TS-EDIT: A MODEL EDITING FRAMEWORK FOR ADVERSARIAL DEFENSE IN TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series foundation models (TSFMs) report strong zero-shot forecasting performance, but where they store factual associations internally that drive a particular forecast, and whether the association can be corrected after a malicious attack are open questions. While these models are increasingly deployed in industrial settings such as power grids, traffic and finance, they remain vulnerable to careful backdoor injections inducing catastrophic operational hazards. To address this, we introduce TS-Edit, a novel model editing framework for adversarial defense in TSFMs. We extend model interpretability approaches built for language models such as causal tracing, ROME, sequential ROME and MEMIT on decoder-only transformer models (TimesFM 2.5, Toto 2.0) and evaluate the framework on popular time-series benchmarks. We use masked white-box attacks (FGSM, BIM and PGD) to corrupt a single segment of the context window; causal tracing then localizes the damage across various MLP layers and a closed-form rank-one update to the identified feed-forward weight then repairs the corrupted forecast. For instance, a single edit after a PGD attack on the UCI Electricity data reduces the corrupted-forecast NRMSE by $9.2\times$ while preserving the model's behavior on clean inputs. The same localize-then-edit procedure, applied per attack, also repairs BIM and FGSM by $3.1\times$ and $1.5\times$ at negligible drift on clean-input predictions (specificity) on the edited model. Subsequent iterative editing exposes a restoration-versus-specificity trade-off. Sequential ROME gives the most consistent restoration across attacks, whereas MEMIT attains both the lowest error and the lowest specificity under the strongest attack, PGD. In every case the clean-input drift remains negligible. We further verify the robustness of the model using cross quarter attacks with random starts and varying attack types. Together, the results demonstrate that both the key-value memory interpretation of feed-forward layers and their localized editability can extend beyond language models. The experiments also demonstrate that attack salience within TSFMs can be localized, diagnosed, and repaired through surgical edits to individual weights.

1 INTRODUCTION

Time-series foundation models (TSFMs) extend the "pretrain once, forecast anywhere" paradigm to forecasting. TimesFM (Das et al., 2024), Toto (Cohen et al., 2024) Chronos (Ansari et al., 2024), Moirai (Woo et al., 2024), and recent mixture-of-expert and long-context variants (Gao et al., 2024; Liu et al., 2025; Shi et al., 2025) achieve strong zero-shot performance and are increasingly deployed in domains such as power grids, traffic, and finance. In these settings a single compromised sensor readings can distort an entire forecast horizon and its downstream decisions that depend on it, yet almost nothing is known about where inside such a model a corrupted forecast is produced.

We pose a question that has previously not been explored for TSFMs: when an adversarial perturbation corrupts the input context, where does that corruption localize within the model, and can it be removed surgically without retraining? In language models similar questions are addressed by causal tracing and rank-one editing (Meng et al., 2022; 2023), which localize a behavior to specific feed-forward layers and rewrite it in closed form, motivated by the key-value memory interpretation of those layers (Geva et al., 2021). This machinery, however does not transfer mechanically to TSFMs. It

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.

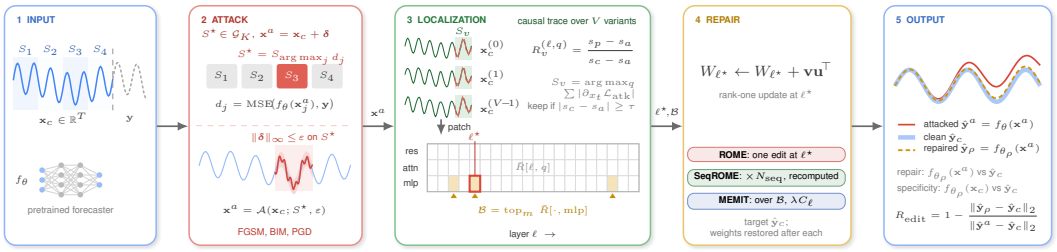


Figure 1: TS-Edit pipeline: segment selection, causal localization, and localized model editing.

presumes discrete subject tokens to locate, a factual target to write, and a vocabulary over which to define the editing objective. Establishing whether feed-forward layers in a numeric forecaster are localized and locally editable in the same sense is therefore an open question, and establishing this is the central contribution of the present work.

We introduce TS-Edit, a framework for diagnosing and repairing adversarially corrupted forecasts through localized weight edits. TS-Edit selects a context segment by its attack-induced forecast deviation, designates the clean forecast as the edit target, and utilizes NRMSE as the edit loss. Causal tracing then localizes the responsible feed-forward layer, which is edited using ROME, Sequential ROME, or MEMIT. The answer to the question above is affirmative and the effect is substantial: on UCI Electricity, a single rank-one update to one weight matrix recovers 97% of the error introduced by a PGD attack while perturbing clean-input predictions by less than 3×10^{-4}.

We establish these findings on two independently pretrained TSFMs that differ in feed-forward parameterization, normalization, forecast head, and decoding, namely TimesFM 2.5 (Das et al., 2024) and Toto 2.0 (Cohen et al., 2024), the latter at both released scales, collectively spanning 20 to 48 transformer blocks and an order of magnitude in parameter count. We evaluate FGSM (Goodfellow et al., 2015), BIM (Kurakin et al., 2017), and PGD (Madry et al., 2018) on UCI Electricity (Trindade, 2015), ETTh1 (Zhou et al., 2021), and the LSTNet Traffic benchmark (Lai et al., 2018), localizing with PGD and testing whether the identified layer transfers across attack families. Figure 1 summarizes the pipeline.

1.1 Related work

Time-series foundation models. Following deep forecasters such as Informer (Zhou et al., 2021), PatchTST (Nie et al., 2023), and simple linear baselines (Zeng et al., 2023), a newer class is pretrained once on large corpora and forecasts unseen domains zero-shot: TimesFM (Das et al., 2024), Toto (Cohen et al., 2024) Chronos (Ansari et al., 2024), Moirai (Woo et al., 2024), and MOMENT (Goswami et al., 2024), with recent mixture-of-expert and long-context successors (Shi et al., 2025; Liu et al., 2025; Gao et al., 2024) surveyed Liang et al. (2024). This line optimizes accuracy and transfer; how a particular forecast is produced internally, and how it behaves under attack, is seldom studied. We take that internal computation as the object of repair.

Adversarial attacks and defenses for time series. Gradient attacks from image classification, FGSM (Goodfellow et al., 2015), BIM (Kurakin et al., 2017), PGD (Madry et al., 2018), and C&W (Carlini & Wagner, 2017), transfer to sequential inputs, and time-series robustness has centered on classification (Fawaz et al., 2019; Karim et al., 2021); attacks on forecasting (Dang-Nhu et al., 2020) and backdoor threats (Gu et al., 2017; Jiang et al., 2023) are comparatively rare. The standard defenses are global, whether adversarial training (Madry et al., 2018; Zhang et al., 2019) or input-level smoothing (Cohen et al., 2019). We instead locate and undo one corruption in place, leaving the rest of the model as pretrained.

Model editing and interpretability. Reading feed-forward layers as key-value memories (Geva et al., 2021) motivated methods that locate and rewrite specific behaviors: ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023), hypernetwork editors such as MEND (Mitchell et al., 2022), and recent work on locality and lifelong stability including WISE (Wang et al., 2024) and null-space constrained AlphaEdit (Fang et al., 2025), surveyed by Zhang et al. (2024). The localization step

draws on causal mediation (Vig et al., 2020), activation patching (Zhang & Nanda, 2024), transformer circuits (Elhage et al., 2021), and sparse-autoencoder features (Cunningham et al., 2024; Templeton et al., 2024). This literature is almost entirely about language models and aims to edit knowledge. We repurpose localize-then-edit for a numeric forecaster and use it as an adversarial defense, redefining the subject, target, and objective for continuous context windows so the closed-form update applies unchanged.

2 METHODOLOGY

2.1 MODEL AND DATASETS

We evaluate TimesFM 2.5 (Das et al., 2024), a decoder-only transformer with 20 blocks and $d_{\text{model}} = d_{\text{ff}} = 1280$, whose rank-one target is the feed-forward projection `ff1` of shape $(1280, 1280)$, and Toto 2.0 (Cohen et al., 2024) at both released scales (313M: 24 blocks, $d_{\text{model}} = 1024$, $d_{\text{ff}} = 2736$; 2.5B: 48 blocks, 2048, 5464), whose fused SwiGLU block is edited at the down projection `ffn.fc2` of shape $(d_{\text{model}}, d_{\text{ff}})$. Toto is multivariate-capable but is driven one series at a time, so both models receive identical univariate inputs. Both tokenize into 32-point patches, giving 32 patches for a 1024-step context. We evaluate UCI ElectricityLoadDiagrams (Trindade, 2015), ETTh1 (Zhou et al., 2021), and Traffic (Lai et al., 2018) with `MT_253`, `OT`, and a high-variance Traffic sensor as focal series, a 1024-step context, and a 128-step horizon; the 2.5B checkpoint is run on Electricity only, as a scale check. Preprocessing and channel-selection details are provided in Appendix A.3.

2.2 RESEARCH QUESTIONS

Our experiments are organized around three research questions.

RQ1: Do adversarial perturbations change where forecast-relevant information is causally localized inside the model?

RQ2: Can a localized model edit recover the clean forecast after an attack, and how do ROME, SeqROME, and MEMIT differ in restoration and specificity?

RQ3: Do these localization and repair effects remain consistent across attack families and across UCI Electricity, ETTh1, and Traffic?

These questions separate the mechanistic claim from the repair claim and the question of generalization.

2.3 SUBJECT SELECTION AND EXPERIMENTAL PROTOCOL

Causal tracing requires a subject region whose perturbation changes the forecast. Unlike language ROME, time-series inputs have no explicit subject tokens, so we select the subject empirically. We divide the 1024-step context into four equal segments, apply masked PGD independently to each, and select the segment producing the largest NRMSE against the clean forecast. We use $N = 4$ segments, as fewer segments produce broad model-wide disruption and more segments reduce the per-segment budget until PGD is near-inert. This segment is corrupted during subsequent localization and editing. The next sub-section reports the focal-channel diagnostics.

2.4 PER-SEGMENT ATTACK DIAGNOSTIC

Table 1 reports the forecast error produced by attacking each of the four context segments independently. The attack has its largest effect on Q2 according to both MSE and NRMSE against the held-out future. Figure 7 provides the corresponding forecast-level comparison across the four attacked segments. Together, these results show that the effect of a localized perturbation varies substantially with its temporal position and motivates selecting the subject segment empirically rather than fixing it in advance. Figure 7 in Appendix A.2 gives an extended look of the per-channel diagnostics.

Table 1: Per-segment masked PGD attack on the focal Electricity channel `MT_253` at the primary absolute perturbation budget $\varepsilon = 0.2$.

Metric	Q0	Q1	Q2	Q3
MSE vs. future	0.0123	0.3000	**1.7055**	0.4412
NRMSE vs. future	0.0926	0.4577	**1.0913**	0.5550

2.5 Adversarial corruption

We apply masked white-box FGSM, BIM, and PGD attacks to the selected subject segment using

$$\mathbf{x}_{t+1} = \Pi_{B_\varepsilon(\mathbf{x}_0)}[\mathbf{x}_t + \alpha\, \mathbf{m}_{S^\star} \odot \operatorname{sign}(\nabla_{\mathbf{x}} \mathcal{L}_{\mathrm{atk}}(\mathbf{x}_t))], \tag{1}$$

where $\mathcal{L}_{\mathrm{atk}}(\mathbf{x}) = \mathrm{MSE}(\hat{\mathbf{y}}(\mathbf{x}), \mathbf{y})$ is measured against the true $H = 128$-step future; using the clean forecast instead would yield a zero gradient at clean initialization. The mask $\mathbf{m}_{S^\star}$ restricts perturbations to the subject segment, and $\Pi_{B_\varepsilon(\mathbf{x}_0)}$ enforces the ℓ_∞ budget. Complete attack configurations are provided in Appendix A.1.

2.6 Causal localization and model editing

Causal tracing caches clean residual, attention, and MLP activations and patches them into the corrupted pass at each layer ℓ, site $q \in \{\mathrm{resid}, \mathrm{attn}, \mathrm{mlp}\}$, and subject position. For tracing variant v, let $s_{\mathrm{clean},v}$, $s_{\mathrm{adv},v}$, and $s^{(\ell,q)}_{\mathrm{patched},v}$ denote scalar summaries of the clean, corrupted, and patched forecasts. The causal restoration fraction is:

$$R_v^{(\ell,q)} = \frac{s^{(\ell,q)}_{\mathrm{patched},v} - s_{\mathrm{adv},v}}{s_{\mathrm{clean},v} - s_{\mathrm{adv},v}}. \tag{2}$$

Here, $R = 1$ denotes full recovery of the clean scalar summary, $R = 0$ no effect, and $R < 0$ degradation. Following Meng et al. (2022), we retain variants satisfying $|s_{\mathrm{clean},v} - s_{\mathrm{adv},v}| \geq \tau$, average their restoration values into $\bar{R}[\ell, q]$, and select $\ell^\star = \arg\max_\ell \bar{R}[\ell, \mathrm{mlp}]$.

At $\ell^\star$, ROME constructs a normalized key from the mean feed-forward input $\bar{\mathbf{k}}$ over the corrupted subject positions:

$$\mathbf{u} = \frac{\bar{\mathbf{k}}}{\|\bar{\mathbf{k}}\|_2}. \tag{3}$$

It then optimizes a free vector $\boldsymbol{\delta}$ added to the feed-forward output:

$$\mathcal{L}_{\mathrm{ROME}}(\boldsymbol{\delta}) = \mathrm{MSE}(\hat{\mathbf{y}}(\boldsymbol{\delta}), \hat{\mathbf{y}}_c) + \beta \frac{\|\boldsymbol{\delta}\|_2^2}{\|\mathbf{z}_0\|_2^2}. \tag{4}$$

We minimize Equation equation 4 for 30 Adam steps with learning rate 0.5 and a norm clamp, then compute $\mathbf{v} = \boldsymbol{\delta}/(\bar{\mathbf{k}}^\top \mathbf{u})$. The resulting rank-one update is:

$$W_{\ell^\star} \leftarrow W_{\ell^\star} + \mathbf{v}\mathbf{u}^\top. \tag{5}$$

2.7 Repeated editing and evaluation

Sequential ROME applies eight rank-one edits at the same layer ℓ^*, recomputing $\mathbf{u}$ and $\mathbf{v}$ after each edit. *Iterative MEMIT* (Meng et al., 2023) distributes the correction over the top-m MLP-restoration layers $\mathcal{B}$ using

$$\Delta W_{r,\ell} = R_{r,\ell} K_{r,\ell}^\top \left(\lambda C_{0,\ell} + K_{r,\ell} K_{r,\ell}^\top\right)^{-1}, \quad \text{for } \ell \in \mathcal{B}. \tag{6}$$

Here, $C_{0,\ell}$ is the preserve covariance, λ its regularization strength, $K_{r,\ell}$ the editing keys, and $R_{r,\ell}$ the residual changes assigned to layer ℓ. We evaluate MEMIT after one and five iterations. Causal restoration in Equation equation 2 is used only for localization. We evaluate forecast repair using the NRMSE-based edit-restoration fraction

$$R_{\mathrm{edit}} = 1 - \frac{\mathrm{NRMSE}\,(\hat{\mathbf{y}}_{\mathrm{edit}}, \hat{\mathbf{y}}_c)}{\mathrm{NRMSE}\,(\hat{\mathbf{y}}_{\mathrm{adv}}, \hat{\mathbf{y}}_c)}. \tag{7}$$

Thus, $R_{\mathrm{edit}} = 1$ denotes complete recovery, $R_{\mathrm{edit}} = 0$ no improvement, and $R_{\mathrm{edit}} < 0$ degradation. We report R_{edit}, NRMSE against clean forecast, and specificity between clean-input predictions from edited and unedited models.

2.8 ALGORITHMIC SUMMARY

Algorithm 1 summarizes the complete TS-Edit pipeline, from empirical subject selection and adversarial corruption to causal localization and model editing.

Algorithm 1 Causal Localization and Adversarial Forecast Repair

Input: Forecaster f_θ, clean context $\mathbf{x}_c$, held-out future $\mathbf{y}$, segments $\mathcal{G}_K$, tracing variants $\mathcal{V}$, layers $\mathcal{L}$, sites $\mathcal{Q} = \{\text{resid}, \text{attn}, \text{mlp}\}$, attack a, editor ρ, iterations N_ρ, MEMIT width m with preserve contexts $\mathcal{X}_{\text{pre}}$, threshold τ
Output: Subject $S^\star$, restoration profile $\bar{R}$, localized layer $\ell^\star$, edited parameters θ_ρ

$\hat{\mathbf{y}}_c \leftarrow f_\theta(\mathbf{x}_c)$; attack each $S_j \in \mathcal{G}_K$ with masked PGD against $\mathbf{y}$ and score $d_j \leftarrow \text{MSE}(f_\theta(\mathbf{x}_j^a), \mathbf{y})$
$S^\star \leftarrow S_{\arg\max_j d_j}$
Generate $\mathbf{x}^a$ on $S^\star$ using $a \in \{\text{PGD}, \text{FGSM}, \text{BIM}\}$
for all $v \in \mathcal{V}$ **do**
    Form $\mathbf{x}_c^{(v)}$ ($\mathbf{x}_c$ for $v = 0$, a noised copy otherwise), select S_v by absolute MSE-gradient mass, attack S_v with masked PGD, and record the scalar summaries $s_{c,v}, s_{a,v}$
    Keep v only if $|s_{c,v} - s_{a,v}| \geq \tau$; then cache the clean activations and patch each $(\ell, q) \in \mathcal{L} \times \mathcal{Q}$ over S_v to obtain $R_v^{(\ell,q)}$
end for
$\bar{R}[\ell, q] \leftarrow \text{mean}_{v:\,\text{effective}}\, R_v^{(\ell,q)}$; $\ell^\star \leftarrow \arg\max_\ell \bar{R}[\ell, \text{mlp}]$
if $\rho \in \{\text{ROME}, \text{SeqROME}\}$ **then**
    for $r = 1, \ldots, N_\rho$ **do** {$N_\rho = 1$ for ROME}
        Recompute $(\mathbf{u}_r, \mathbf{v}_r)$ under the current weights and apply $W_{\ell^\star} \leftarrow W_{\ell^\star} + \mathbf{v}_r \mathbf{u}_r^\top$
    end for
else
    Select the top-m layers $\mathcal{B}$ by $\bar{R}[\ell, \text{mlp}]$ and estimate the fixed preserve covariances $\{C_\ell\}_{\ell \in \mathcal{B}}$
    for $r = 1, \ldots, N_\rho$ **do**
        Optimize the deepest-layer target $\mathbf{z}_r^\star$ and spread residual MEMIT updates across $\mathcal{B}$
    end for
end if
return $S^\star, \bar{R}, \ell^\star, \theta_\rho$

3 RESULTS

We present the results by research question rather than by dataset. UCI Electricity `MT_253` remains the running example. ETTh1 and Traffic are then used as independent checks of the same procedure. For each research question, the main text shows the representative mechanism and the compact cross-dataset evidence needed to support the claim. The complete per-dataset breakdown is retained in the appendix (Tables 8 and 9).

We organize our evaluation around three research questions (RQs) and compare the corrupted forecast, single-shot ROME at the traced layer, iterated MEMIT over the top-three MLP-restoration layers, and Sequential ROME (SeqROME). We use Electricity channel `MT_253` under PGD as the running example, then repeat the comparison across FGSM, BIM, and PGD on ETTh1 and Traffic. Each attack uses 10 seeds at $\varepsilon = 0.2$. Results are reported as mean $\pm$ standard deviation.

3.1 RQ 1: ARE CAUSAL LOCALIZATION PATTERNS ROBUST TO ADVERSARIES?

Attack and segment selection. The attack drives the horizon away from the clean forecast, with pre-repair $\text{NRMSE}_{\text{adv}} = 0.889 \pm 0.078$ and $\text{MAE}_{\text{adv}} = 0.909 \pm 0.075$ over 10 seeds (Table 2), while the clean context outside Q2 is unchanged. Figure 7 (Appendix A.2) confirms that of the four segments, only the gradient-selected one produces a large forecast shift when actually corrupted, supporting the segment-scoring rule from Section 2.

Causal tracing. Figure 2 plots the restoration curve across all 20 blocks for the residual, attention, and MLP sites at the subject segment Q2, averaged over six clean-context noise variants. Restoration is maximized at the MLP site of block 4 (consensus across variants); this becomes the ROME target layer, and $\{2, 4, 17\}$ becomes the MEMIT band (top-3 MLP-restoration).

Table 2: Editing results under PGD at $\varepsilon = 0.2$ for TimesFM 2.5 and Toto 2.0 (313M), as mean $\pm$ standard deviation over 10 attack seeds. Attacked is the unedited model on the perturbed context. Forecast metrics are computed against each model's own clean-context forecast, not against ground truth. R_{edit} is the fraction of attack-induced error removed; Spec is the drift on unattacked context.

		Restoration	Forecast discrepancy			Locality
Dataset	Editor	R_{edit} ↑	NRMSE↓	MAE↓	WQL↓	Spec↓
TimesFM 2.5						
Electricity	Attacked	–	0.8890 ± 0.0780	0.9096 ± 0.0753	0.3764 ± 0.0301	–
	ROME	0.9738 ± 0.0030	0.0231 ± 0.0013	0.0205 ± 0.0011	0.0132 ± 0.0004	$(2.57 \pm 0.88) \times 10^{-4}$
	SeqROME	0.9809 ± 0.0033	0.0168 ± 0.0023	0.0145 ± 0.0022	0.0130 ± 0.0005	$(1.74 \pm 0.53) \times 10^{-4}$
	MEMIT	0.9834 ± 0.0037	0.0145 ± 0.0020	0.0121 ± 0.0018	0.0121 ± 0.0007	$(5.25 \pm 1.93) \times 10^{-5}$
ETTh1	Attacked	–	0.0824 ± 0.0004	0.3901 ± 0.0023	0.3615 ± 0.0006	–
	ROME	0.6356 ± 0.0057	0.0300 ± 0.0005	0.1393 ± 0.0030	0.3336 ± 0.0005	$(3.26 \pm 0.09) \times 10^{-2}$
	SeqROME	0.7069 ± 0.0133	0.0242 ± 0.0011	0.0975 ± 0.0055	0.2930 ± 0.0034	$(2.59 \pm 0.11) \times 10^{-2}$
	MEMIT	0.2201 ± 0.0065	0.0643 ± 0.0005	0.3041 ± 0.0022	0.3406 ± 0.0011	$(1.15 \pm 0.08) \times 10^{-3}$
Traffic	Attacked	–	0.8352 ± 0.5238	0.2938 ± 0.2010	0.1124 ± 0.0700	–
	ROME	0.6392 ± 0.1877	0.2254 ± 0.0634	0.0791 ± 0.0270	0.0285 ± 0.0084	$(8.66 \pm 9.25) \times 10^{-5}$
	SeqROME	0.8525 ± 0.0358	0.1110 ± 0.0568	0.0385 ± 0.0223	0.0186 ± 0.0055	$(1.19 \pm 0.67) \times 10^{-4}$
	MEMIT	0.8702 ± 0.0154	0.1085 ± 0.0688	0.0388 ± 0.0272	0.0184 ± 0.0065	$(4.02 \pm 2.11) \times 10^{-5}$
Toto 2.0 (313M)						
Electricity	Attacked	–	0.8355 ± 0.0563	0.8516 ± 0.0567	0.3429 ± 0.0224	–
	ROME	0.9350 ± 0.0024	0.0542 ± 0.0024	0.0494 ± 0.0024	0.0255 ± 0.0027	$(2.68 \pm 0.89) \times 10^{-4}$
	SeqROME	0.9938 ± 0.0018	0.0053 ± 0.0018	0.0041 ± 0.0011	0.0186 ± 0.0009	$(1.31 \pm 0.26) \times 10^{-4}$
	MEMIT	0.9922 ± 0.0037	0.0064 ± 0.0026	0.0058 ± 0.0027	0.0168 ± 0.0009	$(1.12 \pm 0.86) \times 10^{-4}$
ETTh1	Attacked	–	0.1810 ± 0.0019	0.9570 ± 0.0083	0.4581 ± 0.0032	–
	ROME	0.7176 ± 0.0791	0.0510 ± 0.0137	0.2423 ± 0.0726	0.3650 ± 0.0153	$(3.05 \pm 0.43) \times 10^{-2}$
	SeqROME	0.9246 ± 0.0068	0.0137 ± 0.0013	0.0595 ± 0.0053	0.2242 ± 0.0183	$(7.53 \pm 2.40) \times 10^{-3}$
	MEMIT	0.6685 ± 0.0055	0.0600 ± 0.0013	0.3222 ± 0.0062	0.3470 ± 0.0024	$(3.60 \pm 0.20) \times 10^{-2}$
Traffic	Attacked	–	18.5660 ± 0.8801	7.2775 ± 0.3350	3.9036 ± 0.3287	–
	ROME	0.9902 ± 0.0007	0.1821 ± 0.0048	0.0779 ± 0.0034	0.0297 ± 0.0010	$(2.97 \pm 1.01) \times 10^{-5}$
	SeqROME	0.99968 ± 0.00003	0.0059 ± 0.0007	0.0021 ± 0.0003	0.0123 ± 0.0005	$(3.63 \pm 0.87) \times 10^{-5}$
	MEMIT	0.9968 ± 0.0011	0.0596 ± 0.0207	0.0140 ± 0.0054	0.0117 ± 0.0007	$(3.43 \pm 2.47) \times 10^{-6}$

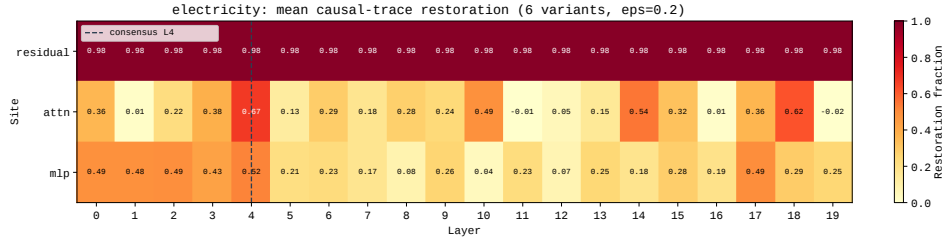


Figure 2: Layer-wise causal-trace restoration at $\varepsilon = 0.2$.

3.2 RQ 2: Can localized model edits recover the clean forecast?

ROME repair under PGD ($\varepsilon = 0.2$). A single ROME edit at L4 recovers most of the clean trajectory. It reduces $\text{NRMSE}_{\text{vs clean}}$ from 0.889 ± 0.0780 before repair to 0.0231 ± 0.0013 after repair. Similar improvements are observed in MAE and WQL, as summarized in Table 2. The edited model's prediction on the *clean* input remains visually close to that of the unedited model, indicating that the edit introduces little drift on the clean context (Figure 3).

Table 2 presents the PGD evaluation across Electricity, ETTh1, and Traffic for both TimesFM 2.5 and Toto 2.0, including restoration and specificity. The corresponding FGSM and BIM results are deferred to Appendix Tables 8 and 9. Additionally we ran the 2.5B-parameter Toto 2.0 checkpoint on Electricity as a scale check and its results are reported in Appendix D.1.

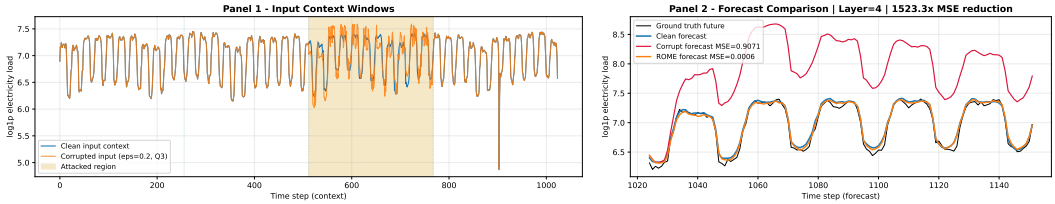


Figure 3: ROME repair at $\varepsilon = 0.2$. Left: clean and adversarial input contexts, with the attacked segment highlighted. Right: the adversarial forecast and the ROME-corrected forecast at the L4.

Iterative editors: SeqROME vs. MEMIT. Under PGD, both iterative editors improve upon single-shot ROME. SeqROME reduces NRMSE from 0.889 ± 0.078 to 0.0168 ± 0.0023, while MEMIT achieves the lowest NRMSE of 0.0145 ± 0.0020. MEMIT also obtains the lowest MAE (0.0121 ± 0.0018) and WQL (0.0121 ± 0.0007), with a mean R_{edit} of 0.9834, making it the strongest of the three editors under PGD. Additional convergence curves and forecast-level comparisons are provided in Appendix C.1.

Cross-channel repair on Electricity. Table 3 repeats the localize-and-edit pipeline independently on the seven quality-filtered UCI Electricity clients. The traced layer $L^\star$ varies widely across channels, yet a single ROME edit lowers NRMSE, MAE, and WQL for every channel and attack, with the largest gains under PGD and BIM and smaller ones under the near-inert FGSM. The repair therefore does not depend on the corruption converging to one shared editing layer. The corresponding ETTh1 and Traffic channel results are in Appendix D.3.

Table 3: Per-channel single-shot ROME for TimesFM 2.5 repair at $\varepsilon = 0.2$ on the seven quality-filtered UCI Electricity clients. $L^\star$ is the causally-traced target layer (varies by channel); NRMSE and MAE are against the clean forecast, WQL against the held-out ground truth.

Channel	$L^\star$	Attack	$\text{NRMSE}_{\text{adv}}$	$\text{NRMSE}_{\text{rome}}$	MAE_{adv}	MAE_{rome}	WQL_{adv}	WQL_{rome}
MT_002	17	PGD	3.12	0.372	1.88	0.190	0.777	0.064
		FGSM	0.17	0.094	0.09	0.045	0.047	0.039
		BIM	0.68	0.202	0.37	0.120	0.162	0.056
MT_061	17	PGD	1.21	0.129	1.04	0.115	0.431	0.065
		FGSM	0.10	0.065	0.09	0.057	0.075	0.063
		BIM	0.15	0.088	0.14	0.080	0.089	0.068
MT_129	0	PGD	3.15	0.356	1.04	0.120	0.468	0.078
		FGSM	0.39	0.264	0.13	0.086	0.074	0.061
		BIM	1.05	0.445	0.37	0.153	0.188	0.091
MT_197	4	PGD	2.10	0.077	1.53	0.053	0.628	0.039
		FGSM	0.18	0.141	0.14	0.108	0.074	0.063
		BIM	0.33	0.161	0.27	0.130	0.127	0.068
MT_253	4	PGD	0.87	0.022	0.90	0.020	0.393	0.019
		FGSM	0.02	0.016	0.02	0.014	0.024	0.023
		BIM	0.35	0.029	0.36	0.028	0.173	0.027
MT_306	15	PGD	1.08	0.041	0.64	0.030	0.272	0.033
		FGSM	0.06	0.039	0.04	0.027	0.039	0.035
		BIM	0.27	0.037	0.20	0.028	0.104	0.034
MT_369	18	PGD	4.23	0.289	1.81	0.125	0.859	0.090
		FGSM	0.28	0.221	0.11	0.087	0.075	0.067
		BIM	1.78	0.274	0.75	0.118	0.358	0.089

3.3 Restoration scales with attack magnitude

The fraction of error an edit removes scales with how far the attack moved the forecast, which explains the dataset ordering in Table 2. At $\varepsilon = 0.2$ PGD displaces ETTh1 $10.8\times$ less than Electricity on TimesFM 2.5 and $4.6\times$ less on Toto 2.0, making it the most attack-resistant benchmark at a fixed budget. Because R_{edit} measures the share of displacement removed (Equation equation 7), less displacement leaves less to recover, though the repair itself remains effective: ETTh1's post-edit NRMSE of 0.0300 and 0.0510 is comparable to the Electricity residuals. Figure 4 confirms the relationship across 43 windows on both models (Spearman $\rho = 0.63$ to 0.87, $p < 10^{-5}$), within datasets as well as across them.

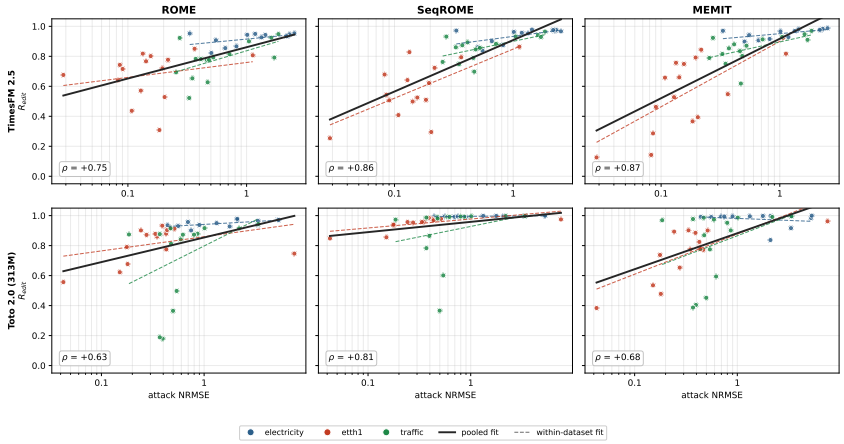


Figure 4: R_{edit} against the realised attack NRMSE under PGD at $\varepsilon = 0.2$, for TimesFM 2.5 (top) and Toto 2.0 (bottom). Each point is one of the 43 windows in the population sweep; solid lines are the pooled fit in $\log_{10}$ attack strength, dashed lines the within-dataset fits. ETTh1 occupies the low-NRMSE region on both models, reflecting its resistance to a perturbation of fixed magnitude.

3.4 RQ 3: Do these effects remain consistent across attacks and datasets?

We test if the Electricity findings remain consistent across attacks, strengths, and datasets.

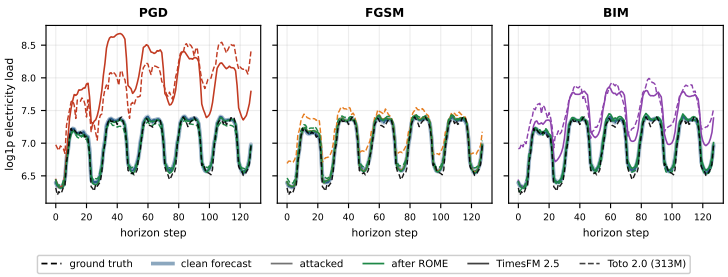
Cross-attack comparison. Using the PGD-localized layer without re-tracing, we apply ROME under PGD, BIM, and FGSM at $\varepsilon = 0.2$. On Electricity, ROME reduces mean NRMSE from 0.889 to 0.023 under PGD, from 0.337 to 0.028 under BIM, and from 0.019 to 0.016 under FGSM, with mean specificity at most 2.57×10^{-4}. PGD produces the largest shift, whereas FGSM is nearly inert at this budget. The PGD column of this comparison appears in Table 2; complete results for all datasets, attacks, and editors are reported in Appendix Tables 8 and 9.

Figure 5 confirms this at the forecast level for both models. The corrected forecast moves back toward the clean forecast under all three attacks, and the pinball loss improves across quantiles $\tau = 0.1, \ldots, 0.9$, so the repair is not confined to one attack family, one region of the predictive distribution, or one model.

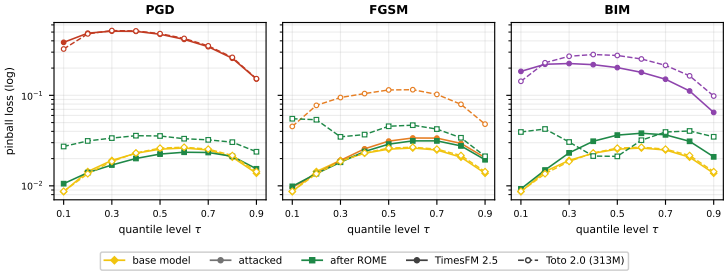
Budget sweep. Appendix Tables 12-14 evaluate single-seed ROME repair for $\varepsilon \in \{0.1, 0.2, 0.3, 0.4\}$. PGD and BIM generally become more damaging as the budget increases, whereas FGSM remains comparatively weak. ROME maintains low post-repair error across evaluated budgets.

Cross-dataset comparison. Across Electricity, ETTh1, and Traffic, ROME consistently reduces attack-induced forecast error, while SeqROME and MEMIT often reduce the error further. Neither of the iterative editor dominates every dataset and attack combination: their NRMSE and specificity advantages vary with the dataset and attack. Thus, localized editing transfers across evaluated settings, but the preferred editor remains setting-dependent (Table 2; Appendix Tables 8 and 9).

Robustness. Table 4 tests whether a single frozen edit generalizes to adversarial inputs it was not fitted to. The edit is computed once at the traced layer and evaluated without refitting against five fresh PGD random starts and against the BIM and FGSM families on the same subject segment. Across held-out PGD starts the repair holds, with recovery of 0.93 to 0.99 on Electricity and 0.70 to 0.99 on ETTh1 for both models; TimesFM 2.5 on Traffic is the exception at 0.38 to 0.48. Transfer across attack families is more dataset-dependent: BIM recovery exceeds 0.81 on Electricity for both models but falls to 0.35 for TimesFM 2.5 on Traffic, and FGSM is weakest throughout, with the Electricity attack near-inert (Pristine ≤ 0.044) and its ratio uninformative, including one negative value. A layer localized once under the strongest attack therefore transfers to other attacks without additional tracing, though the strength of that transfer varies with dataset and attack severity.



(a) ROME-corrected forecast across three attacks



(b) Pinball loss across quantile levels τ

Figure 5: ROME repair on `MT_253` at $\varepsilon = 0.2$ for TimesFM 2.5 (solid) and Toto 2.0 313M (dashed). **(a)** Forecasts under PGD, FGSM, and BIM before and after a single edit, with the per-model MSE reduction inset. **(b)** Per-quantile pinball loss against the held-out future.

Table 4: Generalization of a frozen TS-Edit repair to held-out adversarial inputs at ε=0.2. *Pristine* and *Edited* are the MSE between the attacked and clean forecasts, for the unedited and frozen edited model respectively. Recovery is $1 - \text{Edited}/\text{Pristine}$. PGD entries are mean $\pm$ std over five random starts. † marks rows with Pristine MSE below 0.05, where the ratio is ill-conditioned.

		TimesFM 2.5 (200M)			Toto 2.0 (313M)		
Dataset	Attack / Editor	Pristine	Edited	Recovery	Pristine	Edited	Recovery
Electricity	PGD / ROME	0.866± 0.048	0.015± 0.003	0.982	1.726± 1.100	0.082± 0.001	0.928
	BIM / ROME	0.359	0.009	0.976	0.427	0.078	0.817
	FGSM / ROME†	0.000	0.000	0.044	0.044	0.066	−0.480
ETTh1	PGD / ROME	0.157± 0.004	0.018± 0.001	0.883	0.966± 0.014	0.289± 0.002	0.701
	BIM / ROME	0.146	0.012	0.918	0.866	0.303	0.651
	FGSM / ROME	0.113	0.013	0.884	0.647	0.368	0.430
Traffic	PGD / ROME	1.096± 0.301	0.530± 0.091	0.479	10.523± 12.216	0.033± 0.002	0.992
	BIM / ROME	0.481	0.314	0.346	8.106	0.028	0.997
	FGSM / ROME†	0.006	0.005	0.241	0.019	0.018	0.048

4 Conclusion

We studied adversarial robustness in TSFMs as a localize-and-repair problem rather than only as a forecasting error problem. Across UCI Electricity, ETTh1, and Traffic, and under FGSM, BIM, and PGD, the experiments address three questions. First, causal tracing can identify a small set of feed-forward layers whose clean activations restore an attacked forecast. Second, editing at the localized site can recover much of the clean forecast, with no single editor dominating: MEMIT attains the lowest residual error under PGD on Electricity and Traffic, while SeqROME is strongest on ETTh1 and most consistent across attack families. Third, the repair effect persists across datasets and attack families, although specificity is more sensitive to the dataset and editing method; the ETTh1 results make this distinction especially clear for MEMIT. Together, these findings suggest that causal localization and model editing provide a useful way to study how adversarial perturbations propagate through time-series foundation models, while also showing that successful restoration and local, non-disruptive editing should be evaluated separately.

AI USE STATEMENT

In this work, we have used generative AI tools to assist wtih software developmentation tasks that are limited to debugging code written by the authors, adapting and extending the authors' core experimental pipeline to additional datasets and model architectures, and understanding implementation constraints and limitations specific to foundation models.

Generative AI tools were also used to assist with the consolidation of experimental outputs into tables, and with editing and polishing parts of the manuscript for readibility. The numerical results that are used in the tables were produced by the experimental pipelines and were not generated by the AI tools.

We did not use generative AI tools to generate synthetic datasets, formulate mathematical claims or proofs, develop theoretical models, or to perform qualitative or thematic analysis. We also did not rely on generative AI to originate the core research question or methodology.

All AI assisted code and manuscript content were reviewed by the authors. AI assisted code modifications were inspected and tested within the experimental pipelines, and the generated tables were checked against the underlying experimental results. The authors take full responsibility for the final content of this work, including all texts, claims, code, results, and artifacts that were produced with the aid of generative AI.

ETHICS STATEMENT

All the experiments conducted in this work have made use of publicly available benchmark datasets. No human subjects were involved. No personally identifiable or sensitive data was used. No new data was collected as part of this research.

The adversarial attacks that have been evaluated in this work, including the standard gradient-based methods, are already established techniques from existing literature and have been used to assess the robustness of time-series foundation models under controlled experimental conditions. The use of adversarial attack techniques in this work is explicitly defensive. They can, however, have dual-use implications. The objective of this work is purely to characterize model vulnerabilities and to evaluate methods for improving robustness against adversarial perturbations. The proposed framework does not introduce, or seek to introduce, a new attack mechanism, neither does it seek to facilitate attacks against deployed systems.

The experiments have been conducted on research benchmarks and pretrained models in controlled environments. We believe that studying these failure modes and developing mechanisms to mitigate them can contribute to the safer and more reliable deployment of time-series forecasting systems.

REPRODUCIBILITY STATEMENT

We provide the methodological, experimental, and implementation details that are required to reproduce the results that have been reported in this work. The evaluated models, datasets, focal series, context and forecast lengths, and primary preprocessing choices are described in Section 2.1, with additional dataset preprocessing and channel selection details provided in Appendix A.3. Subject region selection and the experimental protocol are described in Section 2.3. The adversarial attack formulation is given in Section 2.5, while the complete model and dataset-specific attack configurations, including perturbation budgets, step counts, restart settings, and clipping constraints, are reported in Appendix A.1. Causal tracing, editing-layer selection, and the ROME update are described in Section 2.6, while Sequential ROME, MEMIT, and the repeated-editing and evaluation procedures are described in Section 2.7. Model specific adaptations required for applying the editing procedures to Toto 2.0, including differences in normalization, activation scaling, and weight updates, are documented in Appendix A.4.

All datasets used in this work are publicly available and are cited in Section 2.1. The anonymous supplementary material provides the experimental code, model-specific implementation libraries, dataset-specific notebooks, saved experimental results, and instructions for environment setup and execution. Experiment configurations and random seeds are explicitly specified in the accompanying code. Because some attack settings differ across model & dataset combinations, the configurations

reported in Appendix A.1 and the corresponding notebooks should be used rather than assuming a uniform configuration across all experiments. Together, the main paper, appendix, and supplementary material provide the information and implementation artifacts needed to reproduce the reported experiments.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, 2017.

Ben Cohen, Emaad Khwaja, Kan Wang, Charles Masson, Elise Ramé, Youssef Doubli, and Othmane Abou-Amal. Toto: Time series optimized transformer for observability. *arXiv preprint arXiv:2407.07874*, 2024.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Raphaël Dang-Nhu, Gagandeep Singh, Pavol Bielik, and Martin Vechev. Adversarial attacks on probabilistic autoregressive forecasting models. In *International Conference on Machine Learning (ICML)*, 2020.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2024.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Xiang Wang, Xiangnan He, and Tat-Seng Chua. Alphaedit: Null-space constrained knowledge editing for language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Adversarial attacks on deep neural networks for time series classification. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2019.

Shanghua Gao, Teddy Koker, Owen Queen, Thomas Hartvigsen, Theodoros Tsiligkaridis, and Marinka Zitnik. Units: A unified multi-task time series model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2021.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. In *International Conference on Machine Learning (ICML)*, 2024.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

Yujing Jiang, Xingjun Ma, Sarah M. Erfani, and James Bailey. Backdoor attacks on time series: A generative approach. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023.

Fazle Karim, Somshubra Majumdar, and Houshang Darabi. Adversarial attacks on time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3309–3320, 2021.

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 (ICLR)*, 2022.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR) Workshop Track*, 2017.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long- and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR)*, pp. 95–104, 2018.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-xl: Long-context transformers for unified time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2025.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations (ICLR)*, 2022.

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

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.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jermyn, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

Artur Trindade. ElectricityLoadDiagrams20112014 data set. UCI Machine Learning Repository, 2015.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI Conference on Artificial Intelligence*, 2023.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, et al. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*, 2024.

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*, 2021.

A Additional background and method details

A.1 Attack configuration

All attacks are restricted to the empirically selected subject segment using a binary mask; values outside this segment remain unchanged. FGSM performs one deterministic full-budget update with $T = 1$ and $\alpha = \varepsilon$. BIM starts from the clean input and performs $T = 10$ updates with $\alpha = \varepsilon/8$. PGD performs $T = 50$ updates with $\alpha = \varepsilon/20$ and uses five restarts initialized with masked uniform noise from $[-\varepsilon, \varepsilon]$. We retain the restart with the highest final attack loss.

The perturbation budget ε is a literal ℓ_∞ radius in the preprocessed input space and is not rescaled by the context standard deviation. For Electricity, the attack is applied after the $\log(1+x)$ transformation; TimesFM performs its normalization internally. Gradients are computed through the differentiable decoder, while reported predictions use the configured forecasting path of the respective model.

The primary experiments use $\varepsilon = 0.2$ and repeat the complete localization-and-editing pipeline over 10 seeds for each attack. PGD varies across seeds because of its random initialization and restart selection. FGSM and BIM are deterministic for a fixed window, so their across-seed variation arises from subsequent localization and editing stages rather than from attack generation.

This configuration is the default but is not uniform across every model-dataset pair, and Table 5 records the deviations. Traffic clips perturbations to the data domain $[0, 1]$ on both models, since occupancy rates are bounded. Toto 2.0 313M uses a reduced PGD schedule on Traffic with a matched BIM step count, and the Toto 2.0 2.5B scale check on Electricity uses $\varepsilon = 0.15$. The step size is derived from the step count as $\alpha = 2.5\,\varepsilon/T$ for PGD and $\alpha = 1.25\,\varepsilon/T$ for BIM, so the values quoted above correspond to the default schedule and a changed T changes α accordingly.

Table 5: Attack configuration by model and dataset. The step size follows $\alpha = 2.5\,\varepsilon/T$ for PGD and $\alpha = 1.25\,\varepsilon/T$ for BIM, so a changed step count also changes α.

Model	Dataset	ε	PGD $T\times$restarts	BIM T	Clip
TimesFM 2.5	Electricity, ETTh1	0.2	50×5	10	–
	Traffic	0.2	50×5	10	$[0, 1]$
Toto 2.0 (313M)	Electricity, ETTh1	0.2	50×5	10	–
	Traffic	0.2	20×2	20	$[0, 1]$
Toto 2.0 (2.5B)	Electricity	0.15	50×5	10	–

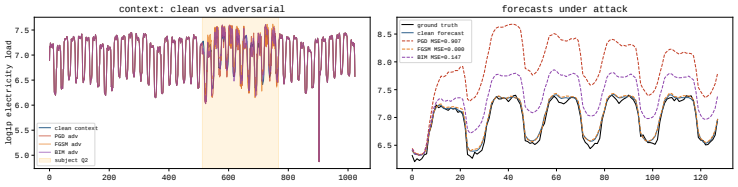


Figure 6: Left: Clean and adversarial contexts. Right: Resulting attacked forecast w.r.t clean forecast.

A.2 Extended per-segment attack diagnostic

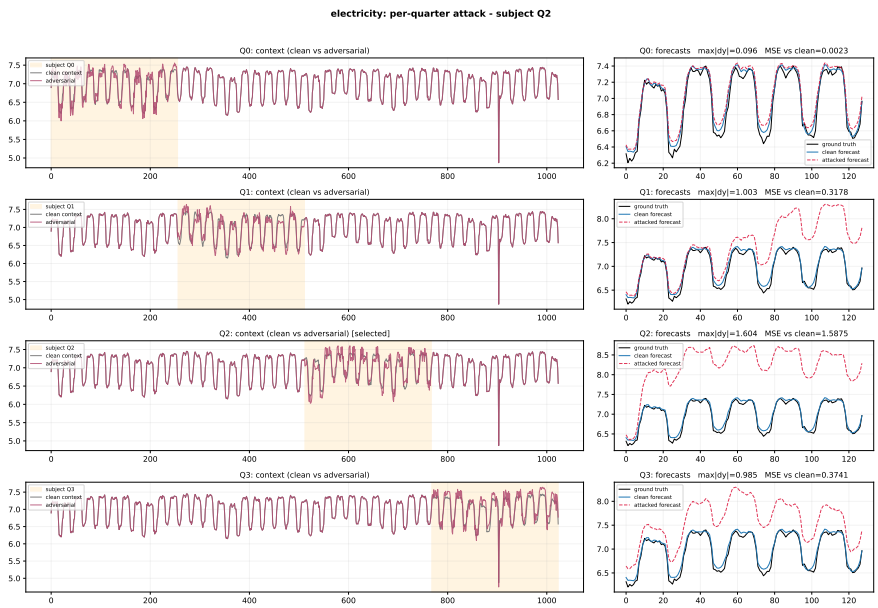


Figure 7: Per-segment attack diagnostic on the focal Electricity channel. Each panel runs the same attack on a different segment of the 1024-step context. The resulting forecast deviations show that attack sensitivity varies substantially with temporal position.

A.3 Model and dataset details

Both models process each univariate series independently in non-overlapping 32-point patches, so each 1024-step context produces 32 input patches followed by a 128-step forecast horizon.

UCI ElectricityLoadDiagrams (Trindade, 2015) contains 15-minute load measurements from 370 clients. We aggregate the measurements hourly, apply a $\log(1+x)$ transformation before the model's internal normalization (RevIN (Kim et al., 2022) for TimesFM 2.5, a patched causal standard scaler followed by an asinh transform for Toto 2.0), and filter channels using mean load, coefficient of variation, and median proximity. Seven channels are sampled evenly from the retained set, with `MT_253` used as the focal series.

For ETTh1 (Zhou et al., 2021), we use the oil-temperature variable `OT` as the focal series without additional filtering or transformation. For Traffic (Lai et al., 2018), we rank sensors by variance and select the highest-variance sensor whose initial 1024-step window has non-trivial variability and no missing values. Occupancy rates are clipped to $[0, 1]$.

A.4 Model editing configuration

Both models use the same editing hyperparameters: 30 Adam steps, weight decay $\beta = 0.1$, and a norm clamp at $3\times$ the baseline feed-forward output norm. Sequential ROME applies eight rank-one edits at $\ell^\star$, recomputing $\mathbf{u}$ and $\mathbf{v}$ after each, and MEMIT runs eight rounds over the top-three MLP-restoration layers with $\lambda = 15000$ and a ridge of 10^{-4} on the solve matrix. TimesFM 2.5 uses the learning rate $\texttt{lr} = 0.5$ of the original ROME formulation directly.

That absolute step size does not transfer to Toto 2.0, whose feed-forward activations are roughly three orders of magnitude larger per coordinate under its unit-scaling conventions. The same `lr` then moves $\boldsymbol{\delta}$ only a small fraction of the admissible ball, the clamp never binds, and the edit is effectively inert. We therefore generalize the learning rate to be expressed in units of the activation scale,

$$\texttt{lr}_{\text{eff}} = \texttt{lr} \cdot \frac{\|\mathbf{z}_0\|_2}{\sqrt{d}}, \tag{8}$$

where $\|\mathbf{z}_0\|_2$ is the measured feed-forward output norm at the subject position and d the value-space dimension. This is a generalization of the original rule rather than a new constant: evaluating it for TimesFM recovers 0.313, essentially its hardcoded 0.5. It also transfers unchanged between the two Toto checkpoints, since the norm is measured at run time. Because each edit writes into the same activation the norm is read from, we measure it once on pristine weights and hold it fixed across iterations.

Three smaller adaptations follow from Toto's architecture. Its down projection is unit-scaled, so every update is divided by the measured factor $1/\sqrt{d_{\text{ff}}}$; its post-SwiGLU keys are collinear, so we use the full ROME left vector $\mathbf{u} \propto C^{-1}\mathbf{k}$ rather than the normalized raw key; and its MEMIT preserve bank is sized to at least $d_{\text{ff}}/32$ clean windows to span the key space. We also clamp the causal scaler variance before its square root to keep gradients finite.

B Additional results for RQ1: causal tracing

The main text localizes the corruption on the Electricity running example (Figure 2). Figures 8 and 9 repeat the causal-trace diagnostic for TimesFM 2.5 on ETTh1 and Traffic, averaging restoration over the six clean-context tracing variants for the residual, attention, and MLP sites at each of its 20 transformer blocks. On both datasets, the MLP restoration profile concentrates in the earliest blocks and peaks at block 0, which is selected as the editing layer. Attention contributes little, while the residual stream is restored in the early layers because patching it copies the clean state wholesale. The peak layer differs from the Electricity example, confirming that the editing site is dataset-specific, but the qualitative localization pattern remains stable.

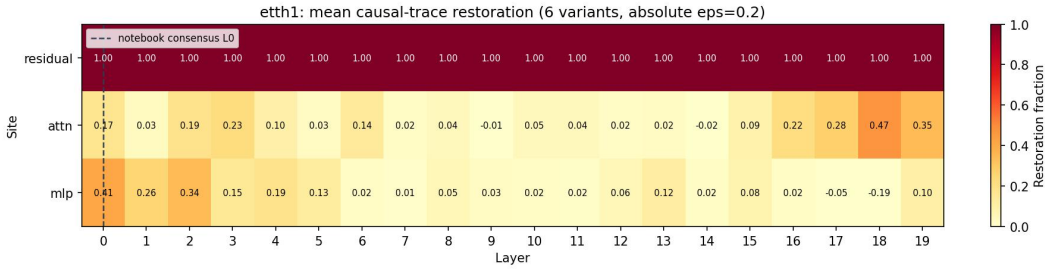


Figure 8: Layer-wise causal-trace restoration for TimesFM on ETTh1 (OT) at $\varepsilon = 0.2$, averaged over six tracing variants, for the residual, attention, and MLP sites across all 20 blocks. The MLP profile peaks at block 0, which is selected as the ROME editing layer.

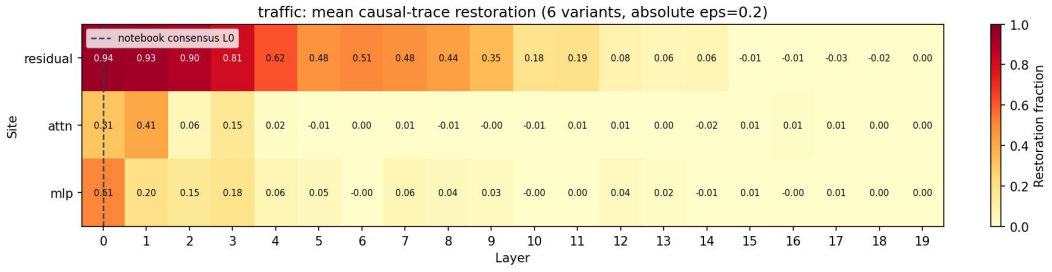


Figure 9: Layer-wise causal-trace restoration for TimesFM on Traffic (highest-variance sensor) at $\varepsilon = 0.2$, in the same format as Figure 8. The MLP restoration again concentrates in the earliest blocks and peaks at block 0.

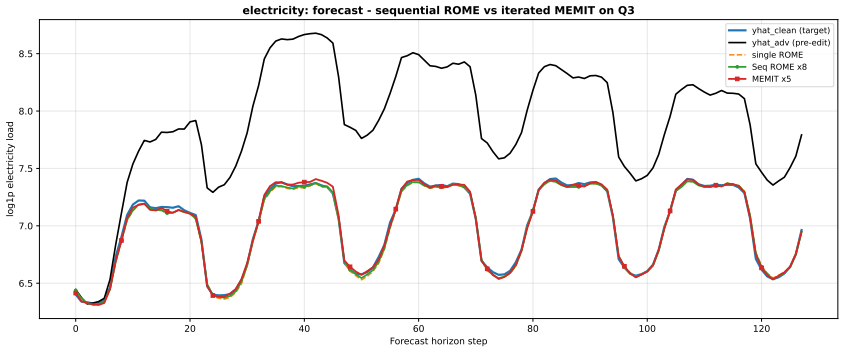


Figure 10: Forecast comparison under PGD at the primary absolute perturbation budget $\varepsilon = 0.2$. The clean target and pre-edit adversarial forecast are shown alongside the forecasts after single ROME, sequential ROME, and iterated MEMIT.

C Additional results for RQ2: model editing

C.1 Electricity editor convergence and attack-wise repair

Figure 11 tracks Sequential ROME and iterated MEMIT across editing iterations on the focal Electricity example, and Figure 10 overlays the resulting forecasts. Sequential ROME reaches near-complete restoration in a single iteration and then holds a small, fixed specificity cost, whereas iterated MEMIT restores more gradually but ultimately reaches a comparable or slightly lower error while remaining at least as specific (Tables 8 and 9). The balance between the two is dataset-specific, and Sequential ROME is the more reliable editor overall (Figures 12 and 13).

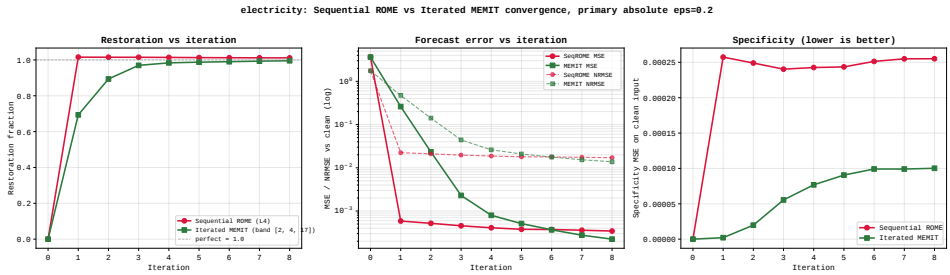


Figure 11: Sequential ROME vs. iterated MEMIT convergence on the focal Electricity example (`MT_253`) under PGD at the primary absolute perturbation budget $\varepsilon = 0.2$. From left to right: restoration fraction, forecast error against the clean prediction (log scale), and specificity error (clean-input MSE) across editing iterations. SeqROME reaches $R \approx 1$ in a single iteration and holds a fixed specificity cost, whereas iterated MEMIT restores more slowly but crosses below SeqROME in forecast error around iteration 6 and keeps a lower specificity error. Both methods approach complete restoration.

C.2 Editors across datasets

Figures 12 and 13 repeat the convergence diagnostic on ETTh1 and Traffic, each reporting restoration, forecast error against the clean prediction, and specificity across editing iterations. A single ROME edit already removes the bulk of the corruption in one step, gaining most under the stronger PGD and BIM attacks and little under the near-inert FGSM (Table 7). Sequential ROME repeats the update at the same layer, driving restoration to $R \approx 1$ on every dataset while keeping specificity low and flat, which gives it the lowest final error and the most consistent behavior. Iterated MEMIT is the most dataset-dependent: it matches Sequential ROME on Traffic, but on ETTh1 the band update fails to take hold and its specificity climbs, so its restoration and edit locality must be read together.

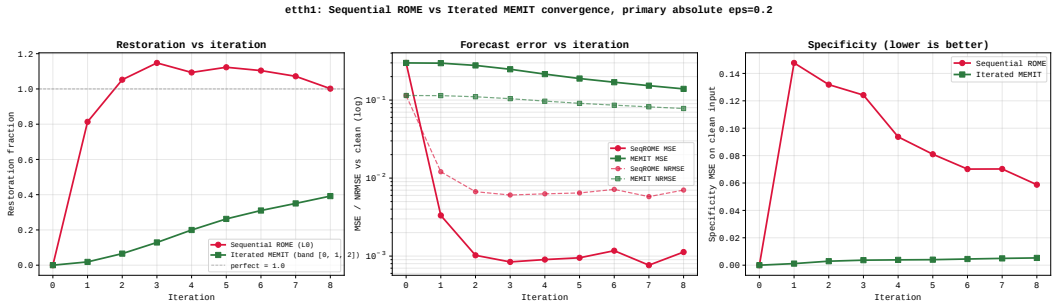


Figure 12: Sequential ROME vs. iterated MEMIT on ETTh1 (OT) under PGD at $\varepsilon = 0.2$. Left to right: restoration fraction, forecast error vs. the clean forecast (log scale), and specificity error, all across editing iterations. SeqROME converges to $R \approx 1$ with low, flat specificity, whereas iterated MEMIT plateaus at a much lower restoration and its specificity spikes by roughly an order of magnitude, mirroring the Electricity behavior in Figure 11.

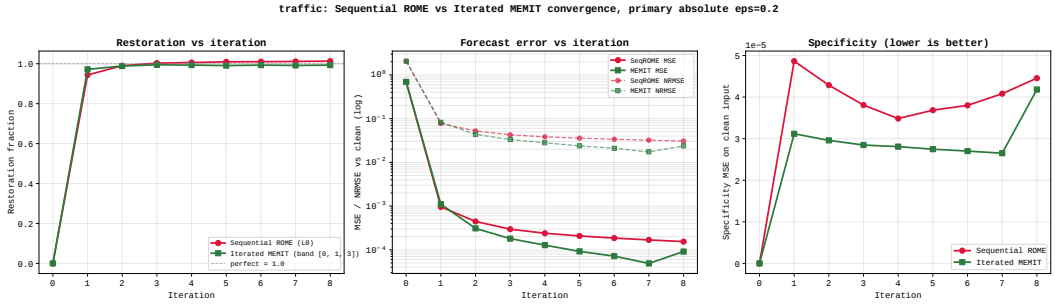


Figure 13: Sequential ROME vs. iterated MEMIT on Traffic (highest-variance sensor) under PGD at $\varepsilon = 0.2$, in the same format as Figure 12. Both editors reach $R \approx 1$ within a few iterations, but SeqROME attains lower final forecast error and holds specificity flat while iterated MEMIT's specificity climbs across iterations.

D Additional results for RQ3: Generalization

D.1 Model-scale results: Toto 2.0 (2.5B)

To test whether the procedure holds at larger scale, we repeat the Electricity pipeline on the 2.5B-parameter Toto 2.0 checkpoint, which has 48 transformer blocks, $d_{\text{model}} = 2048$, and $d_{\text{ff}} = 5464$ against the 24 blocks, 1024, and 2736 of the 313M model. Causal tracing places the peak MLP restoration at L5, with the top-three band $\{5, 6, 9\}$, so the editing site stays in the early layers as on the 313M model. This sweep uses $\varepsilon = 0.15$ rather than the $\varepsilon = 0.2$ used elsewhere, so Table 6 is a standalone scale check rather than a row comparable to Table 2.

All three editors transfer without modification. Under PGD, SeqROME attains the highest restoration at $R_{\text{edit}} = 0.9984$ and the lowest residual NRMSE at 0.0029, with MEMIT close behind at 0.9959; single-shot ROME already removes 96.2% of the attack-induced error. The ordering on the weaker attacks matches the 313M model, with the iterative editors adding most where ROME leaves the largest residual: on FGSM, R_{edit} rises from 0.8522 to 0.9771 and 0.9785. Specificity stays at or below 4×10^{-4} for every editor and attack, and MEMIT records no failures in any of the 30 runs.

D.2 Cross-dataset results

Tables 8 and 9 give the complete evaluation behind the focal main-paper result (Table 2), covering UCI Electricity, ETTh1, and Traffic under PGD, FGSM, and BIM, with one table per model. Alongside NRMSE, MAE, WQL, and R_{edit} they report specificity, the clean-input drift each editor introduces.

Table 6: Editing results for Toto 2.0 (2.5B) on Electricity (`MT_253`) at $\varepsilon = 0.15$, traced layer L5, MEMIT band $\{5, 6, 9\}$, 10 attack seeds. PGD is mean $\pm$ std; FGSM and BIM are deterministic. Metrics follow Table 2.

Editor	Metric	PGD	FGSM	BIM
Attacked	NRMSE	1.8432 ± 0.3476	0.1192	0.3598
	MAE	1.6366 ± 0.2987	0.1100	0.3812
	WQL	0.6172 ± 0.0973	0.0436	0.1477
ROME	NRMSE	0.0694 ± 0.0148	0.0176	0.0162
	MAE	0.0561 ± 0.0108	0.0161	0.0147
	WQL	0.0246 ± 0.0029	0.0181	0.0175
	$\mathbf{R}_{\text{edit}}$	0.9624 ± 0.0030	0.8522	0.9551
	Spec	$(7.0 \pm 4.0) \times 10^{-6}$	3.0×10^{-6}	3.32×10^{-4}
SeqROME	NRMSE	0.0029 ± 0.0003	0.0027	0.0027
	MAE	0.0025 ± 0.0004	0.0024	0.0024
	WQL	0.0159 ± 0.0012	0.0159	0.0138
	$\mathbf{R}_{\text{edit}}$	0.9984 ± 0.0005	0.9771	0.9926
	Spec	$(4.1 \pm 1.8) \times 10^{-5}$	5.0×10^{-6}	3.96×10^{-4}
MEMIT	NRMSE	0.0074 ± 0.0071	0.0026	0.0036
	MAE	0.0066 ± 0.0063	0.0022	0.0032
	WQL	0.0171 ± 0.0016	0.0165	0.0176
	$\mathbf{R}_{\text{edit}}$	0.9959 ± 0.0034	0.9785	0.9901
	Spec	$(7.0 \pm 5.0) \times 10^{-6}$	9.0×10^{-6}	1.20×10^{-4}

PGD values are mean $\pm$ standard deviation over 10 seeds; FGSM and BIM are deterministic across seeds and reported as single values. The main paper reports only the PGD column of these tables.

Table 7 recasts these errors as an NRMSE reduction factor ($\text{NRMSE}_{\text{adv}}/\text{NRMSE}_{\text{edit}}$), normalizing for differing attack severities. The factor is largest on Electricity and Traffic, which the attacks damage most, and smallest on ETTh1; the iterative editors add the most on the datasets where a single edit leaves the largest residual. No editor dominates uniformly, and the ranking depends on both dataset and attack family.

Table 7: Cross-dataset repair expressed as the NRMSE reduction factor $\text{NRMSE}_{\text{adv}}/\text{NRMSE}_{\text{edit}}$ at $\varepsilon = 0.2$; larger factors indicate greater reduction of the attack-induced forecast error. PGD factors are computed from the 10-seed mean NRMSE values in Tables 8 and 9; FGSM and BIM factors use their deterministic single-run values.

Dataset	Attack	$\text{NRMSE}_{\text{adv}}$	ROME	SeqROME	MEMIT
Electricity	PGD	0.889	$38.5\times$	$52.9\times$	$61.3\times$
	FGSM	0.019	$1.2\times$	$1.4\times$	$1.3\times$
	BIM	0.337	$12.2\times$	$22.9\times$	$19.4\times$
ETTh1	PGD	0.082	$2.7\times$	$3.4\times$	$1.3\times$
	FGSM	0.070	$3.5\times$	$2.9\times$	$1.2\times$
	BIM	0.078	$2.9\times$	$2.7\times$	$1.2\times$
Traffic	PGD	0.835	$3.7\times$	$7.5\times$	$7.7\times$
	FGSM	0.130	$1.1\times$	$1.6\times$	$1.8\times$
	BIM	0.292	$1.9\times$	$5.3\times$	$4.7\times$

Table 8: Full cross-dataset and cross-attack editing results for TimesFM 2.5 at the fixed absolute perturbation $\varepsilon = 0.2$. PGD entries are the mean over 10 attack seeds with the per-seed standard deviation; FGSM and BIM are deterministic across seeds and are reported as point values. Metrics follow Table 2: forecast metrics are computed against the clean-context forecast of the unedited model, and Spec is written in scientific notation throughout.

Dataset	Editor	Metric	PGD	FGSM	BIM
Electricity	Attacked	NRMSE	0.8890 ± 0.0780	0.0189	0.3367
		MAE	0.9096 ± 0.0753	0.0174	0.3456
		WQL	0.3764 ± 0.0301	0.0143	0.1466
	ROME	NRMSE	0.0231 ± 0.0013	0.0158	0.0275
		MAE	0.0205 ± 0.0011	0.0141	0.0259
		WQL	0.0132 ± 0.0004	0.0137	0.0144
		$\mathbf{R}_{\text{edit}}$	0.9738 ± 0.0030	0.1631	0.9183
		Spec	$(2.57 \pm 0.88) \times 10^{-4}$	2.51×10^{-6}	1.26×10^{-4}
	SeqROME	NRMSE	0.0168 ± 0.0023	0.0135	0.0147
		MAE	0.0145 ± 0.0022	0.0119	0.0134
		WQL	0.0130 ± 0.0005	0.0134	0.0127
		$\mathbf{R}_{\text{edit}}$	0.9809 ± 0.0033	0.2855	0.9563
		Spec	$(1.74 \pm 0.53) \times 10^{-4}$	2.02×10^{-5}	1.37×10^{-4}
	MEMIT	NRMSE	0.0145 ± 0.0020	0.0144	0.0174
		MAE	0.0121 ± 0.0018	0.0131	0.0147
		WQL	0.0121 ± 0.0007	0.0137	0.0133
		$\mathbf{R}_{\text{edit}}$	0.9834 ± 0.0037	0.2355	0.9485
		Spec	$(5.25 \pm 1.93) \times 10^{-5}$	9.78×10^{-7}	1.55×10^{-4}
ETTh1	Attacked	NRMSE	0.0824 ± 0.0004	0.0698	0.0781
		MAE	0.3901 ± 0.0023	0.3286	0.3689
		WQL	0.3615 ± 0.0006	0.3589	0.3591
	ROME	NRMSE	0.0300 ± 0.0005	0.0199	0.0270
		MAE	0.1393 ± 0.0030	0.0820	0.1225
		WQL	0.3336 ± 0.0005	0.3308	0.3310
		$\mathbf{R}_{\text{edit}}$	0.6356 ± 0.0057	0.7147	0.6550
		Spec	$(3.26 \pm 0.09) \times 10^{-2}$	4.15×10^{-2}	3.39×10^{-2}
	SeqROME	NRMSE	0.0242 ± 0.0011	0.0241	0.0290
		MAE	0.0975 ± 0.0055	0.0976	0.1173
		WQL	0.2930 ± 0.0034	0.2927	0.2878
		$\mathbf{R}_{\text{edit}}$	0.7069 ± 0.0133	0.6549	0.6292
		Spec	$(2.59 \pm 0.11) \times 10^{-2}$	2.65×10^{-2}	2.40×10^{-2}
	MEMIT	NRMSE	0.0643 ± 0.0005	0.0604	0.0625
		MAE	0.3041 ± 0.0022	0.2841	0.2953
		WQL	0.3406 ± 0.0011	0.3429	0.3387
		$\mathbf{R}_{\text{edit}}$	0.2201 ± 0.0065	0.1348	0.2002
		Spec	$(1.15 \pm 0.08) \times 10^{-3}$	2.07×10^{-3}	1.25×10^{-3}
Traffic	Attacked	NRMSE	0.8352 ± 0.5238	0.1302	0.2924
		MAE	0.2938 ± 0.2010	0.0348	0.0930
		WQL	0.1124 ± 0.0700	0.0183	0.0320
	ROME	NRMSE	0.2254 ± 0.0634	0.1135	0.1580
		MAE	0.0791 ± 0.0270	0.0303	0.0440
		WQL	0.0285 ± 0.0084	0.0176	0.0164
		$\mathbf{R}_{\text{edit}}$	0.6392 ± 0.1877	0.1283	0.4594
		Spec	$(8.66 \pm 9.25) \times 10^{-5}$	7.25×10^{-6}	7.82×10^{-5}
	SeqROME	NRMSE	0.1110 ± 0.0568	0.0804	0.0555
		MAE	0.0385 ± 0.0223	0.0219	0.0165
		WQL	0.0186 ± 0.0055	0.0158	0.0125
		$\mathbf{R}_{\text{edit}}$	0.8525 ± 0.0358	0.3828	0.8101
		Spec	$(1.19 \pm 0.67) \times 10^{-4}$	5.81×10^{-5}	1.51×10^{-4}
	MEMIT	NRMSE	0.1085 ± 0.0688	0.0723	0.0620
		MAE	0.0388 ± 0.0272	0.0198	0.0181
		WQL	0.0184 ± 0.0065	0.0150	0.0128
		$\mathbf{R}_{\text{edit}}$	0.8702 ± 0.0154	0.4448	0.7879
		Spec	$(4.02 \pm 2.11) \times 10^{-5}$	8.39×10^{-6}	3.89×10^{-5}

Table 9: Full cross-dataset and cross-attack editing results for Toto 2.0 (313M) at the fixed absolute perturbation $\varepsilon = 0.2$. PGD entries are the mean over 10 attack seeds with the per-seed standard deviation; FGSM and BIM are deterministic across seeds and are reported as point values. Metrics follow Table 2: forecast metrics are computed against the clean-context forecast of the unedited model, and Spec is written in scientific notation throughout.

Dataset	Editor	Metric	PGD	FGSM	BIM
Electricity	Attacked	NRMSE	0.8355 ± 0.0563	0.1961	0.4877
		MAE	0.8516 ± 0.0567	0.1854	0.5074
		WQL	0.3429 ± 0.0224	0.0704	0.1932
	ROME	NRMSE	0.0542 ± 0.0024	0.0469	0.0359
		MAE	0.0494 ± 0.0024	0.0460	0.0295
		WQL	0.0255 ± 0.0027	0.0307	0.0319
		$\mathbf{R}_{\text{edit}}$	0.9350 ± 0.0024	0.7608	0.9264
		Spec	$(2.68 \pm 0.89) \times 10^{-4}$	2.20×10^{-4}	4.83×10^{-4}
	SeqROME	NRMSE	0.0053 ± 0.0018	0.0048	0.0042
		MAE	0.0041 ± 0.0011	0.0040	0.0036
		WQL	0.0186 ± 0.0009	0.0197	0.0224
		$\mathbf{R}_{\text{edit}}$	0.9938 ± 0.0018	0.9757	0.9914
		Spec	$(1.31 \pm 0.26) \times 10^{-4}$	1.93×10^{-4}	4.30×10^{-4}
	MEMIT	NRMSE	0.0064 ± 0.0026	0.0047	0.0048
		MAE	0.0058 ± 0.0027	0.0042	0.0040
		WQL	0.0168 ± 0.0009	0.0236	0.0193
		$\mathbf{R}_{\text{edit}}$	0.9922 ± 0.0037	0.9759	0.9902
		Spec	$(1.12 \pm 0.86) \times 10^{-4}$	1.55×10^{-5}	1.24×10^{-4}
ETTh1	Attacked	NRMSE	0.1810 ± 0.0019	0.1468	0.1678
		MAE	0.9570 ± 0.0083	0.7830	0.8940
		WQL	0.4581 ± 0.0032	0.4250	0.4459
	ROME	NRMSE	0.0510 ± 0.0137	0.1062	0.0429
		MAE	0.2423 ± 0.0726	0.5427	0.1964
		WQL	0.3650 ± 0.0153	0.3964	0.3660
		$\mathbf{R}_{\text{edit}}$	0.7176 ± 0.0791	0.2762	0.7441
		Spec	$(3.05 \pm 0.43) \times 10^{-2}$	5.28×10^{-2}	2.55×10^{-2}
	SeqROME	NRMSE	0.0137 ± 0.0013	0.0136	0.0094
		MAE	0.0595 ± 0.0053	0.0602	0.0393
		WQL	0.2242 ± 0.0183	0.2073	0.2058
		$\mathbf{R}_{\text{edit}}$	0.9246 ± 0.0068	0.9076	0.9439
		Spec	$(7.53 \pm 2.40) \times 10^{-3}$	1.68×10^{-2}	7.94×10^{-3}
	MEMIT	NRMSE	0.0600 ± 0.0013	0.0464	0.0570
		MAE	0.3222 ± 0.0062	0.2494	0.3075
		WQL	0.3470 ± 0.0024	0.3378	0.3456
		$\mathbf{R}_{\text{edit}}$	0.6685 ± 0.0055	0.6842	0.6606
		Spec	$(3.60 \pm 0.20) \times 10^{-2}$	3.77×10^{-2}	3.94×10^{-2}
Traffic	Attacked	NRMSE	18.5660 ± 0.8801	0.2744	2.3432
		MAE	7.2775 ± 0.3350	0.1159	1.0274
		WQL	3.9036 ± 0.3287	0.0381	0.4038
	ROME	NRMSE	0.1821 ± 0.0048	0.0394	0.1476
		MAE	0.0779 ± 0.0034	0.0137	0.0489
		WQL	0.0297 ± 0.0010	0.0149	0.0202
		$\mathbf{R}_{\text{edit}}$	0.9902 ± 0.0007	0.8563	0.9370
		Spec	$(2.97 \pm 1.01) \times 10^{-5}$	1.84×10^{-6}	5.99×10^{-5}
	SeqROME	NRMSE	0.0059 ± 0.0007	0.0050	0.0154
		MAE	0.0021 ± 0.0003	0.0018	0.0040
		WQL	0.0123 ± 0.0005	0.0122	0.0136
		$\mathbf{R}_{\text{edit}}$	0.99968 ± 0.00003	0.9817	0.9934
		Spec	$(3.63 \pm 0.87) \times 10^{-5}$	2.74×10^{-6}	1.98×10^{-5}
	MEMIT	NRMSE	0.0596 ± 0.0207	0.0073	0.2056
		MAE	0.0140 ± 0.0054	0.0025	0.0381
		WQL	0.0117 ± 0.0007	0.0132	0.0213
		$\mathbf{R}_{\text{edit}}$	0.9968 ± 0.0011	0.9734	0.9122
		Spec	$(3.43 \pm 2.47) \times 10^{-6}$	4.95×10^{-7}	1.39×10^{-4}

The restoration behavior also holds beyond the single focal window. Figures 14 and 15 aggregate restoration and MSE recovery for TimesFM 2.5 over the population sweep (N=15 windows) on ETTh1 and Traffic: SeqROME restores most reliably across all three attacks, ROME tracks it closely, and MEMIT (failures excluded) shows the largest variance, consistent with the per-dataset editor behavior in Appendix C.

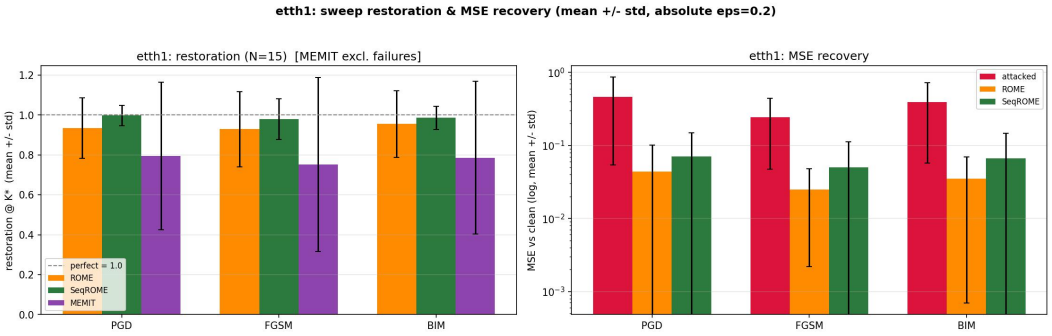


Figure 14: ETTh1 restoration and MSE recovery for TimesFM 2.5 aggregated over the population sweep (N=15 windows, $\varepsilon = 0.2$, MEMIT excluding flagged failures). Left: restoration fraction at $K^\star$ (mean $\pm$ std) for ROME, SeqROME, and MEMIT under PGD, FGSM, and BIM; the dashed line marks perfect restoration R=1. Right: MSE vs. the clean forecast (log scale) for the attacked, ROME, and SeqROME forecasts.

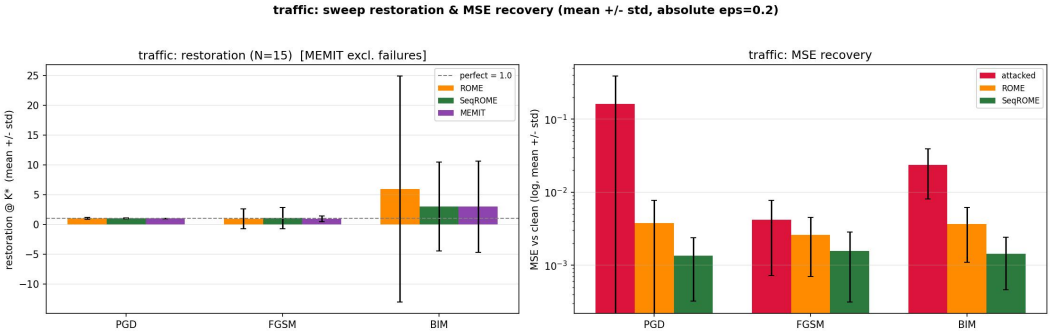


Figure 15: Traffic restoration and MSE recovery for TimesFM 2.5 aggregated over the population sweep (N=15 windows, $\varepsilon = 0.2$, MEMIT excluding flagged failures), in the same format as Figure 14. SeqROME restores most reliably; ROME and MEMIT show larger variance under BIM.

D.3 Cross-channel results

Tables 10 and 11 report the per-channel single-shot ROME results for TimesFM 2.5 on ETTh1 and Traffic; the corresponding Electricity results are given in the main text (Table 3). For each channel, the editing layer $L^\star$ is identified through causal tracing and then reused across PGD, FGSM, and BIM. NRMSE and MAE are measured against the clean forecast, while WQL is measured against the held-out ground truth. These results test whether the localize-and-repair behavior observed on the focal series extends to other channels within each dataset.

ETTh1. Table 10 extends this to the seven ETTh1 variables, whose editing layers concentrate in the earliest blocks. ROME reduces the forecast error on most variables; the smaller gain on MULL shows that a single rank-one edit helps most when the attack-induced behavior is strongly localized for a series.

Table 10: Per-channel single-shot ROME repair on TimesFM 2.5 at $\varepsilon = 0.2$ on the seven ETTh1 variables. $L^\star$ is the causally traced target layer, which varies by channel. NRMSE and MAE are measured against the clean forecast, while WQL is measured against the held-out ground truth.

Channel	Attack	$L^\star$	$\text{NRMSE}_{\text{adv}}$	$\text{NRMSE}_{\text{rome}}$	MAE_{adv}	MAE_{rome}	WQL_{adv}	WQL_{rome}
HUFL	PGD	0	0.20	0.067	1.16	0.395	1.131	0.915
HUFL	FGSM	0	0.19	0.060	1.08	0.351	1.107	0.905
HUFL	BIM	0	0.20	0.066	1.15	0.387	1.129	0.913
HULL	PGD	0	0.21	0.048	0.37	0.078	0.323	0.295
HULL	FGSM	0	0.11	0.049	0.18	0.085	0.313	0.299
HULL	BIM	0	0.16	0.048	0.28	0.079	0.320	0.298
MUFL	PGD	0	0.09	0.038	0.51	0.208	0.992	0.882
MUFL	FGSM	0	0.09	0.038	0.48	0.206	0.983	0.883
MUFL	BIM	0	0.09	0.038	0.50	0.209	0.990	0.883
MULL	PGD	0	0.11	0.076	0.18	0.130	0.326	0.306
MULL	FGSM	0	0.11	0.073	0.16	0.119	0.323	0.309
MULL	BIM	0	0.11	0.075	0.17	0.125	0.325	0.307
LUFL	PGD	1	0.18	0.086	0.46	0.179	0.390	0.262
LUFL	FGSM	1	0.10	0.055	0.23	0.123	0.326	0.270
LUFL	BIM	1	0.18	0.101	0.44	0.220	0.386	0.288
LULL	PGD	0	3.15	0.328	0.98	0.099	0.407	0.080
LULL	FGSM	0	0.49	0.286	0.15	0.081	0.093	0.078
LULL	BIM	0	1.01	0.260	0.33	0.074	0.145	0.075
OT	PGD	0	0.08	0.030	0.39	0.137	0.511	0.465
OT	FGSM	0	0.07	0.020	0.33	0.083	0.504	0.461
OT	BIM	0	0.08	0.027	0.37	0.122	0.508	0.463

Traffic. Table 11 reports seven representative Traffic sensors, most localizing to layer 0. ROME again gives large reductions under PGD and BIM; sensor 840 is the hardest case, improved under all three attacks but with the highest residual error.

Table 11: Per-channel single-shot ROME on TimeFM repair at $\varepsilon = 0.2$ on seven representative Traffic sensors. $L^\star$ is the causally traced target layer, which varies by sensor. NRMSE and MAE are measured against the clean forecast, while WQL is measured against the held-out ground truth.

Channel	Attack	$L^\star$	$\text{NRMSE}_{\text{adv}}$	$\text{NRMSE}_{\text{rome}}$	MAE_{adv}	MAE_{rome}	WQL_{adv}	WQL_{rome}
sensor_406	PGD	0	0.335	0.064	0.103	0.018	0.044	0.020
sensor_406	FGSM	0	0.130	0.113	0.035	0.030	0.025	0.024
sensor_406	BIM	0	0.292	0.158	0.093	0.044	0.037	0.023
sensor_617	PGD	1	0.562	0.059	0.245	0.018	0.099	0.012
sensor_617	FGSM	1	0.042	0.039	0.011	0.010	0.010	0.009
sensor_617	BIM	1	0.109	0.025	0.021	0.006	0.014	0.009
sensor_701	PGD	0	0.184	0.092	0.070	0.023	0.037	0.019
sensor_701	FGSM	0	0.106	0.067	0.031	0.022	0.023	0.020
sensor_701	BIM	0	0.175	0.085	0.064	0.025	0.032	0.019
sensor_632	PGD	0	0.758	0.173	0.308	0.057	0.125	0.026
sensor_632	FGSM	0	0.042	0.038	0.012	0.011	0.012	0.011
sensor_632	BIM	0	0.229	0.033	0.069	0.009	0.033	0.011
sensor_413	PGD	0	0.426	0.022	0.198	0.007	0.077	0.006
sensor_413	FGSM	0	0.014	0.014	0.005	0.005	0.006	0.006
sensor_413	BIM	0	0.021	0.020	0.006	0.006	0.006	0.006
sensor_840	PGD	14	5.537	0.179	0.165	0.005	0.071	0.007
sensor_840	FGSM	14	0.048	0.048	0.001	0.001	0.005	0.005
sensor_840	BIM	14	1.436	0.497	0.032	0.011	0.016	0.008
sensor_470	PGD	0	0.548	0.022	0.207	0.006	0.086	0.006
sensor_470	FGSM	0	0.038	0.018	0.008	0.005	0.005	0.004
sensor_470	BIM	0	0.114	0.012	0.022	0.004	0.012	0.004

Across datasets the editing layer varies both between and within datasets, yet the repair pattern is consistent: single-shot ROME helps most when the attack shifts the forecast substantially (PGD, BIM) and least for near-inert FGSM. The site must be localized per series, but the resulting rank-one repair generalizes across channels, datasets, and attacks.

D.4 Cross-attack results

Across all three datasets the pre-repair severity ordering at $\varepsilon = 0.2$ is PGD $>$ BIM $>$ FGSM (Attacked rows of Tables 8 and 9). The editing site is localized once from PGD and then held fixed for FGSM and BIM, with the edit recomputed per attack, testing whether a PGD-localized site transfers across attack families. Relative recovery tracks the initial severity: a single edit gains most on the strong PGD and BIM attacks and little on near-inert FGSM, where the attacked forecast is already close to clean (e.g. Traffic FGSM), and iterative editing closes more of the residual gap under the strongest attacks (Table 7).

D.5 Attack-strength analysis

Tables 12–14 sweep the attack budget $\varepsilon \in \{0.1, 0.2, 0.3, 0.4\}$ with a single-seed ROME edit located at the primary budget. Pre-repair error grows steadily with ε while post-repair error stays low across the range, so the edit located at one budget continues to work as the attack strengthens. FGSM's scalar-slope restoration $R_{K^\star}$ is unstable at the weakest attacks because its baseline forecast shift is near zero, which ill-conditions the normalized ratio even when the MSE is already tiny.

Table 12: ε sweep for TimesFM 2.5 on `MT_253` (Electricity). MSE and WQL are measured against the clean forecast; $R_{K^\star}$ is the scalar-slope restoration at the maximum-deviation forecast step. FGSM's $R_{K^\star}$ is unstable because its baseline forecast shift is near zero.

Attack	ε	MSE_{adv}	MSE_{rome}	WQL_{rome}	$R_{K^\star}$
PGD	0.10	0.1237	0.0003	0.0121	+1.00
	0.20	0.9071	0.0006	0.0125	+1.02
	0.30	2.5542	0.0009	0.0148	+1.02
	0.40	6.0584	0.0004	0.0118	+1.00
FGSM	0.10	0.0009	0.0004	0.0131	+0.51
	0.20	0.0004	0.0003	0.0137	+0.08
	0.30	0.0004	0.0003	0.0136	+0.16
	0.40	0.0006	0.0004	0.0139	+0.36
BIM	0.10	0.0379	0.0002	0.0107	+0.98
	0.20	0.1470	0.0010	0.0152	+0.96
	0.30	0.3296	0.0003	0.0131	+1.00
	0.40	0.8223	0.0008	0.0174	+1.02

Table 13: Attack-budget sweep for TimesFM 2.5 on ETTh1 (`OT`), single-seed ROME repair, extending the Electricity sweep in Table 12 to $\varepsilon \in \{0.1, 0.2, 0.3, 0.4\}$. All MSE/WQL values are against the clean forecast; $R_{K^\star}$ is scalar-slope restoration at the max-deviation step.

Attack	ε	MSE_{adv}	MSE_{rome}	WQL_{rome}	$R_{K^\star}$
PGD	0.10	0.036	0.004	0.334	+1.02
	0.20	0.156	0.020	0.333	+0.81
	0.30	0.367	0.028	0.330	+1.08
	0.40	0.683	0.056	0.337	+1.14
FGSM	0.10	0.033	0.004	0.335	+1.11
	0.20	0.112	0.009	0.331	+0.97
	0.30	0.221	0.023	0.329	+0.83
	0.40	0.347	0.032	0.334	+1.00
BIM	0.10	0.035	0.004	0.335	+1.01
	0.20	0.141	0.017	0.331	+0.80
	0.30	0.334	0.028	0.330	+0.97
	0.40	0.616	0.041	0.335	+1.20

Table 14: Attack-budget sweep on Traffic for TimesFM (highest-variance sensor), single-seed ROME repair, extending the Electricity sweep in Table 12 to $\varepsilon \in \{0.1, 0.2, 0.3, 0.4\}$. All MSE/WQL values are against the clean forecast; $R_{K^\star}$ is scalar-slope restoration at the max-deviation step. FGSM's $R_{K^\star}$ is ill-conditioned because its baseline forecast shift is near zero.

Attack	ε	MSE_{adv}	MSE_{rome}	WQL_{rome}	$R_{K^\star}$
PGD	0.10	0.016	0.000	0.012	+0.99
	0.20	0.017	0.001	0.014	+0.94
	0.30	0.022	0.006	0.022	+0.89
	0.40	0.020	0.008	0.024	+1.04
FGSM	0.10	0.001	0.001	0.012	+0.60
	0.20	0.003	0.002	0.018	+0.32
	0.30	0.004	0.003	0.023	+0.20
	0.40	0.004	0.003	0.028	+0.19
BIM	0.10	0.011	0.001	0.012	+0.90
	0.20	0.013	0.004	0.016	+0.29
	0.30	0.017	0.001	0.014	+0.94
	0.40	0.016	0.001	0.016	+0.64