CIRCUITTS: DISCOVERING FUNCTIONAL TEMPORAL CIRCUITS IN TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Time Series Foundation Models (TSFMs) achieve strong zero-shot forecasting, yet their internal temporal mechanisms remain poorly understood. Existing TSF interpretability mainly reveals *what* temporal information is encoded, while conventional circuit analysis identifies *where* a behavior is supported; neither jointly characterizes a temporal function by its causal support and internal state. We introduce **CircuitTS**, which defines a functional temporal circuit as a computational support together with its function-specific causal state. CircuitTS comprises two components. **Counterfactual Circuit Discovery** constructs controlled counterfactuals for Period, Trend, and Copy, localizes candidate supports through intervention responsiveness, function selectivity, and behavioral alignment, and verifies them through sufficiency, necessity, and specificity. **Circuit Utilization** evaluates cross-dataset generalization and aligns causal states for cross-model transplantation. Experiments on three heterogeneous TSFMs and nine real-world benchmarks show that temporal functions can exhibit different anatomical realizations while retaining causal roles, generalize to unseen datasets, and transfer across independently pretrained models. CircuitTS thus moves TSFM interpretability from representation analysis toward function-centered causal mechanisms. Codes are available at `https://anonymous.4open.science/r/circuitts-F2F0/`.

1 INTRODUCTION

Time Series Foundation Models (TSFMs) have demonstrated strong zero-shot forecasting across heterogeneous domains (Woo et al., 2024; Liu et al., 2024; Goswami et al., 2024; He et al., 2026; Fu et al., 2026), yet how their predictions are internally computed remains poorly understood. Existing TSF interpretability mainly studies input relevance or whether pretrained representations encode temporal concepts such as periodicity and trend (Kacprzyk et al., 2024; Wiliński et al., 2025). Moving beyond representation analysis requires identifying the internal mechanisms that causally implement temporal computation.

Mechanistic interpretability offers two complementary perspectives. Circuit discovery localizes the internal components that causally support a behavior, providing a view of *where* computation occurs (Conmy et al., 2023; Wang et al., 2023; Marks et al., 2025). In contrast, work on function vectors shows that task-level functions can be captured by compact causal internal states that can be reused to trigger corresponding behavior, providing a view of *how* a function is internally represented (Todd et al., 2024). As illustrated in Fig. 1, these views characterize complementary aspects of a mechanism, but neither alone jointly specifies its causal support and function-specific state.

Bridging these two views raises two challenges: how to quantitatively identify the computational support and causal state of a temporal function, and how to reuse them across datasets and heterogeneous model representations. To address the first challenge, we introduce the *functional temporal circuit*, defined by a computational support s_ϕ indicating *where* the function is implemented and a function-specific causal state γ_ϕ describing *how* it is causally expressed. We operationalize γ_ϕ through function-specific displacements induced by controlled temporal interventions (Geiger et al., 2021; Meng et al., 2022), allowing the same functional identity to admit different anatomical realizations across architectures.

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.

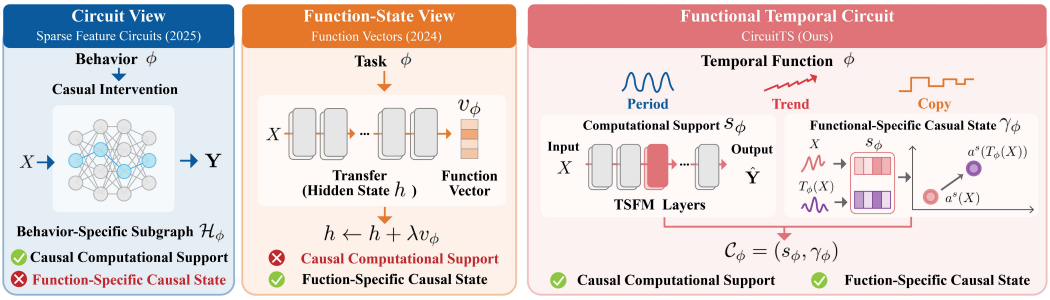

Figure 1: Three views of internal mechanisms. (a) The **Circuit View** identifies *where* a behavior is causally supported. (b) The **Function-State View** identifies *how* a function is represented as an internal state. (c) **CircuitTS** combines both by defining a functional temporal circuit as a computational support s_ϕ together with a function-specific causal state γ_ϕ.

Building on this formulation, we propose **CircuitTS** to identify functional temporal circuits in frozen TSFMs. CircuitTS constructs function-conditioned counterfactuals by selectively suppressing the defining historical evidence of each temporal function while preserving other temporal structure (Geiger et al., 2022; Zhang & Nanda, 2024). We instantiate this principle on **Period**, **Trend**, and **Copy**, covering recurrence, directional extrapolation, and local persistence (Oreshkin et al., 2020; Wiliński et al., 2025). From these counterfactuals, CircuitTS localizes candidate supports using Intervention Responsiveness, Function Selectivity, and Behavioral Alignment, and verifies them through causal sufficiency, counterfactual necessity, and functional specificity to obtain s_ϕ and its associated state family γ_ϕ.

To address the second challenge, we further study two forms of circuit utilization: reusing discovered supports on unseen datasets and aligning function-specific causal states across independently pretrained TSFMs. Experiments on three heterogeneous TSFMs and nine real-world datasets show that supports discovered without access to a target dataset remain causally effective on unseen data, while aligned causal states can recover disrupted forecasts across models. These results suggest that temporal mechanisms can preserve functional correspondence despite different internal realizations.

Our contributions are threefold:

- We introduce *functional temporal circuits*, jointly characterized by computational support and function-specific causal state, unifying the *where* and *how* of temporal computation.
- We propose **CircuitTS**, a function-conditioned counterfactual framework that discovers Period, Trend, and Copy circuits through function-contrastive localization based on Intervention Responsiveness, Function Selectivity, and Behavioral Alignment, followed by causal verification.
- We demonstrate two forms of circuit utilization across heterogeneous TSFMs and real-world datasets: cross-dataset support generalization and cross-model causal-state transplantation.

2 Related Work

2.1 Time Series Foundation Models and Interpretability

Recent advances in large-scale pretraining have led to a rapidly growing family of TSFMs (Rasul et al., 2023; Das et al., 2024; Woo et al., 2024; Goswami et al., 2024). Existing TSFMs broadly follow different modeling paradigms, including tokenization-based forecasting, generative sequence modeling, and lightweight structure-aware modeling. Chronos (Ansari et al., 2024), Timer (Liu et al., 2024), and SEMPO (He et al., 2026) are representative examples of these directions. Despite substantial differences in architecture and pretraining, they collectively demonstrate that transferable forecasting knowledge can emerge from large and heterogeneous time-series corpora. Yet, understanding what temporal concepts and computations are internally learned and reused remains underexplored (Wiliński et al., 2025; Pandey et al., 2025).

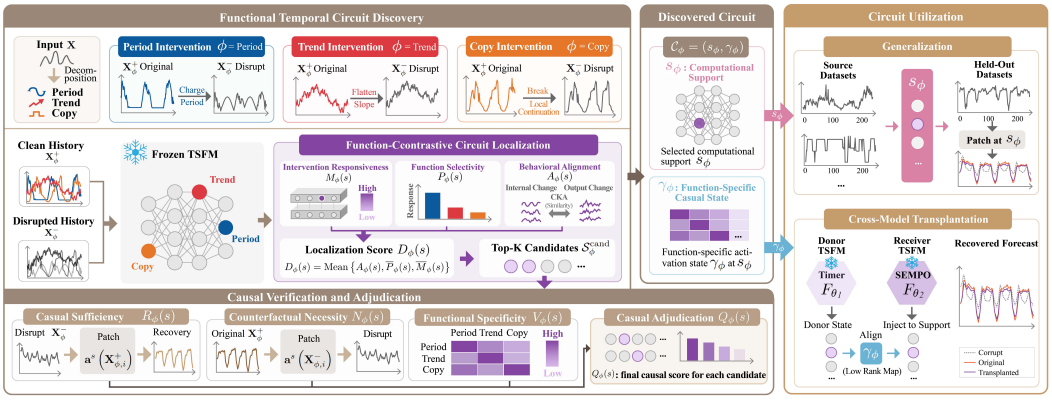

Figure 2: Overview of CircuitTS. The left panel discovers a functional temporal circuit through function-specific counterfactuals, function-contrastive localization, and causal verification. The middle panel summarizes the discovered circuit $\mathcal{C}_\phi = (s_\phi, \gamma_\phi)$. The right panel reuses s_ϕ for cross-dataset generalization and aligns γ_ϕ for cross-model transplantation.

Existing TSF interpretability methods can likewise be grouped into several levels. Prediction-level methods explain how historical observations influence forecasts, such as TIMEVIEW (Kacprzyk et al., 2024). Representation-level studies examine what temporal concepts are encoded inside TSFMs, with (Wiliński et al., 2025) identifying steerable representations related to periodicity and trend. More recent feature-level approaches use sparse representations to uncover and causally intervene on interpretable latent features (Mishra, 2026; Lakhanpal, 2026). These methods respectively explain input relevance, encoded concepts, and interpretable features, whereas CircuitTS focuses on the circuit level: identifying the computational supports and causal states through which temporal functions are implemented.

2.2 Circuit Discovery

Mechanistic interpretability explains model behavior through the internal computations that causally support it. Existing circuit studies span localization-based methods that identify behaviorally important components and connections (Conmy et al., 2023; Zhang & Nanda, 2024), feature-level causal circuits built from sparse interpretable features (Marks et al., 2025), and distributed or reusable mechanisms shared across behaviors (Yao et al., 2024; Todd et al., 2024; Merullo et al., 2024). Related concept-level work in vision further traces and describes concept circuits, with CoE explicitly addressing the polysemanticity of internal visual concepts (Yu et al., 2025). Together, these studies show that circuit explanations can operate at multiple levels of abstraction, from components to features and semantic concepts.

However, causal support alone need not provide a unique or complete description of the underlying mechanism. Patching results can depend on intervention design (Zhang & Nanda, 2024), causally effective subspaces may differ from the mechanisms naturally used by the model (Makelov et al., 2024), and multiple internal realizations may support the same behavior (Méloux et al., 2025). CircuitTS addresses this ambiguity by pairing the computational support of a temporal function with its function-specific causal state. This formulation enables temporal mechanisms to be compared across datasets and heterogeneous TSFMs even when their anatomical realizations differ.

3 CircuitTS

As illustrated in Fig. 2, CircuitTS first discovers a functional temporal circuit $\mathcal{C}_\phi = (s_\phi, \gamma_\phi)$ through function-specific counterfactuals, function-contrastive localization, and causal verification, and then utilizes the discovered support s_ϕ and causal state γ_ϕ for cross-dataset generalization and cross-model transplantation. Secs. 3.1–3.3 describe circuit discovery, while Sec. 3.4 presents cross-dataset support generalization and cross-model causal-state transplantation.

3.1 FUNCTIONAL TEMPORAL CIRCUITS

Consider a pretrained TSF model parameterized by θ,

$$\hat{\mathbf{Y}} = F_\theta(\mathbf{X}), \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{N\times C}$ is an input sequence with look-back length N and C variables, and $\hat{\mathbf{Y}} \in \mathbb{R}^{H\times C}$ is the forecast over horizon H. We consider a set of architecture-compatible internal sites $\mathcal{S} = \{s_1, \ldots, s_M\}$ whose activations can be observed and intervened on, and denote the activation at site s by $\mathbf{a}^s(\mathbf{X})$.

We study three representative temporal forecasting functions,

$$\Phi = \{\text{Period}, \text{Trend}, \text{Copy}\}. \tag{2}$$

Period captures continuation of recurring temporal structure, Trend captures extrapolation of the local direction of change, and Copy captures continuation of a recent local pattern. Following controlled counterfactual intervention paradigms used for causal localization (Geiger et al., 2022; Zhang & Nanda, 2024), we construct a history-only intervention T_ϕ that selectively suppresses the evidence associated with each function. Related TSFM studies have likewise shown that temporal concepts such as periodicity and trend can be manipulated through targeted interventions (Wiliński et al., 2025). Specifically, Period removes the characteristic-period harmonic, Trend removes the fitted local linear component, and Copy replaces the latest local block with its immediately preceding block. None of these interventions accesses future targets.

For a candidate site s, we define the function-specific state displacement induced by T_ϕ as

$$\Delta^s_\phi(\mathbf{X}) = \mathbf{a}^s(\mathbf{X}) - \mathbf{a}^s\left(T_\phi(\mathbf{X})\right). \tag{3}$$

The displacement captures how the internal state changes when temporal evidence associated with ϕ is selectively disrupted.

Given the computational support s_ϕ identified for function ϕ, we define its function-specific causal state as the input-conditioned displacement family

$$\gamma_\phi = \left\{ \Delta^{s_\phi}_\phi(\mathbf{X}) \;\middle|\; \mathbf{X} \in \mathcal{X} \right\}, \tag{4}$$

where $\mathcal{X}$ denotes the space of admissible forecasting histories. The corresponding functional temporal circuit is

$$\mathcal{C}_\phi = (s_\phi, \gamma_\phi). \tag{5}$$

Thus, s_ϕ specifies *where* the temporal function is implemented, while γ_ϕ characterizes its *function-specific causal state*. This separates functional identity from a fixed anatomical realization. We next describe how s_ϕ is localized.

3.2 FUNCTION-CONTRASTIVE CIRCUIT LOCALIZATION

For each discovery window $\mathbf{X}_i$, we contrast the original history with its function-disrupted counterpart:

$$\mathbf{X}^+_{\phi,i} = \mathbf{X}_i, \qquad \mathbf{X}^-_{\phi,i} = T_\phi(\mathbf{X}_i), \qquad \hat{\mathbf{Y}}^\pm_{\phi,i} = F_\theta\left(\mathbf{X}^\pm_{\phi,i}\right). \tag{6}$$

The corresponding behavioral and internal displacements are

$$\begin{aligned} \Delta\mathbf{y}_{\phi,i} &= \text{vec}\left(\hat{\mathbf{Y}}^+_{\phi,i} - \hat{\mathbf{Y}}^-_{\phi,i}\right), \\ \Delta^s_\phi(\mathbf{X}_i) &= \mathbf{a}^s\left(\mathbf{X}^+_{\phi,i}\right) - \mathbf{a}^s\left(\mathbf{X}^-_{\phi,i}\right), \end{aligned} \tag{7}$$

where $\text{vec}(\cdot)$ vectorizes the forecast matrix.

Across B discovery windows, we collect the response matrices

$$\begin{aligned} \Delta\mathbf{A}^s_\phi &= \text{stack}^B_{i=1}\, \text{vec}\left(\Delta^s_\phi(\mathbf{X}_i)\right), \\ \Delta\mathbf{Y}_\phi &= \text{stack}^B_{i=1}\, \Delta\mathbf{y}_{\phi,i}. \end{aligned} \tag{8}$$

A function-specific support should exhibit more than a large activation change. Motivated by representation analyses that characterize meaningful internal structure through systematic variation (Wiliński et al., 2025), we assess each site along three complementary perspectives: whether it responds to disruption of the target function, whether this response is selective to that function, and whether its internal variation tracks the resulting forecasting behavior.

Intervention Responsiveness. A site involved in function ϕ should react when its temporal evidence is disrupted. We measure the strength of this response by

$$M_\phi(s) = \mathrm{RMS}\left(\Delta \mathbf{A}^s_\phi\right). \tag{9}$$

A large $M_\phi(s)$ indicates strong sensitivity to T_ϕ, but responsiveness alone does not imply functional exclusivity: individual units may be polysemantic, and computational components can be reused across distinct behaviors (Dalvi et al., 2019; Merullo et al., 2024; Yu et al., 2025).

Function Selectivity. We therefore contrast the target response with those induced by other temporal interventions:

$$P_\phi(s) = \frac{M_\phi(s)}{\frac{1}{|\Phi|-1}\sum_{\psi\in\Phi,\psi\neq\phi} M_\psi(s) + \epsilon}. \tag{10}$$

$P_\phi(s)$ favors sites whose response is preferentially associated with ϕ, filtering out generic or function-shared locations.

Behavioral Alignment. Responsiveness and selectivity characterize internal reactions but do not establish whether they track the resulting forecast change. We therefore measure

$$A_\phi(s) = \mathrm{CKA}\left(\Delta \mathbf{A}^s_\phi, \Delta \mathbf{Y}_\phi\right), \tag{11}$$

where CKA captures whether the structure of intervention-induced internal variation across histories aligns with the corresponding behavioral variation (Kornblith et al., 2019; Wiliński et al., 2025).

Together, these criteria ask whether a site responds to the target disruption, does so selectively, and tracks the resulting behavior. We combine the three complementary signals with equal weight to obtain the function-contrastive localization score

$$D_\phi(s) = \mathrm{Mean}\left\{A_\phi(s), \overline{P}_\phi(s), \overline{M}_\phi(s)\right\}, \tag{12}$$

where $\overline{P}_\phi$ and $\overline{M}_\phi$ are min–max normalized across candidate sites. The top-ranked sites form the candidate set

$$\mathcal{S}^{\mathrm{cand}}_\phi = \mathrm{TopK}_{s\in\mathcal{S}}\, D_\phi(s). \tag{13}$$

Thus, D_ϕ provides the primary criterion for localizing candidate computational supports, which are then subjected to causal verification.

3.3 Causal Verification and Adjudication

We causally probe each $s \in \mathcal{S}^{\mathrm{cand}}_\phi$ through sufficiency, counterfactual necessity, and functional specificity. For compactness, we define the forecast distance

$$d(\mathbf{Y}_1, \mathbf{Y}_2) = \sqrt{\frac{1}{HC}\left\|\mathbf{Y}_1 - \mathbf{Y}_2\right\|^2_F}. \tag{14}$$

Causal Sufficiency. For each candidate $s \in \mathcal{S}^{\mathrm{cand}}_\phi$, we patch the original activation into the function-disrupted forward pass:

$$\widetilde{\mathbf{Y}}^{\mathrm{rec}}_{\phi,s,i} = F_\theta\left(\mathbf{X}^-_{\phi,i}; \mathbf{a}^s \leftarrow \mathbf{a}^s\left(\mathbf{X}^+_{\phi,i}\right)\right), \tag{15}$$

and measure recovery by

$$R_\phi(s) = \frac{1}{B}\sum_{i=1}^{B}\left[1 - \frac{d(\widetilde{\mathbf{Y}}^{\mathrm{rec}}_{\phi,s,i}, \hat{\mathbf{Y}}^+_{\phi,i})}{d(\hat{\mathbf{Y}}^-_{\phi,i}, \hat{\mathbf{Y}}^+_{\phi,i}) + \epsilon}\right]. \tag{16}$$

Counterfactual Necessity. We test the reverse direction by patching the function-disrupted activation into the original forward pass:

$$\widetilde{\mathbf{Y}}^{\mathrm{cf}}_{\phi,s,i} = F_\theta\left(\mathbf{X}^+_{\phi,i}; \mathbf{a}^s \leftarrow \mathbf{a}^s\left(\mathbf{X}^-_{\phi,i}\right)\right), \tag{17}$$

with

$$N_\phi(s) = \frac{1}{B}\sum_{i=1}^{B}\left[1 - \frac{d\left(\widetilde{\mathbf{Y}}^{\mathrm{cf}}_{\phi,s,i}, \hat{\mathbf{Y}}^-_{\phi,i}\right)}{d\left(\hat{\mathbf{Y}}^+_{\phi,i}, \hat{\mathbf{Y}}^-_{\phi,i}\right) + \epsilon}\right]. \tag{18}$$

Functional Specificity. Unlike the response-level selectivity P_ϕ, functional specificity tests whether the causal state preferentially restores its matched temporal function. For source ϕ and target ψ, we inject

$$\widetilde{\mathbf{Y}}_{\phi\to\psi,s,i} = F_\theta\left(\mathbf{X}^-_{\psi,i}; \mathbf{a}^s \leftarrow \mathbf{a}^s\left(\mathbf{X}^-_{\psi,i}\right) + \Delta^s_\phi(\mathbf{X}_i)\right). \tag{19}$$

and define

$$G_{\phi\to\psi,i}(s) = 1 - \frac{d\left(\widetilde{\mathbf{Y}}_{\phi\to\psi,s,i}, \hat{\mathbf{Y}}^+_{\psi,i}\right)}{d\left(\hat{\mathbf{Y}}^-_{\psi,i}, \hat{\mathbf{Y}}^+_{\psi,i}\right) + \epsilon}. \tag{20}$$

with the matched-transfer margin

$$V_\phi(s) = \frac{1}{B}\sum_{i=1}^{B}\left[G_{\phi\to\phi,i}(s) - \max_{\psi\in\Phi,\psi\neq\phi} G_{\phi\to\psi,i}(s)\right]. \tag{21}$$

Higher $R_\phi(s)$, $N_\phi(s)$, and $V_\phi(s)$ indicate stronger sufficiency, counterfactual necessity, and function-specific recovery, respectively.

Candidate Adjudication. The localization score $D_\phi(s)$ provides the primary criterion for circuit discovery, while causal verification is applied only to the shortlisted candidates. When a single representative support is required, we combine the localization and verification evidence to adjudicate among them:

$$Q_\phi(s) = \lambda_D D_\phi(s) + \lambda_R\left[R_\phi(s)\right]_+ + \lambda_N\left[N_\phi(s)\right]_+ + \lambda_V\left[V_\phi(s)\right]_+, \tag{22}$$

where $[x]_+ = \max(0, x)$. The final computational support is selected from the candidate set as

$$s_\phi = \arg\max_{s\in\mathcal{S}^{\text{cand}}_\phi} Q_\phi(s). \tag{23}$$

Together with its state-displacement family γ_ϕ, the selected support s_ϕ forms the functional temporal circuit $\mathcal{C}_\phi$ defined in Sec. 3.1.

3.4 Circuit Utilization

Cross-Dataset Support Generalization. We first examine whether the discovered support s_ϕ generalizes beyond its discovery data. We discover and freeze s_ϕ using source training datasets only, and directly reuse it to restore the corresponding disrupted function on a held-out dataset, without target-data adaptation.

Cross-Model Functional Alignment and Transplantation. We next ask whether function-specific causal states can be aligned across independently pretrained TSFMs. Let s^d_ϕ and s^r_ϕ denote the supports of function ϕ in donor model d and receiver model r, with corresponding state displacements Δ^d_ϕ and Δ^r_ϕ.

Because their representation spaces are incompatible, we represent each displacement using d_z-dimensional coordinates $\mathbf{z}^m_\phi$ derived from normalized training displacement prototypes. For paired histories from the same channel and forecasting origin, each model uses its native context, and we learn the ridge map

$$\mathbf{W}^{d\to r}_\phi = \arg\min_{\mathbf{W}} \sum_i \left\|\mathbf{z}^r_{\phi,i} - \mathbf{z}^d_{\phi,i}\mathbf{W}\right\|_2^2 + \lambda_{\text{map}}\|\mathbf{W}\|_F^2. \tag{24}$$

The mapped coordinates are reconstructed in the receiver's prototype space to obtain the transferred state displacement $\widetilde{\Delta}^r_\phi$. We inject this displacement into the corresponding function-disrupted receiver state,

$$\mathbf{a}^{s^r_\phi} \leftarrow \mathbf{a}^{s^r_\phi} + \widetilde{\Delta}^r_\phi, \tag{25}$$

and measure whether it restores the disrupted temporal behavior.

Both TSFMs remain frozen throughout alignment, and only the low-dimensional coordinate mapping is fitted using training histories without future targets. Functional correspondence is supported when matched transplantation consistently outperforms, indicating that architecture-specific causal states can encode transferable temporal functionality.

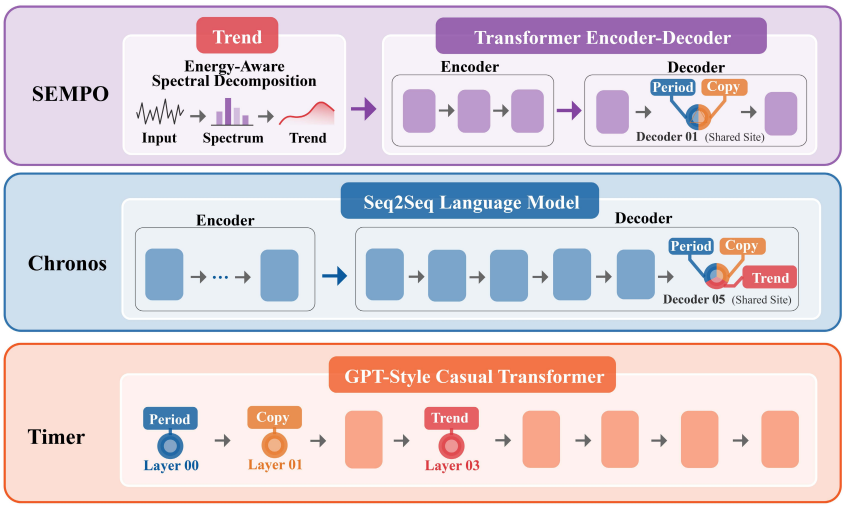

Figure 3: Function-specific localization across SEMPO, Chronos, and Timer, with bold-colored circles marking the final supports selected from the Top-K candidates.

4 Experiments

4.1 Experimental Setup

Models and datasets. We evaluate CircuitTS on three frozen TSFMs with heterogeneous architectures: **SEMPO**, **Chronos**, and **Timer**. Experiments cover nine benchmarks: ETTh1, ETTh2, ETTm1, ETTm2, Weather, Electricity, Traffic, Exchange, and Solar. Unless otherwise specified, we use $H = 96$ and study three temporal functions: *Period*, *Trend*, and *Copy*. Multivariate datasets follow channel-independent forecasting. All TSFM parameters remain frozen; only the lightweight cross-model alignment map is fitted when explicitly specified.

Evaluation protocol. Circuit localization and candidate adjudication are performed exclusively on training windows: D_ϕ constructs the Top-2 candidate set, from which Q_ϕ selects the final support. The selected support is then frozen, and sufficiency, necessity, and functional specificity are recomputed on held-out validation windows. No validation statistic or future target is used for support selection. For cross-dataset generalization, each dataset is held out in turn, with supports discovered from the remaining eight training datasets and evaluated on the test split of the unseen dataset.

Implementation details. The main discovery setting uses 24 training windows per dataset. Cross-model alignment represents causal states with $d_z = 8$ prototype coordinates and fits a ridge map using paired training histories, with evaluation on disjoint test windows. Forecast-grounded Timer→SEMPO transplantation is evaluated on five datasets and four horizons. Further implementation details and cross-horizon results are provided in the Appendix.

4.2 Function-Contrastive Circuit Localization and Verification

Circuit Localization. We first examine how temporal functions are localized across TSFMs. As shown in Fig. 3, their anatomical organization varies substantially across architectures. SEMPO isolates Trend in its spectral pathway while Period and Copy share decoder support; Chronos concentrates all three functions at the same late decoder block; and Timer separates Period, Copy, and Trend across different depths. These patterns indicate that temporal functions may either share or separate anatomical support depending on the architecture, making location alone insufficient to identify functional mechanisms. Detailed results are provided in Appendix.

Causal Verification. We next test whether the localized supports causally mediate their corresponding temporal functions. As shown in Table 1, the selected supports exhibit consistently positive sufficiency, necessity, and functional specificity across models and functions. This indicates that

Table 1: Causal verification of the discovered temporal circuits across nine datasets. We report mean **sufficiency R**, **necessity N**, and **functional specificity V** for each model–function pair.

	Period			Trend			Copy			Average		
Model	R	N	V	R	N	V	R	N	V	R	N	V
SEMPO	0.765	0.783	**1.732**	0.911	0.921	1.539	0.852	0.850	1.155	0.842	0.851	1.475
Chronos	**0.948**	**0.960**	1.661	**0.987**	**0.989**	**4.098**	**0.971**	**0.970**	1.035	**0.968**	**0.973**	**2.265**
Timer	0.814	0.841	1.608	0.947	0.958	2.759	0.928	0.928	**1.157**	0.896	0.909	1.841

Table 2: Ablation of localization criteria, averaged across models, datasets, and temporal functions. Only the criterion used to construct the Top-K candidate set is varied.

Criterion	**Sufficiency R ↑**	**Necessity N ↑**	**Functional specificity V ↑**
Intervention Responsiveness M	0.889	0.898	**1.885**
Function Selectivity P	0.836	0.838	1.497
Behavioral Alignment A	0.858	0.855	1.614
localization score D	**0.902**	**0.911**	1.860

they not only mediate the disrupted forecasting behavior, but also carry causal states preferentially associated with the corresponding temporal function.

Ablation of Localization Criteria. We further ablate the three signals used for candidate localization. The Top-K candidate set is constructed using Intervention Responsiveness M, Function Selectivity P, Behavioral Alignment A, or their equal-weight combination D, while the subsequent causal adjudication remains unchanged. The selected supports are evaluated on held-out windows using causal sufficiency R, counterfactual necessity N, and functional specificity V.

As shown in Table 2, the individual criteria capture different aspects of candidate quality. Responsiveness M yields the strongest functional specificity, whereas P and A are weaker when used independently. Their combination D achieves the highest sufficiency and necessity while remaining close to the best specificity. These results support combining responsiveness, selectivity, and behavioral alignment for candidate localization rather than relying on any single criterion.

4.3 Generalization to Unseen Datasets

We next examine whether the discovered computational supports generalize to unseen datasets. In a leave-one-dataset-out setting, each target dataset is excluded from circuit discovery, while the support is identified using only the remaining eight training datasets and then frozen for evaluation on the target test split. We disrupt each temporal function on the unseen dataset and restore the clean state only at the transferred support.

Table 3 reports SEMPO as a representative example. Supports discovered without the target dataset consistently recover the corrupted forecasts for Period and Copy, and improve Trend in eight of nine datasets. In most cases, the restored prediction approaches the original forecast despite the support being identified entirely from other datasets. These results indicate that the discovered circuits capture temporal functions that generalize beyond the datasets used for localization. Results for Chronos and Timer are provided in the Appendix.

4.4 Cross-Model Functional Alignment and Transplantation

Finally, we ask whether function-specific causal states can be aligned across independently pretrained TSFMs. We learn lightweight mappings from donor to receiver circuit states using paired training histories, while keeping both models frozen, and test whether the mapped state can restore the corresponding function in the receiver. Full results are provided in Appendix.

As shown in Table 4, Timer→SEMPO transplantation consistently improves forecasts degraded by function-specific corruption across datasets and horizons. Period improves all 20 dataset–horizon cases, while Trend produces predictions closest on average to the original receiver. Copy also improves 19 of 20 cases. These results provide causal evidence that function-specific states can be

Table 3: Cross-dataset generalization of SEMPO circuits at $H = 96$. Each function reports MSE as *Corrupt→Restore*, where the support is discovered without access to the target dataset. Original denotes the unmodified forecast; lower is better.

Dataset	Original	Period	Trend	Copy
ETTh1	0.390	0.572→**0.398**	0.400→**0.390**	0.461→**0.390**
ETTh2	0.276	0.294→**0.276**	0.309→**0.278**	0.317→**0.277**
ETTm1	0.521	0.641→**0.543**	0.525→**0.520**	0.710→**0.509**
ETTm2	0.192	0.196→**0.194**	0.259→**0.193**	0.223→**0.191**
Weather	0.141	0.149→**0.142**	0.196→**0.145**	0.157→**0.141**
Electricity	0.179	0.534→**0.231**	0.185→**0.180**	0.280→**0.181**
Traffic	0.422	0.782→**0.450**	**0.421**→0.422	0.546→**0.422**
Exchange	0.120	0.123→**0.121**	0.399→**0.153**	0.143→**0.121**
Solar	0.458	0.674→**0.475**	0.460→**0.457**	0.650→**0.458**
Improved	–	**9/9**	**8/9**	**9/9**

Table 4: Forecast-grounded functional transplantation from Timer to SEMPO. MSE is averaged over $H \in \{96, 192, 336, 720\}$. MSE Wins counts dataset–horizon cells in which transplantation improves over the corrupted receiver.

Function	Variant	ETTh1	ETTh2	ETTm1	ETTm2	Weather	Avg.	MSE Wins
–	Original	0.414	0.341	0.517	0.289	0.248	0.362	–
Period	Corrupt	0.557	0.357	0.602	0.296	0.258	0.414	–
	Transplant	0.437	0.342	0.590	0.293	0.255	**0.384**	**20/20**
Trend	Corrupt	0.419	0.354	0.516	0.321	0.295	0.381	–
	Transplant	0.416	0.342	0.516	0.293	0.258	**0.365**	**16/20**
Copy	Corrupt	0.462	0.365	0.606	0.310	0.267	0.402	–
	Transplant	0.460	0.357	0.578	0.303	0.262	**0.392**	**19/20**

aligned and transferred across heterogeneous TSFMs despite differences in their internal representations and computational supports.

5 Discussion and Conclusion

5.1 Discussion and Limitations

Our results separate functional identity from anatomical realization: temporal functions may share or shift computational supports, while their causal roles remain transferable across datasets and models. Cross-model transplantation further suggests that corresponding functions can be represented in different internal coordinate systems. This compatibility supports comparing temporal mechanisms through function-specific causal states, but does not imply identical underlying algorithms. CircuitTS currently considers three temporal functions and three TSFMs, and its conclusions remain conditioned on the chosen interventions and candidate sites. Extending the functional vocabulary, architectural coverage, and criteria for stronger cross-model equivalence remains an important direction for future work.

5.2 Conclusion

We introduced **CircuitTS**, a framework for discovering functional temporal circuits in pretrained TSFMs. By combining controlled temporal interventions, function-contrastive localization, and causal verification, CircuitTS identifies the computational supports and function-specific causal states of Period, Trend, and Copy. The discovered circuits generalize beyond their discovery data, while their causal states can be aligned and transferred across independently pretrained models. These findings advance TSFM interpretability from describing represented temporal information toward understanding how temporal computations are causally implemented and reused.

AI USE STATEMENT

In this work, we used generative AI tools for idea completion and execution, finding related works, and polishing writings. We have not used generative AI tools for drafting all sections of the paper, generating synthetic datasets and proving mathematical claims are not applicable to this work. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey. LLM-generated code was verified and tested for correctness by the authors. All writings were checked and finalized by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate reproducibility. The experimental protocol, including dataset splits, preprocessing, forecasting settings, frozen TSFM backbones, candidate internal sites, temporal interventions, and the separation between circuit discovery and held-out evaluation, is described in Sec. 4.1 and Appendix A.1. The complete circuit localization and causal adjudication procedure, together with the corresponding hyperparameters and candidate-level results, is provided in Secs. 3.2–3.3 and Appendix A.2. Appendix A.2.1 reports sensitivity to the discovery-set size, while Appendices A.3–A.5 provide the full cross-dataset, cross-model, and cross-horizon results. All circuit discovery and cross-model alignment procedures use only training histories and do not access future targets; future observations are used only for final forecasting evaluation where explicitly stated. We will provide the complete implementation, experiment scripts, configuration files, and instructions for reproducing the reported results as anonymized supplementary material, together with references to the publicly available pretrained checkpoints and datasets used in this work.

REFERENCES

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle Maddix Robinson, Hao Wang, Michael Mahoney, Kari Torkkola, Andrew Wilson, Michael Bohlke-Schneider, and Yuyang (Bernie) Wang. Chronos: Learning the language of time series. *Transactions of Machine Learning Research*, 2024.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6309–6317, 2019.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Forty-first International Conference on Machine Learning*, 2024.

Xinghong Fu, Yanhong Li, GEORGIOS PAPAIOANNOU, and Yoon Kim. Reverso: Efficient time series foundation models for zero-shot forecasting. In *2nd ICML Workshop on Foundation Models for Structured Data*, 2026. URL https://openreview.net/forum?id=gsuqQs0G0K.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in neural information processing systems*, 34:9574–9586, 2021.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning*, pp. 7324–7338. PMLR, 2022.

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

Hui He, Kun Yi, Yuanchi Ma, Qi Zhang, Zhendong Niu, and Guansong Pang. Sempo: Lightweight foundation models for time series forecasting. *Advances in Neural Information Processing Systems*, 38:161846–161879, 2026.

Krzysztof Kacprzyk, Tennison Liu, and Mihaela van der Schaar. Towards transparent time series forecasting. In *International Conference on Learning Representations*, volume 2024, pp. 41452–41487, 2024.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Vedika Sanjeev Lakhanpal. TimeSAE: Mechanistic interpretability for time-series foundation models. In *1st ICLR Workshop on Time Series in the Age of Large Models*, 2026. URL https://openreview.net/forum?id=Ojd6YjHpyE.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=bYRYb7DMNo.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, volume 2024, pp. 26486–26515, 2024.

Samuel Marks, Can Rager, Eric Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, volume 2025, pp. 23888–23923, 2025.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *International Conference on Learning Representations*, volume 2025, pp. 32900–32923, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. In *International Conference on Learning Representations*, volume 2024, pp. 18349–18377, 2024.

Anurag Mishra. Dissecting chronos: Sparse autoencoders reveal causal feature hierarchies in time series foundation models. In *1st ICLR Workshop on Time Series in the Age of Large Models*, 2026. URL https://openreview.net/forum?id=Q5ox302DLN.

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: Neural Basis Expansion Analysis for Interpretable Time Series Forecasting. In *International Conference on Learning Representations*, 2020. URL https://mlanthology.org/iclr/2020/oreshkin2020iclr-nbeats/.

Atharva Pandey, Abhilash Neog, and Gautam Jajoo. On the internal semantics of time-series foundation models. In *Recent Advances in Time Series Foundation Models Have We Reached the 'BERT Moment'?*, 2025. URL https://openreview.net/forum?id=m3J83J1xhT.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Arian Khorasani, George Adamopoulos, Rishika Bhagwatkar, Marin Biloš, Hena Ghonia, Nadhir Hassen, Anderson Schneider, et al. Lag-llama: Towards foundation models for time series forecasting. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pp. 17282–17333, 2024.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023.

Michał Wiliński, Mononito Goswami, Willa Potosnak, Nina Żukowska, and Artur Dubrawski. Exploring representations and interventions in time series foundation models. In *Forty-second International Conference on Machine Learning*, 2025.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*, 2024.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. Knowledge circuits in pretrained transformers. *Advances in Neural Information Processing Systems*, 37:118571–118602, 2024.

Wenlong Yu, Qilong Wang, Chuang Liu, Dong Li, and Qinghua Hu. Coe: Chain-of-explanation via automatic visual concept circuit description and polysemanticity quantification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4364–4374, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

A Appendix

A.1 Implementation Details and Reproducibility

Data Processing and Sampling. We use nine real-world forecasting benchmarks: ETTh1, ETTh2, ETTm1, ETTm2, Weather, Electricity, Traffic, Exchange, and Solar. Following standard long-term forecasting protocols, the ETT datasets use the predefined 12/4/4-month train/validation/test split, while the remaining datasets use a 7:1:2 split. Each channel is standardized using statistics computed from the training split and evaluated independently under the channel-independent setting.

The main experiments use forecasting horizon $H = 96$. To preserve the native inference setting of each TSFM, SEMPO, Chronos, and Timer receive context lengths of 512, 2048, and 672, respectively. All models share the same forecasting origin and channel, while function-specific interventions are restricted to the latest $W = 336$ observations. Main circuit discovery uses 24 training windows per dataset, and held-out causal verification uses 24 validation windows per dataset.

Frozen TSFM Backbones and Candidate Sites. All pretrained model parameters remain frozen. SEMPO is instantiated from its official zero-shot checkpoint with a 512-step context, and candidate sites include its spectral patch representation and all encoder and decoder blocks. Chronos uses the pretrained Chronos-Bolt checkpoint with a 2048-step context, with all exposed encoder and decoder blocks considered as candidate sites. Timer follows the TSFM-Bench implementation with eight Transformer layers, $d_{\text{model}} = 1024$, eight attention heads, and a token length of 96; all eight layers are included as candidate sites. For module outputs containing temporal or token dimensions, we mean-pool the non-feature dimensions to obtain one activation vector per input history before computing intervention-induced state displacements.

Temporal Interventions. All interventions modify only the latest $W = 336$ observations and use full dose. For Period, a dataset-level characteristic period is estimated exclusively from the training split using detrended ACF–FFT consensus over at most 16 channels, with

$$\mathcal{P} = \{4, 6, 8, 12, 16, 24, 32, 48, 72, 96, 120, 144, 168\}.$$

Given the selected period p, we fit an intercept, centered linear trend, and sinusoidal pair at frequency $2\pi/p$, and remove only the fitted harmonic component. Trend removes the centered least-squares linear component while preserving the local mean. Copy replaces the last

$$k = \max(8, \text{round}(0.05W))$$

observations with the immediately preceding block of the same length; for $W = 336$, $k = 17$. Period estimation, intervention construction, and circuit localization never access future targets.

Cross-Dataset Generalization. For each held-out dataset, circuit discovery is repeated using only the remaining eight training datasets. Each source dataset contributes 8 discovery windows, followed by causal adjudication on 24 training histories. The resulting support is frozen before evaluation on the unseen dataset. Forecast-grounded results in Sec. 4.3 use 256 deterministic test origin–channel pairs at $H = 96$; future targets are accessed only after circuit discovery and intervention construction are complete, solely for computing MSE and MAE.

Table 5: Key implementation settings used throughout CircuitTS.

Setting	Value
Models	SEMPO / Chronos / Timer
Main horizon	96
Native contexts	512 / 2048 / 672
Intervention window	336
Discovery windows	24 per dataset
Validation windows	24 per dataset
Intervention dose	1.0
Localization weights	$1/3,\ 1/3,\ 1/3$
Adjudication weights	$0.45,\ 0.25,\ 0.15,\ 0.15$

Table 6: Detailed localization and causal adjudication of the Top-K candidates. $\overline{M}$ and $\overline{P}$ denote site-wise normalized Intervention Responsiveness and Function Selectivity, respectively, and A denotes Behavioral Alignment. The three signals are combined with equal weight to form the function-contrastive localization score $D = (A+\overline{P}+\overline{M})/3$. R, N, and V denote causal sufficiency, counterfactual necessity, and functional specificity, respectively. Q is used only to adjudicate within the Top-K shortlist. The final selected support is shown in bold.

Model	Function	Site	$\overline{M}$	$\overline{P}$	A	D	R	N	V	Q
SEMPO	Period	**decoder_01**	1.000	1.000	0.461	0.820	0.756	0.780	1.939	0.966
		decoder_00	0.259	0.988	0.581	0.609	0.719	0.751	1.959	0.860
	Trend	encoder_00	0.472	0.873	0.499	0.615	0.816	0.821	2.495	0.978
		spectral_patch	0.116	1.000	0.628	0.581	0.937	0.950	2.398	0.998
	Copy	**decoder_01**	1.000	0.581	0.524	0.702	0.850	0.854	0.907	0.792
		decoder_02	0.000	1.000	0.603	0.534	0.915	0.915	0.982	0.754
Chronos	Period	**decoder_05**	1.000	0.143	0.427	0.523	0.958	0.970	1.689	0.874
		decoder_04	0.592	0.305	0.524	0.473	0.757	0.674	1.172	0.679
	Trend	**decoder_05**	1.000	0.244	0.290	0.511	0.984	0.988	2.231	0.959
		encoder_04	0.039	1.000	0.352	0.464	0.984	0.988	1.474	0.824
	Copy	**decoder_05**	1.000	0.771	0.257	0.676	0.980	0.981	0.835	0.822
		decoder_03	0.249	0.809	0.650	0.569	0.477	0.422	0.362	0.493
Timer	Period	**layer_00**	1.000	1.000	0.658	0.886	0.817	0.849	2.321	1.078
		layer_01	0.902	0.633	0.674	0.736	0.817	0.849	2.231	0.997
	Trend	layer_04	1.000	1.000	0.307	0.769	0.956	0.969	2.733	1.140
		layer_03	0.956	0.830	0.323	0.703	0.956	0.969	3.140	1.172
	Copy	**layer_01**	1.000	0.777	0.346	0.708	0.921	0.921	0.958	0.831
		layer_02	0.968	0.515	0.368	0.617	0.921	0.921	0.930	0.786

Cross-Model Alignment. For the forecast-grounded transplantation experiment, we evaluate ETTh1, ETTh2, ETTm1, ETTm2, and Weather over $H \in \{96, 192, 336, 720\}$. Donor and receiver states are extracted from paired histories sharing the same dataset, channel, and forecasting origin, while each model retains its native context. Each function-specific state is represented using $d_z = 8$ normalized training displacement prototypes. Coordinate projection uses ridge coefficient 10^{-4}, and the standardized donor-to-receiver coordinate map uses ridge coefficient 10^{-2}. The map is fitted using 16 training windows per dataset and horizon and is then frozen. Test evaluation uses 32 temporal origins per dataset and evaluates all channels at each origin. Neither basis construction nor map fitting uses future targets; ground-truth futures are used only for final forecasting metrics.

A.2 Detailed Localization and Candidate Adjudication

Figure 3 in the main text visualizes the architectural locations of the selected temporal-function supports. Here, we provide the candidate-level evidence underlying the final support selection. For each model–function pair, D_ϕ first determines the Top-K candidate supports using training windows only. Causal sufficiency R_ϕ, counterfactual necessity N_ϕ, and functional specificity V_ϕ are then evaluated for these shortlisted candidates and combined with D_ϕ through Q_ϕ to select the final support.

Table 6 reports the candidate-level localization and adjudication scores for the Top-K shortlist. The three localization signals capture complementary properties: $\overline{M}_\phi$ measures intervention responsiveness, $\overline{P}_\phi$ measures function selectivity, and A_ϕ measures behavioral alignment. Their equal-weight aggregation forms D_ϕ, while R_ϕ, N_ϕ, and V_ϕ provide subsequent causal evidence within the shortlist. This decomposition makes explicit how the final computational support is obtained without using held-out validation statistics.

The shortlisted sites generally exhibit different strengths across the three localization signals, confirming that no single criterion fully determines the candidate set. The final support is therefore selected from candidates that are jointly strong in function-contrastive localization and causal veri-

Table 7: Sensitivity to the number of discovery windows. Support agreement is measured against the supports discovered using 48 windows per dataset. R and N are recomputed on the same held-out validation set for the supports obtained under each discovery budget. Positive cells count model–dataset–function combinations with positive sufficiency and necessity out of 81 total cells.

Discovery windows	Support agreement	$R \uparrow$	$N \uparrow$	Positive R/N cells
6	0.889	0.902	0.911	81/81
12	0.778	**0.908**	**0.917**	81/81
24	0.778	0.902	0.911	81/81
48	1.000	0.889	0.898	81/81

fication, rather than from the largest activation response alone. Importantly, all quantities in Table 6 used for support selection are computed from training windows; the held-out validation results reported in the main text are recomputed only after the support has been frozen.

A.2.1 Sensitivity to Discovery-Set Size

We examine whether circuit discovery is sensitive to the number of training windows used for localization. We vary the discovery budget over $\{6, 12, 24, 48\}$ windows per dataset while keeping the intervention design, candidate sites, adjudication procedure, and held-out validation protocol unchanged. For each budget, the resulting supports are frozen and evaluated on the same held-out validation windows. Support agreement is measured against the largest discovery budget of 48 windows per dataset.

As shown in Table 7, causal validity is stable across discovery budgets. All settings retain positive sufficiency and necessity in all 81 model–dataset–function cells, while the mean R and N vary only moderately. In particular, increasing the discovery budget from 24 to 48 windows does not improve held-out causal scores, suggesting that the default 24-window setting already provides sufficient evidence for identifying causally effective supports.

A.3 Full Cross-Dataset Generalization Results

The main text reports SEMPO as a representative example of cross-dataset generalization. Here, we provide the complete results for Chronos and Timer under the same leave-one-dataset-out protocol. For each target dataset, the computational support is discovered using only the remaining eight training datasets and is then frozen before evaluation on the target test split. For each temporal function, we report the forecasting MSE before and after restoring the clean causal state at the transferred support. An improvement is counted when restoration reduces the MSE relative to the corresponding corrupted forecast.

A.3.1 Chronos

Table 8 reports the complete results for Chronos. Period and Copy exhibit particularly stable cross-dataset generalization: restoration improves the corrupted forecast on all nine held-out datasets. The effect is especially pronounced when the intervention substantially degrades the original forecast. Trend restoration improves seven of nine datasets, with ETTh1 and Traffic being the two exceptions.

The restored forecasts frequently approach the clean Chronos predictions despite the target dataset being completely absent during support discovery. In several cases, restoration even yields an MSE slightly below the clean forecast, which is possible because the causal intervention is evaluated against the ground-truth future whereas support discovery itself does not use future targets. Overall, these results indicate that the localized Chronos supports capture temporal functionality that is not specific to the datasets on which they were discovered.

A.3.2 Timer

Table 9 shows the corresponding results for Timer. The same overall pattern is observed. Period and Copy restoration improves all nine held-out datasets, including large recoveries on Electricity

Table 8: Cross-dataset generalization of Chronos circuits at $H = 96$. Each function reports MSE as *Corrupt→Restore*, where the support is discovered without access to the target dataset. Bold values indicate improvements over the corrupted forecast. Clean denotes the unmodified forecast; lower is better.

Dataset	Clean	Period	Trend	Copy
ETTh1	0.380	0.601→**0.397**	0.389→0.392	0.458→**0.389**
ETTh2	0.272	0.287→**0.273**	0.310→**0.289**	0.321→**0.295**
ETTm1	0.301	0.408→**0.342**	0.328→**0.316**	0.565→**0.399**
ETTm2	0.151	0.167→**0.157**	0.232→**0.159**	0.198→**0.175**
Weather	0.128	0.142→**0.125**	0.208→**0.141**	0.152→**0.133**
Electricity	0.120	0.585→**0.177**	0.139→**0.127**	0.382→**0.140**
Traffic	0.361	0.784→**0.395**	0.365→0.366	0.645→**0.381**
Exchange	0.085	0.091→**0.089**	0.311→**0.112**	0.110→**0.095**
Solar	0.268	0.538→**0.277**	0.452→**0.243**	0.412→**0.272**
Improved	–	**9/9**	**7/9**	**9/9**

Table 9: Cross-dataset generalization of Timer circuits at $H = 96$. Each function reports MSE as *Corrupt→Restore*, where the support is discovered without access to the target dataset. Bold values indicate improvements over the corrupted forecast. Clean denotes the unmodified forecast; lower is better.

Dataset	Clean	Period	Trend	Copy
ETTh1	0.398	0.622→**0.392**	0.400→**0.398**	0.484→**0.398**
ETTh2	0.279	0.292→**0.278**	0.323→**0.281**	0.316→**0.280**
ETTm1	0.647	0.676→**0.649**	0.655→**0.647**	0.823→**0.643**
ETTm2	0.192	0.196→**0.194**	0.246→**0.194**	0.220→**0.192**
Weather	0.147	0.148→**0.146**	0.205→**0.151**	0.165→**0.146**
Electricity	0.138	0.587→**0.159**	0.144→**0.138**	0.349→**0.138**
Traffic	0.390	0.871→**0.390**	0.387→0.389	0.600→**0.390**
Exchange	0.123	0.126→**0.124**	0.327→**0.150**	0.138→**0.124**
Solar	0.515	0.645→**0.521**	0.510→0.515	0.697→**0.516**
Improved	–	**9/9**	**7/9**	**9/9**

and Traffic. Trend again exhibits greater variation, improving seven of nine datasets. The two non-improving cases, Traffic and Solar, already have corrupted MSEs slightly below their corresponding clean forecasts, making reduction of forecasting error an especially stringent criterion for restoration.

Overall Cross-Dataset Generalization. Together with the SEMPO results in Table 3, the complete evaluation reveals a consistent pattern across all three architectures. Period and Copy restoration improves forecasting MSE on every held-out dataset for SEMPO, Chronos, and Timer, whereas Trend improves $8/9$, $7/9$, and $7/9$ datasets, respectively. Importantly, these supports are identified without any training window from the target dataset and remain fixed during evaluation.

These results provide complementary evidence to the in-distribution causal verification in the main text. The discovered supports do not merely mediate function-specific behavior on the datasets used for localization; they remain causally effective under dataset shift. At the same time, the greater variability of Trend suggests that some temporal functions may depend more strongly on dataset-specific realization than others. Overall, the results support our central distinction between *functional identity* and *anatomical realization*: a computational support can preserve its function-specific causal role across unseen datasets without requiring the underlying data distribution to remain unchanged.

A.4 Full Cross-Model Alignment and Ablations

The main text reports Timer→SEMPO as a representative forecast-grounded transplantation experiment. Here, we provide the complete results over all six directed model pairs and additional controls that test whether successful transplantation depends on function-matched cross-model alignment rather than generic state injection.

Table 10: Forecast-grounded cross-model transplantation over all directed TSFM pairs at $H = 720$. MSE is averaged over five datasets. *Wins* counts the number of datasets in which transplantation improves over the corrupted receiver. Bold transplantation values indicate lower MSE than the corresponding corrupted forecast.

Donor → Receiver	Function	Clean	Corrupt	Transplant	Wins
SEMPO → Chronos	Period	0.435	0.507	**0.468**	**2/5**
	Trend	0.435	0.451	0.591	0/5
	Copy	0.435	0.464	**0.452**	**4/5**
SEMPO → Timer	Period	0.467	0.501	**0.471**	**2/5**
	Trend	0.467	0.474	**0.468**	**5/5**
	Copy	0.467	0.487	**0.483**	**4/5**
Chronos → SEMPO	Period	0.425	0.472	**0.446**	**5/5**
	Trend	0.425	0.433	**0.424**	**4/5**
	Copy	0.425	0.444	**0.440**	**4/5**
Chronos → Timer	Period	0.467	0.501	**0.475**	**3/5**
	Trend	0.467	0.474	**0.468**	**5/5**
	Copy	0.467	0.487	**0.485**	**3/5**
Timer → SEMPO	Period	0.425	0.472	**0.446**	**5/5**
	Trend	0.425	0.433	**0.425**	**4/5**
	Copy	0.425	0.444	**0.437**	**5/5**
Timer → Chronos	Period	0.435	0.507	**0.469**	**3/5**
	Trend	0.435	0.451	0.631	0/5
	Copy	0.435	0.464	**0.457**	**4/5**

For every donor–receiver pair, the two TSFMs remain frozen. Function-specific state displacements are represented using $d_z = 8$ prototype coordinates, and only the lightweight donor-to-receiver ridge map is fitted from paired training histories. Evaluation is performed on disjoint test histories. Future targets are used only for the forecast-grounded MSE and MAE evaluation and are never used to fit the prototype spaces or cross-model maps.

A.4.1 Full Directed-Pair Forecast Transplantation

Table 10 extends the representative Timer→SEMPO experiment in the main text to all six directed model pairs at $H = 720$. Results are averaged over five datasets: ETTh1, ETTh2, ETTm1, ETTm2, and Weather. *Wins* counts the number of datasets for which transplantation reduces MSE relative to the function-corrupted receiver.

The full directed-pair results show that forecast recovery is generally transferable across heterogeneous TSFMs, but the strength of transfer is direction- and function-dependent. The clearest pattern is observed when SEMPO is the receiver: both Chronos→SEMPO and Timer→SEMPO improve Period on all five datasets and improve Trend and Copy on most or all datasets. Transfers into Timer are also consistently beneficial, particularly for Trend.

Transfers into Chronos are more heterogeneous. Period and Copy usually improve over the corrupted receiver, whereas Trend does not improve for SEMPO→Chronos or Timer→Chronos at $H = 720$. This asymmetry indicates that functional correspondence does not imply equally easy transplantation in every direction. Instead, the effectiveness of transfer depends on how well the donor state can be represented in the receiver's function-specific coordinate system.

A.4.2 Ablation of Functional Alignment

We further ablate the two key ingredients of cross-model functional alignment: paired histories and function correspondence. All model parameters, circuit supports, prototype spaces, and mapping capacities are kept unchanged; only the correspondence used to learn or apply the cross-model mapping is modified.

We compare three settings. *Matched* denotes the proposed alignment, where donor and receiver states are obtained from paired histories and the same temporal function. *Shuffled* randomly per-

Table 11: Ablation of cross-model functional alignment. Entries report median output-space recovery on held-out test histories. *Matched* uses correctly paired histories and the corresponding temporal function; *Shuffled* removes training-history correspondence; and *Mismatch* reports the stronger recovery obtained from either incorrect temporal function. Higher is better. Best results are shown in **bold**.

Donor → Receiver	Function	Matched	Shuffled	Mismatch
SEMPO → Chronos	Period	**0.228**	-0.015	0.039
	Trend	**0.587**	0.193	0.011
	Copy	**0.109**	-0.009	0.069
SEMPO → Timer	Period	**0.360**	-0.004	0.001
	Trend	**0.461**	0.068	-0.344
	Copy	**0.056**	-0.100	0.051
Chronos → SEMPO	Period	**0.179**	-0.056	0.003
	Trend	**0.610**	-0.134	-0.175
	Copy	**0.235**	-0.070	0.042
Chronos → Timer	Period	**0.259**	0.016	0.042
	Trend	**0.433**	0.022	-0.127
	Copy	**0.123**	0.007	0.043
Timer → SEMPO	Period	**0.280**	-0.026	-0.003
	Trend	**0.755**	0.247	-0.511
	Copy	**0.421**	-0.130	0.024
Timer → Chronos	Period	**0.241**	0.012	0.051
	Trend	**0.588**	0.294	0.010
	Copy	**0.133**	-0.027	0.078

mutes the correspondence between donor and receiver training histories before fitting the mapping, thereby preserving the same samples and mapping capacity while removing sample-level alignment. *Mismatch* transfers a state associated with an incorrect temporal function; for each test example, we report the stronger recovery among the two mismatched functions, making this a conservative function-specificity ablation.

As shown in Table 11, matched functional alignment achieves the highest median recovery in all 18 donor–receiver–function combinations. Removing sample correspondence through shuffling substantially reduces recovery and frequently results in values near or below zero. This shows that the learned mapping cannot be explained merely by global similarity between the donor and receiver state spaces.

The mismatched-function ablation yields a similarly clear separation. Even though we conservatively retain the better of the two incorrect temporal functions for each test example, matched transfer remains superior in every model–function pair.

A.5 Cross-Horizon Analysis

We further examine whether the computational supports discovered at the default forecasting horizon remain functionally valid when the prediction horizon changes. Importantly, circuits are not rediscovered at each horizon. Instead, the support s_ϕ identified by the GLOBAL9 discovery experiment at $H = 96$ is frozen and reused for $H \in \{96, 192, 336, 720\}$. At each target horizon, the clean and function-disrupted states are recomputed at this fixed support, allowing us to separate the stability of *where* a function is implemented from the numerical invariance of its state displacement.

To ensure a paired comparison, the same validation origins and channels are used across all horizons. For each model–dataset–function combination, we recompute causal sufficiency R_ϕ and counterfactual necessity N_ϕ at the frozen support. We additionally perform the full 3×3 cross-function state-transfer test and report *functional specificity*, defined as the fraction of held-out examples for which the matched function yields greater recovery than both mismatched functions. No future targets are used in this analysis.

Table 12: Cross-horizon causal stability of circuits discovered at $H = 96$. R and N denote mean sufficiency and counterfactual necessity, respectively, averaged over three models, nine datasets, and three temporal functions. Positive cells count model–dataset–function combinations with positive mean causal scores out of 81 total cells. Specificity denotes the fraction of held-out examples for which matched function transfer exceeds both mismatched functions. The circuit supports are fixed from $H = 96$ and are not rediscovered at longer horizons.

Horizon	$R \uparrow$	$N \uparrow$	Positive R	Positive N	Specificity $\uparrow$
96	0.800	0.793	**81/81**	**81/81**	0.955
192	0.803	0.801	**81/81**	**81/81**	0.945
336	**0.811**	**0.810**	**81/81**	**81/81**	0.892
720	0.790	0.794	**81/81**	**81/81**	0.762

Table 13: Model-wise cross-horizon circuit stability. For each model and horizon, R and N are averaged over nine datasets and three temporal functions, while Specificity measures matched-function dominance in the full cross-function transfer test. All supports are those originally discovered at $H = 96$.

Model	Horizon	$R \uparrow$	$N \uparrow$	Specificity $\uparrow$
SEMPO	96	**0.860**	**0.869**	0.998
	192	0.845	0.861	0.997
	336	0.847	0.857	**1.000**
	720	0.833	0.848	**1.000**
Chronos	96	0.642	0.600	**0.880**
	192	0.676	0.643	0.863
	336	0.717	0.693	0.747
	720	**0.770**	**0.758**	0.580
Timer	96	**0.896**	**0.909**	**0.988**
	192	0.887	0.901	0.977
	336	0.869	0.882	0.931
	720	0.768	0.776	0.705

Table 12 shows that the causal role of the $H = 96$ supports remains remarkably stable as the forecasting horizon grows. Mean sufficiency varies only from 0.790 to 0.811, while mean necessity ranges from 0.793 to 0.810. More importantly, all 81 model–dataset–function cells retain positive mean sufficiency and necessity at every horizon. Relative to $H = 96$, mean recovery at $H = 720$ retains 98.8% of its original value, while mean necessity remains essentially unchanged.

Functional specificity exhibits a different pattern. The matched-function rate remains high at $H = 192$ (0.945) and $H = 336$ (0.892), but decreases to 0.762 at $H = 720$. Thus, longer-horizon forecasting does not eliminate the causal relevance of the discovered supports, but makes their function-specific state geometry less sharply separated from competing temporal functions.

The model-wise results in Table 13 reveal that cross-horizon stability is not restricted to a single architecture. SEMPO maintains consistently strong causal recovery and nearly perfect functional specificity across all four horizons. Timer also preserves high sufficiency and necessity through $H = 336$, with a larger decline only at $H = 720$. Chronos shows a complementary pattern: its causal sufficiency and necessity increase with the forecasting horizon, whereas functional specificity gradually decreases from 0.880 at $H = 96$ to 0.580 at $H = 720$. The consistently positive sufficiency and necessity results show that support identity is highly stable across horizons, while the reduction in specificity at long horizons suggests increasing overlap among the causal states associated with different temporal functions. Cross-horizon generalization therefore appears stronger at the level of causal support than at the level of fine-grained functional state separation.