FORECAST PRIVATELY, CORRECT MINIMALLY: BOUNDED RELATIONAL INTERVENTIONS FOR MULTIVARIATE TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-variable information can improve multivariate forecasting, yet an unrestricted mixer can also replace reliable channel-private dynamics and propagate local faults to healthy channels. We ask a narrower question: how much relational intervention is sufficient? We introduce the Minimum-Intervention Relational Forecaster (MIRF), which first produces a complete channel-private temporal representation and then applies one identity-start residual correction. A $C \to r \to C$ factorization limits the correction to r channel directions, while spectral caps and a bounded scalar gain limit its magnitude. This design uses one ordinary MAE objective and no auxiliary routing or decomposition losses. Under a fixed look-back of 96 and four forecast horizons, MIRF ranks first in 12 of 16 dataset-average MSE/MAE cells across eight benchmarks and a compact recent baseline set. Controlled relation misalignment and sensor-outage tests show why the constraint matters. On Traffic with 50% channel outage, the selected bounded correction incurs 14.11% untouched-channel MSE spillover, compared with 41.90% for an unbounded low-rank correction and 111.72% for an unbounded dense correction. At 862 channels, rank four uses 207.8$\times$ fewer relational parameters and runs 37.7$\times$ faster at inference than the matched dense correction. These results support a minimum-intervention principle: relational structure is most useful as a controlled amendment to a private forecast, rather than as its unrestricted replacement.

1 INTRODUCTION

Multivariate time-series forecasting must reconcile two facts. Each variable has temporal structure that can often be forecast independently, while synchronized variables can provide additional predictive information. Channel-independent models protect the first fact (Zeng et al., 2023; Nie et al., 2022). Channel-dependent models exploit the second (Liu et al., 2024; Ye et al., 2026). The usual design question is therefore how to extract richer cross-variable relationships.

We study a different failure mode. A learned relationship can improve clean forecasts yet become harmful when synchrony changes, a sensor fails, or the training correlation is incidental. Dense mixing then gives one corrupted channel many paths into otherwise healthy predictions. The same problem appears during ordinary training: a sufficiently expressive relation module can overwrite a competent private predictor instead of correcting its residual error. More interaction capacity is therefore not automatically safer or more accurate.

Figure 1 makes this distinction concrete. A private predictor keeps a local fault local, but it cannot exploit useful synchrony. An unrestricted mixer can improve clean prediction, yet it also gives that fault direct routes into healthy outputs. MIRF takes the middle ground. It preserves one complete forecast per channel, then permits only a low-rank, gain-bounded relational amendment. The two arrows above panel (c) state the resulting design objective: preserve reliable private dynamics while controlling cross-channel fault propagation.

Recent work makes the tension explicit. CoRA represents time-varying and time-invariant correlations with low-rank adapters for pretrained forecasters (Cheng et al., 2026). VPNet argues that local cross-variate interactions can replace global attention while scaling linearly in the channel count (Liu et al., 2026a). ChannelTokenFormer targets dependency, asynchronous sampling, and missing blocks

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

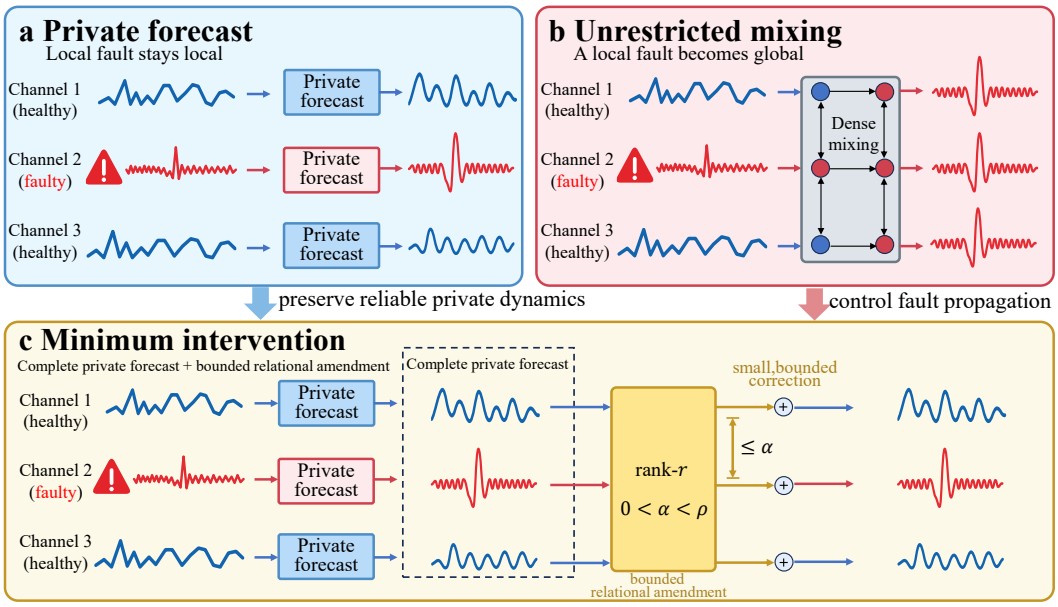


Figure 1: **Why relational information should amend rather than replace a private forecast.** (a) Channel-private forecasting isolates a faulty channel, so healthy outputs remain unaffected, but useful cross-channel information is unavailable. (b) Unrestricted mixing creates dense propagation paths; a fault originating in channel 2 can therefore distort channels 1 and 3. (c) MIRF first retains a complete private forecast for every channel, then adds a small correction constrained to r channel directions and gain $0 < \alpha < \rho$. The design preserves reliable private dynamics while limiting fault propagation.

jointly (Jang et al., 2026). The 2025–2026 Pattern Recognition literature likewise explores dynamic graphs, multi-period interaction, frequency separation, and collaborative variable analysis (Hou et al., 2025; Tan et al., 2025; Ye et al., 2026; Hu et al., 2025; Tan et al., 2026). These approaches improve relationship representations. They do not by themselves answer how strongly those relationships should be allowed to alter a complete private forecast.

Our starting hypothesis is deliberately restrictive. Cross-variable modeling should be treated as a budgeted intervention, not an alternative forecasting path. A model should first preserve a complete private representation. Its relational module should start as the identity, modify only a low-dimensional set of channel directions, and have an explicit gain ceiling. This hypothesis is falsifiable. It fails if a dense or unbounded correction gives a better clean-accuracy and relation-shift trade-off, or if a small rank cannot retain competitive benchmark accuracy.

We operationalize this hypothesis with the Minimum-Intervention Relational Forecaster (MIRF). A shared temporal MLP is applied separately to each channel. One $C \to r \to C$ residual then mixes channels at every latent coordinate. The expansion factor is zero initialized, making the initial network exactly equal to the private predictor. Each factor is spectrally capped, and the residual gain is parameterized as $\rho\sigma(g)$. The channel-private decoder maps the corrected latent to the horizon; its output difference from decoding the uncorrected latent defines the relational amendment. The final model uses instance normalization as a standard component, but no patching, decomposition, router, or auxiliary loss.

Our evidence closes the loop from problem to mechanism. First, MIRF achieves the best displayed dataset-average value in 12 of 16 MSE/MAE cells across eight datasets. Second, controlled channel misalignment and outage tests isolate error propagation into untouched channels. Third, matched dense, unbounded, private-only, rank, and initialization controls identify which constraint addresses which failure. Finally, channel-scaling profiles verify the intended $O(Cr)$ relational cost. The residual spectrum itself is not rank four, so we interpret rank as an intervention budget rather than a data-generating assumption.

Our contributions are:

- We formulate cross-variable forecasting as a minimum-intervention problem and distinguish a low-rank update budget from a low-rank covariance assumption.
- We introduce an identity-start, rank- and gain-bounded relational correction on top of a complete channel-private forecaster, with an exact Jacobian rank statement for the correction map.
- We test the claimed mechanism directly through relation misalignment, sensor outage, matched dense and unbounded controls, rank and initialization studies, and measured channel scaling.

2 Related work

Private temporal modeling. DLinear showed that simple temporal projections remain strong long-horizon baselines (Zeng et al., 2023). PatchTST combines channel independence with patch tokens (Nie et al., 2022), while RevIN reduces instance-level distribution shift (Kim et al., 2021). SRSNet learns a selective patch representation (Wu et al., 2026), and Amplifier emphasizes low-energy frequency components (Fei et al., 2025). These methods motivate a competent private path. In MIRF, instance normalization is a standard component, not a contribution, and the final validation gate removed patching and multiresolution synthesis.

Cross-variable modeling. iTransformer treats variables as tokens for global attention (Liu et al., 2024). Recent Pattern Recognition methods model dynamic graphs, signed relationships, periodic interactions, and global context (Hou et al., 2025; Ye et al., 2026; Liu et al., 2026b; Tan et al., 2026). Delayformer converts each variable into a delay-embedded state, cross-represents those states with a shared ViT encoder, and uses variable-specific linear decoders (Wang et al., 2026). VPNet instead uses local convolutions over a variate-patch field (Liu et al., 2026a). Our distinction is not the first use of channel dependence, channel-independent backbones, or linear channel scaling. We constrain the *update to a complete private representation* and evaluate the resulting propagation behavior.

Low-rank and robust forecasting. Low-rank forecasting is also not new. MixLinear uses rank-constrained spectral filters for temporal compression (Ma et al., 2026), and CoRA factorizes correlation adapters for foundation models (Cheng et al., 2026). Their ranks compress a temporal or correlation representation. Our rank bounds the channel directions that the relational residual can change. The gain ceiling adds a separate magnitude constraint. SEER replaces low-quality patches (Qiu et al., 2026), while ChannelTokenFormer addresses missing blocks and asynchrony (Jang et al., 2026). MIRF does not claim a complete missing-data solution. Its interventions are controlled diagnostics of cross-channel fault propagation.

3 Minimum-intervention relational forecasting

Figure 2 follows the complete forward pass and separates MIRF into three named modules. **Channel-Private Temporal Forecasting** constructs a full forecast without channel exchange. **Directed Low-Rank Relation Correction** computes a bounded latent amendment through a $C \to r \to C$ path. **Bounded Residual Fusion** expresses the corrected output as the private forecast plus the amendment induced by that latent change. This order is important: relational information can refine a valid prediction, but it never becomes the only route from history to forecast.

3.1 Channel-Private Temporal Forecasting

Let $X \in \mathbb{R}^{B\times L\times C}$ contain B windows, L historical steps, and C channels. The target is $Y \in \mathbb{R}^{B\times H\times C}$. We normalize each input channel independently,

$$X'_{b,:,c} = \frac{X_{b,:,c} - \mu_{b,c}}{s_{b,c}}, \quad \mu_{b,c} = \frac{1}{L}\sum_{t} X_{b,t,c}, \tag{1}$$

where $s_{b,c}$ is the corresponding standard deviation with 10^{-5} stabilization. The statistics are detached and reversed after decoding.

A shared temporal encoder $E : \mathbb{R}^L \to \mathbb{R}^d$ is applied independently to every channel,

$$Z_c = E(X'_{:,c}) = \mathrm{Drop}\big(\mathrm{LN}\big(\mathrm{GELU}(W_e X'_{:,c} + b_e)\big)\big). \tag{2}$$

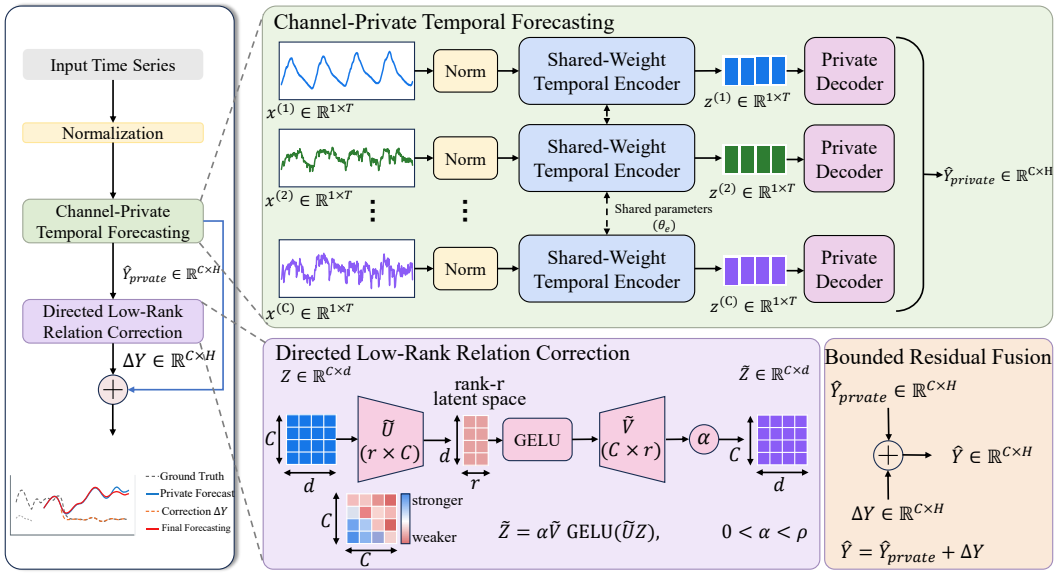


Figure 2: **MIRF architecture and the correspondence between overview and modules.** The left column shows the end-to-end flow; the enlarged panels use the same three module names. **Channel-Private Temporal Forecasting** normalizes, encodes, and decodes each channel independently with shared weights, producing the complete forecast $\widehat{Y}_{\text{private}}$. **Directed Low-Rank Relation Correction** compresses the channel axis from C to r, applies GELU, and expands it back to C; spectral caps limit both factors, and $\alpha < \rho$ limits the residual gain. **Bounded Residual Fusion** adds the induced output correction ΔY to $\widehat{Y}_{\text{private}}$. Zero initialization of the expansion factor makes $\Delta Y = 0$ at the start of training.

The same encoder parameters are reused for all channels, but no channel is present in another channel's computation. A shared decoder $D : \mathbb{R}^d \to \mathbb{R}^H$ is likewise executed channel by channel, giving a complete private forecast $\widehat{Y}'_{\text{private},c} = D(Z_c)$. Weight sharing controls parameter growth; private execution guarantees that corrupting channel i cannot change $\widehat{Y}'_{\text{private},j}$ for $i \neq j$.

For a concrete Traffic example with $B = 32$, $L = 96$, $C = 862$, $d = 200$, and $H = 96$, the input has shape $32 \times 96 \times 862$. Independent encoding yields $Z \in \mathbb{R}^{32\times 862\times 200}$, and private decoding yields $\widehat{Y}'_{\text{private}} \in \mathbb{R}^{32\times 96\times 862}$. Thus, the private path already solves the forecasting task before any relation module is used.

3.2 Directed Low-Rank Relation Correction

The second module receives $Z \in \mathbb{R}^{B\times C\times d}$. Define the source projection $U \in \mathbb{R}^{r\times C}$ and target projection $V \in \mathbb{R}^{C\times r}$. Their order gives an explicit direction: U gathers information from C source channels into r shared coordinates, while V distributes those coordinates back to C target channels. We cap each factor without increasing its spectral norm:

$$\widehat{U} = \frac{U}{\max(1, \|U\|_2)}, \qquad \widehat{V} = \frac{V}{\max(1, \|V\|_2)}. \tag{3}$$

The directed latent amendment and corrected representation are

$$\Delta Z = \alpha\, \widehat{V}\ \mathrm{GELU}\Big(\widehat{U}\,\mathcal{N}(Z)\Big), \qquad Z^+ = Z + \Delta Z, \qquad \alpha = \rho\,\sigma(g), \tag{4}$$

where $\mathcal{N}$ is channel-separable LayerNorm over the latent width. Equation 4 is evaluated independently at every latent coordinate. In the Traffic example, U maps $32 \times 862 \times 200$ to $32 \times 4 \times 200$, and V restores $32 \times 862 \times 200$. Rank therefore limits how many shared channel directions can be changed, whereas ρ limits the change magnitude. We initialize $V = 0$, so $\Delta Z = 0$ and $Z^+ = Z$ exactly before optimization.

Table 1: Problem–design–test map. Each constraint has a distinct failure mode and control.

Observed risk	Design response	Structural effect	Matched evidence
distribution shift	instance normalization	per-window stabilization	standard component
private dynamics overwritten	complete private path	forecast exists without mixing	private-only control
too many propagation routes	rank-r residual	direction budget	rank/dense controls
large harmful corrections	spectral caps + gain	magnitude budget	bounded/unbounded tests
unstable optimization start	zero expansion	exact identity at step zero	initialization control

For a fixed coordinate k, consider the correction as a function of $q_k = \mathcal{N}(Z)_{:,k}$. Its Jacobian is

$$J_{\mathrm{corr},k} = \alpha \widehat{V} D_k \widehat{U}, \qquad D_k = \mathrm{diag}\Big(\mathrm{GELU}'(\widehat{U} q_k)\Big). \tag{5}$$

Therefore $\mathrm{rank}(J_{\mathrm{corr},k}) \le r$ and $\|J_{\mathrm{corr},k}\|_2 \le \rho L_{\mathrm{GELU}}$. Across all d coordinates, the correction rank is at most rd. This statement concerns the map from the normalized latent $q = \mathcal{N}(Z)$; it does not assert a global Lipschitz constant through LayerNorm.

Table 1 states the model logic: every nonstandard component has one structural role and one matched diagnostic. The table also separates standard instance normalization from the claimed contribution.

3.3 Bounded Residual Fusion

The final module decodes the corrected latent while retaining the private forecast as an explicit reference. Using the same channel-private decoder D, define

$$\widehat{Y}'_{\mathrm{corr},c} = D(Z_c^{+}), \qquad \Delta Y'_c = \widehat{Y}'_{\mathrm{corr},c} - \widehat{Y}'_{\mathrm{private},c}, \qquad \widehat{Y}'_c = \widehat{Y}'_{\mathrm{private},c} + \Delta Y'_c. \tag{6}$$

This is an exact decomposition, not a second forecasting head: $\widehat{Y}'_c = D(Z_c + \Delta Z_c)$. It makes the intervention measurable in output space. At initialization, $\Delta Z = 0$ implies $\Delta Y' = 0$, and the full model equals the private forecaster. We reverse Equation 1 to obtain $\widehat{Y}$. Training minimizes only $\mathcal{L} = \|Y - \widehat{Y}\|_1/(BHC)$, without contrastive, spectral, decomposition, or routing losses.

The relational parameter count is $2Cr + 1$, compared with C^2 for a dense channel map. Its channel-mixing multiply-accumulate count is $2Crd$ rather than $C^2 d$. Spectral capping requires singular-value computation and is therefore not free; Section 5.2 reports the measured overhead.

4 Experimental protocol

4.1 Datasets, metrics, and splits

We evaluate ETTh1, ETTh2, ETTm1, ETTm2, Weather, Solar, Electricity, and Traffic. Every experiment uses multivariate-to-multivariate forecasting, $L = 96$, stride one, and $H \in \{96, 192, 336, 720\}$. We report MSE and MAE after reversing input normalization.

ETT follows the canonical 12/4/4-month boundaries: 8,640/2,880/2,880 hourly points and 34,560/11,520/11,520 15-minute points. The other datasets use chronological splits close to 70/10/20. Exact point and window counts are provided in Appendix B. Normalization is fitted or computed without target leakage.

4.2 Model selection and optimization

The architecture uses $r = 4$ throughout. Temporal width is 100 for ETT and Weather, and 200 for Solar, Electricity, and Traffic. The hidden multiplier is two at $H = 96$ and one otherwise. These rules were frozen by capacity gates before the final matrix. We search only $\rho \in \{1, 2, 4\}$ per dataset. One complete four-horizon row is selected by its mean validation loss; test metrics are not consulted and horizons are never spliced across configurations. Selected ρ values are 1,1,2,1,4,1,4,2 in dataset order.

We use Adam, learning rate 10^{-3} with the repository's type-3 decay, batch size 32, at most 30 epochs, and patience five. The training seed is 2021.

Table 2: Dataset-average multivariate forecasting results for $L = 96$ and $H \in \{96, 192, 336, 720\}$. Lower is better. Red and blue denote the best and second-best displayed values; ties are counted for every tied model.

Models	MIRF (ours)		MixLinear (2026)		Delayformer (2026)		SRSNet (2025)		Amplifier (2025)		PatchTST (2023)		DLinear (2023)	
Metrics	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
ETTh1	0.441	0.430	0.448	0.435	0.438	0.438	0.442	0.433	0.449	0.437	0.469	0.455	0.461	0.458
ETTh2	0.374	0.394	0.383	0.402	0.377	0.399	0.376	0.401	0.389	0.411	0.387	0.407	0.563	0.519
ETTm1	0.391	0.387	0.441	0.423	0.377	0.398	0.385	0.395	0.383	0.396	0.387	0.400	0.403	0.408
ETTm2	0.277	0.318	0.296	0.336	0.281	0.328	0.284	0.330	0.280	0.326	0.281	0.326	0.354	0.402
Weather	0.239	0.261	0.283	0.303	0.245	0.275	0.249	0.277	0.247	0.275	0.259	0.280	0.265	0.315
Solar	0.237	0.234	0.534	0.449	0.248	0.295	0.250	0.280	0.222	0.256	0.270	0.307	0.330	0.401
Electricity	0.179	0.267	0.224	0.297	0.188	0.286	0.189	0.276	0.173	0.268	0.216	0.304	0.225	0.319
Traffic	0.474	0.293	0.636	0.380	0.518	0.319	0.494	0.307	0.485	0.315	0.555	0.361	0.625	0.383
1st Count	4	8	0	0	2	0	0	0	2	0	0	0	0	0

4.3 Baselines and evidence tiers

The compact main set contains MixLinear (ICLR 2026), Delayformer (Advanced Science 2026), SRSNet (NeurIPS 2025), Amplifier (AAAI 2025), PatchTST (ICLR 2023), and DLinear (AAAI 2023). MixLinear's paper uses $L = 720$, so we rerun its official code at $L = 96$ with validation-selected hyperparameters and its official seed 2023. Delayformer contributes its complete reported Table 1 benchmark: the same eight datasets, $L = 96$, and $H \in \{96, 192, 336, 720\}$ under ordinary per-dataset supervised training; its separate cross-domain foundation-model experiment is not used.

4.4 Mechanism tests

We evaluate how local channel faults affect untouched channels at $H = 96$. A fixed fraction of channels is selected with mask seeds 11, 29, and 47. *Circular shift* moves each selected input trajectory in time, preserving its marginal values while disrupting synchrony. *Mean outage* replaces the selected input with its training mean. MSE spillover is the percentage change on untouched channels relative to their clean error under the same mask. We compare the private model, bounded low-rank variants, an unbounded low-rank correction, and an unbounded dense correction.

For scaling, we profile matched relation modules at $C \in \{7, 21, 137, 321, 862\}$ on one RTX 4090. CUDA events measure inference and training-step latency after warm-up. The bounded-low-rank and bounded-dense variants differ only in relation rank. The bound-overhead comparison uses the same low-rank factorization with and without spectral caps.

5 Results

5.1 Forecasting accuracy

Table 2 reports arithmetic means over all four horizons. MIRF is best in 12 of 16 dataset-metric cells (75%) and has an average rank of 1.44. It leads both metrics on ETTh2, ETTm2, Weather, and Traffic, and leads MAE on ETTh1, ETTm1, Solar, and Electricity. The largest displayed MAE improvement over the strongest baseline is 8.59% on Solar. Its four losses are ETTh1 MSE, ETTm1 MSE, Electricity MSE, and Solar MSE. The maximum deficit is 6.76% on Solar MSE, so the result is broad but not uniform.

The 2026 comparisons confirm that MIRF's gains are not an artifact of comparing only older methods. MIRF outperforms the official-code MixLinear rerun on all 16 dataset-average cells and Delayformer on 14 of 16. The complete horizon table in Appendix C exposes every retained cell rather than only favorable averages.

5.2 Controlled relation shift

Figure 3 tests the mechanism on Traffic. The bounded $\rho = 2$ model obtains 0.4366 clean MSE, compared with 0.4742 for the private control. At 50% circular shift, its untouched-channel MSE

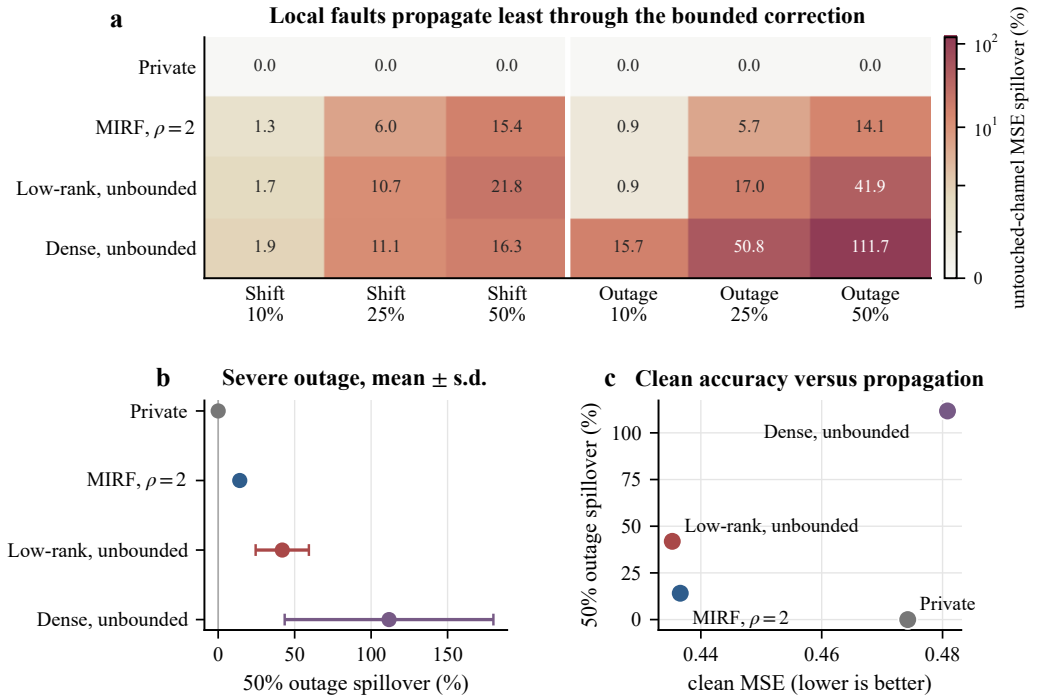


Figure 3: **Mechanism evidence on Traffic.** (a) Mean untouched-channel MSE spillover across relation shifts and sensor outages; each cell reports the mean over three fixed masks. (b) Mean and sample standard deviation under 50% outage. (c) Clean MSE versus severe-outage propagation. Private forecasting has zero spillover but higher clean error; MIRF retains most of the clean benefit while reducing propagation relative to unrestricted alternatives.

spillover is 15.43%, below 21.76% for the unbounded low-rank correction. At 50% mean outage, the separation grows to 14.11% versus 41.90%. The unbounded dense correction reaches 111.72% spillover and is also worse on clean data.

The private model has zero spillover by construction because it never mixes channels. It therefore remains the safest severe-corruption control, but its clean error is higher. The Pareto panel makes the intended claim precise: bounded mixing retains most of the clean benefit of cross-channel information while reducing propagation relative to unrestricted alternatives. It does not dominate the private model under every fault severity.

Electricity gives a complementary boundary. At 50% outage, $\rho = 1$ limits spillover to 19.48%, while the selected clean-accuracy setting $\rho = 4$ reaches 211.84%. A single dataset-level gain cannot adapt to sample-specific reliability. This finding motivates abstention or reliability-conditioned gains, but we do not add a router after observing the test behavior.

5.3 Rank, Initialization, and Scaling

On Weather at $H = 96$, ranks 2, 4, 8, and 16 produce MSE values 0.1539, 0.1489, 0.1501, and 0.1520. The intermediate rank four is best on both MSE and MAE. This result supports finite intervention capacity, not the claim that the residual covariance has rank four. Indeed, only 37.0% and 40.0% of the Electricity and Traffic residual energy lies in the first four singular directions. Their participation ratios are 22.1 and 16.3.

Figure 4 separates three claims that would otherwise be easy to conflate. The rank sweep tests whether intervention capacity should grow monotonically. The initialization control tests the identity-start optimization prior. The residual spectrum tests whether rank four can be interpreted as an intrinsic property of the data. The three panels support a budgeted correction while rejecting an intrinsic low-rank covariance story.

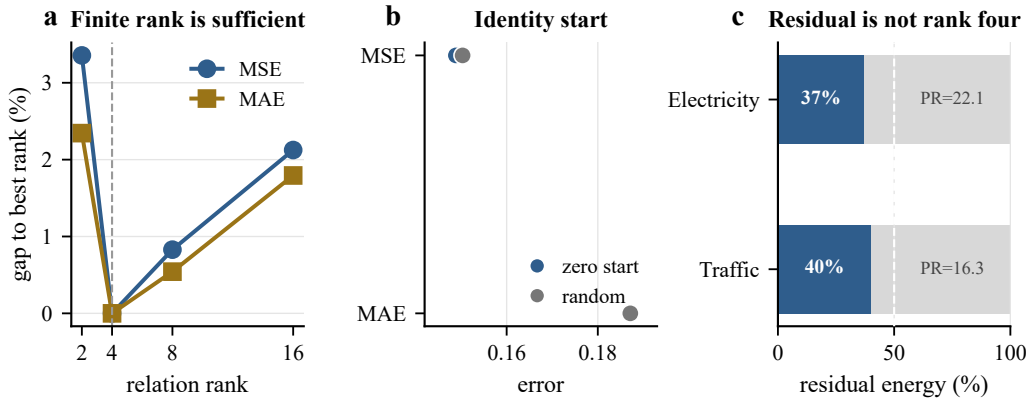


Figure 4: **Matched controls identify what the rank and identity start do.** (a) Weather $H = 96$ error relative to the best tested rank; both metrics select $r = 4$. (b) Paired errors for zero and random expansion initialization under the same architecture and protocol. (c) The first four singular directions contain only 37–40% of private-model residual energy; PR denotes participation ratio. The rank therefore acts as an intervention budget, not an estimated residual rank. All controls use one training seed.

Zero expansion initialization gives 0.1489/0.1870 MSE/MAE on Weather, compared with 0.1503/0.1871 for matched random initialization. The difference is modest, but the zero start supplies an exact optimization prior and never requires the private path to be relearned at initialization.

At $C = 862$, rank four uses 207.8× fewer relational parameters and 215.5× fewer channel-mixing operations than the matched dense correction. Measured inference and training-step speedups are 37.7× and 21.0× (Figure 5c). Spectral capping makes bounded low rank 3.39× slower at inference than the matched unbounded factorization. The method trades some factorization speed for a testable gain bound.

Figure 5 joins the main boundary and the intended computational mechanism. On Electricity, reducing the gain ceiling suppresses severe-outage propagation, although the clean-validation setting prefers $\rho = 4$. Across channel counts, the factorized correction increasingly separates from dense mixing in parameters, operations, and measured latency. These panels show both sides of the design contract: the gain is a safety budget, while the rank factorization is the scaling budget.

6 Discussion and Limitations

The evidence supports a narrow principle. A complete private predictor is a useful default, and cross-variable structure can be added as a controlled residual. Rank and gain solve different problems. Rank restricts how many channel directions can be changed, whereas the gain ceiling restricts the amplitude of those changes. Zero initialization addresses early optimization by preserving the private model exactly. The private decoder prevents later temporal projection from introducing new channel mixing.

The experiments also reject stronger stories. Rank four is not an empirical rank estimate of the data. A bounded correction is not always more accurate than an unbounded one on clean inputs. Private forecasting remains preferable under sufficiently severe faults, and Electricity shows that a gain chosen for clean validation can propagate outage errors. The model also does not address asynchronous sampling, arbitrary missing blocks, or online concept drift.

The most direct extension is a reliability-aware gain that can abstain from relation mixing at inference. Such a mechanism must be selected using validation evidence and evaluated without target access. We leave it to future work rather than adding a post hoc router to the present model.

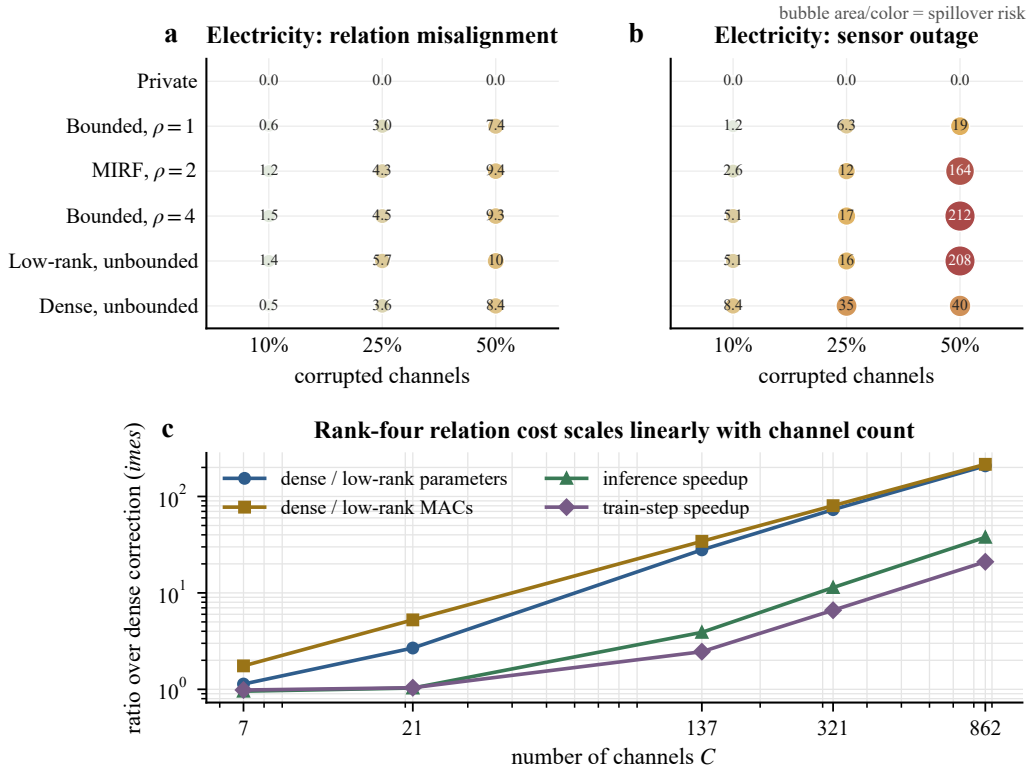


Figure 5: **Boundary condition and matched scaling.** (a,b) Electricity untouched-channel spillover under circular shift and mean outage; bubble area, colour, and labels encode the mean over three fixed masks. The clean-validation setting $\rho = 4$ can propagate a severe outage, whereas $\rho = 1$ is more conservative. (c) Dense-over-low-rank parameter/MAC ratios and measured low-rank speedups as channel count increases. Timing uses CUDA events after warm-up on one RTX 4090.

7 Conclusion

MIRF treats cross-variable structure as a minimal intervention on a complete private forecast. Its identity start, rank budget, and gain ceiling have separate structural roles and corresponding controls. Across eight long-horizon benchmarks, the model leads 12 of 16 dataset-average cells in the compact recent comparison. Controlled faults and matched scaling show that the same design reduces error propagation and relation cost. The conclusion is therefore conditional but useful: when private temporal dynamics are strong, relational modeling should correct them with explicit budgets rather than replace them without constraint.

Ethics and Broader Impact

This work uses public benchmark time series and introduces no human-subject or personally identifiable data. More robust forecasting can support resource planning, but benchmark accuracy does not establish safety in critical infrastructure. Deployment should monitor distribution change, missingness, and channel reliability, especially because a fixed relational gain can propagate severe sensor faults.

AI Use Statement

Generative AI assisted with literature discovery and LaTeX preparation. The authors verified the sources, code, numerical artifacts, figures, and claims, and remain responsible for the submitted content.

REFERENCES

Hanyin Cheng, Xingjian Wu, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Guo Guo. Cora: Boosting time series foundation models for multivariate forecasting through correlation-aware adapter. In *International Conference on Learning Representations*, volume 2026, pp. 106657–106680, 2026.

Jingru Fei, Kun Yi, Wei Fan, Qi Zhang, and Zhendong Niu. Amplifier: Bringing attention to neglected low-energy components in time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 11645–11653, 2025.

Mingjie Hou, Zhenyu Liu, Guodong Sa, Yueyang Wang, Jiacheng Sun, Zhinan Li, and Jianrong Tan. Parallel multi-scale dynamic graph neural network for multivariate time series forecasting. *Pattern Recognition*, 158:111037, 2025.

Ao Hu, Liangjian Wen, Jiang Duan, Yong Dai, Dongkai Wang, Shudong Huang, Jun Wang, and Zenglin Xu. Fdnet: high-frequency disentanglement network with information-theoretic guidance for multivariate time series forecasting. *Pattern Recognition*, pp. 112810, 2025.

Jinkwan Jang, Hyungjin Park, Jinmyeong Choi, and Taesup Kim. Towards robust real-world multivariate time series forecasting: a unified framework for dependency, asynchrony, and missingness. In *International Conference on Learning Representations*, volume 2026, pp. 153495–153528, 2026.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations*, 2021.

Kun Liu, Renjun Jia, Ruifeng Yang, Xirui Zeng, Yuqi Liang, and Cen Chen. Are global dependencies necessary? scalable time series forecasting via local cross-variate modeling. In *International Conference on Learning Representations*, volume 2026, pp. 102794–102817, 2026a.

Xiangsen Liu, Jinchang Ren, Hongming Zhang, and Erlei Zhang. Gcmnet: A global context mamba network for long-term time series forecasting. *Pattern Recognition*, pp. 113287, 2026b.

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*, volume 2024, pp. 11116–11140, 2024.

Aitian Ma, Dongsheng Luo, and Mo Sha. Mixlinear: Extreme low resource multivariate time series forecasting with $0.1k$ parameters. In *International Conference on Learning Representations*, volume 2026, pp. 77764–77789, 2026.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Xiangfei Qiu, Xvyuan Liu, Tianen Shen, Xingjian Wu, Hanyin Cheng, Bin Yang, and Jilin Hu. Seer: Transformer-based robust time series forecasting via automated patch enhancement and replacement. *arXiv preprint arXiv:2602.00589*, 2026.

Cheng Tan, Wenbin Zhang, and Mingjing Du. Dpia: Dynamic periodic interaction and adaptation for multivariate time series forecasting. *Pattern Recognition*, pp. 114310, 2026.

Gang Tan, Yueyang Wang, Ziyi Xiao, Dandan He, and Guodong Sa. From a multi-period perspective: A periodic dynamics forecasting network for multivariate time series forecasting. *Pattern Recognition*, 167:111760, 2025.

Zijian Wang, Peng Tao, and Luonan Chen. Delayformer: Spatiotemporal transformation for predicting high-dimensional dynamics. *Advanced Science*, 13(3):e14671, 2026.

Xingjian Wu, Xiangfei Qiu, Hanyin Cheng, Zhengyu Li, Jilin Hu, Chenjuan Guo, and Bin Yang. Enhancing time series forecasting through selective representation spaces: A patch perspective. *Advances in Neural Information Processing Systems*, 38:23328–23354, 2026.

Junjie Ye, Chengli Zhou, Xiaojun Zhou, Yaqun Huang, and Chunna Zhao. Cvacl-ma: Comprehensive variate analysis and collaborative learning with multi-adapter for multivariate time series forecasting. *Pattern Recognition*, 169:111936, 2026.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

Table 3: Dataset point splits used by MIRF.

Dataset	Channels	Train	Validation	Test
ETTh1/ETTh2	7	8,640	2,880	2,880
ETTm1/ETTm2	7	34,560	11,520	11,520
Weather	21	36,886	5,270	10,540
Solar	137	36,792	5,256	10,512
Electricity	321	18,412	2,631	5,261
Traffic	862	12,280	1,755	3,509

Table 4: Exact train/validation/test window counts for $L = 96$. Each entry is train/validation/test.

Dataset	$H = 96$	$H = 192$	$H = 336$	$H = 720$
ETTh1/ETTh2	8,449/2,785/2,785	8,353/2,689/2,689	8,209/2,545/2,545	7,825/2,161/2,161
ETTm1/ETTm2	34,369/11,425/11,425	34,273/11,329/11,329	34,129/11,185/11,185	33,745/10,801/10,801
Weather	36,695/5,175/10,445	36,599/5,079/10,349	36,455/4,935/10,205	36,071/4,551/9,821
Solar	36,601/5,161/10,417	36,505/5,065/10,321	36,361/4,921/10,177	35,977/4,537/9,793
Electricity	18,221/2,536/5,166	18,125/2,440/5,070	17,981/2,296/4,926	17,597/1,912/4,542
Traffic	12,089/1,660/3,414	11,993/1,564/3,318	11,849/1,420/3,174	11,465/1,036/2,790

A Structural properties

Identity start. With $V = 0$, its capped form $\widehat{V}$ is also zero. Equation 4 then gives $\Delta Z = 0$ and $Z^+ = Z$ for every input. Consequently, $\Delta Y = 0$, and both the output and private-path gradients match the channel-private model at initialization.

Rank bound. For normalized coordinate $q_k \in \mathbb{R}^C$, the correction is $c_k(q_k) = \alpha \widehat{V} \phi(\widehat{U} q_k)$. Differentiating gives $J_{c,k} = \alpha \widehat{V} D_k \widehat{U}$. The rank of a product cannot exceed the minimum factor rank, hence $\mathrm{rank}(J_{c,k}) \le r$. Stacking d coordinate maps yields rank at most rd. Channel-separable LayerNorm mixes latent coordinates inside each channel, so the statement is deliberately made with respect to its output q.

Gain bound. Because $\left\|\widehat{U}\right\|_2 \le 1$, $\left\|\widehat{V}\right\|_2 \le 1$, and $0 < \alpha < \rho$, submultiplicativity gives $\|J_{c,k}\|_2 \le \rho \|D_k\|_2 \le \rho L_{\mathrm{GELU}}$. This bound constrains the relational correction, not the identity-plus-correction map or the preceding normalization.

B Exact data splits and sample counts

Table 3 records the chronological point boundaries used by MIRF. The ETT boundaries are the canonical 12/4/4-month partitions; the other four datasets use the repository's approximately 70/10/20 chronological rule. These counts are reported explicitly because a nominal split ratio is insufficient to reconstruct the available windows.

After fixing $L = 96$, the forecast horizon determines how many complete input–target windows remain in each partition. Table 4 gives the exact counts used by every MIRF run; longer horizons reduce all three partitions without moving their chronological boundaries.

The official MixLinear loader rounds non-ETT boundaries in a different order. This changes some split endpoints by one observation. We classify those comparisons as substantively matched and preserve the official implementation instead of modifying its data code.

C Complete horizon-level results

Table 5 expands the dataset averages in the main paper into all 64 dataset–horizon combinations and reports an arithmetic-mean row for every dataset. Red and blue denote the best and second-best displayed value, respectively; ties are retained after rounding to three decimals. The final row counts

first-place entries separately for MSE and MAE, so both horizon-level performance and overall coverage are visible without hiding cells on which MIRF does not lead.

Table 5: Full multivariate forecasting results for $L = 96$. Avg is the arithmetic mean over $H \in \{96, 192, 336, 720\}$. Lower is better. Red and blue denote the best and second-best displayed values; ties are counted for every tied model.

Models		MIRF (ours)		MixLinear (2026)		Delayformer (2026)		SRSNet (2025)		Amplifier (2025)		PatchTST (2023)		DLinear (2023)	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	0.378	0.391	0.384	0.391	0.379	0.400	0.383	0.395	0.376	0.393	0.414	0.419	0.397	0.412
	192	0.431	0.421	0.457	0.441	0.428	0.434	0.433	0.422	0.442	0.430	0.460	0.445	0.446	0.441
	336	0.474	0.444	0.484	0.450	0.476	0.451	0.476	0.446	0.478	0.446	0.501	0.466	0.489	0.467
	720	0.480	0.465	0.465	0.460	0.468	0.465	0.474	0.471	0.501	0.479	0.500	0.488	0.513	0.510
	Avg	0.441	0.430	0.448	0.435	0.438	0.438	0.442	0.433	0.449	0.437	0.469	0.455	0.461	0.458
ETTh2	96	0.289	0.334	0.304	0.348	0.290	0.338	0.296	0.345	0.298	0.347	0.302	0.348	0.340	0.394
	192	0.369	0.386	0.389	0.397	0.369	0.391	0.369	0.392	0.378	0.401	0.388	0.400	0.482	0.479
	336	0.417	0.422	0.421	0.428	0.423	0.427	0.413	0.425	0.428	0.437	0.426	0.433	0.591	0.541
	720	0.421	0.436	0.420	0.437	0.427	0.441	0.425	0.444	0.452	0.460	0.431	0.446	0.839	0.661
	Avg	0.374	0.394	0.383	0.402	0.377	0.399	0.376	0.401	0.389	0.411	0.387	0.407	0.563	0.519
ETTm1	96	0.313	0.343	0.383	0.391	0.320	0.362	0.319	0.358	0.318	0.356	0.329	0.367	0.346	0.374
	192	0.370	0.373	0.413	0.405	0.356	0.382	0.359	0.381	0.362	0.381	0.367	0.385	0.382	0.391
	336	0.404	0.395	0.471	0.442	0.388	0.406	0.391	0.404	0.393	0.404	0.399	0.410	0.410	0.415
	720	0.476	0.436	0.496	0.453	0.445	0.440	0.470	0.436	0.460	0.442	0.454	0.439	0.473	0.451
	Avg	0.391	0.387	0.441	0.423	0.377	0.398	0.385	0.395	0.383	0.396	0.387	0.400	0.403	0.408
ETTm2	96	0.172	0.252	0.198	0.279	0.175	0.260	0.181	0.267	0.178	0.261	0.175	0.259	0.193	0.293
	192	0.238	0.294	0.258	0.315	0.245	0.308	0.243	0.306	0.243	0.303	0.241	0.302	0.284	0.361
	336	0.299	0.334	0.316	0.349	0.304	0.344	0.306	0.346	0.305	0.344	0.305	0.343	0.382	0.429
	720	0.398	0.392	0.414	0.402	0.401	0.400	0.407	0.399	0.393	0.397	0.402	0.400	0.558	0.525
	Avg	0.277	0.318	0.296	0.336	0.281	0.328	0.284	0.330	0.280	0.326	0.281	0.326	0.354	0.402
Weather	96	0.149	0.187	0.207	0.250	0.160	0.210	0.167	0.214	0.165	0.210	0.177	0.218	0.195	0.252
	192	0.203	0.239	0.253	0.283	0.208	0.252	0.215	0.255	0.212	0.253	0.225	0.259	0.237	0.295
	336	0.261	0.282	0.301	0.316	0.265	0.293	0.270	0.294	0.267	0.293	0.278	0.297	0.282	0.331
	720	0.343	0.336	0.370	0.361	0.346	0.346	0.346	0.344	0.344	0.342	0.354	0.348	0.345	0.382
	Avg	0.239	0.261	0.283	0.303	0.245	0.275	0.249	0.277	0.247	0.275	0.259	0.280	0.265	0.315
Solar	96	0.205	0.218	0.515	0.446	0.213	0.271	0.216	0.258	0.186	0.232	0.234	0.286	0.290	0.378
	192	0.223	0.225	0.543	0.458	0.246	0.295	0.247	0.280	0.231	0.264	0.267	0.310	0.320	0.398
	336	0.252	0.244	0.554	0.458	0.266	0.304	0.268	0.294	0.234	0.263	0.290	0.315	0.353	0.415
	720	0.269	0.250	0.523	0.433	0.267	0.309	0.268	0.290	0.238	0.265	0.289	0.317	0.357	0.413
	Avg	0.237	0.234	0.534	0.449	0.248	0.295	0.250	0.280	0.222	0.256	0.270	0.307	0.330	0.401
Electricity	96	0.146	0.237	0.210	0.282	0.158	0.260	0.161	0.252	0.149	0.245	0.195	0.285	0.210	0.302
	192	0.169	0.257	0.206	0.282	0.172	0.271	0.172	0.261	0.165	0.260	0.199	0.289	0.210	0.305
	336	0.183	0.273	0.218	0.295	0.191	0.289	0.190	0.279	0.176	0.271	0.215	0.305	0.223	0.319
	720	0.217	0.300	0.261	0.331	0.230	0.323	0.231	0.313	0.204	0.296	0.256	0.337	0.258	0.350
	Avg	0.179	0.267	0.224	0.297	0.188	0.286	0.189	0.276	0.173	0.268	0.216	0.304	0.225	0.319
Traffic	96	0.437	0.274	0.663	0.397	0.504	0.316	0.471	0.295	0.450	0.295	0.544	0.359	0.650	0.396
	192	0.463	0.289	0.611	0.368	0.513	0.319	0.480	0.300	0.489	0.311	0.540	0.354	0.598	0.370
	336	0.480	0.295	0.617	0.368	0.510	0.309	0.496	0.306	0.484	0.321	0.551	0.358	0.605	0.373
	720	0.518	0.315	0.654	0.388	0.547	0.332	0.531	0.328	0.517	0.333	0.586	0.375	0.645	0.394
	Avg	0.474	0.293	0.636	0.380	0.518	0.319	0.494	0.307	0.485	0.315	0.555	0.361	0.625	0.383
1st Count		20	37	2	2	7	0	2	1	11	2	0	0	0	0

The horizon-level view confirms that the average gains are not produced by suppressing difficult horizons. It also shows where rankings switch with forecast length, motivating the use of four-horizon averages for the headline comparison rather than a single favorable horizon.

C.1 Dataset-average comparison

Figure 6 shows the signed gap to the strongest eligible compact-set baseline for every dataset and metric. Negative values indicate lower error for MIRF. The figure makes the four MSE deficits visible alongside the twelve leading cells, complementing the exact horizon-level table without replacing it.

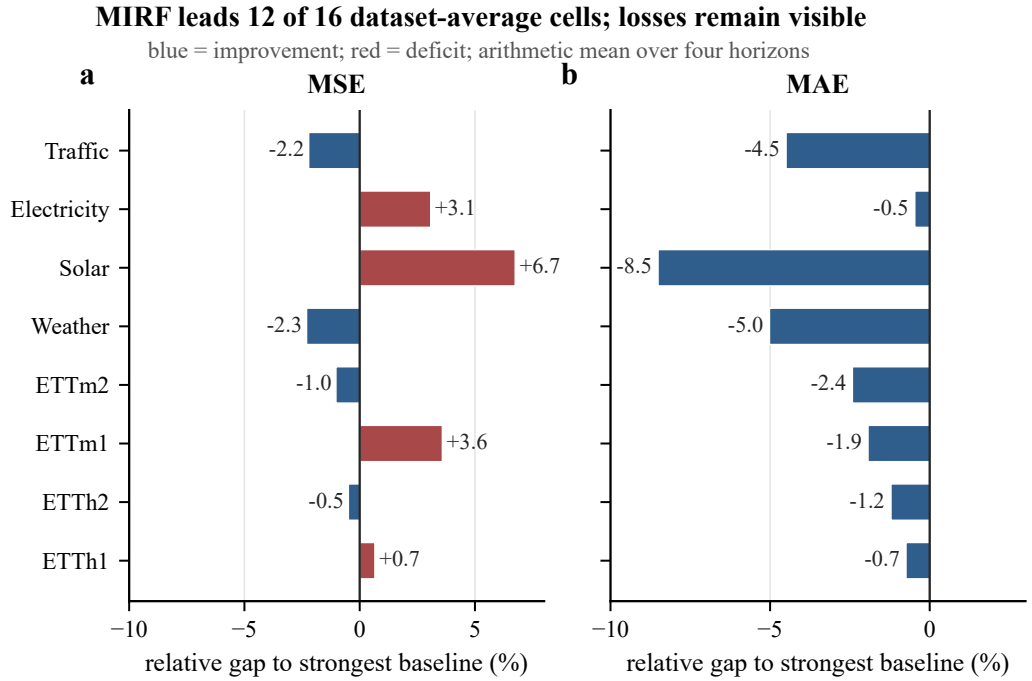


Figure 6: **Breadth of dataset-average accuracy.** Signed relative gaps compare MIRF with the strongest available compact-set baseline for each dataset and metric. Blue bars indicate improvement and red bars indicate a deficit. Values are arithmetic means over $H \in \{96, 192, 336, 720\}$ from one training seed.

D HYPERPARAMETERS AND VALIDATION SELECTION

Table 6 lists the final architecture and optimization settings. Rank, width rules, decoder multiplier, objective, and seed were frozen before the final matrix; only the dataset-level gain ceiling was selected from the declared grid.

Table 6: Frozen MIRF configuration.

Component	Value
Look-back / horizons	96 / {96,192,336,720}
Relation rank	4
Encoder width	100 (ETT, Weather); 200 (others)
Decoder hidden multiplier	2 at $H = 96$; 1 otherwise
Normalization	per-instance mean and variance
Expansion initialization	zero
Gate initialization	$g = -1$
Dropout	0
Objective	MAE
Optimizer / learning rate	Adam / 10^{-3}
Batch / epochs / patience	32 / 30 / 5
Seed	2021

The gain search never splices horizons. For each dataset, one ρ is selected by the mean validation loss across the complete four-horizon row, then that row is transferred unchanged to test evaluation. Table 7 records both the chosen value and the selection statistic.

Table 7: Dataset-level gain selection. One ρ is selected by the mean validation loss over all four horizons.

Dataset	selected ρ	validation mean
ETTh1	1	0.697863
ETTh2	1	0.402141
ETTm1	2	0.504696
ETTm2	1	0.295052
Weather	4	0.333598
Solar	1	0.196109
Electricity	4	0.242408
Traffic	2	0.259031

The selected ceilings vary from one to four, which is allowed because the rule is dataset-level, validation-only, and fixed across horizons. This choice improves fit to dataset scale and dependence strength without consulting any test metric.

E Additional mechanism evidence

E.1 Rank and initialization controls

Figure 4 in the main paper visualizes the three matched controls. Tables 8–10 provide their exact values. The U-shaped rank response rejects both "smaller is always safer" and "more relation capacity is always better" as explanations for the selected rank.

Table 8: Weather $H = 96$ rank control.

Rank	MSE	MAE
2	0.153857	0.191428
4	**0.148861**	**0.187042**
8	0.150094	0.188057
16	0.152024	0.190398

Table 9 isolates initialization while holding the architecture and protocol fixed. The numerical advantage is modest, so the paper relies on the exact identity property rather than claiming a large empirical gain.

Table 9: Matched expansion initialization on Weather $H = 96$.

Initialization	MSE	MAE
Zero (identity start)	**0.148861**	**0.187042**
Random	0.150333	0.187134

E.2 Residual spectrum is not rank four

Table 10 quantifies the spectrum summarized in Figure 4c. Top-four energy remains below 40% and the participation ratios are much larger than four, so the data do not justify an intrinsic rank-four covariance story.

Table 10: Spectrum of private-model residuals. Top-4 energy is too low to justify an intrinsic rank-four claim.

Dataset	top-4 energy	stable rank	participation ratio
Electricity	37.0%	6.31	22.1
Traffic	40.0%	4.85	16.3

E.3 ELECTRICITY RELATION SHIFT

Electricity exposes the principal boundary of a fixed dataset-level gain. Figure 5a–b in the main paper compares gain ceilings under two interventions, while panel c verifies the intended high-dimensional scaling. The clean-validation choice $\rho = 4$ is competitive without intervention but is not the safest setting under severe outage.

Table 11 supplies the exact intervention means. The private control has zero propagation by construction, and the dense/unbounded alternative is not uniformly worst; this is why the claim is a controlled trade-off rather than universal robustness.

Table 11: Electricity untouched-channel MSE spillover (%) at $H = 96$. Values are means across three fixed masks.

	Circular shift			Mean outage		
Variant	10%	25%	50%	10%	25%	50%
Private	0.00	0.00	0.00	0.00	0.00	0.00
Low-rank, unbounded	1.35	5.74	10.30	5.06	15.56	207.61
Low-rank, $\rho = 1$	**0.55**	**2.96**	**7.40**	**1.22**	**6.26**	**19.48**
Low-rank, $\rho = 2$	1.16	4.26	9.43	2.60	12.06	164.10
Low-rank, $\rho = 4$	1.47	4.46	9.30	5.08	16.55	211.84
Dense, unbounded	0.49	3.60	8.42	8.36	34.56	40.03

At 50% mean outage, $\rho = 1$ limits spillover to 19.48%, compared with 211.84% for the validation-selected $\rho = 4$. This failure motivates an input-reliability-conditioned or abstaining gain, but no such router is introduced after observing the test interventions.