VISION PRETRAINING FOR TIME-SERIES FORECASTING DEPENDS ON MORE THAN UNDIRECTED GRID TOPOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

VisionTS poses an unusual transfer problem: a vision model pretrained on natural images forecasts a numerical sequence by reconstructing a masked image. We test whether undirected grid structure explains why those pretrained weights help. Phase, cycle, and joint reflections preserve observed values and undirected adjacency exactly while changing how temporal structure occupies the visual coordinate system. Across 11 datasets, 71 series, 2,272 origins, and three masked-image checkpoints, reflections reduced pretrained-weight advantage by 0.113–0.494 under MSE. The effect persisted across random controls and under MAE, MASE, and sMAPE. A post-hoc loss decomposition shows that reflection increased pretrained losses while leaving architecture-matched random losses nearly unchanged. Undirected topology alone therefore cannot explain the transfer benefit. A registered directed-geometry hypothesis did not meet its cross-dataset criterion and did not generalize to held-out transformations; direction, absolute position, and boundary anchoring remain unresolved.

1 INTRODUCTION

VisionTS poses an unusual transfer-learning puzzle. It turns a numerical history into a period-aligned image, masks the future, and uses a masked autoencoder pretrained on natural images to complete the image (Chen et al., 2025; He et al., 2022). Why should weights learned from photographs help with a rendered time series?

One possibility is that the image exposes useful undirected grid structure: neighboring pixels represent observations that are nearby in time or occupy corresponding positions in different cycles. But the same rendering also assigns a direction to those relations, places observations at absolute image coordinates, and anchors them relative to the visible–masked boundary. Standard corruptions alter several of these properties simultaneously, so they cannot reveal which one matters. Moreover, a transformed image may simply be harder for any network, irrespective of pretraining.

We address both ambiguities with a paired intervention. For every rendering, we compare a pretrained checkpoint with an architecture-matched random initialization and define their log loss ratio as pretrained-weight advantage. We then reflect the visible texture across its phase axis, cycle axis, or both (Figure 1). Reflection preserves the observations and their undirected grid adjacency exactly while changing orientation, absolute position, or boundary anchoring. If pretrained-weight advantage changes, undirected topology cannot by itself explain the benefit of natural-image pretraining.

Across 11 datasets and three masked-image checkpoints, all three reflections reduce pretrained-weight advantage. This supplies a counterexample: preserving undirected affinity does not preserve the contribution of pretrained weights. A separately proposed directed-displacement score shows a positive aggregate association but fails its registered cross-dataset criterion and does not generalize to held-out transformation families. The study contributes a controlled topology-preserving intervention, an estimand that separates pretrained transfer from difficulty shared by an architecture-matched random model, and cross-dataset evidence that useful visual transfer depends on more than undirected adjacency. It establishes sensitivity to the pretrained coordinate system without identifying direction, absolute position, or boundary anchoring as the mechanism. The evaluation plan and

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.

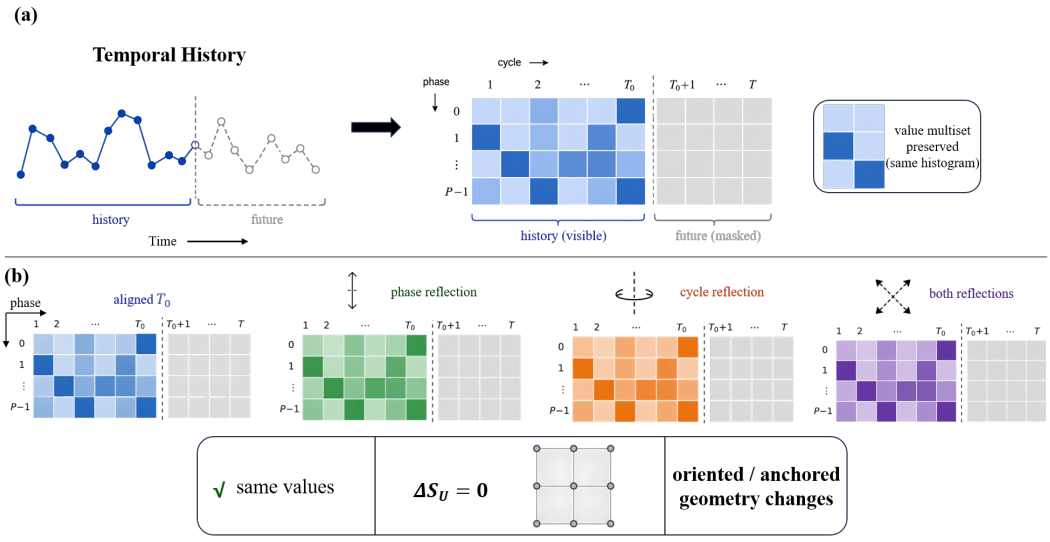


Figure 1: A reflection isolates what undirected topology cannot explain. (a) The history becomes a period-by-cycle texture and the future occupies a masked block. (b) Reversing the phase axis, cycle axis, or both preserves every observed value and the undirected grid affinity ($\Delta S_U = 0$), but changes orientation, absolute position, or position relative to the mask boundary.

data selection were fixed before these outcomes were examined; Appendix A records the protocol history and safeguards.

2 Related Work

Vision models as time-series forecasters. VisionTS renders a univariate history as a period-based image and reconstructs its masked future with an image-pretrained MAE (Chen et al., 2025; He et al., 2022). Subsequent work studies alternative image encodings and temporal perturbations (Zhao et al., 2025), multiple decomposed views (Shen et al., 2025a), time-series continual pretraining (Shen et al., 2025b), and adaptation across spectral and structural modality gaps (Zhang et al., 2026). TimesNet instead learns intra- and inter-period 2D structure directly (Wu et al., 2023), while numerical foundation models train on large sequence corpora (Ansari et al., 2024; Das et al., 2024; Woo et al., 2024). Input perturbation can reveal whether a forecast depends on an encoding, but it need not reveal whether the dependence comes from pretrained weights or from the architecture itself. We ask whether a specific invariant preserves the incremental benefit of vision pretraining under a matched pretrained-versus-random comparison.

Isolating transfer from task difficulty. Transfer-learning studies distinguish target-task performance from the gain due to pretrained initialization (Yosinski et al., 2014; Kornblith et al., 2019; Raghu et al., 2019; He et al., 2019). Matched pretrained and random models have been used to separate feature reuse from low-level image statistics (Neyshabur et al., 2020) and to estimate transferability (Sorkhei et al., 2025). Cross-modal forecasting work likewise compares pretrained and random language models, tokenizers, and perturbations (Tan et al., 2024; Riachi et al., 2025). We combine this matched comparison with transformations that preserve one candidate invariant exactly.

Coordinates, invariance, and symmetry. Vision Transformers process spatial patches together with positional information (Dosovitskiy et al., 2021). MAE uses an asymmetric encoder–decoder to reconstruct missing patches (He et al., 2022); SimMIM reconstructs masked pixels with a lightweight head and can use hierarchical Swin backbones (Xie et al., 2022; Liu et al., 2021). Absolute and relative position encodings affect recognition and can represent directed relative distances (Wu et al., 2021). Work on equivariant networks formalizes responses to symmetry transformations (Cohen & Welling, 2016; Tai et al., 2019), while permutation-invariant models discard

order by construction (Zaheer et al., 2017). For a rendered time series, moving values can change local neighborhoods, displacement direction, absolute patch position, and distance to the reconstruction boundary. Our controls keep undirected grid adjacency fixed and leave the other properties as competing explanations.

Learned invariances can help or hinder transfer (Speicher et al., 2024), and transformation probes can reveal soft equivariance in generic sequence interfaces (Li & Cheng, 2026). Our outcome is not representation invariance itself, but the incremental benefit of pretrained weights. The resulting counterexample combines an exact invariant with a paired architecture-matched comparison: $\Delta S_U = 0$ while $\Delta A < 0$.

3 What Must the Rendering Preserve?

The central comparison asks whether a transformation changes the benefit of pretraining, rather than merely changing forecasting difficulty. Let dataset d provide a scalar history $x_{1:L}$ and future $y_{1:H}$. A rendering T maps history index t to grid coordinate $p_t^{(T)} = (r_t^{(T)}, c_t^{(T)})$; the history is visible and the future region is masked. For model architecture m, let $L_{P,d,T}$ and $L_{R,d,T}$ be the mean losses from the image-pretrained and architecture-matched random models. We define pretrained-weight advantage as

$$A_{d,T} = \log L_{R,d,T} - \log L_{P,d,T}, \tag{1}$$

so $A > 0$ means that pretrained weights help. Because both models see the same rendering, the matched contrast controls for rendering effects shared by the pretrained and random members of the architecture pair.

For a reflection T_r paired with the aligned rendering T_0, define the matched change in pretrained-weight advantage as

$$\Delta A_d(T_r) = A_{d,T_r} - A_{d,T_0}. \tag{2}$$

Equivalently,

$$\Delta A = \underbrace{\Delta \log L_R}_{\text{change for random weights}} - \underbrace{\Delta \log L_P}_{\text{change for pretrained weights}}. \tag{3}$$

For every reflection, $\Delta S_U = 0$ exactly. A systematic nonzero ΔA therefore rules out undirected affinity as a sufficient characterization of pretrained-weight advantage. It does not, by itself, distinguish among the coordinate properties changed by reflection.

We also ask whether preserving the direction of relevant grid displacements predicts A. Scores are defined relative to the aligned rendering T_0. Dataset-specific lag weights $w_d(\ell)$ use only the first half of each selected series and the lag set $\mathcal{L}_d = \{1, P_d\}$, where P_d is the fixed period; they normalize median absolute autocorrelations across selected series. The directed displacement score is

$$S_D(T; d) = \sum_{\ell \in \mathcal{L}_d} w_d(\ell) \frac{1}{L-\ell} \sum_{t=1}^{L-\ell} \exp\left[- \left\| \Delta p_t^{(T,\ell)} - \Delta p_t^{(T_0,\ell)} \right\|_1 \right], \tag{4}$$

where $\Delta p_t^{(T,\ell)} = p_{t+\ell}^{(T)} - p_t^{(T)}$. The undirected affinity S_U discards displacement direction; S_P and S_B instead measure agreement with aligned absolute coordinates and with horizontal distance from the history–future boundary. Full definitions appear in Appendix B.

The primary regression is $A_{d,T} = \alpha_d + \beta_D S_D(T; d) + \epsilon_{d,T}$, with dataset fixed effects. The dataset is the scientific unit; origins and transformation seeds are not treated as independent units. A positive β_D is an aggregate association, not a causal estimate. Separately, the paired reflection test compares renderings with the same value multiset and identical S_U. A consistent nonzero ΔA under these controls contradicts sufficiency of undirected affinity.

4 Experimental Setup

Datasets and windows. We evaluate 11 public GluonTS datasets (Alexandrov et al., 2020) spanning economic, energy, mobility, and natural domains (Appendix C). Eligible series are finite and at least 512 observations long. A deterministic hash ordering selects at most eight series per dataset,

yielding 71 series. Each contributes 32 evenly spaced forecast origins in its second half, for 2,272 origins. The selection, transformations, and analysis were fixed before any evaluation outcome was computed; hashes and full protocol details are reported in Appendices A and F.

From series to forecast. Each origin supplies $L = 96$ history values and $H = 192$ future values. The history is centered and scaled using its own values, multiplied by a fixed constant of 0.2, and arranged as a $P_d \times \lceil L/P_d \rceil$ texture. Each P_d was an explicit dataset constant in the outcome-blind manifest-generation code and was frozen before any prospective loss was computed; Appendix C lists the values, which are rendering choices rather than claims of a uniquely optimal physical period. The retained records do not establish a single metadata-derived selection rule beyond those frozen constants. Bilinear interpolation maps the texture to the visible columns of an RGB canvas, while the remaining columns are masked. A frozen pretrained model or its architecture-matched random counterpart reconstructs the image. We resize the predicted future region, flatten it in temporal order, and restore the original scale. No model is fine-tuned.

Models and interventions. The primary pair is MAE-Base (111.9M parameters). MAE-Large (329.5M) tests scale, and SimMIM/Swin-Base (89.9M) tests a second architecture family. Each random model has exactly the configuration of its pretrained checkpoint. Random initialization seed 2031 defines the primary analysis; seeds 2032 and 2033 test robustness. The MAE-Base panel contains 47 raw layouts: aligned; cycle, phase, and joint rewiring at four intensities and three transformation seeds; three reflections; and seven held-out families. Each replication panel contains 14 layouts spanning the aligned control, reflections, held-out families, and full rewiring. We average transformation seeds before analysis.

Analysis. Datasets are the scientific and bootstrap units; origins and transformation seeds are not treated as independent observations. The directed-score regression was the registered primary endpoint; reflection contrasts were prespecified secondary tests of the separate topology-sufficiency claim. The primary test uses MAE-Base, MSE, random seed 2031, and 20,000 dataset-cluster bootstrap resamples with seed 5201. It requires $\hat{\beta}_D > 0$, a 95% interval above zero, positive slopes in at least 9/11 datasets, and positive estimates in all 11 leave-one-dataset-out (LODO) fits. Alternative metrics, held-out transformations, replication checkpoints, and random seeds complement but do not redefine this test. We label new pooled displays and decompositions introduced after the registered analysis as post-hoc.

5 RESULTS

Reflections consistently reduce the benefit of pretrained weights even though they leave undirected affinity unchanged. Table 1 leads with this counterexample, then reports the registered directed-score endpoint and its held-out boundary. Branch-completeness checks are reported in Appendix F.

Table 1: Scientific findings across three masked-image checkpoints. Reflection rows are descriptive three-seed summaries of prespecified secondary contrasts and preserve $\Delta S_U = 0$. The registered primary endpoint is the MAE-Base directed-score regression with random seed 2031.

Finding	Statistic	MAE-Base	MAE-Large	SimMIM/Swin
A. Topology-preserving reflections ($\Delta S_U = 0$)	Phase reflection ΔA	-0.302 [-0.539, -0.093]	-0.323 [-0.570, -0.106]	-0.401 [-0.715, -0.128]
	Cycle reflection ΔA	-0.113 [-0.167, -0.064]	-0.148 [-0.241, -0.066]	-0.130 [-0.169, -0.093]
	Both reflections ΔA	-0.383 [-0.600, -0.179]	-0.403 [-0.641, -0.175]	-0.494 [-0.774, -0.241]
	Seed robustness	9/9 CIs < 0	9/9 CIs < 0	9/9 CIs < 0
B. Directed-geometry hypothesis	$\hat{\beta}_D$ [95% CI]	0.169 [0.026, 0.332]	—	—
	Positive datasets	8/11 (required: 9/11)	—	—
	LODO	11/11; min 0.113	—	—
	Cross-dataset criterion	**not met**	—	—
C. Held-out transformations	$\hat{\beta}_D$ [95% CI]	0.076 [-0.054, 0.192]	0.005 [-0.261, 0.206]	0.149 [-0.002, 0.302]

Standard numerical forecasting baselines are reported in Appendix D to provide error-scale context; because their supervised training regimes differ from the frozen visual models, they are not part of the controlled transfer comparison.

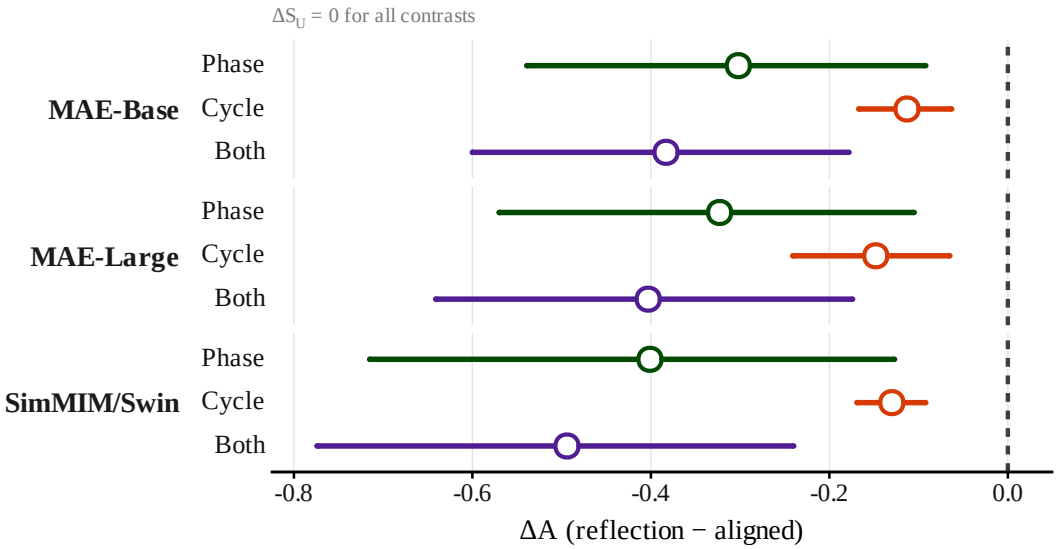


Figure 2: Reflections reduce pretrained-weight advantage across all three checkpoints. Points show mean ΔA relative to the aligned rendering, with dataset-cluster bootstrap 95% intervals after aggregating random losses over seeds 2031–2033. All effects are negative although undirected grid affinity is unchanged ($\Delta S_U = 0$).

5.1 Reflections rule out sufficiency of undirected grid adjacency

All nine checkpoint–reflection contrasts reduced pretrained-weight advantage while keeping $\Delta S_U = 0$ (Figure 2). Mean reductions were 0.302–0.401 for phase reflection, 0.113–0.148 for cycle reflection, and 0.383–0.494 when both axes were reflected. Since $\exp(\Delta A) = (L_R/L_P)_{\text{reflection}}/(L_R/L_P)_{\text{aligned}}$, the reflected layouts retained 61.0–89.3% of the aligned random-to-pretrained loss ratio, a reduction of 10.7–39.0%. Every dataset-cluster interval excluded zero in the three-seed aggregate and for each random-control seed (Appendix Table 12).

Exact enumeration of all 2^{11} dataset-level sign assignments gave two-sided sign-flip $p \leq 0.0352$ for every contrast (Appendix Table 13); this sensitivity analysis was added after the primary analysis. Because A compares pretrained and random models under the same transformation, a layout becoming uniformly harder cannot explain the change. Identical undirected affinity need not preserve the benefit of visual pretraining.

5.2 Reflection acts on pretrained losses, and the effect is broadly distributed

Equation 3 makes the interaction concrete. In a post-hoc decomposition using the same three-seed random aggregation as Figure 2, reflection increased pretrained log loss by 0.114–0.516 across checkpoint–reflection pairs. Random-model changes were much smaller (-0.0005 to 0.0215; Figure 3). Thus the negative ΔA is not an opaque ratio artifact: in these frozen models, reflection disproportionately worsens the pretrained member of the matched pair. The small nonzero random changes, particularly for SimMIM/Swin, are retained rather than assumed away.

The dataset-level MAE-Base view is heterogeneous but not consistent with a single-dataset explanation (Figure 4). Cycle reflection reduces A in all 11 datasets. Phase and joint reflection are negative in 8/11 and 10/11 datasets, respectively, with the largest reductions on Australian Electricity, NN5 Daily, and Pedestrian Counts. Sunspot has a positive phase effect, and Temperature–Rain has a small positive joint effect. These cases remain visible in the display. The corresponding grids for all checkpoints appear in Appendix Figure 5.

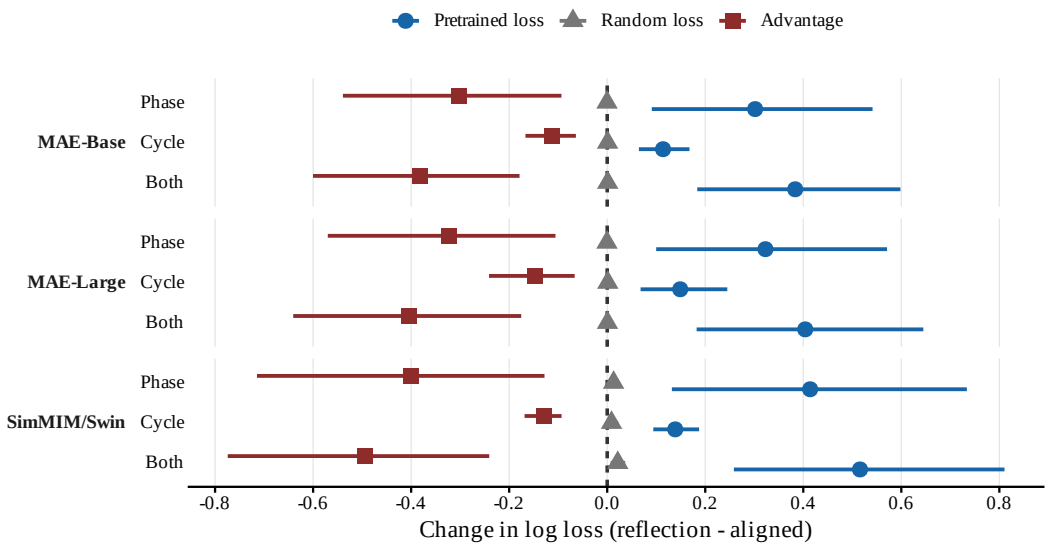


Figure 3: Post-hoc decomposition of the MSE reflection interaction. Positive $\Delta \log L$ means higher loss than the aligned layout. Across checkpoints, pretrained loss increases while random-model loss remains near zero; their difference is the negative ΔA from Figure 2. Points are dataset means and lines are dataset-bootstrap 95% intervals; random losses are geometrically aggregated over seeds 2031–2033.

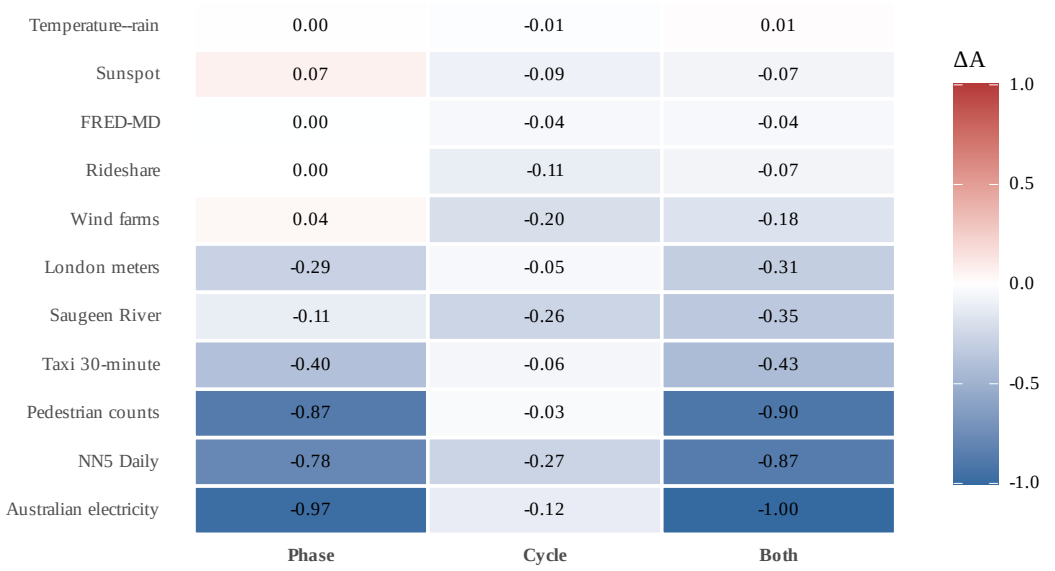


Figure 4: Dataset-level MSE reflection effects for MAE-Base. Cells show post-hoc descriptive ΔA values using the three-seed random aggregation; blue is a reduction and red an increase. The aggregate result is broadest for cycle reflection (11/11 negative), while phase and joint reflections retain visible heterogeneity.

5.3 Directed displacement is not a sufficient replacement

In MAE-Base, pretrained-weight advantage increased with the directed score in aggregate ($\hat{\beta}_D = 0.169$, dataset-cluster 95% interval $[0.026, 0.332]$). The association was not consistent enough across datasets: only 8/11 slopes were positive, below the registered requirement of 9/11, so the

Table 2: Reflection metric robustness. Cells are ranges of ΔA across the three checkpoints using registered random seed 2031. All individual intervals exclude zero.

Metric	Phase ΔA	Cycle ΔA	Both ΔA
MSE	$[-0.386, -0.303]$	$[-0.147, -0.114]$	$[-0.478, -0.383]$
MAE	$[-0.247, -0.184]$	$[-0.100, -0.070]$	$[-0.308, -0.233]$
MASE	$[-0.242, -0.181]$	$[-0.111, -0.072]$	$[-0.319, -0.230]$
sMAPE	$[-0.210, -0.147]$	$[-0.070, -0.045]$	$[-0.256, -0.180]$

criterion failed (Table 1). Appendix Table 9 reports alternative metrics, Appendix Table 11 reports cross-checkpoint aggregates, and Appendix Figure 6 shows the three negative slopes.

5.4 The reflection effect is stable across random controls and metrics

Seeds 2032 and 2033 add 1,650 architecture-matched random-control branches while reusing the same pretrained outputs and inputs. For every seed, all nine checkpoint–reflection intervals are negative (27/27 overall; Appendix Table 12). Thus the reflection pattern is not specific to random seed 2031.

The conclusion is also not MSE-specific. With the registered random seed 2031, every aggregate reflection effect remains negative under MAE, MASE, and sMAPE, and all corresponding dataset-bootstrap intervals exclude zero. Table 2 gives the range across the three checkpoints; alternative metrics attenuate the log-ratio magnitude but preserve its sign and the ordering in which joint reflection is largest and cycle reflection smallest. Appendix Table 10 reports every estimate and interval. Reflection and alternative-metric analyses were prespecified secondary families; their joint tabulation and bootstrap computation are post-hoc additions generated from saved losses.

5.5 Held-out transformations leave the mechanism unresolved

The seven transformation families excluded from score construction did not support generalization. Three-seed held-out-family slopes were 0.076 $[-0.054, 0.192]$ for MAE-Base, 0.005 $[-0.261, 0.206]$ for MAE-Large, and 0.149 $[-0.002, 0.302]$ for SimMIM/Swin. All intervals included zero.

Dataset-held-out prediction also failed to identify a replacement mechanism (Table 3). Relative to an intensity-only model, the directed score gave MSE ratios of 1.038, 1.085, and 1.019 and improved none of the three models. Absolute-position affinity performed modestly better, with ratios of 0.959, 0.959, and 0.954. Results for the combined directed, position, and boundary model depended on the architecture (0.963, 1.034, 0.939). Condition numbers of 1.59–1.85 and variance-inflation factors below 1.41 exclude severe numerical collinearity, but these correlated predictors cannot establish causality. Appendix Table 14 reproduces the values.

Table 3: No candidate geometry gives a consistent replacement explanation under dataset-held-out prediction. Entries are three-seed predictive MSE ratios relative to an intensity-only model; lower is better.

Predictor	MAE-Base	MAE-Large	SimMIM/Swin
S_D	1.038	1.085	1.019
S_U	1.146	1.246	1.218
S_P	0.959	0.959	0.954
S_B	1.148	1.139	1.169
Changed-index fraction	0.997	1.004	1.011
S_D, S_P, S_B	0.963	1.034	0.939

6 DISCUSSION

The experiment establishes that undirected grid topology is not sufficient for the transfer benefit in these rendered time series. Reflected and aligned layouts have identical values and undirected affinity, yet their pretrained-weight advantages differ. The decomposition sharpens the interpretation: pretrained losses increase under reflection while matched random losses remain nearly fixed. The recurrence across MAE-Base, MAE-Large, SimMIM/Swin, three random seeds, and four metrics argues against an isolated checkpoint, control seed, or loss definition, while not establishing universality beyond the evaluated masked-image families.

What replaces undirected topology remains unresolved. Phase reflection changes directed displacement and absolute row position while preserving horizontal boundary distance. Cycle reflection changes directed displacement and boundary-relative position; joint reflection changes all three. The directed score has a positive aggregate slope, but fails the registered dataset-consistency criterion and yields intervals containing zero on held-out transformation families. Absolute-position affinity performs modestly better in dataset-held-out prediction, but no candidate is consistently predictive across architectures. These correlated scores cannot identify a unique mechanism without interventions that vary temporal orientation, canvas position, and boundary distance independently.

More broadly, a visual rendering is not merely a lossless or topology-preserving encoding of a sequence. It places temporal relationships into the coordinate system on which a vision model was pretrained. Local adjacency may still be useful for forecasting, but preserving the observations and their undirected local relations does not preserve the contribution of pretrained visual weights. Future sequence-to-image studies should therefore separate two questions: whether a rendering is intrinsically easy for an architecture, and whether it presents structure in a form that pretrained weights know how to exploit. The matched pretrained/random interaction supports this distinction directly. Our evidence concerns masked-image reconstruction of univariate value textures on the public datasets studied here; other visual encodings and pretraining schemes may rely on different structure. The conclusion is conditional on the frozen period-based rendering family, and sensitivity to alternative period choices remains unresolved.

7 CONCLUSION

Value-preserving reflections reduce pretrained-weight advantage while leaving undirected grid adjacency unchanged because they disproportionately increase pretrained, rather than random-model, loss. Cross-modal transfer therefore depends on more than which temporal relationships a rendering preserves: it is sensitive to how those relationships are embedded in the pretrained visual coordinate system. The registered directed score fails its criterion and does not generalize to held-out transformations, so direction, absolute position, and boundary anchoring remain unresolved.

AI USE STATEMENT

Generative AI tools assisted with conceptual framing, experimental-design review, code development and debugging, literature discovery, manuscript editing, reference formatting, and artifact preparation. AI-generated synthetic inputs were used only for software smoke tests, not for scientific analysis; the tools did not generate evaluation data or alter model outputs. The authors verified the protocol, code, results, sources, claims, and prose and take responsibility for the manuscript and artifacts.

ETHICS STATEMENT

This study uses public benchmark time series and released pretrained checkpoints; it involves neither human participants nor deployment. Forecasts in energy or mobility require domain validation before operational use. The main computational cost is inference through large vision backbones, so model outputs were cached to avoid repeated inference.

Reproducibility statement

Appendix A specifies selection, rendering, transformations, scores, and analyses. Appendix C lists the dataset units; Appendix D gives numerical-baseline configurations and per-dataset losses; and Appendix F records checkpoint revisions, seeds, branch counts, and hashes. The supplement provides the manifest, selected series and origins, code, metadata, and sufficient statistics needed to reproduce every table. No private data or outcome-tuned hyperparameters are used.

References

Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C. Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, Lorenzo Stella, Ali Caner Türkmen, and Yuyang Wang. GluonTS: Probabilistic and neural time series modeling in python. *Journal of Machine Learning Research*, 21(116):1–6, 2020. URL https://www.jmlr.org/papers/v21/19-820.html.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=gerNCVqqtR.

Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. VisionTS: Visual masked autoencoders are free-lunch zero-shot time series forecasters. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8979–9007. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25be.html.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 2990–2999. PMLR, 2016. URL https://proceedings.mlr.press/v48/cohenc16.html.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10148–10167. PMLR, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4918–4927, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/He_Rethinking_ImageNet_Pre-Training_ICCV_2019_paper.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2661–2671, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Kornblith_Do_Better_ImageNet_Models_Transfer_Better_CVPR_2019_paper.html.

Weitao Li and Gong Cheng. Generator-aligned representation interfaces for diagnostic soft equivariance. *arXiv preprint arXiv:2607.25988*, 2026. URL https://arxiv.org/abs/2607.25988.

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://proceedings.iclr.cc/paper_files/paper/2024/hash/2ea18fdc667e0ef2ad82b2b4d65147ad-Abstract-Conference.html.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 512–523. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/0607f4c705595b911a4f3e7a127b44e0-Abstract.html.

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.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/eb1e78328c46506b46a4ac4a1e378b91-Abstract.html.

Roland Riachi, Kashif Rasul, Arjun Ashok, Prateek Humane, Alexis Roger, Andrew R. Williams, Yuriy Nevmyvaka, and Irina Rish. Random initialization can't catch up: The advantage of language model transfer for time series forecasting. *arXiv preprint arXiv:2506.21570*, 2025. URL https://arxiv.org/abs/2506.21570.

ChengAo Shen, Wenchao Yu, Ziming Zhao, Dongjin Song, Wei Cheng, Haifeng Chen, and Jingchao Ni. Multi-modal view enhanced large vision models for long-term time series forecasting. In *Advances in Neural Information Processing Systems*, volume 38, pp. 93269–93296, 2025a. doi: 10.52202/085713-3119. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/860c1c657deafe09f64c013c2888bd7b-Abstract-Conference.html.

Lefei Shen, Mouxiang Chen, Xu Liu, Han Fu, Xiaoxue Ren, Jianling Sun, Zhuo Li, and Chenghao Liu. VisionTS++: Cross-modal time series foundation model with continual pre-trained vision backbones. *arXiv preprint arXiv:2508.04379*, 2025b. URL https://arxiv.org/abs/2508.04379.

Moein Sorkhei, Christos Matsoukas, Johan Fredin Haslum, Emir Konuk, and Kevin Smith. k-NN as a simple and effective estimator of transferability. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=hGlkjP1zHc.

Till Speicher, Vedant Nanda, and Krishna P. Gummadi. Understanding the role of invariance in transfer learning. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=spJI4LSPIU.

Kai Sheng Tai, Peter Bailis, and Gregory Valiant. Equivariant transformer networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 6086–6095. PMLR, 2019. URL https://proceedings.mlr.press/v97/tai19a.html.

Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? In *Advances in Neural Information*

Processing Systems, volume 37, pp. 60162–60191. Curran Associates, Inc., 2024. doi: 10.52202/079017-1922. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6ed5bf446f59e2c6646d23058c86424b-Abstract-Conference.html.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 53140–53164. PMLR, 2024. URL https://proceedings.mlr.press/v235/woo24a.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. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10033–10041, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Wu_Rethinking_and_Improving_Relative_Position_Encoding_for_Vision_Transformer_ICCV_2021_paper.html.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Xie_SimMIM_A_Simple_Framework_for_Masked_Image_Modeling_CVPR_2022_paper.html.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, volume 27, pp. 3320–3328, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/532a2f85b6977104bc93f8580abbb330-Abstract.html.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Mingrui Zhang, Hanchen Yang, Wengen Li, Xudong Jiang, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. SSDA: Bridging spectral and structural gaps via dual adaptation for vision-based time series forecasting. *arXiv preprint arXiv:2605.12550*, 2026. URL https://arxiv.org/abs/2605.12550.

Ziming Zhao, ChengAo Shen, Hanghang Tong, Dongjin Song, Zhigang Deng, Qingsong Wen, and Jingchao Ni. From images to signals: Are large vision models useful for time series analysis? *arXiv preprint arXiv:2505.24030*, 2025. URL https://arxiv.org/abs/2505.24030.

A Protocol and Analysis Details

A.1 Discovery-to-confirmation separation

The v1 discovery protocol evaluated six datasets and an undirected compatibility score. Its registered all-layout Spearman correlation was 0.400, below the 0.50 gate, so the study did not advance. Reflection failures observed in v1 motivated S_D, whose post-hoc discovery correlation was 0.638. Neither value is confirmatory because S_D was defined after inspecting the outcomes. We froze the v2 manifest on 27 August 2026, before inspecting any v2 pretrained-versus-random forecasting result. The supplementary protocol retains the v1 failure, the post-hoc origin of S_D, and the v2 decision rule.

A.2 Series selection and origins

Eligible targets contain only finite values and have length at least 512. At most eight series per dataset are selected in ascending order of `SHA-256('SAFE-Route-v2-series-v1' | dataset | repository-index)`. Datasets with fewer than eight eligible series retain all of them. Only the first $\lfloor n/2 \rfloor$ observations are used to estimate lag weights. The 32 origins are distinct rounded integers on an even grid from $\max(96, \lfloor n/2 \rfloor)$ through $n - 192$. The protocol does not permit removing a candidate on the basis of model loss.

A.3 Rendering transformations

The aligned source grid has period rows and $\lceil 96/P_d \rceil$ cycle columns. Cycle rewiring permutes columns, phase rewiring permutes rows, and joint rewiring nests both operations. Intensities 0.25, 0.50, 0.75, and 1.00 change a frozen fraction of locations under transformation seeds 4101–4103. The reflections reverse rows, columns, or both. Held-out families are row roll by one, row roll by a quarter period, column roll by one, column roll by half, temporal roll by one, temporal roll by one period, and a serpentine row scan. Padding uses a deterministic source index, and every mapping is hashed.

B Score Definitions

For lag ℓ, let $D_t^{(T,\ell)} = \|p_{t+\ell}^{(T)} - p_t^{(T)}\|_1$. With exponential temperature one grid cell, the undirected affinity is

$$S_U(T; d) = \sum_{\ell \in \mathcal{L}_d} w_d(\ell) \frac{1}{L - \ell} \sum_t \exp[-\max\{D_t^{(T,\ell)} - 1, 0\}]. \tag{5}$$

The absolute-position affinity is

$$S_P(T) = \frac{1}{L} \sum_t \exp[-\|p_t^{(T)} - p_t^{(T_0)}\|_1]. \tag{6}$$

If the last visible history column is $C - 1$, the boundary-anchor affinity is

$$S_B(T) = \frac{1}{L} \sum_t \exp\{-|[(C-1) - c_t^{(T)}] - [(C-1) - c_t^{(T_0)}]|\}. \tag{7}$$

No score coefficient is learned from v1 or v2 forecasting outcomes. Nominal transformation intensity and changed-index fraction are outcome-independent baselines.

C Datasets and Models

Table 4: Dataset-level units. Series counts follow the hash selection rule, and every retained series contributes 32 origins.

Dataset	Period	Series	Dataset	Period	Series
NN5 Daily	7	8	Saugeen River	7	1
London Smart Meters	48	8	Sunspot	7	1
Wind Farms	60	8	Temperature–Rain	7	8
Pedestrian Counts	24	8	Rideshare	24	8
FRED-MD	12	8	Australian Electricity	48	5
Taxi 30-minute	48	8			

Table 5: Model pairs. Each architecture-matched random initialization uses the pretrained checkpoint's exact configuration.

Checkpoint	Parameters	Image/patch	Role
ViT-MAE-Base	111.9M	224/16	primary
ViT-MAE-Large	329.5M	224/16	scale replication
Swin-Base SimMIM	89.9M	192/4	family replication

D POST-HOC FORECASTING CONTEXT

The numerical baselines in Table 6 were added after the confirmatory analysis to show the scale of the forecasting errors. Seasonal Naive repeats the final dataset-specific period within each history window. DLinear (Zeng et al., 2023), PatchTST (Nie et al., 2023), and iTransformer (Liu et al., 2024) use implementations vendored with the VisionTS evaluation repository. Each model is trained separately for each dataset on pooled windows from the selected series. Training targets end by $\lfloor n/2 \rfloor$, and all 32 evaluation origins per series begin at or after that cutoff. The numerical and visual evaluations therefore share the same 11 datasets, 71 series, 2,272 origins, context length 96, and horizon 192.

All numerical models use seed 2031, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch size 128, and six fixed epochs. We use no validation set, early stopping, or per-dataset hyperparameter tuning, and retain up to 512 evenly spaced training windows per series. Each window is normalized by the mean and standard deviation of its history and restored before loss evaluation. DLinear uses a moving-average width of 25. PatchTST uses $d_{\text{model}} = 128$, $d_{\text{ff}} = 256$, eight heads, two encoder layers, patch length 16, stride 8, and dropout 0.1. iTransformer uses the same width, feed-forward size, number of heads, layers, and dropout. The artifacts record source and runner SHA-256 hashes, library versions, training-window counts, and unrounded losses. Because these supervised regimes are not matched to the frozen visual models, the comparison is not an architecture ranking.

Table 6: Post-hoc forecasting-error scale on the evaluation windows. Relative MSE and MAE are geometric means of per-dataset ratios to Seasonal Naive (lower is better). The supervised numerical models and frozen visual models have different training regimes and are not directly ranked.

Method	Rel. MSE	Rel. MAE	Regime
Seasonal Naive	1.000	1.000	deterministic
DLinear	0.814	0.955	supervised, post-hoc
PatchTST	1.321	1.172	supervised, post-hoc
iTransformer	1.143	1.128	supervised, post-hoc
MAE-Base (pretrained)	0.984	1.060	frozen visual
MAE-Large (pretrained)	1.046	1.089	frozen visual
SimMIM/Swin (pretrained)	0.901	0.965	frozen visual

Table 7: Per-dataset MSE for the post-hoc forecasting context. Visual entries are aligned-rendering pretrained checkpoints; numerical models are supervised as documented above. Raw errors are not averaged across heterogeneous dataset scales.

Dataset	Seasonal Naive	DLinear	PatchTST	iTransformer	MAE-Base	MAE-Large	SimMIM/Swin
NN5 Daily	51.909	44.894	40.849	42.018	55.349	66.994	42.022
London Smart Meters	0.0598	0.0392	0.0394	0.0393	0.0482	0.0484	0.0448
Wind Farms	902.047	723.829	7629.1	6389.1	928.611	927.781	919.574
Pedestrian Counts	4.728×10^5	3.166×10^5	3.201×10^5	3.264×10^5	5.934×10^5	5.578×10^5	4.614×10^5
FRED-MD	1.268×10^6	1.317×10^6	1.330×10^6	1.586×10^6	1.530×10^6	1.536×10^6	1.467×10^6
Saugeen River	1804.3	2360.8	20808.2	7488.1	2658.5	2075.3	2176.4
Sunspot	2993.3	2131.1	2256.5	2219.4	2496.2	4041.7	2486.7
Temperature–Rain	308.023	160.329	175.095	166.336	170.568	216.206	165.097
Rideshare	125.300	124.832	219.289	132.107	117.211	116.753	128.738
Australian Electricity	3.356×10^5	2.928×10^5	2.793×10^5	2.887×10^5	3.837×10^5	3.947×10^5	3.338×10^5
Taxi 30-minute	60.219	47.006	57.431	53.434	51.592	53.861	49.976

Table 8: Per-dataset MAE for the same post-hoc forecasting context. The normalized main-text summary first forms a ratio to Seasonal Naive within each dataset and then takes the geometric mean of the 11 ratios.

Dataset	Seasonal Naive	DLinear	PatchTST	iTransformer	MAE-Base	MAE-Large	SimMIM/Swin
NN5 Daily	4.791	4.675	4.404	4.438	5.235	5.856	4.370
London Smart Meters	0.1274	0.1197	0.1221	0.1200	0.1215	0.1234	0.1095
Wind Farms	9.497	8.613	23.609	24.763	9.784	9.740	9.749
Pedestrian Counts	316.521	298.248	315.926	313.827	443.286	426.406	336.827
FRED-MD	419.088	420.030	362.490	444.876	462.104	455.961	454.683
Saugeen River	24.527	30.052	82.240	57.373	33.963	28.657	28.100
Sunspot	39.866	34.267	35.047	35.076	36.385	45.686	36.127
Temperature–Rain	9.011	6.850	6.865	6.642	7.419	8.596	6.641
Rideshare	5.332	5.548	7.354	5.781	5.299	5.297	5.471
Australian Electricity	350.114	341.668	319.098	324.083	399.928	400.329	360.371
Taxi 30-minute	4.756	4.510	5.106	4.920	4.588	4.688	4.297

E Additional Results

E.1 Directed-score metric and checkpoint robustness

Table 9: MAE-Base primary-score robustness. Intervals use the same dataset-cluster bootstrap.

Metric	Slope	95% interval	Positive datasets
MSE (primary)	0.169	$[0.026, 0.332]$	8/11
MAE	0.115	$[0.025, 0.216]$	8/11
MASE	0.116	$[0.027, 0.216]$	8/11
sMAPE	0.099	$[0.029, 0.184]$	9/11

E.2 Reflection metric robustness and dataset heterogeneity

Table 10: Reflection robustness across metrics with registered random seed 2031. Cells give mean ΔA and dataset-bootstrap 95% intervals. Reflection and alternative-metric analyses were prespecified secondary families; this joint tabulation and its deterministic bootstrap computation were produced post hoc from saved branch losses.

Checkpoint	Metric	Phase	Cycle	Both
MAE-Base	MSE	$-0.303\,[-0.535, -0.091]$	$-0.114\,[-0.169, -0.064]$	$-0.383\,[-0.604, -0.182]$
	MAE	$-0.184\,[-0.320, -0.058]$	$-0.070\,[-0.106, -0.035]$	$-0.233\,[-0.360, -0.114]$
	MASE	$-0.181\,[-0.314, -0.061]$	$-0.072\,[-0.113, -0.032]$	$-0.230\,[-0.354, -0.112]$
	sMAPE	$-0.147\,[-0.261, -0.045]$	$-0.045\,[-0.075, -0.021]$	$-0.180\,[-0.293, -0.079]$
MAE-Large	MSE	$-0.323\,[-0.573, -0.103]$	$-0.147\,[-0.242, -0.067]$	$-0.403\,[-0.641, -0.177]$
	MAE	$-0.198\,[-0.348, -0.063]$	$-0.100\,[-0.160, -0.048]$	$-0.259\,[-0.393, -0.127]$
	MASE	$-0.195\,[-0.338, -0.065]$	$-0.111\,[-0.173, -0.054]$	$-0.264\,[-0.391, -0.138]$
	sMAPE	$-0.154\,[-0.272, -0.048]$	$-0.070\,[-0.114, -0.031]$	$-0.196\,[-0.307, -0.094]$
SimMIM/Swin	MSE	$-0.386\,[-0.687, -0.112]$	$-0.129\,[-0.166, -0.094]$	$-0.478\,[-0.755, -0.237]$
	MAE	$-0.247\,[-0.433, -0.083]$	$-0.084\,[-0.119, -0.049]$	$-0.308\,[-0.475, -0.157]$
	MASE	$-0.242\,[-0.420, -0.080]$	$-0.099\,[-0.165, -0.045]$	$-0.319\,[-0.477, -0.171]$
	sMAPE	$-0.210\,[-0.363, -0.072]$	$-0.068\,[-0.117, -0.028]$	$-0.256\,[-0.404, -0.116]$

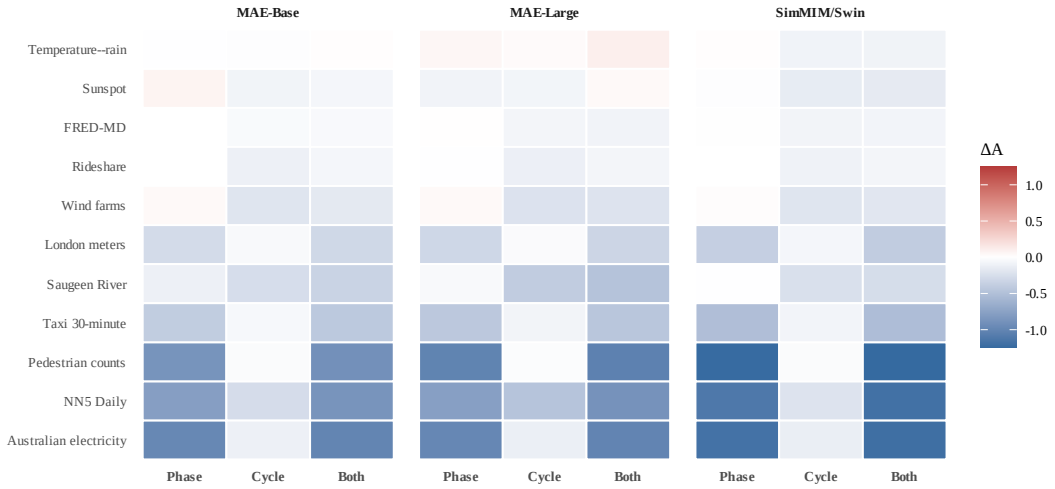


Figure 5: Post-hoc descriptive dataset-level MSE reflection effects for all three checkpoints using the three-seed random aggregation. Every dataset and sign is retained. The magnitude varies substantially, but cycle reflection is the broadest pattern and the aggregate effects in Figure 2 are not attributable to a single dataset.

E.3 Dataset-level heterogeneity

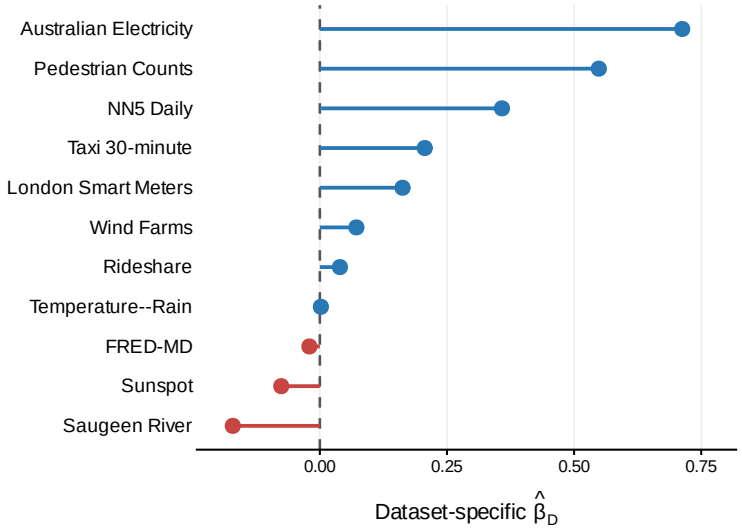


Figure 6: The directed-score association varies by dataset. Eight of 11 MAE-Base slopes are positive; FRED-MD, Sunspot, and Saugeen River are negative. The registered analysis uses MSE and random seed 2031.

Table 11: Directed-score slopes after aggregating random-control seeds 2031–2033. The primary analysis uses MAE-Base, MSE, and seed 2031.

Model	Slope	95% interval	Positive datasets
MAE-Base	0.169	$[0.028, 0.331]$	8/11
MAE-Large	0.149	$[-0.038, 0.348]$	8/11
SimMIM/Swin	0.257	$[0.078, 0.463]$	9/11

E.4 PER-SEED REFLECTION CONTRASTS

Table 12: Reflection effects are negative for every checkpoint and random-control seed. Cells report mean ΔA with dataset-cluster bootstrap 95% intervals; all 27 intervals exclude zero.

Checkpoint	Seed	Phase	Cycle	Both
MAE-Base	2031	-0.303 [-0.537, -0.089]	-0.114 [-0.168, -0.065]	-0.383 [-0.603, -0.184]
	2032	-0.302 [-0.536, -0.088]	-0.113 [-0.168, -0.064]	-0.382 [-0.601, -0.183]
	2033	-0.302 [-0.534, -0.088]	-0.113 [-0.167, -0.063]	-0.382 [-0.603, -0.183]
MAE-Large	2031	-0.323 [-0.570, -0.099]	-0.147 [-0.241, -0.066]	-0.403 [-0.641, -0.183]
	2032	-0.323 [-0.570, -0.099]	-0.149 [-0.243, -0.067]	-0.403 [-0.641, -0.183]
	2033	-0.324 [-0.569, -0.101]	-0.147 [-0.243, -0.066]	-0.404 [-0.645, -0.182]
SimMIM/Swin	2031	-0.386 [-0.682, -0.114]	-0.129 [-0.166, -0.094]	-0.478 [-0.750, -0.237]
	2032	-0.403 [-0.713, -0.124]	-0.125 [-0.167, -0.085]	-0.492 [-0.776, -0.238]
	2033	-0.414 [-0.741, -0.121]	-0.135 [-0.180, -0.094]	-0.512 [-0.805, -0.255]

E.5 POST-HOC FINITE-SAMPLE SENSITIVITY

Table 13: Exact dataset-level sign-flip sensitivity for the three-seed reflection means. Cells are unadjusted two-sided p values from all $2^{11} = 2{,}048$ sign assignments under the sharp zero-effect null and dataset-level sign exchangeability.

Checkpoint	Phase	Cycle	Both
MAE-Base	0.0254	0.0010	0.0020
MAE-Large	0.0273	0.0029	0.0078
SimMIM/Swin	0.0352	0.0010	0.0010

E.6 GENERALIZATION AND RIVAL PREDICTORS

Table 14: Three-seed dataset-held-out predictive MSE relative to intensity-only regression. Values below one improve on intensity.

Predictor	MAE-Base	MAE-Large	SimMIM/Swin
S_D	1.038	1.085	1.019
S_U	1.146	1.246	1.218
S_P	0.959	0.959	0.954
S_B	1.148	1.139	1.169
Changed-index fraction	0.997	1.004	1.011
S_D, S_P, S_B	0.963	1.034	0.939

F REPRODUCIBILITY AND INTEGRITY CHECKS

The frozen manifest has SHA-256 `de38480e18ebb2e1feb1993b763cd6cf273457d99c6d12c9e48e42bfb446a22e`. Checkpoint revisions are `25b184bea5538bf5c4c852c79d221195fdd2778d` (MAE-Base), `142cb8c25e1b1bc1769997a919aa1b5a2345a6b8` (MAE-Large), and `d5c0bac35edfdcf83e0c9ffe063435970651990a` (SimMIM/Swin). The primary and replication run completed all 1,650 expected branches. The seed-2032/2033 robustness run completed all 1,650 additional random branches. No missing or unexpected branch entered the analysis. A failed path-resolution launch was archived separately and excluded from all summaries and inference.

G SCOPE OF THE CLAIM

The primary panel supports a positive aggregate association between S_D and pretrained-weight advantage, while the reflection controls show that undirected affinity is insufficient under the evaluated protocol. The results do not confirm S_D prospectively, establish generalization to unseen transformation families, identify direction as the unique mechanism, or show universality across vision architectures. Routing, fusion, and deployed system utility are outside the study.