1 risk-reduction point estimate remains positive for all three mixing weights, although its magnitude varies with α. Finally, Figure 8(c) shows 100% agreement in the SUMO operational-regime assignments across all tested settings. Together, these results indicate that the main qualitative conclusions are robust to the choice of α within the evaluated range.

A.8 Reproducibility Checklist

All forecasting experiments use the fixed seeds declared in the released configuration. H2 uses model seeds $\{0, 1, 2, 3, 4\}$ with the prespecified intervention seeds in controlled synthetic dynamics.

Released audit artifacts. Table 17 summarizes the unit-level artifacts needed to reconstruct the primary H1 and prespecified H2 analyses. The release includes a run manifest with inclusion/exclusion reasons, seeds, and checkpoint hashes; an H1 synthetic prediction-unit table with weak-error histories, calibrated/raw response summaries, controls, and strong targets; an H2 synthetic paired steering table with placebo assignments and before/after error; and a calibration manifest recording thresholds and magnitude normalizers.

Table 17: Released audit artifacts used to reconstruct the primary H1 and prespecified H2 analyses. Each artifact is versioned and keyed to the run manifest.

Artifact	Primary unit	Core fields	Release role
Run manifest	run	model, setting, seed, checkpoint hash, status	Inclusion and provenance
H1 unit table	prediction unit	weak-RD summary, DMIS, controls, strong target	Primary H1 reconstruction
H2 paired table	steering pair	condition, placebo block, before/after IPE	H2 and placebo reconstruction
Calibration manifest	setting-family	thresholds, distance map, frozen hyperparameters	Frozen evaluation settings

The artifact roles in Table 17 are keyed to the run manifest so that every reported aggregate can be regenerated from stable identifiers rather than local filenames.

Audit consistency checks. Every H1 and H2 row carries stable model, intervention-family, intensity, and seed identifiers that link back to the run manifest. An automated audit verifies that strong-test rows never appear in calibration files, that no excluded run contributes to a reported aggregate, that each H1 prediction unit has exactly one strong target, and that every reported table entry can be regenerated from the released unit-level files. File checksums and the configuration commit hash are stored with the manifest so that regenerated summaries can be compared against the submitted version without relying on local filenames or directory structure.