ANCHORMIX: DISTILLING HETEROGENEOUS FORECASTERS INTO A SINGLE ROUTED TIME-SERIES MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Heterogeneous time-series forecasters capture complementary inductive biases, but exploiting them through runtime ensembles multiplies latency, memory, and deployment dependencies. We introduce AnchorMix, a two-stage framework that converts this diversity into one CycleRoutedTQNet checkpoint. Stage 1 learns a routed temporal-query anchor from labels and cached Chronos-2 forecasts. Stage 2 warm-starts the same architecture and trains it against ground truth plus a convex target mixing the frozen anchor with an architecture-heterogeneous teacher. The anchor preserves established student behavior while the external teacher supplies complementary corrections; every teacher is removed after training. Across five benchmarks, four horizons, and three seeds, the single student improves 29 of 40 MSE/MAE cells relative to the selected external teachers and records lower values in 39 of 40 cells against four recent published comparison tables. A full-array diagnostic finds a positive convex-combination advantage on all 20 dataset–horizon tasks, and target-composition ablations on ETTh2 and ETTm2 identify the roles of anchor preservation and heterogeneous correction. The deployed model contains 4.06–5.36M parameters and averages 1.19 ms per batch of 32. On matched RTX 5090 profiles, replacing the corresponding runtime ensemble yields up to a $181.7\times$ speedup and 80.7% lower peak allocated memory. AnchorMix therefore turns heterogeneous forecast diversity from a deployment burden into a training-time resource for one compact forecaster.

INTRODUCTION

Long-term multivariate time series forecasting (MTSF) predicts a future trajectory for many correlated variables from a fixed history. Recent progress has not converged to one architecture. Temporal-query models efficiently represent periodic cross-channel structure Lin et al. (2025); patch models capture local shape and trend Nie et al. (2023); Hu et al. (2026); and heterogeneous mixture-of-experts models route inputs through temporal and spectral specialists Ortigossa et al. (2026). Recent phase and graph-spectral designs further improve efficiency or channel-patch adaptation Niu et al. (2026); Li et al. (2026a). The resulting diversity is useful: models fail on different samples, channels, and horizons.

A direct ensemble can exploit this complementarity, but it is a poor deployment endpoint. It requires every expert at test time, couples the system to multiple codebases, and makes latency scale with the teacher bank. TimeFuse learns sample-wise weights over a model zoo Liu et al. (2025b), while TSCOMP searches component configurations and deploys a top-5 ensemble Liang et al. (2026); both demonstrate the value of diversity but retain multiple predictors. Ordinary one-teacher distillation Hinton et al. (2015) removes this dependency, yet a student can inherit the teacher's weaknesses and forget a strong initialization. This is especially visible when a teacher has lower MSE but worse MAE, or when its advantage changes across horizons.

We introduce AnchorMix, an anchor-preserving form of heterogeneous forecast distillation. The deployed model is always CycleRoutedTQNet, a routed bank of five internal forecasting branches. We first obtain a strong anchor by training this model with cached Chronos-2 forecasts. For the second stage, we pair the frozen anchor with an external teacher whose architecture is deliberately

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.

Method	Source	Anchor	#Deploy
TimeDistill	teacher	no	1
DistilTS	TSFM	no	1
TimeFuse	13-model zoo	no	13
TSCOMP	top-5 configs	no	5
AnchorMix	anchor + hetero.	yes	1

Table 1: Method positioning Ni et al. (2026); Li et al. (2026b); Liu et al. (2025b); Liang et al. (2026). "Anchor" means an explicit frozen student-family target; #Deploy counts retained forecasters, not internal branches.

different: MoHETS for datasets where heterogeneous temporal/spectral experts are strongest, and PMDformer where patch-mean decoupling is strongest. Their predictions form a convex training target. The student is initialized from the anchor and optimized against both ground truth and the mixed target. This turns an expensive training-time combination into one checkpoint rather than an inference ensemble.

Our contributions are: (1) an anchor-mixed target that preserves a strong local forecaster while introducing corrections from an architecture-heterogeneous teacher; (2) a TQNet/CycleNet-derived routed student with scale, multi-cycle, refinement, and cycle-residual branches plus a sample-adaptive router that absorbs this target without any teacher dependency at inference; and (3) an error-risk analysis and deployment-centered evaluation showing 29/40 controlled teacher-cell improvements, 39/40 improvements against recent published tables, and up to a $181.7\times$ speedup over the corresponding runtime ensemble with one checkpoint.

Related Work

Efficient multivariate forecasting. Transformer forecasting moved from sparse temporal attention Zhou et al. (2021) toward patch, inverted, and query-based representations. PatchTST Nie et al. (2023) tokenizes local temporal patches, iTransformer Liu et al. (2024a) treats variables as tokens, and TQNet Lin et al. (2025) uses cycle-indexed temporal queries. PhaseFormer replaces patches with compact phase embeddings Niu et al. (2026), whereas xCPD routes graph-spectral channel-patch interactions Li et al. (2026a). Our student builds directly on TQNet and CycleNet Lin et al. (2025; 2024), then adds multiple internal branches and a sample-adaptive router.

Heterogeneous and patch-decoupled teachers. PMDformer separates patch means from local shapes and restores trend information through attention Hu et al. (2026). MoHETS combines shared convolutional and routed Fourier experts and can incorporate exogenous covariates Ortigossa et al. (2026). These models encode different priors from a temporal-query student, making them useful sources of complementary errors rather than interchangeable copies of the same teacher. TimeFuse Liu et al. (2025b) performs adaptive sample-level fusion over heterogeneous forecasters, and TSCOMP Liang et al. (2026) recommends forecasting components from a large performance corpus. Unlike these inference-time ensembles, our teachers are removed after distillation.

Foundation models and distillation. Pretrained forecasters such as TimesFM Das et al. (2024), Timer Liu et al. (2024b), Chronos-2 Ansari et al. (2025), and Sundial Liu et al. (2025a) learn broad temporal priors, while Time-LLM Jin et al. (2024) and Time-VLM Zhong et al. (2025) adapt language or vision-language models to forecasting. TimeDistill transfers multi-scale and frequency information across forecasting architectures Ni et al. (2026). DistilTS adds horizon weighting and temporal representation alignment for distilling time-series foundation models Li et al. (2026b). AnchorMix differs in its target construction: it retains a frozen student-family anchor inside the teacher target and introduces a second, architecturally distinct expert only as a correction.

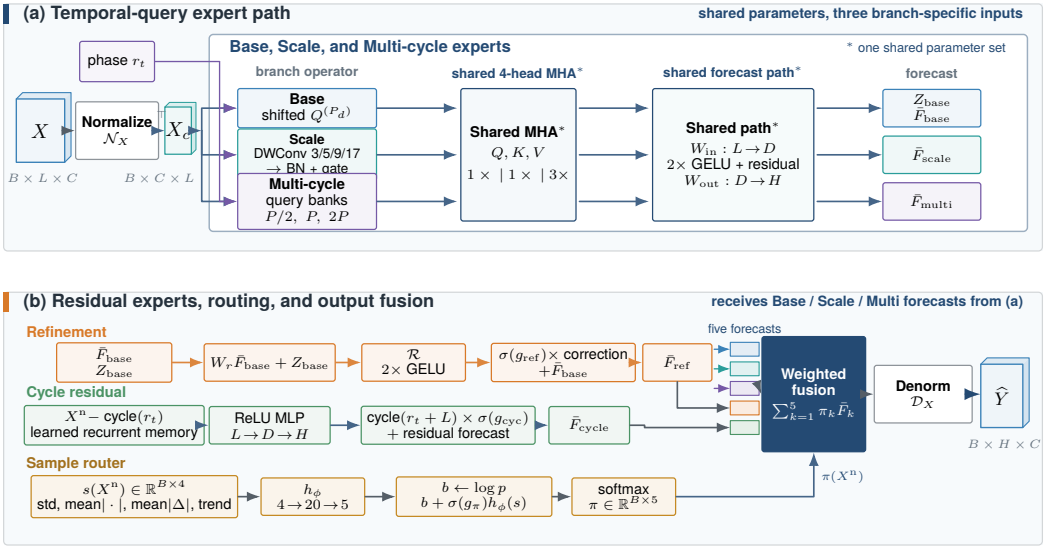

Figure 1: CycleRoutedTQNet inference architecture, split into two readable paths. (a) Base, Scale, and Multi-cycle apply branch-specific operators while sharing one attention module and forecast path. (b) Base-conditioned Refinement and Cycle residual join the three forecasts; the sample router supplies five fusion weights before denormalization. The deployed graph is one checkpoint and contains no external teacher.

METHOD

PROBLEM SETUP

Let $X \in \mathbb{R}^{B\times L\times C}$ be a batch of histories, where B is batch size, L is lookback length, and C is the number of variables. The target is $Y \in \mathbb{R}^{B\times H\times C}$ for horizon H. For a forecast origin t, one history is $(x_{t-L}, \ldots, x_{t-1})$ and its first target is x_t. For dataset d, let P_d denote its cycle length and $r_t = t \bmod P_d$ the phase of that first forecast point. The goal is a single predictor $S_\theta(X, r_t)$ that minimizes MSE and MAE without evaluating a teacher at test time. We suppress r_t in subsequent equations when it is not the focus.

ROUTED TEMPORAL-QUERY STUDENT

Architecturally and in code, CycleRoutedTQNet is derived from the official TQNet implementation Lin et al. (2025). Figure 1 summarizes its single-checkpoint inference graph. Its base branch retains TQNet's cycle-indexed temporal query, multi-head attention, and MLP forecast path, while its cycle branch uses CycleNet's residual-cycle forecasting operator Lin et al. (2024). Our architectural additions are the scale-enhanced, multi-cycle, and refinement branches and the five-way sample-adaptive router described below. Let $\mathcal{N}_X$ denote the reversible per-instance normalization determined by X, let $\mathcal{D}_X$ be its matching denormalization, and set $X^{\text{n}} = \mathcal{N}_X(X)$ and $X_c = (X^{\text{n}})^\top \in \mathbb{R}^{B\times C\times L}$. Internally, CycleRoutedTQNet produces five normalized channel-first forecasts $\bar{F}_k(X^{\text{n}}, r_t) \in \mathbb{R}^{B\times C\times H}$: base temporal query, scale-enhanced temporal query, multi-cycle query, forecast refinement, and recurrent cycle residual. Let $Q^{(P_d)} \in \mathbb{R}^{P_d\times C}$ be the learned cycle-query bank. For each sample, the base branch gathers L consecutive phases modulo P_d and transposes the gathered query into channel-first form,

$$\left[\widetilde{Q}^{(P_d)}(r_t)\right]_{c,\tau} = Q^{(P_d)}_{(r_t+\tau) \bmod P_d, c},$$
$$\widetilde{Q}^{(P_d)}(r_t) \in \mathbb{R}^{C\times L}.$$

Here $c \in \{1, \ldots, C\}$ and $\tau \in \{0, \ldots, L-1\}$; after stacking samples, the query has shape $B \times C \times L$. The base branch applies channel attention,

$$U_{\text{base}} = \text{MHA}\Big(\widetilde{Q}^{(P_d)}(r_t), X_c, X_c\Big) \in \mathbb{R}^{B \times C \times L}. \tag{1}$$

The scale branch softmax-mixes depthwise convolutions of widths $3, 5, 9, 17$, batch-normalizes the result, and adds it to X_c through a sigmoid gate to obtain X_{scale}; it then reuses the base P_d query bank with keys and values X_{scale}. The multi-cycle branch uses three separate query banks with periods $P_d/2$, P_d, and $2P_d$ and averages their attended channel representations with keys and values X_c. All five attention calls—Base, Scale, and the three Multi-cycle calls—reuse the same MHA parameters. The base, scale, and multi-cycle branches share one projection, two-layer GELU MLP $\mathcal{G}$, and output head; let D denote their hidden width. Writing $X_{\text{base}} = X_{\text{multi}} = X_c$, and U_k for branch k's attended representation, the implemented shared path is

$$\begin{aligned} Z_k &= W_{\text{in}}(X_k + U_k), \\ \bar{F}_k &= W_{\text{out}}(\mathcal{G}(Z_k) + Z_k)\,, \\ k &\in \{\text{base}, \text{scale}, \text{multi}\}. \end{aligned}$$

The refinement branch is explicitly conditioned on both the base forecast and its projected state: $q = W_r \bar{F}_{\text{base}}$, $u = \mathcal{R}(q + Z_{\text{base}})$, and $\bar{F}_{\text{ref}} = \bar{F}_{\text{base}} + \sigma(g_{\text{ref}}) W_h(u+q)$. Here $\mathcal{R}$ is the implemented two-layer GELU refinement MLP. The cycle branch subtracts a learned recurrent cycle indexed from r_t, forecasts the residual with an $L \to D \to H$ ReLU MLP, and adds $\sigma(g_{\text{cyc}})$ times the future cycle indexed from $(r_t + L) \bmod P_d$.

For each sample, the router reduces X^{n} over the indicated lookback and channel axes; Δ_L denotes a first difference along the lookback axis. Stacking the four statistics gives

$$\begin{aligned} s(X^{\text{n}}) &= [s_1, s_2, s_3, s_4] \in \mathbb{R}^{B \times 4}, \\ s_1 &= \text{std}_{L,C}(X^{\text{n}}), \\ s_2 &= \text{mean}_{L,C}|X^{\text{n}}|, \\ s_3 &= \text{mean}_{L,C}|\Delta_L X^{\text{n}}|, \\ s_4 &= \text{mean}_{C}(X^{\text{n}}_L - X^{\text{n}}_1). \end{aligned} \tag{2}$$

Given an initial normalized prior $p \in \mathbb{R}^5_+$ with $\sum_{k=1}^5 p_k = 1$, the router initializes a learnable bias $b \leftarrow \log p$ and computes

$$\pi(X^{\text{n}}) = \text{softmax}(b + \sigma(g_\pi) h_\phi(s(X^{\text{n}}))) \in \mathbb{R}^{B \times 5}, \tag{3}$$

and returns

$$S_\theta(X, r_t) = \mathcal{D}_X\left(\left[\sum_{k=1}^{5} \pi_k(X^{\text{n}}) \bar{F}_k(X^{\text{n}}, r_t)\right]^\top\right). \tag{4}$$

The prior is either uniform or mapped from a validation simplex over four clean experts into the five internal branches; the exact mapping is given in the supplement. The bias b, router h_ϕ, and gate g_π are optimized jointly. The sigmoid of g_π starts near zero, allowing the model to begin near the initial prior and learn sample-specific deviations.

Stage 1: Distilled Routed Anchor

Figure 2 summarizes the training-only graph. We first train the same CycleRoutedTQNet architecture as an anchor A_{θ_A}. Chronos-2 Ansari et al. (2025) predictions $Z_{\text{Chr},H}(X)$ are generated once for training windows and stored in index-aligned caches. For two forecast tensors T_1, T_2, define

$$\ell_\mu(T_1, T_2) = \text{MSE}(T_1, T_2) + \mu\, \text{MAE}(T_1, T_2). \tag{5}$$

The anchor objective is

$$\mathcal{L}_A = \ell_{\mu_A}(A_{\theta_A}(X), Y) + \lambda_{\text{Chr}} \ell_{\mu_A}(A_{\theta_A}(X), Z_{\text{Chr},H}(X)). \tag{6}$$

Here μ_A weights MAE within each Stage-1 loss term, and λ_{Chr} weights the Chronos-cache term. Chronos is absent after cache construction. This stage provides a stable routed-student initialization and an anchor prediction with useful temporal-query behavior.

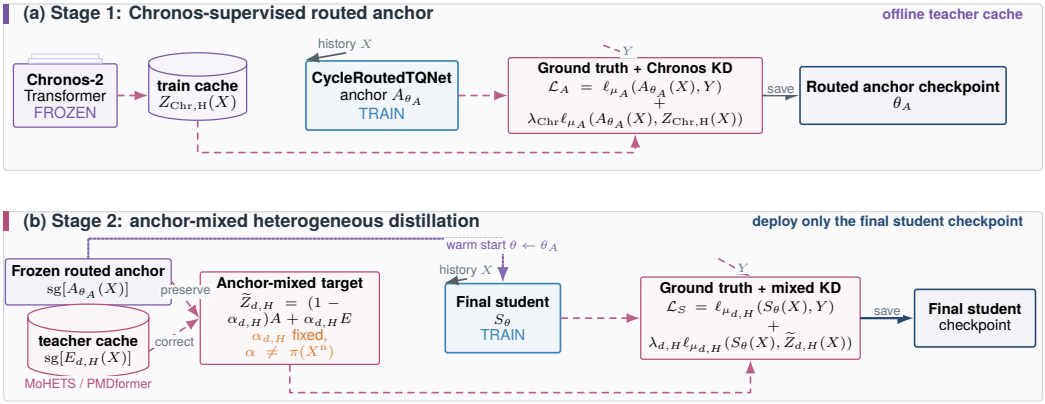

Figure 2: AnchorMix training, split by optimization stage. (a) Chronos-2 is evaluated offline to supervise a routed anchor. (b) The frozen anchor and cached heterogeneous teacher form the fixed-α target for a warm-started student. Dashed arrows are training-only supervision and dotted arrows denote initialization; only the final student checkpoint is deployed.

Stage 2: Anchor-Mixed Heterogeneous Target

Let $E_{d,H}$ be the external teacher configured for dataset d and horizon H. We use teachers with inductive biases distinct from the routed student: MoHETS Ortigossa et al. (2026) contributes routed temporal/spectral experts, whereas PMDformer Hu et al. (2026) decouples patch-level trend and residual shape. For each training window, the frozen anchor and external teacher form

$$\widetilde{Z}_{d,H}(X) = (1 - \alpha_{d,H})\,\mathrm{sg}[A_{\theta_A}(X)] + \alpha_{d,H}\,\mathrm{sg}[E_{d,H}(X)]\,, \tag{7}$$

where sg stops gradients and $\alpha_{d,H} \in [0, 1]$ controls the external correction. Only training-window external forecasts are loaded by the student optimizer.

The final student is initialized with $\theta \leftarrow \theta_A$ and minimizes

$$\mathcal{L}_S = \ell_{\mu_{d,H}}(S_\theta(X), Y) + \lambda_{d,H}\ell_{\mu_{d,H}}(S_\theta(X), \widetilde{Z}_{d,H}(X)). \tag{8}$$

Here $\mu_{d,H}$ is the Stage-2 MAE weight and $\lambda_{d,H}$ is the Stage-2 distillation weight for task (d, H). Equations 7–8 serve two purposes. The anchor term prevents a strong routed solution from being replaced wholesale by an imperfect external teacher. The external term exposes corrections that are unavailable from another copy of the same architecture. Warm starting makes the second stage a local projection of heterogeneous knowledge into one routed student.

Why Heterogeneous Mixing Can Help

Fix a task (d, H) and write $\alpha = \alpha_{d,H}$ and $E = E_{d,H}$ in this analysis. Let $e_A = A_{\theta_A}(X) - Y$ and $e_E = E(X) - Y$ denote anchor and external-teacher errors. For an $H \times C$ forecast, let $\langle T_1, T_2\rangle_m = (HC)^{-1}\sum_{j,c}(T_1)_{j,c}(T_2)_{j,c}$ and $\|T_1\|_m^2 = \langle T_1, T_1\rangle_m$, and define $R_A = \mathbb{E}\|e_A\|_m^2$, $R_E = \mathbb{E}\|e_E\|_m^2$, and $\Gamma_{AE} = \mathbb{E}\langle e_A, e_E\rangle_m$. The MSE risk of the mixed target in Equation 7 is

$$R(\alpha) = (1 - \alpha)^2 R_A + \alpha^2 R_E + 2\alpha(1 - \alpha)\Gamma_{AE}. \tag{9}$$

Its constrained minimizer is

$$\alpha^* = \mathrm{clip}_{[0,1]}\frac{R_A - \Gamma_{AE}}{R_A + R_E - 2\Gamma_{AE}}. \tag{10}$$

Thus teacher strength alone is insufficient: the cross-error moment determines whether the second source contributes a useful correction. This motivates selecting structurally distinct teachers and retaining the anchor rather than replacing it.

Single-Model Inference

For every dataset and horizon, deployment loads one final CycleRoutedTQNet checkpoint. Inference consists only of RevIN-style reversible instance standardization Kim et al. (2022), the five internal branches, the router in Equation 4, and the matching denormalization. Chronos-2, the frozen anchor, MoHETS, PMDformer, teacher caches, and simplex fitting are not part of the inference graph.

Training and Evaluation Protocol

Training consumes only train ground truth and index-aligned train teacher caches. Both stages select checkpoints with their supervised validation criterion; teacher predictions and distillation terms are excluded from validation. The task-specific coefficients and optimization recipes were established during preliminary benchmark development, which included test-metric inspection, and then frozen before the reported three-seed replay. The replay does not construct a test loader during training and evaluates test only after each checkpoint is fixed; complete split and cache checks are provided in the supplement.

Experiments

Setup

We evaluate ETTh1, ETTh2, ETTm1, ETTm2, and Weather under the common long-horizon setting $L = 720$ and $H \in \{96, 192, 336, 720\}$. The ETT variants cover two transformer stations at hourly and 15-minute resolutions, while Weather supplies a higher-dimensional meteorological domain; we follow the benchmark data and splits released with Informer for ETT Zhou et al. (2021) and Autoformer for Weather Wu et al. (2021). Cycle lengths are 24 for hourly ETT, 96 for minute-level ETT, and 144 for Weather. **Every number in every main result table is the mean over three independent random seeds** (seed IDs 2024, 2025, and 2026). Each seed is an independently trained model from a different random initialization, with its own validation-selected checkpoint per seed and per horizon, giving 60 selected runs in total; the reported spread is the sample standard deviation across these three seeds. Only ablations that deliberately use a single seed are run that way, and every such case is explicitly labeled as single-seed. The strongest selected comparisons are MoHETS on ETTh1, ETTm1, and Weather, and PMDformer on ETTh2 and ETTm2. We evaluated released MoHETS checkpoints and retrained PMDformer from its official code on the same local CSV files before constructing teacher caches. Both stages use Adam and batch size 32; Stage 1/2 use epoch cap/patience 20/5 and 15/4, and learning rates 10^{-3} and 10^{-4}, respectively; all horizon-specific settings are in the supplement.

Single-Checkpoint Forecasting Quality

Table 2 compares the three-seed student mean and sample standard deviation with the strongest selected external teacher. AnchorMix improves seven of ten dataset-average metrics, including both MSE and MAE on ETTh2, ETTm1, and ETTm2. It also reduces ETTh1 MAE, while both Weather differences remain at the 2.5×10^{-4} scale. These results establish that one routed checkpoint can absorb heterogeneous corrections while retaining teacher-level forecasting quality across the full benchmark suite.

Three-Seed Stability

Table 2 reports the frozen-recipe replay over seeds 2024, 2025, and 2026 directly alongside the main comparison. ETTm1 and ETTm2 retain double-metric advantages in every seed, and ETTh2 does so in two of three. Dataset-average standard deviations remain at most 0.002020 MSE and 0.001227 MAE. A 20,000-replicate hierarchical moving-block bootstrap further yields MAE intervals favoring AnchorMix on ETTh1, ETTm1, and ETTm2; full horizon-level intervals are in the supplement.

Dataset	Teacher	External teacher		AnchorMix ($n = 3$)		Δ ours−teacher	
		MSE	MAE	MSE ± s.d.	MAE ± s.d.	MSE	MAE
ETTh1	M	**0.380232**	0.416165	0.385810 ±0.000550	**0.413664** ±0.000305	+0.005578	-0.002501
ETTh2	P	0.337869	0.382207	**0.336509** ±0.002020	**0.379123** ±0.001227	-0.001360	-0.003084
ETTm1	M	0.329178	0.364814	**0.327456** ±0.000178	**0.361182** ±0.000067	-0.001722	-0.003632
ETTm2	P	0.246052	0.305035	**0.245170** ±0.000206	**0.302954** ±0.000183	-0.000882	-0.002081
Weather	M	**0.216019**	**0.248080**	0.216261 ±0.001103	0.248190 ±0.000910	+0.000241	+0.000110

Table 2: Four-horizon averages. AnchorMix entries are mean ± sample standard deviation over three independently trained seeds (2024, 2025, and 2026); Δ uses the mean. Lower is better. M denotes fixed released MoHETS checkpoints Ortigossa et al. (2026); P denotes fixed official-release PMDformer result-file values Hu et al. (2026), checked by local retraining in the supplement. Each AnchorMix run deploys one CycleRoutedTQNet checkpoint.

Dataset	Variant	MSE	MAE
ETTh2	Full	**.336599**	**.379269**
	External only	.337906	.382212
	Anchor only	.345929	.382475
ETTm2	Full	**.244945**	**.303119**
	External only	.245599	.304115
	Anchor only	.256044	.310112

Table 3: Seed-2024 target-composition ablation, averaged over four horizons.

Why Anchor Mixing Matters

Table 3 isolates the two sources in the mixed target under the same seed-2024 validation-selected protocol. The complete target is best on both metrics for both datasets. Using only the external teacher raises ETTh2 MSE/MAE by $0.001308/0.002943$ and ETTm2 by $0.000654/0.000996$; using only the anchor raises them by as much as $0.011100/0.006993$. The result rules out either single source as the explanation and directly supports anchor preservation plus heterogeneous correction. The supplement separately reports the initialization ablation.

Router counterfactual. A checkpoint-preserving counterfactual disables only the input-dependent router correction while retaining learned priors and experts. Dynamic routing improves both metrics in 14/20 tasks and MAE in 16/20, with its largest contribution on ETTm1, where AnchorMix improves all eight teacher-comparison cells. Complete task-level results are in the supplement.

Recent Published Comparisons

Table 4 broadens the comparison beyond the two reproduced teachers. Our averages are lower on both metrics for 19 of 20 dataset–method pairs and in 39 of 40 MSE/MAE cells. The pair/cell counts are 19/20 and 39/40 for seeds 2024 and 2025 individually, and 20/20 and 40/40 for seed 2026. The table preserves each source's reported lookback or search range; notably, TimeFuse and TSCOMP retain multiple predictors whereas AnchorMix deploys one. Table 2 provides the complementary controlled comparison with locally aligned teacher predictions.

Horizon-Level Results

Table 5 shows that the average gains are not produced by one favorable horizon. Across three-seed means, AnchorMix wins 29 of 40 metric cells. ETTm1 contributes all eight wins, ETTm2 seven, and ETTh2 six, demonstrating that the strongest gains span short, medium, and long horizons.

Dataset	TF (96)	PF (720)	xCPD (512)	TC (48–512)	AnchorMix (720)
ETTh1	.430/.433	.403/.415	.401/.416	.407/.424	**.386/.414**
ETTh2	.365/.395	.347/.389	.345/.392	**.336**/.384	.337/**.379**
ETTm1	.369/.388	.347/.374	.343/.371	.341/.371	**.327/.361**
ETTm2	.274/.319	.251/.314	.251/.312	.246/.306	**.245/.303**
Weather	.240/.270	.223/.261	.221/.259	.222/.256	**.216/.248**

Table 4: Published four-horizon averages (MSE/MAE): TF = TimeFuse Table 2 Liu et al. (2025b), PF = PhaseFormer Table 6 Niu et al. (2026), xCPD = TSMixer+xCPD Table 2 Li et al. (2026a), and TC = TSCOMP Table 6b Liang et al. (2026). Parentheses give lookback or search range; protocols and deployment forms differ.

Dataset	H	Teacher	T MSE	Ours MSE	T MAE	Ours MAE
ETTh1	96	M	0.345119	**0.340174**	0.384344	**0.377534**
ETTh1	192	M	**0.373406**	0.375287	0.408466	**0.403219**
ETTh1	336	M	**0.390189**	0.403276	0.424530	**0.421133**
ETTh1	720	M	**0.412211**	0.424503	**0.447315**	0.452772
ETTh2	96	P	0.269863	**0.266598**	0.329138	**0.325635**
ETTh2	192	P	0.333992	**0.332104**	0.373941	**0.368456**
ETTh2	336	P	**0.357216**	0.363158	**0.396494**	0.398501
ETTh2	720	P	0.390404	**0.384176**	0.429254	**0.423900**
ETTm1	96	M	0.269313	**0.264227**	0.325420	**0.319889**
ETTm1	192	M	0.306912	**0.306808**	0.351844	**0.348709**
ETTm1	336	M	0.340353	**0.339861**	0.374132	**0.370924**
ETTm1	720	M	0.400134	**0.398927**	0.407860	**0.405208**
ETTm2	96	P	**0.155494**	0.155726	0.240500	**0.239779**
ETTm2	192	P	0.213820	**0.213362**	0.282372	**0.280754**
ETTm2	336	P	0.266960	**0.266237**	0.320755	**0.318156**
ETTm2	720	P	0.347934	**0.345356**	0.376512	**0.373125**
Weather	96	M	0.139776	**0.138819**	0.182040	**0.180492**
Weather	192	M	0.182775	**0.182772**	0.225109	**0.224363**
Weather	336	M	**0.233944**	0.235061	**0.266344**	0.267109
Weather	720	M	**0.307583**	0.308390	**0.318828**	0.320798

Table 5: Three-seed-mean horizon comparison with the selected external teacher. P values are official-release PMDformer result-file metrics Hu et al. (2026); M values are locally evaluated from released MoHETS checkpoints Ortigossa et al. (2026). Bold marks the lower value in each metric pair.

Mechanism Evidence: Complementarity and Projection

Equation 9 is supported by a full-array diagnostic over 20 dataset–horizon pairs. Every analytically optimized convex pair improves on its better single predictor. The mean centered error correlation is 0.9453, yet the mean oracle MSE gain is 0.003868 (range 0.000737–0.008298), showing that small departures from perfectly aligned errors are consequential at current benchmark margins. The exact Stage-2 coefficients produce nearly identical static risk; their largest gap to the analytic optimum is 1.21×10^{-4} MSE.

Table 6 asks how much of a two-model convex upper bound survives single-checkpoint projection. Relative to the routed anchor, AnchorMix recovers 81.1%/69.3% of the ETTh2 upper-bound MSE/MAE improvement and 98.4%/103.3% on ETTm2. The student therefore preserves most or all of the measurable combination benefit while removing the second predictor.

Single-Checkpoint Deployment

Across the 20 dataset–horizon settings, the final model has 4.06–5.36M parameters (4.56M mean) and does not retain either Stage-2 teacher. Its average RTX 5090 latency is 1.19 ms per batch of 32. Table 7

Data	Predictor	#	MSE	MAE
ETTh2	Routed anchor	1	0.347038	0.383160
	PMDformer	1	0.337869	0.382207
	Two-model bound	2	0.334049	0.377337
	AnchorMix	1	**0.336509**	**0.379123**
ETTm2	Routed anchor	1	0.256137	0.310248
	PMDformer	1	0.246052	0.305035
	Two-model bound	2	0.244991	0.303184
	AnchorMix	1	**0.245170**	**0.302954**

Table 6: Projection diagnostic. # is the number of deployed predictors. Bold marks the best single checkpoint; underline marks the overall best. The two-model row is a convex diagnostic bound, while AnchorMix deploys one validation-selected CycleRoutedTQNet checkpoint.

Case	Graph	#	Params	ms/batch	MiB
Stage 1†	Chronos-2 cache	0	119.478	114.657	1471.9
ETTh1–M	AnchorMix	1	5.344	1.198	47.0
	Teacher (M)	1	2.396	214.961	222.2
	Student+M	2	7.740	217.614	243.8
ETTh2–P	AnchorMix	1	5.344	1.203	46.4
	Teacher (P)	1	0.568	0.528	20.1
	Student+P	2	5.912	1.852	50.8

Table 7: Matched-hardware teacher and deployment profile. M = MoHETS; P = PMDformer. †Chronos-2 is the offline cache adapter (median wall time over 11 calls after three warm-ups); its # is zero because it is absent from deployment. Other rows use forward-only CUDA-event timing; Student+M/P loads and executes both predictors, whereas AnchorMix loads only its distilled checkpoint. Data loading is excluded.

accounts for the offline Chronos-2 cache teacher, one AnchorMix checkpoint, each external teacher alone, and the corresponding two-predictor runtime graph under matched FP32, $L = H = 720$, and batch-size-32 conditions. Chronos-2 has 119.478M parameters and requires 114.657 ms and 1471.9 MiB for one ETT-shaped cache batch, but this offline Stage-1 cache phase contributes zero deployed predictors. Removing the runtime ensemble yields a $181.7\times$ speedup and 80.7% lower peak allocated memory for ETTh1–MoHETS, and a $1.54\times$ speedup with 8.6% lower memory for ETTh2–PMDformer. The teacher-only rows localize this cost: MoHETS accounts for most of the ETTh1 runtime, while the lightweight PMDformer case provides a conservative deployment comparison. In both cases the deployed graph remains the same single-student architecture.

LIMITATIONS

The external teacher and coefficient are configured per dataset–horizon task, so coverage is scoped to the five benchmarks and four horizons studied here rather than a full Cartesian sweep. Equations 9–10 give an inexpensive validation criterion for prospective teacher screening; both directions act only during training and preserve AnchorMix's single-checkpoint inference graph. We do not claim complete conditional calibration beyond the reported interval diagnostics.

CONCLUSION

We presented AnchorMix, a two-stage framework that compresses a Chronos-supervised anchor and an architecture-heterogeneous teacher into one routed temporal-query checkpoint: the anchor-mixed target retains established student behavior while exposing complementary corrections, and all teachers disappear after training. Across 20 tasks and three seeds, the resulting model improves 29/40 controlled teacher cells and 39/40 cells against four recent published tables and delivers up to a $181.7\times$ matched-ensemble speedup with 80.7% lower peak allocated memory—heterogeneous forecasting knowledge exploited without deploying a heterogeneous forecasting system.

Reproducibility statement

All 60 selected runs are backed by array-level metric recomputation, unique checkpoint hashes, ordered teacher-cache indices, and recorded validation selections; the supplement provides teacher reproduction, split intervals, complete uncertainty estimates, configuration tables, and matched-profile details, and the anonymous code package contains the corresponding launchers and machine-readable reports. Generative AI tools were used only for code implementation and debugging, and for copy-editing and language polishing of the manuscript. The authors take full responsibility for all research content, claims, results, figures, tables, citations, and artifacts.

AI use statement

Generative AI tools were used only for code implementation and debugging, and for copy-editing and language polishing of the manuscript. The authors take full responsibility for all research content, claims, results, figures, tables, citations, and artifacts.

Ethics statement

This work uses public time-series forecasting benchmarks and released model checkpoints. No human subjects or personally identifiable information are involved. Because improved forecasting accuracy can influence operational decisions in domains such as energy and finance, reported gains should be read together with the deployment-cost, calibration, and fairness considerations noted in the limitations discussion.

References

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. URL https://arxiv.org/abs/2510.15821.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Ao Hu, Liangjian Wen, Jiang Duan, Yong Dai, Yan He, Dongkai Wang, Jun Wang, Yukun Zhang, Ruoxi Jiang, and Zenglin Xu. PMDformer: Patch-mean decoupling information transformer for long-term forecasting. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rfJ41gK9Ct.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-llm: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/680b2a8135b9c71278a09cafb605869e-Abstract-Conference.html.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cGDAkQo1C0p.

Dongyuan Li, Shun Zheng, Chang Xu, Jiang Bian, and Renhe Jiang. Routing channel-patch dependencies in time series forecasting with graph spectral decomposition. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=uIPAuyno4Z.

Yuqi Li, Kuiye Ding, Chuanguang Yang, Szu-Yu Chen, and Yingli Tian. Distilling time series foundation models for efficient forecasting. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4631–4635, 2026b. doi: 10.1109/ICASSP55912.2026.11460474. URL https://arxiv.org/abs/2601.12785.

Shuang Liang, Chaochuan Hou, Xu Yao, Shiping Wang, Hailiang Huang, Songqiao Han, and Minqi Jiang. Beyond holistic models: Systematic component-level benchmarking of deep multivariate time-series forecasting. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2*, 2026. doi: 10.1145/3770855.3817551. URL https://arxiv.org/abs/2605.26562.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. Cyclenet: Enhancing time series forecasting through modeling periodic patterns. In *Advances in Neural Information Processing Systems*, volume 37, pp. 106315–106345. Curran Associates, Inc., 2024. doi: 10.52202/079017-3373. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bfe7998398779dde03cad7a73b1f81b6-Abstract-Conference.html.

Shengsheng Lin, Haojun Chen, Haijie Wu, Chunyun Qiu, and Weiwei Lin. Temporal query network for efficient multivariate time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 37797–37814. PMLR, 2025. URL https://proceedings.mlr.press/v267/lin25e.html.

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*, 2024a. URL https://openreview.net/forum?id=JePfAI8fah.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32369–32399. PMLR, 2024b. URL https://proceedings.mlr.press/v235/liu24cb.html.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317. PMLR, 2025a. URL https://proceedings.mlr.press/v267/liu25be.html.

Zhining Liu, Ze Yang, Xiao Lin, Ruizhong Qiu, Tianxin Wei, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. Breaking silos: Adaptive model fusion unlocks better time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40022–40042. PMLR, 2025b. URL https://proceedings.mlr.press/v267/liu25cm.html.

Juntong Ni, Zewen Liu, Shiyu Wang, Ming Jin, and Wei Jin. Timedistill: Efficient long-term time series forecasting with mlp via cross-architecture distillation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.1*, pp. 1113–1124, 2026. doi: 10.1145/3770854.3780218. URL https://arxiv.org/abs/2502.15016.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Yiming Niu, Jinliang Deng, and Yongxin Tong. Phaseformer: From patches to phases for efficient and effective time series forecasting. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Lk9SqMQzhX.

Evandro S. Ortigossa, Guy Lutsker, and Eran Segal. MoHETS: Long-term time series forecasting with mixture-of-heterogeneous-experts. *arXiv preprint arXiv:2601.21866*, 2026. URL https://arxiv.org/abs/2601.21866.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22419–22430. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 78478–78497. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhong25a.html.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325. URL https://ojs.aaai.org/index.php/AAAI/article/view/17325.