Equal Accuracy, Unequal Cost: Channel Dependence in PatchTST under Controlled Coupling

Anonymous authors
Paper under double-blind review

Abstract

Channel-independent (CI) and channel-dependent (CD) variants of PatchTST are typically compared on observational benchmarks where correlation strength, dimensionality, and temporal dynamics all vary at once, leaving no clean way to attribute either variant win to any single cause. We isolate one factor at a time. A factorial AR(1) experiment varies contemporaneous correlation ρ and variate count C while Granger non-causality holds across channels by construction, and a leader-follower VAR(1) experiment introduces lag-1 coupling of controllable strength γ. On the synthetic grid we detect no consistent CD advantage at the 1% level: the grand-mean CD$-$CI difference is $+0.0013$ MSE (95% CI $[-0.0002, +0.0028]$) over five seeds, with a pooled within-cell detection half-width of 0.0063 MSE, so any global CD gain (averaged across the nine AR(1) cells) above roughly 0.63% of the CI mean would have moved the interval off zero, and none did. Under the original early-stopping protocol an apparent CD deficit emerges under lagged coupling and grows with γ, but an instrumented single-trajectory diagnostic at the leader-follower $\gamma = 0.6$ cell shows this deficit to be largely a selection artefact rather than an architectural failure: CD reaches a sharper, earlier validation minimum than CI and degrades past it, and selecting each model at its own validation minimum erases the gap (CD$-$CI of -0.0031 MSE, 95% CI $[-0.0109, +0.0046]$). A cross-variate prediction head ties CI at every γ and removes the coupling-dependent widening, locating that widening in the shared head rather than the encoder. Matching the gradient-update budget across modes removes the deficit, which rules out the per-epoch step-count asymmetry as its source, and a block-covariance family carries the same no-advantage finding beyond compound-symmetry correlation. Across every tested synthetic cell (AR(1), leader-follower, and block-covariance) the absolute CD$-$CI difference stays inside a pre-registered 1% relative-MSE band in both directions. What does not equalise is cost: CD requires smaller feasible batches and far more gradient steps per epoch, its batch size collapses to one at $C = 84$, and at $C = 321$ (ECL) the flattened-token formulation does not finish under a single-T4 budget while CI trains stably. On the observational ETTh1 ($C = 7$), a matched-budget comparison still gives CI a modest but reliable edge (pooled CD$-$CI of $+0.0147$ MSE, 95% CI $[+0.0047, +0.0246]$, significant at horizons 96 and 720). On the constructed AR(1), leader-follower, and block-covariance processes, neither mode shows an accuracy gap that clears our 1% bar once each is selected fairly, and CD stays substantially more expensive to train. CI is the better default on accuracy per unit compute. All findings are scoped to PatchTST's flattened-token CD formulation, the generative families studied, and a single-T4 training budget.

1 Introduction

Whether channel-independent (CI) or channel-dependent (CD) architectures forecast better has been judged almost entirely on observational benchmarks, where many properties of the data move together across datasets. A win there names the stronger model on those series without isolating its cause. Here we control the generative process: synthetic AR(1) and leader-follower VAR(1) families that vary instantaneous

correlation, lagged cross-channel dependence, and variate count independently, with ETTh1 and ECL as real-data context rather than primary evidence.

The first experiment generates data from an AR(1) process with diagonal transition and compound-symmetry covariance, so the only cross-channel structure is instantaneous correlation. The second introduces lagged coupling through a leader-follower VAR(1) process where coupling strength γ is a controllable parameter. No point in the tested coupling-structure space yielded a consistent CD advantage. Section 2 describes the experimental design, Section 3 presents the AR(1) grid and real-data results, Section 4 covers the leader-follower experiments, Section 5 presents the fair-protocol robustness controls, and the remaining sections discuss limitations, related work, and conclusions.

1.1 Research Questions and Scope

CI architectures like PatchTST have become strong baselines for multivariate time series forecasting despite ignoring cross-channel interactions in their attention layers (Nie et al., 2023). At the same time, iTransformer, LIFT, and Abdelmalak et al. each report settings where modelling channel dependence explicitly yields clear gains (Liu et al., 2024; Zhao & Shen, 2024; Abdelmalak et al., 2026). None of these results are directly comparable: the architectures, datasets, and notions of cross-channel structure all differ across papers. There is no account of which property of the data tips the balance.

We focus on one specific part of that question: multivariate processes in which channels share *instantaneous* correlation but have no lagged cross-channel dependence in the dynamics. Concretely, we use a vector AR(1) process with a diagonal transition matrix and compound-symmetry covariance, where ρ controls contemporaneous correlation while Granger non-causality holds across all channel pairs by construction. Two questions frame the work:

> **RQ1.** In a setting where channels share purely instantaneous correlation, even at high ρ and high dimensionality, and no lagged cross-channel dependence exists by construction, does a channel-dependent PatchTST variant achieve meaningfully better forecast accuracy than its channel-independent counterpart?
>
> **RQ2.** When lagged cross-channel coupling is introduced at controlled strength γ via a leader-follower VAR(1) process, does CD outperform CI consistently as γ grows, and does this depend on the patch resolution used?

RQ1 is answered by a $C \times \rho \times$ mode factorial experiment on AR(1) data where Granger non-causality holds by construction. RQ2 is answered by two companion experiments: a γ sweep at fixed patch size $P = 16$, and a joint sweep over $\gamma \in \{0.0, 0.3, 0.6, 0.9\}$ and $P \in \{2, 4, 8, 16\}$ to test whether finer patch resolution concentrates lag-1 coupling within individual patches and thereby gives CD's attention a detectable signal. Within these regimes both RQ1 and RQ2 come back null at our 1% practical threshold: small differences exist, but none clears the bar. A set of fair-protocol controls (Section 5) then shows that the apparent CD deficit under lagged coupling is largely an artefact of training protocol and selection rather than an architectural limit.

Our conclusions are scoped accordingly: we make no broad claim that CD is unnecessary. We show that purely instantaneous correlation, regardless of strength or variate count, is not sufficient for CD to gain an advantage over CI under PatchTST's architecture, and that this null result extends to lag-1 coupling at the strengths and patch resolutions we tested.

2 Experimental Setup

2.1 Synthetic Data

We generate multivariate time series from a vector AR(1) process:

$$x_t = \phi\, x_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0,\, \Sigma), \tag{1}$$

where $\phi = 0.8$ and Σ has compound-symmetry structure: $\Sigma_{ii} = 1$ and $\Sigma_{ij} = \rho$ for $i \neq j$. The parameter ρ directly controls pairwise contemporaneous correlation while the transition matrix remains diagonal. We discard 1000 burn-in steps before recording observations.

For each configuration we generate 14,400 timesteps and split them 60/20/20 into train, validation, and test sets. All splits are normalised with per-channel z-score statistics fit on the training set only. We vary three factors independently: number of channels $C \in \{7, 21, 84\}$, correlation strength $\rho \in \{0.1, 0.5, 0.9\}$, and model mode $\in \{\text{CI}, \text{CD}, \text{DLinear}\}$, with five random seeds $\{42, 123, 456, 789, 1011\}$ per cell, yielding 135 runs in total. We report the empirical pairwise Pearson correlation of the training data alongside the nominal ρ to confirm intended data-generating process behaviour.

Dynamical structure. Because the transition operator $\phi\mathbf{I}$ is diagonal, each variate's next value depends only on its own past. Granger non-causality holds exactly across all channel pairs: all cross-variate dependence arises from the contemporaneous covariance Σ, not from the dynamics. The stationary lagged cross-covariance between channels $i \neq j$ at any lag $k > 0$ is zero, because $A^k = (\phi\mathbf{I})^k = \phi^k\mathbf{I}$ has zero off-diagonal entries at every power. Contemporaneous correlation in Σ does not provide a predictive signal because innovations are unobserved at prediction time. The leader-follower experiments in Section 4 introduce the complementary case where the transition operator structurally links one variate's future to another's past.

Covariance design. Compound-symmetry assigns the same pairwise correlation ρ to every channel pair, giving CD no differential block structure or correlation hierarchy to exploit. This is the configuration most hostile to CD under instantaneous-only coupling: if CD cannot gain an advantage here, contemporaneous correlation alone will not be sufficient at any other covariance structure within this family.

Prediction horizon. All synthetic AR(1) runs use pred_len = 96, matching the shortest horizon in the ETTh1 experiments.

2.2 Real-World Datasets

ETTh1. Hourly measurements of 7 variates (oil temperature and six power-load features) from a Chinese electrical transformer. We use the standard train/val/test split of 8640/2880/2880 timesteps. The mean absolute off-diagonal Pearson correlation at lag 0 across the training split is 0.311 (median 0.233), indicating moderate contemporaneous cross-variate correlation.

ECL (Electricity Consuming Load). Hourly electricity consumption for 321 clients over 26,304 timesteps. (The iTransformer paper (Liu et al., 2024) reports 26,352 rows; the difference reflects a minor dataset version discrepancy and does not affect the experimental protocol.) We apply proportional 60/20/20 splitting following Liu et al., yielding 15,811 train / 5,246 validation / 5,247 test rows. The mean absolute off-diagonal Pearson correlation at lag 0 is 0.503 (median 0.483), substantially higher than ETTh1. All data are normalised with per-channel z-score statistics fit on the training set only.

2.3 Models

PatchTST is the backbone for both CI and CD modes because the two modes differ in exactly one thing: whether self-attention operates within each variate's patch sequence independently or across all variates' patch sequences jointly. Patch extraction, linear projection, Transformer encoder, training procedure, and hyperparameters are otherwise shared, making this the most direct available test of the CI/CD distinction without entangling it with other architectural choices. iTransformer would confound the comparison with its inverted embedding design; Crossformer with its two-stage routing attention. All findings are conditioned on PatchTST's patch-based representation and the lookback window and patch granularity used.

We implement PatchTST (Nie et al., 2023) from scratch in PyTorch, supporting both modes.

Channel-Independent (CI). Each variate is processed separately. An input of shape (B, L, C) is reshaped to $(B \cdot C, L)$, split into patches of length P with stride S, and passed through a shared Transformer encoder. The encoder sees N patches per variate.

Channel-Dependent (CD). All variates are processed jointly. After patching, tokens from all variates are concatenated along the sequence dimension before the encoder, giving an input of shape $(B, C \cdot N, D)$. Self-attention operates over all $C \cdot N$ tokens simultaneously, allowing the model to attend across variates. After encoding, the output is split back into C per-variate representations before a shared linear head projects each variate to the prediction horizon. The encoder mixes cross-variate information; the head does not.

For the synthetic experiments and ETTh1 we use $L = 512$, $P = 16$, $S = 8$, giving

$$N = \left\lfloor \frac{L - P}{S} \right\rfloor + 1 = \left\lfloor \frac{512 - 16}{8} \right\rfloor + 1 = 63 \tag{2}$$

patches per variate. For ECL we use $L = 96$ following iTransformer (Liu et al., 2024), giving $N = 11$ patches per variate and $C \cdot N = 321 \times 11 = 3531$ tokens for CD. All remaining architectural hyperparameters are listed in Table 1.

Table 1: Architectural hyperparameters. Synthetic and ETTh1 share one configuration; ECL uses a wider model to accommodate 321 variates.

Hyperparameter	Synthetic + ETTh1	ECL
d_{model}	64	128
Attention heads	8	16
Encoder layers	3	3
Dropout	0.2	0.2
Sequence length L	512	96
Patch size P	16	16
Stride S	8	8
Patches per variate N	63	11

We include DLinear (Zeng et al., 2023) as a linear baseline. DLinear decomposes each variate's lookback window into trend (moving-average filter with kernel size 25) and remainder, then applies separate channel-independent linear maps from L to H for each component, summing the two outputs. It is the most competitive non-Transformer baseline on standard benchmarks and sets a natural lower bar for what qualifies as learned temporal structure. The DLinear values in Tables 3 and 6 are produced by a dedicated re-run using this full decomposition; CI and CD values in Table 6 are from a separate set of runs using the same windowing and seeds.

2.4 Training Regime

PatchTST (CI and CD) is trained with AdamW (lr $= 10^{-4}$, weight_decay $= 10^{-4}$), a cosine learning-rate schedule with linear warm-up, gradient clipping at max_norm $= 1.0$, and FP16 mixed precision (autocast + GradScaler). DLinear uses the same optimiser without warm-up. Early stopping is applied on validation MSE with patience 10 and a maximum of 100 epochs for ETTh1 and 50 epochs for all other experiments. All experiments run on a single NVIDIA T4 16 GB GPU.

Batch sizes. T4 memory constraints force different batch sizes between CI and CD at large C, creating an asymmetry in per-epoch gradient update counts. Table 2 reports these counts for each configuration.

For ETTh1, CI uses batch size 128 and CD uses batch size 32. For ECL (CI only), batch size is 8 with 1,940 steps per epoch.

Step-count confound. At $C = 84$, CD runs with batch size 1, giving 31.9× more gradient updates per epoch than CI. The cause is a hardware memory limit, not a deliberate experimental choice. We log `total_steps` for every run and report it alongside results. The significance of this asymmetry for interpreting the null result is discussed in Section 3.

Matched-budget ETTh1 protocol. Both CI and CD use batch size 32, warmup_epochs $= 10$, and a minimum of 15 epochs before early stopping may fire. The identical batch size gives both modes the same

Table 2: Batch sizes and gradient update counts per epoch by mode and C. The Ratio column is CD steps / CI steps. Step counts are those of the runs for seeds $\{42, 123, 456\}$; seeds 789 and 1011 trained on a marginally larger window count (8,033 rather than 7,433 windows at $C = 84$) with the same batch sizes and the same CD/CI step ratios.

C	CI batch	CD batch	CI steps/epoch	CD steps/epoch	Ratio
7	128	64	59	117	2.0×
21	128	8	59	930	15.8×
84	32	1	233	7433	31.9×

number of gradient updates per epoch at every horizon (252, 249, 244, and 232 at $H \in \{96, 192, 336, 720\}$ respectively), removing the step-count confound present in earlier runs. The minimum-epoch floor prevents a degenerate epoch-one stop without inflating test MSE, since the best validation checkpoint is still used for evaluation.

ECL CD infeasibility. At $C = 321$ and $L = 96$, a single CD training epoch takes approximately 5000 seconds on the T4, so a complete early-stopping run exceeds the 12-hour session limit on wall-clock time long before the 16 GB VRAM becomes binding. Throughout this paper, *commodity GPU budget* refers to single-T4 training with a 12-hour wall-clock limit, as in a standard Kaggle notebook session.

3 Results

3.1 Synthetic AR(1) Grid

Table 3 shows mean test MSE over five seeds for each cell. CI and CD track each other almost exactly across the grid. The largest mean difference is 0.0054 MSE units at $C = 84$, $\rho = 0.9$, and the direction of the difference is not consistent across the grid: CI leads at $C = 7$ across all ρ, at $C = 21$ for $\rho \in \{0.1, 0.5\}$, and at $C = 84$ for $\rho \in \{0.5, 0.9\}$; CD leads marginally at $C = 21$, $\rho = 0.9$ and at $C = 84$, $\rho = 0.1$. At $C \in \{21, 84\}$ DLinear is consistently worse than both PatchTST variants, confirming that the Transformer representations add value once dimensionality increases. At $C = 7$, DLinear is within roughly 1% of the PatchTST variants across ρ and occasionally slightly better; the CI/CD comparison remains the central question and both variants behave identically regardless of the DLinear baseline's position.

Table 3: Mean test MSE over seeds $\{42, 123, 456, 789, 1011\}$ for the synthetic AR(1) grid. CD/CI above 1.0 favours CI; below 1.0 favours CD.

C	ρ	CI	CD	DLinear	CD/CI
7	0.1	1.0282	1.0298	1.0404	1.0016
7	0.5	1.0404	1.0412	1.0557	1.0008
7	0.9	1.0642	1.0653	1.1064	1.0011
21	0.1	0.9931	0.9950	1.0013	1.0019
21	0.5	1.0179	1.0189	1.0339	1.0010
21	0.9	1.0327	1.0311	1.0595	0.9984
84	0.1	0.9832	0.9831	0.9831	0.9999
84	0.5	0.9752	0.9764	0.9781	1.0012
84	0.9	0.9759	0.9813	0.9942	1.0056

The $C = 84$, $\rho = 0.9$ cell deserves closer attention. Seeds 42, 123, and 456 all stopped at best epoch 1 or 2 (7433 / 14866 / 7433 total gradient updates), while the same seeds' CI runs stopped with only 233–466 updates. At $n = 5$, CD mean is 0.9813 and CI mean is 0.9759. CD therefore consumed up to 32× more compute per epoch and more epochs before stopping, yet landed at essentially the same test MSE. At $\rho \in \{0.1, 0.5\}$ at $C = 84$, CD consistently stops earlier than CI while CI takes 9–28 epochs. The two modes

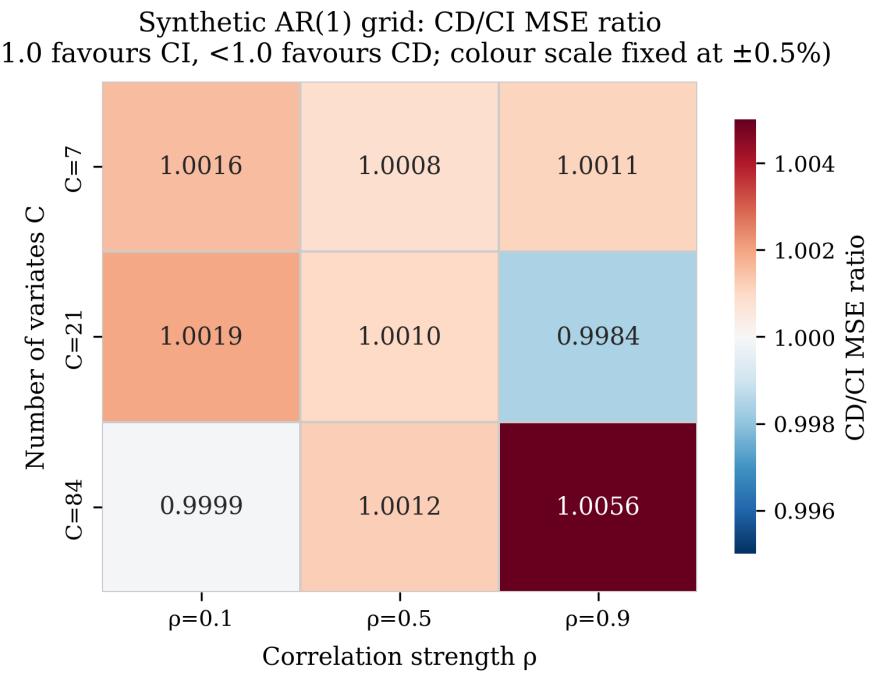

Figure 1: Synthetic AR(1) grid: CD/CI test-MSE ratio by variate count C and contemporaneous correlation ρ, averaged over seeds $\{42, 123, 456, 789, 1011\}$. Values above 1.0 favour CI, below 1.0 favour CD; the colour scale is fixed at $\pm 0.5\%$. No cell departs from unity by more than 0.56%, and the sign of the deviation is not consistent across the grid.

follow very different optimisation trajectories at this scale yet converge to the same test MSE, and the extra gradient updates buy CD nothing when there is no lagged cross-channel signal for the attention to act on.

Statistical analysis. For each (C, ρ) cell we compute the per-seed difference $d_s = e_s^{\mathrm{CD}} - e_s^{\mathrm{CI}}$ over seeds $\{42, 123, 456, 789, 1011\}$ and derive its mean, standard deviation, and 95% confidence interval from a t-distribution with four degrees of freedom ($t_{0.025,4} = 2.776$). Across all nine cells the mean CD$-$CI difference stays within ± 0.0054 MSE units, at most 0.56% of the corresponding CI mean, and seven of nine 95% CIs include zero. The two that exclude zero are at $C = 7$, $\rho = 0.1$ and $C = 21$, $\rho = 0.1$; both are positive (CD worse) and below 0.20% of the CI mean in their respective cells, well within the detection bound. Pooling the 45 matched pairs as a simple unweighted average over all differences, the grand-mean CD$-$CI is $+0.0013$ MSE (95% CI $[-0.0002, +0.0028]$, 0.13% of the mean CI MSE), and regressing the paired difference on C (categorical) and ρ finds no dependence on either factor (F-test $p = 0.81$, $R^2 = 0.02$). These intervals are a bound rather than a mere failure to reject: pooling the within-cell variances gives a 95% CI half-width of 0.0063 MSE at $n = 5$, roughly 0.63% of the mean CI MSE. Any uniform CD advantage larger than 0.63% of the CI mean would have pushed the grand-mean CI entirely above zero; none of the observed differences reached that threshold.

3.2 Real-Data Experiments

ETTh1 ($C = 7$). Under the matched-budget protocol, CI matches or beats CD at every horizon (Table 4). CD is significantly worse at $H = 96$ (paired 95% CI $[+0.0013, +0.0618]$) and $H = 720$ (95% CI $[+0.0041, +0.0367]$); the $H = 192$ and $H = 336$ differences are not significant. Pooled across all 20 pairs (four horizons, five seeds), the CD$-$CI difference is $+0.0147$ MSE (95% CI $[+0.0047, +0.0246]$). The $H = 192$ ratio sits below one (CD 0.5194 vs. CI 0.5229), the only horizon where CD leads, but the 95% CI $[-0.0227, +0.0159]$ includes zero and CD's lead is not reliable. Steps per epoch are identical across modes at every horizon, so neither the update count nor the warmup schedule can explain the CI advantage at the significant horizons.

Table 4: ETTh1 mean test MSE over seeds $\{42, 123, 456, 789, 1011\}$, matched-budget protocol (batch 32, warmup 10, min epochs 15 for both modes). CD/CI above 1.0 favours CI. SIG = paired 95% CI excludes zero.

H	CI	CD	CD/CI	SIG
96	0.4599	0.4914	1.0686	SIG
192	0.5229	0.5194	0.9935	
336	0.5789	0.5890	1.0174	
720	0.7062	0.7266	1.0289	SIG

ECL ($C = 321$). CD training is infeasible at this scale (Section 2): each CD epoch takes about 5000 s on the T4, so a run exceeds the 12-hour session limit well before the 16 GB VRAM binds. CI results across all four horizons appear in Table 5, with negligible cross-seed variance (CV below 0.5% at every horizon). No CI/CD significance test is reported, because the CD numbers do not exist. ECL is included as evidence of the scaling boundary of this CD formulation rather than as a performance benchmark, and the infeasibility is itself a scoped finding, specific to PatchTST's flattened-token tokenisation rather than to channel dependence as a concept.

Table 5: ECL CI test MSE per seed and mean over seeds $\{42, 123, 456\}$. CD is excluded due to computational infeasibility ($C \cdot N = 3531$ tokens; ≈5000 s per epoch on T4).

H	seed=42	seed=123	seed=456	Mean	Std
96	0.1754	0.1754	0.1756	0.1754	0.0001
192	0.1846	0.1830	0.1841	0.1839	0.0008
336	0.2026	0.2019	0.2029	0.2025	0.0005
720	0.2449	0.2448	0.2453	0.2450	0.0003

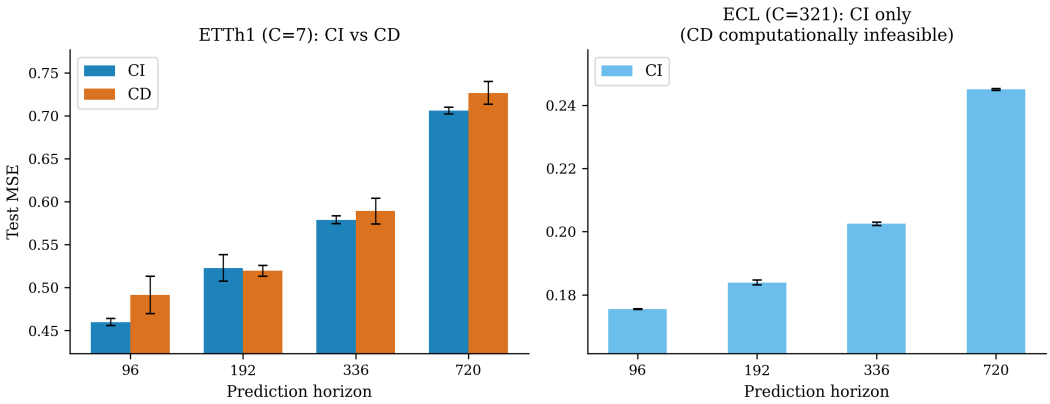

Figure 2: Real-data test MSE across prediction horizons $H \in \{96, 192, 336, 720\}$, with error bars showing the cross-seed standard deviation. Left: ETTh1 ($C = 7$), CI versus CD under the matched-budget protocol, mean over seeds $\{42, 123, 456, 789, 1011\}$. CI is significantly better at $H \in \{96, 720\}$ and never worse than CD across all horizons. Right: ECL ($C = 321$), CI only over seeds $\{42, 123, 456\}$, since the concatenated-token CD formulation is computationally infeasible at this variate count.

Correlation and lead-lag structure. For ETTh1, the lag-0 mean absolute Pearson correlation is 0.311 (median 0.233) and the mean of each pair's maximum absolute correlation over lags is 0.349; 43% of channel pairs gain at least 5 percentage points of absolute correlation from some non-zero lag. For ECL the corresponding numbers are higher: lag-0 mean 0.503 (median 0.483), mean max-over-lags 0.596, and 49% of

pairs gaining at least 5pp from non-zero lags. In both datasets lag 0 is the strongest single lag, but lagged structure adds meaningful signal for a substantial fraction of pairs.

Both datasets sit between the AR(1) grid (instantaneous correlation only) and the leader-follower regime (explicit structural lagged coupling). The ETTh1 and ECL results are contextual evidence about CI/CD behaviour in realistic settings, not controlled tests of the coupling-structure hypothesis.

4 Analysis: Leader-Follower Coupling Experiments

The AR(1) null result is easy to explain: when the data has no lagged cross-channel signal by construction, a model attending across channels has nothing additional to exploit beyond what channel-independent attention already sees. RQ2 asks whether this changes under structured lagged coupling.

4.1 Leader-Follower VAR(1) Design

We generate data from a $C = 21$ VAR(1) process with the following channel structure: channels 0–9 are *leaders* (pure AR(1), no cross-channel dependence); channels 10–19 are *followers*, where follower $i+10$ depends on leader i with lag-1 coefficient γ; channel 20 is an *isolate* (pure AR(1), internal negative control). The transition matrix $A(\gamma)$ is lower-triangular with diagonal entries $\phi = 0.8$ and off-diagonal entries $A[i+10, i] = \gamma$ for $i \in \{0, \ldots, 9\}$. Lower-triangular structure guarantees spectral radius ϕ at all finite γ, so the process is stationary by construction.

At $\gamma = 0$ the process reduces to the original AR(1) grid at $C = 21$, $\rho = 0.5$, providing a cross-experiment sanity check. We sweep $\gamma \in \{0.0, 0.3, 0.6, 0.9\}$ with five seeds and CI, CD, and DLinear modes, all at patch size $P = 16$.

4.2 Patch-Resolution Sweep

The null result at $P = 16$ has a natural mechanistic explanation: with stride $S = 8$, each patch spans 16 timesteps. A lag-1 cross-channel relationship means follower $i + 10$ at time t depends on leader i at time $t - 1$. Both timesteps likely fall within the same patch, but whether the cross-variate attention can resolve a single-step offset within a 16-timestep window is unclear. Smaller patches would place lag-1 coupling more prominently within the patch structure and give CD's attention a more concentrated signal.

We therefore run a joint sweep over $\gamma \in \{0.0, 0.3, 0.6, 0.9\}$ and $P \in \{2, 4, 8, 16\}$ with $S = P/2$ (50% overlap), modes CI and CD, five seeds. At $P = 2$, each patch spans exactly two timesteps and stride $S = 1$, so a lag-1 relationship falls across adjacent patches, the finest resolution the architecture supports. CD is excluded at $P = 4$ because the sequence length $C \cdot N = 21 \times 255 = 5355$ tokens exceeds T4 VRAM at any viable batch size.

4.3 Results

Table 6 reports mean CI, CD, and DLinear MSE at $P = 16$ across the γ sweep ($C = 21$, pred_len $= 96$, five seeds). The CD/CI ratio stays within 1% of unity throughout and never falls below it, but it does climb with γ under this early-stopping protocol: regressing the paired CD$-$CI difference on γ gives a slope of $+0.0072$ MSE per unit γ (95% CI $[+0.0027, +0.0117]$, $p = 0.004$, $R^2 = 0.38$). That climb comes from how early stopping selects CD rather than from the coupling itself. Section 5.1 ties it to CD's earlier validation minimum: taking each model at its own minimum shrinks the gap below the threshold and can reverse its sign. DLinear stays within 0.004 MSE of CI at every γ, so at this C a channel-independent linear map is about as accurate as patch attention, though Table 3 shows the parity breaks at larger C or ρ. These DLinear runs reached the epoch budget without early stopping and should be read as budget-limited reference points rather than fully converged optima.

Table 7 shows the CD/CI ratio across the (P, γ) grid. At $P \in \{8, 16\}$ the entry is a genuine CD/CI ratio; at $P \in \{2, 4\}$ CD is infeasible, so the cell carries the CI-only test MSE where a run exists and a dash otherwise ($P = 4$ at $\gamma \in \{0.0, 0.3\}$, and $P = 2$ at $\gamma \in \{0.0, 0.3\}$, were not run under compute constraints). This joint

Table 6: Leader-follower VAR(1) at $P = 16$. Mean test MSE and CD/CI ratio over seeds $\{42, 123, 456, 789, 1011\}$. CI and CD match the AR(1) grid at $\gamma = 0.0$ to within seed variance. The CD/CI ratio stays within 1% of unity across γ; under this early-stopping protocol it drifts upward with γ, an effect Section 5 attributes largely to selection rather than to coupling.

γ	CI	CD	DLinear	CD/CI
0.0	1.0270	1.0287	1.0301	1.0016
0.3	1.0326	1.0393	1.0330	1.0065
0.6	1.0298	1.0377	1.0301	1.0077
0.9	1.0278	1.0362	1.0284	1.0082

sweep is a separate set of runs from the γ-sweep in Table 6; its $P = 16$ ratios sit nearer 1.001 than the 1.007 reported there, a difference of a few thousandths of MSE that is comparable to the per-seed spread in both experiments. The two agree on the substantive point.

Table 7: CD/CI MSE ratio in the (P, γ) boundary experiment. At $P \in \{8, 16\}$ both modes run and the entry is a CD/CI ratio (above 1.0 favours CI). At $P \in \{2, 4\}$ CD exceeds T4 memory, so no ratio is defined; those cells carry the CI-only test MSE where a run exists and a dash where none was performed.

P	$\gamma = 0.0$	$\gamma = 0.3$	$\gamma = 0.6$	$\gamma = 0.9$
16	1.0012	1.0004	1.0005	1.0010
8	1.0014	1.0009	1.0009	1.0004
4	—	—	CI: 0.9757	CI: 0.9743
2	—	—	CI: 0.9806	CI: 0.9801

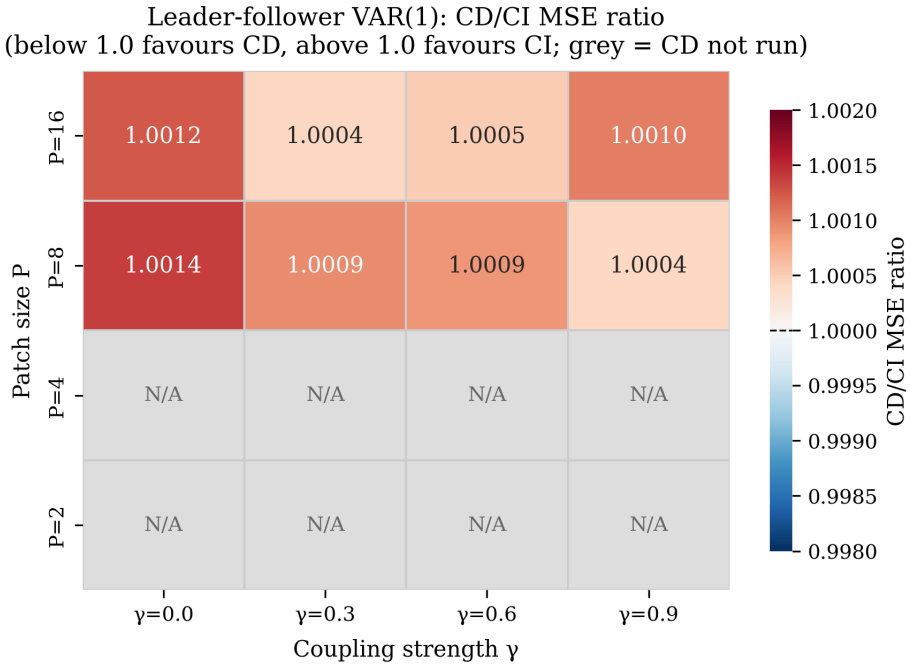

Figure 3: Leader-follower VAR(1): CD/CI test-MSE ratio across the (P, γ) grid, mean over seeds $\{42, 123, 456, 789, 1011\}$. Values above 1.0 favour CI. Coloured cells are the configurations where both modes were run ($P \in \{8, 16\}$); grey cells mark where CD was not run ($P \in \{2, 4\}$, infeasible on the T4). No cell favours CD, and the ratio stays within 0.15% of unity throughout.

4.4 Mechanistic Interpretation

The two leader-follower experiments differ in how sensitive the gap looks to γ. The boundary sweep over $P \in \{8, 16\}$ gives a flat trend (slope -0.0006 MSE per unit γ, 95% CI $[-0.0021, +0.0009]$, $p = 0.44$, five seeds per cell), whereas the dedicated $P = 16$ sweep shows a larger apparent trend under the original early-stopping protocol. A seed-clustered version of the same regression, treating each seed as a cluster, gives an equally flat slope (95% CI $[-0.0029, +0.0018]$, $p = 0.62$), so the result does not depend on treating seeds as independent. They agree that CD never holds a practically relevant lead; they differ on whether the gap drifts with γ. Section 5 resolves this difference. The apparent trend in the dedicated sweep is largely a selection artefact: CD overtrains, and the original protocol selects it from the descending edge of its validation curve rather than at its minimum (Section 5.1). The part of the gap that does track γ is a property of the shared prediction head rather than the encoder (Section 5.3). These are two separate findings, each supported by its own experiment, and we do not use either to stand in for the other.

A representational reading of why CD gains nothing from the coupling is also available, though it is secondary to the protocol explanation. At $P = 16$ and $S = 8$, a lag-1 cross-channel relationship maps to the same patch position in both the leader and the follower sequence rather than to a phase offset between them, so the cross-variate attention has no distinct inter-patch signal to route: position k of variate $j + 10$ sits beside position k of variate j in the $C \cdot N = 21 \times 63 = 1323$-token sequence, but the lag-1 information is already aggregated within the shared patch position. CD and CI test MSEs stay within seed variance in every cell where both modes run ($P \in \{8, 16\}$). The finest resolution, $P = 2$ with stride 1, where lag-1 coupling would fall across adjacent patches and a phase offset would be visible, is CI-only because CD exceeds T4 memory there, so this representational account cannot be tested directly and remains a conjecture rather than a result.

5 Fair-Protocol Robustness Controls

The leader-follower $P = 16$ sweep raises a question the main protocol cannot answer on its own: the apparent CD$-$CI gap grows with γ (Section 4), which could mean CD is genuinely worse under stronger coupling, or could mean the early-stopping protocol selects CD at a disadvantage. This section dissects that gap with four controls, each isolating one confound, and then summarises the resulting effect sizes against a pre-registered threshold. We fix the practical-equivalence threshold at 1% of the CI mean test MSE, chosen before these controls were run; differences smaller than this in either direction are treated as practically irrelevant regardless of statistical significance.

5.1 Overtraining and Selection

To see why the gap appears, we instrument a single training trajectory per (mode, seed) at the leader-follower $\gamma = 0.6$, $C = 21$ cell, logging validation MSE five times per epoch and training past the early-stopping horizon for all five seeds $\{42, 123, 456, 789, 1011\}$. Because mid-epoch validation runs in evaluation mode, it draws no random numbers and does not perturb the weight trajectory, so a single faithful run supports every post-hoc selection rule we compare.

The curves rule out a cadence effect. Selecting CD at its best epoch boundary against its best mid-epoch update changes the result by $+0.0005$ MSE (95% CI $[-0.0049, +0.0059]$), and the same quantity for CI is -0.0007 MSE, so neither mode hides a better checkpoint between epoch boundaries. What the curves do show is overtraining in CD. Averaged over seeds, CD reaches its validation minimum at epoch 5 to 6 and then climbs from 0.986 to 1.200 MSE by the end of training, while CI reaches its minimum later, at epoch 15 to 22, and stays near it (0.986 to 1.003). Figure 4 shows the asymmetry.

This is what drives the headline gap. Under the original protocol the CD$-$CI difference in this cell is $+0.0079$ MSE (95% CI $[+0.0046, +0.0113]$). Taking each model at its own validation minimum brings it to -0.0031 MSE (95% CI $[-0.0109, +0.0046]$), a tie leaning slightly to CD, and replaying early stopping with patience 10 on the same curves gives -0.0019 MSE (95% CI $[-0.0105, +0.0066]$), again a tie. The deficit under the original protocol is therefore largely a matter of selecting CD from the rising tail of its validation curve. We have full trajectories for this one cell only, but the early validation minima it depends on recur in the

cross-variate-head and block-covariance runs below, where CD's best epoch is 4 to 7. Whether the same quantitative overtraining holds across other tasks is something these data cannot settle.

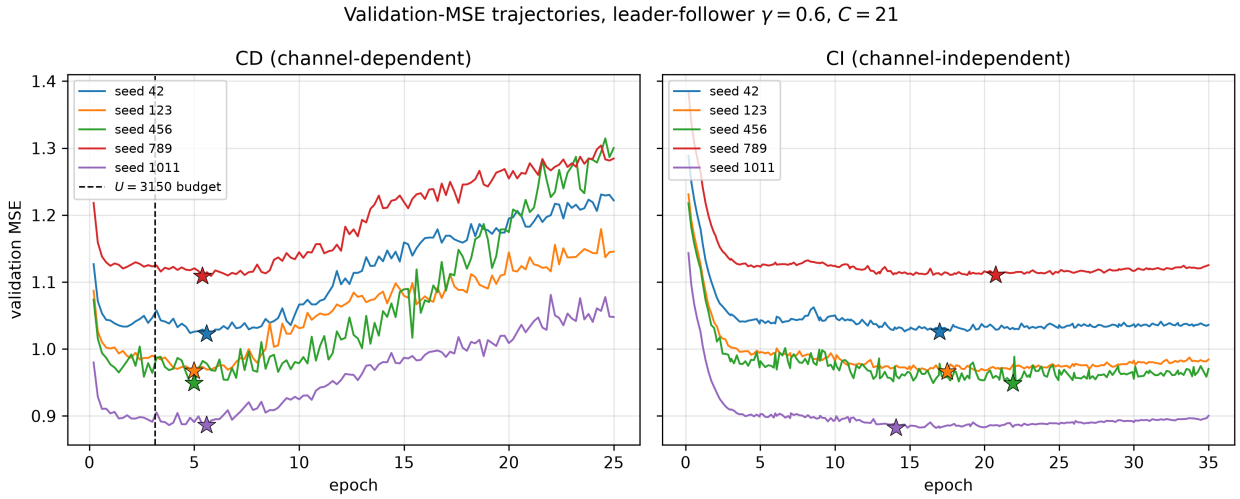

Figure 4: Validation-MSE trajectories at the leader-follower $\gamma = 0.6$, $C = 21$ cell, one faithful run per seed, drawn to each mode's diagnostic horizon. Stars mark each seed's validation minimum. CD (left) reaches a sharp minimum near epoch 5 to 6 and then degrades, while CI (right) settles into a flat basin. The dashed line on the CD panel marks the equal-compute update budget $U = 3150$, which falls before CD's minimum and therefore selects CD from its descending edge.

5.2 Matched Compute

At $C = 21$ the original protocol gives CD many more gradient updates per epoch than CI (Table 2), which invites the objection that CD is simply trained for longer and pushed past its optimum. Capping both modes at an identical update budget and taking the best checkpoint within it answers that objection. The deficit does not survive the cap: the CD−CI difference is -0.0036 MSE (95% CI $[-0.0059, -0.0012]$) at leader-follower $\gamma = 0.0$ and -0.0053 MSE (95% CI $[-0.0092, -0.0014]$) at $\gamma = 0.6$, both favouring CD, and -0.0028 MSE (95% CI $[-0.0138, +0.0082]$) at the AR(1) $C = 84$, $\rho = 0.9$ cell, a tie. The per-epoch step-count asymmetry is not what produces the apparent deficit. We treat this protocol as a sensitivity check rather than a canonical schedule, because the budget $U = 3150$ is itself too tight for CD at leader-follower $\gamma = 0.6$: set below CD's validation minimum, it selects from the descending edge and the sign turns positive ($+0.0059$ MSE, 95% CI $[+0.0003, +0.0114]$), which is the selection effect of Section 5.1 rather than a separate disadvantage.

5.3 Head Fairness

The CD architecture in the main experiments shares CI's per-variate prediction head, so a residual objection is that the head, not the encoder, limits CD. We rerun the leader-follower sweep with a cross-variate head (CD_Head) that pools across variates before projecting, leaving the encoder unchanged. The cross-variate head ties CI at every coupling strength: CD_Head−CI is $+0.0016$ ($\gamma = 0.0$), $+0.0030$ ($\gamma = 0.6$), and $+0.0023$ MSE ($\gamma = 0.9$), all within 0.3% of the CI mean and all with 95% CIs spanning zero. Pooled across γ, the CD_Head−CI grand mean is $+0.0023$ MSE ($+0.23$% of the CI mean) with a 95% CI whose lower bound sits just above zero, so the cross-variate head is detectably though negligibly worse than CI in pooled terms, well inside the 1% band. The plain shared-head CD over the same runs is $+0.0017$, $+0.0079$, and $+0.0085$ MSE, carrying the γ-dependent widening that the main sweep showed. Swapping only the head removes that widening while leaving the encoder fixed, which locates the coupling-dependent widening in the shared head rather than in the cross-variate encoder. This is a head effect specifically; it is separate evidence from the overtraining phenomenon of Section 5.1, and we do not use one to explain the other.

5.4 Covariance Family

The AR(1) grid uses compound-symmetry covariance, where every channel pair shares the same correlation. To check that the null is not specific to that structure, we test a block-covariance family in which within-block correlation $\rho_{\text{in}} \in \{0.5, 0.9\}$ is concentrated in groups of seven variates and cross-block correlation is zero, giving CD a block structure to exploit that compound symmetry does not offer. CD gains no practically relevant advantage there either: at $C = 21$ the CD−CI difference is +0.0028 MSE (+0.27%, 95% CI $[+0.0015, +0.0040]$) at $\rho_{\text{in}} = 0.5$ and −0.0062 MSE (−0.60%, 95% CI $[-0.0137, +0.0013]$) at $\rho_{\text{in}} = 0.9$; at $C = 84$, $\rho_{\text{in}} = 0.5$ it is +0.0005 MSE (+0.05%, 95% CI $[-0.0005, +0.0016]$), and at $C = 84$, $\rho_{\text{in}} = 0.9$ it is +0.0000 MSE (+0.00%, 95% CI $[-0.0010, +0.0010]$), a tie. Regressing the paired difference on variate count and ρ_{in} across the four cells gives a significant ρ_{in} slope (−0.0118 MSE per unit ρ_{in}, $p = 0.013$, $R^2 = 0.35$), so CD's standing relative to CI improves slightly as within-block correlation rises, while variate count carries no effect ($p = 0.26$). Every cell sits within the 1% threshold, and the $C = 84$ cells again ran CD at batch size 1.

5.5 Practical-Equivalence Summary

Table 8 collects the controlled synthetic effect sizes with the selection rule used for each, so that cells evaluated under different protocols are never pooled. Across all tested AR(1), leader-follower, and block-covariance cells, the absolute CD−CI difference stays within the pre-registered 1% relative-MSE threshold in both directions. For the AR(1) grid this holds cell by cell as well as in aggregate: no single cell exceeds 0.56% of its CI mean (Section 3), so the grand-mean row of Table 8 does not average over any cell that breaches the threshold. Some differences are statistically detectable in one direction or the other, but none reaches the practical threshold. The conclusion is narrow and symmetric: on these controlled synthetic processes, neither mode holds a practically relevant accuracy advantage under fair selection.

Table 8: Controlled synthetic effect sizes, CD−CI in test MSE, each with the selection rule under which it was computed. Cells under different rules are never pooled. "Rel" is the difference as a percentage of the CI mean; the final column reports whether the absolute difference stays within the pre-registered 1% threshold.

Family	Cell	Selection rule	CD−CI	Rel	$\leq 1\%$
AR(1) grid	grand mean (9 cells)	early-stop	+0.0013	+0.13%	yes
Leader-follower	$\gamma = 0.6$	early-stop	+0.0079	+0.77%	yes
Leader-follower	$\gamma = 0.6$	val-minimum	−0.0031	−0.30%	yes
Leader-follower	$\gamma = 0.6$	matched-U	−0.0053	−0.52%	yes
Leader-follower	$\gamma = 0.9$	early-stop	+0.0085	+0.82%	yes
Leader-follower	$\gamma = 0.6$, cross-var. head	early-stop	+0.0030	+0.29%	yes
Leader-follower	$\gamma = 0.9$, cross-var. head	early-stop	+0.0023	+0.23%	yes
AR(1)	C=84, ρ=0.9	matched-U	−0.0028	−0.27%	yes
Block-cov.	C=21, ρ_{in}=0.5	early-stop	+0.0028	+0.27%	yes
Block-cov.	C=21, ρ_{in}=0.9	early-stop	−0.0062	−0.60%	yes
Block-cov.	C=84, ρ_{in}=0.5	early-stop	+0.0005	+0.05%	yes
Block-cov.	C=84, ρ_{in}=0.9	early-stop	+0.0000	+0.00%	yes

6 Limitations

Every conclusion here is conditional on the generative families studied. The AR(1) process has a diagonal transition, Gaussian innovations, and compound-symmetry covariance; the leader-follower extension adds one specific coupling topology with a single scalar γ at $C = 21$. We do not explore leader-follower structures at higher variate counts (for example, multiple leader-follower blocks embedded in $C = 84$); the leader-follower results are strictly at $C = 21$. The covariance families tested, compound-symmetry on the AR(1) grid and block-diagonal in Section 5.4, do not exhaust the space; hierarchical or factor-structured correlations could behave differently even without lagged dependence and are not tested here. Richer coupling graphs, non-Gaussian noise, and non-stationary dynamics are also excluded. The AR(1) coefficient is fixed at $\phi = 0.8$

and all synthetic experiments use pred_len = 96, so the conclusions are specific to this dynamic regime and horizon.

The CD architecture used in the main experiments applies a per-variate head after the joint encoder, so cross-variate information from the encoder is not explicitly pooled during prediction. For RQ1 this is not a concern: the AR(1) design contains no lagged cross-channel signal by construction, so if the encoder attending across all $C \cdot N$ tokens produces no CI/CD gap, no head design could manufacture one from a signal that is absent. For the leader-follower experiments the head-fairness control in Section 5.3 addresses this objection in part: a cross-variate head ties CI at every γ and removes the coupling-dependent widening, which places that widening in the head rather than the encoder. What the control does not establish is that some other head design could turn the lagged signal into a genuine CD advantage; it shows only that the shared head, not an encoder bottleneck, drives the widening seen under the original protocol.

The $P = 2$ CD case was not evaluated: at $C = 21$ and $N = 511$, the CD sequence length of $21 \times 511 = 10{,}731$ tokens exceeds T4 VRAM at any batch size. Only CI runs are available at $P = 2$, and only for $\gamma \in \{0.6, 0.9\}$; the remaining $P = 2$ cells were not run under compute constraints. This is the configuration where CD's attention resolution most closely matches the lag-1 coupling scale, since one patch spans two timesteps and a lag-1 relationship falls across adjacent patches, so the presence or absence of a CD advantage there remains empirically untested. A variate-token CD implementation, which attends over C tokens rather than $C \cdot N$, would avoid this memory constraint and would allow repeating the boundary sweep without the infeasibility gap. Implementing such a variant was outside the scope of this work.

All architectural conclusions are confined to PatchTST with flattened-token channel dependence. Other ways of building channel dependence, such as hierarchical or graph-based encoders, may behave differently, and the results are in any case specific to the patch sizes and lookback windows used here.

7 Related Work

PatchTST (Nie et al., 2023) introduced patch-based channel-independent Transformers for long-horizon forecasting and showed competitive or superior performance against models that attend across variates. The benchmark results leave open the mechanism: the datasets confound patch size, weight sharing, and cross-variate correlation simultaneously, so the source of any CI advantage cannot be isolated from the architecture alone.

DLinear (Zeng et al., 2023) established a surprisingly competitive baseline: a decomposition-linear model that splits each series into a moving-average trend and a residual, each mapped linearly from lookback to horizon. A Transformer that cannot beat it is not demonstrably learning useful temporal structure beyond what a linear filter captures.

iTransformer (Liu et al., 2024) inverts the embedding: each variate's full history becomes a single token and attention operates across variates. It performs well on large multivariate benchmarks, but those benchmarks mix correlation strength, dimensionality, and temporal structure in ways that make attributing any gain to cross-variate attention specifically very difficult.

LIFT (Zhao & Shen, 2024) identifies leader-follower pairs empirically and prepends leader context to each variate's input before a CI model, recovering cross-variate signal without full joint attention. It is the closest antecedent to the leader-follower design used here. Its limitation in this context is that the datasets it uses are observational, so lag structure and instantaneous correlation cannot be decoupled.

Abdelmalak et al. (Abdelmalak et al., 2026) show that standard benchmarks favour CI because they are dominated by within-channel periodicity, while datasets generated from chaotic ODEs with genuine dynamical coupling favour CD. The two papers are complementary rather than competing: where Abdelmalak et al. contrast dataset categories, we fix the generative model analytically and sweep coupling strength continuously. The near-unity CD/CI ratios across all nine AR(1) cells (0.9984–1.0056) indicate that the ODE-dataset advantage of CD does not extend to the instantaneous-correlation regime. The leader-follower experiment probes the boundary from the other side and finds the null result extends there too, at least within the patch sizes and coupling strengths tested.

Read together, these five lines of work leave two gaps. None runs a factorial experiment varying cross-variate correlation and dimensionality independently under a known generative model, so none can attribute performance differences to a single factor. And none analytically separates instantaneous correlation from dynamical coupling: LIFT probes lead-lag empirically on observational data; Abdelmalak et al. build in strong ODE coupling but have no instantaneous-only baseline. The two synthetic experiments here address both gaps within a specific generative family, under the PatchTST architecture and training conditions described above.

8 Conclusion

We set out to find where in the coupling-structure space a channel-dependent PatchTST starts to beat its channel-independent counterpart. Within the ranges we tested it does not, and the fair-protocol controls show that the open question is one of cost rather than accuracy. On the AR(1) grid the grand-mean CD−CI difference is +0.0013 MSE with a pooled within-cell detection half-width of 0.0063 MSE over five seeds, and every cell difference falls inside that bound. Under leader-follower coupling a CD deficit appears to grow with γ on the original early-stopping protocol, but the diagnostic at $\gamma = 0.6$ attributes it largely to selection: CD reaches an earlier, sharper validation minimum than CI and worsens past it, so taking each model at its own minimum closes the gap (-0.0031 MSE, 95% CI $[-0.0109, +0.0046]$). Matching the gradient-update budget rules out the per-epoch step-count asymmetry as the cause. Swapping in a cross-variate head keeps CD_Head within 0.3% of CI at every γ and removes the γ-dependent widening, which places that widening in the shared head rather than the encoder, and a block-covariance family carries the no-advantage finding past compound-symmetry correlation. No tested synthetic cell shows an absolute CD−CI difference outside a pre-registered 1% relative-MSE threshold in either direction, so on these controlled processes neither mode holds a practically relevant accuracy advantage once it is selected fairly.

The two modes diverge in cost. Making PatchTST channel-dependent by concatenating every variate's patches into one attention sequence gives $C \cdot N$ tokens and an attention cost of $\mathcal{O}(C^2N^2)$ against $\mathcal{O}(CN^2)$ for CI. That cost forces CD onto smaller batches and many more gradient steps per epoch, drops its batch size to one at $C = 84$, and at $C = 321$ (ECL) pushes a single training epoch to roughly 5000 s on a T4, enough for a full run to exceed the 12-hour session budget on wall-clock time well before memory binds. CI trains without trouble across the same range. Variate-token designs such as iTransformer (Liu et al., 2024) attend over C tokens instead of $C \cdot N$ and avoid the wall, so what we report is specific to PatchTST's flattened-token CD rather than to channel dependence in general. On a single-T4 budget CI is the better default on accuracy per unit compute: it matches CD wherever both run, and it still trains where this CD formulation cannot.

The observational results point the same way, though they sit outside the equivalence claim. On ETTh1 ($C = 7$) under a matched-budget protocol, CI holds a small but reliable edge: significantly better at two of four horizons, never worse, with a pooled CD−CI of +0.0147 MSE, about 2.6% of the CI mean. That figure is above the 1% band we apply to the synthetic cells, so we report it as a genuine if modest CI advantage and keep it separate from the equivalence result.

Several things are left open by design. The generative families are narrow: diagonal-transition AR(1) and one leader-follower topology with a scalar γ, the block-covariance extension, all at a single horizon (pred_len = 96) and at $C = 21$ for the leader-follower experiments. The $P = 2$ CD configuration, where attention resolution most closely matches the lag-1 coupling scale, is the most direct comparison left undone because its $C \cdot N = 10{,}731$ token sequence exceeds T4 memory. The overtraining we document is a phenomenon we can measure but not yet explain: CD reaches its validation minimum earlier and worsens past it under this protocol and data, and we do not establish why. A variate-token CD implementation would remove the memory wall and let the boundary sweep and the $P = 2$ comparison be run directly; whether the null survives that formulation is the natural next test.

Broader Impact Statement

This work is a controlled empirical study of forecasting architectures on synthetic processes and two public benchmark datasets. It introduces no new data collection, no application-specific model, and no capability

with a plausible path to misuse. Its practical implication is guidance on architecture choice under constrained compute, which, where adopted, reduces the energy spent training multivariate forecasting models.

References

Ibram Abdelmalak, Kiran Madhusudhanan, Jungmin Choi, Christian Klötergens, Vijaya Krishna Yalavarthi, Maximilian Stubbemann, and Lars Schmidt-Thieme. Channel dependence, limited lookback windows, and the simplicity of datasets: How biased is time series forecasting? In *Advances in Knowledge Discovery and Data Mining (PAKDD 2026)*, volume 16598 of *Lecture Notes in Computer Science*, pp. 585–597. Springer, 2026. doi: 10.1007/978-981-92-1462-4_46. URL https://arxiv.org/abs/2502.09683.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06625. ICLR.

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. URL https://openreview.net/forum?id=Jbdc0vTOcol. ICLR.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317. URL https://arxiv.org/abs/2205.13504. AAAI.

Lifan Zhao and Yanyan Shen. Rethinking channel dependence for multivariate time series forecasting: Learning from leading indicators. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2401.17548. ICLR.