# SPATIAL HETEROGENEITY THAT MATTERS: AN APPLICATION-ORIENTED BENCHMARK FOR MOBILE TRAFFIC MODELING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Mobile traffic modeling is important for understanding spatial urban functions and corresponding human activity patterns, particularly the spatial heterogeneity. But it can be ignored by evaluations focusing on citywide error-based metrics aggregated across spatial and temporal dimensions, and capturing aggregate scale alone does not establish usefulness for region-level network management and planning within cities. To support these practical applications, we introduce NETSPACE-BENCH, an application-oriented benchmark that identifies modeling weaknesses of existing methods and guides the further direction by jointly evaluating modeling quality, especially the spatial context-related heterogeneity and decision utility. We achieve these through spatial reference and dispersion checks, five complementary evaluation perspectives, and five operational decision tasks. Specifically, our evaluation across 12 models and three cities reveals that 18 of 24 eligible main generation runs retain less than 5% of observed region-mean dispersion and 19 yield energy policies approximately no better than inaction. These findings motivate explicit evaluation of spatial heterogeneity for practical applications, and the operational value of controllability through spatial context remains an open question.

## 1 INTRODUCTION

Human activities are spatially heterogeneous because different areas within a city serve different functions (Yuan et al., 2012; Louail et al., 2014; González et al., 2008; Rong et al., 2026). For example, people live in residential areas, seek entertainment and shop in leisure districts, and work in employment areas (Yuan et al., 2012; Louail et al., 2014). These different activities generate distinct demands for mobile network services (Zhang et al., 2026b; Chai et al., 2024b; Mishra et al., 2022). Understanding how the spatiotemporal variation in these demands relates to urban regional functions can help mobile service providers optimize resource allocation, improve service quality, and reduce cost (Marquez et al., 2017; Zhang et al., 2023; 2026c).

Existing works use multimodal urban contextual information, including socioeconomics, satellite imagery, points of interest (POIs), functional-region semantics, and spatial relations, to profile urban regions and guide traffic generation or prediction (Xu et al., 2021; Chai et al., 2025a;b; Zhang et al., 2025; Wang et al., 2026; Zhang et al., 2026a). These models are typically evaluated by comparing their outputs with observed data and aggregating errors across regions and time into city-level scores (Chai et al., 2024a; 2025a; Wen et al., 2023; Gneiting, 2011). However, when training and evaluation are driven primarily by average error, models tend to prioritize prevalent patterns over the distinctive demands of a small number of critical regions, risking mode collapse in generative modeling (Srivastava et al., 2017; Metz et al., 2016; Salimans et al., 2016; Kong et al., 2024; Fan et al., 2026). In mobile traffic modeling, such a failure can manifest spatially as overly homogenized demand patterns across regions with different urban functions, which we refer to as spatial mode collapse (Yang et al., 2021; Steininger et al., 2021; Xu et al., 2016). Recent cellular traffic generation approaches increasingly adopt spatio-temporal diffusion models to better capture complex traffic distributions and spatio-temporal dependencies (Liu et al., 2025; Chai et al., 2025a). Nevertheless, although existing studies introduce spatial metrics such as Spatial RMSE (Qi et al., 2025b), these still aggregate errors in time-averaged regional traffic into an overall score, without directly revealing performance across different functional regions or whether their distinctive demand patterns are



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.

preserved (Kynkäänniemi et al., 2019; Naeem et al., 2020). This limitation has direct operational implications: a small number of critical regions, such as major commercial districts, may experience high traffic volumes and substantial demand fluctuations, so misrepresenting their patterns can lead to inefficient capacity provisioning, resource allocation, and congestion management (Chai et al., 2026b; 2025b; Wu et al., 2015; Bega et al., 2019; Panitsas et al., 2026; Wan et al., 2026). Meanwhile, attention to these critical regions should not come at the expense of modeling accuracy elsewhere; evaluation should account for both the demand characteristics and the operational importance of different regions (Elmachtoub & Grigas, 2022; Mandi et al., 2024; Maier-Hein et al., 2024). Evaluation must therefore go beyond overall fidelity to explicitly assess whether demand patterns associated with regional functions are preserved and whether preserving this spatial heterogeneity translates into practical value for region-level network management and planning (Liang et al., 2023; Feng et al., 2025). Table 1 summarizes which evaluation dimensions representative mobile traffic studies report.

To achieve this goal, we introduce NETSPACE-BENCH, an application-oriented benchmark for mobile traffic generation and forecasting that evaluates overall modeling quality, with a focus on whether model outputs preserve spatial heterogeneity and support practical decisions. **At the modeling level**, we distinguish the preservation of regional variation from the accuracy of spatial allocation. We introduce the Spatial Dispersion Ratio (SDR) to quantify how much of the observed relative variation in regional mean traffic is preserved in model outputs. Complementing SDR, $\mathrm{SAD}^{\star}$ calibrates spatial allocation error against a city-specific blindness floor: the score assigned to uniform regional shares, determined by the observed demand distribution and urban geography. This reference makes explicit how a model's spatial error compares with that of an output whose regional mean demand is entirely undifferentiated, rather than treating a small numerical error as sufficient evidence of preserved spatial structure. Together, the two measures assess allocation accuracy alongside the extent of regional homogenization. We further develop five complementary metrics, grounded in a decomposition of the mean traffic field into overall volume, temporal distribution, spatial distribution, and region-time interactions, with a fifth evaluation perspective capturing fluctuations across time windows. This structure helps distinguish different sources of mismatch that an overall error score can obscure. **To assess practical value**, we use five downstream tasks—capacity planning, energy saving, resource allocation, congestion warning, and training with synthetic data—with clearly defined losses and reference policies to measure the benefits and costs of using models.

Overall, we benchmark 12 mobile traffic models: eight for generation and four for forecasting. These include statistical and spatial baselines, GAN-based models, diffusion models, and Transformer-based models. Each model is evaluated in three cities: New York City (NYC), Washington, D.C. (DC), and Baltimore (BM), yielding 24 evaluations for generation and 12 for forecasting. In 18 of the 24 generation evaluations, the generated traffic retains less than 5% of the observed relative variation in regional mean traffic, indicating substantial spatial homogenization. In contrast, the four forecasting models preserve substantially more regional variation, with SDR across cities ranging from 0.87 to 1.01. Across the 24 generation evaluations, the conventional metrics examined show only weak rank correlations with SDR ($|\rho| \leq 0.23$), indicating that their scores provide limited insight into the loss of regional variation. The downstream tasks reveal further limitations: in 19 of the 24 generation evaluations, the resulting energy-saving policies perform approximately no better than keeping all cells active. Model rankings also differ across accuracy metrics and operational tasks, showing that lower aggregate error does not consistently correspond to better decisions. Together, these findings highlight the need to evaluate both the preservation of spatial heterogeneity and task-specific decision value, rather than relying on overall accuracy alone.

In a nutshell, the contributions of this work can be summarized as follows.

- We introduce NETSPACE-BENCH, an application-oriented benchmark for mobile traffic generation and forecasting that evaluates spatial heterogeneity alongside overall modeling quality. It further assesses practical value through five downstream tasks.
- We systematically evaluate eight generation models and four forecasting models. Evaluating each model across three major cities in the U.S., allowing us to compare their strengths and limitations across modeling quality and downstream performance.
- Our findings reveal that, although many evaluated models incorporate spatial context, substantial gaps remain in preserving regional heterogeneity and supporting practical decisions. These findings highlight the need to improve spatial fidelity in terms of regional distinctness and task-specific utility beyond only aggregate accuracy.

# 2 PRELIMINARIES

## 2.1 NOTATIONS AND PROBLEM DEFINITIONS

In this section, we introduce the problem formulation of mobile traffic modeling, from both generation and prediction perspectives.

**Definition 1** (Urban Regions)**.** A city is represented by a set of $R$ regions, $\mathcal{R} = \{r_1, \ldots, r_R\}$. Each region is a spatial unit at which mobile network traffic is recorded.

**Definition 2** (Regional Spatial Context)**.** Regional spatial context, denoted by $\mathcal{C}$, describes the characteristics of urban regions and the relations among them. In this work, the available context includes geographic coordinates, population distributions, points of interest, and a regional knowledge graph with pretrained embeddings, which is described in Appendix A.2.

**Definition 3** (Mobile Traffic)**.** Let $X[w, r, t] \geq 0$ denote the mobile traffic observed in region $r$ at hour $t$ of time window $w$. Traffic observations over $W$ windows, $R$ regions, and $H$ hours per window form a tensor

$$X \in \mathbb{R}_{\geq 0}^{W \times R \times H}. \tag{1}$$

We write $X_w \in \mathbb{R}_{\geq 0}^{R \times H}$ for the region-time field within a single window and use $\widehat{X}$ to denote model outputs. The mean traffic field across windows is

$$\bar{X}[r, t] = \frac{1}{W} \sum_{w=1}^{W} X[w, r, t], \tag{2}$$

which describes the average traffic level in each region at each relative hour of the window.

PROBLEM 1. ***Mobile Traffic Generation.*** Given observed training traffic and regional spatial context $\mathcal{C}$, the task is to learn a conditional traffic distribution and generate region–time fields:

$$\widehat{X}_w \sim p_\theta^{\text{gen}}(\cdot \mid \mathcal{C}), \qquad \widehat{X}_w \in \mathbb{R}_{\geq 0}^{R \times H_{\text{gen}}}, \tag{3}$$

where $\theta$ denotes the model parameters and $H_{\text{gen}}$ is the generation window length. The goal is to reproduce the spatial and temporal patterns of observed traffic, including differences across regions and variation across windows.

PROBLEM 2. ***Mobile Traffic Prediction.*** Given an observed traffic history $X_w^{\text{hist}} \in \mathbb{R}_{\geq 0}^{R \times H_{\text{in}}}$ and available regional spatial context $\mathcal{C}$, the task is to predict traffic over the following $H_{\text{pred}}$ hours. The prediction problem can be expressed through the conditional distribution

$$p_\theta^{\text{pred}}\big(X_w \mid X_w^{\text{hist}}, \mathcal{C}\big), \qquad X_w \in \mathbb{R}_{\geq 0}^{R \times H_{\text{pred}}}, \tag{4}$$

where $H_{\text{in}}$ is the history length. A model may provide a point forecast or a predictive distribution. Each forecast is associated with a specific observed future window, allowing its regional and temporal predictions to be compared with the corresponding observations.

## 2.2 EXISTING EVALUATION AND BENCHMARKS

Mobile traffic modeling has moved from task-specific forecasting to shared models for generation and prediction (Chai et al., 2025b; Qi et al., 2025a), and spatiotemporal benchmarks have broadened data coverage and evaluation scope (Tan et al., 2023; Liu et al., 2023; Ma et al., 2025; Martin et al., 2026). However, none of them evaluates whether model outputs preserve the regional heterogeneity induced by urban context. Urban representation learning characterizes regional differences from multimodal data and spatial relations, and its benchmarks test the learned representations through downstream prediction and spatial generalization (Hao et al., 2025; Sun et al., 2026; Wu et al., 2024; Moska et al., 2025; Liu et al., 2026). These benchmarks, however, evaluate the representations rather than the mobile traffic generated or predicted from them. NETSPACE-BENCH complements both lines by directly assessing whether regional demand differences are preserved in model outputs; Appendix A.1 gives an extended discussion.

# 3 METHODOLOGY

In this section, we present the three components of NETSPACE-BENCH (Figure 1): spatial heterogeneity evaluation (Section 3.1), multi-perspective quality evaluation (Section 3.2), and downstream utility evaluation (Section 3.3), which measures the benefits and costs of model outputs through explicit losses and task-specific references (Gneiting & Raftery, 2007; Elmachtoub & Grigas, 2022).

Table 1: A comparison of evaluation scope in representative mobile traffic works.

| Work | Fidelity | | Structure | | | Operational value | Spatial audit |
|---|---|---|---|---|---|---|---|
| | Numerical agreement | Distributional agreement | Temporal profile | Spatial allocation | Cross-window variation† | Decision utility | Regional collapse† |
| OpenDiff (Chai et al., 2024a) | ✓ | ✓ | | | | ✓ | |
| KG-Diff (Chai et al., 2024b) | ✓ | ✓ | ✓ | | | | |
| CANDLE (Qi et al., 2024) | ✓ | ✓ | | | | ✓ | |
| STK-Diff (Chai et al., 2025a) | ✓ | ✓ | ✓ | ✓ | | ✓ | |
| UoMo (Chai et al., 2025b) | ✓ | ✓ | | | | ✓ | |
| MobiGPT (Qi et al., 2025a) | ✓ | ✓ | | | | | |
| MCDiff (Chai et al., 2026b) | ✓ | ✓ | | ✓ | | ✓ | |
| ZoomDiff (Qi et al., 2025b) | ✓ | | | ✓ | | | |
| MobiFM (Chai et al., 2026a) | ✓ | | | | | ✓ | |
| NETSPACE-BENCH | ✓ LBE | ✓ distribution | ✓ TPD | ✓ SAD and CLE | ✓ DLD | ✓ five tasks | ✓ SDR and floor |

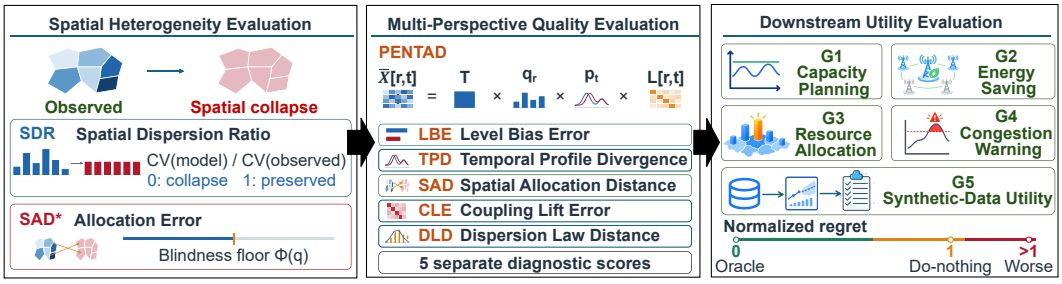


Figure 1: An overview of the framework of our NetSpace-Bench.

### 3.1 SPATIAL HETEROGENEITY EVALUATION

Preserving regional variation and accurately allocating traffic across regions are related but distinct requirements: a model may reproduce the amount of regional variation while placing high demand in the wrong regions. We therefore assess the two properties separately and interpret allocation error against a city-specific reference.

**Preservation of regional variation.** Let $\mu_r = H^{-1}\sum_{t=1}^{H}\bar{X}[r,t]$ be the observed mean traffic of region $r$, and $\widehat{\mu}_r$ its counterpart computed from the mean field of model outputs. For fields with positive total traffic, the *Spatial Dispersion Ratio* (SDR) is

$$\mathrm{SDR} = \frac{\mathrm{CV}_r(\widehat{\mu}_r)}{\mathrm{CV}_r(\mu_r)}, \qquad \mathrm{CV}_r(z_r) = \frac{\mathrm{sd}_r(z_r)}{|\,\mathrm{mean}_r(z_r)|}, \tag{5}$$

where the standard deviation and mean are taken across regions, which removes the effect of overall traffic scale. SDR near zero indicates nearly uniform regional means, and SDR near one a similar degree of relative variation to the observations; values above one indicate greater variation, not necessarily better accuracy (Willmott & Matsuura, 2006; Hodson, 2022). SDR does not establish whether the remaining differences are assigned to the correct regions.

**Accuracy of spatial allocation.** To evaluate where traffic is allocated independently of its volume, we normalize regional means into shares $q_r = \mu_r / \sum_j \mu_j$ and $\widehat{q}_r = \widehat{\mu}_r / \sum_j \widehat{\mu}_j$ and measure their 1-Wasserstein distance (Villani et al., 2009; Rubner et al., 2000),

$$\mathrm{SAD}(\widehat{X}, X) = W_1(\widehat{q}, q; D) = \min_{\pi \in \Pi(\widehat{q},q)} \sum_{r,r'} \pi_{rr'} D_{rr'}, \tag{6}$$

where $D_{rr'}$ is the haversine distance between regions $r$ and $r'$, and $\Pi(\widehat{q}, q)$ contains nonnegative transport plans with marginals $\widehat{q}$ and $q$. SAD is computed by linear programming and reported in kilometers, so allocating demand to a nearby region costs less than allocating it to a distant one.

**A city-specific reference for spatial error.** The magnitude of SAD depends on both the observed demand distribution and the city's geography. To make it interpretable, consider a spatially blind output $\widehat{X}[w,r,t] = a[w,t]$, which assigns the same time series to every region. Its regional shares are necessarily uniform, $u = (1/R, \ldots, 1/R)$, giving

$$\mathrm{SAD}(\widehat{X}, X) = W_1(u, q; D) =: \Phi(q). \tag{7}$$

All such outputs therefore receive the same spatial score, regardless of their temporal pattern or overall volume; the same fixed-reference property holds for any share-based discrepancy $d$ with finite $d(u, q)$. We call $\Phi(q)$ the *blindness floor*: a reference for uniform allocation, not a lower bound on the error of all models. We compute it for each city and report the floor-normalized spatial error

$$\mathrm{SAD}^\star = \frac{\mathrm{SAD}(\widehat{X}, X)}{\Phi(q)}, \qquad \Phi(q) > 0, \tag{8}$$

where values below, equal to, or above one indicate lower, equal, or higher allocation error than uniform allocation; the normalization is dimensionless and preserves model rankings within each city. Appendix A.8 locates $\Phi(q)$ within the attainable error range. We report SDR and $\mathrm{SAD}^\star$ together rather than combining them: matching the uniform reference does not prove spatial blindness, just as matching observed dispersion does not prove accurate allocation, and both concern aggregate regional demand, whose region-specific time patterns are examined in Section 3.2.

### 3.2 Multi-Perspective Quality Evaluation

Beyond regional mean demand, mobile traffic also varies over time and across observation windows. We therefore develop PENTAD, which assesses traffic from five complementary perspectives organized by a decomposition of the mean field. For a field with positive total traffic, let $T = \sum_{r,t} \bar{X}[r, t]$ be its total traffic, $P[r, t] = \bar{X}[r, t]/T$ its joint region–time distribution, and $q_r = \sum_t P[r, t]$ and $p_t = \sum_r P[r, t]$ its spatial and temporal shares. On cells where $q_r p_t > 0$,

$$\bar{X}[r, t] = T\, q_r\, p_t\, L[r, t], \qquad L[r, t] = \frac{P[r, t]}{q_r p_t}, \tag{9}$$

where the lift $L[r, t]$ describes the region–time interaction and equals one when the joint share matches the product of the marginal shares. With $\widehat{T}$, $\widehat{P}$, $\widehat{q}$, and $\widehat{p}$ defined analogously for model outputs, the four factors motivate the first four perspectives, and the fifth examines the fluctuations across windows that the mean field does not retain (full definitions in Appendix A.6).

**Level Bias Error (LBE).** $\mathrm{LBE} = |\widehat{T} - T|/T$ captures over- or underestimation of total demand, independently of how it is distributed; we also retain its signed counterpart, since the two directions of error have different operational consequences.

**Temporal Profile Divergence (TPD).** TPD is the Jensen–Shannon divergence (base 2) between the temporal shares $p$ and $\widehat{p}$ (Lin, 1991). It evaluates when citywide demand occurs, independently of total volume and of spatial reallocations that preserve the citywide temporal profile.

**Spatial Allocation Distance (SAD).** SAD (Equation (6)) compares regional shares under geographic transport costs after removing differences in total volume; we report its normalized form $\mathrm{SAD}^\star$ (Equation (8)) together with SDR (Section 3.1).

**Coupling Lift Error (CLE).** Correct spatial and temporal shares do not guarantee correct regional time patterns: a model may preserve regional totals and the citywide temporal profile while placing a commercial district's peak at wrong hours. For strictly positive distributions, CLE is the total-variation distance $\frac{1}{2} \sum_{r,t} |S(\widehat{P})[r, t] - S(P)[r, t]|$, where $S(\cdot)$ is the Sinkhorn normalization that rescales rows and columns to uniform marginals (Cuturi, 2013; Sinkhorn & Knopp, 1967). Because $S$ is invariant to positive row and column scaling, CLE does not count such rescaling as an interaction error.

**Dispersion Law Distance (DLD).** The preceding perspectives describe the mean field but not how demand at the same region and relative hour varies across windows. DLD compares the observed and modeled distributions of these cellwise fluctuations, normalized by their mean scale, using the one-dimensional Wasserstein distance on sorted samples (Rubner et al., 2000); sorting makes it invariant to window permutations, as required when generated windows have no observed counterparts. DLD detects changes in fluctuation distributions, including loss of variability, but does not fully characterize dependence across regions.

Where defined, all five scores are zero for identical fields and increase with the discrepancy. We report them separately rather than as a composite; the decomposition motivates complementary perspectives without implying that five scalar scores fully describe the traffic field.

### 3.3 Downstream Utility Evaluation

Beyond modeling quality, we evaluate whether model outputs support network decisions (Elmachtoub & Grigas, 2022; Mandi et al., 2024). Each task applies a fixed decision rule to model outputs, selects actions on a decision block of test windows, and pays the task loss on a disjoint evaluation block; the same rule applied to observed traffic defines the reference policy. The five tasks cover complementary operational needs (details in Appendix A.13):

- **G1 Capacity planning.** Each region is provisioned at the $\alpha$-quantile of its modeled demand (Koenker & Bassett Jr, 1978), and the loss $\lambda_{\mathrm{u}}(y-c)_+ + \lambda_{\mathrm{o}}(c-y)_+$ charges unmet and unused capacity asymmetrically.
- **G2 Energy saving.** A region-hour is put to sleep when its normalized modeled load falls below the threshold implied by the energy-versus-SLA loss; keeping all cells active is the inaction baseline.
- **G3 Resource allocation.** In each hour, resources go to the $k$ regions with the highest modeled demand, and the loss is the peak demand left uncovered.
- **G4 Congestion warning.** An alarm is raised when modeled demand exceeds a high quantile of observed traffic; missed events are charged by severity and false alarms at a fixed cost, and never alarming is the inaction baseline.
- **G5 Synthetic-data utility.** A ridge next-day forecaster is trained on model outputs and tested on real traffic (TSTR) (Esteban et al., 2017); the reference is the same forecaster trained on real data.

We report each task's regret relative to its reference,

$$\mathrm{Regret}_j = \frac{C_j^{\mathrm{model}} - C_j^{\mathrm{ref}}}{Z_j}, \qquad Z_j = \begin{cases} |C_j^{\mathrm{ref}}|, & j \in \{1,3,5\}, \\ C_j^{\mathrm{base}} - C_j^{\mathrm{ref}}, & j \in \{2,4\}, \end{cases} \tag{10}$$

where energy saving and congestion warning use range normalization because their reference costs can change sign or be zero; for these two tasks, 0 matches the reference and 1 matches inaction. Lower is better, and the five regrets are reported separately.

## 4 Benchmark Evaluation

### 4.1 Experimental Setup

**Datasets.** We use hourly regional mobile traffic and the regional spatial context of Section 2.1 for New York City (NYC), Washington, D.C. (DC), and Baltimore (BM). Regions are split spatially into disjoint training, validation, and test sets, with 61, 59, and 210 test regions that are never seen during training (Appendix A.3). Generation samples are exchangeable 168-hour weekly windows, and forecasting samples are 84-hour horizons paired with their history.

**Models.** The eight generation models are the baselines SeasonalMean and GeoKNN (Kanevski et al., 2008), the recurrent GAN 5GT-GAN (Pandey et al., 2023), the diffusion models KG-Diff (Chai et al., 2024b), MCDiff (Chai et al., 2026b), STK-Diff (Chai et al., 2025a), and ZoomDiff (Qi et al., 2025b), and the diffusion-Transformer foundation model MobiGPT (Qi et al., 2025a); the four forecasting models are LastDay (Zhu & Laptev, 2017), DiffSTG (Wen et al., 2023), PatchTST (Nie et al., 2023), and MobiFM (Chai et al., 2026a). Each model is evaluated once per city, giving 24 generation and 12 forecasting evaluations. Implementations are labeled as adapted author-released code (AR), paper-based reproductions (PF), or benchmark-native baselines (BN). Five of the six learned generation models are PF, so the results characterize these implementations under our training setup rather than the best achievable performance of the original methods (Appendix A.4). One additional model, which copied target-region traffic into a conditioning channel, is excluded.

**Evaluation protocol.** Generation metrics compare the mean field of generated windows with that of observed test windows, and DLD compares sorted cross-window fluctuations; forecasting metrics use paired horizons, and the two tracks are never pooled. Each decision task selects actions on one block of test windows and measures cost on a disjoint block. For comparison, we also compute eight conventional metrics: MAE, RMSE, sMAPE, peak error, congestion miss cost, capacity-planning regret (as defined in prior work, distinct from G1), hotspot migration distance, and volatility error (Appendix A.7). Full details are given in Appendix A.5.

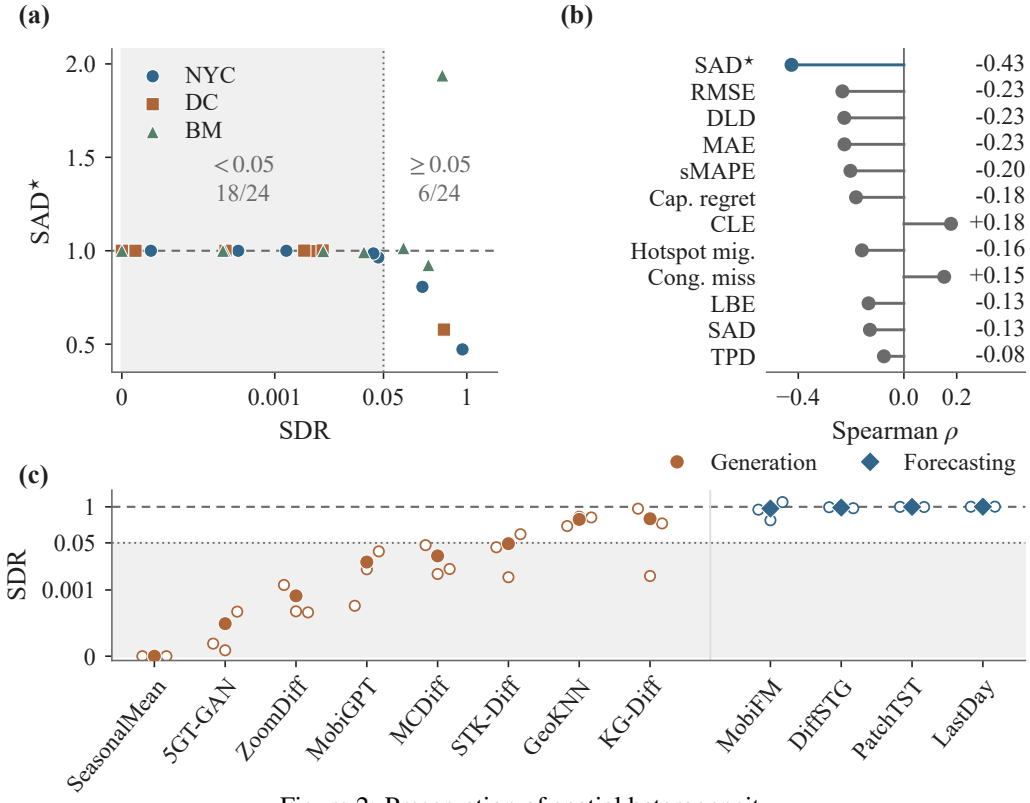


Figure 2: Preservation of spatial heterogeneity.

Table 2: Main results as three-city means per model.

| | | | | PENTAD (↓) | | | | | Decision regret (↓) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Src | RMSE | SDR | LBE | TPD | SAD⋆ | CLE | DLD | Cap. | Ener. | Alloc. | Warn. | TSTR |
| *Generation (168-hour weekly windows, exchangeable)* | | | | | | | | | | | | | |
| SeasonalMean | BN | 55.8 | 1.6e-16 | 0.16 | 0.010 | 1.000 | 0.129 | 0.477 | 3.01 | 1.013 | 0.629 | 0.965 | 2.32 |
| 5GT-GAN | AR | 1,534 | 6.1e-5 | 52.90 | 0.067 | 1.000 | 0.129 | 0.476 | 25.54 | 1.000 | 0.631 | 0.090 | 71.29 |
| ZoomDiff | PF | 884.7 | 6.1e-4 | 34.33 | 0.069 | 1.000 | 0.129 | 0.473 | 15.97 | 1.000 | 0.572 | 0.090 | 35.42 |
| MobiGPT | PF | 885.1 | 0.010 | 34.09 | 0.069 | 0.998 | 0.129 | 0.370 | 22.09 | 1.000 | 0.578 | 0.099 | 17.39 |
| MCDiff | PF | 787.3 | 0.017 | 30.59 | 0.067 | 0.987 | 0.132 | 0.327 | 35.49 | 1.000 | 0.555 | 0.253 | 38.13 |
| STK-Diff | PF | 1,210 | 0.047 | 47.73 | 0.069 | 1.000 | 0.129 | 0.439 | 24.48 | 1.000 | 0.574 | 0.090 | 36.03 |
| GeoKNN | BN | 72.8 | 0.352 | 0.79 | 0.005 | 1.108 | 0.127 | 0.477 | 2.51 | 0.904 | 0.461 | 0.597 | 0.54 |
| KG-Diff | PF | 3,022 | 0.370 | 123.38 | 0.056 | 0.798 | 0.129 | 0.263 | 141.58 | 0.828 | 0.439 | 0.300 | 146.40 |
| *Forecasting (84-hour paired horizons)* | | | | | | | | | | | | | |
| MobiFM | AR | 50.8 | 0.867 | 0.70 | 0.014 | 1.134 | 0.237 | 0.351 | 1.81 | 0.973 | 0.070 | 0.288 | 1.08 |
| DiffSTG | AR | 38.4 | 0.939 | 0.07 | 0.012 | 0.061 | 0.105 | 0.203 | 0.24 | 0.180 | 0.015 | 0.240 | 0.18 |
| PatchTST | AR | 27.7 | 0.995 | 0.01 | 0.002 | 0.040 | 0.091 | 0.213 | 0.13 | 0.058 | 0.007 | 0.119 | 0.15 |
| LastDay | BN | 44.1 | 1.011 | 0.04 | 9.3e-4 | 0.036 | 0.024 | 0.080 | 0.00 | −0.002 | −1.3e-4 | 0.173 | 0.28 |

### 4.2 Do models preserve spatial heterogeneity?

**Regional variation collapses in generation but is largely preserved in forecasting.** In 18 of the 24 generation evaluations, SDR is below 0.05: the output retains less than 5% of the observed relative dispersion of regional mean traffic. Six of the eight generation models have a three-city mean SDR below this threshold, and only GeoKNN and KG-Diff retain a substantial share of regional variation, with mean SDR of 0.352 and 0.370 (Table 2 and Figure 2c; details of both are given in Appendix A.15). In contrast, the four forecasting models have mean SDR between 0.867 and 1.011, and no forecasting evaluation falls below 0.05. Most of the collapsed generation models are conditioned on spatial context, yet in the present implementations this context leaves no visible difference in regional means. Per-city results are given in Appendix A.9, and sanity checks confirming that the collapse is not an evaluation artifact in Appendix A.14.

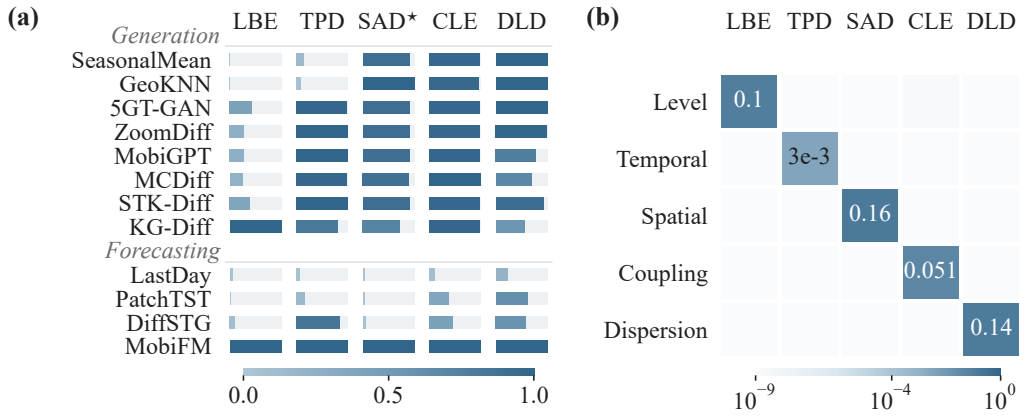


Figure 3: PENTAD scores of all models (a) and responses to RMSE-matched perturbations (b).

**Collapsed outputs sit at the uniform-allocation reference, which raw spatial error does not reveal.** The 18 low-SDR evaluations have $\mathrm{SAD}^{\star}$ between 0.964 and 1.002, within 3.6% of the uniform-allocation reference (Figure 2a). On the raw scale this is hard to see, because the reference occupies only 6.6%, 17.5%, and 24.7% of the attainable SAD range in BM, NYC, and DC (Appendix A.8), so an output with no regional differentiation looks acceptable. SDR and $\mathrm{SAD}^{\star}$ are also not interchangeable: an output can preserve dispersion while placing it in the wrong regions, or align with the observed pattern while retaining almost no dispersion (examples in Appendix A.9).

**Conventional metrics do not reveal the collapse.** Across the 24 generation evaluations, none of the conventional metrics in Figure 2b has a Spearman rank correlation (Spearman, 1961) with SDR exceeding 0.23 in absolute value, whereas normalizing by the city-specific reference raises the correlation of $\mathrm{SAD}^{\star}$ with SDR to $-0.43$. The most direct evidence is the uniform baseline itself: SeasonalMean, which assigns the same profile to every region, achieves the lowest generation RMSE in all three cities, ahead of every learned model. Selecting or tuning models by these metrics would therefore favor outputs that are homogeneous across regions; Figure 5 in Appendix A.7 illustrates why matched RMSE or JSD admits opposite operational outcomes.

### 4.3 Where does modeling quality break down?

**Generation errors concentrate in total volume.** The six learned generation models have LBE between 30.59 and 123.38, overestimating total traffic by 30 to 123 times, whereas the baselines SeasonalMean and GeoKNN have LBE of 0.16 and 0.79 (Table 2 and Figure 3a; details in Appendix A.15). TPD is small for all models, so the citywide temporal profile is learned. In contrast, $\mathrm{SAD}^{\star}$ and CLE hardly separate the models: $\mathrm{SAD}^{\star}$ stays near the uniform reference for the collapsed models (Section 4.2), and CLE lies between 0.127 and 0.132 for all eight generation models. DLD ranges from 0.263 for KG-Diff to 0.477 for the two deterministic baselines, which produce no cross-window fluctuation. In short, these models learn *when* traffic occurs but not *where* or *how much*, and no model leads on all five perspectives (Appendix A.14).

**In forecasting, paired error and mean-field structure rank models differently.** PatchTST has the lowest paired RMSE (27.7), yet LastDay, which repeats the previous day, has the lowest value on four of the five PENTAD perspectives, and MobiFM is the worst on all five despite an SDR of 0.867. A forecaster with low pointwise error can therefore still depart from the observed average structure; a per-model analysis is given in Appendix A.14.

**The five perspectives are complementary, but they do not describe the full field.** Across the 24 generation evaluations, the eight conventional metrics have an effective dimension of 2.21 out of 8, and MAE and RMSE rank the models identically, whereas the PENTAD perspectives reach 3.45 out of 5 (Appendix A.10). When perturbations of the observed field are calibrated to the same RMSE, each changes its targeted perspective by at least $3.0 \times 10^{-3}$ and the other four by at most $2.1 \times 10^{-9}$ (Figure 3b; Appendix A.11), so PENTAD separates failure modes that RMSE cannot. Matching the four mean-field perspectives is not sufficient for decisions: a control field that keeps them fixed but replaces the cross-window fluctuations of the 12 forecasting outputs yields median relative regret gaps of 52%, 46%, and 37% for TSTR, warning, and capacity planning (Appendix A.12).

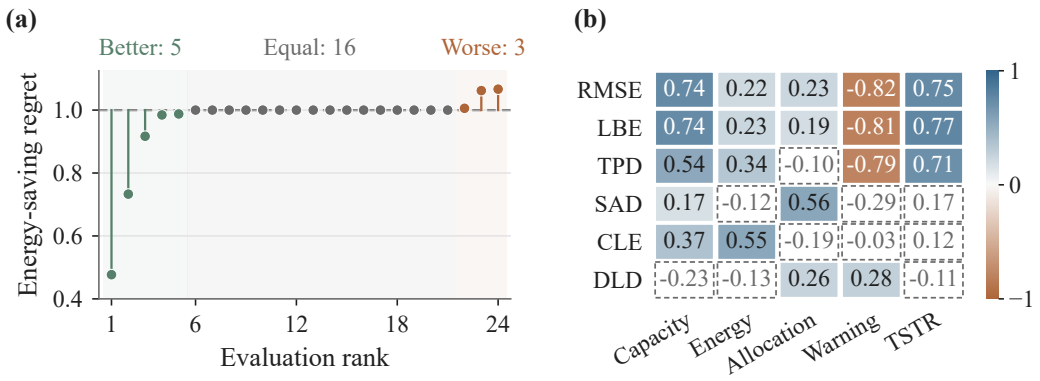


Figure 4: Decision utility of model outputs.

### 4.4 Do outputs support operational decisions?

**Most generation outputs yield decisions no better than inaction.** In 19 of the 24 generation evaluations, the energy-saving regret is at least 0.999, and five models have three-city means that round to 1.000: letting them decide which cells to put to sleep is almost no different from keeping every cell active, and SeasonalMean, at 1.013, is worse than inaction (Table 2, Figure 4a; details in Appendix A.15). Only KG-Diff (0.828) and GeoKNN (0.904), the two models with the highest SDR, are clearly below one; we report this association as suggestive, since the evaluation cannot separate the collapse from other differences between models. Capacity and TSTR regrets grow with the volume error: the two baselines with the smallest LBE have capacity regrets of 2.51 and 3.01, whereas the six learned models range from 15.97 to 141.58. Task-by-task results, including why models that overestimate traffic obtain low warning regret, are discussed in Appendix A.14.

**Forecasting outputs support most decisions, but preserving regional variation is not sufficient.** PatchTST and DiffSTG have regrets of at most 0.15 and 0.24 across the five tasks. MobiFM retains regional variation (SDR 0.867), yet its level bias of 0.70 brings its energy regret to 0.973, close to inaction, with capacity and TSTR regrets of 1.81 and 1.08 (Appendix A.14).

**Rankings change with the task, and which aspect of modeling quality matters depends on the decision.** Averaging within-city rank correlations over the three cities, RMSE correlates with capacity regret at $+0.74$ but with warning regret at $-0.82$ on the generation track, so improving a model by RMSE moves the two decisions in opposite directions (Figure 4b). Individual perspectives align with particular decisions: $\mathrm{SAD}^{\star}$ correlates with resource-allocation regret at $+0.56$ and CLE with energy regret at $+0.55$, against $+0.23$ and $+0.22$ for RMSE, and the five regrets have an effective dimension of 3.16 out of 5 (Appendix A.10). No single score orders all five decisions. These associations suggest a simple correspondence: capacity planning and TSTR are governed mainly by total volume, resource allocation by spatial shares, energy saving by the region–time interaction, and congestion warning by the direction of the volume error. A benchmark should therefore report the perspectives separately, and users should choose the perspective that matches the target decision.

## 5 Discussion and Conclusion

**Limitations.** First, the three cities share one upstream data source, which limits external validity. Second, each decision task is instantiated at one representative operating point (Table 12); a systematic sweep over operating points is left to future work, and the tasks remain offline proxies rather than decisions in a live network. Third, PENTAD summarizes selected discrepancies, and equal scores on the four mean-field perspectives do not imply equal distributions.

**Conclusion.** Existing evaluations of mobile traffic models reduce errors across regions and time to citywide scores, obscuring spatial heterogeneity and decision utility. NetSpace-Bench evaluates whether regional differences are preserved, where errors arise, and how well outputs support decisions. Most generation models produce nearly uniform regional patterns, conventional metrics fail to detect this collapse, and different decisions depend on distinct aspects of model quality, motivating benchmarks that go beyond aggregate accuracy to preserve spatial heterogeneity and decision utility.

### AI USE STATEMENT

We used large language models to assist with manuscript editing and translation, figure and table preparation, and the implementation and debugging of supporting code. We also used these tools to polish the language and improve the clarity and readability of the manuscript. The authors take full responsibility for the final text, code, analyses, and conclusions.

### REPRODUCIBILITY STATEMENT

We provide an anonymous code repository containing the evaluation implementation, experiment configurations, and scripts for reproducing the reported results: `https://anonymous.4open.science/r/netspace-bench-review-AE71/`. The processed datasets and archived model outputs are available through an anonymous read-only link: `https://osf.io/df2et/overview?view_only=0b57aabb09724f8ba8292fdb2a10f78c`. Dataset processing, model configurations, evaluation protocols, and decision-task definitions are documented in the appendix and the accompanying README files.

## REFERENCES

Ivana Balažević, Carl Allen, and Timothy Hospedales. Tucker: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 5185–5194, 2019.

Dario Bega, Marco Gramaglia, Marco Fiore, Albert Banchs, and Xavier Costa-Perez. Deepcog: Cognitive network management in sliced 5g networks with deep learning. In *IEEE INFOCOM 2019-IEEE conference on computer communications*, pp. 280–288. IEEE, 2019.

Haoye Chai, Tao Jiang, and Li Yu. Diffusion model-based mobile traffic generation with open data for network planning and optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4828–4838, New York, NY, USA, 2024a. Association for Computing Machinery. doi: 10.1145/3637528.3671544.

Haoye Chai, Tong Li, Fenyu Jiang, Shiyuan Zhang, and Yong Li. Knowledge guided conditional diffusion model for controllable mobile traffic generation. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 851–854, New York, NY, USA, 2024b. Association for Computing Machinery. doi: 10.1145/3589335.3651530.

Haoye Chai, Xiaoqian Qi, and Yong Li. Spatio-temporal knowledge driven diffusion model for mobile traffic generation. *IEEE Transactions on Mobile Computing*, 24(6):4939–4953, 2025a.

Haoye Chai, Shiyuan Zhang, Xiaoqian Qi, Baohua Qiu, and Yong Li. UoMo: A universal model of mobile traffic forecasting for wireless network optimization. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4308–4319, New York, NY, USA, 2025b. Association for Computing Machinery. doi: 10.1145/3711896.3737272.

Haoye Chai, Xiaoqian Qi, Yibo Ma, Zhaocheng Wang, Lei Yue, and Yong Li. MobiFM: A foundation model for mobile data forecasting. *IEEE Journal on Selected Areas in Communications*, 44, 2026a.

Haoye Chai, Baohua Qiu, Xiaobin Mo, Raoyuan Pan, Zhaocheng Wang, and Yong Li. MCDiff: Mobile traffic and user generation with multimodal context-aware diffusion model. *IEEE Transactions on Networking*, 34, 2026b.

Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3): 1247–1250, 2014.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

JT Edwards and DJ Thouless. Numerical studies of localization in disordered systems. *Journal of Physics C: Solid State Physics*, 5(8):807–820, 1972.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

Dayu Fan, Rui Meng, Xiaodong Xu, Yiming Liu, Guoshun Nan, Chenyuan Feng, Shujun Han, Song Gao, Bingxuan Xu, Dusit Niyato, et al. Generative diffusion models for wireless networks: Fundamental, architecture, and state-of-the-art. *IEEE Communications Surveys & Tutorials*, 2026.

Jie Feng, Jun Zhang, Junbo Yan, et al. CityBench: Evaluating the capabilities of large language models as world model. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025. doi: 10.1145/3711896.3737375.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Marta C. González, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008. doi: 10.1038/nature06958.

Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. Urbanvlp: Multi-granularity vision-language pretraining for urban socioeconomic indicator prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28061–28069, 2025.

Timothy O Hodson. Root mean square error (rmse) or mean absolute error (mae): When to use them or not. *Geoscientific Model Development Discussions*, 2022:1–10, 2022.

Mikhail Kanevski, A Pozdnukhov, and Vadim Timonin. Machine learning algorithms for geospatial data. applications and software tools. 2008.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Z Jonny Kong, Nathan Hu, Y Charlie Hu, Jiayi Meng, and Yaron Koral. High-fidelity cellular network control-plane traffic generation without domain knowledge. In *Proceedings of the 2024 ACM on Internet Measurement Conference*, pp. 530–544, 2024.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Percy Liang, Rishi Bommasani, Tony Lee, et al. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023. doi: 10.1111/nyas.15007.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Junyuan Liu, Xinglei Wang, Zichao Zeng, Jiazhuang Feng, Quan Qin, Ilya Ilyankou, Guangsheng Dong, and Tao Cheng. Cityrep: A unified benchmark for urban representations across cities, tasks, and modalities. *arXiv preprint arXiv:2605.26036*, 2026.

Xiaosi Liu, Xiaowen Xu, Zhidan Liu, Zhenjiang Li, and Kaishun Wu. Spatio-temporal diffusion model for cellular traffic generation. *IEEE Transactions on Mobile Computing*, 25(1):257–271, 2025.

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. Largest: A benchmark dataset for large-scale traffic forecasting. *Advances in neural information processing systems*, 36:75354–75371, 2023.

Thomas Louail, Maxime Lenormand, Oliva G Cantu Ros, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *Scientific reports*, 4(1):5276, 2014.

Jingtian Ma, Jingyuan Wang, Wayne Xin Zhao, Guoping Liu, and Xiang Wen. Spatio-temporal data enhanced vision-language model for traffic scene understanding. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

Lena Maier-Hein, Annika Reinke, Patrick Godau, et al. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, 2024. doi: 10.1038/s41592-023-02151-z.

Jayanta Mandi, James Kotary, Senne Berden, et al. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80: 1623–1701, 2024. doi: 10.1613/jair.1.15320.

Cristina Marquez, Marco Gramaglia, Marco Fiore, Albert Banchs, Cezary Ziemlicki, and Zbigniew Smoreda. Not all apps are created equal: Analysis of spatiotemporal heterogeneity in nationwide mobile service usage. In *Proceedings of the 13th international conference on emerging networking EXperiments and Technologies*, pp. 180–186, 2017.

Kenneth Martin, Simon Heilig, Asja Fischer, Michel FC Haddad, Adam M Sykulski, and Moshe Eliasof. A critical audit of spatiotemporal forecasting benchmark datasets and baselines. *arXiv preprint arXiv:2608.20980*, 2026.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Sachit Mishra, Zbigniew Smoreda, and Marco Fiore. Second-level digital divide: A longitudinal study of mobile traffic consumption imbalance in france. In *Proceedings of the ACM Web Conference 2022*, pp. 2532–2540, 2022.

Julia Moska, Oleksii Furman, Kacper Kozaczko, Szymon Leszkiewicz, Jakub Polczyk, Piotr Gramacki, and Piotr Szymański. Obsr: Open benchmark for spatial representations. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, pp. 670–681, 2025.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7176–7185, 2020.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

Chandrasen Pandey, Vaibhav Tiwari, Agbotiname Lucky Imoize, Chun-Ta Li, Cheng-Chi Lee, and Diptendu Sinha Roy. 5gt-gan: Enhancing data augmentation for 5g-enabled mobile edge computing in smart cities. *IEEE access*, 11:120983–120996, 2023.

Ioannis Panitsas, Tolga O. Atalay, Dragoslav Stojadinovic, Angelos Stavrou, and Leandros Tassiulas. 5gc-bench: A framework for stress-testing and benchmarking 5g core vnfs. In *2026 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, 2026. doi: 10.1109/WCNC65185.2026.11555639.

Xiaoqian Qi, Haoye Chai, Li Yu, Yong Li, and Zhaocheng Wang. Regional features conditioned diffusion models for 5g network traffic generation. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, pp. 396–409, New York, NY, USA, 2024. Association for Computing Machinery. doi: 10.1145/3678717.3691312.

Xiaoqian Qi, Haoye Chai, and Yong Li. MobiGPT: A foundation model for mobile wireless networks. *arXiv preprint arXiv:2509.18166*, 2025a.

Xiaoqian Qi, Haoye Chai, Sichang Liu, Lei Yue, Raoyuan Pan, Yue Wang, and Yong Li. Denoising refinement diffusion models for simultaneous generation of multi-scale mobile traffic. *arXiv preprint arXiv:2511.17532*, 2025b.

Can Rong, Jingtao Ding, Meng Li, and Yong Li. A global intra-city commuting origin-destination flow dataset for urban sustainable development. *Scientific Data*, 2026.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000. doi: 10.1023/A:1026543900054.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Charles Spearman. The proof and measurement of association between two things. 1961.

Akash Srivastava, Lazar Valkov, Chris Russell, Michael Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30, 2017.

Michael Steininger, Konstantin Kobs, Padraig Davidson, Anna Krause, and Andreas Hotho. Density-based weighting for imbalanced regression. *Machine Learning*, 110(8):2187–2211, 2021.

Fengze Sun, Egemen Tanin, Shanika Karunasekera, Zuqing Li, Flora D Salim, and Jianzhong Qi. Urbanverse: Learning urban region representation across cities and tasks. *arXiv preprint arXiv:2602.15750*, 2026.

Cheng Tan, Siyuan Li, Zhangyang Gao, Wenfei Guan, Zedong Wang, Zicheng Liu, Lirong Wu, and Stan Z Li. Openstl: A comprehensive benchmark of spatio-temporal predictive learning. *Advances in Neural Information Processing Systems*, 36:69819–69831, 2023.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Haoran Wan, Yaxiong Xie, and Kyle Jamieson. Neuralemu: in situ measurement-driven, ml-based, high-fidelity 5g network emulation. *arXiv preprint arXiv:2604.26080*, 2026.

Hongjun Wang, Jiyuan Chen, and Xuan Song. Mobilellm: Semantic-enhanced large language model for multimodal cellular traffic prediction in urban networks. *IEEE Transactions on Mobile Computing*, 2026.

Haomin Wen, Youfang Lin, Yutong Xia, Huaiyu Wan, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models. In *Proceedings of the 31st ACM international conference on advances in geographic information systems*, pp. 1–12, 2023.

Cort J Willmott and Kenji Matsuura. On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. *International journal of geographical information science*, 20(1):89–102, 2006.

Jingjin Wu, Yujing Zhang, Moshe Zukerman, and Edward Kai-Ning Yung. Energy-efficient base-stations sleep-mode techniques in green cellular networks: A survey. *IEEE communications surveys & tutorials*, 17(2):803–826, 2015.

Nemin Wu, Qian Cao, Zhangyu Wang, Zeping Liu, Yanlin Qi, Jielu Zhang, Joshua Ni, Xiaobai Yao, Hongxu Ma, Lan Mu, et al. Torchspatial: A location encoding framework and benchmark for spatial representation learning. *Advances in Neural Information Processing Systems*, 37:81437–81460, 2024.

Fengli Xu, Yong Li, Huandong Wang, Pengyu Zhang, and Depeng Jin. Understanding mobile traffic patterns of large scale cellular towers in urban environment. *IEEE/ACM transactions on networking*, 25(2):1147–1161, 2016.

Kai Xu, Rajkarn Singh, Marco Fiore, Mahesh K Marina, Hakan Bilen, Muhammad Usama, Howard Benn, and Cezary Ziemlicki. Spectragan: Spectrum based generation of city scale spatiotemporal mobile network traffic data. In *Proceedings of the 17th International Conference on Emerging Networking EXperiments and Technologies*, pp. 243–258, 2021.

Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *International conference on machine learning*, pp. 11842–11851. PMLR, 2021.

Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 186–194, 2012.

Shiyuan Zhang, Tong Li, Shuodi Hui, Guangyu Li, Yanping Liang, Li Yu, Depeng Jin, and Yong Li. Deep transfer learning for city-scale cellular traffic generation through urban knowledge graph. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4842–4851, 2023.

Shiyuan Zhang, Tong Li, Zhu Xiao, Hongyang Du, and Kaibin Huang. LSDM: LLM-enhanced spatio-temporal diffusion model for service-level mobile traffic prediction. *arXiv preprint arXiv:2507.17795*, 2025.

Shiyuan Zhang, Jiale Du, Yuanwei Liu, Kaibin Huang, and Hongyang Du. Netspatial: Spatially conditional traffic generation for cellular planning and operations. *arXiv preprint arXiv:2603.13839*, 2026a.

Shiyuan Zhang, Yilai Liu, Yuwei Du, Ruoxuan Yang, Dong In Kim, and Hongyang Du. U-MASK: User-adaptive spatio-temporal masking for personalized mobile AI applications. *arXiv preprint arXiv:2601.06867*, 2026b.

Xiting Zhang, Xudong Wang, Lijun Sun, and Luis Miranda-Moreno. Reconstruction of od demand using mobile phone data: A regression model with spatially varying coefficients. *Journal of Transport Geography*, 134:104677, 2026c.

Lingxue Zhu and Nikolay Laptev. Deep and confident prediction for time series at uber. In *2017 IEEE international conference on data mining workshops (ICDMW)*, pp. 103–110. IEEE, 2017.

# A Appendix

## A.1 Extended Related Work

**Mobile Traffic Modeling and Evaluation.** Mobile traffic modeling has expanded from task-specific forecasting to shared models for generation and prediction. UoMo (Chai et al., 2025b) unifies these tasks across cities, while MobiGPT (Qi et al., 2025a) extends shared modeling to multiple mobile-network data types. In the broader spatiotemporal literature, benchmarks have expanded both data coverage and evaluation scope. OpenSTL (Tan et al., 2023) standardizes predictive learning across several domains, while LargeST (Liu et al., 2023) provides multi-year road-sensor data for large-scale traffic forecasting. More recently, STContext (Ma et al., 2025) combines nine crowd-mobility datasets with ten contextual features and evaluates performance under rare conditions and across central and non-central regions. A recent benchmark by Martin et al. (2026) further questions whether commonly used forecasting datasets adequately distinguish spatial models from strong spatially unaware baselines. All these studies neglect the evaluation of the spatial heterogeneity caused by urban contexts.

**Spatial Context Representation.** Urban representation learning uses multimodal data and spatial relations to characterize regional differences. UrbanVLP (Hao et al., 2025) combines satellite imagery, street-view images, and text, while UrbanVerse (Sun et al., 2026) learns regional representations for transfer across cities and tasks. Corresponding benchmarks increasingly examine spatial bias and generalization alongside predictive accuracy. TorchSpatial (Wu et al., 2024) introduces LocBench to compare location encoders and explicitly measures geographic bias. OBSR (Moska et al., 2025)

broadens evaluation to region- and trajectory-level tasks using seven datasets from cities across three continents. CityRep (Liu et al., 2026) unifies heterogeneous urban representations across eight cities and eight tasks, using spatially blocked splits to reduce spatial leakage. These benchmarks evaluate spatial representations through downstream prediction tasks; This work complementarily focuses on the resulting mobile traffic, directly assessing whether regional demand differences are preserved.

### A.2 Details of Regional Knowledge Graph

The regional spatial context of Definition 2 is shared by all models and consists of geographic coordinates, population, points of interest, and a regional knowledge graph with pretrained embeddings. This section describes the entities and relations of the graph and the procedure used to obtain the embeddings. The graph contains two entity types, namely urban regions and points of interest, together with five relation types. *NearBy* links regions within a proximity threshold, *BorderBy* links regions that share a boundary, *SimilarFunc* links regions of similar functional composition, *LocateAt* places a point of interest in a region, and *Cate1Of* assigns a point of interest to its first-level category. The three graphs contain 67,418, 33,112, and 47,884 triples over 26,604, 11,988, and 16,579 entities for NYC, DC, and BM respectively. Each region additionally carries centroid coordinates, a point-of-interest count, population, education level, income, unemployment rate, and crime count. Entity embeddings are pretrained with TuckER (Balažević et al., 2019) under a cross-entropy objective, and the resulting 32-dimensional vectors are held fixed and shared by every model, so no model gains an advantage from a different graph representation.

### A.3 Datasets

The three cities come from one upstream source of hourly regional mobile traffic. Each city is divided into regions at which traffic is recorded, and the regions are split spatially into disjoint training, validation, and test sets, so that test regions are never seen during training. Generation samples are 168-hour weekly windows taken at a one-day stride, so a city with $D$ days provides $D - 6$ windows; forecasting samples are 84-hour horizons paired with their preceding history, each forecasting sample takes the first 84 hours of a 168-hour window as history and predicts the following 84 hours, so the history and the horizon are of equal length. Table 3 lists the sizes, the splits, and the coefficient of variation of observed regional mean traffic on the test regions, which is the denominator of SDR. All models receive the same regional spatial context, namely geographic coordinates, population, points of interest, and the regional knowledge graph with pretrained embeddings described in Appendix A.2. None of the datasets includes satellite imagery, per-region user counts, wireless channel state, quality-of-service alarms, weather, or network actions; models whose original versions use such inputs are adapted as described in Appendix A.4.

Table 3: Datasets. Region splits are disjoint. Windows are 168-hour generation windows at a one-day stride, and the coefficient of variation (CV) is computed over the regional mean traffic of the test regions.

| City | Days | Regions | Train | Val. | Test | Windows | Region-mean CV |
|---|---|---|---|---|---|---|---|
| NYC | 182 | 280 | 175 | 44 | 61 | 176 | 0.693 |
| DC | 151 | 237 | 142 | 36 | 59 | 145 | 1.840 |
| BM | 151 | 403 | 154 | 39 | 210 | 145 | 1.328 |

### A.4 Models and Training Configuration

Table 4 lists the twelve models with their track, their main mechanism for using spatial information, and their implementation source. AR denotes an adaptation of author-released code, PF a reproduction written from the paper and checked against it, and BN a benchmark-native baseline that involves no learning. To run on the shared inputs, MCDiff omits the satellite-image and user-count channels of its original version, and ZoomDiff uses a deterministic conversion from regions to a grid and back. These adaptations make evaluation on common data possible; they do not claim to reproduce the results reported on the original data.

Table 4: Models on the two tracks. AR: adaptation of author-released code; PF: paper-based reproduction; BN: benchmark-native baseline.

| Model | Track | Main spatial mechanism | Src |
|---|---|---|---|
| KG-Diff | Generation | knowledge-graph conditioned diffusion | PF |
| MCDiff | Generation | multimodal conditional diffusion | PF |
| MobiGPT | Generation | prompted diffusion transformer | PF |
| STK-Diff | Generation | temporal and graph knowledge | PF |
| ZoomDiff | Generation | multiscale refinement | PF |
| 5GT-GAN | Generation | recurrent GAN | AR |
| SeasonalMean | Generation | hour-of-week profile | BN |
| GeoKNN | Generation | inverse-distance spatial transfer | BN |
| MobiFM | Forecasting | memory and mixture-of-experts diffusion transformer | AR |
| DiffSTG | Forecasting | graph diffusion | AR |
| PatchTST | Forecasting | patched temporal transformer | AR |
| LastDay | Forecasting | previous-day repetition | BN |

All learned models except KG-Diff and MCDiff, which keep their native configurations, share the settings in Table 5. Model selection uses the validation loss only. Normalization statistics are fitted on the training regions and training windows, and every output is transformed back to the physical scale before any metric is computed. Each configuration is trained once under the fixed seed reported in Table 5, applied identically across all models to keep the 162-run archive, at about 38.1 GPU-hours in total, directly comparable.

Table 5: Shared training configuration of the learned models.

| Setting | Value |
|---|---|
| Batch size | 128 |
| Hidden width | 64 |
| Number of layers | 2 |
| Diffusion steps (training) | 80 |
| Diffusion steps (inference) | 40 |
| Training length | 160 epochs |
| Optimizer | AdamW, weight decay $10^{-4}$ |
| Learning rate | $10^{-3}$ |
| Model selection | validation loss |
| Random seed | 20260706 (one per configuration) |
| Hardware | 4× NVIDIA RTX 5090 |

### A.5 Detailed Experimental Setup

**Datasets.** We evaluate on three U.S. cities: New York City (NYC), Washington, D.C. (DC), and Baltimore (BM). Each city provides hourly regional mobile traffic together with the regional spatial context defined in Section 2.1. Regions are partitioned spatially into disjoint training, validation, and test sets; only training regions are used for model fitting, and test regions remain entirely unseen during training. The three cities contain 61, 59, and 210 test regions, respectively. For generation, each sample is a 168-hour weekly window, with windows shifted by one day and treated as exchangeable; generated windows are therefore not paired one-to-one with observed windows. For forecasting, each sample consists of an 84-hour future window paired with its corresponding history. Dataset sizes, split statistics, coefficients of variation of observed regional mean traffic, and city-specific blindness references are reported in Table 3. Both the coefficients of variation and the reference values vary substantially across cities, with the latter differing by a factor of 4.9; we therefore normalize both SDR and $\mathrm{SAD}^{\star}$ within each city.

**Models.** We evaluate eight generation models and four forecasting models, covering the four methodological families introduced in Section 1. Each model is evaluated once in each city, yielding 24 generation evaluations and 12 forecasting evaluations. The generation models include two transpar-

ent baselines, SeasonalMean, which averages traffic by hour-of-week, and GeoKNN (Kanevski et al., 2008), which transfers traffic from training regions using inverse-distance weighting; one recurrent GAN, 5GT-GAN (Pandey et al., 2023); four diffusion models incorporating spatial knowledge or multi-scale structure, KG-Diff (Chai et al., 2024b), MCDiff (Chai et al., 2026b), STK-Diff (Chai et al., 2025a), and ZoomDiff (Qi et al., 2025b); and the diffusion-Transformer foundation model MobiGPT (Qi et al., 2025a). The forecasting models include the LastDay baseline (Zhu & Laptev, 2017), which repeats the previous day, the graph diffusion model DiffSTG (Wen et al., 2023), the patch-based temporal Transformer PatchTST (Nie et al., 2023), and the foundation model MobiFM (Chai et al., 2026a). We annotate each implementation by source: AR denotes adaptations of author-released code, PF denotes paper-based reproductions, and BN denotes benchmark-native baselines. Five of the six learned generation models are PF implementations; accordingly, the results in Section 4 characterize these implementations with the training used in our experiments rather than the upper-bound performance of the original methods. To operate under a shared input setting, some models omit modalities unavailable in our benchmark, such as satellite imagery and user-count channels for MCDiff, while ZoomDiff uses a deterministic region-to-grid conversion. Model mechanisms, implementation sources, and adaptation details are provided in Table 4. During auditing, we identified one model that copied target-region traffic into a conditioning channel; all results from this model are excluded.

**Evaluation protocol.** All learned models use the same number of training epochs and the same network size, except KG-Diff and MCDiff, for which we retain their native configurations. Model selection is based solely on validation loss. Normalization parameters are fitted only on training regions and training windows, and all outputs are transformed back to the physical scale before any metric is computed; full configurations are given in Table 5. All metrics are computed directly from saved model outputs, without retraining. For generation, SDR, $\mathrm{SAD}^{\star}$, and PENTAD are computed between the mean field of generated windows and that of observed test windows. Cross-window variation is evaluated by comparing sorted distributions, making the evaluation invariant to window ordering. Forecasting metrics are instead computed on paired future windows. We do not compare the two tracks on a shared numerical scale: generation profile errors range from 26.8 to 8932.4, whereas paired forecasting errors range only from 12.8 to 86.4, making pooled comparisons misleading. For the five downstream tasks, test windows are partitioned into non-overlapping decision and evaluation blocks: the former are used to select actions, and the latter to compute costs. Task-specific operating points are listed in Table 12. Task-specific operating points are listed in Table 12, and training follows the single fixed-seed protocol in Table 5 for every model. To examine whether existing evaluation practices capture losses of spatial heterogeneity, we additionally compute eight metrics commonly used in prior mobile traffic studies: four fit-oriented metrics, MAE, RMSE, sMAPE, and peak error; and four deployment-oriented metrics: congestion miss cost, capacity-planning regret, hotspot migration distance, and volatility error. The capacity-planning regret follows the definition used in prior work and differs from the G1 task in Section 3.3 in both its loss function and reference policy; the two should therefore not be used interchangeably. Full metric definitions are provided in Appendix A.7.

### A.6 Details of the PENTAD Perspectives

This section gives the full definitions of the perspectives summarized in Section 3.2.

**Level Bias Error (LBE).** We measure the relative difference in overall traffic volume using

$$\mathrm{LBE} = \frac{|\widehat{T} - T|}{T}. \tag{11}$$

This score captures over- or underestimation of total demand, independently of how that demand is distributed across regions and time. We also retain the signed counterpart $(\widehat{T} - T)/T$ to distinguish the two directions of error, which can have different operational consequences.

**Temporal Profile Divergence (TPD).** We compare the observed and modeled temporal shares using Jensen–Shannon divergence (Lin, 1991):

$$\begin{aligned} \mathrm{TPD} &= \frac{1}{2}\mathrm{KL}_2(p\|m) + \frac{1}{2}\mathrm{KL}_2(\widehat{p}\|m), \\ m &= \frac{p+\widehat{p}}{2}, \end{aligned} \tag{12}$$

where $\mathrm{KL}_2$ denotes Kullback–Leibler divergence with base-2 logarithms. TPD evaluates when citywide demand occurs, independently of total traffic volume. It is also unchanged by spatial reallocations that preserve the citywide temporal profile.

**Coupling Lift Error (CLE).** For a strictly positive joint distribution $P$, Sinkhorn normalization alternately rescales its rows and columns to obtain (Cuturi, 2013; Sinkhorn & Knopp, 1967)

$$S(P) = \mathrm{diag}(u)\, P\, \mathrm{diag}(v), \tag{13}$$

where $u$ and $v$ are positive scaling vectors chosen so that every row sums to $1/R$ and every column sums to $1/H$. The resulting matrix is unique and satisfies

$$S(\mathrm{diag}(a)\, P\, \mathrm{diag}(b)) = S(P) \tag{14}$$

for any positive row and column scaling vectors $a$ and $b$. For strictly positive observed and modeled distributions, we define

$$\mathrm{CLE} = \frac{1}{2}\sum_{r=1}^{R}\sum_{t=1}^{H}\left|S(\widehat{P})[r,t] - S(P)[r,t]\right|. \tag{15}$$

This total-variation distance compares region–time interactions without treating positive row or column rescaling as an interaction error.

### A.7 Conventional Metrics

Section 4 compares the benchmark scores with eight metrics that prior mobile traffic studies commonly report. Figure 5 illustrates why matching one of these scores does not settle a deployment question. Four of them measure fit: MAE, RMSE, and sMAPE (Hodson, 2022; Chai & Draxler, 2014) are computed over region–time cells, against the mean profile on the generation track and on paired horizons on the forecasting track, and peak error compares the maximum traffic of each region. Four are deployment-oriented summaries: congestion miss cost, capacity-planning regret, hotspot migration distance, and volatility error. Table 6 summarizes what each metric measures. Capacity-planning regret here follows its definition in prior work and is distinct from the G1 regret of Section 3.3, which uses the loss and reference policy of Table 12.

Let $X$ and $\widehat{X}$ denote the observed and modeled fields over the evaluated region–time cells, and write $(\cdot)_+ = \max(\cdot, 0)$. The four deployment-oriented metrics are defined as follows.

**Congestion miss cost.** A congestion level $\tau$ is fixed at the 90th percentile of the observed field. Each cell contributes its observed exceedance above $\tau$ whenever the modeled value falls below $\tau$, and the total is normalized by the observed exceedance mass,

$$\mathrm{CMC} = \frac{\sum_c (X_c - \tau)_+ \mathbb{1}\left[\widehat{X}_c < \tau\right]}{\sum_c (X_c - \tau)_+} \in [0,1]. \tag{16}$$

A severe overload that goes undetected therefore costs more than a marginal one, and the score is comparable across cities of different traffic scale.

**Capacity-planning regret.** Capacity is provisioned per region at the $\alpha = 0.95$ quantile of the predicted demand, $c_r = Q_\alpha(\widehat{X}_{\cdot,r,\cdot})$, and charged with an asymmetric cost whose shortage weight $\lambda_\mathrm{u} = 19$ and surplus weight $\lambda_\mathrm{o} = 1$ satisfy $\lambda_\mathrm{u}/(\lambda_\mathrm{u}+\lambda_\mathrm{o}) = \alpha$, so that the $\alpha$ quantile minimizes the cost,

$$C(c) = \mathrm{mean}_c\left[\lambda_\mathrm{u}\,(X_c - c)_+ + \lambda_\mathrm{o}\,(c - X_c)_+\right]. \tag{17}$$

Writing $c^*$ for the same rule applied to observed demand, the reported score is the relative excess cost,

$$\mathrm{CPR} = \frac{(C(c) - C(c^*))_+}{C(c^*)}. \tag{18}$$

**Hotspot migration distance.** At each window $w$ and hour $t$, the traffic-weighted centroid of the top $k = \lceil 0.10R \rceil$ regions is computed for both fields,

$$m(Z)_{w,t} = \frac{\sum_{r \in \mathcal{T}_k(Z_{w,\cdot,t})} Z_{w,r,t}\, \mathbf{p}_r}{\sum_{r \in \mathcal{T}_k(Z_{w,\cdot,t})} Z_{w,r,t}}, \tag{19}$$

where $\mathcal{T}_k$ returns the top-$k$ region indices and $\mathbf{p}_r$ is the coordinate pair of region $r$. The metric averages the haversine distance $d_{\mathrm{hav}}$ between the two centroid trajectories,

$$\mathrm{HMD} = \mathrm{mean}_{w,t}\, d_{\mathrm{hav}}\Big(m(\widehat{X})_{w,t},\, m(X)_{w,t}\Big), \tag{20}$$

reported in kilometres. The centroid is recomputed at every hour, so averaging a whole week before locating one hotspot would not measure migration.

**Volatility error.** Short-term variability is compared through the first difference along time, $\Delta Z_{\cdot,\cdot,t} = Z_{\cdot,\cdot,t+1} - Z_{\cdot,\cdot,t}$,

$$\mathrm{VE} = \sqrt{\mathrm{mean}\left[\left(\Delta\widehat{X} - \Delta X\right)^2\right]}. \tag{21}$$

This charges errors in the rate of change rather than in the level, so a model that matches the average level but produces an over-smoothed profile is still penalized.

Table 6: The eight conventional metrics used for comparison.

| Metric | Type | What it measures |
|---|---|---|
| MAE | fit | mean absolute error over region–time cells |
| RMSE | fit | root mean squared error over region–time cells |
| sMAPE | fit | symmetric mean absolute percentage error over region–time cells |
| Peak error | fit | relative error of the maximum traffic of each region, averaged over regions |
| Congestion miss cost | deployment | cost of hours in which observed traffic exceeds a congestion threshold but modeled traffic does not |
| Capacity-planning regret | deployment | extra cost when capacity is provisioned from modeled rather than observed peak demand |
| Hotspot migration distance | deployment | distance between the highest-traffic regions under modeled and observed traffic |
| Volatility error | deployment | difference between modeled and observed short-term variability of traffic |

### A.8 City-Specific References for Spatial Allocation Error

For each city, we compute the uniform-allocation reference $\Phi(q)$ of Equation 7 once, by solving the transport problem of Equation 6 between uniform regional shares and the observed shares as a linear program with haversine ground distances. To locate this reference on the attainable scale, we also compute the largest transport cost obtained when all traffic share is placed in a single region,

$$\mathrm{SAD}_{\max} = \max_{r} W_1(\delta_r, q; D), \tag{22}$$

where $\delta_r$ assigns all traffic share to region $r$. Table 7 reports both values together with the observed coefficient of variation of regional mean traffic. The reference occupies between 6.6% (BM) and 24.7% (DC) of the attainable range, so an output with uniform regional shares already lies in the lowest quarter of the raw error scale in every city. It also varies by a factor of 4.9 across cities, because both the observed shares and the geography differ. These two properties are why Section 3.1 reports $\mathrm{SAD}^\star$ relative to the city-specific reference rather than raw SAD.

### A.9 Per-City Results of the Generation Track

Table 8 lists the 24 generation evaluations by city. Region-mean correlation complements SDR by measuring alignment rather than amplitude: in NYC, MCDiff reaches a correlation of 0.412

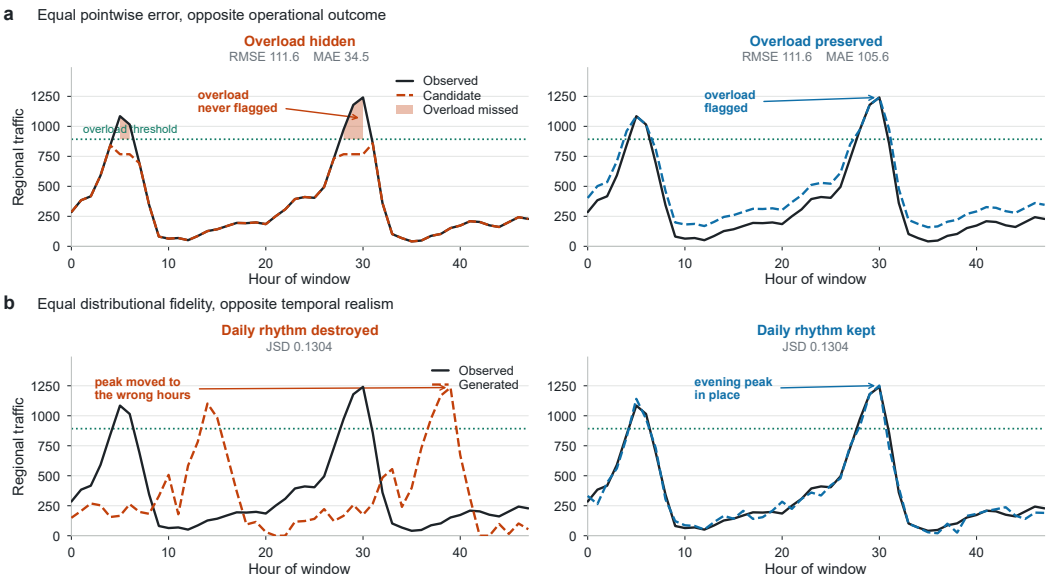


Figure 5: Conventional fidelity metrics do not determine operational outcome. **(a)** Identical RMSE to $2.8 \times 10^{-14}$, yet the left candidate leaves five overload hours undetected (shaded) and the right leaves none. **(b)** Jensen–Shannon divergence matched to $5.1 \times 10^{-6}$, yet correlation with the observed series is $-0.340$ and $+0.991$. The observed series is a real NYC regional profile from the archive and the threshold is its 90th percentile. *The four candidate series are constructed to hold one metric fixed while the consequence differs, and are not model outputs.*

Table 7: City-specific reference values for spatial allocation error. $\mathrm{SAD}_{\mathrm{floor}} = \Phi(q)$ is the error of uniform regional shares (Equation 7) and $\mathrm{SAD}_{\max}$ is the largest attainable error (Equation 22). Region counts refer to test regions.

| City | Test regions | Region-mean CV | $\mathrm{SAD}_{\mathrm{floor}}$ (km) | $\mathrm{SAD}_{\max}$ (km) | Floor as % of range |
|---|---|---|---|---|---|
| NYC | 61 | 0.693 | 0.496 | 2.827 | 17.5% |
| DC | 59 | 1.840 | 1.931 | 7.825 | 24.7% |
| BM | 210 | 1.328 | 2.443 | 36.800 | 6.6% |

despite an SDR of 0.041, whereas MobiGPT reaches $-0.027$, so its small remaining variation is not aligned with the observed field. Conversely, GeoKNN in BM retains an SDR of 0.415 but has an $\mathrm{SAD}^{\star}$ of 1.939 and a negative correlation, so matching dispersion alone does not establish accurate allocation. Evaluations with SDR below 0.05 cluster at $\mathrm{SAD}^{\star} \approx 1$, and their CLE lies within 3.7% of the uniform reference. In DC, the raw SAD of the seven collapsed models lies between 1.928 and 1.935 km, while GeoKNN reaches 1.116 km, although the level biases of the collapsed models differ by more than two orders of magnitude; this is the fixed-reference property discussed in Section 3.1. Cross-city comparisons on the raw scale are more misleading still: a nearly uniform output in NYC receives 0.496 km, whereas KG-Diff in BM, which retains about a quarter of the observed dispersion, receives 2.256 km.

### A.10 Effective Dimension of Metric Families

For each family of metrics, we compute the Spearman rank-correlation matrix across the evaluations of one track and summarize its redundancy by the effective dimension $(\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$, where $\lambda_i$ are the eigenvalues of the matrix (Edwards & Thouless, 1972). This quantity equals $m$ for $m$ mutually uncorrelated metrics and one when all metrics rank the models identically. On the generation track, the eight conventional metrics carry an effective dimension of 2.21, with MAE and RMSE ranking the models identically, whereas the five PENTAD perspectives carry 3.45 and the five decision regrets 3.16 (Table 9). The forecasting rows rest on four models and are reported for completeness only.

Table 8: Generation evaluations per city, sorted by $\mathrm{SAD}^\star$ within each city. Corr. is the correlation between modeled and observed regional means and is undefined for SeasonalMean, whose regional means are identical.

| City | Model | SDR | Corr. | SAD (km) | $\mathrm{SAD}^\star$ | CLE | RMSE |
|---|---|---|---|---|---|---|---|
| NYC | KG-Diff | 0.856 | 0.647 | 0.234 | 0.472 | 0.121 | 93.6 |
| | GeoKNN | 0.202 | 0.231 | 0.400 | 0.807 | 0.123 | 111.3 |
| | MCDiff | 0.041 | 0.412 | 0.478 | 0.964 | 0.131 | 679.7 |
| | STK-Diff | 0.035 | 0.162 | 0.489 | 0.986 | 0.128 | 747.5 |
| | MobiGPT | 2.7e-4 | −0.027 | 0.496 | 1.000 | 0.128 | 803.4 |
| | ZoomDiff | 1.5e-3 | 0.071 | 0.496 | 1.000 | 0.128 | 737.9 |
| | 5GT-GAN | 1.0e-5 | 0.070 | 0.496 | 1.000 | 0.128 | 1532.0 |
| | SeasonalMean | 4.0e-16 | – | 0.496 | 1.000 | 0.128 | 87.9 |
| DC | GeoKNN | 0.439 | 0.367 | 1.116 | 0.578 | 0.121 | 78.1 |
| | MCDiff | 3.8e-3 | −0.087 | 1.928 | 0.998 | 0.124 | 943.2 |
| | KG-Diff | 3.2e-3 | −0.192 | 1.929 | 0.999 | 0.123 | 8932.4 |
| | 5GT-GAN | 1.0e-5 | −0.077 | 1.931 | 1.000 | 0.119 | 2295.2 |
| | ZoomDiff | 1.7e-4 | 0.015 | 1.931 | 1.000 | 0.119 | 1145.0 |
| | SeasonalMean | 6.3e-17 | – | 1.931 | 1.000 | 0.119 | 52.6 |
| | STK-Diff | 2.9e-3 | 0.068 | 1.933 | 1.001 | 0.120 | 2075.4 |
| | MobiGPT | 5.6e-3 | −0.135 | 1.935 | 1.002 | 0.121 | 1018.5 |
| BM | KG-Diff | 0.251 | 0.479 | 2.256 | 0.924 | 0.144 | 41.4 |
| | MobiGPT | 0.025 | −0.067 | 2.424 | 0.992 | 0.138 | 833.4 |
| | MCDiff | 5.7e-3 | −0.083 | 2.439 | 0.998 | 0.141 | 739.0 |
| | SeasonalMean | 0.0 | – | 2.443 | 1.000 | 0.138 | 26.8 |
| | 5GT-GAN | 1.7e-4 | −0.061 | 2.443 | 1.000 | 0.138 | 774.1 |
| | ZoomDiff | 1.6e-4 | −0.022 | 2.443 | 1.000 | 0.138 | 771.2 |
| | STK-Diff | 0.103 | −0.064 | 2.477 | 1.014 | 0.138 | 806.6 |
| | GeoKNN | 0.415 | −0.061 | 4.735 | 1.939 | 0.137 | 29.0 |

Table 9: Effective dimension of each metric family. The conventional family contains MAE, RMSE, sMAPE, peak error, congestion miss cost, capacity-planning regret, hotspot migration distance, and volatility error. Efficiency is the effective dimension divided by the number of metrics $m$, and mean $|\rho|$ is the mean absolute pairwise rank correlation within the family.

| Track | Family | $m$ | Eff. dim. | Efficiency | Mean $\lvert\rho\rvert$ |
|---|---|---|---|---|---|
| Generation | Conventional | 8 | 2.21 | 28% | 0.51 |
| | PENTAD | 5 | 3.45 | 69% | 0.29 |
| | Decision regret | 5 | 3.16 | 63% | 0.32 |
| Forecasting | Conventional | 8 | 1.30 | 16% | 0.83 |
| | PENTAD | 5 | 1.95 | 39% | 0.59 |
| | Decision regret | 5 | 1.86 | 37% | 0.57 |

### A.11 Perturbation-Response Test for PENTAD

For each city, we apply five perturbations to the observed traffic field, each designed to alter one factor of the decomposition in Equation 9 or the cross-window fluctuations. Volume scaling multiplies the field by a constant and is designed to change only the total volume $T$; temporal reshaping is designed to change the citywide temporal shares $p_t$; spatial reallocation is designed to change the regional shares $q_r$; interaction flattening is designed to move the region–time interaction $L[r, t]$ toward one while keeping both marginals; and dispersion collapse is designed to shrink the fluctuations of each cell across windows toward its mean. Write the mean field as the product of total mass, region share, time share, and a Sinkhorn canonical interaction, and let $s \in (0, 1]$ be the intervention strength. *Level gain* multiplies total mass by $1 + s$ and leaves both marginals and the interaction untouched. *Temporal reshape* moves the time-of-week marginal a fraction $s$ towards a fixed Dirichlet target. *Spatial reallocation* moves the region marginal a fraction $s$ towards a random permutation of itself, so mass changes location without changing the multiset of shares. *Coupling flattening* interpolates

the canonical interaction towards the uniform table and reprojects it, which preserves both marginals exactly. *Dispersion collapse* shrinks every window towards the mean week, leaving the mean field fixed and reducing only cross-window fluctuation. All randomness uses seed 20260901.

The strength of each perturbation is chosen by bisection so that all five reach the same pointwise RMSE against the observed field. A perturbation is attributed correctly when its targeted perspective shows the largest change among the five. Table 10 reports the results. Interaction flattening has the lowest response ratio because the Sinkhorn normalization is solved iteratively to a tolerance of $10^{-12}$; the changes of about $10^{-9}$ in the remaining perspectives are consistent with this numerical tolerance rather than with a genuine response.

Table 10: Response of PENTAD to RMSE-matched perturbations. Hits counts the cities in which the targeted perspective shows the largest change. The response ratio is the change in the targeted perspective divided by the largest change among the remaining four perspectives. All 15 city–perturbation combinations identify the targeted perspective.

| Perturbation | Targeted perspective | Hits | Median response ratio | Max RMSE mismatch |
|---|---|---|---|---|
| Volume scaling | LBE | 3/3 | $1.0 \times 10^{11}$ | $7.0 \times 10^{-5}$ |
| Temporal reshaping | TPD | 3/3 | $2.6 \times 10^{9}$ | $6.6 \times 10^{-5}$ |
| Spatial reallocation | SAD$^{\star}$ | 3/3 | $1.7 \times 10^{11}$ | $9.3 \times 10^{-5}$ |
| Interaction flattening | CLE | 3/3 | $2.3 \times 10^{7}$ | $7.3 \times 10^{-5}$ |
| Dispersion collapse | DLD | 3/3 | $1.5 \times 10^{11}$ | $4.7 \times 10^{-5}$ |
| Total | | 15/15 | $9.5 \times 10^{10}$ | $9.3 \times 10^{-5}$ |

### A.12 Matched-Control Test of the Mean-Field Perspectives

For each of the 12 forecasting outputs, we construct a matched control field whose total volume, temporal shares, spatial shares, and region–time interaction equal those of the output, so that the four mean-field perspectives agree with the original to within $2.6 \times 10^{-9}$ in CLE, and whose cross-window fluctuations are drawn from the observed field, so that the median DLD falls from 0.154 to zero. We then run the five decision tasks on the control field and on the original output and report the median relative gap in regret over the 12 evaluations (Table 11). The gap is smallest for energy saving, whose decision rule depends mainly on the mean load of each region-hour, and largest for TSTR and congestion warning, which depend on how traffic fluctuates across windows. The construction changes dispersion and possibly cross-window dependence, so it does not isolate the effect of DLD alone, nor does it test whether all five perspectives together are sufficient.

Table 11: Median relative gap in decision regret between each forecasting output and its matched control field.

| Task | Capacity | Energy | Allocation | Warning | TSTR |
|---|---|---|---|---|---|
| Median relative gap | 0.37 | 0.008 | 0.11 | 0.46 | 0.52 |

### A.13 Decision-Task Specifications

Beyond modeling quality, we evaluate whether model outputs support effective network decisions and downstream predictions. We consider five tasks, each with a specified decision rule, task loss, and observed-data reference. The protocol separates action selection on a decision block from cost measurement on a disjoint evaluation block. This allows us to assess the consequences of using model outputs rather than inferring practical value from prediction error alone.

**Capacity Planning (G1).** This task evaluates whether modeled demand supports appropriate capacity allocation across regions. We provision each region at an upper quantile $\alpha$ of its modeled demand (Koenker & Bassett Jr, 1978). Given allocated capacity $c$ and observed demand $y$, the loss is

$$\ell_{\mathrm{cap}}(c, y) = \lambda_{\mathrm{u}}(y - c)_+ + \lambda_{\mathrm{o}}(c - y)_+, \tag{23}$$

where $(z)_+ = \max(z, 0)$, and $\lambda_u$ and $\lambda_o$ specify the costs of insufficient and unused capacity, respectively. The observed-data reference applies the corresponding quantile rule to observed demand. This task distinguishes underprovisioning from overprovisioning, even when the two produce similar absolute prediction errors.

**Energy Saving (G2).** This task evaluates whether model outputs identify low-demand periods in which cells can be put to sleep. A region-hour is assigned to sleep when its normalized modeled load falls below the threshold prescribed by the task's energy-versus-service loss. The cost balances energy savings against penalties for violating the service-level agreement (SLA). The observed-data reference minimizes the same loss separately for each region-hour, rather than forcing a fixed proportion of cells to sleep. Keeping all cells active provides the inaction baseline.

**Resource Allocation (G3).** This task evaluates whether limited resources are directed toward the regions with the greatest demand. At each hour, we select the $k$ regions with the highest modeled demand, where $k$ specifies the resource budget. The loss measures peak demand left uncovered by the selected regions. The reference selects the true top-$k$ regions using observed demand. This task therefore assesses whether the model identifies the locations that should receive priority, rather than whether it only reproduces citywide demand.

**Congestion Warning (G4).** This task evaluates whether modeled traffic supports timely warnings for high-demand events. A warning is issued when modeled demand exceeds a threshold defined by a high quantile of observed traffic (Elmachtoub & Grigas, 2022; Mandi et al., 2024). Missed events are penalized according to their exceedance severity, while false alarms incur a fixed cost. We compare the resulting cost with an observed-data reference and a never-alarm baseline. The task captures the different consequences of overlooking severe demand peaks and issuing unnecessary warnings.

**Synthetic-Data Utility (G5).** We assess whether model-produced traffic can support learning through a train-on-synthetic, test-on-real (TSTR) task (Esteban et al., 2017). A ridge forecaster is trained on synthetic traffic to predict next-day demand and evaluated on real traffic. A ridge forecaster trained on real data provides the reference. This reference is a comparison predictor, not a guaranteed optimum. The task evaluates whether the generated data preserve patterns that are useful for prediction on real observations.

**Task-Specific Normalization.** Let $C_j^{\text{model}}$ and $C_j^{\text{ref}}$ denote the evaluation costs of the model-based and reference procedures for task $j$. For capacity planning, resource allocation, and synthetic-data utility, we report

$$\text{Regret}_j = \frac{C_j^{\text{model}} - C_j^{\text{ref}}}{|C_j^{\text{ref}}|}, \qquad j \in \{1, 3, 5\}, \tag{24}$$

when the reference cost is nonzero.

Energy-saving costs can change sign, and the warning reference has zero cost under the stated task. For these tasks, we instead normalize by the cost difference between the inaction baseline and the reference:

$$\text{Regret}_j = \frac{C_j^{\text{model}} - C_j^{\text{ref}}}{C_j^{\text{base}} - C_j^{\text{ref}}}, \qquad j \in \{2, 4\}, \tag{25}$$

where $C_2^{\text{base}}$ is the never-sleep cost and $C_4^{\text{base}}$ is the never-alarm cost. With a positive denominator, a score of zero matches the reference, a score of one matches inaction, and a score above one indicates greater cost than inaction. Lower scores indicate better task performance. We report the five scores separately because they reflect different operational objectives.

Each task selects actions on a decision block and measures cost on a disjoint evaluation block of the test windows. The observed-data reference policies apply the same decision rule to observed traffic; for G5 the reference is a ridge forecaster trained on real data, which is a reference predictor rather than an optimum. Table 12 lists the decision rules, losses, and operating points. Regret is normalized in two ways, as described in Section 3.3: for capacity planning, resource allocation, and synthetic-data utility, the cost difference from the reference is divided by the reference cost; for energy saving and congestion warning, whose reference cost is zero or can change sign, it is divided

by the cost difference between inaction and the reference, so that zero matches the reference and one matches inaction.

The reference policies themselves were checked. An earlier energy-saving rule put a fixed 30% of cells to sleep; because that rule does not minimize the task loss, a spatially scrambled field outperformed it with a regret of $-2.13$, which exposed the reference as suboptimal. The final rule therefore takes the cellwise minimizer of the same loss on which the models are scored, so the number of sleeping cells is an outcome rather than a parameter. The held-out windows are split into two disjoint halves by a deterministic permutation under seed $20260901$. The decision block is used only to choose the action and the evaluation block only to pay for it, so a model cannot be credited for a decision validated on the same windows that scored it. For congestion warning the congestion level is fixed once from the observed decision block, so the model and the reference are judged against the same physical overload definition and a model cannot game the task by rescaling its own output. Each field then alarms on its own exceedance of that level. Misses are charged at $10$ per unit of severity above the level and false alarms at a flat cost of $1$. Because an oracle that sees the truth attains exactly zero cost on this task, the regret is normalized against the range between the never-alarm baseline and the reference rather than against the reference cost. We also compared the two-block protocol with a single-block variant in which actions are selected and costs are measured on the same windows: the median absolute difference in regret is 0.037 for capacity planning and 0.004 for resource allocation, and the two-block value is lower in 84% and 74% of the evaluations, respectively, so the single-block variant would report higher regret in most cases.

### A.14 Extended Result Analysis

**The collapse is not an artifact of the evaluation.** Three checks support this. All metrics are computed by grouping each run's outputs by its own region index, without reshaping any array; the observed targets retain their full regional variation; and the same procedure yields SDR close to one on the forecasting track, with mean region-mean correlations between 0.900 and 0.999, so the measure is not saturated. These findings describe the outputs saved after the training used in our experiments; whether further training or author-released code would change them is a separate question.

**Complementarity of the perspectives.** The two generation models that retain regional dispersion show that the perspectives complement each other: KG-Diff has the best $\mathrm{SAD}^{\star}$ and DLD among generation models but the worst LBE, whereas GeoKNN has the best LBE and TPD but an $\mathrm{SAD}^{\star}$ above one.

**Forecasting errors on the mean field are small overall, but the rankings by perspective differ from the ranking by paired error.** PatchTST has the lowest paired RMSE at 27.7, followed by DiffSTG at 38.4, LastDay at 44.1, and MobiFM at 50.8. Yet LastDay, which repeats the previous day, has the lowest value on four of the five PENTAD perspectives: TPD $9.3 \times 10^{-4}$, $\mathrm{SAD}^{\star}$ 0.036, CLE 0.024, and DLD 0.080; only LBE is led by PatchTST, at 0.01 against 0.04. While reducing pointwise error, the three learned models alter the region–time interaction and the distribution of cross-window fluctuations: their CLE lies between 0.091 and 0.237 and their DLD between 0.203 and 0.351, against 0.024 and 0.080 for LastDay. MobiFM is the worst on all five perspectives, with LBE 0.70 and $\mathrm{SAD}^{\star}$ 1.134, even though its SDR reaches 0.867: it preserves the amount of regional variation but misplaces both total volume and allocation. The lead of LastDay does not mean that a simple baseline is better. Its mean field is nearly the observed mean field shifted by one day, and the periodicity of traffic gives it a natural advantage on mean-field perspectives. The comparison shows instead that paired error and mean-field structure measure different things: a forecaster with low RMSE can still depart from the observations in its average structure.

**Decision results by task on the generation track.** KG-Diff, which overestimates total traffic by 123 times, incurs a capacity cost more than 140 times that of the reference policy and a TSTR regret of 146.40, whereas GeoKNN has a TSTR regret of 0.54. Resource-allocation regret lies between 0.439 and 0.631 for all generation models; KG-Diff and GeoKNN, which retain regional variation, are the lowest but still far from the reference. Congestion warning needs a separate explanation. 5GT-GAN, ZoomDiff, and STK-Diff, which overestimate total traffic by 34 to 53 times, reach the best warning regret of 0.090 because their outputs exceed the threshold almost everywhere and therefore raise an

alarm in almost every hour; under the present miss penalty and false-alarm cost (Table 12), alarming continuously costs about 9% of never alarming. SeasonalMean, at 0.965, is close to never alarming. This is not a flaw of the warning task but a sign that the task is sensitive to the direction of the error.

**Decision results on the forecasting track.** PatchTST has regrets between 0.007 and 0.15 across the five tasks, and DiffSTG between 0.015 and 0.24. LastDay is almost indistinguishable from the reference policy on capacity planning, energy saving, and resource allocation, at 0.00, $-0.002$, and about zero, but falls behind PatchTST on warning, 0.173 against 0.119, and on TSTR, 0.28 against 0.15: repeating the previous day is enough for decisions that depend on average structure, but not for tasks that depend on events or on learning. On this track, CLE reaches a rank correlation of $+1.00$ with both energy-saving and resource-allocation regret, against $+0.47$ for paired RMSE, although these correlations rest on only four models.

### A.15 Figure and Table Details

**Figure 2.** (a) $\mathrm{SAD}^{\star}$ versus SDR for the 24 generation evaluations; cities have distinct colors and markers. The dashed line marks uniform allocation ($\mathrm{SAD}^{\star} = 1$). Paired labels count 18 runs below SDR $0.05$ (shaded) and 6 at or above it. (b) Spearman correlation with SDR, ordered by absolute magnitude, with $\mathrm{SAD}^{\star}$ highlighted. (c) Three-city mean SDR (filled markers) and individual city values (open markers), grouped by track; the dashed line marks observed dispersion. In (a) and (c), dotted lines mark SDR $= 0.05$. SDR axes are linear below $10^{-5}$ and logarithmic above it.

**Table 2.** Results are three-city means per model. SDR measures preserved regional dispersion, where 1 matches the observations; $\mathrm{SAD}^{\star}$ is spatial allocation error relative to the uniform-allocation reference, where 1 equals uniform allocation. PENTAD scores and decision regrets are lower-is-better; energy and warning regrets use range normalization, where 1 matches inaction. RMSE is computed against the mean profile for generation and on paired horizons for forecasting, so the two tracks are not comparable in this column. Very small values are written in e-notation. Src: AR author-released code, PF paper-based reproduction, BN benchmark-native baseline. One seed per configuration.

**Figure 3.** (a) Three-city mean PENTAD scores for eight generation and four forecasting models. Each column is divided by its maximum within its track; longer, darker bars indicate larger relative error, and pale tracks indicate the full 0–1 range. Colors are comparable only within a perspective and track, not across them. The spatial column uses $\mathrm{SAD}^{\star}$, as in Table 2. (b) City-averaged responses to five RMSE-matched perturbations, using raw SAD (km). Numbers show targeted responses; unlabelled off-target responses are at most $2.1 \times 10^{-9}$, versus a minimum targeted response of $3.0 \times 10^{-3}$. Colors use a separate logarithmic scale with a floor of $10^{-9}$.

**Figure 4.** (a) Energy-saving regret for all 24 generation evaluations, sorted in ascending order; each point is one model–city evaluation. The dashed line marks inaction (regret 1): 5 evaluations improve on it, 16 tie exactly, and 3 perform worse. (b) Within-city Spearman correlations between RMSE or a PENTAD score and each of the five decision regrets on the generation track, averaged over the three cities; the color scale is shown at right. Dashed hollow cells mark correlations whose sign differs across cities; their means are still reported. SAD and $\mathrm{SAD}^{\star}$ have identical within-city ranks.

Table 12: The five decision tasks and their operating points. Regret is normalized by the reference cost for G1, G3, and G5 (Equation 24) and by the cost range between inaction and the reference for G2 and G4 (Equation 25).

| Task | Decision rule | Loss and reference | Normalization |
|---|---|---|---|
| G1 Capacity planning | provision each region at the 0.95 quantile of modeled demand | $\lambda_{\mathrm{u}}(y-c)_+ + \lambda_{\mathrm{o}}(c-y)_+$ with $\lambda_{\mathrm{u}}/\lambda_{\mathrm{o}} = 19$; the same quantile rule on observed demand | reference |
| G2 Energy saving | sleep a region-hour when normalized load is below $1/4$ | energy saved minus SLA penalty, penalty weight 4; cellwise minimizer on observed demand | range |
| G3 Resource allocation | select the top $\lceil 0.10R \rceil$ regions per hour | uncovered peak demand; observed top-$k$ regions | reference |
| G4 Congestion warning | alarm when modeled demand exceeds the observed 0.90 quantile | misses weighted by exceedance, false alarms at a fixed cost; observed exceedances | range |
| G5 Synthetic-data utility | ridge next-day forecaster trained on model output | forecast error on real data; ridge forecaster trained on real data | reference |