UNDERSTANDING AND IMPROVING ZERO-SHOT FORECASTING THROUGH ATTENTION-HEAD FUNCTIONS IN TIME-SERIES FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformer-based Time-Series Foundation Models (T²SFMs) enable zero-shot forecasting across diverse domains by inferring temporal characteristics from unseen input contexts. However, how such capability is internally organized and utilized for forecasting remains poorly understood. Since self-attention performs input-dependent computation, attention heads provide a natural unit for investigating this question. We therefore introduce **FBC**, a Frequency-feature Based Causal framework for characterizing attention-head functions, revealing that individual heads are multifunctional yet functionally redundant, their functions exhibit both cross-domain stability and variation, and scaling functional heads causally alters forecasts. Together, these findings reveal how attention-head functions are organized and support zero-shot forecasting. Building on these, we propose **SHIFT**, a two-stage Soft-gated Head-wise Intervention Framework for improving T²SFMs' zero-shot forecasting capability. In calibration stage, SHIFT's soft-gated head-wise calibration module learns intervention vectors across observed domains and constructs a source-domain centroid bank. In inference stage, centroid-based retrieval module identifies a compatible source domain for unseen targets, and transfers its calibrated intervention vectors to modulate attention-head outputs, thereby enhancing zero-shot forecasting. Experiments on six T²SFMs across seven datasets demonstrate consistent zero-shot improvements, with average MSE and MASE reductions of up to 14.5% and 65.8%, respectively.

1 INTRODUCTION

Transformer-based time series foundation models (T^2SFMs) acquire zero-shot forecasting capability through pre-training on large and heterogeneous collections of time series (Das et al., 2023; Rasul et al., 2023; Ansari et al., 2024; Woo et al., 2024; Wang et al., 2025a), making them particularly appealing in data-scarce settings where task-specific training data are limited (Nie et al., 2022; Liu et al., 2024a; Dong et al., 2023; Shabani et al., 2022; Liu et al., 2024b; Zhou et al., 2021). Zero-shot forecasting relies on a model's ability to infer temporal characteristics from an unseen input context and adjust its predictions accordingly. However, how this capability is internally organized and utilized for forecasting remains poorly understood (Wiliński et al., 2024; Bao et al., 2026).

In Transformer architectures, self-attention performs input-dependent computation by dynamically aggregating information from the current context (Vaswani et al., 2017). Studies of Transformer language models further show that individual attention heads can support in-context behavior and carry task-relevant functional representations (Olsson et al., 2022; Todd et al., 2024). These observations motivate us to explore how attention heads contribute to zero-shot forecasting in T^2SFMs.

To investigate this question, we introduce **FBC**, a **F**requency-feature **B**ased **c**ausal framework for characterizing attention-head functions. Specifically, FBC measures the causal influence of individual heads and characterizes their roles in modeling temporal components using a proposed functional score. Using FBC, our analysis reveals three key properties. *Entanglement:* individual heads contribute to multiple temporal components while exhibiting a dominant role, and similar functions are redundantly distributed across heads. *Adaptivity:* head functions exhibit cross-domain stability alongside domain-dependent variation, with some heads shifting their dominant roles across do-

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.

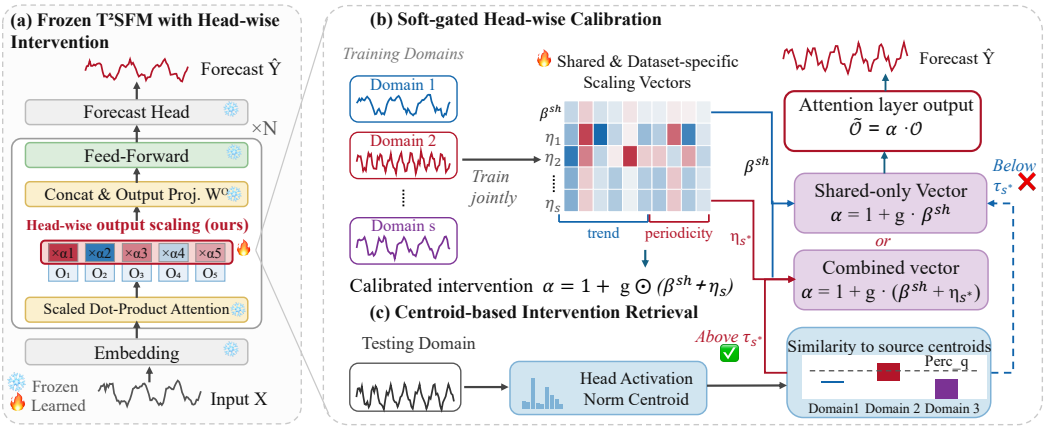

Figure 1: **Overview of SHIFT. (a)** A pretrained T²SFM is kept frozen, while learnable head-wise scaling is applied to attention-head outputs. **(b)** In the calibration stage, SHIFT learns a shared intervention vector β^{sh}, domain-specific vectors η_s, and soft gates $g_{l,h}$ from observed source domains. **(c)** In the inference stage, SHIFT computes a target-domain activation centroid and retrieves a compatible source-domain intervention when similarity exceeds a source-specific threshold.

mains. *Controllability:* scaling functional heads causally alters forecasts in accordance with their temporal roles, providing a direct interface for modulating T²SFM predictions. Together, these findings reveal how attention-head functions are organized and support zero-shot forecasting.

Building on these findings, we propose **SHIFT**, a two-stage **S**oft-gated **H**ead-wise **I**ntervention **F**ramework for improving **T**²SFMs' zero-shot forecasting capability (Figure 1). SHIFT leverages the identified head-level properties to learn lightweight and transferable interventions while keeping the pretrained T²SFM frozen. In the calibration stage, its soft-gated head-wise calibration module learns transferable head-wise interventions across observed source domains. In the inference stage, its centroid-based retrieval module selects a compatible calibrated intervention based on target–source similarity and transfers it to modulate attention-head outputs, thereby improving the T²SFM's zero-shot forecasting on the target domain. This enables direct deployment on unseen domains without target labels or target-specific optimization. We evaluate SHIFT on six representative T²SFMs across seven unseen datasets, achieving average MSE and MASE reductions of up to 14.5% and 65.8%, respectively. Further analysis shows that the learned interventions mitigate structured forecasting errors through functionally aligned head modulation. Our contributions are as follows:

- We introduce **FBC**, a Frequency-feature based causal framework for characterizing attention-head functions in T²SFMs, and identify three properties relevant to zero-shot forecasting: *entanglement*, *adaptivity*, and *controllability*. These findings provide a functional perspective on how attention heads support context-dependent forecasting behavior.
- We propose **SHIFT**, a lightweight two-stage head-wise intervention framework that improves zero-shot forecasting on unseen domains while keeping the pretrained T²SFM frozen. Its design enables transferable head-wise modulation without requiring target labels or target-specific optimization.
- Extensive experiments on six T²SFMs across seven unseen target datasets demonstrate consistent zero-shot improvements across diverse models and domains, with average MSE and MASE reductions of up to 14.5% and 65.8%, respectively.

2 Preliminary

We first formalize the attention-head representation used throughout our analysis. Building on this representation, we then present **FBC**, our frequency-feature based causal framework for characterizing attention-head functions. FBC quantifies the causal influence of individual heads through mean ablation and analyzes the spectral structure of their forecast contributions using a proposed functional score, supporting the subsequent analyses of *entanglement*, *adaptivity*, and *controllability*. For our analysis, we treat each dataset as a separate domain, as datasets from the same application area may still exhibit distinct temporal characteristics, such as periodicity and trend.

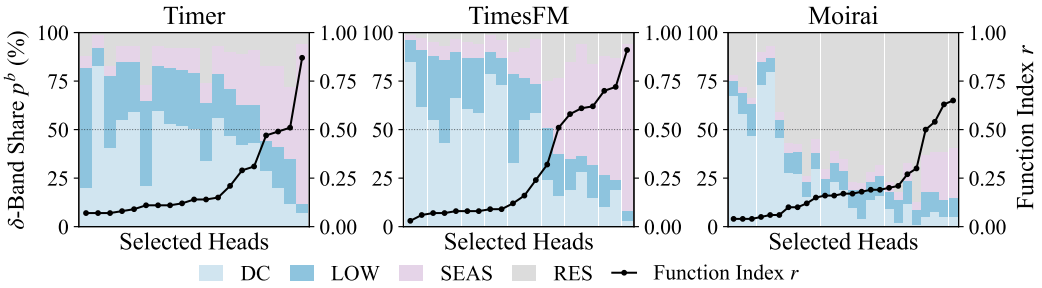

Figure 2: **Attention-head functions exhibit entanglement.** Each bar shows the normalized energy distribution $p^b_{l,h}$ across DC, LOW, SEAS, and RES bands, with heads ordered by functional score $r_{l,h}$. Multi-band contributions indicate multifunctionality, while similar profiles across heads indicate functional redundancy.

2.1 Attention-Head Representation

Consider a Transformer-based T^{2}SFM $\mathcal{M}_\theta$ with L attention layers and H_l heads in layer l. Let $X_l \in \mathbb{R}^{n_l \times d_l}$ denote the sequence representation entering the attention module of layer l, where n_l is the input context length and d_l is the hidden dimension. For head $h \in \{1, \ldots, H_l\}$, the query, key, and value representations are $Q_{l,h} = X_l W^Q_{l,h}$, $K_{l,h} = X_l W^K_{l,h}$, and $V_{l,h} = X_l W^V_{l,h}$, where $W^Q_{l,h}$, $W^K_{l,h}$, and $W^V_{l,h}$ are the head-specific projection matrices. The output of head (l, h) is

$$O_{l,h} = \mathrm{softmax}\left(\frac{Q_{l,h}K_{l,h}^\top}{\sqrt{d_{l,h}}}\right) V_{l,h}, \tag{1}$$

where $d_{l,h}$ is the head hidden dimension. The outputs of all heads are concatenated and mapped through the output projection:

$$\mathrm{MHA}_l(X_l) = [O_{l,1} \,\|\, \cdots \,\|\, O_{l,H_l}]\, W^O_l. \tag{2}$$

The resulting multi-head attention output is incorporated into the residual stream and propagated through the subsequent feed-forward computation, allowing changes to individual head outputs $O_{l,h}$ to influence later representations and ultimately the forecast.

2.2 Mean Ablation for Causal Head Analysis

We quantify the causal influence of individual attention heads on forecasting using mean ablation (Wang et al., 2022). For each head (l, h) and each domain $\mathcal{D}_s$, we replace its input-dependent output $O_{l,h}(x)$ with the domain-specific mean activation $\bar{O}_{s,l,h} = \frac{1}{|\mathcal{R}_s|}\sum_{x \in \mathcal{R}_s} O_{l,h}(x)$, where $\mathcal{R}_s$ is a reference set sampled from the corresponding domain. The resulting ablated forecast is

$$\hat{y}^{\,\mathrm{ablate}(l,h)}(x) = \mathcal{M}_\theta\big(x;\, O_{l,h}(x) \leftarrow \bar{O}_{s,l,h}\big)\,, \tag{3}$$

where all other model computations remain unchanged. We quantify the causal forecasting effect as

$$\Delta\mathrm{MSE}_{l,h} = \mathrm{MSE}\Big(\hat{y}^{\,\mathrm{ablate}(l,h)}\Big) - \mathrm{MSE}(\hat{y})\,, \tag{4}$$

where $\hat{y}$ denotes the original forecast. Using a fixed prediction length H of 96, we rank heads by the magnitude of their causal effect, $|\Delta\mathrm{MSE}_{l,h}|$, and retain the top-12 heads from each domain for subsequent functional characterization. Their union forms the model-specific analysis set $\mathcal{H}$, containing $19/64$ heads for Timer (Liu et al., 2024c), $20/320$ for TimesFM (Das et al., 2023), and $25/384$ for Moirai (Woo et al., 2024) as shown in Appendix Tables 4–6. We use $\mathcal{H}$ throughout the following analyses of temporal functions, cross-domain variation, and causal controllability.

2.3 Frequency-Feature based Functional Characterization

To characterize the temporal functions of heads in $\mathcal{H}$, we analyze the spectral structure of their forecast contributions. For head (l, h), we define the ablation-induced forecast difference as $\delta_{l,h}(x) = \hat{y}(x) - \hat{y}^{\,\mathrm{ablate}(l,h)}(x)$. Let $A_f(\delta)$ denote the Fourier amplitude at frequency f.

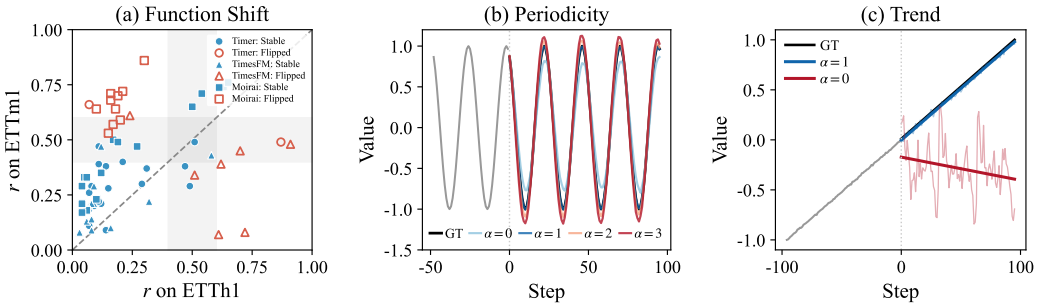

Figure 3: **Adaptivity and controllability of attention-head functions. (a)** Functional scores $r_{l,h}$ on ETTh1 and ETTm1, showing cross-domain stability and variation. **(b)** Scaling the TimesFM periodicity-dominant head $L19H12$ alters the amplitude and phase of sinusoidal forecasts. **(c)** Scaling the trend-dominant head $L0H11$ alters the predicted trend and level.

We first apply the fast Fourier transform to $\delta_{l,h}(x)$ and partition the spectrum into four bands, $b \in \{\mathrm{DC}, \mathrm{LOW}, \mathrm{SEAS}, \mathrm{RES}\}$, representing the signal level, low-frequency variation, dominant seasonal frequencies and their harmonics, and the remaining frequencies (Xu et al., 2024; Wang et al., 2025b). The normalized energy in band b is

$$p_{l,h}^{b} = \left\langle \frac{\sum_{f \in b} A_f(\delta_{l,h})^2}{\sum_f A_f(\delta_{l,h})^2} \right\rangle_x, \tag{5}$$

where $\langle \cdot \rangle_x$ denotes averaging over input contexts. We then define the functional score

$$r_{l,h} = \frac{p_{l,h}^{\mathrm{SEAS}}}{p_{l,h}^{\mathrm{DC}} + p_{l,h}^{\mathrm{LOW}} + p_{l,h}^{\mathrm{SEAS}}}. \tag{6}$$

Larger $r_{l,h}$ indicates a stronger periodic contribution and corresponds to a *periodicity-dominant* head, whereas smaller $r_{l,h}$ indicates a stronger level or low-frequency contribution and corresponds to a *trend-dominant* head. Heads near $r_{l,h} = 0.5$ are treated as neutral. We exclude the RES band here since it captures high-frequency noise. Using this framework, we analyze head functions on two datasets with distinct temporal characteristics, ETTh1 and ETTm1 (Zhou et al., 2021), which exhibit dominant periodicities of 24 and 96 steps (Lin et al., 2024).

Finding 1—Entanglement: Attention-head functions are multifunctional and redundant. Figure 2 reveals two complementary aspects of attention-head functionality in ETTh1. First, individual heads are multifunctional rather than specialized to a single temporal component. Periodicity-dominant heads ($r_{l,h} \to 1$) primarily affect dominant seasonal frequencies while still contributing to low-frequency components. Conversely, trend-dominant heads ($r_{l,h} \to 0$) primarily affect the signal level and low-frequency variation while still contributing to seasonal frequencies. Second, functional redundancy exists across heads within each T^{2}SFM. Multiple heads exhibit similar spectral profiles and share the same dominant role, including several periodicity-dominant and trend-dominant heads within the same model. We refer to this combination of within-head multifunctionality and across-head functional redundancy as *functional entanglement*.

Finding 2—Adaptivity: Head functions exhibit cross-domain stability and variation. We compare the functional score $r_{l,h}$ of each head between ETTh1 and ETTm1. As shown in Figure 3(a), many heads remain functionally stable across domains, with $|\Delta r_{l,h}| < 0.2$ ($13/19$ for Timer, $12/20$ for TimesFM, and $9/25$ for Moirai), while others preserve the same dominant role but exhibit larger shifts in functional strength. Meanwhile, a smaller subset undergoes a functional flip. We count a flip only when a head is clearly biased toward one function on one domain ($r_{l,h} < 0.4$ or $r_{l,h} > 0.6$) and crosses to the opposite side of $r = 0.5$ on the other. Under this criterion, flips occur for $2/19$ heads in Timer, $7/20$ in TimesFM, and $10/25$ in Moirai. Appendix Table 9 reports results for all six models. These results show that attention-head functions exhibit both cross-domain stability and domain-dependent variation, which we refer to as *functional adaptivity*.

Finding 3—Controllability: Scaling functional heads causally modulates forecasts. We test whether directly scaling a functional head changes the final forecast while keeping the rest of the model fixed. For head (l, h), we apply

$$\widetilde{O}_{l,h} = \alpha_{l,h} O_{l,h}, \tag{7}$$

where $\alpha_{l,h} = 1$ recovers the original model and $\alpha_{l,h} = 0$ removes the head output. Figure 3(b) shows that scaling the periodicity-dominant TimesFM head $L19H12$ systematically changes the amplitude and phase of forecasts on a synthetic sinusoidal sequence. Likewise, scaling the trend-dominant head $L0H11$ alters the extrapolated trend and overall level on a synthetic trend sequence (Figure 3(c)). These interventions demonstrate that individual heads can causally modulate forecasts in ways consistent with their dominant temporal functions, establishing their *controllability*.

Overall, these findings show that attention heads in T^{2}SFMs are functionally entangled, exhibit both functional stability and variation across domains, and can be causally modulated through scaling. Together, they provide a functional view of how attention heads contribute to zero-shot forecasting.

3 METHODOLOGY

Motivated by the findings above, we propose **SHIFT**, a **S**oft-gated **H**ead-wise **I**ntervention **F**ramework for improving **T**^{2}SFMs' zero-shot forecasting capability (Figure 1). SHIFT consists of two stages: soft-gated head-wise calibration on observed source domains and activation-based intervention retrieval for unseen targets at inference time.

3.1 SOFT-GATED HEAD-WISE CALIBRATION

Let $\mathcal{D}_{\text{src}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_{n_s}\}$ denote the observed source domains, where each $\mathcal{D}_s$ contains forecasting pairs (x, y). Given a pretrained T^{2}SFM $\mathcal{M}_\theta$, SHIFT learns only a small number of head-wise intervention parameters from $\mathcal{D}_{\text{src}}$, while keeping all pretrained model parameters frozen.

Head-wise intervention design. Our controllability analysis in Finding 3 shows that attention-head outputs can be causally modulated through multiplicative scaling (Eq. 7). Building on this intervention form, SHIFT assigns each head (l, h) in source domain $\mathcal{D}_s$ a domain-dependent intervention coefficient $\alpha_{s,l,h}$. Our analysis further shows that attention-head functions exhibit *adaptivity*, with both cross-domain stability and domain-dependent variation. We therefore decompose $\alpha_{s,l,h}$ into shared and domain-specific intervention components. We first define a shared intervention vector

$$\boldsymbol{\beta}^{\text{sh}} = [\beta^{\text{sh}}_{l,h}]_{(l,h)\in\mathcal{A}}, \tag{8}$$

where $\mathcal{A} = \{(l, h) \mid 1 \le l \le L,\ 1 \le h \le H_l\}$ denote all attention heads, with $N_H = \sum_{l=1}^{L} H_l$. $\boldsymbol{\beta}^{\text{sh}}$ is jointly learned across all observed source domains. For each source domain $\mathcal{D}_s$, we further define a domain-specific intervention vector

$$\boldsymbol{\eta}_s = [\eta_{s,l,h}]_{(l,h)\in\mathcal{A}}. \tag{9}$$

The shared component $\boldsymbol{\beta}^{\text{sh}}$ captures intervention patterns jointly learned across all observed domains, while $\boldsymbol{\eta}_s$ captures domain-specific variation. Meanwhile, the observed *entanglement* of head functions suggests that restricting intervention to a fixed subset of heads may fail to capture complementary and redundant functional contributions across heads. We therefore assign every head a learnable soft gate $g_{l,h}$ to controls the strength of its intervention, and the head set $\mathcal{H}$ identified in Section 2.2 is used for initialization, with $g^{(0)}_{l,h} \approx 1$ for $(l, h) \in \mathcal{H}$ and $g^{(0)}_{l,h} \approx 0$ otherwise.

Overall, the intervention coefficient for head (l, h) in source domain $\mathcal{D}_s$ is

$$\alpha_{s,l,h} = 1 + g_{l,h} \cdot \left(\beta^{\text{sh}}_{l,h} + \eta_{s,l,h}\right). \tag{10}$$

Calibration procedure. During calibration, each mini-batch from source domain $\mathcal{D}_s$ updates the shared intervention vector $\boldsymbol{\beta}^{\text{sh}}$, its corresponding domain-specific vector $\boldsymbol{\eta}_s$, and the parameter of soft gates γ, while the domain-specific vectors of other source domains remain fixed. Consequently, $\boldsymbol{\beta}^{\text{sh}}$ is optimized across all observed domains, whereas each $\boldsymbol{\eta}_s$ is learned only from its corresponding domain. For a mini-batch $\mathcal{B}_s$ from $\mathcal{D}_s$, we optimize

$$\mathcal{L} = \frac{1}{|\mathcal{B}_s|} \sum_{(x,y)\in\mathcal{B}_s} (\|y - \hat{y}_s(x)\|_2^2 + \lambda_g \sum_{(l,h)\in\mathcal{H}} g_{l,h}), \tag{11}$$

where $\hat{y}_s(x)$ denotes the forecast produced using the intervention defined in Eq. 10. The first term minimizes the forecasting error, while the second encourages sparse head-wise intervention. All pretrained T^{2}SFM parameters remain frozen throughout calibration.

Source activation bank. Each source domain has a calibrated intervention. To transfer these interventions to unseen targets without target-specific optimization, we construct a source activation bank in the attention-head space. Since SHIFT directly intervenes on attention-head outputs, we use their activation norms to characterize the model responses to different domains. We then use the similarity between these head-activation patterns as a proxy for determining whether a source-domain intervention is suitable for an unseen target. For each source domain $\mathcal{D}_s$, we sample a set of input contexts $\mathcal{C}_s$. Given a context $x_s \in \mathcal{C}_s$ and a selected head $(l, h) \in \mathcal{H}$, we compute its average output norm across input positions as

$$\xi_{l,h}^{(x_s)} = \frac{1}{n_l} \sum_{t=1}^{n_l} \|O_{l,h}(x_s)_t\|_2 , \qquad \boldsymbol{\xi}^{(x_s)} = \left(\xi_{l_1,h_1}^{(x_s)}, \xi_{l_2,h_2}^{(x_s)}, \dots\right) \in \mathbb{R}^{|\mathcal{H}|}. \tag{12}$$

where $|\mathcal{H}|$ denotes the number of selected heads in $\mathcal{H}$, and $\boldsymbol{\xi}^{(x_s)}$ collects the activation norms of the selected heads for input context x_s. We then summarize each source domain by its activation-norm centroid:

$$\boldsymbol{\mu}_s = \frac{1}{|\mathcal{C}_s|} \sum_{x_s \in \mathcal{C}_s} \boldsymbol{\xi}^{(x_s)}. \tag{13}$$

To determine whether a target domain is sufficiently compatible with a source domain, we further define a source-specific similarity threshold using the q-th percentile of within-source sample-to-centroid similarities:

$$\tau_s = \mathrm{Perc}_q \left(\left\{ \cos\left(\boldsymbol{\xi}^{(x_s)}, \boldsymbol{\mu}_s\right) \,\middle|\, x_s \in \mathcal{C}_s \right\} \right), \tag{14}$$

where $\mathrm{Perc}_q(\cdot)$ denotes the q-th percentile. The resulting source centroids and compatibility thresholds are stored together with the corresponding calibrated domain-specific interventions for inference-time retrieval.

3.2 Activation-Centroid-Based Intervention Retrieval

Before forecasting an unseen target domain, SHIFT performs an offline domain-level retrieval step to determine whether a calibrated domain-specific intervention can be transferred. Given a set of historical input contexts $\mathcal{C}_{\mathrm{tgt}}$, we compute the target activation-norm centroid as

$$\boldsymbol{\mu}_{\mathrm{tgt}} = \frac{1}{|\mathcal{C}_{\mathrm{tgt}}|} \sum_{x_{\mathrm{tgt}} \in \mathcal{C}_{\mathrm{tgt}}} \boldsymbol{\xi}^{(x_{\mathrm{tgt}})}, \tag{15}$$

where $\boldsymbol{\xi}^{(x_{\mathrm{tgt}})}$ is constructed from the same selected head set $\mathcal{H}$. $\mathcal{C}_{\mathrm{tgt}}$ contains only historical input contexts available before forecasting; no target labels or future values are used to construct $\boldsymbol{\mu}_{\mathrm{tgt}}$. We then compare $\boldsymbol{\mu}_{\mathrm{tgt}}$ with the pre-computed source-domain centroids $\{\boldsymbol{\mu}_s\}_{s=1}^{n_s}$ and retrieve the most similar source:

$$s^\star = \underset{s \in \{1,\dots,n_s\}}{\arg\max} \cos(\boldsymbol{\mu}_{\mathrm{tgt}}, \boldsymbol{\mu}_s) . \tag{16}$$

We further apply the source-specific compatibility threshold $\tau_{s^\star}$ to determine whether the retrieved domain-specific intervention should be transferred. For each attention head $(l, h) \in \mathcal{A}$, the target intervention coefficient is

$$\alpha_{\mathrm{tgt},l,h} = \begin{cases} 1 + g_{l,h} \cdot \left(\beta_{l,h}^{\mathrm{sh}} + \eta_{s^\star,l,h}\right), & \cos(\boldsymbol{\mu}_{\mathrm{tgt}}, \boldsymbol{\mu}_{s^\star}) \geq \tau_{s^\star}, \\ 1 + g_{l,h} \cdot \beta_{l,h}^{\mathrm{sh}}, & \text{otherwise.} \end{cases} \tag{17}$$

Thus, when the retrieved source passes the compatibility threshold, the target uses both the shared and retrieved domain-specific intervention components; otherwise, it uses only the shared component. No model or intervention parameters are updated during target forecasting.

4 Experiments

4.1 Experimental settings

We evaluate six representative TSFMs—Timer (Liu et al., 2024c), TimesFM (Das et al., 2023), Chronos (Ansari et al., 2024), Moirai (Woo et al., 2024), Time-MoE (Shi et al., 2025), and

Table 1: TSLib zero-shot transfer results measured by MSE (↓) at prediction length $H \in \{96, 192, 336, 720\}$. "Vanilla" denotes the original frozen TSFM, while "Ours" denotes the proposed head-wise scaling method. $\Delta(\%) = 100(\text{Ours} - \text{Vanilla})/\text{Vanilla}$, where negative values indicate improvement. "–" denotes results excluded due to pre-training data overlap.

		Timer			TimesFM			Chronos			Moirai			Time-MoE			SEMPO		
Dataset	H	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)
ETTm2	96	0.191	**0.176**	**-7.9**	0.180	**0.168**	**-6.7**	0.167	**0.158**	**-5.4**	0.191	**0.175**	**-8.4**	0.184	**0.171**	**-7.1**	0.237	**0.224**	**-5.5**
	192	0.245	**0.229**	**-6.5**	0.247	**0.231**	**-6.5**	0.238	**0.215**	**-9.7**	0.257	**0.239**	**-7.0**	0.254	**0.228**	**-10.2**	0.305	**0.283**	**-7.2**
	336	0.296	**0.278**	**-6.1**	0.310	**0.286**	**-7.7**	0.306	**0.263**	**-14.1**	0.325	**0.299**	**-8.0**	0.315	**0.279**	**-11.4**	0.354	**0.330**	**-6.8**
	720	0.384	**0.356**	**-7.3**	0.411	**0.371**	**-9.7**	0.416	**0.327**	**-21.4**	0.432	**0.388**	**-10.2**	0.370	**0.350**	**-5.4**	0.430	**0.411**	**-4.4**
	Avg	0.279	**0.260**	**-6.8**	0.287	**0.264**	**-8.0**	0.282	**0.241**	**-14.5**	0.301	**0.275**	**-8.6**	0.281	**0.257**	**-8.5**	0.331	**0.312**	**-5.7**
ETTh2	96	0.280	**0.277**	**-1.1**	0.281	**0.266**	**-5.3**	0.272	**0.257**	**-5.5**	0.286	**0.278**	**-2.8**	0.265	**0.261**	**-1.5**	0.338	**0.315**	**-6.8**
	192	0.328	**0.323**	**-1.5**	0.342	**0.317**	**-7.3**	0.333	**0.310**	**-6.9**	0.334	**0.323**	**-3.3**	0.314	**0.305**	**-2.9**	0.426	**0.392**	**-8.0**
	336	0.339	**0.329**	**-2.9**	0.351	**0.326**	**-7.1**	0.357	**0.325**	**-9.0**	0.345	**0.331**	**-4.1**	0.336	**0.316**	**-6.0**	0.454	**0.421**	**-7.3**
	720	0.390	**0.381**	**-2.3**	0.370	**0.351**	**-5.1**	0.396	**0.361**	**-8.8**	0.398	**0.362**	**-9.0**	0.393	**0.360**	**-8.4**	0.456	**0.438**	**-3.9**
	Avg	0.334	**0.327**	**-2.1**	0.336	**0.315**	**-6.3**	0.339	**0.313**	**-7.7**	0.341	**0.324**	**-5.0**	0.327	**0.310**	**-5.2**	0.418	**0.391**	**-6.5**
Weather	96	0.169	0.171	+1.2	0.151	**0.149**	**-1.3**	0.157	**0.153**	**-2.5**	0.180	**0.176**	**-2.2**	0.156	**0.155**	**-0.6**	0.204	**0.197**	**-3.4**
	192	0.217	**0.213**	**-1.8**	0.202	**0.196**	**-3.0**	0.204	**0.196**	**-3.9**	0.221	**0.218**	**-1.4**	0.210	0.210	+0.0	0.268	**0.256**	**-4.5**
	336	0.271	**0.260**	**-4.1**	0.269	**0.256**	**-4.8**	0.260	**0.243**	**-6.5**	0.277	**0.276**	**-0.4**	0.271	**0.270**	**-0.4**	0.317	**0.304**	**-4.1**
	720	0.356	**0.327**	**-8.1**	0.370	**0.347**	**-6.2**	0.345	**0.310**	**-10.1**	0.360	0.364	+1.1	0.351	**0.347**	**-1.1**	0.377	**0.367**	**-2.7**
	Avg	0.253	**0.243**	**-4.0**	0.248	**0.237**	**-4.4**	0.241	**0.225**	**-6.6**	0.260	**0.259**	**-0.4**	0.247	**0.246**	**-0.4**	0.292	**0.281**	**-3.8**
Exchange	96	0.093	**0.091**	**-2.2**	0.083	**0.082**	**-1.2**	0.095	**0.092**	**-3.2**	0.085	**0.084**	**-1.2**	–	–	–	0.123	0.123	+0.0
	192	0.178	**0.172**	**-3.4**	0.174	**0.168**	**-3.4**	0.196	**0.189**	**-3.6**	0.169	**0.168**	**-0.6**	–	–	–	0.254	**0.249**	**-2.0**
	336	0.300	**0.283**	**-5.7**	0.327	**0.311**	**-4.9**	0.364	**0.349**	**-4.1**	0.298	**0.294**	**-1.3**	–	–	–	0.428	**0.417**	**-2.6**
	720	0.669	**0.600**	**-10.3**	0.877	0.879	+0.2	0.964	**0.934**	**-3.1**	0.739	**0.697**	**-5.7**	–	–	–	1.016	**1.008**	**-0.8**
	Avg	0.310	**0.286**	**-7.7**	0.365	**0.360**	**-1.4**	0.405	**0.391**	**-3.5**	0.323	**0.311**	**-3.7**	–	–	–	0.455	**0.449**	**-1.3**

SEMPO (He et al., 2026)—on TSLib (Wu et al., 2022) and GIFT-Eval (Aksu et al., 2024). For TSLib, ETTm1, ETTh1, and ECL are used as observed datasets, while ETTm2, ETTh2, Weather, and Exchange are used as unseen targets. For GIFT-Eval, BizITObs-L2C, BizITObs-Service, M_DENSE, and Loop Seattle form the observed set, while BizITObs_Application, SZ-Taxi, and BizITObs_Fast-Storage are used for zero-shot evaluation. We randomly sample 3,000 training windows from each observed dataset to learn the head-wise scaling parameters. Appendix Table 3 summarizes the complete dataset we adopted. Our evaluation focuses on zero-shot forecasting on unseen target datasets, and we report MSE and MAE on TSLib, and MASE and CRPS on GIFT-Eval.

4.2 Results on the TSLib benchmark

Table 1 reports the zero-shot forecasting results on unseen TSLib datasets, with corresponding MAE results provided in Appendix Table 12. Without using target training data or target-specific optimization, the proposed head-wise scaling consistently improves forecasting performance across the six TSFMs. On ETTm2, average MSE is reduced by 5.7%–14.5% across all models, while ETTh2 also shows consistent improvements. Gains are further observed on Weather and Exchange, with average MSE reductions of up to 6.6% and 7.7%, respectively. These results demonstrate that head-wise adjustments learned from a limited collection of observed datasets can effectively transfer to unseen targets and improve the zero-shot forecasting capability of pretrained TSFMs. We also report the forecasting results on the observed source datasets in Appendix Table 10.

4.3 Results on GIFT-Eval

Table 2 reports the zero-shot forecasting results on GIFT-Eval across diverse sampling frequencies and forecast lengths, with corresponding CRPS results provided in Appendix Table 14. Without target training data or target-specific optimization, the proposed head-wise scaling improves forecasting performance across multiple unseen datasets. On BizITObs-Application, activation-norm retrieval selects BizITObs-Service as the source dataset, yielding average MASE reductions of 33.1%, 65.8%, 37.2%, and 45.3% for Timer, Chronos, Moirai, and Time-MoE, respectively. Additional gains are observed on SZ-Taxi and Bitbrains-Fast-Storage, where Timer reduces average MASE by 6.6% and 2.8%, respectively. These results further demonstrate that head-wise adjustments learned from observed datasets can transfer to unseen targets and improve zero-shot forecasting across diverse datasets. We also report forecasting results on the observed datasets in Appendix Table 11.

Table 2: GIFT-Eval zero-shot transfer results measured by MASE (↓) over short-, medium-, and long-horizon forecasts. "Vanilla" denotes the original frozen TSFM, while "Ours" denotes the proposed head-wise scaling method. $\Delta(\%) = 100(\text{Ours} - \text{Vanilla})/\text{Vanilla}$, where negative values indicate improvement. "–" denotes results excluded due to pre-training data overlap.

		Timer			TimesFM			Chronos			Moirai			Time-MoE			SEMPO		
Dataset	H	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)
BizITObs Application	Short	4.684	**4.082**	**-12.9**	1.166	**1.155**	**-1.0**	5.522	**1.939**	**-64.9**	3.166	**2.825**	**-10.8**	3.506	**2.457**	**-29.9**	7.842	**7.807**	**-0.4**
	Medium	12.614	**7.721**	**-38.8**	2.112	**2.094**	**-0.8**	9.577	**2.979**	**-68.9**	7.920	**4.454**	**-43.8**	8.438	**5.118**	**-39.3**	11.395	**11.357**	**-0.3**
	Long	12.384	**8.045**	**-35.0**	3.355	**3.286**	**-2.0**	10.231	**3.741**	**-63.4**	8.317	**4.914**	**-40.9**	13.825	**6.514**	**-52.9**	10.959	**10.585**	**-3.4**
	Avg	9.894	**6.616**	**-33.1**	2.211	**2.178**	**-1.5**	8.443	**2.887**	**-65.8**	6.467	**4.064**	**-37.2**	8.590	**4.696**	**-45.3**	10.065	**9.916**	**-1.5**
SZ-Taxi	Short	0.598	**0.585**	**-2.2**	0.545	0.545	+0.0	0.548	0.548	+0.0	–	–	–	–	–	–	0.623	**0.621**	**-0.3**
	Medium	0.665	**0.605**	**-8.9**	0.554	**0.550**	**-0.8**	0.562	0.562	+0.0	–	–	–	–	–	–	0.596	**0.595**	**-0.2**
	Long	0.638	**0.564**	**-11.5**	0.525	**0.520**	**-0.9**	0.548	**0.544**	**-0.7**	–	–	–	–	–	–	0.562	**0.560**	**-0.4**
	Avg	0.633	**0.585**	**-7.7**	0.541	**0.538**	**-0.6**	0.553	**0.551**	**-0.3**	–	–	–	–	–	–	0.594	**0.592**	**-0.3**
Bitbrains Fast-Storage	Short	0.959	**0.936**	**-2.4**	0.702	**0.691**	**-1.7**	0.743	**0.737**	**-0.9**	0.889	**0.868**	**-2.4**	0.770	**0.768**	**-0.2**	1.327	**1.308**	**-1.4**
	Medium	1.320	**1.263**	**-4.3**	1.037	**0.991**	**-4.6**	1.049	**1.048**	**-0.1**	1.108	**1.089**	**-1.7**	1.175	**1.157**	**-1.5**	1.442	**1.430**	**-0.9**
	Long	1.219	**1.147**	**-5.9**	0.942	**0.895**	**-5.2**	0.952	**0.946**	**-0.6**	0.979	**0.974**	**-0.5**	1.081	**1.054**	**-2.5**	1.383	**1.368**	**-1.1**
	Avg	1.166	**1.116**	**-4.3**	0.894	**0.859**	**-4.0**	0.915	**0.910**	**-0.5**	0.992	**0.977**	**-1.5**	1.009	**0.993**	**-1.5**	1.384	**1.369**	**-1.1**

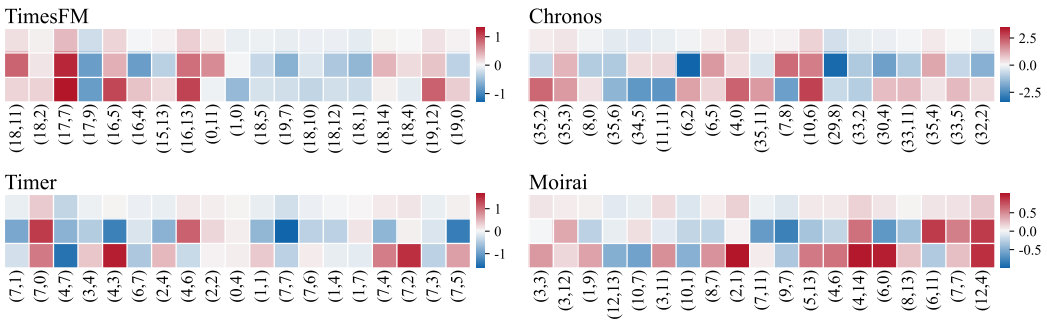

Figure 4: **Gated head-wise intervention components for four T^{2}SFMs.** Rows show the shared, ETTh1-specific, and ETTm1-specific components, and columns show selected heads ordered by the ETTm1 functional score $r_{l,h}$. Domain-specific components exhibit larger and more distinct head-wise patterns than the shared component.

4.4 Mechanistic Analysis

To understand the working mechanism of our framework, we further analyze the learned shared and dataset-specific adjustments, $g_{l,h}\beta^{\text{sh}}_{l,h}$ and $g_{l,h}\eta_{s,l,h}$, together with the functional index r defined in Section 2. Head functions are profiled on ETTh1 and ETTm1, with complete per-head statistics reported in Appendix Scetion 6.4.

The shared vector provides mild cross-dataset recalibration. As shown in Figure 4, the mean absolute shared adjustment ranges from 0.008 to 0.052 across five TSFMs, approximately 3–6$\times$ smaller than the corresponding dataset-specific adjustments, with no entry exceeding 0.55. The shared adjustments are concentrated on trend-dominant heads: Timer and TimesFM allocate 59% and 66% of their adjustment mass, respectively, to heads with $r < 0.4$ on both ETT datasets. Their largest shared adjustments also occur primarily on such heads, whereas the strongest periodicity-dominant heads receive adjustments of at most $+0.14$. This suggests that the shared vector mainly performs a mild recalibration of trend-dominant heads.

Dataset-specific vectors capture source-dependent temporal structure. The learned dataset-specific vectors differ substantially between datasets in Figure 4. In ETTm1 and ETTh1 dataset, their cosine similarity is -0.05 for Timer, 0.00 for Chronos, and $+0.05$ for Moirai. Moreover, $8/35$, $4/20$, and $8/25$ of the function-profiled heads in Timer, TimesFM, and Moirai, respectively, receive adverse adjustments from the two datasets when both adjustment magnitudes exceed 0.1. These differences align with dataset-dependent changes in head function. For Moirai, $\boldsymbol{\eta}_{\text{ETTm1}}$ allocates 70% of its total adjustment mass, and 77% of its positive adjustment mass, to heads that are periodicity-dominant on the 15-minute ETTm1 data. It amplifies $L2H1$ ($+1.00$), $L4H14$ ($+0.98$), and $L6H0$ ($+0.96$), whose r_{ETTm1} values range from 0.64 to 0.86. Conversely, $\boldsymbol{\eta}_{\text{ETTh1}}$ emphasizes heads that are periodicity-dominant on the hourly ETTh1 data, including $L6H11$ ($+0.82$) and $L7H7$ ($+0.53$),

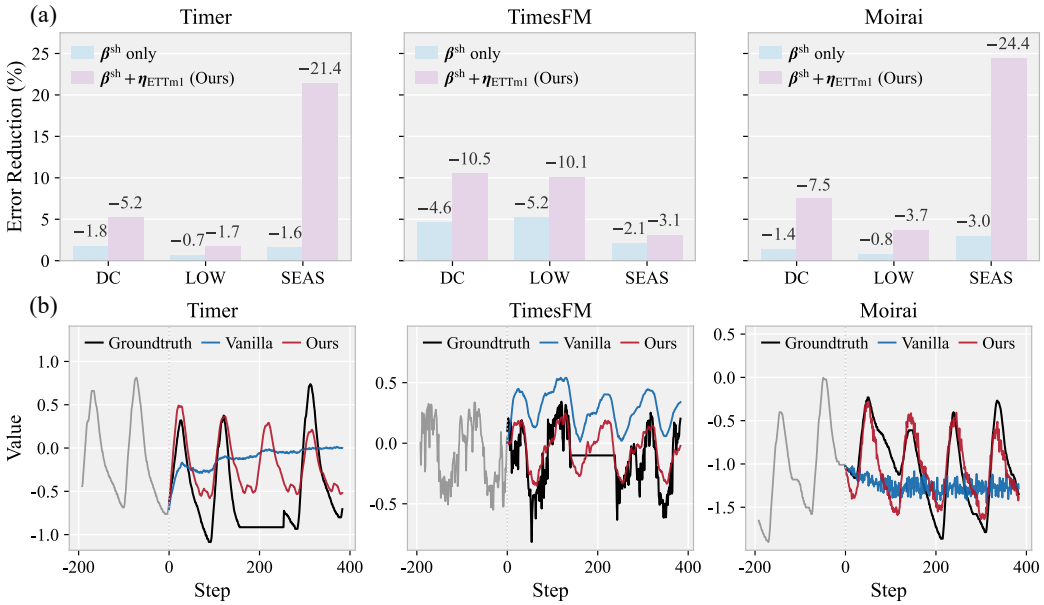

Figure 5: Mechanistic case study on ETTm2 under zero-shot transfer. **(a)** Frequency-wise error reduction relative to the vanilla model for shared-only and full transfer. Adding the retrieved $\boldsymbol{\eta}_{\mathrm{ETTm1}}$ yields larger corrections across structured frequency bands. **(b)** Representative examples. Scaling restores daily periodicity for Timer and Moirai and reduces trend-related deviation for TimesFM.

while suppressing $L6H0$ (-0.65), which is amplified by the ETTm1-specific vector. Timer and TimesFM also exhibits a similar pattern.

Case study: zero-shot transfer to ETTm2. We examine how the learned head-wise adjustments improve zero-shot forecasting on ETTm2, for which activation-norm retrieval selects ETTm1 as the source dataset. Under forecasting horizon $H = 384$, we decompose forecast-error energy into the DC, LOW, and SEAS bands defined in Section 2.3, and compare the vanilla model, shared-only scaling, and full transfer with the retrieved ETTm1-specific vector. In Figure 5a, **our full method reduces errors across all three bands for the three backbones**. Timer and Moirai achieve seasonal-error reductions of 21.4% and 24.4%, respectively, while TimesFM reduces DC and LOW errors by 10.5% and 10.1%. The corresponding forecasts in Figure 5b show restored daily oscillations for Timer and Moirai and reduced trend deviation for TimesFM. **These improvements are consistent with the functions of the adjusted heads**. For Moirai, the five largest positive adjustments occur on periodicity-dominant heads ($r = 0.64$–0.86), with 58% of the total absolute adjustment concentrated on heads with $r > 0.6$. In contrast, TimesFM places 82% of its total absolute adjustment on trend-dominant heads, including $L17H7$ ($+1.64$), $L16H13$ ($+1.25$), and $L16H5$ ($+1.21$), all with $r \leq 0.2$, while Timer adjusts both periodicity- and trend-dominant heads. Overall, the dominant head adjustments align with the temporal components most improved in the resulting forecasts.

5 Conclusion

We investigate how attention-head functions contribute to zero-shot forecasting in pretrained T^{2}SFMs. We introduce **FBC**, a frequency-feature-based causal framework that reveals three key properties of attention-head functions: *entanglement*, *adaptivity*, and *controllability*. Building on these findings, we propose **SHIFT**, a lightweight two-stage head-wise intervention framework that transfers calibrated interventions from observed source domains to unseen targets while keeping the pretrained T^{2}SFM frozen. Across six T^{2}SFMs and seven unseen target datasets from TSLib and GIFT-Eval, SHIFT achieves average MSE and MASE reductions of up to 14.5% and 65.8%, respectively. These results provide a functional view of attention heads in zero-shot forecasting and demonstrate their potential as a lightweight intervention interface for T^{2}SFMs.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the implementation of parts of the proposed method. We additionally used generative AI tools for literature review and manuscript editing, including identifying and summarizing relevant literature and improving the clarity and readability of the text. All AI-assisted code was manually reviewed and tested by the authors, and all referenced literature and citations were verified against the original sources. We have reviewed and revised all AI-assisted writing and take full responsibility for the final content, claims, experiments, and artifacts presented in this work.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details required to reproduce our results throughout the paper and appendix. Section 3 specifies the proposed head-wise scaling mechanism, training objective, and activation-norm-based transfer procedure. Section 4.1 describes the evaluated TSFMs, datasets, forecasting settings, data usage, and evaluation metrics. The appendix further reports implementation details, hyperparameters, dataset configurations, and complete experimental results to facilitate reproduction of the reported findings. We view head-wise scaling as a step from interpretability findings toward deployable adaptation; richer per-head interventions, broader source pools, and finer-grained retrieval are natural directions for future work.

REFERENCES

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Anthony Bao, Venkata Hasith Vattikuti, Jeffrey Lai, and William Gilpin. Universal redundancies in time series foundation models. *arXiv preprint arXiv:2602.01605*, 2026.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Abhimanyu Das, Matthew Faw, Rajat Sen, and Yichen Zhou. In-context fine-tuning for time-series foundation models. *arXiv preprint arXiv:2410.24087*, 2024.

Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Mingsheng Long. Simmtm: A simple pre-training framework for masked time-series modeling. *Advances in Neural Information Processing Systems*, 36:29996–30025, 2023.

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.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Yuze Li and Wei Zhu. Trace: Time series parameter efficient fine-tuning. *Neurocomputing*, pp. 132098, 2025.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. *Advances in Neural Information Processing Systems*, 37:106315–106345, 2024.

Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-tao Xia. Timebridge: Non-stationarity matters for long-term time series forecasting. *arXiv preprint arXiv:2410.04442*, 2024a.

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*, volume 2024, pp. 11116–11140, 2024b.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. *arXiv preprint arXiv:2402.02368*, 2024c.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, et al. Lag-llama: Towards foundation models for probabilistic time series forecasting. *arXiv preprint arXiv:2310.08278*, 2023.

Amin Shabani, Amir Abdi, Lili Meng, and Tristan Sylvain. Scaleformer: Iterative multi-scale refining transformers for time series forecasting. *arXiv preprint arXiv:2206.04038*, 2022.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. In *International conference on learning representations*, volume 2025, pp. 34635–34667, 2025.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12694–12702, 2025a.

Hao Wang, Lichen Pan, Yuan Shen, Zhichao Chen, Degui Yang, Yifei Yang, Sen Zhang, Xinggao Liu, Haoxuan Li, and Dacheng Tao. Fredf: Learning to forecast in the frequency domain. In *International Conference on Learning Representations*, volume 2025, pp. 6893–6922, 2025b.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Michał Wiliński, Mononito Goswami, Willa Potosnak, Nina Żukowska, and Artur Dubrawski. Exploring representations and interventions in time series foundation models. *arXiv preprint arXiv:2409.12915*, 2024.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2402.02592*, 2024.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

Zhijian Xu, Ailing Zeng, and Qiang Xu. Fits: Modeling time series with $10k$ parameters. In *International Conference on Learning Representations*, volume 2024, pp. 26295–26318, 2024.

Fangcong Yin, Xi Ye, and Greg Durrett. Lofit: Localized fine-tuning on llm representations. *Advances in Neural Information Processing Systems*, 37:9474–9506, 2024.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

6 Appendix

6.1 Related Work

6.1.1 Time-Series Foundation Models and Adaptation

Transformer-based time-series foundation models (T^2SFMs), including TimesFM (Das et al., 2023), Timer (Liu et al., 2024c), Chronos (Ansari et al., 2024), Moirai (Woo et al., 2024), Time-MoE (Shi et al., 2025), and SEMPO (He et al., 2026), enable zero-shot forecasting across diverse domains through large-scale pretraining. When target-domain examples are available, their performance can be further improved through fine-tuning, parameter-efficient adaptation (Li & Zhu, 2025), or in-context adaptation (Das et al., 2024). These approaches adapt a pretrained model using information from the target domain, either through optimization or inference-time examples.

Our setting instead focuses on improving zero-shot forecasting on unseen domains without target labels or target-specific optimization. SHIFT calibrates lightweight head-wise interventions across observed source domains while keeping the pretrained T^2SFM frozen, and transfers a compatible calibrated intervention to an unseen target based only on its available historical input contexts. This separates source-domain calibration from target-domain deployment and avoids adapting model parameters for each new target.

6.1.2 Attention-Head Analysis and Intervention

Attention heads have been widely studied as functional and intervention units in Transformer language models. ITI (Li et al., 2023) identifies behavior-relevant heads and intervenes on their activations, Function Vectors (Todd et al., 2024) associate attention heads with task-relevant functional representations, and LoFiT (Yin et al., 2024) learns head-level interventions for downstream tasks. In time-series foundation models, prior work has primarily analyzed temporal representations at the layer level, including interpretable structures associated with periodicity and trend (Wiliński et al., 2024). Our work studies temporal functions at the granularity of individual attention heads. We introduce **FBC**, a **F**requency-feature-**B**ased **C**ausal framework that combines causal ablation with frequency-domain characterization to analyze how individual heads contribute to forecasting. This analysis reveals *entanglement*, *adaptivity*, and *controllability* in attention-head functions. Building on these properties, **SHIFT** uses soft-gated head-wise interventions and activation-based retrieval to improve zero-shot forecasting on unseen domains without modifying the pretrained T^2SFM.

6.2 Datasets

TSLib. We use seven datasets from the standard long-term forecasting benchmark (Wu et al., 2022): ETTm1, ETTh1, ECL, ETTm2, ETTh2, Weather, and Exchange. Following the benchmark protocol, all datasets use chronological train/validation/test splits and forecast horizons $H \in \{96, 192, 336, 720\}$. ETTm1, ETTh1, and ECL are used as observed source datasets, while ETTm2, ETTh2, Weather, and Exchange are held out for zero-shot transfer.

GIFT-Eval. We additionally evaluate on uncontaminated streams from GIFT-Eval (Aksu et al., 2024), covering IT observability, datacenter telemetry, and transportation. Evaluation follows the official rolling-window protocol over short-, medium-, and long-horizon forecasts. BizITObs-L2C, BizITObs-Service, M-Dense, and Loop Seattle are used as observed sources; BizITObs-Application, SZ-Taxi, and Bitbrains-Fast-Storage are held out as zero-shot targets. Dataset categories and sampling frequencies are summarized in Table 3.

The source–target split includes both closely related and more distinct transfer cases. In TSLib, ETTm2 and ETTh2 share sampling granularities with observed ETT sources, whereas Weather and

Exchange differ in application domain. In GIFT-Eval, BizITObs-Application belongs to the same application family as the observed BizITObs streams, while SZ-Taxi and Bitbrains-Fast-Storage provide transportation and datacenter targets. This variation allows us to examine whether source-specific interventions are transferred selectively rather than applied uniformly to every unseen target.

Table 3: Datasets used in the TSLib and GIFT-Eval experiments. Observed sources are used to learn the head-wise scaling parameters, while held-out targets are used only for cross-dataset transfer evaluation.

Benchmark	Dataset	Category	Freq.	Role
TSLib	ETTm1	Electricity	15 min	Observed source
	ETTh1	Electricity	1 h	Observed source
	ECL	Electricity	1 h	Observed source
	ETTm2	Electricity	15 min	Held-out target
	ETTh2	Electricity	1 h	Held-out target
	Weather	Weather	10 min	Held-out target
	Exchange	Finance	1 day	Held-out target
GIFT-Eval	BizITObs-L2C	Web/CloudOps	5 min	Observed source
	BizITObs-Service	Web/CloudOps	10 s	Observed source
	M_DENSE	Transport	1 h	Observed source
	Loop Seattle	Transport	5 min	Observed source
	BizITObs-Application	Web/CloudOps	10 s	Held-out target
	SZ-Taxi	Transport	15 min	Held-out target
	Bitbrains-Fast-Storage	Web/CloudOps	5 min	Held-out target

6.3 Experimental Settings and Implementation Details

Backbones. We evaluate six publicly released TSFMs: Timer (`thuml/timer-base-84m`), TimesFM-2.5 (`google/timesfm-2.5-200m-transformers`), Chronos-Bolt (`amazon/chronos-bolt-base`), Moirai (`Salesforce/moirai-1.0-R-large`), Time-MoE (`Maple728/TimeMoE-200M`), and SEMPO. Their respective attention configurations are 8×8, 20×16, 36×12, 24×16, 12×12, and 12 heads. All pretrained model parameters remain frozen, and each model retains its original normalization, decoding, and autoregressive rollout procedure.

Head-wise scaling. For each attention head, we scale its output $O_{l,h}$ before the output projection W_l^O. The trainable parameters consist of a shared scaling vector $\boldsymbol{\beta}^{\mathrm{sh}}$, a source-specific residual $\boldsymbol{\eta}_s$ for each observed dataset, and gate logits $\boldsymbol{\gamma}$. This introduces approximately $(n_s + 2)N$ trainable scalars for a model with N heads and n_s sources, corresponding to only 10^{-4}–10^{-3} of the backbone parameters.

Optimization. The scaling parameters are optimized jointly with Adam using source-dataset training windows only. We sample 3,000–6,000 training windows and 400–1,200 validation windows per source. Learning rates are 10^{-3} for $\boldsymbol{\beta}^{\mathrm{sh}}$ and $\boldsymbol{\gamma}$ and 5×10^{-3} for $\boldsymbol{\eta}_s$. Batch sizes range from 4 to 32, with gradient accumulation for larger backbones, and training runs for 4,000–12,000 steps with early stopping on pooled source-validation loss.

The gates are initialized using the mean-ablation ranking from Section 2.2: selected heads are initialized with $\gamma = +m$ and the remaining heads with $\gamma = -m$, where $m \in \{2, 3\}$. Importantly, this ranking is used only for initialization; all head gates remain trainable. After calibration, the learned intervention parameters are fixed for all held-out targets; no target labels are used for optimization or early stopping. Retrieval uses only target input contexts, as detailed in Section 6.9. All experiments are conducted on a single RTX 4090 GPU.

6.4 Cross-Dataset Head-Function Statistics

Tables 4–8 provide the per-head profiles underlying the aggregate cross-domain analysis. For each of the five band-profiled backbones, we report the functional score r on ETTh1 and ETTm1 together with the resulting role assignment. Reporting the full profiles makes it possible to distinguish

small score changes that preserve a head's dominant role from genuine functional flips across domains. These values support the adaptivity analysis in Section 2.3 and the mechanistic analysis of Section 4.4. Timer is dominated by trend-oriented heads across both ETT domains, while two selected heads undergo functional flips. The supplementary rows additionally report heads used in the shared-vector analysis, providing a broader view of how the learned intervention distributes over the model.

Table 4: Per-head function profiles for Timer's selected head set $\mathcal{H}$. For each head we report the functional score r of Section 2 on ETTh1 and ETTm1 (higher r = more periodicity-dominant, lower = more trend-dominant) and the resulting role. Roles: "trend-dominant" ($r<0.4$ on both datasets), "periodicity-dominant" (periodicity-leaning on both), "flip" (a functional flip under the criterion of Table 9), and "neutral" otherwise. The supplementary block lists additional heads profiled for the shared-vector analysis of Section 4.4.

Head	r (ETTh1)	r (ETTm1)	Role
L7H2	0.87	0.49	flip: periodicity→trend
L7H6	0.47	0.38	neutral
L7H3	0.51	0.49	neutral
L1H1	0.31	0.37	trend-dominant
L1H7	0.21	0.40	neutral
L0H4	0.29	0.30	trend-dominant
L7H7	0.14	0.38	trend-dominant
L4H6	0.49	0.29	neutral
L7H5	0.07	0.66	flip: trend→periodicity
L7H4	0.11	0.47	neutral
L1H4	0.11	0.39	trend-dominant
L7H0	0.07	0.11	trend-dominant
L7H1	0.14	0.09	trend-dominant
L3H4	0.12	0.21	trend-dominant
L2H2	0.08	0.29	trend-dominant
L4H7	0.09	0.20	trend-dominant
L4H3	0.11	0.22	trend-dominant
L2H4	0.15	0.28	trend-dominant
L6H7	0.07	0.26	trend-dominant
Supplementary heads			
L6H6	0.10	0.16	trend-dominant
L5H1	0.06	0.17	trend-dominant
L2H7	0.21	0.40	neutral
L5H4	0.09	0.23	trend-dominant
L6H3	0.07	0.25	trend-dominant
L2H1	0.20	0.38	trend-dominant
L5H7	0.08	0.24	trend-dominant
L4H5	0.08	0.27	trend-dominant
L5H0	0.10	0.24	trend-dominant
L5H2	0.10	0.07	trend-dominant
L3H3	0.11	0.34	trend-dominant
L4H4	0.11	0.36	trend-dominant
L1H3	0.17	0.53	flip: trend→periodicity
L2H6	0.22	0.49	neutral
L4H0	0.11	0.30	trend-dominant
L4H1	0.07	0.27	trend-dominant

Table 9 summarizes the cross-domain role changes across the five profiled backbones, while SEMPO is included only for head-selection statistics. Most selected heads preserve their dominant function between ETTh1 and ETTm1, but every profiled backbone also contains heads whose dominant function changes across datasets. The number of such functional flips ranges from $2/19$ for Timer to $10/25$ for Moirai.

These results suggest that head behavior contains both cross-dataset regularities and dataset-dependent variation. This observation motivates our decomposition into shared scaling and source-specific residuals rather than a single dataset-independent adjustment.

Table 5: Per-head function profiles for TimesFM. For each head we report the functional score r of Section 2 on ETTh1 and ETTm1 (higher r = more periodicity-dominant, lower = more trend-dominant) and the resulting role. Roles: "trend-dominant" ($r<0.4$ on both datasets), "periodicity-dominant" (periodicity-leaning on both), "flip" (a functional flip under the criterion of Table 9), and "neutral" otherwise.

Head	r (ETTh1)	r (ETTm1)	Role
L19H12	0.91	0.48	flip: periodicity→trend
L18H1	0.58	0.43	neutral
L18H12	0.62	0.39	flip: periodicity→trend
L18H14	0.70	0.45	flip: periodicity→trend
L18H10	0.51	0.34	flip: periodicity→trend
L19H0	0.24	0.61	flip: trend→periodicity
L18H2	0.72	0.08	flip: periodicity→trend
L18H11	0.61	0.07	flip: periodicity→trend
L18H4	0.12	0.47	neutral
L18H5	0.32	0.22	trend-dominant
L19H7	0.09	0.29	trend-dominant
L0H11	0.09	0.21	trend-dominant
L18H0	0.03	0.08	trend-dominant
L16H5	0.06	0.13	trend-dominant
L17H7	0.08	0.09	trend-dominant
L16H13	0.07	0.20	trend-dominant
L17H9	0.16	0.10	trend-dominant
L1H0	0.08	0.21	trend-dominant
L16H4	0.08	0.14	trend-dominant
L15H13	0.07	0.19	trend-dominant

Table 6: Per-head function profiles for Moirai. For each head we report the functional score r of Section 2 on ETTh1 and ETTm1 (higher r = more periodicity-dominant, lower = more trend-dominant) and the resulting role. Roles: "trend-dominant" ($r<0.4$ on both datasets), "periodicity-dominant" (periodicity-leaning on both), "flip" (a functional flip under the criterion of Table 9), and "neutral" otherwise.

Head	r (ETTh1)	r (ETTm1)	Role
L0H12	0.06	0.18	trend-dominant
L1H9	0.05	0.33	trend-dominant
L2H1	0.10	0.64	flip: trend→periodicity
L2H3	0.17	0.50	neutral
L3H3	0.04	0.17	trend-dominant
L3H11	0.19	0.49	neutral
L3H12	0.10	0.23	trend-dominant
L3H15	0.06	0.33	trend-dominant
L4H6	0.19	0.70	flip: trend→periodicity
L4H14	0.16	0.71	flip: trend→periodicity
L5H13	0.16	0.68	flip: trend→periodicity
L6H0	0.54	0.71	periodicity-dominant
L6H11	0.63	0.74	periodicity-dominant
L7H7	0.65	0.76	periodicity-dominant
L7H11	0.18	0.64	flip: trend→periodicity
L7H14	0.04	0.21	trend-dominant
L8H7	0.20	0.59	flip: trend→periodicity
L8H13	0.21	0.72	flip: trend→periodicity
L9H7	0.50	0.65	periodicity-dominant
L10H1	0.17	0.57	flip: trend→periodicity
L10H7	0.27	0.47	neutral
L10H14	0.15	0.53	flip: trend→periodicity
L12H4	0.30	0.86	flip: trend→periodicity
L12H13	0.12	0.35	trend-dominant
L17H13	0.04	0.29	trend-dominant

Table 7: Per-head function profiles for Chronos. For each head we report the functional score r of Section 2 on ETTh1 and ETTm1 (higher r = more periodicity-dominant, lower = more trend-dominant) and the resulting role. Roles: "trend-dominant" ($r<0.4$ on both datasets), "periodicity-dominant" (periodicity-leaning on both), "flip" (a functional flip under the criterion of Table 9), and "neutral" otherwise.

Head	r (ETTh1)	r (ETTm1)	Role
L0H0	0.29	0.22	trend-dominant
L1H8	0.02	0.05	trend-dominant
L4H0	0.14	0.24	trend-dominant
L6H2	0.16	0.17	trend-dominant
L6H5	0.08	0.17	trend-dominant
L7H8	0.31	0.25	trend-dominant
L8H0	0.08	0.08	trend-dominant
L9H3	0.09	0.16	trend-dominant
L10H5	0.02	0.19	trend-dominant
L10H6	0.15	0.27	trend-dominant
L11H0	0.11	0.33	trend-dominant
L11H1	0.08	0.20	trend-dominant
L11H11	0.63	0.15	flip: periodicity→trend
L29H8	0.02	0.30	trend-dominant
L30H4	0.05	0.52	flip: trend→periodicity
L32H2	0.71	0.72	periodicity-dominant
L33H2	0.50	0.50	neutral
L33H5	0.68	0.63	periodicity-dominant
L33H11	0.68	0.52	periodicity-dominant
L34H5	0.25	0.12	trend-dominant
L34H6	0.28	0.24	trend-dominant
L34H10	0.09	0.36	trend-dominant
L35H2	0.69	0.04	flip: periodicity→trend
L35H3	0.07	0.04	trend-dominant
L35H4	0.72	0.63	periodicity-dominant
L35H5	0.40	0.23	neutral
L35H6	0.44	0.10	neutral
L35H9	0.02	0.03	trend-dominant
L35H11	0.02	0.24	trend-dominant

Table 8: Per-head function profiles for Time-MoE. For each head we report the functional score r of Section 2 on ETTh1 and ETTm1 (higher r = more periodicity-dominant, lower = more trend-dominant) and the resulting role. Roles: "trend-dominant" ($r<0.4$ on both datasets), "periodicity-dominant" (periodicity-leaning on both), "flip" (a functional flip under the criterion of Table 9), and "neutral" otherwise.

Head	r (ETTh1)	r (ETTm1)	Role
L0H1	0.41	0.25	neutral
L0H4	0.41	0.10	neutral
L0H5	0.30	0.19	trend-dominant
L0H6	0.28	0.23	trend-dominant
L3H7	0.22	0.17	trend-dominant
L3H11	0.33	0.21	trend-dominant
L4H8	0.13	0.12	trend-dominant
L4H10	0.13	0.07	trend-dominant
L4H11	0.19	0.18	trend-dominant
L5H4	0.09	0.12	trend-dominant
L5H5	0.07	0.32	trend-dominant
L5H6	0.09	0.30	trend-dominant
L5H9	0.07	0.26	trend-dominant
L6H2	0.17	0.13	trend-dominant
L6H7	0.19	0.35	trend-dominant
L8H1	0.12	0.42	neutral
L8H7	0.38	0.14	trend-dominant
L9H5	0.55	0.12	flip: periodicity→trend
L10H11	0.69	0.17	flip: periodicity→trend
L11H8	0.46	0.31	neutral

Table 9: Cross-dataset stability of selected attention-head functions. "Non-flipped" and "flipped" indicate whether a head preserves or changes its dominant function between ETTh1 and ETTm1 under the criterion in Section 2. "–" denotes models for which functional profiling is not reported.

Model	$\lvert\mathcal{H}\rvert$	non-flipped	flipped
Timer	19 / 64	17/19	2/19
TimesFM	20 / 320	13/20	7/20
Chronos	29 / 432	26/29	3/29
Moirai	25 / 384	15/25	10/25
Time-MoE	20 / 144	18/20	2/20
SEMPO	6 / 12	–	–

6.5 Few-Shot Results on Observed Source Datasets

Tables 10 and 11 report performance on the observed source datasets used to learn the scaling parameters. These results complement the zero-shot transfer experiments in the main paper: they show how the same parameterization adapts to observed sources, whereas the main zero-shot results directly evaluate whether the learned adjustments transfer to unseen datasets.

TSLib sources. Table 10 shows that calibration generally improves the observed ETT and ECL sources across backbones. The largest gains occur on ETTm1 for several models, while ETTh1 and ECL typically show smaller but consistent reductions. This variation is expected because each source-specific intervention is optimized only on its corresponding domain while sharing the common intervention component.

GIFT-Eval sources. Table 11 exhibits a wider range of calibration gains. BizITObs-L2C and BizITObs-Service benefit strongly for several backbones, whereas M_DENSE and Loop Seattle show more moderate changes and contain unavailable entries for contaminated model–dataset combinations. These results emphasize that the calibrated interventions are domain dependent rather than a uniform global rescaling.

Table 10: Few-shot forecasting results on the observed TSLib source datasets. We report MSE (↓) for horizons $H \in \{96, 192, 336, 720\}$. $\Delta(\%) = 100(\text{Ours} - \text{Vanilla})/\text{Vanilla}$. "–" denotes results excluded due to pre-training contamination.

		Timer			TimesFM			Chronos			Moirai			Time-MoE			SEMPO		
Dataset	H	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)
ETTm1	96	0.310	**0.305**	**-1.6**	0.352	**0.298**	**-15.3**	0.310	**0.291**	**-6.1**	0.383	**0.318**	**-17.0**	0.278	**0.253**	**-9.0**	0.520	**0.500**	**-3.8**
	192	0.354	**0.343**	**-3.1**	0.411	**0.344**	**-16.3**	0.364	**0.325**	**-10.7**	0.412	**0.351**	**-14.8**	0.355	**0.304**	**-14.4**	0.518	**0.498**	**-3.9**
	336	0.384	**0.372**	**-3.1**	0.448	**0.375**	**-16.3**	0.408	**0.354**	**-13.2**	0.427	**0.368**	**-13.8**	0.411	**0.342**	**-16.8**	0.555	**0.535**	**-3.6**
	720	0.430	**0.420**	**-2.3**	0.510	**0.434**	**-14.9**	0.500	**0.409**	**-18.2**	0.461	**0.399**	**-13.4**	0.483	**0.409**	**-15.3**	0.574	**0.558**	**-2.8**
	Avg	0.369	**0.360**	**-2.4**	0.430	**0.363**	**-15.6**	0.396	**0.345**	**-12.9**	0.421	**0.359**	**-14.7**	0.382	**0.327**	**-14.4**	0.542	**0.523**	**-3.5**
ETTh1	96	0.366	**0.364**	**-0.5**	0.385	**0.356**	**-7.5**	0.368	**0.350**	**-4.9**	0.390	**0.380**	**-2.6**	0.354	**0.338**	**-4.5**	0.425	**0.415**	**-2.4**
	192	0.391	0.391	+0.0	0.421	**0.393**	**-6.7**	0.424	**0.391**	**-7.8**	0.431	**0.419**	**-2.8**	0.416	**0.381**	**-8.4**	0.453	**0.443**	**-2.2**
	336	0.406	0.406	+0.0	0.431	**0.404**	**-6.3**	0.468	**0.415**	**-11.3**	0.461	**0.444**	**-3.7**	0.476	**0.414**	**-13.0**	0.466	**0.450**	**-3.4**
	720	0.417	**0.413**	**-1.0**	0.412	**0.390**	**-5.3**	0.461	**0.427**	**-7.4**	0.459	**0.433**	**-5.7**	0.523	**0.451**	**-13.8**	0.475	**0.461**	**-2.9**
	Avg	0.395	**0.394**	**-0.3**	0.412	**0.386**	**-6.3**	0.430	**0.396**	**-7.9**	0.436	**0.419**	**-3.9**	0.442	**0.396**	**-10.4**	0.455	**0.442**	**-2.9**
ECL	96	0.142	**0.133**	**-6.3**	0.133	**0.129**	**-3.0**	0.122	**0.121**	**-0.8**	0.152	**0.150**	**-1.3**	–	–	–	0.177	**0.172**	**-2.8**
	192	0.161	**0.151**	**-6.2**	0.150	**0.146**	**-2.7**	0.141	**0.139**	**-1.4**	0.168	**0.166**	**-1.2**	–	–	–	0.192	**0.188**	**-2.1**
	336	0.178	**0.169**	**-5.1**	0.168	**0.163**	**-3.0**	0.159	**0.156**	**-1.9**	0.185	**0.182**	**-1.6**	–	–	–	0.204	**0.201**	**-1.5**
	720	0.217	**0.211**	**-2.8**	0.207	**0.203**	**-1.9**	0.203	**0.198**	**-2.5**	0.227	**0.223**	**-1.8**	–	–	–	0.242	**0.240**	**-0.8**
	Avg	0.174	**0.166**	**-4.6**	0.165	**0.160**	**-3.0**	0.156	**0.153**	**-1.9**	0.183	**0.180**	**-1.6**	–	–	–	0.204	**0.200**	**-2.0**

Table 11: Few-shot forecasting results on the observed GIFT-Eval source datasets. We report MASE (↓) over short-, medium-, and long-horizon forecasts. $\Delta(\%) = 100(\text{Ours} - \text{Vanilla})/\text{Vanilla}$. "–" denotes results excluded due to pre-training contamination.

		Timer			TimesFM			Chronos			Moirai			Time-MoE			SEMPO		
Dataset	H	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)
BizITObs L2C	Short	0.336	**0.314**	**-6.5**	0.283	**0.278**	**-1.8**	0.278	**0.266**	**-4.3**	0.283	0.313	+10.6	0.334	**0.333**	**-0.3**	0.415	**0.404**	**-2.7**
	Medium	0.904	**0.826**	**-8.6**	0.974	**0.608**	**-37.6**	0.876	**0.598**	**-31.7**	0.877	**0.729**	**-16.9**	1.164	**1.065**	**-8.5**	1.049	**1.043**	**-0.6**
	Long	1.302	**1.143**	**-12.2**	1.393	**0.662**	**-52.4**	1.242	**0.682**	**-45.1**	1.274	**1.167**	**-8.4**	1.583	**1.396**	**-11.8**	1.325	**1.308**	**-1.3**
	Avg	0.847	**0.761**	**-10.2**	0.884	**0.516**	**-41.6**	0.799	**0.516**	**-35.4**	0.811	**0.736**	**-9.2**	1.027	**0.931**	**-9.3**	0.930	**0.918**	**-1.3**
BizITObs Service	Short	2.495	**2.097**	**-16.0**	0.808	**0.803**	**-0.6**	3.304	**0.953**	**-71.2**	2.196	**1.371**	**-37.6**	1.808	**1.344**	**-25.7**	4.490	**4.352**	**-3.1**
	Medium	6.447	**3.792**	**-41.2**	1.060	1.060	+0.0	4.941	**1.345**	**-72.8**	4.097	**2.341**	**-42.9**	4.640	**2.769**	**-40.3**	7.070	**6.875**	**-2.8**
	Long	6.484	**3.953**	**-39.0**	1.437	**1.415**	**-1.6**	5.209	**1.600**	**-69.3**	4.190	**2.380**	**-43.2**	4.937	**3.288**	**-33.4**	7.203	**6.955**	**-3.4**
	Avg	5.142	**3.281**	**-36.2**	1.102	**1.093**	**-0.8**	4.485	**1.300**	**-71.0**	3.494	**2.031**	**-41.9**	3.795	**2.467**	**-35.0**	6.254	**6.061**	**-3.1**
M_DENSE	Short	0.844	0.847	+0.4	0.803	0.803	+0.0	0.775	0.787	+1.5	–	–	–	0.804	**0.776**	**-3.5**	1.203	**1.148**	**-4.6**
	Medium	0.756	**0.748**	**-1.1**	0.701	**0.694**	**-1.0**	0.907	**0.788**	**-13.1**	–	–	–	0.864	**0.737**	**-14.7**	1.075	**1.037**	**-3.5**
	Long	0.782	**0.772**	**-1.3**	0.700	**0.695**	**-0.8**	0.995	**0.938**	**-5.7**	–	–	–	0.922	**0.750**	**-18.7**	1.061	**1.023**	**-3.6**
	Avg	0.794	**0.789**	**-0.6**	0.735	**0.731**	**-0.6**	0.892	**0.837**	**-6.2**	–	–	–	0.863	**0.754**	**-12.6**	1.113	**1.069**	**-4.0**
Loop Seattle	Short	0.582	**0.577**	**-0.9**	0.559	**0.550**	**-1.6**	0.628	**0.575**	**-8.4**	–	–	–	–	–	–	0.707	**0.697**	**-1.4**
	Medium	0.866	**0.850**	**-1.8**	0.860	**0.848**	**-1.4**	1.116	**0.993**	**-11.0**	–	–	–	–	–	–	1.159	**1.141**	**-1.6**
	Long	0.916	**0.894**	**-2.4**	0.931	**0.918**	**-1.4**	1.205	**1.052**	**-12.7**	–	–	–	–	–	–	1.231	**1.221**	**-0.8**
	Avg	0.788	**0.774**	**-1.8**	0.783	**0.772**	**-1.5**	0.983	**0.873**	**-11.2**	–	–	–	–	–	–	1.032	**1.020**	**-1.2**

6.6 TSLib Mean Absolute Error

Tables 12 and 13 provide MAE results corresponding to the zero-shot and source-domain MSE experiments, respectively. The MAE results broadly follow the same trend as MSE: 38 of the 40 model–dataset averages improve after head-wise scaling. The two exceptions are small increases for Chronos on ETTh1 and ECL. Overall, the improvements are therefore not specific to the squared-error metric used in the main evaluation.

Zero-shot targets. Table 12 shows that the MAE reductions persist across the held-out ETT, Weather, and Exchange datasets. The gains are generally strongest on ETTm2, where the retrieved intervention is consistently drawn from the closely related ETTm1 source, while the more distinct targets receive smaller corrections.

Observed sources. Table 13 reports the corresponding MAE on calibration domains. The pattern is consistent with the source-domain MSE results: improvements are larger on ETTm1 for several backbones and more modest on ECL, while the small Chronos increases on ETTh1 and ECL illustrate that optimizing squared error does not guarantee monotonic improvement under every auxiliary metric.

Table 12: TSLib zero-shot transfer results measured by MAE (↓), corresponding to the MSE results in Table 1.

		Timer			TimesFM			Chronos			Moirai			TimeMoE			SEMPO		
Dataset	*H*	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)
ETTm2	96	0.280	**0.264**	**-5.4**	0.244	**0.238**	**-2.4**	0.236	0.238	+0.9	0.261	**0.244**	**-6.7**	0.269	**0.257**	**-4.3**	0.312	**0.305**	**-2.3**
	192	0.319	**0.301**	**-5.6**	0.288	**0.281**	**-2.7**	0.284	**0.280**	**-1.2**	0.304	**0.286**	**-5.8**	0.323	**0.300**	**-7.0**	0.350	**0.339**	**-3.2**
	336	0.355	**0.335**	**-5.7**	0.328	**0.318**	**-3.1**	0.328	**0.316**	**-3.6**	0.345	**0.326**	**-5.6**	0.362	**0.335**	**-7.4**	0.377	**0.366**	**-3.1**
	720	0.413	**0.385**	**-6.7**	0.387	**0.372**	**-3.8**	0.393	**0.360**	**-8.2**	0.409	**0.383**	**-6.2**	0.400	**0.381**	**-4.6**	0.419	**0.410**	**-2.1**
	Avg	0.342	**0.322**	**-5.9**	0.312	**0.302**	**-3.1**	0.310	**0.299**	**-3.6**	0.330	**0.310**	**-6.0**	0.338	**0.318**	**-5.9**	0.364	**0.355**	**-2.7**
ETTh2	96	0.339	**0.336**	**-1.0**	0.321	**0.315**	**-2.0**	0.312	**0.312**	**-0.1**	0.324	**0.320**	**-1.4**	0.329	**0.324**	**-1.4**	0.386	**0.372**	**-3.5**
	192	0.375	**0.369**	**-1.5**	0.361	**0.352**	**-2.5**	0.352	**0.349**	**-0.9**	0.361	**0.354**	**-1.9**	0.369	**0.359**	**-2.5**	0.436	**0.418**	**-4.2**
	336	0.391	**0.383**	**-1.9**	0.376	**0.367**	**-2.2**	0.376	**0.367**	**-2.5**	0.378	**0.368**	**-2.5**	0.392	**0.376**	**-4.2**	0.455	**0.438**	**-3.8**
	720	0.433	**0.427**	**-1.4**	0.395	**0.392**	**-0.7**	0.408	**0.397**	**-2.8**	0.419	**0.398**	**-4.9**	0.436	**0.412**	**-5.5**	0.462	**0.453**	**-2.1**
	Avg	0.384	**0.379**	**-1.5**	0.363	**0.357**	**-1.8**	0.362	**0.356**	**-1.7**	0.371	**0.360**	**-2.8**	0.381	**0.368**	**-3.5**	0.435	**0.420**	**-3.4**
Weather	96	0.225	0.233	+3.4	0.185	0.185	+0.0	0.191	0.192	+0.5	0.209	**0.205**	**-1.8**	0.202	**0.202**	**-0.0**	0.255	**0.251**	**-1.7**
	192	0.271	0.272	+0.2	0.234	**0.231**	**-1.2**	0.235	**0.233**	**-0.7**	0.251	**0.247**	**-1.4**	0.254	**0.254**	**-0.1**	0.301	**0.294**	**-2.3**
	336	0.313	**0.306**	**-2.2**	0.282	**0.275**	**-2.3**	0.276	**0.270**	**-2.1**	0.293	**0.291**	**-1.0**	0.300	**0.299**	**-0.3**	0.334	**0.327**	**-2.0**
	720	0.373	**0.352**	**-5.4**	0.344	**0.336**	**-2.6**	0.332	**0.318**	**-4.2**	0.349	**0.348**	**-0.5**	0.355	**0.354**	**-0.3**	0.375	**0.370**	**-1.5**
	Avg	0.295	**0.291**	**-1.6**	0.261	**0.257**	**-1.7**	0.258	**0.253**	**-2.0**	0.276	**0.273**	**-1.1**	0.278	**0.277**	**-0.2**	0.316	**0.311**	**-1.9**
Exchange	96	0.221	**0.218**	**-1.2**	0.202	**0.200**	**-0.9**	0.214	**0.211**	**-1.4**	0.202	**0.202**	**-0.0**	–	–	–	0.246	0.247	+0.3
	192	0.310	**0.304**	**-1.9**	0.295	**0.291**	**-1.4**	0.310	**0.305**	**-1.8**	0.292	**0.291**	**-0.3**	–	–	–	0.361	**0.358**	**-0.8**
	336	0.409	**0.398**	**-2.8**	0.408	**0.402**	**-1.5**	0.434	**0.424**	**-2.2**	0.395	**0.390**	**-1.3**	–	–	–	0.482	**0.476**	**-1.3**
	720	0.614	**0.587**	**-4.4**	0.703	0.704	+0.2	0.739	**0.724**	**-2.0**	0.645	**0.624**	**-3.3**	–	–	–	0.745	**0.745**	**-0.1**
	Avg	0.388	**0.377**	**-3.0**	0.402	**0.399**	**-0.7**	0.424	**0.416**	**-1.9**	0.384	**0.377**	**-1.8**	–	–	–	0.459	**0.456**	**-0.5**

Table 13: TSLib MAE (↓) on observed source datasets, corresponding to the MSE results in Table 10.

		Timer			TimesFM			Chronos			Moirai			TimeMoE			SEMPO		
Dataset	*H*	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)	Vanilla	Ours	Δ(%)
ETTm1	96	0.357	**0.356**	**-0.1**	0.342	**0.327**	**-4.4**	0.325	0.332	+2.1	0.367	**0.341**	**-7.0**	0.337	**0.317**	**-5.9**	0.479	**0.468**	**-2.3**
	192	0.384	**0.380**	**-1.1**	0.377	**0.357**	**-5.3**	0.357	**0.357**	**-0.2**	0.390	**0.364**	**-6.5**	0.393	**0.353**	**-10.4**	0.484	**0.471**	**-2.6**
	336	0.406	**0.401**	**-1.2**	0.402	**0.378**	**-5.9**	0.384	**0.378**	**-1.3**	0.404	**0.381**	**-5.6**	0.433	**0.380**	**-12.4**	0.501	**0.489**	**-2.4**
	720	0.438	**0.434**	**-0.9**	0.436	**0.411**	**-5.7**	0.430	**0.412**	**-4.2**	0.427	**0.405**	**-5.1**	0.477	**0.420**	**-11.8**	0.513	**0.504**	**-1.8**
	Avg	0.396	**0.393**	**-0.8**	0.389	**0.368**	**-5.4**	0.374	**0.370**	**-1.1**	0.397	**0.373**	**-6.0**	0.410	**0.367**	**-10.4**	0.494	**0.483**	**-2.3**
ETTh1	96	0.391	**0.388**	**-0.6**	0.388	**0.379**	**-2.2**	0.375	0.381	+1.6	0.393	**0.391**	**-0.6**	0.390	**0.375**	**-3.8**	0.435	**0.429**	**-1.4**
	192	0.407	**0.405**	**-0.7**	0.407	**0.399**	**-1.9**	0.404	0.406	+0.5	0.413	**0.411**	**-0.6**	0.433	**0.401**	**-7.4**	0.452	**0.446**	**-1.2**
	336	0.418	**0.415**	**-0.7**	0.415	**0.408**	**-1.9**	0.425	**0.423**	**-0.6**	0.426	**0.423**	**-0.7**	0.477	**0.422**	**-11.5**	0.460	**0.452**	**-1.7**
	720	0.438	**0.433**	**-1.2**	0.421	**0.415**	**-1.4**	0.441	0.449	+1.8	0.439	**0.434**	**-1.0**	0.525	**0.456**	**-13.1**	0.478	**0.472**	**-1.2**
	Avg	0.414	**0.410**	**-0.8**	0.408	**0.400**	**-1.8**	0.411	0.415	+0.8	0.418	**0.415**	**-0.7**	0.456	**0.413**	**-9.3**	0.456	**0.450**	**-1.4**
ECL	96	0.238	**0.229**	**-4.0**	0.220	**0.218**	**-1.0**	0.208	0.210	+0.5	0.241	**0.240**	**-0.5**	–	–	–	0.284	**0.278**	**-2.3**
	192	0.256	**0.245**	**-4.0**	0.237	**0.234**	**-1.0**	0.226	0.228	+0.6	0.256	**0.254**	**-0.6**	–	–	–	0.295	**0.292**	**-1.2**
	336	0.273	**0.264**	**-3.3**	0.254	**0.252**	**-0.7**	0.246	0.247	+0.7	0.272	**0.270**	**-0.6**	–	–	–	0.306	**0.302**	**-1.1**
	720	0.308	**0.302**	**-1.8**	0.287	**0.287**	**-0.2**	0.285	0.287	+0.9	0.304	**0.302**	**-0.7**	–	–	–	0.335	**0.334**	**-0.3**
	Avg	0.269	**0.260**	**-3.2**	0.250	**0.248**	**-0.7**	0.241	0.243	+0.7	0.268	**0.267**	**-0.6**	–	–	–	0.305	**0.301**	**-1.2**

6.7 GIFT-Eval Probabilistic Accuracy

Tables 14 and 15 report CRPS for TimesFM, Moirai, and Chronos, the three evaluated backbones that produce probabilistic forecasts. Timer, Time-MoE, and SEMPO output point forecasts and are therefore excluded. CRPS generally follows the corresponding MASE trends, although the magnitude of improvement varies across datasets. Since our training objective optimizes point-forecast accuracy rather than distributional quality, we use CRPS as a complementary evaluation rather than a directly optimized metric.

For configurations in which retrieval selects the identity/shared fallback, the corresponding predictive distribution is unchanged, explaining the unchanged CRPS entries for several TimesFM targets.

Zero-shot targets. Table 14 shows that the strongest distributional improvements occur on BizITObs-Application, where retrieval selects the related BizITObs-Service intervention for several probabilistic backbones. In contrast, SZ-Taxi and Bitbrains-Fast-Storage receive smaller changes, consistent with the more conservative retrieval decisions reported in Section 6.9.

Observed sources. Table 15 provides the source-domain counterpart. The largest CRPS reductions again occur on the BizITObs streams for several models, whereas the transportation streams show more moderate changes. This confirms that improvements are visible not only in point metrics but also, for probabilistic backbones, in the predicted distribution.

6.8 Ablation: Hard Head Selection versus Learned Gating

Our method uses a learnable gate $g_{l,h} = \sigma(\gamma_{l,h})$ for every attention head. We compare this design with a hard-selection variant that fixes $g_{l,h} = 1$ for heads in the selected set $\mathcal{H}$ and $g_{l,h} = 0$ elsewhere, while keeping all other training and inference settings unchanged.

Table 14: GIFT-Eval zero-shot transfer CRPS (↓) for probabilistic TSFMs. "–" denotes excluded pre-training-contaminated streams. Timer, Time-MoE, and SEMPO are omitted because they only generate point forecasts.

		TimesFM			Moirai			Chronos		
Dataset	len	zs	ours	Δ(%)	zs	ours	Δ(%)	zs	ours	Δ(%)
bizitobs_application/10S	short	0.011	0.011	+0.3	0.039	**0.035**	**-10.6**	0.054	**0.022**	**-59.7**
	medium	0.037	**0.037**	**-0.3**	0.093	**0.051**	**-45.8**	0.100	**0.041**	**-58.7**
	long	0.059	**0.057**	**-2.3**	0.098	**0.061**	**-37.7**	0.105	**0.052**	**-50.5**
	Avg	0.035	**0.035**	**-1.4**	0.077	**0.049**	**-36.4**	0.086	**0.038**	**-55.6**
sz_taxi/15T	short	0.201	0.201	+0.0	–	–	–	0.202	0.202	+0.0
	medium	0.234	**0.232**	**-0.7**	–	–	–	0.222	0.222	+0.0
	long	0.235	**0.233**	**-0.8**	–	–	–	0.229	**0.228**	**-0.1**
	Avg	0.223	**0.222**	**-0.5**	–	–	–	0.218	0.218	+0.0
bitbrains_fast_storage/5T	short	0.414	**0.403**	**-2.9**	0.424	0.424	+0.0	0.451	**0.436**	**-3.3**
	medium	0.967	**0.901**	**-7.3**	0.668	**0.665**	**-0.5**	0.672	0.679	+0.9
	long	1.044	**0.879**	**-18.8**	0.708	**0.706**	**-0.3**	0.755	**0.745**	**-1.3**
	Avg	0.808	**0.728**	**-11.1**	0.600	**0.598**	**-0.3**	0.626	**0.620**	**-1.0**

Table 15: GIFT-Eval CRPS (↓) on observed source streams.

		TimesFM			Moirai			Chronos		
Dataset	len	zs	ours	Δ(%)	zs	ours	Δ(%)	zs	ours	Δ(%)
bizitobs_l2c/5T	short	0.077	**0.076**	**-1.5**	0.077	0.085	+10.5	0.074	0.081	+9.6
	medium	0.503	**0.273**	**-45.7**	0.377	**0.338**	**-10.3**	0.415	**0.286**	**-31.2**
	long	0.834	**0.327**	**-60.8**	0.632	**0.598**	**-5.4**	0.713	**0.348**	**-51.1**
	Avg	0.472	**0.226**	**-52.2**	0.362	**0.340**	**-6.0**	0.401	**0.239**	**-40.5**
bizitobs_service/10S	short	0.011	0.011	+0.9	0.038	**0.024**	**-35.4**	0.050	**0.016**	**-68.7**
	medium	0.028	**0.027**	**-2.9**	0.081	**0.049**	**-40.0**	0.094	**0.037**	**-61.0**
	long	0.056	**0.055**	**-1.1**	0.100	**0.061**	**-38.6**	0.110	**0.055**	**-50.1**
	Avg	0.032	0.031	**-0.3**	0.073	**0.045**	**-38.6**	0.085	**0.036**	**-57.8**
m_dense/H	short	0.133	0.133	+0.5	–	–	–	0.125	0.130	+4.2
	medium	0.123	**0.121**	**-1.5**	–	–	–	0.156	**0.149**	**-4.2**
	long	0.120	**0.119**	**-0.7**	–	–	–	0.171	**0.168**	**-1.7**
	Avg	0.125	**0.125**	**-0.6**	–	–	–	0.151	**0.149**	**-0.9**
loop_seattle/5T	short	0.049	**0.048**	**-1.9**	–	–	–	0.055	**0.052**	**-4.4**
	medium	0.083	**0.081**	**-1.5**	–	–	–	0.107	**0.092**	**-14.0**
	long	0.092	**0.090**	**-1.3**	–	–	–	0.117	**0.097**	**-17.4**
	Avg	0.074	**0.073**	**-1.5**	–	–	–	0.093	**0.080**	**-13.5**

Tables 16 and 17 show that learned gating typically matches or improves upon hard selection, with larger gains for several Chronos and Time-MoE settings. The result supports using head-function analysis to initialize the scaling module without permanently restricting adaptation to a fixed subset of heads. Allowing all gates to remain trainable provides additional flexibility when the most useful head configuration varies across datasets.

The zero-shot comparison also shows that learned gating is not uniformly better for every model–target pair. For example, hard selection is stronger for Timer on ETTh2, whereas learned gating gives clearer gains for TimesFM, Moirai, and Time-MoE on several targets. This behavior supports treating the mean-ablation set as an informative initialization rather than a fixed structural constraint.

Across observed sources, the same trend is visible: learned gates can move beyond the initial hard subset when source-domain calibration favors a different head configuration. Together with the zero-shot results, this ablation supports the soft-gating design used in SHIFT.

Table 16: Hard head selection versus learned gating under zero-shot transfer. Both variants use the same initialization, sources, and inference settings. $\Delta(\%) = 100(\text{Gated} - \text{Hard})/\text{Hard}$

		Timer			TimesFM			Chronos			Moirai			Time-MoE		
Dataset	H	Hard	Gated	Δ(%)	Hard	Gated	Δ(%)	Hard	Gated	Δ(%)	Hard	Gated	Δ(%)	Hard	Gated	Δ(%)
ETTm2	96	0.177	**0.176**	**-0.2**	0.167	0.168	+0.6	0.157	0.158	+0.8	0.177	**0.175**	**-0.7**	0.174	**0.171**	**-1.6**
	192	0.229	**0.229**	**-0.1**	0.231	0.231	+0.1	0.214	0.215	+0.5	0.241	**0.239**	**-0.9**	0.232	**0.228**	**-1.6**
	336	0.278	**0.278**	**-0.3**	0.287	**0.286**	**-0.4**	0.263	**0.263**	**-0.0**	0.303	**0.299**	**-1.1**	0.283	**0.279**	**-1.6**
	720	0.358	**0.356**	**-0.6**	0.371	**0.371**	**-0.0**	0.329	**0.327**	**-0.7**	0.393	**0.388**	**-1.3**	0.359	**0.350**	**-2.3**
	Avg	0.261	**0.260**	**-0.3**	0.264	0.264	+0.0	0.241	**0.241**	**-0.0**	0.278	**0.275**	**-1.1**	0.262	**0.257**	**-1.8**
ETTh2	96	0.274	0.277	+1.0	0.270	**0.266**	**-1.5**	0.257	**0.257**	**-0.1**	0.281	**0.278**	**-1.0**	0.262	**0.261**	**-0.2**
	192	0.318	0.323	+1.6	0.321	**0.317**	**-1.3**	0.310	0.310	+0.1	0.326	**0.323**	**-1.0**	0.306	**0.305**	**-0.4**
	336	0.321	0.329	+2.2	0.332	**0.326**	**-1.7**	0.326	**0.325**	**-0.3**	0.335	**0.331**	**-1.0**	0.321	**0.316**	**-1.5**
	720	0.360	0.381	+5.9	0.355	**0.351**	**-1.0**	0.366	**0.361**	**-1.5**	0.364	**0.362**	**-0.4**	0.374	**0.360**	**-4.0**
	Avg	0.318	0.327	+2.8	0.320	**0.315**	**-1.4**	0.315	**0.313**	**-0.5**	0.326	**0.324**	**-0.9**	0.316	**0.310**	**-1.7**
Weather	96	0.171	0.171	+0.0	0.155	**0.149**	**-3.7**	0.152	0.153	+0.4	0.177	**0.176**	**-0.7**	0.155	0.155	+0.1
	192	0.214	**0.213**	**-0.3**	0.201	**0.196**	**-2.5**	0.197	**0.196**	**-0.5**	0.220	**0.218**	**-0.5**	0.209	0.210	+0.3
	336	0.261	**0.260**	**-0.5**	0.261	**0.256**	**-1.9**	0.244	**0.243**	**-0.5**	0.277	**0.276**	**-0.5**	0.269	0.270	+0.3
	720	0.331	**0.327**	**-1.2**	0.351	**0.347**	**-1.2**	0.311	**0.310**	**-0.5**	0.366	**0.364**	**-0.4**	0.346	0.347	+0.3
	Avg	0.244	**0.243**	**-0.6**	0.242	**0.237**	**-2.1**	0.226	**0.225**	**-0.4**	0.260	**0.259**	**-0.5**	0.245	0.246	+0.3
Exchange	96	0.091	0.091	+0.2	0.082	**0.082**	**-0.1**	0.093	**0.092**	**-0.6**	0.084	0.084	+0.6	–	–	–
	192	0.172	**0.172**	**-0.1**	0.168	**0.168**	**-0.3**	0.190	**0.189**	**-0.4**	0.167	0.168	+0.5	–	–	–
	336	0.284	**0.283**	**-0.3**	0.313	**0.311**	**-0.5**	0.350	**0.349**	**-0.3**	0.294	0.294	+0.0	–	–	–
	720	0.606	**0.600**	**-1.0**	0.872	0.879	+0.8	0.938	**0.934**	**-0.4**	0.710	**0.697**	**-1.8**	–	–	–
	Avg	0.288	**0.286**	**-0.6**	0.359	0.360	+0.3	0.393	**0.391**	**-0.4**	0.314	**0.311**	**-0.9**	–	–	–

Table 17: Hard head selection versus learned gating on observed source datasets. Conventions follow Table 16.

		Timer			TimesFM			Chronos			Moirai			Time-MoE		
Dataset	H	Hard	Gated	Δ(%)	Hard	Gated	Δ(%)	Hard	Gated	Δ(%)	Hard	Gated	Δ(%)	Hard	Gated	Δ(%)
ETTm1	96	0.305	**0.305**	**-0.2**	0.304	**0.298**	**-2.0**	0.290	0.291	+0.3	0.320	**0.318**	**-0.7**	0.259	**0.253**	**-2.5**
	192	0.344	**0.343**	**-0.5**	0.350	**0.344**	**-1.7**	0.334	**0.325**	**-2.6**	0.353	**0.351**	**-0.7**	0.317	**0.304**	**-4.2**
	336	0.374	**0.372**	**-0.5**	0.381	**0.375**	**-1.6**	0.372	**0.354**	**-4.8**	0.371	**0.368**	**-0.9**	0.364	**0.342**	**-5.9**
	720	0.422	**0.420**	**-0.4**	0.446	**0.434**	**-2.8**	0.454	**0.409**	**-10.0**	0.403	**0.399**	**-0.9**	0.451	**0.409**	**-9.3**
	Avg	0.362	**0.360**	**-0.4**	0.370	**0.363**	**-2.1**	0.363	**0.345**	**-4.9**	0.362	**0.359**	**-0.8**	0.348	**0.327**	**-6.0**
ETTh1	96	0.365	0.367	+0.4	0.357	**0.356**	**-0.3**	0.358	**0.350**	**-2.3**	0.384	**0.380**	**-0.8**	0.337	0.338	+0.4
	192	0.406	0.407	+0.2	0.394	**0.393**	**-0.2**	0.410	**0.391**	**-4.5**	0.422	**0.419**	**-0.8**	0.378	0.381	+0.9
	336	0.427	0.429	+0.3	0.405	**0.404**	**-0.1**	0.446	**0.415**	**-7.0**	0.449	**0.444**	**-1.0**	0.410	0.414	+0.9
	720	0.421	0.429	+1.9	0.393	**0.390**	**-0.6**	0.475	**0.427**	**-10.1**	0.435	**0.433**	**-0.3**	0.453	**0.451**	**-0.5**
	Avg	0.405	0.408	+0.7	0.387	**0.386**	**-0.3**	0.422	**0.396**	**-6.3**	0.422	**0.419**	**-0.7**	0.394	0.396	+0.4

6.9 Representation-Based Retrieval Outcomes

For an unseen target dataset, we first compute its mean head-activation fingerprint and compare it with the centroid of each observed source. Let $s^\star$ denote the most similar source. Its source-specific residual $\boldsymbol{\eta}_{s^\star}$ is used only when the target–source similarity exceeds a threshold estimated from that source's validation windows; otherwise, prediction uses the shared scaling alone. The threshold percentile q is selected using source data only. The tables below expose the raw target–source similarities, the source-specific thresholds, and the resulting deployment decisions. These decisions depend only on target input-context activations; no target labels or forecast-horizon observations are used.

Table 18 reports the TSLib decisions. The two same-family ETT targets are consistently matched to their corresponding observed sources across all backbones: ETTm2 retrieves ETTm1 and ETTh2 retrieves ETTh1. In contrast, Weather falls below the retrieval threshold for all evaluated models and therefore uses shared-only scaling. Exchange also uses shared-only scaling for most backbones, with Moirai being the only model that retrieves the ETTm1-specific residual.

Table 19 shows a similar pattern on GIFT-Eval. BizITObs-Application is sufficiently close to BizITObs-Service to retrieve its source-specific residual for Timer, Chronos, Moirai, and Time-MoE, whereas TimesFM and SEMPO fall below their respective thresholds. The remaining held-out streams use shared-only scaling across the evaluated backbones.

Overall, retrieval behaves conservatively: source-specific adjustments are transferred when a target closely matches an observed source, while more distinct targets rely on the shared component. This prevents every unseen dataset from being forced to inherit a source-specific configuration.

Table 18: TSLib retrieval outcomes. We report cosine similarity between each target fingerprint and source centroids, the best-matching source, its source-specific threshold, and the resulting deployment decision.

Model	Target	Cosine to source centroid			Best	Perc_q	Decision
		ETTm1	ETTh1	ECL			
Timer (q=10)	ETTm2	+0.63	−0.09	−0.64	ETTm1	0.478	borrow ETTm1
	ETTh2	−0.07	+0.65	−0.23	ETTh1	0.023	borrow ETTh1
	Weather	+0.25	−0.19	−0.18	ETTm1	0.478	shared-only
	Exchange	+0.31	−0.14	−0.27	ETTm1	0.478	shared-only
TimesFM (q=50)	ETTm2	+0.66	+0.06	−0.62	ETTm1	0.648	borrow ETTm1
	ETTh2	+0.11	+0.70	−0.53	ETTh1	0.462	borrow ETTh1
	Weather	+0.27	+0.08	−0.29	ETTm1	0.648	shared-only
	Exchange	+0.15	+0.15	−0.24	ETTh1	0.462	shared-only
Chronos (q=10)	ETTm2	+0.59	+0.26	−0.62	ETTm1	0.359	borrow ETTm1
	ETTh2	+0.20	+0.63	−0.42	ETTh1	−0.010	borrow ETTh1
	Weather	+0.32	+0.27	−0.38	ETTm1	0.359	shared-only
	Exchange	+0.45	+0.34	−0.53	ETTm1	0.359	shared-only
Moirai (q=10)	ETTm2	+0.55	+0.06	−0.57	ETTm1	0.345	borrow ETTm1
	ETTh2	+0.01	+0.49	−0.29	ETTh1	0.156	borrow ETTh1
	Weather	+0.32	+0.10	−0.37	ETTm1	0.345	shared-only
	Exchange	+0.35	+0.17	−0.44	ETTm1	0.345	borrow ETTm1
Time-MoE (q=10)	ETTm2	+0.21	−0.21	–	ETTm1	0.212	borrow ETTm1
	ETTh2	−0.47	+0.47	–	ETTh1	−0.006	borrow ETTh1
	Weather	+0.21	−0.21	–	ETTm1	0.212	shared-only
SEMPO (q=25)	ETTm2	+0.76	+0.26	−0.90	ETTm1	0.757	borrow ETTm1
	ETTh2	+0.10	+0.78	−0.44	ETTh1	0.389	borrow ETTh1
	Weather	+0.66	+0.34	−0.83	ETTm1	0.757	shared-only
	Exchange	+0.08	+0.38	−0.25	ETTh1	0.389	shared-only

The TSLib outcomes illustrate the intended behavior of the retrieval rule: targets that are close to a source in the head-activation space borrow its domain-specific intervention, whereas targets that do not clear the calibrated threshold remain on the shared component. Importantly, the decision is made once at the domain level and is then fixed for all target forecasts.

Table 19: GIFT-Eval retrieval outcomes. Source-specific scaling is used only when the best target–source cosine similarity exceeds the corresponding source threshold; otherwise shared-only scaling is deployed.

Model	Target	Cosine to source centroid				Best	Perc_q	Decision
		service	l2c	loop	m_dense			
Timer (q=10)	application/10S	+0.68	+0.29	−0.03	−0.50	service	0.379	borrow service
	sz_taxi/15T	−0.03	+0.20	+0.10	−0.19	l2c	0.579	shared-only
	fast_storage/5T	−0.25	−0.27	−0.06	+0.34	m_dense	0.803	shared-only
TimesFM (q=50)	application/10S	+0.57	+0.08	−0.16	−0.31	service	0.591	shared-only
	sz_taxi/15T	−0.30	+0.14	−0.08	+0.08	l2c	0.860	shared-only
	fast_storage/5T	−0.44	+0.53	−0.49	+0.04	l2c	0.860	shared-only
Chronos (q=10)	application/10S	+0.68	+0.24	+0.02	−0.25	service	0.359	borrow service
	sz_taxi/15T	+0.09	−0.10	+0.26	+0.10	loop	0.414	shared-only
	fast_storage/5T	+0.12	−0.09	+0.21	+0.10	loop	0.414	shared-only
Moirai (q=10)	application/10S	+0.71	+0.20	−0.24	−0.39	service	0.242	borrow service
	sz_taxi/15T	−0.02	−0.16	+0.00	+0.16	m_dense	0.639	shared-only
	fast_storage/5T	+0.14	+0.27	−0.35	−0.05	l2c	0.665	shared-only
Time-MoE (q=10)	application/10S	+0.72	+0.22	+0.04	−0.61	service	0.475	borrow service
	sz_taxi/15T	+0.25	+0.05	−0.00	−0.17	service	0.475	shared-only
	fast_storage/5T	+0.02	+0.07	+0.18	−0.19	loop	0.477	shared-only
SEMPO (q=25)	application/10S	+0.82	+0.32	−0.44	−0.48	service	0.865	shared-only
	sz_taxi/15T	+0.15	−0.64	+0.36	+0.11	loop	0.621	shared-only
	fast_storage/5T	+0.15	−0.60	+0.38	+0.07	loop	0.621	shared-only