ANCHOR BEFORE YOU LEARN: ZERO-PARAMETER CLIMATOLOGY BEATS LEARNED NORMALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Trainable per-instance normalization is the usual answer to non-stationary demand, and it spends every origin relearning a periodic level profile that one grouped average over the training split hands back exactly. Put a unit-by-hour-of-week climatology table in that slot, pass the network only the residual, then add the table back. The alternative costs nothing. Fix one causal origin-frozen rolling table before any test number, keep it across four backbones from 27k to 682k parameters, five demand systems in two cities and five horizons, run both sides to convergence, and all 100 cells fall to it against the strongest learned normalizer, by 10.55 to 74.26 percent of MAE, median 36.19, each cell clearing a five-seed screen, surviving Holm correction, ahead on RMSE. Inside it sits the frequency-adaptive normalizer built around periodicity, beaten in 100 of 100 at a median 44.67 percent and the pool's best member in 23. Measured against a threshold fixed before the second city was scored, 39 of its 40 cells reproduce and none fails. Independence is thinner than the count implies, five overlapping panels resting on two lattices and one split design. Grade each head on its own table and almost no credit remains. Under the pinned rolling table the 682k patch transformer returns 1.20 percent over fifteen long-horizon cells, the 27k recurrent head 0.21, two of four below nothing. A cost-free rolling climatology keeps 38 of the 100, the other 62 worth a median 0.56 percent. Work out the climatology floor before any normalization block joins a periodic-data comparison.

1 INTRODUCTION

Parameters, gradients and inference-time bookkeeping inside a trainable per-instance normalizer all go on a quantity that a database query returns for nothing. Each Manhattan taxi zone, each Chicago community area and each subway station complex holds Tuesday-morning pickups at a characteristic level. Fitting is not what recovers it. Averaging per unit within hour-of-week produces it outright on the training split, and it accounts for 89.9 percent of New York taxi variance out of sample, 94.4 of subway and 90.9 of ride-hail, while the trainable schemes chase the same quantity across one short window after another (Kim et al., 2022; Fan et al., 2023; Ye et al., 2024). Set the two in one slot and the asymmetry becomes measurable.

Fix a zero-parameter climatology table before any test number has come into view, carry it untouched over four backbones, five demand systems in two cities and five horizons, and it ends ahead of the strongest learned per-instance normalizer on its own backbone in all 100 cells, spanning 10.55 to 74.26 percent MAE about a median of 36.19, each of the 100 clearing a five-seed variance screen, holding after Holm–Bonferroni correction over the grid, ahead on pooled RMSE too. Nothing shifts when the opponent sharpens. Frequency-adaptive normalization, the family member built for periodic non-stationarity (Ye et al., 2024), concedes all 100 cells at a median of 44.67 percent and heads the learned pool in merely 23. One city does not own the outcome either. Chicago taxi and ride-hail at community-area level, every protocol constant held at its New York setting, return 39 reproductions, one weaker cell and no failure at all against a bar committed to before scoring, though Chicago's margins do come out smaller, a median of 32.69 percent beside New York's 38.51.

Point that observation inward and bookkeeping is what it becomes. The matched-anchor residual gain $R(h)$ scores every network against the frozen table it carries, and the choice of reference dictates the answer. Under the causal origin-frozen rolling-8 table behind every headline number,

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.

fifteen long-horizon cells per backbone hand 1.20 percent to the 682k patch transformer, 0.21 to the 27k recurrent head, -0.96 to the 57k linear decomposition and -4.36 to the 309k inverted transformer; swap in the static table, which none of these arms deploys, and three of the four go positive. Referenced to the table a practitioner would ship, swapping the interface strips a median 36.19 percent off the learned arm's error while the residual head restores next to none of it.

Where the finding gives out is reported as precisely as where it holds. Lift the model-free climatology predictors out of the straw-man row and seat them as genuine arms, and cells begin going their way. A parameter-free origin-frozen rolling climatology claims 38 of the 100, while the 62 the trained head keeps are worth a median of 0.56 percent MAE, with 43 of them above seed noise beside 24 of the 38 losses. Convergence raised that tally from 52 of 100 at the published cap and shifted the boundary not at all; at day-ahead the 27k recurrent head hands two of the five systems to a predictor holding no parameters. That floor is not a competing interface but SCAN minus its residual head, and because the headline is confined to learned normalizers the two facts bound one another rather than collide. The measurement protocol that exposed the asymmetry is what survives. Our contributions:

- **A comparison protocol for per-instance normalization, and the zero-parameter instrument that runs it.** The per-unit hour-of-week climatology floor belongs in any comparison on periodic data, computed and seated as an arm before a normalization block joins it. Seasonal Climatology ANchoring (SCAN) fills that role: take the climatology out, fit what remains, put the climatology back. Its causal origin-frozen variants read nothing dated later than the forecast origin, learn nothing and take no per-system tuning, which is what lets one unaltered table serve every cell.
- $R(h)$**, matched-anchor residual accounting, reported per anchor.** Grade a trained head on the very table it was frozen onto and "the anchor or the network?" stops being rhetorical and turns into a measurement. Inside one backbone this is a clean deflation test; over four heterogeneous ones it fixes no ordering, and we report it that way. Its reference has to be the table the recommended configuration ships.
- **The benchmarking result, on evidence that is selection-free, budget-controlled and replicated in a second city.** One table pinned ahead of time over all 100 cells; a learned pool broadened to take in the periodicity-aware normalizer; eight validation-selected configurations per learned arm, swept on the three New York systems, set against the two ours chooses between; three model-free climatology floors seated as first-class arms, the 38 cells they take printed beside the rest; a replication threshold set before the second city was scored; and two validation protocols whose disagreement goes into print.

2 Related Work

2.1 Learned per-instance normalization

One reply to drifting input distributions has held for years. Normalize each instance, then restore its statistics downstream. It started as a preprocessing step (Ogasawara et al., 2010) and set into an architectural block along the normalize-then-restore pattern of batch normalization (Ioffe & Szegedy, 2015), with reversible instance normalization taking a window mean and variance off at the encoder for the decoder to replace (Kim et al., 2022). Several lines of work have since loosened the premise that a single window captures the shift: Dish-TS splits input from output distributions via coefficient networks (Fan et al., 2023), frequency-adaptive normalization works in the spectral domain (Ye et al., 2024), contrastive objectives pull drifting factors apart from stationary ones (Pang et al., 2027), and series stationarization hands attention statistics back to the model interior (Liu et al., 2022). This paper runs the spectral member rather than merely citing it (Section 4.3); its published gains sit on other benchmarks and go uncontested here. Long-horizon architectures carry the block as standard (Nie et al., 2023; Wu et al., 2023; Liu et al., 2024; Luo & Wang, 2024; Wang et al., 2024b), and so do applied photovoltaic and decomposition-based forecasters (Wang et al., 2023; 2024a); two of our backbones embed it, so every arm has it removed and the slot handed to whatever is under test. Throughout that literature a learned normalizer meets either a second learned normalizer or an unnormalized control, never a fixed seasonal profile estimated on the training split, precisely what the block pursues implicitly here.

2.2 CLIMATOLOGICAL REFERENCE FORECASTS

In meteorology skill is scored against a reference forecast, climatology by convention, on the unglamorous ground that beating nothing establishes nothing (Murphy, 1988; Murphy & Epstein, 1989). Forecasting picked up that habit through scale-free measures and seasonal naïve references (Hyndman & Koehler, 2006); it passed into deep probabilistic and basis-expansion models (Oreshkin et al., 2020; Salinas et al., 2020), and it sharpens wherever elaborate learned methods fail to displace plain statistical references (Makridakis et al., 2018; 2020). Meanwhile compact models have drawn level with far larger ones on long-horizon benchmarks (Zeng et al., 2023; Xu et al., 2024), while the scaling branch of that literature goes on finding capacity that buys accuracy (Shi et al., 2025) and a recent deconstruction finds no monotone link between size and quality (Wen et al., 2026). Seasonality is no secret in urban demand research either, where a stable diurnal and weekly signature runs through subway ridership (Chen et al., 2009; He et al., 2018) and hour-of-week structure through taxi and ride-hail demand (Liao et al., 2018; Grahn et al., 2020; Wang & Noland, 2021). Deep forecasting papers nevertheless park the seasonal reference in a straw-man row at the foot of a table (Hsu & Chen, 2020; Wang, 2024), denying it what meteorology grants, an arm allowed to win, and reported when it does.

2.3 EVALUATION PRACTICE IN URBAN DEMAND FORECASTING

The third gap is procedural. Architectural elaboration in spatio-temporal demand work has outpaced the evidentiary standard it answers to, a record Appendix K follows through the model lineage, the one-metric one-horizon one-run habit beneath most comparisons, and fixes that neighbouring fields settled long ago. Benchmarking reform is still unfinished (Hahn et al., 2023), and the exemplars nearest our setting either print bare point estimates or banish multi-seed dispersion to an appendix. Section 3.3 responds by declaring comparator budgets explicitly and fixing both the selection rule and the replication bar before the test grid is examined.

3 SCAN: SEASONAL CLIMATOLOGY ANCHORING

3.1 ANCHORS AND RESIDUAL-SPACE LEARNING

Write $y_{u,t} \in \mathbb{R}_{\geq 0}$ for demand at spatial unit $u \in \{1, \dots, U\}$ during hour t, and $\phi(t) \in \{0, \dots, 167\}$ for the hour-of-week index. Given a window $\mathbf{y}_{u,t-L+1:t}$ of $L = 168$ hours the target is $y_{u,t+h}$, and every arm emits 168 hours at once. Average by unit and by hour-of-week over $\mathcal{T}_{\text{train}}$ and the *static anchor* results,

$$a_u^{\text{stat}}(\tau) \;=\; \frac{1}{\left|\{t \in \mathcal{T}_{\text{train}} : \phi(t) = \tau\}\right|} \sum_{t \in \mathcal{T}_{\text{train}},\, \phi(t)=\tau} y_{u,t}, \qquad \tau \in \{0, \dots, 167\}. \tag{1}$$

a $U \times 168$ table built once, carrying no fitted parameter. Deployment also wants a table that follows slow level drift without looking forward, and the *origin-frozen rolling anchor* delivers that by averaging the W weeks preceding the origin and then freezing the table for the whole prediction,

$$a_u^{\text{roll}}(\tau \mid t) \;=\; \frac{1}{W} \sum_{k=1}^{W} y_{u,\, t-\delta(t,\tau)-168(k-1)}, \qquad \delta(t,\tau) = \big(\phi(t) - \tau\big) \bmod 168, \tag{2}$$

with $W = 8$ throughout. The freeze at t buys causality, since one table covers every horizon from that origin and nothing dated after t reaches it. Zero learned parameters is not zero design choices. Hour-of-week granularity and $W = 8$ weeks were both fixed before the grid ran and went into all 100 cells unswept, so nothing reported rests on a tuned W and nothing shows 8 weeks to be best.

Let a denote whichever anchor is in force. The interface is three lines: take the anchor off the window, pass the residual to any backbone f_θ, put the anchor back at the target time.

$$r_{u,s} = y_{u,s} - a_u(\phi(s)), \qquad \hat{r}_{u,t+h} = f_\theta(\mathbf{r}_{u,t-L+1:t})_h\,, \qquad \hat{y}_{u,t+h} = a_u(\phi(t+h)) + \hat{r}_{u,t+h}. \tag{3}$$

Training minimises $\mathcal{L}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{(u,t)\in\mathcal{B}} \sum_h |\hat{r}_{u,t+h} - r_{u,t+h}|$ across a minibatch $\mathcal{B}$ of unit-origin pairs; the anchor cancels, so this is absolute error in demand units. Figure 1 places the interface next to the block it replaces along one path, and that contrast is the whole method. A learned normalizer

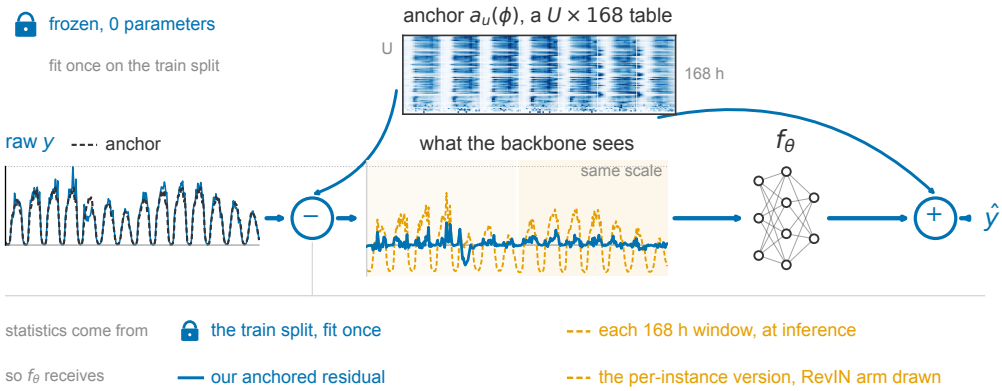


Figure 1: The anchoring interface, on real data. The frozen hour-of-week anchor a_u lies dashed over the raw series, comes off at the minus operator and returns at the plus one, so f_θ sees the residual alone. One panel holds both methods on a single vertical scale. Ours is flat; the dashed trace, what per-instance normalization hands the same backbone, keeps the weekly cycle. Of RevIN, Dish-TS and frequency-adaptive normalization the arm drawn is RevIN-style per-window location and scale, and only that slot differs between arms. Curves are New York taxi at its busiest Manhattan zone over the median-dispersion Monday-aligned fortnight of the training split.

reads location and scale off the input window at inference and undoes them at the output, while the anchor is a lookup decided before training starts.

No constraint falls on f_θ from Equation 3, so one unaltered anchor stands equally ahead of a recurrent model, a linear one or a transformer, and a table covering 69 zones, 77 community areas or 424 complexes stays the same kind of object. History at the unit is what it demands – that unit's training split for Equation 1, eight weeks of it at the origin for Equation 2 – so a new station, a re-drawn zone or a short-record unit has no anchor at all, while every learned normalizer measured against it here takes its statistics off the input window instead (Appendix M).

3.2 Matched-anchor residual gain

The anchor costs nothing, so the question is whether a trained head improves on the anchor it was handed. Grading against a *different* seasonal reference would blend two accounts of seasonality, so every network is scored on its own frozen table over one shared set of test origins:

$$R(h) \;=\; \frac{\mathrm{MAE}_a(h) - \mathrm{MAE}_{f_\theta \circ a}(h)}{\mathrm{MAE}_a(h)} \times 100, \tag{4}$$

where $\mathrm{MAE}_a(h)$ is the anchor's own error and $\mathrm{MAE}_{f_\theta \circ a}(h)$ belongs to the anchored network built atop it. A positive $R(h)$ says the parameters improved on the table they received; an $R(h)$ near 0 says they did not. The result is meteorology's skill score (Murphy, 1988) with its reference pinned to the model's own climatology, turning a margin into an accounting. Every $R(h)$ aggregate below obeys one rule. A cell is a single system-backbone-anchor-horizon combination carrying a five-seed mean over one shared origin set, and an aggregate is the unweighted mean of such cells, never a pooling over origins or seeds. That makes 25 cells per backbone within each anchor type, 15 once $h \in \{24, 72, 168\}$ is imposed. Anchor types stay separate, because a pooled mean folds in a static-anchor reference the recommended configuration never ships.

Inside one backbone $R(h)$ is a clean deflation test, because arms scored on the same frozen table differ only in what their parameters drew out of the same residual. Over backbones it is nothing of the sort, since the four networks of Section 5.1 differ in tokenization, architecture family, covariates and decoding as much as in parameter count, so no ordering $R(h)$ yields there isolates capacity, and nowhere do we claim otherwise.

3.3 A selection-free evidence design

The anchor is fixed a priori. Every headline number comes from the causal origin-frozen rolling-8 variant of Equation 2, identical over all 100 cells. Refit at each origin, it assumes nothing

about stationarity across a test window running months past the training split, which makes it the deployable variant and the pinned one at once; no reported number rests on an anchor chosen per cell, per system or per backbone.

One configuration is recommended; its single free choice is settled on validation. A grid helps no practitioner, so we name one configuration, the fixed rolling-8 anchor under a PatchTST residual head. Its only degree of freedom is the backbone, decided by a rule written down beforehand – within each system, take seed-averaged MAE at $h{=}24$ over that system's static-climatology-floor MAE at $h{=}24$ on the same block, average unweighted across systems, lowest wins, all of it on the published block adjacent to training. Nothing from test enters, and the static floor serves as denominator because it alone is untouched by an arm's own anchor choice.

Model-free floors are arms, and the comparator receives two budgets. Three deterministic predictors stand level with any trained arm: the static anchor by itself, the origin-frozen rolling anchor by itself, and a seasonal naïve predictor that repeats a week-old value – all three on the headline table's rows, their wins printed as plainly as their losses. The comparator is whichever of RevIN, Dish-TS and frequency-adaptive normalization does better on the identical backbone, taken per cell on test MAE so no baseline ends up understated (Ferrari Dacrema et al., 2019), then swept over 1, 2, 4 and 8 validation-selected configurations on the three New York systems (Section 4.5).

The replication bar precedes the second city. A Chicago cell replicates when its margin is positive, clears combined seed noise, and reaches at least half the New York median margin in the matching backbone-by-horizon cell; positive and screened but under that bar is weaker; non-positive, or inside seed noise, is a failure. The fraction and the three categories alike were settled before any Chicago test number was read.

The verdict rule. A win requires three conditions together: a mean margin above zero, a five-seed variance gate passed, $|\bar{m}| > \sqrt{\sigma_A^2 + \sigma_B^2}$, and a week-block bootstrap 95 percent interval clear of zero. Here A is the anchored arm and B its comparator in the same cell, $\bar{m}$ the mean margin of A over B, and σ_A, σ_B the population standard deviations of their MAE over seeds 42 to 46. The gate screens, it does not test, and it runs roughly 12 percent laxer than a reader assuming a sample standard deviation would expect (Appendix M). Overlapping test origins destroy the independence a Wilcoxon signed-rank test assumes (Diebold & Mariano, 1995), so those p-values remain descriptive while inference rests on block-bootstrap intervals (Künsch, 1989), on the gate, and on paired seed-level tests Holm-corrected over the grid (Holm, 1979). The word significant appears nowhere without a corrected test that passed on the quantity named.

4 Experiments

4.1 Setup

Data. Five demand systems across two cities ride on one interface (Table A1): in New York, yellow-taxi pickups over 69 Manhattan taxi zones, for-hire vehicle (Uber and Lyft) pickups over that same lattice, and subway entries at 424 station complexes; in Chicago, taxi and ride-hail pickups over 77 community areas. Each panel spans 17544 hours, splitting is chronological with a validation block between train and test, and test origins stride 6 hours apart from the subway, whose stride is 24. Independence is weaker than the count of five suggests. A single 69-zone lattice serves New York taxi and ride-hail and a single 77-area lattice both Chicago systems, split boundaries and the test calendar are shared within each city, and every one of the five horizons comes out of one 168-step forecast – 100 comparisons over five overlapping panels, not 100 experiments. A single zero-versus-missing inclusion rule, set out in Appendix A and applied identically to every arm, fixes which unit-hours enter each anchor mean and each metric.

Arms. Nothing varies but the normalization interface. Anchored arms get the static or the origin-frozen rolling-8 table; learned arms get reversible instance normalization (Kim et al., 2022), Dish-TS (Fan et al., 2023) or frequency-adaptive normalization (Ye et al., 2024); an unnormalized control and three model-free floors round out the set, and the whole crosses with the four backbones of Table A1 (Cho et al., 2014; Nie et al., 2023; Zeng et al., 2023; Liu et al., 2024) at $h \in \{1, 6, 24, 72, 168\}$ hours. The spectral arm's one structural hyperparameter, the number K of retained dominant frequency components, has no defensible universal default, so it is picked per cell on the adjacent validation

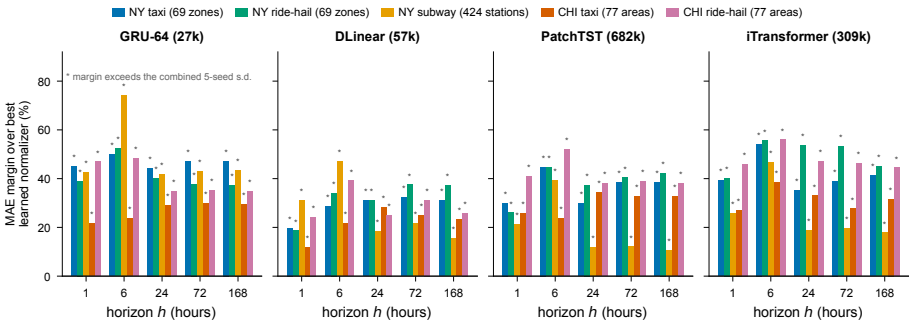


Figure 2: MAE margin of the zero-parameter climatology anchor over whichever of RevIN, Dish-TS and frequency-adaptive normalization fares better on the matching backbone, read at the converged epoch budget across four backbones, five demand systems in two cities and five horizons. A single fixed origin-frozen rolling-8 table spans all 100 cells, not one margin falls below zero, and an asterisk on a cell marks a margin exceeding the combined five-seed standard deviation. Higher is better.

block from $\{3, 5\}$ crossed with the two learning rates – four configurations beside the two every other arm gets, an asymmetry leaning towards the comparator, the only safe direction. Defined in the released configuration but never swept is the wider $\{2, 3, 5, 8\}$ bracket (Appendix D). Arms on a shared backbone share hyperparameters, splits, input length, origin sets and seeds 42 to 46; each dispersion below is a population standard deviation over those five runs; and all of it ran on one NVIDIA TITAN RTX (23.6 GB) under torch 2.2.0 with CUDA 12.1. Parity ends at the backbone boundary, where calendar covariates go to the recurrent head alone, which alone decodes autoregressively (Table A2).

Provenance of the comparators. RevIN alone is the released layer taken verbatim; our adaptation with a shared two-layer coefficient head supplies the Dish-TS arm; the frequency-adaptive arm is our reimplementation from the paper and its released reference code, auxiliary loss held at the reference weight of 1.0; and both transformer backbones we rebuilt. Not one of them is calibrated against a published benchmark number, so telling a weak normalizer from a weak rebuild rests on our description and released code (Appendix A). The two validation protocols pick different arms, for a structural reason laid out in Appendix I.

4.2 The advantage over learned normalization

Table 1 sits at Section 4.5's converged budget, every arm trained under a cap six times the published one and stopped by patience. Nowhere does selection enter. Over a grid four backbones by five systems by five horizons wide, the single anchor in use is Equation 2's origin-frozen rolling-8 climatology. All 100 cells return a positive margin against the best learned per-instance normalizer on their own backbone, from 10.55 percent for PatchTST at New York subway $h{=}168$ up to 74.26 for the recurrent head at subway $h{=}6$ about a median of 36.19, all 100 clearing the five-seed variance gate, holding under Holm–Bonferroni correction over the grid, and ahead on pooled RMSE (Figure 2). At the published cap the same grid runs 13.50 to 78.54 percent about a median of 37.62, again 100 of 100.

No single backbone, system or city owns the effect. Backbone medians come to 28.30 percent for DLinear, 37.37 for PatchTST, 40.30 for iTransformer and 41.66 for the recurrent head; system medians come to 23.80 on the New York subway, 27.98 on Chicago taxi, 38.81 on New York taxi, 38.93 on Chicago ride-hail and 39.59 on New York ride-hail, with the New York 60 at 38.51 and the Chicago 40 at 32.69. A backbone we rebuilt ourselves holds one of the largest medians, and the arm it beats most often is our own Dish-TS adaptation, which puts some of the grid's widest margins on the least externally verified code here. Both transformers ship a learned normalization block of their own and both ran with that block removed, turning the test into a contest between two tenants of a slot their designers built. On the static anchor a second selection-free view returns 100 of 100 positive and screened (Appendix B).

The first two columns of Table 1 hold what a practitioner would actually deploy, the fixed rolling-8 anchor under the PatchTST residual head that Section 3.3's pre-declared rule selects. Across its 25 cells the margin over the best learned normalizer runs 10.55 to 51.98 percent about a median of 37.37, all 25 screened and ahead on RMSE (Table B1). At the decision cell nominated in advance, day-ahead New York taxi, it returns 11.496 MAE where Dish-TS returns 16.395, a cut

Table 1: Selection-free headline grid at the converged epoch budget, carrying one fixed anchor, the causal origin-frozen rolling-8 climatology, across all 100 cells. Columns 1-2 hold the recommended configuration and the best learned normalizer on the same backbone over seeds 42-46. Every margin block gives the margin over that best learned normalizer next to the margin of the *same* arm in the *same* cell over the free rolling climatology floor, whose 38 negative entries mark cells a parameter-free groupby forecasts better.

	Recommended arm vs. best learned		△MAE % vs. best learned				△MAE % vs. rolling floor			
h	SCAN+PatchTST (↓)	Learned (↓)	GRU	DLin	iTrans	Patch	GRU	DLin	iTrans	Patch
Taxi (New York)										
1	**10.293** ± 0.175	RevIN 14.70 ± 1.12	+45.01	+19.78	+39.30	+29.96	+6.10	+19.11	+7.03	+14.46
6	**11.353** ± 0.022	RevIN 20.56 ± 0.92	+49.89	+28.50	+54.21	+44.79	+1.65	+6.71	+1.21	+3.35
24	**11.496** ± 0.184	Dish-TS 16.39 ± 1.18	+44.33	+31.18	+35.16	+29.88	+1.52	+0.68	-5.33	+2.72
72	**11.910** ± 0.133	Dish-TS 19.40 ± 1.81	+47.13	+32.36	+39.01	+38.60	+0.59	-3.63	-13.88	+0.32
168	**12.337** ± 0.231	Dish-TS 20.04 ± 1.52	+47.19	+30.99	+41.26	+38.43	+0.53	-1.19	-6.37	-1.10
Ride-hail (New York)										
1	**19.285** ± 0.464	RevIN 26.07 ± 2.02	+38.98	+18.58	+40.30	+26.02	+5.64	+23.83	+11.36	+17.29
6	**21.246** ± 0.102	Dish-TS 38.32 ± 1.23	+52.50	+34.05	+55.52	+44.56	+0.93	+8.68	+4.25	+6.32
24	**22.173** ± 0.285	Dish-TS 35.40 ± 1.12	+40.21	+30.97	+53.70	+37.37	+1.06	-1.19	-9.26	+2.49
72	**23.543** ± 0.218	Dish-TS 39.46 ± 1.04	+37.80	+37.45	+53.20	+40.33	+0.36	-3.98	-6.12	-1.91
168	**24.664** ± 0.662	Dish-TS 42.74 ± 3.98	+37.27	+37.11	+44.95	+42.29	-0.37	-4.90	-7.66	-5.13
Subway (New York)										
1	**13.511** ± 0.348	RevIN 17.17 ± 1.19	+42.59	+31.08	+25.89	+21.32	-1.61	+18.88	-3.95	+11.81
6	**14.075** ± 0.338	RevIN 23.16 ± 1.24	+74.26	+47.09	+46.76	+39.22	-1.12	+5.45	-9.67	+2.47
24	**22.859** ± 0.275	RevIN 25.90 ± 0.38	+41.66	+18.30	+18.93	+11.76	-0.26	+8.13	-1.06	+8.53
72	**24.290** ± 0.522	RevIN 27.61 ± 0.46	+42.94	+21.71	+19.52	+12.02	-0.26	+3.92	-2.36	+4.00
168	**25.372** ± 0.385	RevIN 28.36 ± 1.29	+43.58	+15.45	+17.79	+10.55	-0.23	+3.49	-2.65	+1.49
Taxi (Chicago)										
1	**2.496** ± 0.048	FAN 3.36 ± 0.24	+21.71	+11.93	+27.13	+25.67	+1.26	+17.49	+9.09	+9.05
6	**2.610** ± 0.021	Dish-TS 3.42 ± 0.19	+23.70	+21.63	+38.43	+23.60	+0.14	+5.73	+3.01	+2.82
24	**2.587** ± 0.025	FAN 3.93 ± 0.09	+29.23	+28.30	+32.95	+34.24	+0.05	+4.03	+3.90	+4.49
72	**2.727** ± 0.040	FAN 4.05 ± 0.18	+29.86	+24.90	+27.66	+32.67	+0.10	-3.91	+0.05	+0.30
168	**2.770** ± 0.024	FAN 4.12 ± 0.24	+29.39	+23.33	+31.67	+32.70	+0.14	-4.66	-0.39	+0.11
Ride-hail (Chicago)										
1	**15.497** ± 0.157	Dish-TS 26.28 ± 1.52	+46.90	+24.27	+46.04	+41.03	+3.34	+19.38	+9.56	+14.48
6	**17.064** ± 0.099	Dish-TS 35.54 ± 2.05	+48.20	+39.11	+55.94	+51.98	+0.11	+5.98	+3.12	+4.38
24	**17.670** ± 0.137	Dish-TS 28.55 ± 1.74	+34.71	+25.05	+47.19	+38.12	-0.05	-0.64	-1.78	+2.28
72	**18.614** ± 0.170	Dish-TS 30.40 ± 1.65	+35.27	+31.26	+46.19	+38.76	+0.02	-5.52	-6.25	-0.37
168	**19.059** ± 0.290	Dish-TS 30.87 ± 1.36	+34.78	+25.91	+44.45	+38.27	-0.02	-5.04	-6.21	-0.25
Cells won (of 25)			**25**	**25**	**25**	**25**	17	15	10	20
Median margin %			+41.66	+28.30	+40.30	+37.37				

of 29.88 percent. Since beating a learned normalizer weak in absolute terms counts for little, the parameter-free predictors score these cells too. A lag-168 copy of last week forecasts better than the best learned arm in most of the 100 and still drops all 100 to the anchored rolling arm by a median 17.52 percent – at day-ahead New York taxi under the recurrent head the best learned arm is the spectral one at 20.902 MAE, against 13.793 for the one-week copy.

4.3 THE PERIODICITY-AWARE COMPARATOR

An obvious rejoinder says that RevIN and Dish-TS take away an instance mean and scale, the wrong statistic to take from a periodic series, and that a normalizer built around periodicity would behave otherwise. Frequency-adaptive normalization is that normalizer, and in place of the first two moments it strips the K dominant Fourier components off the input window and carries them forward through a learned head (Ye et al., 2024). It occupies the slot in all 100 cells, with K chosen per cell on adjacent validation and a default configuration budget larger than any other arm's.

The anchor takes 100 of 100 cells, every one of them above the five-seed gate, at a median of 44.67 percent MAE and by 18.25 in the narrowest cell of the grid; system medians run 41.47 on New York taxi, 43.65 on the subway, 54.62 on New York ride-hail, 32.83 on Chicago taxi and 50.39 on Chicago ride-hail. Read a second way, those cells leave the spectral arm at the head of the learned pool in 23 of them and a median 13.93 percent behind whichever of RevIN and Dish-TS fares better (Figure D1, Appendix D). That ordering reverses what its source paper reports on its own benchmarks, and this arm is our rebuild. A weak normalizer and a faulty reimplementation leave one signature between them, nothing here tells them apart (Appendix M), and so the inversion goes on record while no inference about the normalization family is drawn from it. Which reading holds would be settled by a single run of the rebuilt arm against its published number on a benchmark from that paper. The headline waits on no such answer, since the RevIN layer, taken verbatim from its authors, also loses all 100 cells, at a median of 38.45 percent.

4.4 A SECOND CITY

With three systems inside one city the finding could still be a property of New York demand, so the grid runs again on Chicago taxi and Chicago ride-hail at community-area level, every protocol constant pinned to its New York value and the city the only thing altered – a harder test on its face, since a community area carries 9.6 taxi trips in the average hour where a Manhattan zone carries 58.4. The

bar was written down before those cells were scored (Section 3.3), and one grid scores both cities, anchor pinned and spectral comparator present on either side, so the comparison runs like-for-like.

Thirty-nine of the 40 cells reproduce the result, one lands weaker and none fails (Figure C1, Appendix C), which is 19 of 20 on Chicago taxi, 20 of 20 on Chicago ride-hail, all four backbones covered. The weaker cell is Chicago taxi with the recurrent head at $h{=}6$, whose margin of 23.70 percent is positive and above seed noise yet falls short of the 26.25 percent required there, half of New York's 52.50 in that cell. Where the bar was put is what the 39 depends on, and that dependence belongs in print beside it. Chicago taxi with the recurrent head at $h{=}1$ clears its required 21.30 percent with 21.71, so a bar of 0.75 leaves 29 of 40, the observed median ratio of 0.849 leaves 20 and parity with New York leaves 13. No bar at all is needed for the statement that all 40 cells are positive, seed-gated and Holm-significant. Chicago's margins are narrower, a median of 32.69 percent against New York's 38.51 under this like-for-like reading, a ratio of medians of 0.849, the median of per-cell ratios also 0.849. We read that as thinner demand, offered on two cities of evidence as an observation and not a finding, since the denser Chicago system, ride-hail at 122.9 trips per area-hour, records a median of 38.93 percent, within a point of New York taxi. Scored instead against a New York reference column drawn from the three New York systems alone, its learned pool holding RevIN and Dish-TS with the spectral arm absent and its margins larger in 14 of the 20 backbone-by-horizon cells, Chicago gives 38 reproductions, 2 weaker cells and no failures on that stiffer reference. Transfer across data-collection regimes is what none of this establishes, because one provider class, one aggregation convention and one split design cover both cities, so what has been tested is a second city and not a second way of making the data.

4.5 Two budgets for the comparator

Behind a per-cell best learned arm stand one hyperparameter setting and one training budget, so the comparator gets both: up to 8 validation-chosen configurations against the two SCAN chooses between, swept on the three New York systems, and the whole 100-cell grid run twice, at the published epoch cap and at six times it. Truncation was real and not decisive. Under both budgets 100 of 100 margins stay positive and screened, the median settling from 37.62 percent to 36.19, none vanishing and none flipping sign (Appendices H and L, Table L1).

5 What the Parameters Buy

5.1 The capacity ladder

Asking what the parameters contributed means grading every trained head on the frozen table it received, and the reference chosen decides the answer. Table E1 holds that accounting at the converged budget, 25 cells per backbone within each anchor type and 15 of them past a day. Measured on the causal origin-frozen rolling-8 table – the one under every headline number and inside the recommended configuration – only the 682k patch transformer ends above the line, at a mean of 1.20 percent over 10 of its 15 cells, 6 positive after Holm correction beside 2 negative. The 27k recurrent head, at 0.21 percent with 2 corrected positives and no corrected negatives, reproduces the table handed to it, while the 57k linear decomposition lands at -0.96 percent with 9 corrected negatives and the 309k inverted transformer at -4.36 with 11. Swapping in the static table raises three of the four, DLinear as far as $+4.33$ and PatchTST as far as $+4.89$, which is exactly why a figure pooled over anchor types describes nothing any reported arm does. One hour out every head earns something, from 2.95 percent on the recurrent arm up to 19.74 on DLinear; a day out the credit has gone (Figure E1).

At the heart of this lie two denominators, and both deserve naming. Swapping the interface strips a median 36.19 percent off the *learned arm's* MAE, while the strongest residual head returns 1.20 percent of the *anchor's* MAE, itself about half the learned arm's. Past a day the recommended configuration takes roughly one percent from its parameters and all the rest from its interface. Four backbones fix an ordering, not a law. Under the recommended table 27k and 682k stand above zero while 57k and 309k drop below, tokenization, architecture family, covariates and decoding all shifting down those rows with no threshold locatable among them – the same missing monotone link between size and quality that a recent deconstruction of generic transformer baselines reports (Wen et al., 2026). Appendix E carries the accounting cell by cell.

5.2 Where the finding stops: model-free floors as arms

This paper's boundary lives in the floor columns of Table 1 and, cell by cell, in Figure G1 (Appendix G), and convergence leaves it wide open. Against the origin-frozen rolling climatology – a predictor with no parameters, no training and no tuning – the fixed rolling-anchored arms hold 62 of the 100 cells to the free floor's 38, and they hold them by a median of 0.56 percent MAE. Only 43 of those 62 leads rise above seed noise, 24 of the 38 losses do, and 33 of the 100 cells land inside seed noise one way or the other. System by system the arm takes 14, 10, 11, 17 and 10 of 20 on New York taxi, the subway, New York ride-hail, Chicago taxi and Chicago ride-hail; backbone by backbone, in the parameter order of Table 1, 17 of 25 go to the 27k recurrent head, 15 to DLinear, 10 to the inverted transformer and 20 to the patch transformer, of which just 4, 14, 9 and 16 clear seed noise. At the day-ahead horizon it takes 12 of the 20 cells, and there the 27k recurrent head hands two of the five systems to the floor outright, reading $+1.52$, -0.26, $+1.06$, $+0.05$ and -0.05 percent across them. Longer training lifted the tally from 52 of 100 at the published cap and left the boundary precisely where it stood. Our own anchor with everything stripped off it is the hard floor; the milder pair, static climatology and seasonal naïve, surrender 88 of 100 at a median of 9.24 percent and all 100 at a median of 17.52.

One reading fits those 38 cells. That climatology is SCAN minus its residual head, its own zero-parameter ablation and not a rival for the normalization slot, so its record bounds the worth of the *head* rather than of the interface, which is why the claim in Section 4.2 stays confined to learned normalizers. Over its 25 cells the recommended arm leads the rolling floor in 20, the static floor in 23 and seasonal naïve in all 25, its worst loss 5.13 percent at New York ride-hail h=168; and at the nominated cell the head does earn its keep, day-ahead New York taxi giving 11.496 MAE against its free anchor's 11.817 (Table B1).

5.3 Why learned normalization loses

The measurements fit two readings, one far nearer the data. Arithmetic supplies the narrow one. A learned normalizer takes location and scale off a short input window, and for a series whose level swings by an order of magnitude inside a day that is a noisy stand-in for what the training split pins down to a tenth of its variance. Horizon shape agrees, the margin widening near six hours where window and target sit furthest out of phase. Losing is not the same as doing nothing, and the declared unnormalized control settles that. Across the same 100 cells its median MAE is 50.05 beside the best learned normalizer's 26.17, and it takes only 3 of them from that normalizer (Table F1). On these panels the block does substantial real work and is simply beaten by a groupby. From there a broader reading follows that reaches past the evidence. If the block mostly recovers a profile a groupby returns exactly, part of this family's reported progress would be progress at lifting a handicap, recorded by comparisons omitting the arm that removes it for nothing. That goes untested here. All five panels are urban demand under one provider class, one aggregation convention and one split design, in the regime most favourable to the construction, the anchor alone accounting for 89.9 to 94.4 percent of held-out variance and no standard long-horizon benchmark run, which leaves the second reading a hypothesis about other benchmark families rather than a finding (Appendix M). Either way the recommendation costs nothing. Work out the rolling climatology floor before training anything, seat it as an arm, print $R(h)$ beside any improvement claimed.

6 Conclusion and Future Work

Anchor first. Across 100 converged cells four backbones by five systems by two cities by five horizons wide, a climatology table pinned in advance and holding no parameters ends ahead of the strongest learned per-instance normalizer on that same backbone by 10.55 to 74.26 percent MAE, a median of 36.19, all 100 screened over seeds, Holm-corrected and ahead on RMSE; the periodicity-aware normalizer gives up all 100, and a second city reproduces 39 of 40 cells against a bar set beforehand. Past a day the head returns at most 1.20 percent over its own table, while a cost-free rolling climatology keeps 38 of the 100. What generalises is the protocol: work out the floor, seat it as an arm, print $R(h)$ beside any gain claimed.

ETHICS STATEMENT

Public aggregated urban mobility statistics are the only input this study takes. Taxi and for-hire vehicle records arrive aggregated to zone-hours and community-area-hours, subway records to station-complex-hours, so at no point does an individual trip, trajectory or rider attribute enter, and at those aggregations re-identification is neither attempted nor achievable. Demand forecasts shape service allocation, so a forecaster fitted to historical ridership carries historical service patterns forward, inequitable ones among them; our anchor being a readable per-unit table makes that inheritance easier rather than harder to inspect. No human subjects were involved, and the authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

A frozen, SHA-256-manifested artifact set stands behind every number printed here; Table A1 lists the panel hashes and split boundaries. Training ran on one NVIDIA TITAN RTX (23.6 GB) under torch 2.2.0 with CUDA 12.1, cuDNN 8902, Python 3.9.12, numpy 1.26.4, pandas 2.2.0, scipy 1.13.0, Linux 5.4.0. Seeds 42 to 46 served every trained arm, and every dispersion is a population standard deviation (denominator n) over those five runs. Table A2 with the zero-versus-missing inclusion rule in Appendix A is all a reimplementation needs; that appendix also holds the cell, compute and convergence accounting, Appendix J the cross-platform correctness gate. The five-seed rule admits one exception, that in the tuning-budget control of Section 4.5 each learned budget arm is a single run, its SCAN line is the static-anchor arm rather than the origin-frozen rolling-8 arm carrying the headline, and it ran below the published epoch cap. Everything else here is a five-seed mean at the converged budget. Data availability, every input is open, New York panels from the Taxi and Limousine Commission and the Metropolitan Transportation Authority, Chicago panels from the City of Chicago open data portal. Code availability, the anchoring, evaluation and analysis scripts ship with the camera-ready version.

THE USE OF LARGE LANGUAGE MODELS (LLMS)

During the preparation of this work, the authors used ChatGPT 5.5 and Claude Opus 5 for code optimization, proofreading and editing, reformatting, and the preparation of press releases and outreach materials. The tools were used under full human supervision to improve code quality, spelling, grammar, punctuation, clarity, and formatting, and to assist in preparing materials based on content developed by the authors. After using these tools, the authors reviewed and edited all outputs as needed and take full responsibility for the veracity, accuracy, and correctness of the published work.

REFERENCES

Ahmad Ali, Yanmin Zhu, and Muhammad Zakarya. Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flows prediction. *Neural Networks*, 145:233–247, 2022. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2021.10.021. URL https://www.sciencedirect.com/science/article/pii/S0893608021004123.

Christoph Bergmeir, Rob J. Hyndman, and Bonsoo Koo. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83, 2018. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2017.11.003. URL https://www.sciencedirect.com/science/article/pii/S0167947317302384.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In A. Smola, A. Dimakis, and I. Stoica (eds.), *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021. URL https://proceedings.mlsys.org/paper_files/paper/2021/file/0184b0cd3cfb185989f858a1d9f5c1eb-Paper.pdf.

Cynthia Chen, Jason Chen, and James Barry. Diurnal pattern of transit ridership: a case study of the new york city subway system. *Journal of Transport Geography*, 17(3):176–186, 2009. ISSN 0966-6923. doi: https://doi.org/10.1016/j.jtrangeo.2008.09.002. URL https://www.sciencedirect.com/science/article/pii/S0966692308000896.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006. URL https://www.jmlr.org/papers/v7/demsar06a.html.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995. ISSN 07350015. URL http://www.jstor.org/stable/1392185.

Wei Fan, Pengyang Wang, Dongkun Wang, Dongjie Wang, Yuanchun Zhou, and Yanjie Fu. Dish-ts: a general paradigm for alleviating distribution shift in time series forecasting. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i6.25914. URL https://doi.org/10.1609/aaai.v37i6.25914.

Ziquan Fang, Dongen Wu, Lu Pan, Lu Chen, and Yunjun Gao. When transfer learning meets cross-city urban flow prediction: Spatio-temporal adaptation matters. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 2030–2036. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/282. URL https://doi.org/10.24963/ijcai.2022/282. Main Track.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, pp. 101–109, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362436. doi: 10.1145/3298689.3347058. URL https://doi.org/10.1145/3298689.3347058.

Rick Grahn, Corey D. Harper, Chris Hendrickson, Zhen Qian, and H. Scott Matthews. Socioeconomic and usage characteristics of transportation network company (tnc) riders. *Transportation*, 47:3047–3067, 2020. doi: 10.1007/s11116-019-09989-3.

Yannik Hahn, Tristan Langer, Richard Meyes, and Tobias Meisen. Time series dataset survey for forecasting with deep learning. *Forecasting*, 5(1):1–21, March 2023. doi: None. URL https://ideas.repec.org/a/gam/jforec/v5y2023i1p17-335d1087297.html.

Yuxin He, Yang Zhao, and Kwok Leung Tsui. An analysis of factors influencing metro station ridership: Insights from taipei metro. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1598–1603, 2018. doi: 10.1109/ITSC.2018.8569948.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4615733.

Chih-Jung Hsu and Hung-Hsuan Chen. Taxi demand prediction based on lstm with residuals and multi-head attention. In *International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS)*, 2020. doi: 10.5220/0009562000002550.

Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2006.03.001. URL https://www.sciencedirect.com/science/article/pii/S0169207006000239.

Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML'15, pp. 448–456. JMLR.org, 2015.

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

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jangho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022. URL https://api.semanticscholar.org/CorpusID:251647808.

Weiyang Kong, Ziyu Guo, and Yubao Liu. Spatio-temporal pivotal graph neural networks for traffic flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. doi: 10.1609/aaai.v38i8.28707.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2241719.

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

Siyu Liao, Liutong Zhou, Xuan Di, Bo Yuan, and Jinjun Xiong. Large-scale short-term urban taxi demand forecasting using deep learning. In *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 428–433, 2018. doi: 10.1109/ASPDAC.2018.8297361.

Shixuan Liu, Jijun Zhu, Wei Lei, and Ping Zhang. Spatial-temporal attention graph wavenet for traffic forecasting. In *2023 5th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, pp. 1–8, 2023. doi: 10.1109/DOCS60977.2023.10294485.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: exploring the stationarity in time series forecasting. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

Mario Lucic, Karol Kurach, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. Are gans created equal? a large-scale study. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 698–707, Red Hook, NY, USA, 2018. Curran Associates Inc.

Donghao Luo and Xue Wang. ModernTCN: A modern pure convolution structure for general time series analysis. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vpJMJerXHU.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3):e0194889, 2018. doi: 10.1371/journal.pone.0194889.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, 2020. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2019.04.014. URL https://www.sciencedirect.com/science/article/pii/S0169207019301128.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByJHuTgA-.

Allan H. Murphy. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116(12):2417–2424, 1988. doi: 10.1175/1520-0493(1988)116⟨2417:SSBOTM⟩2.0.CO;2.

Allan H. Murphy and Edward S. Epstein. Skill scores and correlation coefficients in model verification. *Monthly Weather Review*, 117(3):572–582, 1989. doi: 10.1175/1520-0493(1989)117⟨0572:SSACCI⟩2.0.CO;2.

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

Eduardo Ogasawara, Leonardo C. Martinez, Daniel de Oliveira, Geraldo Zimbrão, Gisele Lobo Pappa, and Marta Mattoso. Adaptive normalization: A novel data normalization approach for non-stationary time series. In *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2010. URL https://api.semanticscholar.org/CorpusID:5757527.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1ecqn4YwB.

Bowen Pang, Liyi Huang, Xiangnan Feng, and Wei Wei. Colf: Disentangled forecasting for non-stationary time series via contrastive learning. *Pattern Recognition*, 182:114808, 2027. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2026.114808. URL https://www.sciencedirect.com/science/article/pii/S0031320326017723.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tJbbQfw-5wv.

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 338–348, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1035. URL https://aclanthology.org/D17-1035/.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2019.07.001. URL https://www.sciencedirect.com/science/article/pii/S0169207019301888.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations (ICLR)*, 2025.

Gang Wang, Yu Liao, Li Guo, Jiahao Geng, and Xianchao Ma. Dlinear photovoltaic power generation forecasting based on reversible instance normalization. In *2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS)*, pp. 990–995, 2023. doi: 10.1109/DDCLS58216.2023.10166973.

Hao Wang, Dongsheng Zou, Bi Zhao, Yuming Yang, Jiyuan Liu, Naiquan Chai, and Xinyi Song. Rdlinear: A novel time series forecasting model based on decomposition with revin. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2024a. doi: 10.1109/IJCNN60899.2024.10650961.

Haodong Wang. Yellow taxi demand prediction for new york city based on vmd-ssa-lstm. In *Proceedings of the 1st International Conference on Data Science and Engineering - Volume 1: ICDSE*, pp. 434–440. INSTICC, SciTePress, 2024. ISBN 978-989-758-690-3. doi: 10.5220/0012810400004547.

Haoyun Wang and Robert B. Noland. Bikeshare and subway ridership changes during the covid-19 pandemic in new york city. *Transport Policy*, 106:262–270, 2021. ISSN 0967-070X. doi: https://doi.org/10.1016/j.tranpol.2021.04.004. URL https://www.sciencedirect.com/science/article/pii/S0967070X21000974.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024b.

Yunshi Wen, Wesley M. Gifford, Chandra Reddy, Lam M. Nguyen, Jayant Kalagnanam, and Anak Agung Julius. Revisiting the generic transformer: Deconstructing a strong baseline for time series foundation models. *arXiv preprint arXiv:2602.06909*, 2026.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations (ICLR)*, 2023.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, pp. 1907–1913. AAAI Press, 2019. ISBN 9780999241141.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with 10k parameters. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=bWcnvZ3qMb.

Weiwei Ye, Songgaojun Deng, Qiaosha Zou, and Ning Gui. Frequency adaptive normalization for non-stationary time series forecasting. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 3634–3640. AAAI Press, 2018. ISBN 9780999241127.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i9.26317. URL https://doi.org/10.1609/aaai.v37i9.26317.

Can Zhang and Minglong Lei. A survey on spatio-temporal graph neural networks for traffic forecasting. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 1417–1423, 2023. doi: 10.1109/ICDMW60847.2023.00180.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vSVLM2j9eie.

Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. doi: 10.1609/aaai.v34i01.5477.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325.

APPENDIX

A HYPERPARAMETERS AND TRAINING PROTOCOL

Table A1: One interface, five demand systems in two cities and four residual backbones. New York taxi and ride-hail share the 69-zone Manhattan lattice, both Chicago systems the 77-area lattice. Density gives the panel mean of trips per unit-hour and stays blank for the subway, whose series counts station entries instead of trips. Splitting is chronological throughout, the three dates giving the last training hour, the last validation hour and the last test origin, each panel pinned by SHA-256 (first 12 hex digits).

City	System	Units	Density	Test origins	Stride (h)	Train / val / test boundaries	Panel SHA-256
New York	Taxi	69	58.4	1065	6	2024-06-30 / 2024-09-30 / 2025-06-24	`99474e4c69d9`
New York	Ride-hail	69	154.6	1069	6	2023-06-30 / 2023-09-30 / 2024-06-24	`cedc7d0016c3`
New York	Subway	424	—	268	24	2023-06-30 / 2023-09-30 / 2024-06-24	`74ac27744eb3`
Chicago	Taxi	77	9.6	1073	6	2024-12-31 / 2025-03-31 / 2025-12-25	`97185ac2323b`
Chicago	Ride-hail	77	122.9	1073	6	2024-12-31 / 2025-03-31 / 2025-12-25	`4f23e4b524ec`

Backbone	Parameters	Configuration (full detail in Table A2)
GRU-64 / DLinear	27329 / 56784	hidden width 64 / decomposition kernel 25, weights shared across units
iTransformer / PatchTST	308520 / 682280	d_{model} 128, 8 heads, 2 / 3 layers; internal instance norm removed

The zero-versus-missing inclusion rule. Treating an absent (unit, hour) as a real zero or as an absence of measurement decides which cells enter every anchor mean and every MAE and RMSE, so the rule goes on the record rather than staying implied. An absent (unit, hour) is a real zero when that unit reported at least one trip somewhere in the same 168-hour block of the grid; otherwise it is masked, and a masked cell drops out of every anchor average and every reported metric. The rule came over from the subway panel, the one that had to meet raw coverage, and every arm receives it identically, so no comparison here turns on a coverage threshold anyone had to choose. Should a rolling-anchor bin be fully masked across its eight weeks, Equation 2 falls back on that unit's train-split static value from Equation 1 for the same hour-of-week bin. As built, all five panels show a realised mask rate of 0.000, so the rule removes no cell from any number reported here; it is stated because a reimplementation on a sparser panel needs it, and because this study's sparsest panel, Chicago taxi at 9.6 trips per area-hour, is the one that carries the replication claim.

Table A2 and this rule are all a reimplementation needs, and three of its entries matter more than the rest. First comes removal of the internal instance-normalization layers from PatchTST and iTransformer, because leaving them in place puts two tenants in the normalization slot, at which point the comparison stops being between interfaces. Second comes the learning rate, the single quantity chosen per arm on a shared validation proxy so no comparator suffers a rate tuned to the anchored family; that same choice is why a configuration count credits the SCAN arm with two rather than one. Third comes the anchor's complete absence of parameters, which lets the parameter column say plainly what each normalization interface costs.

Comparator fidelity is stated here instead of left implied. The RevIN arm receives the released layer unmodified. The Dish-TS arm adapts the released model, keeping its dual input/output coefficient formulation but routing it through a shared two-layer coefficient head rather than the released per-variate construction. The frequency-adaptive arm follows the paper and its released reference implementation alike: a top-K spectral mask over the input window, a two-head prediction of the future principal component, and the auxiliary regression term that head trains against, held at the reference weight of 1.0. We wrote PatchTST and iTransformer from their papers with channel or variate tokenization and the published prediction head, LayerNorm substituting for BatchNorm inside the PatchTST encoder. Not one is calibrated against a published number on a standard long-horizon benchmark – the check a reader wants before blaming a weak comparator on the method rather than on the rebuild – and Appendix M keeps the item on its list.

Cells, compute and convergence. Behind the headline grid stand 755 recorded cells – 640 trainings, 100 frozen evaluations and 15 model-free floors – across 5.80 hours of recorded GPU compute over five resumed invocations, with 740 convergence records running 9 to 96 epochs, 528 stopping on patience and 212 on the cap. Every panel is re-hashed against the digest in Table A1 before training starts, and the three New York panels carry those same digests in the tuning-budget and epoch-ladder sweeps of Appendices H and L.

Table A2: Hyperparameters in full, shared by every arm on a given backbone unless a row says otherwise. Rows marked "backbone-specific" vary between backbones and hold fixed over the normalization arms inside one. Parameter counts are exact and leave out the calendar embedding, used by the recurrent backbone alone. A lookup table adds none, while each learned normalizer adds its own on top of the backbone.

Setting	Value
Model	
Input length L	168 h (one week), identical for every arm
Output length	168 h emitted at once; scored at $h \in \{1, 6, 24, 72, 168\}$
Recurrent backbone	GRU, hidden width 64, shared across units, 27329 parameters
Calendar covariates (backbone-specific)	consumed by the recurrent backbone only; the other three receive the residual window alone
Decoding (backbone-specific)	autoregressive for the recurrent backbone; direct 168-step projection for the other three
Comparator provenance	RevIN released layer verbatim; Dish-TS adapted with a shared 2-layer coefficient head; frequency-adaptive normalization reimplemented from the paper and reference code with its auxiliary loss at weight 1.0; PatchTST and iTransformer reimplemented by us, PatchTST with LayerNorm in place of the official BatchNorm encoder
Linear backbone	decomposition-linear, kernel 25, shared across units, 56784 parameters
Patch transformer	PatchTST, d_{model} 128, 8 heads, 3 layers, d_{ff} 256, patch 24, stride 12, dropout 0.1, 682280 parameters
Inverted transformer	iTransformer, d_{model} 128, 8 heads, 2 layers, d_{ff} 256, dropout 0.1, 308520 parameters
Internal instance norm	removed from both transformers in every arm
Anchor interface	lookup table, 0 parameters, no inference-time statistics
RevIN / Dish-TS	affine wrapper (up to 138 parameters) / learned coefficient network
Frequency-adaptive normalization	$K \in \{3, 5\}$ retained components crossed with the two learning rates, selected per cell on the adjacent validation block; 4 configurations against 2 for every other arm. The wider $\{2, 3, 5, 8\}$ bracket is defined in the released configuration and was not swept
Zero-versus-missing inclusion rule	an absent (unit, hour) is a real zero when that unit reported at least one trip somewhere in the same 168-hour block, and is masked out of every anchor mean and every MAE and RMSE otherwise; identical for every arm
Optimisation	
Optimiser	Adam
Loss	masked L1 on the residual, equal to absolute error in demand units
Batch size	32
Learning rate	selected per arm from $\{10^{-3}, 3 \times 10^{-4}\}$ on validation 24 h MAE
Early stopping	validation 24 h MAE at h=24, patience 8 at the converged budget
Epoch budget	published cap 16 (taxi, ride-hail, both Chicago systems), 10 (subway); converged budget six times each
Seeds	42, 43, 44, 45, 46 for every trained arm
Data and splits	
Panel length	17544 h per system
Split order	chronological, train then validation then test
Test origin stride	6 h on the four ride-based panels, 24 h on the subway
Rolling anchor window	$W = 8$ weeks, frozen at the forecast origin; fixed a priori and never swept
Tuning-budget grid	8 configurations per learned arm against 2 for SCAN, ranked on the adjacent validation block; wrapper hyperparameters and learning rate only, never the epoch budget
Execution	
Hardware	1 × NVIDIA TITAN RTX, 23.6 GB
Software	torch 2.2.0 + CUDA 12.1, cuDNN 8902, Python 3.9.12, numpy 1.26.4, scipy 1.13.0
Recorded compute	755 cells, 5.80 h across five resumed invocations
Bootstrap	2000 resamples, 168 h blocks, master seed 20260918, per-cell stream = seed + CRC-32(identifier) mod 100000

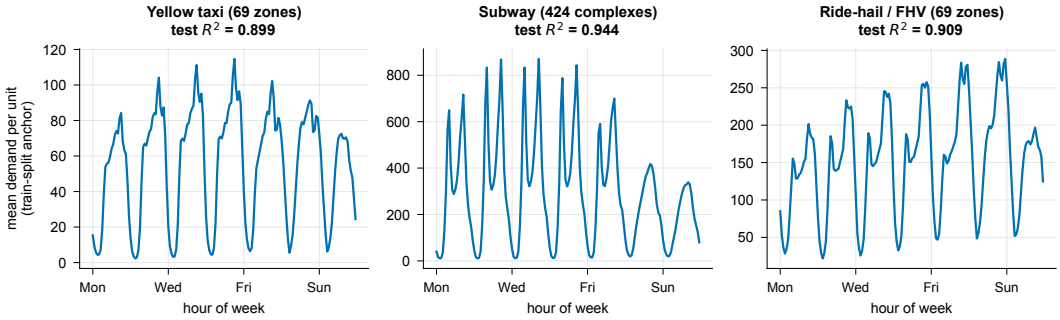


Figure B1: Structure the network never has to learn. The per-unit hour-of-week climatology profile is fit on the training split and labelled with the share of held-out test variance it covers (89.9, 94.4 and 90.9 percent on New York taxi, subway and ride-hail). This panel does not draw test demand itself, and the measurement does not reach the two Chicago panels.

B THE RECOMMENDED CONFIGURATION, CELL BY CELL

Figure B1 draws what the anchor takes away before any network sees the series. Table B1 then gives the recommended configuration across each of its 25 cells: the fixed origin-frozen rolling-8 anchor under a PatchTST residual head, its margin over the best learned normalizer on the identical backbone, the five-seed screen outcome, the pooled RMSE margin, and its record against all three parameter-free floors. All 25 margins over learned normalization come out positive, screened and positive on RMSE. Against the free floors the record reads 20 of 25, 23 of 25 and 25 of 25, a negative entry marking a cell the floor forecasts better. Table B2 aggregates the full 100-cell grid by system and by city. Read on the static anchor instead, those same 100 cells give 100 positive and screened margins running 10.52 to 65.07 percent about a median of 32.57.

Table B1: The recommended configuration in full, one arm over all 25 cells, a fixed origin-frozen rolling-8 anchor under a PatchTST residual head, placed beside its record against each of the three parameter-free floors. Cells whose margin exceeds the combined dispersion over runs 42–46 are ticked in the screen column, and a negative floor margin flags a cell the floor forecasts better. A lower MAE is a better forecast.

		vs. best learned normalizer			ΔMAE % vs. free floor			
System	h	MAE (↓)	ΔMAE %	Screen	ΔRMSE %	Rolling	Static	Seas. naive
Taxi (New York)	1	10.293	+29.96	✓	+22.02	+14.46	+26.18	+26.73
	6	11.353	+44.79	✓	+45.86	+3.35	+17.69	+17.71
	24	11.496	+29.88	✓	+23.89	+2.72	+16.91	+16.65
	72	11.910	+38.60	✓	+29.51	+0.32	+14.14	+13.67
	168	12.337	+38.43	✓	+28.46	-1.10	+11.46	+10.44
Ride-hail (New York)	1	19.285	+26.02	✓	+20.77	+17.29	+18.44	+31.30
	6	21.246	+44.56	✓	+43.32	+6.32	+7.70	+22.32
	24	22.173	+37.37	✓	+33.01	+2.49	+3.74	+18.80
	72	23.543	+40.33	✓	+33.89	-1.91	-1.99	+13.73
	168	24.664	+42.29	✓	+37.89	-5.13	-6.25	+9.65
Subway (New York)	1	13.511	+21.32	✓	+5.66	+11.81	+16.50	+24.60
	6	14.075	+39.22	✓	+26.53	+2.47	+29.97	+15.88
	24	22.859	+11.76	✓	+4.39	+8.53	+17.30	+21.54
	72	24.290	+12.02	✓	+4.61	+4.00	+12.16	+16.75
	168	25.372	+10.55	✓	+4.47	+1.49	+8.22	+12.60
Taxi (Chicago)	1	2.496	+25.67	✓	+31.85	+9.05	+16.71	+26.56
	6	2.610	+23.60	✓	+30.23	+2.82	+12.33	+21.42
	24	2.587	+34.24	✓	+37.78	+4.49	+13.72	+22.67
	72	2.727	+32.67	✓	+36.82	+0.30	+9.33	+19.01
	168	2.770	+32.70	✓	+35.72	+0.11	+8.36	+17.51
Ride-hail (Chicago)	1	15.497	+41.03	✓	+43.40	+14.48	+15.40	+30.54
	6	17.064	+51.98	✓	+56.14	+4.38	+7.26	+22.64
	24	17.670	+38.12	✓	+35.42	+2.28	+4.95	+20.73
	72	18.614	+38.76	✓	+34.09	-0.37	+1.50	+17.92
	168	19.059	+38.27	✓	+33.20	-0.25	+0.20	+16.88
Cells won	of 25		**25**	**25**	**25**	20	23	25

Table B2: The selection-free 100-cell grid at the converged budget, aggregated by system and by city with one fixed origin-frozen rolling-8 anchor throughout. Paired per-run tests take Holm–Bonferroni correction across the whole grid, and the screened column tallies cells whose margin exceeds the combined dispersion over runs 42–46. Higher margins are better.

System	Cells	Positive	Screened	Holm-sig.	RMSE+	Median %	Range %
Taxi (New York)	20	20	20	20	20	+38.81	19.78–54.21
Ride-hail (New York)	20	20	20	20	20	+39.59	18.58–55.52
Subway (New York)	20	20	20	20	20	+23.80	10.55–74.26
Taxi (Chicago)	20	20	20	20	20	+27.98	11.93–38.43
Ride-hail (Chicago)	20	20	20	20	20	+38.93	24.27–55.94
New York subset	60	60	60	60	60	+38.51	10.55–74.26
Chicago subset	40	40	40	40	40	+32.69	11.93–55.94
All five systems	100	100	100	100	100	+36.19	10.55–74.26

C THE SECOND CITY, CELL BY CELL

Table C1 runs Section 3.3's rule over each of the 40 Chicago cells at the converged budget, printing the margin beside the threshold that cell had to clear – half the New York median margin in the matching backbone-by-horizon cell, taken from the same 100-cell grid. Every margin is positive and every one sits above seed noise; the one cell under its threshold is Chicago taxi with the recurrent head at h=6. Replacing the New York reference column with one computed over the three New

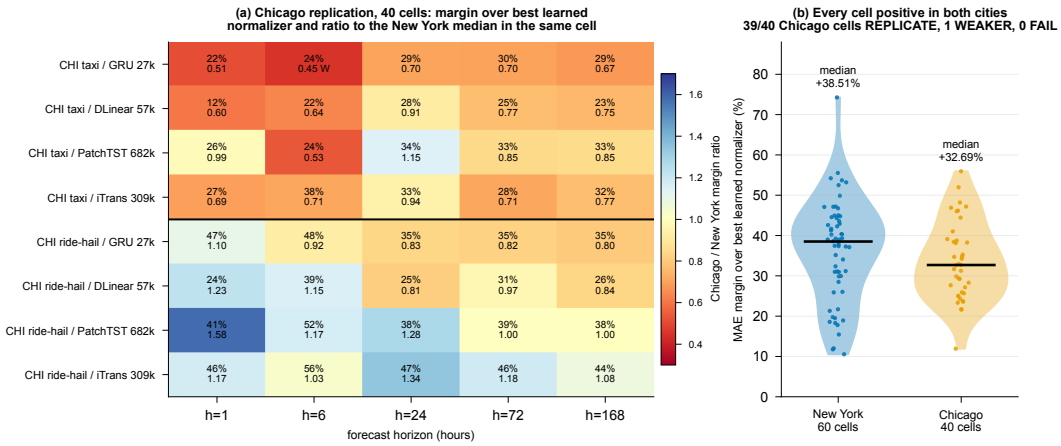


Figure C1: The second city. (a) All 40 Chicago cells at the converged budget, each carrying above it the margin over the best learned normalizer on the matching backbone and below it the ratio of that margin to the New York median in the same backbone-by-horizon cell, with the one weaker cell marked W. (b) Every cell in both cities is positive; ruled medians sit at 38.51 percent over the New York 60 and 32.69 over the Chicago 40, both read off the same 100-cell grid. Higher is better.

York systems alone, from a learned pool holding RevIN and Dish-TS but not the spectral arm and running larger in 14 of the 20 backbone-by-horizon cells, yields the conservative cross-check named in Section 4.4, which gives 38 replications, 2 weaker cells and no failures.

Table C1: Replication verdict across all 40 Chicago cells at the converged budget, run like-for-like. One grid scores both cities, with the anchor fixed and the spectral comparator inside the learned pool on either side. Required means half the New York median margin in the matching backbone-by-horizon cell. Higher margins are better.

System	Backbone	h	Anchor MAE	Learned	Learned MAE	Margin %	Required %	Verdict
CHI taxi	GRU-64	1	2.710	FAN	3.462	+21.71	21.30	REPLICATES
CHI taxi	GRU-64	6	2.682	Dish-TS	3.515	+23.70	26.25	**WEAKER**
CHI taxi	GRU-64	24	2.707	RevIN	3.824	+29.23	20.83	REPLICATES
CHI taxi	GRU-64	72	2.732	RevIN	3.895	+29.86	21.47	REPLICATES
CHI taxi	GRU-64	168	2.769	RevIN	3.922	+29.39	21.79	REPLICATES
CHI taxi	DLinear	1	2.265	RevIN	2.571	+11.93	9.89	REPLICATES
CHI taxi	DLinear	6	2.531	RevIN	3.230	+21.63	17.03	REPLICATES
CHI taxi	DLinear	24	2.599	RevIN	3.625	+28.30	15.48	REPLICATES
CHI taxi	DLinear	72	2.842	RevIN	3.785	+24.90	16.18	REPLICATES
CHI taxi	DLinear	168	2.902	RevIN	3.786	+23.33	15.50	REPLICATES
CHI taxi	iTransformer	1	2.495	FAN	3.424	+27.13	19.65	REPLICATES
CHI taxi	iTransformer	6	2.604	FAN	4.230	+38.43	27.11	REPLICATES
CHI taxi	iTransformer	24	2.602	RevIN	3.881	+32.95	17.58	REPLICATES
CHI taxi	iTransformer	72	2.734	RevIN	3.779	+27.66	19.51	REPLICATES
CHI taxi	iTransformer	168	2.784	RevIN	4.074	+31.67	20.63	REPLICATES
CHI taxi	PatchTST	1	2.496	FAN	3.358	+25.67	13.01	REPLICATES
CHI taxi	PatchTST	6	2.610	Dish-TS	3.416	+23.60	22.28	REPLICATES
CHI taxi	PatchTST	24	2.587	FAN	3.933	+34.24	14.94	REPLICATES
CHI taxi	PatchTST	72	2.727	FAN	4.050	+32.67	19.30	REPLICATES
CHI taxi	PatchTST	168	2.770	FAN	4.116	+32.70	19.21	REPLICATES
CHI ride-hail	GRU-64	1	17.514	FAN	32.980	+46.90	21.30	REPLICATES
CHI ride-hail	GRU-64	6	17.825	Dish-TS	34.409	+48.20	26.25	REPLICATES
CHI ride-hail	GRU-64	24	18.091	RevIN	27.709	+34.71	20.83	REPLICATES
CHI ride-hail	GRU-64	72	18.542	RevIN	28.645	+35.27	21.47	REPLICATES
CHI ride-hail	GRU-64	168	19.016	RevIN	29.157	+34.78	21.79	REPLICATES
CHI ride-hail	DLinear	1	14.608	RevIN	19.290	+24.27	9.89	REPLICATES
CHI ride-hail	DLinear	6	16.778	RevIN	27.553	+39.11	17.03	REPLICATES
CHI ride-hail	DLinear	24	18.198	RevIN	24.281	+25.05	15.48	REPLICATES
CHI ride-hail	DLinear	72	19.569	RevIN	28.468	+31.26	16.18	REPLICATES
CHI ride-hail	DLinear	168	19.970	RevIN	26.953	+25.91	15.50	REPLICATES
CHI ride-hail	iTransformer	1	16.388	RevIN	30.372	+46.04	19.65	REPLICATES
CHI ride-hail	iTransformer	6	17.287	Dish-TS	39.233	+55.94	27.11	REPLICATES
CHI ride-hail	iTransformer	24	18.404	RevIN	34.852	+47.19	17.58	REPLICATES
CHI ride-hail	iTransformer	72	19.705	RevIN	36.616	+46.19	19.51	REPLICATES
CHI ride-hail	iTransformer	168	20.193	RevIN	36.351	+44.45	20.63	REPLICATES
CHI ride-hail	PatchTST	1	15.497	Dish-TS	26.279	+41.03	13.01	REPLICATES
CHI ride-hail	PatchTST	6	17.064	Dish-TS	35.535	+51.98	22.28	REPLICATES
CHI ride-hail	PatchTST	24	17.670	Dish-TS	28.554	+38.12	14.94	REPLICATES
CHI ride-hail	PatchTST	72	18.614	Dish-TS	30.396	+38.76	19.30	REPLICATES
CHI ride-hail	PatchTST	168	19.059	Dish-TS	30.873	+38.27	19.21	REPLICATES

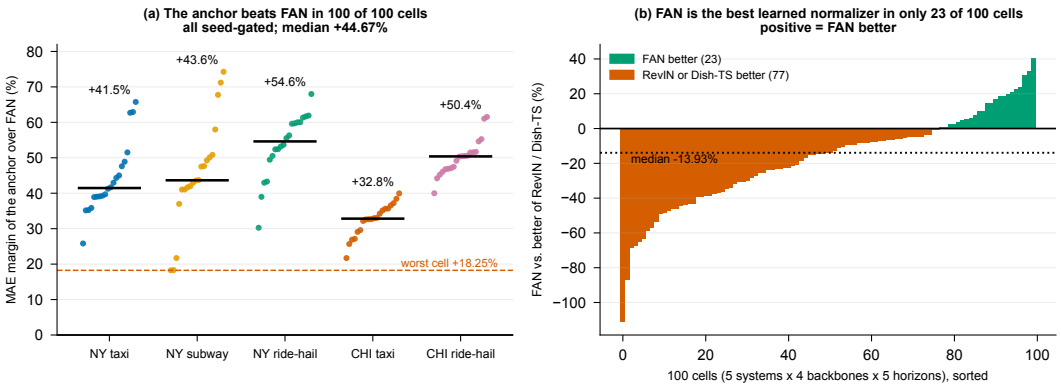


Figure D1: The periodicity-aware comparator over all 100 cells at the converged budget. (a) System by system, the MAE margin of the fixed rolling-8 anchor over frequency-adaptive normalization on the identical backbone, with the per-system median ruled and the grid's narrowest cell marked; higher is better. (b) The same cells rescored against whichever of RevIN and Dish-TS fares better, so a positive bar flags a cell where the spectral normalizer leads both siblings.

D THE PERIODICITY-AWARE COMPARATOR IN FULL

Table D1 lays out the accounting behind Figure D1 system by system: how often the fixed rolling-8 anchor beats the spectral normalizer on the identical backbone, by what median amount, how often that normalizer is itself the strongest of the three learned arms in a cell, and where it stands against whichever of RevIN and Dish-TS fares better. In the final column a negative entry marks a system where the spectral arm trails both siblings at the median.

Four configurations wide, not eight, is the search over K. Every one of the twenty shipped tuning artifacts – five systems by four backbones – holds the same grid, $K \in \{3, 5\}$ crossed with the two learning rates, and adjacent validation at the converged budget selects $K{=}5$ in 14 of those 20 combinations and $K{=}3$ in 6. Defined in the released configuration but never swept is the wider $\{2, 3, 5, 8\}$ bracket, which would put a component on the daily and another on the weekly harmonic of a 168-hour window; the four-against-two configuration asymmetry favouring the comparator stands either way, every other arm drawing two. Two consequences follow, recorded here rather than argued away. One is that the bracket genuinely searched is narrower than a reader of this arm's source paper would assume. The other is that this arm, our reimplementation, heads the learned pool in 23 of 100 cells and sits a median 13.93 percent behind the better of RevIN and Dish-TS – the reverse of the ordering its source paper reports on its own benchmarks. On this evidence a weak normalizer and a weak rebuild cannot be told apart, and the calibration run that would separate them appears in Appendix M as the specific experiment this paper lacks.

Table D1: Frequency-adaptive normalization over all 100 cells at the converged budget, with K chosen per cell on the adjacent validation block out of $\{3, 5\}$. Column 3 tallies cells the anchor takes; the last column gives the median gap between the spectral arm and whichever of RevIN and Dish-TS fares better in that cell, as a percentage of that sibling's MAE, so a negative value puts it behind both.

System	Cells	Anchor > FAN	Median ΔMAE %	FAN best learned	FAN vs. RevIN/Dish-TS %
Taxi (New York)	20	20	+41.47	6	-7.48
Ride-hail (New York)	20	20	+54.62	3	-35.43
Subway (New York)	20	20	+43.65	6	-26.27
Taxi (Chicago)	20	20	+32.83	7	-5.39
Ride-hail (Chicago)	20	20	+50.39	1	-22.86
All five systems	100	**100**	+44.67	23	-13.93

E MATCHED-ANCHOR RESIDUAL ACCOUNTING IN FULL

Table E2 unfolds what Table E1 aggregates, giving residual accounting across all 200 cells at the converged epoch budget, two anchor types crossed with four backbones, five systems and five horizons. Read a backbone block downwards and the origin of each aggregate shows. One hour out sits close enough that departures from climatology still carry information, so the $h{=}1$ column is where the heads collect most of what they collect, from 2.95 percent on the recurrent head to 19.74 on

Table E1: The capacity ladder at the converged epoch budget, ordered by parameter count and reported separately for each anchor type. Each mean is an unweighted mean of per-cell $R(h)$ values over that backbone and anchor's 15 long-horizon cells ($h \in \{24, 72, 168\}$ over five systems), every cell itself a five-seed mean; the + column tallies cells above zero and the two sig columns those whose paired seed-level test survives Holm correction. Higher is better. The rolling-8 block is what bears on the recommended configuration. Downwards no column rises, and capacity stays confounded with tokenization, architecture, covariates and decoding.

		Rolling-8 anchor (recommended), 15 cells				Static anchor, 15 cells				
Backbone	Parameters	Mean R %	+	sig+	sig−	Mean R %	+	sig+	sig−	$R(1)$ %
GRU-64	27329	+0.21	9	2	0	+0.23	5	0	3	+2.95
DLinear	56784	-0.96	5	4	9	**+4.33**	10	10	4	+19.74
iTransformer	308520	-4.36	2	1	11	-0.14	8	5	6	+6.62
PatchTST	682280	**+1.20**	**10**	6	2	**+4.89**	**12**	10	1	+13.42

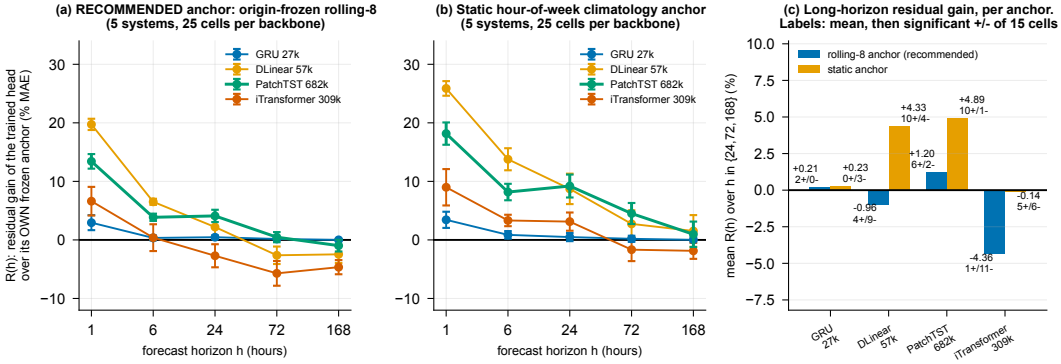


Figure E1: Matched-anchor residual gain $R(h)$ at the converged budget, split by anchor type, 25 cells per backbone per anchor. (a) Under the recommended origin-frozen rolling-8 anchor every backbone collects at h=1, and past a day two of the four fall below zero. (b) Under the static anchor three of the four stay above the line at every horizon. (c) The long-horizon means behind (a) and (b), labelled with counts of corrected positive and negative cells out of 15. Higher is better.

DLinear under the recommended anchor. A day out the information is gone, which is the regime Table E1 reports. Pooled over the two anchor types the figures would read $+0.22$ percent for the recurrent head, $+1.68$ for DLinear, $+3.05$ for PatchTST and -2.25 for iTransformer across each backbone's 30 long-horizon cells, 15 under either anchor; those fold a static-anchor reference into an accounting for arms shipping the rolling one, and they describe no configuration recommended here.

Table E2: Matched-anchor residual gain $R(h)$, percent, by backbone and anchor type at the converged budget, every entry an unweighted mean over the five systems. A positive value says the trained head improved on the frozen anchor handed to it, so higher is better.

Backbone	Anchor	h=1	Long horizon ($h \in \{24, 72, 168\}$, 15 cells)			
		Mean %	Mean %	Median %	Cells > 0	Corrected ±
GRU-64	rolling-8	+2.95	+0.21	+0.05	9	2 / 0
	static	+3.43	+0.23	-0.18	5	0 / 3
DLinear	rolling-8	+19.74	-0.96	-1.19	5	4 / 9
	static	+25.88	+4.33	+2.70	10	10 / 4
iTransformer	rolling-8	+6.62	-4.36	-5.33	2	1 / 11
	static	+8.99	-0.14	+0.11	8	5 / 6
PatchTST	rolling-8	+13.42	+1.20	+0.32	10	6 / 2
	static	+18.16	+4.89	+4.79	12	10 / 1

F THE UNNORMALIZED CONTROL

Table F1 reports the control declared in Section 4.1 – the same four backbones with nothing in the normalization slot – because what it shows moderates Section 5.3. Across the same 100 cells at the converged budget its median MAE is 50.05 beside the best learned normalizer's 26.17, and it ends ahead of that normalizer in 3 cells and ahead of the anchored arm in none. Per-instance normalization is doing substantial real work on these panels, then, stripping a median 59.19 percent off the control's error, most of what the anchored arm's 73.64 percent strips off the same denominator. Broken the comparators are not; beaten by a groupby they are.

Table F1: The unnormalized control at the converged budget, system by system, every entry a median over that system's 20 cells. Learned Δ gives the best learned normalizer's MAE reduction over the control in the same cell, so a large value says the normalization block does real work, while Control wins tallies cells the control takes from that normalizer. A lower MAE is a better forecast.

System	Cells	Control MAE (↓)	Best learned (↓)	Anchored (↓)	Learned Δ %	Control wins
Taxi (New York)	20	46.275	19.717	**11.687**	+53.44	0
Ride-hail (New York)	20	91.217	37.975	**22.755**	+57.15	0
Subway (New York)	20	89.113	28.883	**23.624**	+61.45	0
Taxi (Chicago)	20	8.493	3.785	**2.694**	+52.94	1
Ride-hail (Chicago)	20	94.607	29.764	**18.144**	+65.98	2
All five systems	100	50.05	26.17	**16.16**	+59.19	3

G FLOOR CONTEST IN FULL

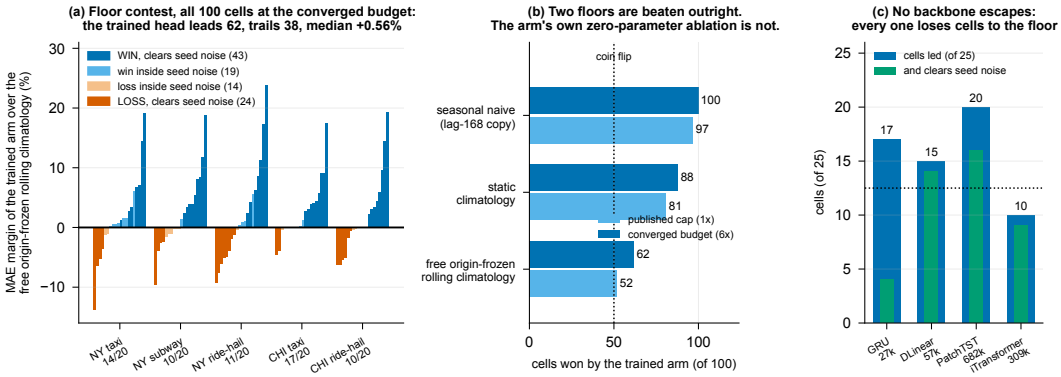


Figure G1: Floor contest at the converged epoch budget, cell by cell. (a) MAE margin of the fixed rolling-anchored arm over the free origin-frozen rolling climatology across all 100 cells, wins and losses in one panel, shaded by whether a cell clears the combined five-seed standard deviation. (b) Cells won against each of the three parameter-free floors, at the published cap and at the converged budget. (c) The same count split by backbone, where every backbone drops cells to a predictor holding no parameters. Higher is better in (a).

Table G1 counts all 600 comparisons between a fixed rolling-anchored arm and a parameter-free floor, at both epoch budgets. Order the three floors by win frequency and Section 5.2 comes back exactly. At the converged budget seasonal naïve loses everywhere, the static climatology table loses over more than five sixths of the grid, and the origin-frozen rolling climatology – the cost-free form of the very anchor those arms carry – holds 38 of 100 even so. Its count falls from 48 to 38 between the published cap and the converged budget while the median margin moves from 0.04 to 0.56 percent, and at both budgets a large share of cells is decided inside seed noise.

Table G1: Floor contest for the fixed rolling-anchored arms across 600 comparisons, three floors by 100 cells by two epoch budgets. A lead means a cell the trained arm forecasts better, and the two noise columns tally leads and losses whose margin clears the combined five-seed standard deviation. Every margin is a percentage of the floor's own MAE, higher favouring the trained arm.

Floor	Budget	Cells	Arm leads	Leads past noise	Losses past noise	Median ΔMAE %
Free origin-frozen rolling climatology	published cap	100	52	34	34	+0.04
	converged	100	62	43	24	+0.56
Static climatology	published cap	100	81	74	16	+6.97
	converged	100	88	80	12	+9.24
Seasonal naïve (lag-168)	published cap	100	97	95	2	+16.46
	converged	100	100	99	0	+17.52

H TUNING BUDGET IN FULL

The tuning-budget sweep covers the three New York systems, with RevIN and Dish-TS filling the learned pool and the spectral arm outside it, and that scope is stated wherever it appears (Figure H1). Its 24 budget-8 cells list, one row per system, backbone and learned normalizer, test MAE at $h{=}24$ for whichever configuration survived 8 validation-selected candidates, set beside the seed-mean test MAE of the single fixed SCAN configuration in that cell. SCAN wins all 24, as it wins the budget 1, 2 and 4 cells, giving 96 of 96. Gaps surviving at budget 8, each read as a percentage of SCAN's own

MAE, run 24.1 to 136.839 percent around a median of 62.915, with per-backbone minima of 24.1 percent on PatchTST, 25.094 on iTransformer, 27.7 on the recurrent head and 33.259 on DLinear. Each learned arm in that sweep receives a grid of 8 points – a learning rate from $\{10^{-3}, 3 \times 10^{-4}\}$ crossed with either a Dish-TS wrapper width or RevIN's affine and subtract-last switches – against a SCAN setting fixed at every budget and drawn from those same 2 learning rates. Neither side varies epoch budget, input length or batch size. Its SCAN line is the static-anchor arm over five seeds while every learned entry is a single run; recomputed against the headline origin-frozen rolling-8 anchor the tally stays 96 of 96, with budget-8 gaps running 29.381 to 118.091 percent around a median of 67.2 and per-backbone minima of 29.381 for PatchTST, 31.477 for iTransformer, 37.814 for DLinear and 47.7 for the recurrent head.

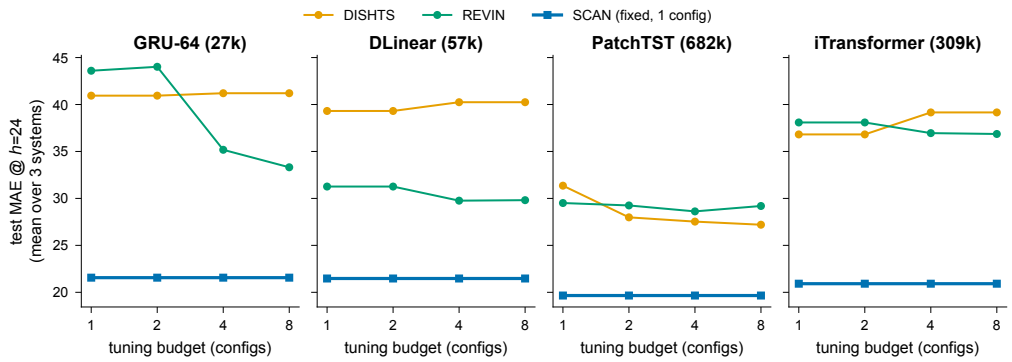


Figure H1: Tuning-budget curves at h=24 over the three New York systems and averaged across them, with RevIN and Dish-TS the learned arms in reach. Each learned arm draws 1, 2, 4 or 8 configurations chosen on the adjacent validation block while SCAN holds one fixed configuration at every budget. SCAN takes all 96 budget cells. A lower MAE is a better forecast.

I Validation protocols and the two alternative selections

Run Section 3.3's rule on the pre-declared adjacent block at the converged budget and the static-anchored PatchTST arm comes first at 0.7781. Run it on an equally long block sitting immediately before the scoring window and the rolling-anchored recurrent arm wins at 0.9652. At the published cap the same two protocols return 0.7966 and 0.9699, picking the same pair. The split is structural, not noise. Refit at every origin, a rolling anchor gains on a static one as distance from training grows, while beside training the static anchor has yet to go stale, so a validation block pressed against training cannot see that gain. Since the adjacent block is both the published protocol and the one we pre-declared, switching now – with the alternative's preference already in view – would itself be post-hoc selection (Figure I1). Fix the anchor family a priori, as Section 3.3 does, let the adjacent rule choose the backbone alone, and the recommended configuration follows, the origin-frozen rolling-8 anchor under the PatchTST residual head. Published-cap rankings reproduce the shipped selection file exactly; converged-budget rankings we recomputed from that budget's validation metrics, for which no selection file was shipped.

The distant block selects the fixed rolling-8 anchor under the 27329-parameter recurrent head, reported here as the pre-declared sensitivity analysis. Against learned normalization it beats the recommended configuration on the New York systems; against the free floors the ordering flips, its residual head returning 0.21 percent over its 15 long-horizon cells under the same anchor (Table E1). Each selection thus trades one quantity for another in opposite directions, and either quoted alone would misdescribe the grid.

J The cross-platform correctness gate

This pipeline's CPU reference run over the three New York systems splits into components that have to agree and components that cannot. The 45 model-free floor values – three systems crossed with three floors and five horizons – all reproduce to 1.57e-08 relative, worst per-system disagreements being 1.32e-08 on taxi, 1.57e-08 on the subway and 6.59e-09 on ride-hail. Panel construction, the split calendar, the anchor tables and the three floors sit inside that span, and together they are all

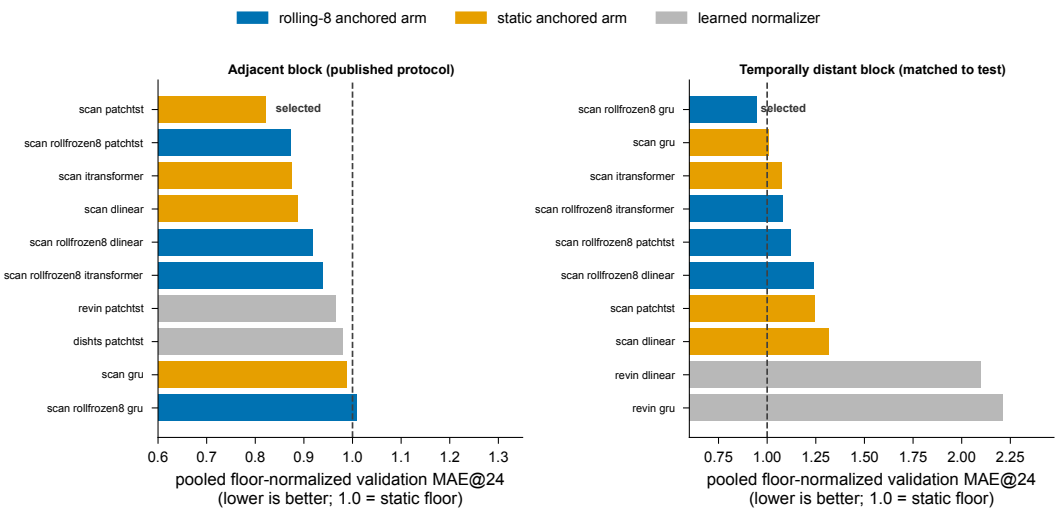


Figure I1: Disagreement between the two validation protocols. Pooled floor-normalized validation MAE at h=24 for the leading candidates, with a block adjacent to training at the left (the pre-declared protocol) and a temporally distant block matched to the test window at the right; lower is better. On the converged budget the adjacent block picks the static-anchored patch transformer at 0.7781, while the distant block flips that ordering and picks the rolling-anchored recurrent arm at 0.9652, the rolling anchor's advantage growing with distance from training.

we call a correctness gate. Trained arms necessarily behave otherwise, with 21 of the 60 values it covers – five arms on the recurrent and linear backbones at $h \in \{24, 168\}$ – differing by more than 5 percent and the worst by 38.48 percent, driven by a differing seed count and by cuDNN kernel choice, the two attributions we make for it. Each trained-arm number here comes out of the GPU run described there, and no claim of cross-platform agreement extends to any trained arm.

K Extended related work: evaluation practice

Spatio-temporal demand and traffic architectures have elaborated steadily. Diffusion-convolutional recurrent networks and graph convolutions (Li et al., 2018; Yu et al., 2018; Wu et al., 2019) yielded to attention over graphs (Zheng et al., 2020; Ali et al., 2022; Liu et al., 2023; Kong et al., 2024) and to general-purpose sequence forecasters imported wholesale (Wu et al., 2021; Zhou et al., 2021; Zhang & Yan, 2023), then to transfer and reprogramming (Fang et al., 2022; Jin et al., 2024). Surveys document that growth alongside how uneven evaluation has stayed (Zhang & Lei, 2023), and dataset surveys record how narrow and how unevenly described the pool of public forecasting panels remains (Hahn et al., 2023). One metric, one horizon and one run carry most comparisons in the area, and this venue's exemplars largely do the same, exiling multi-seed dispersion to an appendix on the occasions it appears at all. Adjacent fields have already shown what the convention hides – language models (Melis et al., 2018), generative models (Lucic et al., 2018), recommendation (Ferrari Dacrema et al., 2019), graph learning under heterophily (Platonov et al., 2023) – and the remedies were settled years ago as well: publish score distributions instead of single runs (Reimers & Gurevych, 2017), budget for benchmark variance (Bouthillier et al., 2021), hold splits in temporal order (Bergmeir et al., 2018), and correct within comparison families (Holm, 1979; Demšar, 2006).

L The epoch-budget ladder

Two rungs make up the ladder on the full five-system grid, the published cap and six times it, behind Section 4.5's summary. Both rungs run under patience 8 on validation MAE at h=24, one ceiling over every arm and each budget's metrics taken from its best validation checkpoint, which is 640 trainings on the same panels, splits, origins, horizons and seeds. Those extra epochs reward the sides unequally – the best learned arm gives a median 5.15 percent of its MAE back to them where the anchored arm gives 0.44 – so the published-cap figures measure truncation in part (Table L1). The cell-level record is 100 of 100 positive and gated at each, the median moving from 37.62 percent to 36.19, 73 cells holding at least four fifths of their cap margin and 27 giving up more, none vanishing

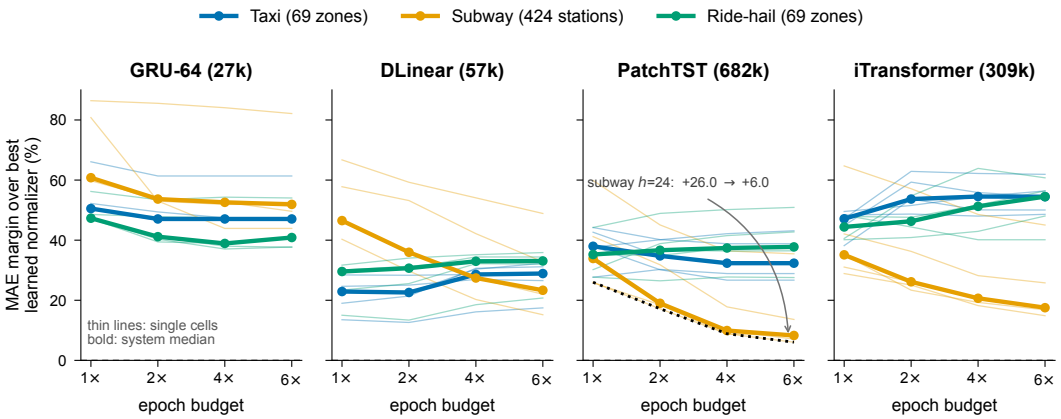


Figure L1: Margin over the best learned normalizer as the epoch budget climbs to six times the published cap, over the three New York systems with RevIN and Dish-TS in the learned pool. All 60 of those cells (thin) appear with the per-system median (bold), patience 8 on validation MAE and one ceiling across both sides. Nowhere does a sign change; the subway transformer column loses most, its $h{=}24$ trace dotted. Higher is better. The two and four times rungs run on these three systems alone.

Table L1: Convergence diagnostics at every epoch budget of the four-rung sweep over the three New York systems, across the arms entering it. "On cap" is the share of runs whose budget expired before patience did; "final-2 gain" is the median validation improvement across the last two epochs, so a value near zero means the run had stopped moving. Lower is better in both columns.

Budget	Arm family	Runs	On cap % (↓)	Final-2 gain % (↓)	Median best epoch	Median epochs seen
1×	Anchored	60	80.00	0.565	12.0	14.0
1×	Learned	120	99.17	1.271	14.0	16.0
1×	Unnormalized	60	100.00	1.285	16.0	16.0
2×	Anchored	60	63.33	0.283	20.0	21.0
2×	Learned	120	90.83	0.334	28.5	32.0
2×	Unnormalized	60	100.00	0.658	31.0	32.0
4×	Anchored	60	41.67	0.160	37.0	40.0
4×	Learned	120	55.83	0.003	40.0	44.0
4×	Unnormalized	60	90.00	0.488	61.0	64.0
6×	Anchored	60	18.33	−0.012	42.0	50.0
6×	Learned	120	25.83	0.118	56.5	60.0
6×	Unnormalized	60	73.33	0.207	66.5	74.5

and none flipping sign, the shrinkage concentrating on the subway whose finer 424-unit panel leaves the learned arms the most room. Intermediate rungs and the per-cell verdicts drawn in Figure L1 run on the three New York systems alone, 60 cells with RevIN and Dish-TS in the learned pool, where 40 persist, 19 shrink, 1 stays unresolved, and no cell changes sign. Those verdicts are computed under the same fixed origin-frozen rolling-8 anchor that carries every headline number, never an anchor picked per cell, so the selection-free design of Section 3.3 holds here without exception. The subway contributes 15 of the 19 shrinking cells. A further check bears on comparability there. Handed the larger budget from the start, a tuner would mostly have kept its learning rate, since re-selecting on the adjacent validation block at every budget recovers the largest-budget pick in 167 of the 192 cell-budgets that sweep holds – three systems by four backbones by four arms by four rungs – and of the 25 that differ, 10 fall on Dish-TS, 10 on our own anchored arm, 4 on RevIN and 1 on the unnormalized control, leaving re-selection risk with neither side alone.

M Limitations and Future Work

Five systems across two cities at hourly resolution carry the evidence, which is more than one city and less than a general claim. The four ride-based panels all come from one provider class, a single aggregation convention covers every system, and one split design runs through all five, so what has been tested is transfer across cities and not transfer across data-collection regimes. Two lattices are each shared by two systems, leaving the five panels less independent than their count suggests, and beyond them nothing – no unit, no city – is held out. An interface of this kind does worst wherever hour-of-week structure is weaker than in these panels.

Benchmark family is a limit of the same kind as city, and it deserves the same plainness. Each panel here is urban demand at hourly resolution where the train-fit climatology by itself covers 89.9 to 94.4 percent of held-out variance, the regime a climatology anchor was built for. Nowhere in this study does a standard long-horizon forecasting benchmark run – not the electricity, traffic, weather or exchange-rate panels the normalization literature reports on. That makes Section 5.3's second reading, that part of the family's reported progress may be progress at lifting a handicap, a hypothesis about those benchmarks and not a measurement of them, with the calibration runs named below supplying the first evidence on that axis.

Per-unit history is what the recommended interface needs and what the normalizers it is measured against do not. Equation 1 asks for that unit's training split and Equation 2 for eight weeks of it at the forecast origin, leaving no anchor to form for a newly opened station, a re-drawn zone or a unit whose record falls short of the window, while RevIN, Dish-TS and frequency-adaptive normalization take their statistics off the input window and apply to an arbitrary new series. Here lies the main practical limit on the recommendation. With a cold start as the operating case the comparison made here does not apply, and the obvious fallback, a neighbourhood or system-level table held until a unit builds its own history, goes untested on this evidence.

The replication in Section 4.4 is ours, not an outside party's. Its threshold was pre-declared before the Chicago cells were scored, true, but the same authors declared it, its reference is the New York grid itself rather than an external registry, and the full five-system grid ran two epoch budgets where the three-system New York sweep ran four. Three planned measurements come out together: the tail quantiles and the seed-dispersion ratios carried by the CPU reference run, neither recomputed for the two transformer backbones or for either Chicago system, plus the per-unit win rate, which goes unreported on this evidence base. Nowhere, then, does this paper make a tail-reliability claim, a seed-stability claim or a per-unit claim.

Four backbones make a short capacity ladder. Over 27k, 57k, 309k and 682k parameters the $R(h)$ column fails to increase. Under the recommended anchor the entries above zero are the 27k recurrent head and the 682k patch transformer, while the 57k linear decomposition and the 309k inverted transformer fall below, so nothing here separates capacity from tokenization, architecture family, covariates or decoder, and no threshold gets located. A ladder varying size with everything else held fixed is the next step.

Answering the training-budget objection rather than deferring it costs part of the margin. A shared epoch cap by itself leaves many learned runs stopping while still descending, so Section 4.5 reruns the whole grid under patience 8 on both sides of every comparison. Under-training was real, and the remedy returns a median 5.15 percent of MAE to the learned arms against 0.44 to the anchored ones; the median margin narrows from 37.62 percent to 36.19, 27 of 100 cells give up a material share, and none turns over. The ceiling stays a bound of its own, since six times a cap is not infinity, 212 of the 740 recorded runs still finish on a cap, and no budget large enough to exhaust every arm in every cell was run.

Comparator coverage is a limit in its own right. Wherever this paper writes "the best learned per-instance normalizer", the set maximised over holds three members: RevIN, Dish-TS and frequency-adaptive normalization. Section 2 names contrastive disentangling (Pang et al., 2027) and series stationarization (Liu et al., 2022), neither of them run, so every margin here reads as a margin over those three and not over the family's strongest conceivable member.

The adjacent gap is comparator fidelity. With both transformer backbones our reimplementations, the Dish-TS arm an adaptation and the frequency-adaptive arm a reimplementation from paper and reference code, one calibration run of each against its published number on a standard long-horizon benchmark would separate a weak comparator from a weak rebuild. No such check sits in this paper, and the backbones in question carry the grid's widest margins. Naming the missing experiment is easy. Run the reimplemented spectral arm, the Dish-TS adaptation and both rebuilt transformers once each on one benchmark drawn from the spectral arm's own paper, at 96, 192, 336 and 720 steps, and print the published number next to ours. Until that table exists, our spectral arm placing behind RevIN and Dish-TS in 77 of 100 cells here, the reverse of the ordering its source paper reports, admits a reading about the method and a reading about our rebuild, and this study cannot decide between them. The headline is untouched either way, since the verbatim RevIN layer loses 100 of 100 cells at a median of 38.45 percent.

The five-seed gate screens, it does not test, and its strength is worth stating. With five seeds and a population standard deviation, the threshold comes to $\sqrt{5} \approx 2.236$ standard errors of the mean margin, applied cell by cell; a sample standard deviation would lift the bar by $\sqrt{5/4} \approx 1.118$, which leaves the screen as run roughly 12 percent laxer than a reader assumes. Wilcoxon p-values over overlapping rollout origins stay descriptive throughout, inference resting with the week-block bootstrap, that gate and Holm-corrected paired seed-level tests; a recheck with our own bootstrap recovers the sign and the exclusion of zero in every cell examined, even where individual interval endpoints move by as much as roughly 2.3 MAE units once an independent random stream is used.