SAME ERROR, DIFFERENT CONSEQUENCES: INFORMATION INTERFACES BETWEEN DETERMINISTIC COMPLETION AND STOCHASTIC TIME-SERIES GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deterministic completion is often used as an intermediate step before stochastic generation for irregular time series, and its quality is typically assessed by pointwise reconstruction errors such as MAE or RMSE. However, two completions with the same reconstruction error need not be equally useful as conditions for a downstream generator. This raises a basic question: *is completion accuracy sufficient to characterize downstream conditioning utility?* We study this question through the information interface between deterministic completion and stochastic generation. To separate *how much* a completion is wrong from *how it is wrong*, we construct controlled interventions that match aggregate missing-region RMSE while changing the temporal organization of completion errors. We then measure how these interventions alter the relative utility of transmitting the full completion versus a coarse temporal representation to the generator. Across four time-series datasets and two downstream generator architectures, we find that aggregate completion RMSE alone is insufficient to predict downstream behavior. Reorganizing learned completion errors with RMS matching can substantially change generation quality and the relative preference between full and coarse interfaces. Importantly, the direction of this effect is not universal: sensor datasets and a prospectively preregistered financial-domain test exhibit different, and in some cases opposite, interface shifts. These results show that *completion accuracy and downstream conditioning utility are distinct quantities*, and suggest that the information interface between deterministic completion and stochastic generation should be treated as a modeling problem in its own right rather than as a passive handoff of the completed sequence.

1 INTRODUCTION

Completion error measures distance to a target, but does not specify where those errors occur in time. Two completed trajectories can therefore have the same aggregate RMSE while providing different temporal cues to a downstream model. This distinction matters in completion-to-generation pipelines: a deterministic model supplies a dense point estimate, which a stochastic model then consumes as a condition (Fadlon et al., 2025). The first stage is evaluated for reconstruction accuracy, whereas the second needs information that supports plausible conditional trajectories. We ask whether the former is sufficient to characterize the latter.

We turn this observation into a controlled question. Let FULL pass the dense completion to the generator and let AVG16 pass a fixed Avg16 temporal average. For a residual intervention T, define

$$\Delta_T = \mathrm{CRPS}_{\mathrm{FULL},T} - \mathrm{CRPS}_{\mathrm{AVG16},T}. \tag{1}$$

Positive Δ_T means that Avg16 is better for that condition. Our primary question is whether the Full-versus-Avg16 comparison moves when the completion residual is time-shuffled:

$$S_{\mathrm{time}} = \Delta_{\mathrm{aligned}} - \Delta_{\text{time-shuffle}}. \tag{2}$$

Our intervention shuffles residuals in time, reapplies the observation mask, and rescales the missing-region residual to match its original aggregate RMS. It holds this scalar error measure fixed, not the

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.

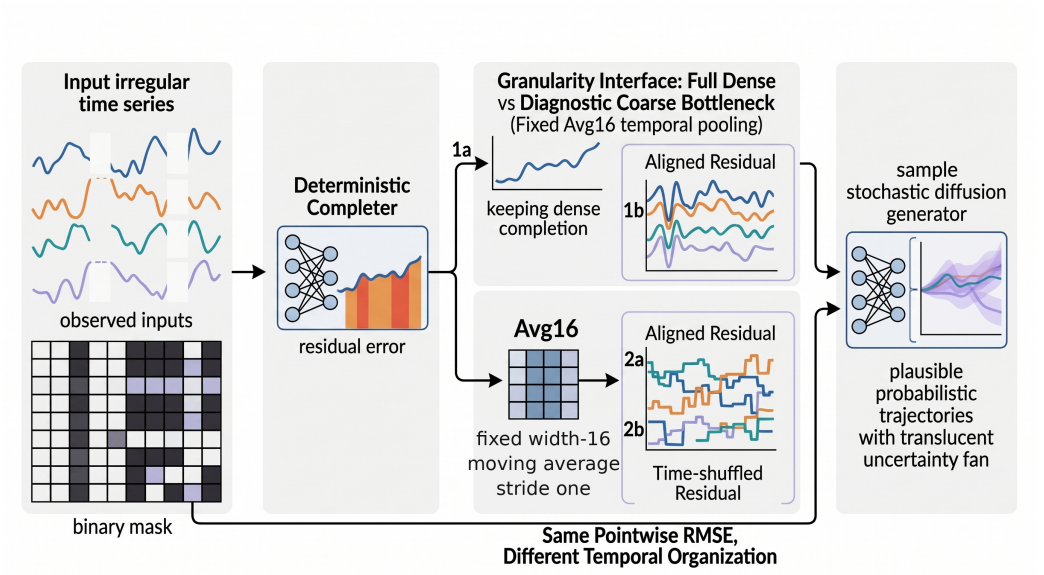

Figure 1: Completion-to-generation interface. Full and Avg16 are compared under both aligned and shuffled residuals; observations and masks remain available to every generator. Avg16 uses stride-one moving averaging, while shuffled residuals are remasked and matched in aggregate missing-region RMS.

entire residual distribution. A positive S_{time} means that the aligned learned residual makes Avg16 relatively more favorable than the shuffled residual; it does not mean that Avg16 is universally better.

Structured conditioning is already central to CSDI and PriSTI (Tashiro et al., 2021; Liu et al., 2023), and ImagenI2R already combines completion with generation (Fadlon et al., 2025). Our question is not whether such conditioning is useful. Ordinary comparisons between completers change accuracy, error structure, and often the consumer simultaneously. They do not determine whether, with observed evidence, generator architecture and budget, and aggregate residual RMSE fixed, the organization of an intermediate completion still change what the generator should trust? We study this question as an intervention problem rather than proposing a new generator or a universal routing rule.

We evaluate four datasets, two downstream denoisers, and frozen completion-pair artifacts. The primary intervention is first tested on the ETT-family datasets and Weather96, then on a prospectively preregistered stock-domain check. The result is clear but bounded. ETTh1 and ETTm1 show positive aligned-versus-time-shuffled shifts under both Conv1D and Transformer denoisers. Weather96 is backbone dependent. The preregistered stock/Conv1D test reverses the hypothesized direction, with $S_{\text{time}} = -0.213$ and zero positive seeds; a later stock/Transformer diagnostic shows an even stronger reversal of -0.436. The experiment therefore establishes that completion error organization can change downstream interface utility, while rejecting a universal sign.

This result changes the claim we make. We do not claim that coarse conditioning is generally better, that larger completion error calls for more compression, or that one residual descriptor already predicts the correct interface. We claim a narrower and testable principle: completion accuracy is not a sufficient description of completion utility when the completion is passed to a finite stochastic learner.

Our contributions are:

1. We formulate deterministic completion followed by stochastic generation as an information-interface problem with an explicit Full-versus-Avg16 statistic.
2. We construct RMS-matched residual interventions to test whether aggregate reconstruction accuracy determines downstream conditioning utility.
3. We show a strong, sign-varying effect across four datasets and two denoisers, including a preregistered independent-domain failure that defines the boundary of the claim.

4. We provide seed-wise, condition-sensitivity, and secondary-transform evidence that makes the negative result part of the scientific conclusion rather than an omitted exception.

2 Related Work

Irregular and incomplete time series. Irregular sampling has motivated models that explicitly represent observation gaps, continuous time, or missingness structure. GRU-D uses time gaps and decay to distinguish stale from recent observations (Che et al., 2018); BRITS performs bidirectional recurrent imputation with a consistency mechanism (Cao et al., 2018); and mTAN represents sparse irregular sequences through continuous-time attention and interpolation (Shukla & Marlin, 2021). Latent ODEs and Neural CDEs instead model dynamics between observations in continuous time (Rubanova et al., 2019; De Brouwer et al., 2019; Kidger et al., 2020). SAITS shows that self-attention can also provide a strong multivariate imputation representation (Du et al., 2023). These works improve the deterministic representation available to a downstream model, but do not isolate what happens after that representation crosses into a separate stochastic generator.

Probabilistic imputation and time-series generation. Diffusion models provide a general stochastic generation framework (Ho et al., 2020; Song et al., 2021). DeepAR provides an earlier autoregressive probabilistic baseline (Salinas et al., 2020); TimeGrad applies autoregressive diffusion to probabilistic forecasting (Rasul et al., 2021), while CSDI conditions score-based diffusion on observed values and masks (Tashiro et al., 2021). SSSD uses structured state-space layers for diffusion-based imputation and forecasting (Lopez Alcaraz & Strodthoff, 2023); TimeDiff studies non-autoregressive conditional diffusion for time-series prediction (Shen & Kwok, 2023); and PriSTI constructs a coarse spatiotemporal prior for conditional imputation (Liu et al., 2023). Diffusion-TS studies interpretable diffusion for general time-series generation (Yuan & Qiao, 2024). These methods establish that the representation given to a stochastic model matters, but their main comparisons change the model or training objective rather than holding the downstream learner fixed while reorganizing an upstream error. TimeGAN provides a non-diffusion reference point for synthetic sequence generation (Yoon et al., 2019); our generator is instead used as a controlled consumer of a completion interface.

Completion-to-generation interfaces. Image-space diffusion models transform sequences into images before generative modelling (Naiman et al., 2024), and ImagenI2R is especially close to our setting: it completes irregular sequences and then uses a masked image-space diffusion model for regular generation (Fadlon et al., 2025). Frequency- and multiscale-aware methods further show that temporal scale and spectral structure can affect missing-data reconstruction (Yang et al., 2024; Fons et al., 2025; Wang & Sugiyama, 2026). Information-bottleneck methods ask which temporal information should be retained or suppressed (Choi & Lee, 2024; Yang et al., 2025). Our question is narrower than these objectives: we treat the completer as fixed, define Full and Avg16 as an observable interface, and intervene on the organization of the completer residual while matching its aggregate RMS. The contribution is not a new imputation architecture; it is an empirical test of whether completion accuracy is enough to characterize downstream conditioning utility.

Accuracy metrics and downstream utility. Patch-based and multiscale forecasters such as PatchTST and TimesNet already show that reorganizing temporal context changes what a finite learner can use (Nie et al., 2023; Wu et al., 2023). A parallel line of work on generative models separates distortion from perceptual or sample quality and formalizes the tradeoff between them (Blau & Michaeli, 2018; 2019), while information-bottleneck arguments distinguish the information a representation carries from the information a finite learner can exploit (Tishby & Zaslavsky, 2015). Our setting makes this tension concrete and narrow: the upstream quantity is a pooled completion RMSE, the downstream quantity is the CRPS of a conditional generator, and the generator's architecture, observed evidence, and training budget are held fixed while only the organization of the intermediate completion error changes. The closest prior systems ask how to improve a completion or how to train a conditional generator; we ask whether aggregate completion accuracy determines the relative value of dense and coarse conditioning once both stages are already present.

3 MEASURING COMPLETION ACCURACY AND CONDITIONING UTILITY

The question in the abstract requires two controls: a measurable downstream comparison and an intervention that changes the condition without changing its aggregate RMSE. We define both explicitly, including the limits of what is held fixed.

3.1 DETERMINISTIC COMPLETION FOLLOWED BY STOCHASTIC GENERATION

Let $X \in \mathbb{R}^{T\times C}$ denote a complete regular multivariate sequence and $M \in \{0,1\}^{T\times C}$ an observation mask. The observed evidence is

$$X_{\text{obs}} = M \odot X. \tag{3}$$

A deterministic completer f_ϕ maps the irregular evidence to a dense estimate,

$$\hat{X}_c = f_\phi(X_{\text{obs}}, M), \qquad E = (1-M)\odot(\hat{X}_c - X), \tag{4}$$

where E is the completion residual on missing coordinates. Observed coordinates are restored exactly before any downstream experiment.

A stochastic generator receives X_{obs}, M, and an interface S. Within a backbone, observations, targets, the diffusion schedule, and the training budget are fixed; each interface/intervention combination is trained separately, so "fixed generator" means a matched architecture and procedure rather than a shared checkpoint.

3.2 FULL AND COARSE COMPLETION INTERFACES

The dense interface is

$$S_{\text{Full}} = \hat{X}_c. \tag{5}$$

The coarse interface applies a fixed temporal smoothing operator P_{16},

$$S_{\text{Avg16}} = P_{16}(\hat{X}_c). \tag{6}$$

In the implementation, P_{16} is a width-16 stride-one moving average with replicated boundary padding that retains the original sequence length; it is not blockwise downsampling. Avg16 was fixed before the final cross-dataset stage.

Because $P_{16}(\hat{X}_c)$ is a deterministic function of $\hat{X}_c$, FULL contains at least the information available to AVG16. Consequently, a Bayes-optimal learner with unlimited data and capacity need not benefit from discarding the dense condition. Any empirical difference between the two interfaces therefore reflects the interaction between the finite downstream learner and the structure of its condition, rather than an information-theoretic dominance claim.

3.3 AN OPERATIONAL DEFINITION OF UTILITY

We measure conditioning utility by held-out missing-region CRPS of the generator trained with that interface: lower CRPS means greater utility under this protocol. This is an achieved performance measure, not a minimum over all possible learners.

3.4 RESIDUAL INTERVENTIONS

Write $Q = 1 - M$ and let $R_Q(E)$ be the RMS pooled over all missing coordinates in one split and pair seed. For a transformation τ, we construct

$$\widetilde{E}_\tau = Q \odot \tau(E), \qquad E_\tau = \frac{R_Q(E)}{R_Q(\widetilde{E}_\tau)}\widetilde{E}_\tau, \qquad \hat{X}_c^{(\tau)} = X + E_\tau. \tag{7}$$

For nonzero residual RMS this gives $R_Q(E_\tau) = R_Q(E)$ up to floating-point precision. The aligned arm uses E directly. The implementation skips rescaling when either RMS is below 10^{-8}. The primary comparison is:

- **Aligned**: retain the learned completer residual.

- **Time shuffle**: use a separate temporal permutation for each window, shared across channels, then remask and RMS-match using Eq. 7.

The permutation alone preserves residual vectors; remasking and scaling do not preserve the final missing-region value multiset or MAE. Individual-window RMSE also need not be equal. The identified contrast is therefore a compound, aggregate-RMSE-matched intervention, not a pure change of order with everything else held fixed. Secondary phase-random and cross-sample arms use the same remasking and RMS-matching procedure.

The interface is illustrated in Figure 1; the case study in Appendix A.3 makes the measurement question concrete.

3.5 Statistics of interest

For each dataset, backbone, seed, and residual intervention, define

$$\Delta = \mathrm{CRPS}_{\mathrm{Full}} - \mathrm{CRPS}_{\mathrm{Avg16}}, \tag{8}$$

so $\Delta > 0$ indicates lower CRPS for the coarse interface. Our original confirmatory contrast was

$$S_{\mathrm{time}} = \Delta_{\mathrm{aligned}} - \Delta_{\mathrm{time\text{-}shuffle}}. \tag{9}$$

A positive value means that preserving the learned residual's temporal organization makes the coarse interface relatively more favorable than time shuffling does.

3.6 What the counterfactual can identify

For a fixed pair seed, aligned and shuffled arms share observations, masks, targets, source residuals, and downstream training settings. Interventions are constructed separately for training, validation, and test splits. Ground truth is required to form residuals and match their RMS, so this is a controlled measurement instrument, not an inference-time procedure available on naturally unobserved values. It tests the sufficiency of aggregate RMSE without identifying which altered distributional property explains the result.

The absolute sign of Δ is not the primary outcome: either sign is compatible with the primary question, and S_{time} allows a positive result even when Full remains the better absolute interface.

Appendix A.2.5 adds a post-hoc in-mask permutation that leaves pooled RMS, per-window RMSE and MAE unchanged, so it bounds how much of S_{time} can be attributed to temporal order rather than to the MAE change that remasking induces.

4 Experimental Design

We verify the error-magnitude control, measure downstream quality and interface gaps, then test the direction across datasets and architectures. Secondary controls probe the scope of the result.

4.1 Datasets and temporal protocol

We evaluate Weather96, ETTh1, ETTm1, and a six-channel daily stock dataset distributed with the ImagenI2R resources (Fadlon et al., 2025); the first three are weather and electricity/sensor data, and stock provides an intentionally different financial domain. All experiments use sequence length $T = 96$, a single contiguous time-block mask shared across channels with mean missing fraction 0.50, chronological train/validation/test splits, windows formed within split boundaries, and train-only normalization statistics.

Table 1 summarizes the final protocol. Weather96/Conv1D was extended to five generator seeds after it emerged as a weak boundary; the remaining primary cells use three seeds. We report seed-wise directions and means rather than asymptotic significance claims from only three seeds.

4.2 Completer and immutable pairs

The deterministic stage is a mask-aware Transformer completer. For each dataset, missingness pattern, and pair seed, we precompute an immutable bundle containing X, M, X_{obs}, and $\hat{X}_c$. A *pair*

Table 1: Final experimental protocol. All settings use sequence length 96, a single contiguous time-block mask shared across channels with mean missing fraction 0.50, pair seeds 0–2, train-only normalization, a mask-aware Transformer completer, and 5120 downstream optimizer steps. Completer MAE and RMSE are test-split missing-region values from Table 5.

Dataset	Ch.	Rows	Windows (tr/va/te)	Missing	Pair seeds	Compl. MAE	Compl. RMSE	Generator seeds
Weather96	21	52696	4096/768/768	0.50	0–2	0.190	0.405	Conv1D (5), Tx (3)
ETTh1	7	17419	4096/768/768	0.50	0–2	0.414	0.606	Conv1D (3), Tx (3)
ETTm1	7	69680	8192/1024/1024	0.50	0–2	0.394	0.602	Conv1D (3), Tx (3)
Stock	6	3685	2484/457/459	0.50	0–2	1.904	2.135	Conv1D (3), Tx (3)

seed fixes the sampled mask and the resulting completer output for a window; a *generator seed* is an independent downstream initialization trained on that same frozen pair. Hashes are stored and checked before downstream training. This prevents the mask/completion mismatch that can arise when a precomputed completion is paired with a regenerated mask. All residual transforms inherit the same observations and masks; the residual transform seed is derived deterministically from the pair seed and the split. Appendix A.1.3 compares the completer with two non-learned baselines: it is clearly better than both on ETTh1 and ETTm1, roughly tied on Weather96, and worse on Stock.

4.3 Downstream stochastic generators

We use a temporal Conv1D residual network and a Transformer denoiser within a conditional DDPM. Both receive noisy target state, diffusion time, X_{obs}, M, and S. Training supervises missing coordinates; sampling restores observations. Runs use 5120 AdamW steps, learning rate 2×10^{-4}, batch size 64, and 100 diffusion steps. Checkpoints are selected by validation noise-prediction loss. Equal steps do not imply equal parameter counts or compute. Evaluation uses 20 samples per window. CRPS averages the empirical score over missing coordinates; MAE uses the sample-mean prediction.

4.4 Evaluation units and uncertainty reporting

The basic evaluation unit is a test window, but neighboring windows overlap and are not treated as independent replications. We therefore summarize each condition by seed-wise means and standard deviations, and emphasize paired seed directions for S_{time}. Error bars in the main figures are across-seed standard deviations, not confidence intervals, and we claim no asymptotic significance from three seeds. Pair seeds also change masks and residual interventions, so across-seed variability is not pure optimizer variability.

The test interface is fixed without using test outcomes: Avg16 is fixed before the final cross-dataset stage, and every condition uses the same interface pair, so S_{time} is a controlled movement of one predefined Full-versus-Coarse comparison.

4.5 Confirmatory sequence and transparency

The experiments evolved from exploratory studies into prospective checks. The ETTm1 Conv1D contrast was preregistered before its generator runs and supported the positive-shift hypothesis. We then preregistered an independent-domain stock/Conv1D replication before any stock run; that test *failed in the opposite direction*, and under the predeclared stop rule we retained the failure as a boundary rather than replacing the dataset or changing the contrast. A stock Transformer run was later added on the same frozen pairs as a descriptive diagnostic, not an independent confirmation.

4.6 Condition-sensitivity sanity check

On Stock/Conv1D we separately train Full, Avg16, blank, sample-shuffled Avg16, and target-derived oracle Avg16 arms. The sample-shuffled interface rolls smoothed conditions across samples within each batch; it is distinct from shuffling residuals along time. This diagnostic asks whether better matched information improves scores. It is neither an interface-selection procedure nor a fixed-checkpoint test of causal reliance on individual condition features.

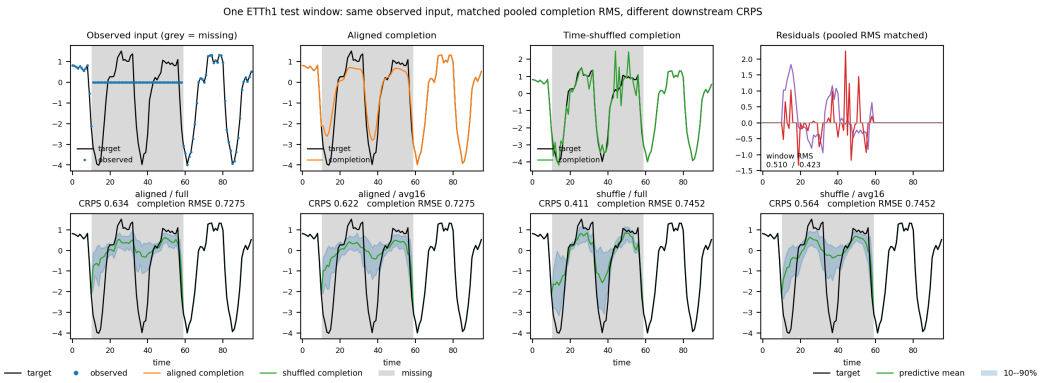

Figure 2: One fixed ETTh1 test window (Conv1D, pair seed 0, sample 0, channel 0). Top row: the observed input with the missing block shaded, the aligned and time-shuffled completions, and their residuals; the two residuals share pooled RMS over the split but not within this window (0.510 vs 0.423). Bottom row: downstream 10–90% intervals and predictive means for all four interface/intervention conditions, annotated with CRPS and completion RMSE. Within each intervention the two interfaces share the same completion RMSE, yet their CRPS differs. This window is illustrative, not a new aggregate estimate.

5 Results

Figure 2 first shows one fixed window in which pooled completion RMS is matched but the downstream predictive bands differ; the aggregate comparisons follow.

5.1 Matched completion RMSE does not fix downstream quality

We first check the control needed to answer the abstract's question. Across four datasets, pair seeds 0–2, and three splits, the frozen metadata give 36 aligned/shuffled comparisons with maximum pooled RMSE discrepancy below 2.4×10^{-7} (Appendix Table 10). MAE is not matched. This supports an aggregate-RMSE comparison, not an assertion that all reconstruction-error statistics are unchanged.

Despite this control, downstream CRPS changes substantially: on ETTh1/Conv1D the shuffled residual lowers Full CRPS by 0.22 but Avg16 by only 0.07, whereas on Stock/Conv1D it raises Full CRPS by 0.13 and lowers Avg16 by 0.08. Equal aggregate completion RMSE therefore does not imply equal downstream quality.

5.2 The interface shift depends on dataset and backbone

Table 2 and Appendix Figure 5 quantify how the Full-versus-Avg16 gap responds to the intervention. ETTh1 gives $S_{\text{time}} = +0.150$ for Conv1D and $+0.152$ for Transformer, with all three seeds positive in both cases. ETTm1 shows the same direction, $+0.042$ and $+0.180$, again with all three seeds positive. Weather96 is backbone dependent: Conv1D gives -0.008 with one of five positive seeds, whereas Transformer gives $+0.080$ with all three seeds positive.

The financial-domain test probes whether that positive direction transfers. The prospective stock/Conv1D test predicted $S_{\text{time}} > 0$ but observed

$$S_{\text{time}} = -0.213, \tag{10}$$

with 0/3 positive seeds. On the same frozen pairs, the later Transformer diagnostic gives -0.436, also 0/3 positive. Thus the positive direction fails on the prospectively tested independent domain; the Transformer result is a post-result diagnostic, not a second independent confirmation.

Pooling seed-level signs shows the same split: the ETT family is positive in 12/12 seed runs (sign test $p \approx 4.9 \times 10^{-4}$) and the stock domain is negative in 6/6 ($p \approx 0.031$), while Weather96 disagrees across backbones. These counts treat seed runs as independent, which they are not, and no individual

Table 2: Primary aligned-vs-time-shuffle shift. $\Delta = \mathrm{CRPS}_{Full} - \mathrm{CRPS}_{Avg16}$ and $S_{time} = \Delta_{aligned} - \Delta_{shuffle}$. Positive S_{time} means aligned residual organization makes Avg16 relatively more favorable than time shuffling. SD is the across-seed standard deviation and Pos. the number of positive seeds; the last column gives the decision under the preregistered rule $S_{time} > 0$.

Dataset	Denoiser	S_{time}	SD	Seeds	Pos.	Decision
Weather96	Conv1D	-0.008	0.016	5	1	not supported
Weather96	Transformer	+0.080	0.027	3	3	supported
ETTh1	Conv1D	+0.150	0.022	3	3	supported
ETTh1	Transformer	+0.152	0.019	3	3	supported
ETTm1	Conv1D	+0.042	0.018	3	3	supported
ETTm1	Transformer	+0.180	0.028	3	3	supported
Stock	Conv1D	-0.213	0.137	3	0	reversed
Stock	Transformer	-0.436	0.304	3	0	reversed

Table 3: Complete primary results, seed by seed. For every cell and generator seed, Δ_a and Δ_s are $\mathrm{CRPS}_{Full} - \mathrm{CRPS}_{Avg16}$ under the aligned and the time-shuffled completion residual, and $S_{time} = \Delta_a - \Delta_s$; positive Δ favors Avg16, and a positive S_{time} supports the preregistered hypothesis. The italic row of each cell gives the across-seed SD, the fraction of positive seeds, the mean S_{time}, and the resulting decision.

		Aligned residual			Time-shuffled residual				
Cell	Seed	Full	Avg16	Δ_a	Full	Avg16	Δ_s	S_{time}	Result
Weather96 / Conv1D	0	0.4708	0.4657	+0.0051	0.4051	0.3803	+0.0248	-0.0197	−
	1	0.4234	0.4170	+0.0063	0.3393	0.3493	-0.0101	+0.0164	+
	2	0.4560	0.4497	+0.0063	0.3726	0.3547	+0.0180	-0.0117	−
	3	0.4529	0.4478	+0.0051	0.3759	0.3485	+0.0275	-0.0223	−
	4	0.4440	0.4422	+0.0018	0.3612	0.3578	+0.0033	-0.0016	−
	mean		*SD 0.016,*	*1/5 positive seeds*				*-0.008*	*not supported*
Weather96 / Transformer	0	0.4671	0.4592	+0.0079	0.2595	0.3592	-0.0997	+0.1076	+
	1	0.4479	0.4372	+0.0107	0.3050	0.3734	-0.0683	+0.0790	+
	2	0.4620	0.4479	+0.0141	0.3101	0.3488	-0.0388	+0.0528	+
	mean		*SD 0.027,*	*3/3 positive seeds*				*+0.080*	*supported*
ETTh1 / Conv1D	0	0.4978	0.5057	-0.0079	0.2951	0.4350	-0.1399	+0.1320	+
	1	0.4901	0.5148	-0.0247	0.2606	0.4601	-0.1995	+0.1748	+
	2	0.5000	0.5103	-0.0103	0.2832	0.4362	-0.1530	+0.1427	+
	mean		*SD 0.022,*	*3/3 positive seeds*				*+0.150*	*supported*
ETTh1 / Transformer	0	0.5233	0.5217	+0.0016	0.2049	0.3711	-0.1662	+0.1678	+
	1	0.5244	0.5112	+0.0132	0.2350	0.3783	-0.1433	+0.1565	+
	2	0.5195	0.5303	-0.0109	0.2224	0.3644	-0.1420	+0.1311	+
	mean		*SD 0.019,*	*3/3 positive seeds*				*+0.152*	*supported*
ETTm1 / Conv1D	0	0.5172	0.5181	-0.0010	0.3468	0.4088	-0.0620	+0.0610	+
	1	0.5235	0.5260	-0.0024	0.3669	0.3945	-0.0276	+0.0252	+
	2	0.5143	0.5188	-0.0045	0.3654	0.4084	-0.0430	+0.0385	+
	mean		*SD 0.018,*	*3/3 positive seeds*				*+0.042*	*supported*
ETTm1 / Transformer	0	0.5396	0.5386	+0.0010	0.2192	0.4299	-0.2107	+0.2117	+
	1	0.5207	0.5202	+0.0005	0.2529	0.4120	-0.1591	+0.1596	+
	2	0.5193	0.5225	-0.0031	0.2391	0.4123	-0.1732	+0.1700	+
	mean		*SD 0.028,*	*3/3 positive seeds*				*+0.180*	*supported*
Stock / Conv1D	0	4.0892	4.0649	+0.0243	4.2810	3.9899	+0.2911	-0.2668	−
	1	3.8197	3.7623	+0.0574	4.0102	3.6382	+0.3720	-0.3146	−
	2	4.0229	3.9783	+0.0447	4.0444	3.9423	+0.1021	-0.0574	−
	mean		*SD 0.137,*	*0/3 positive seeds*				*-0.213*	*not supported*
Stock / Transformer	0	3.7870	3.7431	+0.0439	3.7701	3.1118	+0.6582	-0.6143	−
	1	3.7593	3.6121	+0.1472	3.8508	3.0942	+0.7566	-0.6094	−
	2	4.0587	3.9970	+0.0617	3.6918	3.5447	+0.1471	-0.0854	−
	mean		*SD 0.304,*	*0/3 positive seeds*				*-0.436*	*not supported*

cell reaches two-sided significance; we therefore read the dataset-level sign pattern as the evidence. Appendix A.4.2 adds two further seeds per cell.

5.3 An interface shift is not necessarily a winner reversal

The absolute gap and its change answer different questions. ETTh1/Conv1D has $\Delta_{\mathrm{aligned}} = -0.0143$ and $\Delta_{\mathrm{shuffle}} = -0.1641$: the shift is positive, but Full wins in both arms. Stock/Conv1D has gaps $+0.0421$ and $+0.2551$: its shift is negative, but Avg16 wins in both. Weather96/Transformer does

change winners, from $+0.0109$ to -0.0689. Accordingly, we report relative utility rather than equating every nonzero S_{time} with a reversal of the preferred interface.

S_{time} measures differential response, not the absolute quality of either interface; absolute gaps for every cell are in Table 18.

5.4 Secondary controls and intervention boundaries

Appendix Figure 4 shows non-parallel responses to phase randomization and cross-sample transfer. Aligned-minus-cross-sample shifts are -0.0198 and -0.0303 on ETTm1 (0/3 positive seeds), and $+0.1332$ and $+0.2148$ on Stock (3/3). Weather96/Transformer has a small negative aligned-minus-phase-random shift (-0.0035, 1/3). Time shuffling remains the confirmatory intervention; other transforms are descriptive boundary controls.

Condition sensitivity. Appendix A.4.9 reports the stock sanity arms; they are consistent with benefit from sample-specific completion information.

Together, the results support interface sensitivity under matched pooled RMSE, but not a universal direction.

5.5 Coarsening scale is itself domain dependent

The primary contrast fixes the coarse operator at width 16. Holding the completer and the aligned completion fixed, we sweep $w \in \{2, 4, 8, 16, 32\}$ on three datasets and both denoisers (Appendix Table 20); the $w = 16$ column reproduces the primary aligned gap. The response is domain and backbone dependent: Stock improves monotonically with width under both denoisers, reaching $+0.198$ at $w = 32$ with the Transformer; ETTh1 turns negative beyond $w = 8$ under Conv1D but not under the Transformer; and ETTm1 stays within 0.004 of zero. Coarsening is therefore not a free knob.

6 Discussion and Limitations

Our results distinguish completion accuracy from downstream conditioning utility, but do not identify a universally preferred interface: the stock reversal precludes choosing Full or Avg16 from RMSE alone, and no descriptor has been prospectively validated to predict the sign. The primary intervention matches pooled missing-region RMS only, so it does not isolate temporal order from MAE, sparsity, or higher-order effects; it establishes insufficiency of pooled RMSE, not that temporal order is the sole cause.

Scope is limited to one completer family, one Avg16 operator, two denoisers, and mostly three seeds; overlapping windows are not independent replications. Stock's reversal is unexplained, and marginal CRPS does not fully measure joint trajectory quality. A next step is to preregister and test an interface-selection rule in a new domain.

An exploratory in-mask permutation that preserves pooled RMS, MAE and the missing-value multiset is reported in Appendix A.2.5; it bounds how much of the primary sign contrast is attributable to temporal order rather than to the MAE change induced by remasking.

7 Conclusion

Is completion accuracy sufficient to characterize downstream conditioning utility? For aggregate missing-region RMSE, our results say no: RMS-matched interventions change downstream CRPS and the Full-versus-Avg16 gap across four datasets and two denoisers. Weather96 is backbone dependent and the prospective Stock/Conv1D test reverses the prediction. Completion should therefore be evaluated with its downstream consumer; a general interface-selection rule remains open.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify methods and protocol; the appendix reports seed-wise results and the frozen-pair RMS audit. Configurations, pair hashes, and preregistration timestamps are retained with the artifacts; the stock Transformer arm is diagnostic, not prospective.

AI USE STATEMENT

Language-model assistance was used for experimental planning, code drafting and debugging, result interpretation, literature search, manuscript organization and editing, and conceptual-figure planning and captions. Reported numerical results come from executed experiments and saved evaluation artifacts, not invented language-model outputs. During revision, code-to-text checks corrected the description of residual RMS matching and Avg16 smoothing; a script recomputed the primary contrasts from stored results. The authors are responsible for the final text, references, claims, and artifacts, including all AI-assisted content.

REFERENCES

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 675–685, 2019.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018. doi: 10.1038/s41598-018-24271-9.

MinGyu Choi and Changhee Lee. Conditional information bottleneck approach for time series imputation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/de7858e3e7f9f0f7b2c7bfdc86f6d928-Abstract-Conference.html.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. In *Advances in Neural Information Processing Systems*, 2019.

Wenjie Du, David Cote, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023.

Gal Fadlon, Idan Arbiv, Nimrod Berman, and Omri Azencot. A diffusion model for regular time series generation from irregular data with completion and masking. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2510.06699.

Elizabeth Fons, Alejandro Sztrajman, Yousef El-Laham, Luciana Ferrer, Svitlana Vyetrenko, and Manuela Veloso. Lscd: Lomb–scargle conditioned diffusion for time series imputation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17411–17436, 2025. URL https://proceedings.mlr.press/v267/fons25a.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Mingzhe Liu, Han Huang, Hao Feng, Leilei Sun, Bowen Du, and Yanjie Fu. Pristi: A conditional diffusion framework for spatiotemporal imputation. In *IEEE International Conference on Data Engineering*, pp. 1927–1939, 2023. doi: 10.1109/ICDE55515.2023.00150.

Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *Transactions on Machine Learning Research*, 2023.

Ilan Naiman, Nimrod Berman, Itai Pemper, Idan Arbiv, Gal Fadlon, and Omri Azencot. Utilizing image transforms and diffusion models for generative modeling of short and long time series. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3868.

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

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent odes for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

Lifeng Shen and James T. Kwok. Non-autoregressive conditional diffusion models for time series prediction. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 31016–31029, 2023.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*, 2021.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2015.

Hangtian Wang and Mahito Sugiyama. Time-gated multi-scale flow matching for time-series imputation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/51d317df78eded9eb3c9d3fb1091c279-Abstract-Conference.html.

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

Jie Yang, Kexin Zhang, Guibin Zhang, Philip S. Yu, and Kaize Ding. Glocal information bottleneck for time series imputation. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-3489.

Xinyu Yang, Yu Sun, Xiaojie Yuan, and Xinyang Chen. Frequency-aware generative models for multivariate time series imputation. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-1667.

Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Xue Yuan and Yixuan Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. In *International Conference on Learning Representations*, 2024.

A Additional Results and Protocol Details

This appendix is organized in five parts: experimental details (A.1), the residual interventions and their controls (A.2), a complete worked example on one window (A.3), additional results (A.4), and reproducibility and claim discipline (A.5). Every number quoted below is recomputed from the saved artifacts by the scripts listed in A.5.8.

A.1 Experimental details

A.1.1 Dataset and split details

All windows are formed after chronological splitting, so no window crosses a split boundary. Weather96 contains 21 channels and 52,696 rows; ETTh1 contains 7 channels and 17,419 rows; ETTm1 contains 7 channels and 69,680 rows; and the stock setting contains 6 channels and 3,685 rows. The exact train/validation/test window counts, missingness rate, and seed allocation are summarized in Table 1. ETTm1 uses a larger window pool because its source series is sampled more densely; the sequence length remains 96 in all settings. The three sensor datasets and the six-channel daily stock series are distributed with the ImagenI2R resources (Fadlon et al., 2025); we use the arrays shipped with the reproduction package unchanged and apply train-only normalization.

A.1.2 Missingness pattern

Table 4 records the mask used in every reported run: a single contiguous block per window along time, shared across channels, with a per-split missing fraction of 0.4995–0.4996. ETTh1 and Weather96 use a block of length 48; for ETTm1 and Stock the block position is drawn per window under the same target rate. No window crosses a split boundary.

Table 4: Missingness pattern per dataset. One contiguous block per window, shared across channels along time; the per-split missing fraction is measured on the frozen pairs.

Dataset	Block length	Train	Val	Test	Windows (tr/va/te)
Weather96	48	0.4995	0.4996	0.4996	4096/768/768
ETTh1	48	0.4995	0.4996	0.4996	4096/768/768
ETTm1	—	0.4995	0.4995	0.4996	8192/1024/1024
Stock	—	0.4995	0.4995	0.4996	2484/457/459

A.1.3 Completer architecture and accuracy

The deterministic stage is a mask-aware Transformer with model width 128, four layers, eight attention heads, feed-forward width 256, and dropout 0.1, trained for 40 epochs with batch size 128 and learning rate 3×10^{-4} on a mask distribution over rates $\{0.2, 0.3, 0.5, 0.7\}$ and patterns $\{$random, block$\}$ along time. Table 5 reports test-split missing-region MAE and RMSE of the frozen completer against two non-learned baselines: linear interpolation between the nearest observed values on each channel, and the per-window observation mean, averaged over pair seeds 0–2. The learned completer is substantially better than both baselines on ETTh1 and ETTm1 and roughly tied on Weather96, but on Stock it is an order of magnitude worse than linear interpolation and carries a large constant offset, so the Stock condition is the weakest completion of the four. This does not affect the paired interpretation of S_{time}, which compares two interfaces built from the same frozen completion, but it does bound what the Stock cell can be taken to show.

A.1.4 Per-channel completer accuracy

Table 6 breaks the completer accuracy of Table 5 down by channel. Errors are comparable across channels within a dataset except on Stock, where the five price-like channels all carry the constant offset noted above and the sixth channel is close to constant.

Table 5: Completer accuracy on missing coordinates versus non-learned baselines (test split, averaged over pair seeds 0–2). The learned completer clearly beats both baselines on ETTh1 and ETTm1 and is roughly tied on Weather96; on Stock it is an order of magnitude worse than linear interpolation and carries a large constant offset, so the Stock completion is the weakest of the four conditions.

	MAE			RMSE		
Dataset	Completer	Linear	Mean	Completer	Linear	Mean
Weather96	0.190	0.177	0.281	0.405	0.434	0.515
ETTh1	0.414	0.819	0.677	0.606	1.225	0.939
ETTm1	0.394	0.527	0.851	0.602	0.819	1.208
Stock	1.904	0.212	0.372	2.135	0.287	0.481

Table 6: Per-channel completer accuracy on the test split (missing coordinates only), averaged over pair seeds 0–2. Channels are indexed as stored in the frozen base arrays.

Dataset	Channel	MAE	RMSE
Weather96	0	0.136	0.184
	1	0.112	0.154
	2	0.114	0.156
	3	0.124	0.167
	4	0.183	0.253
	5	0.070	0.099
	6	0.076	0.108
	7	0.080	0.118
	8	0.078	0.111
	9	0.076	0.108
	10	0.124	0.170
	11	0.643	0.897
	12	0.613	0.864
	13	0.299	0.477
	14	0.132	0.579
	15	0.300	0.877
	16	0.183	0.273
	17	0.176	0.263
	18	0.186	0.281
	19	0.122	0.161
	20	0.154	0.215
ETTh1	0	0.535	0.748
	1	0.409	0.544
	2	0.497	0.732
	3	0.372	0.502
	4	0.628	0.849
	5	0.300	0.404
	6	0.155	0.208
ETTm1	0	0.510	0.748
	1	0.389	0.513
	2	0.463	0.721
	3	0.371	0.491
	4	0.641	0.890
	5	0.280	0.378
	6	0.101	0.137
Stock	0	2.158	2.277
	1	2.253	2.368
	2	2.241	2.355
	3	2.242	2.357
	4	2.188	2.305
	5	0.344	0.359

A.1.5 COMPLETER TRAINING

Table 7 and Figure 3 summarize the completer training. All four completers train for 40 epochs, reach their best validation MSE within the last three epochs, and are cheap relative to the downstream runs (23–95 seconds). The checkpoint with the lowest validation MSE is the one frozen into every pair used downstream.

Table 7: Completer training summary. "Best epoch" is the epoch with the lowest validation MSE, which is the checkpoint used for every downstream run.

Dataset	Epochs	Best val. MSE	Best epoch	Final train MSE	Time (s)
Weather96	40	0.0797	39	0.1612	44
ETTh1	40	0.1112	40	0.1444	95
ETTm1	40	0.2837	38	0.0951	65
Stock	40	0.1266	38	0.0836	23

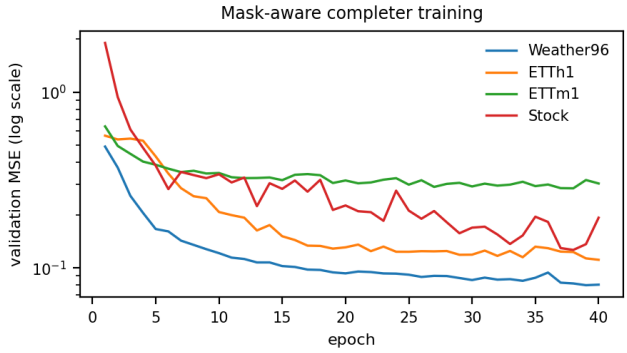

Figure 3: Validation MSE of the mask-aware completer by epoch for the four datasets (log scale).

A.1.6 RESOLVED GENERATOR CONFIGURATIONS

Table 8: Resolved downstream generator configurations shared across confirmatory runs.

Setting	Value
Training steps	5120
Optimizer	AdamW
Learning rate	2×10^{-4}
Weight decay	10^{-4}
Batch size	64
Diffusion steps	100
Evaluation samples	20
Missingness	block, approximately 50%
Sequence length	96
Conv1D denoiser	hidden 128, 4 residual blocks
Transformer denoiser	$d_{\text{model}} = 128$, 4 layers, 4 heads
Transformer feed-forward width	256
Checkpoint selection	lowest validation noise-prediction loss

The two denoisers use the same optimizer, diffusion schedule, training-step budget, batch size, and checkpoint rule. Their parameterizations are not equal-sized: the Conv1D arm uses four residual blocks, whereas the Transformer arm uses four layers with model width 128, four attention heads, and feed-forward width 256. We therefore interpret the architecture comparison as a robustness check across two generator families under a matched optimization budget, not as an equal-parameter scaling experiment. Runs were produced with Python 3.10.19 and PyTorch 2.6.0 with CUDA 12.4 on NVIDIA A40 and V100 GPUs; a 5120-step run takes roughly 3–9 minutes depending on backbone and dataset.

A.1.7 COMPUTE BUDGET AND RUN COUNT

Table 9 reports the parameter count and wall-clock training time of each downstream cell, averaged over the aligned arms and the generator seeds. All arms use the same 5120-step optimization budget, so the comparison is budget-matched rather than parameter-matched. Every number in the tables is produced by 394 downstream generator trainings: 259 primary, secondary and in-mask runs, 72 width-sweep runs, 48 five-seed robustness runs and 15 condition-sensitivity runs.

Table 9: Downstream compute budget. Parameters and wall-clock training time per 5120-step run, averaged over the aligned Full and Avg16 arms and all generator seeds of each cell.

Dataset	Denoiser	Parameters	Mean time (s)	Max time (s)	Runs
Weather96	Conv1D	443,541	325	581	10
Weather96	Transformer	577,173	267	274	6
ETTh1	Conv1D	434,567	265	322	6
ETTh1	Transformer	568,199	322	513	6
ETTm1	Conv1D	434,567	295	339	6
ETTm1	Transformer	568,199	486	490	6
Stock	Conv1D	433,926	195	250	6
Stock	Transformer	567,558	290	429	6

A.1.8 METRIC IMPLEMENTATION

For $K = 20$ stochastic predictions $z_1, \ldots, z_K$ at a missing coordinate with target x, the empirical CRPS is

$$\widehat{\mathrm{CRPS}} = \frac{1}{K}\sum_{k=1}^{K}|z_k - x| - \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{\ell=1}^{K}|z_k - z_\ell|.$$

Scores are averaged over missing coordinates within each window and then over windows. Missing-region MAE uses the sample mean as its point prediction. This coordinatewise score does not by itself evaluate all aspects of the joint distribution across time and channels. Pooled completion RMSE used for matching is a different aggregation from mean per-window completion RMSE.

A.2 RESIDUAL INTERVENTIONS AND CONTROLS

A.2.1 RESIDUAL-TRANSFORM DEFINITIONS

The aligned condition uses the masked learned residual. Time shuffle applies one permutation per sample, shared across channels. Phase randomization changes Fourier phases with amplitudes fixed before the mask is reapplied. Cross-sample transfer reassigns residual trajectories between samples. Every transformed arm is then remasked and rescaled using one RMS factor per split and pair seed. Neither the final value multiset nor the final spectrum is guaranteed to match the source after remasking. Secondary arms are descriptive controls rather than an ordered sequence of "structure destruction."

A.2.2 RESIDUAL INTERVENTION PSEUDOCODE

The following pseudocode describes the split-level operation used to construct the aligned and transformed completion pairs. `mask` is one on observed coordinates and $Q = 1 - \texttt{mask}$ is one on missing coordinates.

```
aligned = (x_comp - x_gt) * Q
if kind == "aligned":
    transformed = aligned
else:
    transformed = transform(aligned, kind, seed)
transformed = transformed * Q
source_rms = sqrt(mean(aligned[Q == 1] ** 2))
```

```
target_rms = sqrt(mean(transformed[Q == 1] ** 2))
if target_rms > 1e-8:
    transformed *= source_rms / target_rms
x_comp_new = x_gt + transformed
x_comp_new[mask == 1] = x_gt[mask == 1]
```

The implementation uses one temporal permutation per sample, shared across channels, for `time_shuffle`; phase randomization operates channel-wise in the Fourier domain; and cross-sample transfer permutes residual trajectories across samples. The pseudocode emphasizes that the mask is reapplied before the single RMS factor is computed.

A.2.3 POOLED RMS CONTROL AND ITS LIMITS

Table 10: Frozen-pair audit. Test-split pooled missing-coordinate RMSE is matched, whereas MAE is not. Values are averaged over available pair seeds 0–2 for each dataset.

Dataset	Aligned RMSE	Shuffled RMSE	Aligned MAE	Shuffled MAE
weather96	0.4047	0.4047	0.1896	0.1337
etth1	0.6060	0.6060	0.4138	0.2924
ettm1	0.6021	0.6021	0.3935	0.2779
stock	2.1345	2.1345	1.9043	1.3475

The audit uses pair seeds 0–2 and all three splits on four datasets: 36 comparisons. All paired mask hashes agree, and the largest recorded pooled RMSE difference is 2.3842×10^{-7}. Weather96's additional generator seeds are included in the primary result but are not part of this 36-pair audit. The code computes a single rescaling factor over each split, not one per window. Thus equality of pooled RMSE does not imply equality of mean window RMSE, MAE, or the missing-coordinate residual distribution.

For example, Stock seed 0 has aligned test pooled RMS 2.1267998. After temporal permutation and remasking, the raw RMS is 1.5026870; a factor of 1.4153312 restores the pooled RMS. This directly illustrates why the intervention must be described as RMS-matched, not value-preserving.

A.2.4 COMPLETE POOLED-RMS AUDIT

Table 11 lists all 36 audited pairs individually rather than in the per-dataset average of Table 10. The rescaling column shows that the shuffled arm is rescaled by a factor between roughly 1.3 and 1.5 on the sensor datasets, confirming that RMS matching is achieved by changing the distribution of the missing-region residual, not its ordering alone.

A.2.5 AN MAE-PRESERVING IN-MASK CONTROL

The primary time shuffle permutes the full window and then remasks, so it also changes the missing-region value multiset and lowers completion MAE: averaged over pair seeds 0–2, test-split completion MAE is 0.4138 aligned versus 0.2924 shuffled on ETTh1, and 1.9043 versus 1.3475 on Stock. To separate temporal order from this induced change, we add a post-hoc *in-mask* permutation that reorders residual values only among time indices that are missing in every channel, using one permutation per sample shared across channels. Values therefore move only between coordinates that are already missing, so pooled RMS, per-window RMSE, MAE and the missing-value multiset are unchanged up to float32 storage precision; only the temporal assignment changes. On Stock/Conv1D pair seed 0, for example, the aligned and in-mask completions share pooled RMS 2.1268 and MAE 1.8960, whereas the remasked time shuffle gives MAE 1.3406 at the same pooled RMS.

Table 12 reports the resulting shifts. Under this strictly MAE-preserving control the ETT cells remain positive, while the two Stock cells no longer reverse: their mean shifts are $+0.058$ and $+0.035$, with five of six seeds positive. Across the six dataset/backbone cells, 16 of 18 seed runs are positive and no cell has a negative mean. The Stock reversal in the primary contrast is therefore attributable to the MAE change induced by remasking rather than to temporal order alone. This arm was not preregistered, uses three generator seeds, and is reported as an exploratory diagnostic that bounds the interpretation of S_{time}; the preregistered primary contrast and its decision are unchanged.

Table 11: Complete pooled missing-coordinate RMS audit at the split level. Every row is one dataset, pair seed and split; the shuffled arm is rescaled by the last column so that the pooled RMS matches the aligned arm to numerical precision, while MAE is not matched.

Dataset	Seed	Split	Aligned RMSE	Shuffled RMSE	$\lvert\Delta\rvert$	Aligned MAE	Shuffled MAE
weather96	0	train	0.4388	0.4388	8.9e-08	0.2100	0.1506
weather96	0	val	0.4090	0.4090	6.0e-08	0.2131	0.1504
weather96	0	test	0.4056	0.4056	3.0e-08	0.1891	0.1329
weather96	1	train	0.4429	0.4429	6.0e-08	0.2101	0.1513
weather96	1	val	0.4170	0.4170	3.0e-08	0.2159	0.1526
weather96	1	test	0.4076	0.4076	6.0e-08	0.1906	0.1348
weather96	2	train	0.4383	0.4383	1.5e-07	0.2083	0.1469
weather96	2	val	0.4074	0.4074	0.0e+00	0.2118	0.1499
weather96	2	test	0.4010	0.4010	1.5e-07	0.1890	0.1336
etth1	0	train	0.4544	0.4544	8.9e-08	0.3156	0.2232
etth1	0	val	0.5376	0.5376	0.0e+00	0.3513	0.2488
etth1	0	test	0.6070	0.6070	6.0e-08	0.4152	0.2931
etth1	1	train	0.4545	0.4545	1.2e-07	0.3154	0.2229
etth1	1	val	0.5392	0.5392	0.0e+00	0.3520	0.2493
etth1	1	test	0.6007	0.6007	6.0e-08	0.4108	0.2907
etth1	2	train	0.4536	0.4536	2.1e-07	0.3150	0.2224
etth1	2	val	0.5356	0.5356	6.0e-08	0.3490	0.2459
etth1	2	test	0.6102	0.6102	6.0e-08	0.4154	0.2933
ettm1	0	train	0.3991	0.3991	8.9e-08	0.2760	0.1953
ettm1	0	val	0.5344	0.5344	6.0e-08	0.3390	0.2391
ettm1	0	test	0.6068	0.6068	6.0e-08	0.3956	0.2791
ettm1	1	train	0.3974	0.3974	2.1e-07	0.2752	0.1945
ettm1	1	val	0.5425	0.5425	6.0e-08	0.3381	0.2389
ettm1	1	test	0.6049	0.6049	6.0e-08	0.3946	0.2788
ettm1	2	train	0.4034	0.4034	1.2e-07	0.2786	0.1967
ettm1	2	val	0.5298	0.5298	6.0e-08	0.3355	0.2373
ettm1	2	test	0.5946	0.5946	6.0e-08	0.3903	0.2758
stock	0	train	0.3032	0.3032	6.0e-08	0.1508	0.1066
stock	0	val	0.3427	0.3427	0.0e+00	0.2753	0.1951
stock	0	test	2.1268	2.1268	0.0e+00	1.8960	1.3406
stock	1	train	0.3049	0.3049	1.5e-07	0.1525	0.1081
stock	1	val	0.3425	0.3425	3.0e-08	0.2758	0.1953
stock	1	test	2.1447	2.1447	2.4e-07	1.9130	1.3516
stock	2	train	0.3052	0.3052	3.0e-08	0.1529	0.1081
stock	2	val	0.3406	0.3406	0.0e+00	0.2716	0.1914
stock	2	test	2.1321	2.1321	2.4e-07	1.9039	1.3503

Table 12: Exploratory MAE-preserving in-mask control. $S_X = \Delta_{aligned} - \Delta_X$ with $\Delta_X = \mathrm{CRPS}_{Full}(X) - \mathrm{CRPS}_{Avg16}(X)$. The in-mask permutation leaves pooled missing-region RMS, per-window RMSE, MAE, and the missing-value multiset unchanged up to float32 storage precision, so it isolates temporal order from the MAE change that the primary remasked time shuffle induces. Pos. is the number of positive seeds. This arm was run after the primary result and is diagnostic, not preregistered.

Dataset	Denoiser	S_{time}	Pos.	S_{inmask}	Pos.
ETTh1	Conv1D	+0.150	3	+0.087	3
ETTh1	Transformer	+0.152	3	+0.063	3
ETTm1	Conv1D	+0.042	3	+0.004	2
ETTm1	Transformer	+0.180	3	+0.026	3
Stock	Conv1D	-0.213	0	+0.058	3
Stock	Transformer	-0.436	0	+0.035	2

A.2.6 SECONDARY CONTROLS

Table 13: Secondary structure contrasts: $\Delta_{aligned} - \Delta_{control}$. These interventions are descriptive and are not treated as interchangeable with the primary time-shuffle contrast.

Dataset	Denoiser	vs. phase random	vs. cross sample
Weather96	Conv1D	-0.005	+0.005
Weather96	Transformer	-0.003	+0.021
ETTh1	Conv1D	+0.103	+0.034
ETTh1	Transformer	+0.063	+0.035
ETTm1	Conv1D	+0.009	-0.020
ETTm1	Transformer	+0.016	-0.030
Stock	Conv1D	-0.016	+0.133
Stock	Transformer	+0.157	+0.215

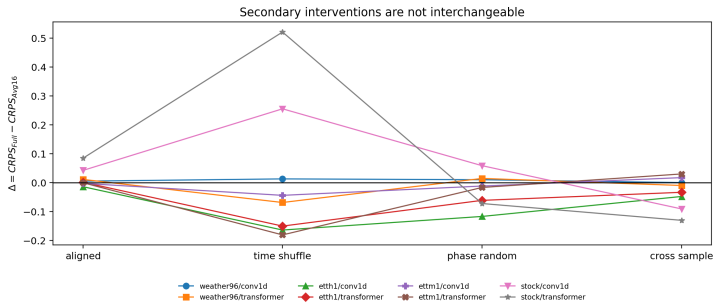

Figure 4: Absolute Full-versus-Avg16 gap Δ under four residual interventions. Positive values favor Avg16; nonparallel responses show the interventions are not interchangeable.

Table 14 reports every seed used in the phase-random and cross-sample controls. The mixed signs are why these arms are not collapsed into a single scalar definition of structure destruction.

Table 14: Seed-wise secondary contrasts $\Delta_{aligned} - \Delta_{control}$. These values are descriptive controls; they are not pooled into the primary claim.

Dataset	Denoiser	Control	Seed-wise values	Mean	Pos.
Weather96	Conv1D	phase random	-0.014,-0.006,+0.003,-0.004,-0.003	-0.005	1/5
Weather96	Conv1D	cross sample	-0.005,+0.010,+0.007,+0.007,+0.004	+0.005	4/5
Weather96	Transformer	phase random	-0.006,+0.008,-0.012	-0.003	1/3
Weather96	Transformer	cross sample	+0.017,+0.022,+0.022	+0.021	3/3
ETTh1	Conv1D	phase random	+0.099,+0.082,+0.126	+0.103	3/3
ETTh1	Conv1D	cross sample	+0.020,+0.025,+0.054	+0.034	3/3
ETTh1	Transformer	phase random	+0.064,+0.087,+0.037	+0.063	3/3
ETTh1	Transformer	cross sample	+0.037,+0.040,+0.026	+0.035	3/3
ETTm1	Conv1D	phase random	+0.005,+0.018,+0.003	+0.009	3/3
ETTm1	Conv1D	cross sample	-0.000,-0.022,-0.035	-0.020	0/3
ETTm1	Transformer	phase random	+0.018,+0.015,+0.014	+0.016	3/3
ETTm1	Transformer	cross sample	-0.031,-0.028,-0.030	-0.030	0/3
Stock	Conv1D	phase random	-0.017,-0.028,-0.003	-0.016	0/3
Stock	Conv1D	cross sample	+0.114,+0.153,+0.131	+0.133	3/3
Stock	Transformer	phase random	+0.347,-0.041,+0.164	+0.157	2/3
Stock	Transformer	cross sample	+0.285,+0.150,+0.209	+0.215	3/3

A.3 A COMPLETE WORKED EXAMPLE

We walk through one cell end to end so that every symbol in Section 3 carries a concrete value. The cell is ETTh1 with the Conv1D denoiser and pair seed 0, evaluated on the test split unless stated otherwise.

Step 1: evidence and mask. The base array supplies X and the frozen pair supplies the observation mask M. The split has 768 test windows, a mean missing fraction of 0.4996, and one contiguous missing block per window shared across channels. Observed coordinates are never modified anywhere in the pipeline.

Step 2: deterministic completion. The completer produces $\hat{X}_c$ and the residual $E = (1 - M) \odot (\hat{X}_c - X)$. Pooled over the missing coordinates of this split, $R_Q(E) = 0.606999$ and the missing-region MAE is 0.415152 (Table 10).

Step 3: residual intervention. The aligned arm uses E directly. The time-shuffle arm permutes the full-window residual, reapplies the mask, and rescales once per split: the permuted and remasked residual has RMS 0.429623, and multiplying by 1.412864 restores 0.606999. Pooled completion RMS is therefore identical across the two arms while MAE is not, falling from 0.415152 to 0.293108 (-29%). Matching is pooled, not per window: for test window 0 (sample id 0, start index 14807) the aligned and shuffled residuals have window RMS 0.510 and 0.423.

Step 4: interfaces. FULL passes $\hat{X}_c$ unchanged. AVG16 passes $P_{16}(\hat{X}_c)$, the width-16 stride-one moving average with replicated boundary padding; both keep length 96 and both receive the same X_{obs} and M.

Step 5: downstream scores. Training the conditional diffusion generator separately for each of the four interface/intervention combinations and drawing $K = 20$ samples per window gives the per-window CRPS and completion RMSE in Table 15, and the predictive bands shown in Figure 2 of the main text. Within each intervention the two interfaces share the same completion RMSE (0.7275 and 0.7452) yet their CRPS differs (0.6337 vs 0.6217, and 0.4109 vs 0.5644). For this window $\Delta_{\text{aligned}} = +0.0120$ and $\Delta_{\text{shuffle}} = -0.1535$, so $S_{\text{time}} = +0.1655$; the aggregate over seeds is $+0.150$ (Table 2). This first saved window is used as an illustration, not as a new aggregate estimate, and shuffled/Full beating shuffled/Avg16 here is compatible with the positive aggregate shift because S_{time} compares interface responses rather than asserting a universal winner.

Table 15: Metrics for the same illustrative ETTh1 test window (Conv1D, seed 0, sample id 0). Lower is better. This is a qualitative case, not a new aggregate estimate.

Condition	CRPS	MAE	Completion RMSE
aligned / full	0.6337	0.7716	0.7275
aligned / avg16	0.6217	0.7649	0.7275
shuffle / full	0.4109	0.5548	0.7452
shuffle / avg16	0.5644	0.6915	0.7452

A.4 ADDITIONAL RESULTS

A.4.1 SEED-WISE PRIMARY SHIFTS

Table 16 lists the seed-level primary shifts. These values are the basis for the "positive seeds" counts in the main text.

Table 16: Seed-wise $S_{time} = \Delta_{aligned} - \Delta_{time\text{-}shuffle}$.

Dataset	Denoiser	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4
Weather96	Conv1D	-0.020	+0.016	-0.012	-0.022	-0.002
Weather96	Transformer	+0.108	+0.079	+0.053	–	–
ETTh1	Conv1D	+0.132	+0.175	+0.143	–	–
ETTh1	Transformer	+0.168	+0.156	+0.131	–	–
ETTm1	Conv1D	+0.061	+0.025	+0.038	–	–
ETTm1	Transformer	+0.212	+0.160	+0.170	–	–
Stock	Conv1D	-0.267	-0.315	-0.057	–	–
Stock	Transformer	-0.614	-0.609	-0.085	–	–

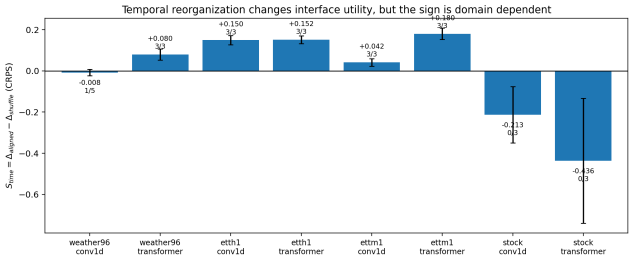

Figure 5: Primary temporal-organization shift S_{time}. Positive means that the aligned learned residual makes Avg16 relatively more favorable than time shuffling; negative means the opposite. Error bars show across-seed standard deviation.

A.4.2 Five-seed robustness of the primary contrast

The primary cells use pair seeds 0–2. To check that the reported signs do not depend on that draw, we build two further pairs per dataset (seeds 3–4) with the same time-block mask and completer procedure, and retrain every interface and transform arm on them. Table 17 reports S over the five seeds. The ETT family is positive in all 20 seed–cell runs and the Stock domain is negative in all 10. Per-cell means move by at most 0.01 relative to the three-seed values except for Stock/Transformer, where the two added seeds are weaker but still negative, and the across-seed standard deviation remains larger for Stock than for the ETT cells. The sign pattern of the primary contrast is therefore stable under two additional draws of the mask and the completion.

Table 17: Five-seed robustness check of the primary contrast. Seeds 0–2 are the primary pair/generator seeds; seeds 3–4 are two further pairs built for this check with the same mask and completion procedure. Each seed gives $S = \Delta_{aligned} - \Delta_{time\text{-}shuffle}$.

Cell	Seeds	Mean S	SD	Pos.
ETTh1 / Conv1D	5	+0.154	0.018	5
ETTh1 / Transformer	5	+0.155	0.021	5
ETTm1 / Conv1D	5	+0.041	0.016	5
ETTm1 / Transformer	5	+0.179	0.020	5
Stock / Conv1D	5	-0.201	0.106	0
Stock / Transformer	5	-0.341	0.286	0

A.4.3 Seed-wise absolute CRPS

Table 3 in the main text exposes the quantities from which every primary Δ is computed, including the per-seed aligned and time-shuffled absolute CRPS with the resulting decisions. It distinguishes a relative shift from an absolute winner: a positive S_{time} can occur while Full remains better in both the aligned and the shuffled condition.

A.4.4 Aligned and shuffled absolute gaps

A.4.5 Per-channel downstream accuracy

Table 19 reports the downstream predictive MAE and RMSE of the aligned Full arm aggregated over channels, together with the best and worst channel. The spread between the best and worst channel is modest on the sensor datasets but larger on Stock, consistent with the per-channel completion offsets noted in Table 6.

A.4.6 Coarsening-scale sweep for both denoisers

Table 20 reports the width sweep for both denoisers on three datasets. The Stock trend is robust across backbones and stronger under the Transformer, where Δ rises from -0.001 at $w = 2$ to $+0.198$ at $w = 32$ with all three seeds positive from $w = 8$ onward; ETTm1 stays within 0.001 of zero

Table 18: Absolute interface gap $\Delta = \mathrm{CRPS}_{Full} - \mathrm{CRPS}_{Avg16}$ for the aligned and time-shuffled residuals. Positive favors Avg16.

Dataset	Denoiser	Aligned Δ	Time-shuffle Δ
Weather96	Conv1D	+0.005	+0.013
Weather96	Transformer	+0.011	-0.069
ETTh1	Conv1D	-0.014	-0.164
ETTh1	Transformer	+0.001	-0.150
ETTm1	Conv1D	-0.003	-0.044
ETTm1	Transformer	-0.001	-0.181
Stock	Conv1D	+0.042	+0.255
Stock	Transformer	+0.084	+0.521

Table 19: Downstream predictive accuracy by cell, averaged over generator seeds and the aligned Full arm. Channel-mean MAE and RMSE are reported together with the best and worst channel, so that the spread across channels is visible without a full per-channel listing.

Dataset	Denoiser	Mean MAE	Worst MAE	Best MAE	Mean RMSE
Weather96	Conv1D	0.528	0.937	0.101	0.644
Weather96	Transformer	0.537	0.963	0.100	0.653
ETTh1	Conv1D	0.624	0.831	0.324	0.835
ETTh1	Transformer	0.644	0.885	0.340	0.858
ETTm1	Conv1D	0.633	0.901	0.265	0.838
ETTm1	Transformer	0.638	0.907	0.268	0.843
Stock	Conv1D	4.401	5.319	0.484	4.430
Stock	Transformer	4.039	4.826	0.552	4.059

under both; and the ETTh1 trend is backbone dependent, turning negative under Conv1D but slightly positive under the Transformer. The scale at which coarsening helps therefore depends on both the domain and the denoiser, and no single smoothing width is preferable across settings.

Table 20: Full-versus-coarse gap by smoothing width for both denoisers. $\Delta(w) = \mathrm{CRPS}_{Full} - \mathrm{CRPS}_{Avgw}$; positive favors the coarse interface. Means over three seeds; the $w = 16$ column is the primary aligned gap in each row.

Dataset	Denoiser	w=2	w=4	w=8	w=16	w=32
ETTh1	Conv1D	0.000	0.000	+0.002	-0.014	-0.060
ETTh1	Transformer	-0.007	-0.006	-0.014	+0.001	+0.009
ETTm1	Conv1D	-0.001	-0.001	-0.002	-0.003	-0.004
ETTm1	Transformer	-0.000	-0.000	-0.001	-0.001	+0.001
Stock	Conv1D	+0.017	+0.027	+0.030	+0.042	+0.058
Stock	Transformer	-0.001	+0.015	+0.038	+0.084	+0.198

A.4.7 Stability to the number of stochastic samples

All CRPS values in the paper use $K = 20$ samples per window. Because the empirical CRPS estimator is biased for small K, we re-evaluated four cells from their frozen checkpoints with a single shared draw of 100 samples per window and computed $\Delta(K)$ from the first K of them. Table 21 shows that the absolute CRPS falls with K, as expected, while the Full-versus-Avg16 gap moves by at most 0.003 between $K = 5$ and $K = 100$. The reported interface comparison is therefore not an artifact of the estimator's sample count. Because this is a fresh evaluation of the frozen checkpoints, its $K = 20$ values can differ slightly from Table 2.

A.4.8 Residual-structure correlates of the interface gap

Descriptors of the frozen completion can be computed without any generator, so it is natural to ask which of them track the interface gap. Table 22 correlates, within each dataset, five residual

Table 21: Stability of the interface gap to the number of stochastic samples K. Each cell reports $\Delta(K) = \mathrm{CRPS}_{Full}(K) - \mathrm{CRPS}_{Avg16}(K)$, computed from one shared draw of 100 samples per window so that only K varies. The absolute CRPS falls with K, as expected for the empirical CRPS estimator, while the interface gap is essentially unchanged. Values come from a fresh evaluation of the frozen checkpoints and can therefore differ slightly from Table 2.

Cell	K=5	K=10	K=20	K=50	K=100
ETTh1 / Conv1D	-0.0149	-0.0144	-0.0142	-0.0141	-0.0141
ETTh1 / Transformer	+0.0013	+0.0013	+0.0012	+0.0011	+0.0012
Stock / Conv1D	+0.0404	+0.0417	+0.0424	+0.0428	+0.0429
Stock / Transformer	+0.0819	+0.0841	+0.0852	+0.0859	+0.0861

descriptors with $\Delta = \mathrm{CRPS}_{\mathrm{FULL}} - \mathrm{CRPS}_{\mathrm{AVG16}}$ across the aligned, time-shuffle, phase-random and cross-sample pairs and pair seeds 0–2 (12 points per dataset). Several descriptors change sign between domains: total variation over RMS correlates -0.59 and -0.63 on ETTh1 and ETTm1 but $+0.61$ on Stock, and lag-1 autocorrelation correlates $+0.59$ and $+0.44$ on the ETT family but -0.59 on Stock. The relationship between residual structure and interface utility is therefore itself domain dependent, which is consistent with the primary result but not a validated selection rule: these are within-dataset correlations on 12 pairs, further reduced by the shared pair seeds, and we do not use any descriptor to choose an interface.

Table 22: Exploratory within-dataset Spearman correlations between frozen-completion residual descriptors and the interface gap $\Delta = \mathrm{CRPS}_{Full} - \mathrm{CRPS}_{Avg16}$, computed across the aligned, time-shuffle, phase-random and cross-sample pairs and pair seeds 0–2 (12 points per dataset). No descriptor is used to select an interface.

Descriptor	Weather96	ETTh1	ETTm1	Stock
lag-1 autocorrelation	-0.21	+0.59	+0.44	-0.59
lag-4 autocorrelation	-0.16	+0.59	+0.33	-0.61
sign-change rate	+0.03	-0.65	-0.63	+0.45
total variation / RMS	+0.15	-0.59	-0.63	+0.61
high-frequency energy ratio	+0.19	-0.50	-0.28	+0.67

A.4.9 Stock condition-sensitivity check

The separately trained stock sanity arms test whether useful completion information improves downstream scores. Mean Conv1D test CRPS is 3.263 for oracle Avg16, 3.938 for Avg16, 3.976 for Full, 4.227 for shuffled Avg16, and 4.251 for blank, consistent with benefit from sample-specific information. The target-derived oracle is not deployable, and all arms are separately trained, so this is not a fixed-checkpoint intervention and does not explain the stock shift.

Table 23: Stock/Conv1D condition-sensitivity sanity check (mean test CRPS over three seeds). Lower is better.

Condition	Test CRPS
Oracle Avg16	3.263
Avg16	3.938
Full	3.976
Shuffled Avg16	4.227
Blank	4.251

A.4.10 Interface families implemented but not reported

The pair-construction and interface code also implements families that this paper does not evaluate: spectral low-pass conditions (`spectral`p), high-pass residual conditions (`detail_avg`k),

reliability-gated adaptive mixtures of the dense and coarse conditions, and circular-shift residual interventions. We ran no generator sweeps for these arms and therefore make no claim about them; we list them only so that the boundary of the reported interface space is explicit. In particular, the width sweep of Table 20 should be read as a one-parameter slice through an interface space larger than what is reported here.

A.5 Reproducibility and claim discipline

A.5.1 Preregistration and interpretation status

The ETTm1/Conv1D contrast was preregistered before its downstream generator runs. The stock/Conv1D independent-domain contrast was also preregistered before any stock generator run and failed in the opposite direction. The stock/Transformer experiment reused the same frozen pair files but was initiated after the Conv1D result was known; it is therefore treated as descriptive diagnostic evidence rather than a fresh confirmatory test. This distinction is preserved in the supplied reproducibility manifest.

A.5.2 Preregistered claims and explicit non-claims

Table 24 restates the frozen contrast of each preregistered cell with its observed outcome. Every cell predicted $S_{time} > 0$; the two ETTm1 cells supported it and the two Stock cells reversed it. The preregistration also froze what the experiment was *not* allowed to claim:

- No preregistered requirement that Avg16 absolutely beats Full.
- No claim that completion RMSE predicts the sign of the Full-versus-Coarse gap.
- No test-time selection of interface or residual transform.

Table 24: Preregistered contrasts as frozen before the corresponding generator runs. The primary claim is the same in every cell: with CRPS and $\Delta = \mathrm{CRPS}_{Full} - \mathrm{CRPS}_{Avg16}$, the aligned-minus-time-shuffle shift is expected to be positive. The last column is the observed outcome.

Cell	Backbone	Created (UTC)	Primary contrast	Outcome
ETTm1 / Conv1D	conv1d	2026-09-18 10:05	$S_{time} > 0$	$+0.042$, supported
ETTm1 / Transformer	transformer	2026-09-18 11:01	$S_{time} > 0$	$+0.180$, supported
Stock / Conv1D	conv1d	2026-09-19 03:21	$S_{time} > 0$	-0.213, reversed
Stock / Transformer	transformer	2026-09-19 07:55	$S_{time} > 0$	-0.436, reversed

A.5.3 Preregistration timeline

Table 25: Preregistration and interpretation timeline. Timestamps are UTC and hashes refer to the frozen preregistration artifacts.

Dataset	Denoiser	Status	Created (UTC)	Interpretation
ETTm1	Conv1D	prospective confirmation	2026-09-18 10:05	Primary contrast preregistered before generator runs; positive shift observed.
ETTm1	Transformer	prospective confirmation	2026-09-18 11:01	Same frozen ETTm1 source pairs; primary contrast preregistered before this arm.
Stock	Conv1D	prospective boundary test	2026-09-19 03:21	Independent-domain test preregistered before any stock generator run; positive-shift hypothesis rejected.
Stock	Transformer	post-result diagnostic	2026-09-19 07:55	Same frozen stock pairs, started after the Conv1D result; not an independent confirmation.

The ETTm1 and stock/Conv1D entries are prospective with respect to their corresponding generator runs. The stock/Transformer timestamp is retained for transparency, but its result is interpreted diagnostically because the stock/Conv1D result was already available.

A.5.4 Reproduction checklist

An anonymous supplement should contain:

1. formal base arrays and JSON metadata for all four datasets;
2. immutable source pairs and derived residual pairs for all four intervention arms;
3. resolved completer and generator configurations, checkpoints or download instructions, and environment reports;
4. run manifests, seed lists, per-sample CSV files, per-channel metrics, and pair-audit outputs;
5. preregistration JSON files and their recorded hashes;
6. scripts for pair construction, generator training, evaluation, table generation, and the RMS audit.

The supplement should remove absolute paths, machine names, usernames, and repository metadata that could compromise double-blind review.

A.5.5 ETHICS STATEMENT

This work uses publicly released, non-personal time-series datasets distributed with the ImagenI2R resources; the six-channel stock series consists of aggregated daily market variables rather than individual-level records. No human subjects or user data were involved. The reported intervention needs ground-truth values at missing coordinates in order to form and RMS-match residuals, so it is a controlled measurement instrument rather than an inference-time procedure on naturally unobserved values, and it is not proposed as a method for recovering private or sensitive values.

A.5.6 SUPPLEMENTARY CODE

An anonymous supplement accompanies the submission. It contains the source for pair construction, completer and generator training, evaluation, the in-mask control, the width sweep, the five-seed robustness check, the descriptor analysis and the completer baselines, together with the analysis outputs from which every table in this appendix is generated. Absolute paths and author-identifying strings have been removed.

A.5.7 CLAIM DISCIPLINE

The paper supports the following statements: (i) pointwise completion RMSE is insufficient to characterize downstream interface utility; (ii) temporal reorganization of learned completion residuals can strongly change the Full-versus-Coarse comparison; (iii) the sign of that effect is dataset dependent; and (iv) there is no universal Full or Avg16 winner. The paper does not claim that coarse conditioning is generally better, that larger completion errors require coarser interfaces, that all structure-breaking transformations behave alike, or that current residual descriptors predict the correct routing policy.

A.5.8 EVIDENCE FILES

The manuscript source includes the primary, secondary, absolute-gap, and seed-wise CSV tables. The numerical audit produces `risk_decomposition.csv` and `rms_control_audit.csv`, and the in-mask, descriptor, sweep, seed-robustness and baseline tables are produced by `setup_inmask_v06.py`, `analyze_sweep_v06.py`, `analyze_seed34_v06.py`, `analyze_fixed_interface_descriptors.py`, and `completer_baselines_v06.py` respectively; their source locations refer to the frozen experiment directories. The audit verifies numerical consistency, not completeness of an independently reproducible training release. A release of the full training artifacts remains distinct from the paper's evidence tables.