ADAPTIVE QUANTILE RESPONSE-BASED KNOWLEDGE DISTILLATION FOR FOUNDATION MODELS IN TIME-SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation has been widely used to transfer knowledge from large foundation models to compact students, yet remains underexplored for time-series foundation models (TSFMs), particularly in multi-teacher probabilistic forecasting. To address this gap, we propose a multi-teacher response-based knowledge distillation framework that consolidates complementary forecasting knowledge from multiple pretrained TSFMs into a single compact student TSFM. Our method introduces Order-Huber Relative KD (OH-RKD) for teacher-guided cross-quantile distillation and Sample-wise Relative Adaptive Weighting (SRAW), which adaptively adjusts the distillation strength of the selected teacher according to its forecasting quality relative to the student. Extensive experiments on FEV-Bench and GIFT-Eval show that our method reduces SQL on FEV-Bench by 9.03% and WQL on GIFT-Eval by 10.96% relative to the matched GT-only student, demonstrating the effectiveness of multi-teacher distillation for TSFMs.

1 INTRODUCTION

Time-series forecasting is a fundamental problem with broad applications in energy, finance, transportation, weather, healthcare, and industrial systems, where accurate predictions of future dynamics are essential for reliable decision-making. Recent advances in large-scale pretraining have shifted time-series forecasting from the conventional paradigm of training a specialized model for each dataset toward general-purpose TSFMs (Liang et al., 2024). Models such as Chronos, TimesFM, Moirai, and GTT demonstrate that pretraining on large and heterogeneous time-series corpora can yield strong zero-shot forecasting capabilities across previously unseen domains and datasets (Ansari et al., 2024; Das et al., 2024; Liu et al., 2025; Feng et al., 2024). Beyond point forecasting, modern TSFMs increasingly provide probabilistic forecasts to characterize predictive uncertainty, while large-scale pretraining enables a single model to operate across diverse frequencies, horizons, and variable configurations. However, the growing scale and complexity of TSFMs also introduce substantial computational and storage costs, motivating knowledge distillation as a promising approach for transferring the forecasting capabilities of large TSFMs into more efficient models.

Knowledge distillation has emerged as an effective paradigm for transferring knowledge from large foundation models to smaller and more efficient models (Hinton et al., 2015). Particularly in large language models (LLMs), response-based knowledge distillation has received extensive attention due to its simplicity and effectiveness (Ko et al., 2024; Agarwal et al., 2024). Instead of requiring access to the internal architectures or intermediate representations of teacher models, response-based knowledge distillation transfers the predictive behaviors encoded in teacher outputs, enabling knowledge transfer between heterogeneous models. Recent studies have explored response-based knowledge distillation for LLM compression and capability transfer, including instruction-following alignment (Gu et al., 2024), reasoning capability preservation (Hsieh et al., 2023; Shridhar et al., 2023), and multi-teacher knowledge aggregation (Tian et al., 2025; Ma et al., 2026). These works demonstrate that the outputs generated by powerful foundation models contain rich information beyond conventional supervised labels, allowing compact students to inherit behaviors from large-scale pretrained models. Furthermore, recent advances suggest that effective distillation requires not only matching teacher responses, but also determining which teacher knowledge should be transferred and how strongly it should influence student learning (Jin et al., 2026; Shing et al., 2025).

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.

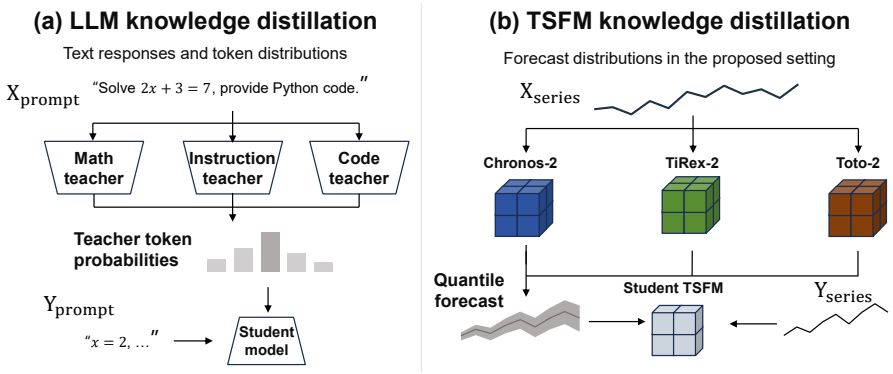


Figure 1: Comparison between response-based knowledge distillation for LLMs and time-series foundation models. While LLM distillation primarily transfers textual responses or token distributions, TSFM distillation requires transferring continuous probabilistic forecasts from heterogeneous forecasting models to a student TSFM.

Despite the rapid progress of response-based KD in LLMs, directly applying these approaches to time-series forecasting remains challenging and relatively underexplored. As illustrated in Figure 1, LLM distillation primarily operates on discrete token distributions or generated textual responses, whereas TSFM distillation must transfer continuous probabilistic forecasts, often represented by multiple quantiles(Miller et al., 2024). Moreover, heterogeneous TSFMs can exhibit complementary forecasting behaviors even on the same dataset, as different models may capture distinct temporal patterns, frequency characteristics, and uncertainty structures (Dey et al., 2026). Consequently, a single dominant teacher or uniform aggregation of multiple teachers may fail to fully exploit the complementary predictive knowledge available across TSFMs. These differences raise fundamental questions regarding which teacher should provide supervision, how probabilistic forecasting responses should be transferred, and how strongly teacher supervision should influence the student.

Who to learn from, what to learn, and how to learn.

To address these fundamental questions, we introduce a multi-teacher response-based knowledge distillation framework for TSFMs. Our contributions are summarized as follows:

- **Multi-Teacher TSFM Distillation.** We propose a unified response-based knowledge distillation framework that transfers complementary forecasting knowledge from multiple heterogeneous pretrained TSFMs into a compact student TSFM.
- **Order-Huber Relative KD.** We introduce OH-RKD, a quantile-based distillation objective that combines robust alignment of corresponding quantiles with teacher-guided cross-quantile constraints.
- **Sample-wise Relative Adaptive Weighting.** We propose SRAW, a sample-wise weighting strategy that dynamically adjusts the distillation strength of the selected teacher according to its forecasting quality relative to the student.

Extensive experiments on FEV-Bench and GIFT-Eval (Shchur et al., 2025; Aksu et al., 2024) demonstrate the effectiveness of our framework. Compared with the matched GT-only student under the same architecture and training configuration, our method reduces SQL on FEV-Bench by 9.03% and WQL on GIFT-Eval by 10.96%. These results demonstrate the effectiveness of multi-teacher response-based knowledge distillation for building compact capable time-series foundation models.

2 Background

2.1 Time-Series Foundation Models

Recent advances in large-scale pretraining have led to a variety of TSFMs with different architectures and modeling paradigms. Chronos-2 (Ansari et al., 2025) adopts an encoder-only Transformer

architecture that maps normalized time-series patches to embeddings and directly produces multi-step quantile forecasts. Its group attention mechanism enables information sharing across related series and covariates. TiRex-2 (Podest et al., 2026) further improves the scalability and efficiency of TSFMs through large-scale pretraining and enhanced probabilistic prediction. Toto-2 (Khwaja et al., 2026) explores general-purpose forecasting by modeling complex temporal dependencies across different variables and forecasting scenarios. Despite their differences in architectures, pretraining datasets, and inductive biases, these TSFMs demonstrate strong but complementary forecasting capabilities, providing diverse sources of predictive knowledge for distillation.

2.2 Response-Based Knowledge Distillation

Knowledge distillation transfers knowledge from one or more teacher models to a student model by encouraging the student to reproduce informative teacher behaviors. (Hinton et al., 2015) introduced response-based KD, in which the student learns from the teacher's output in addition to ground-truth supervision. With the development of foundation models, response-based KD has become an important approach for model compression and capability transfer, particularly in LLMs, where recent studies explore response alignment (Peng & Zhang, 2025), adaptive transfer, and multi-teacher aggregation. Beyond classification tasks, response-based KD has also been extended to regression problems. (Chen et al., 2017) and (Saputra et al., 2019) show that teacher supervision should be adaptively controlled according to teacher reliability. For multi-teacher settings, (Zhang et al., 2022) further highlights the importance of selecting informative knowledge from different teachers.

2.3 Knowledge Distillation for Time-Series Forecasting

Despite the broad success of response-based KD in classification, LLMs, and regression tasks, its application to time-series forecasting remains relatively underexplored. Distillation for TSFMs introduces several additional challenges. Rather than producing a single categorical distribution, TSFMs perform continuous probabilistic forecasting and often generate predictive distributions or quantile forecasts (Woo et al., 2024), requiring distillation methods to preserve forecasting uncertainty and temporal dynamics. Recent studies have begun to investigate KD for TSFMs. (Li et al., 2026) explores forecasting-aware distillation by transferring knowledge from large TSFMs to lightweight models through horizon-aware objectives, while (Dey et al., 2026) introduces adaptive teacher selection and uncertainty-aware distillation for scientific time-series forecasting. These studies demonstrate the potential of KD for improving the efficiency of forecasting models, but primarily focus on compressing individual large models into task-specific lightweight students. In contrast, comparatively little attention has been given to large-scale pretraining that aggregates predictive knowledge from heterogeneous TSFMs into a single student foundation model while preserving probabilistic forecasting capability across diverse datasets, horizons, and variable configurations.

3 Preliminaries

We consider probabilistic time-series forecasting, where X denotes the historical observations and Y denotes the future target. The student TSFM f_θ predicts K quantiles on a shared quantile grid

$$\mathcal{Q} = \{q_k\}_{k=1}^{K} = \{0.1, 0.2, \ldots, 0.9\}. \tag{1}$$

For a residual u, the pinball loss at quantile q is defined as (Koenker & Bassett Jr, 1978)

$$\rho_q(u) = \max\{qu, (q-1)u\}. \tag{2}$$

Training is performed in the normalized space of the student model. For each target series, we apply

$$g(z) = \operatorname{arcsinh}\left(\frac{z-\mu}{\sigma}\right), \tag{3}$$

where μ and σ are computed from the observed history. We denote the normalized ground truth, teacher prediction, and student prediction by $\widetilde{Y}$, $\widetilde{T}$, and $\widetilde{S}$, respectively. Since $g(\cdot)$ is monotonic, it preserves quantile ordering.

For a target series r with forecast horizon H, we define the quantile loss

$$\mathcal{L}_Q(Y_r, S_r) = \frac{2}{H} \sum_{h=1}^{H} \sum_{k=1}^{K} \rho_{q_k} \left(Y_{r,h} - S_{r,k,h} \right). \tag{4}$$

When $\mathcal{L}_Q$ is applied to a complete forecasting window, we use the same loss averaged over all target series in that window.

The ground-truth training objective is then simply

$$\mathcal{L}_{\mathrm{GT}} = \frac{1}{|\mathcal{R}_B|} \sum_{r \in \mathcal{R}_B} \mathcal{L}_Q \left(\widetilde{Y}_r, \widetilde{S}_r \right), \tag{5}$$

where $\mathcal{R}_B$ denotes the target series in the current batch.

4 METHODOLOGY

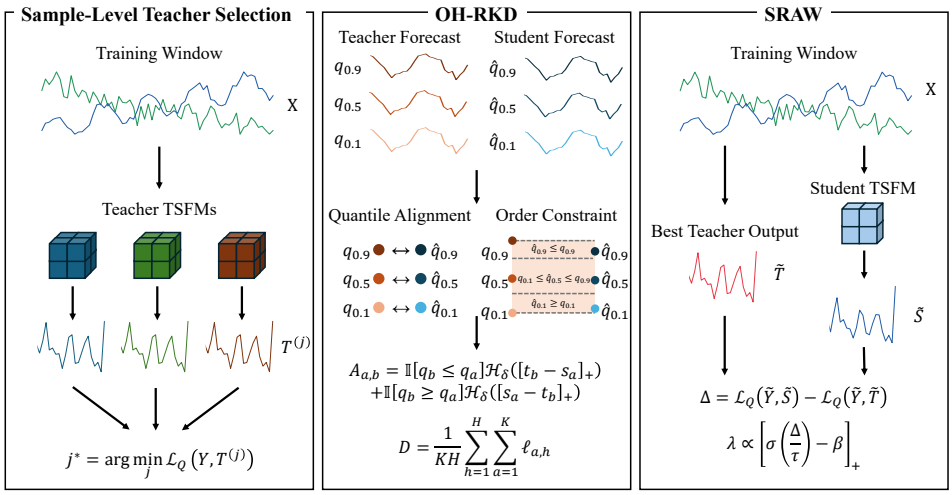


Figure 2: Overview of the proposed multi-teacher response-based distillation framework. For each training window, we select a teacher using ground-truth forecasting loss. OH-RKD aligns corresponding quantiles and adds teacher-guided cross-quantile constraints. SRAW adjusts the distillation weight according to the selected teacher's forecasting loss relative to the student's.

We distill probabilistic forecasts from a pool of heterogeneous TSFMs into a compact student TSFM. As shown in Figure 2, the framework has three components: **Sample-Level Teacher Selection** selects one teacher for each training window, **OH-RKD** transfers quantile forecasts through pointwise alignment and teacher-guided cross-quantile constraints, and **SRAW** adjusts the distillation weight according to the relative forecasting losses of the teacher and student. Teachers are frozen, and all teacher-dependent operations are used only during training.

4.1 SAMPLE-LEVEL TEACHER SELECTION

Different TSFMs exhibit complementary forecasting capabilities, and the teacher with the lowest forecasting error can vary across forecasting windows. We therefore select a teacher independently for each training window.

For window i, let $\widetilde{Y}_i$ denote the normalized future target, and let $\widetilde{T}_i^{(j)}$ denote the normalized quantile forecast from teacher j, where $j \in \{1, \ldots, J\}$. We score each teacher using the ground-truth quantile loss:

$$E_i^{(j)} = \mathcal{L}_Q\left(\widetilde{Y}_i, \widetilde{T}_i^{(j)}\right). \tag{6}$$

The teacher with the lowest loss is selected:

$$j_i^* = \arg\min_j E_i^{(j)}, \qquad \widetilde{T}_i = \widetilde{T}_i^{(j_i^*)}. \tag{7}$$

We use a single teacher for the entire forecasting window rather than selecting different teachers for individual quantiles. Teacher selection uses future targets only during training and is not required at inference.

4.2 Order-Huber Relative Knowledge Distillation

After selecting a teacher, we distill its forecast values and cross-quantile relationships. Independent alignment of corresponding quantiles does not use information about the relative positions of different quantiles. We therefore introduce **Order-Huber Relative Knowledge Distillation (OH-RKD)**.

For clarity, we consider one target series at one forecast step and omit the corresponding indices. Let s_a and t_b denote the normalized student and teacher predictions at quantiles q_a and q_b, respectively. We use the Huber penalty (Huber, 1992)

$$\mathcal{H}_\delta(u) = \begin{cases} \frac{1}{2}u^2, & |u| \le \delta, \\ \delta\left(|u| - \frac{1}{2}\delta\right), & |u| > \delta, \end{cases} \tag{8}$$

where the threshold is adapted to the teacher's predictive range:

$$\delta = 0.1[t_K - t_1]_+ + 10^{-6}. \tag{9}$$

Here, t_1 and t_K correspond to the lowest and highest quantiles, respectively. For student quantile q_a and teacher quantile q_b, we define

$$\begin{aligned} A_{a,b} = {} & \mathbb{I}[q_b \le q_a]\mathcal{H}_\delta\left([t_b - s_a]_+\right) \\ & + \mathbb{I}[q_b \ge q_a]\mathcal{H}_\delta\left([s_a - t_b]_+\right). \end{aligned} \tag{10}$$

The loss for student quantile q_a is

$$\ell_a = \frac{1}{K+1}\sum_{b=1}^{K} A_{a,b}. \tag{11}$$

For student quantile q_a, the pairwise terms impose lower bounds from lower teacher quantiles and upper bounds from higher teacher quantiles. The diagonal term aligns the corresponding teacher and student quantiles. There are a lower-bound terms and $K - a + 1$ upper-bound terms, giving $K + 1$ one-sided terms in total.

$$D_r = \frac{1}{KH}\sum_{h=1}^{H}\sum_{a=1}^{K} \ell_{r,a,h}. \tag{12}$$

OH-RKD combines corresponding-quantile alignment with teacher-guided cross-quantile constraints. The Huber penalty reduces the influence of large residuals, with its behavior controlled by the range-adaptive threshold δ.

4.3 Sample-wise Relative Adaptive Weighting

Even the selected teacher may not provide equally useful supervision for every target series. We therefore introduce **Sample-wise Relative Adaptive Weighting (SRAW)**, which adjusts the distillation strength according to the relative forecasting quality of the teacher and student.

For target series r, we define the student and teacher risks using the quantile loss introduced in Section 3:

$$R_r^S = \mathcal{L}_Q(\widetilde{Y}_r, \widetilde{S}_r), \qquad R_r^T = \mathcal{L}_Q(\widetilde{Y}_r, \widetilde{T}_r). \tag{13}$$

Their relative performance gap is

$$\Delta_r = R_r^S - R_r^T. \tag{14}$$

Thus, $\Delta_r > 0$ means the selected teacher has lower forecasting loss than the student. We additionally use a global schedule

$$\bar{\lambda}(s) = (1 - p_s)\lambda_{\text{start}} + p_s\lambda_{\text{end}}, \qquad p_s = \min\left(1, \frac{s}{S_{\text{train}}}\right), \tag{15}$$

and convert the relative performance gap into a sample-specific weight:

$$a_r = \left[\sigma\left(\frac{\text{sg}(\Delta_r)}{\tau}\right) - \beta\right]_+, \tag{16}$$

$$\lambda_r(s) = \text{sg}\left[\lambda_0(s) + \left(\bar{\lambda}(s) - \lambda_0(s)\right) a_r\right], \tag{17}$$

where

$$\lambda_0(s) = \min\{\lambda_{\min}, \bar{\lambda}(s)\}, \tag{18}$$

$\sigma(\cdot)$ denotes the sigmoid function, and $\text{sg}(\cdot)$ denotes stop-gradient.

Therefore, SRAW assigns stronger distillation when the teacher has a clear advantage over the student and gradually reduces teacher supervision as the student becomes competitive. The specific values of λ_{start}, λ_{end}, $\lambda_{\min}$, τ, and β are provided in Appendix A.

4.4 Overall Training Objective

The final objective combines the ground-truth quantile loss introduced in Section 3 with adaptively weighted OH-RKD:

$$\boxed{\mathcal{L}(\theta) = \mathcal{L}_{\text{GT}}(\theta) + \frac{1}{|\mathcal{V}_B|}\sum_{r\in\mathcal{V}_B}\lambda_r(s)D_r(\theta)} \tag{19}$$

where $\mathcal{V}_B$ denotes the set of target series with available teacher supervision. If no teacher prediction is available for a target series, its distillation term is omitted.

The proposed framework therefore integrates **Sample-Level Teacher Selection** for teacher selection, **OH-RKD** for order-aware probabilistic response transfer, and **SRAW** for adaptive distillation weighting. At inference time, all teacher models and distillation-specific components are discarded, leaving only the compact student TSFM for forecasting.

5 Experiments

5.1 Experimental Setup

Benchmarks and metrics. We evaluate without task-specific fine-tuning on **FEV-Bench** and **GIFT-Eval** (Shchur et al., 2025; Aksu et al., 2024). Our evaluation includes 100 FEV-Bench tasks

and 97 GIFT-Eval dataset–frequency–horizon configurations. We report SQL, MASE, WAPE, and WQL for FEV-Bench, and MASE, ND, WQL, and sMAPE for GIFT-Eval. ND is computed from the median forecast. All metrics are lower-is-better. Evaluation protocols and aggregation details are provided in Appendix B.

Models and training. We use Chronos-2 (Ansari et al., 2025), TiRex-2 (Podest et al., 2026), and Toto-2 (Khwaja et al., 2026) as frozen teachers, with forecasts generated offline. The student uses the Chronos-2 Small architecture and predicts nine quantiles on $\mathcal{Q} = \{0.1, 0.2, \ldots, 0.9\}$. We train for 100,000 steps on four NVIDIA H200 GPUs with a global batch size of 4096. Training data comprise the univariate Chronos pretraining corpus, real multivariate data from Sun et al. (2026), and synthetic multivariate data generated following the Chronos-2 strategy. Teacher selection and weighting use training targets only. Further configurations are in Appendix C.

Baselines. All methods share the student architecture, training data, initialization protocol, optimization settings, and training budget. **GT-only** uses only ground-truth quantile supervision. **Response-based KD** aligns corresponding teacher and student quantiles independently. We also adapt **AIL** (Saputra et al., 2019) and **CA-MKD** (Zhang et al., 2022) to quantile forecasting, using teacher-error-based imitation weighting and confidence-aware teacher response aggregation, respectively. Implementation details are in Appendix D.

Table 1: Forecasting performance on FEV-Bench and GIFT-Eval. Lower is better for all metrics.

Benchmark	Metric	GT-only	Response-based KD	AIL	CA-MKD	Ours
FEV-Bench	SQL	1.488574	1.401028	1.385247	1.425683	1.354216
	MASE	1.790383	1.699547	1.685392	1.732156	1.643362
	WAPE	0.277608	0.277588	0.275518	0.279064	0.273069
	WQL	0.228572	0.227667	0.227043	0.230527	0.225026
GIFT-Eval	MASE	1.919592	1.814306	1.745836	1.845271	1.641016
	ND	0.304579	0.298411	0.286137	0.302084	0.268611
	WQL	0.241482	0.237744	0.230168	0.245093	0.215008
	sMAPE	0.469525	0.469977	0.462319	0.476142	0.448113

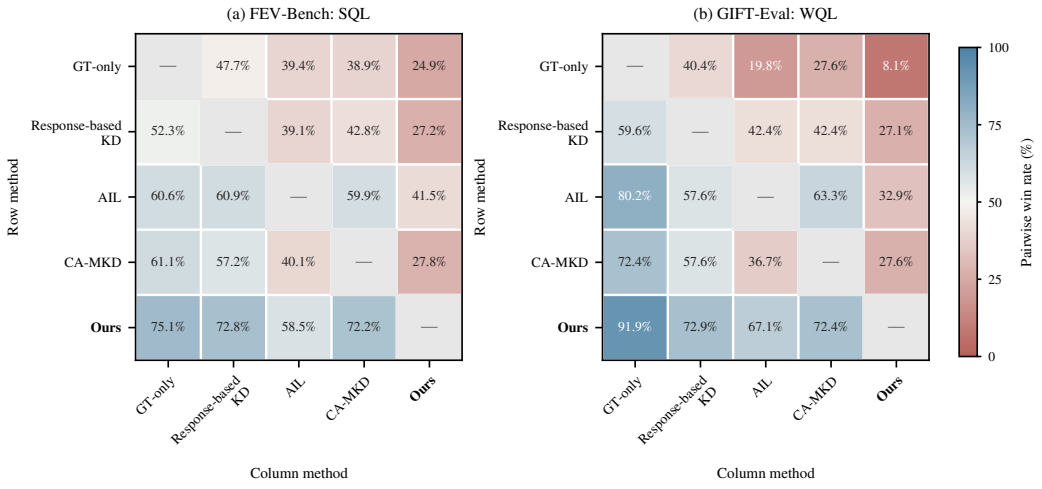


Figure 3: Dataset-balanced pairwise win rates. Each cell compares the row method against the column method; values above 50% favor the row method.

5.2 Main Results

Dataset-level comparison. We report dataset-balanced pairwise win rates, assigning 1 to a win, 0.5 to a tie, and 0 to a loss on each matched task, then averaging within and across datasets.

Table 1 shows that our framework outperforms GT-only and Response-based KD on all reported metrics. Compared with GT-only, it reduces FEV-Bench SQL and GIFT-Eval WQL by 9.03% and 10.96%, respectively. The corresponding reductions over Response-based KD are 3.34% and 9.56%.

In Figure 3, our method achieves win rates above 50% against every baseline on both benchmarks. Against GT-only and Response-based KD, the win rates are 75.1% and 72.8% on FEV-Bench, and 91.9% and 72.9% on GIFT-Eval. AIL is the most competitive baseline under this measure, but our method still achieves win rates of 58.5% and 67.1%, respectively.

Teacher selection. Figure 4 shows that teacher preferences vary across training data categories. Chronos-2 is selected most frequently in each category, while Toto-2 accounts for 40.4% of real multivariate windows and TiRex-2 for 26.1% of univariate windows. These differences motivate sample-level selection rather than reliance on a fixed teacher.

Adaptive weighting. Figure 5 compares the scheduled upper bound with the batch-mean effective KD weight. During the first 5,000 steps, the upper bound decreases from approximately 0.40 to 0.39, while the effective weight drops from approximately 0.33 to 0.22. This early reduction is not explained by the linear schedule alone and reflects the contribution of adaptive weighting. The mean effective weight subsequently decreases more gradually to approximately 0.08.

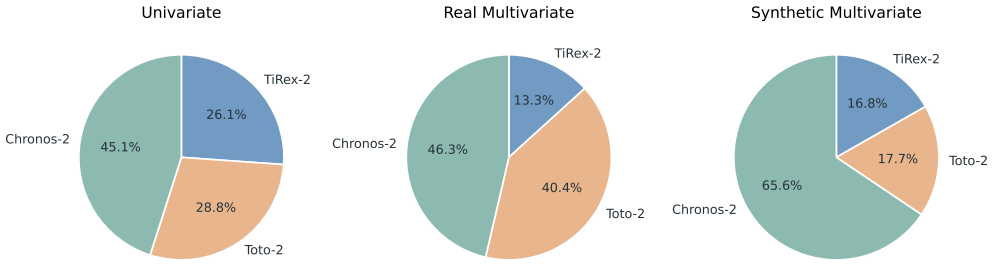


Figure 4: Fractions of training windows assigned to each teacher across data categories.

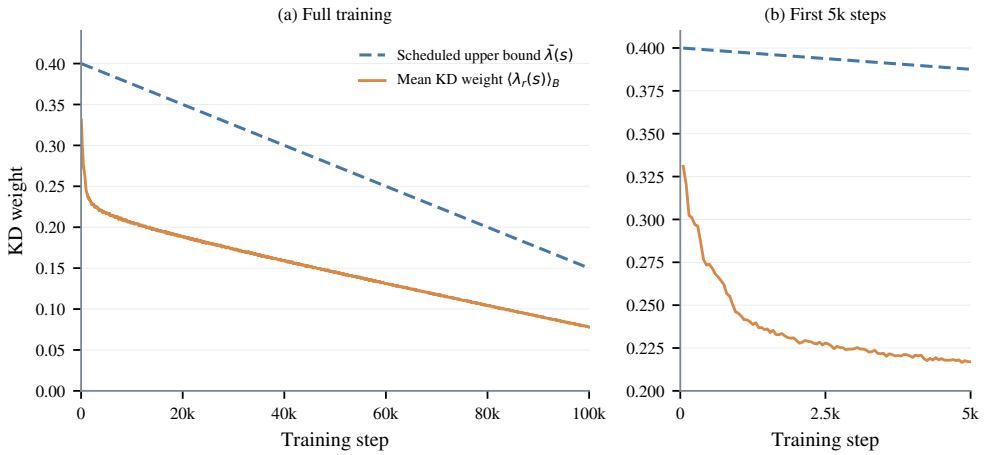


Figure 5: Scheduled upper bound $\bar{\lambda}(s)$ (dashed) and mean effective KD weight $\langle\lambda_r(s)\rangle_B$ (solid), averaged over target series with valid teacher forecasts. The right panel enlarges the first 5,000 steps.

Table 2: Ablations of sample-level teacher selection, OH-RKD, and SRAW. Selection denotes sample-level teacher selection. Lower is better for all metrics.

Variant	Selection	OH-RKD	SRAW	FEV-Bench		GIFT-Eval	
				SQL ↓	MASE ↓	MASE ↓	WQL ↓
GT-only	–	–	–	1.488574	1.790383	1.919592	0.241482
Single-teacher	No	Yes	Yes	1.426558	1.725608	1.708592	0.217923
Quantile-wise KD	Yes	No	Yes	1.401028	1.699547	1.814306	0.237744
Linear Weight	Yes	Yes	No	1.396952	1.697464	1.887329	0.253962
Full framework	Yes	Yes	Yes	**1.354216**	**1.643362**	**1.641016**	**0.215008**

5.3 Ablation Studies

To investigate the contribution of each component, we conduct controlled ablation experiments. We replace one component at a time while keeping the remaining training configuration unchanged. The results are reported in Table 2.

Specifically, the single-teacher variant replaces sample-level teacher selection with one fixed teacher, while retaining OH-RKD and SRAW. The quantile-wise KD variant replaces OH-RKD with independent alignment of corresponding quantiles, while retaining sample-level teacher selection and SRAW. The linear-weight variant replaces SRAW with a global linear schedule while retaining the other components.

Compared with the full framework, using a single fixed teacher increases FEV-Bench SQL from 1.354216 to 1.426558 and GIFT-Eval WQL from 0.215008 to 0.217923. Replacing OH-RKD with quantile-wise KD increases these metrics to 1.401028 and 0.237744, respectively. These results indicate that both selecting a teacher for each training window and modeling teacher-guided cross-quantile relationships contribute to the overall performance.

Replacing SRAW with a global linear weight schedule increases FEV-Bench SQL to 1.396952 and GIFT-Eval WQL to 0.253962. Compared with this variant, SRAW reduces SQL by 3.05% and WQL by 15.34%. The linear-weight variant improves over GT-only on FEV-Bench but performs worse on GIFT-Eval, whereas the full framework improves over GT-only on both benchmarks. This comparison evaluates the complete weighting strategies, including differences in their effective distillation weights.

6 Conclusion and Discussion

We presented a multi-teacher response-based distillation framework for probabilistic time-series forecasting. The framework combines sample-level teacher selection, OH-RKD for quantile alignment with teacher-guided cross-quantile constraints, and SRAW for adapting distillation strength according to the relative forecasting quality of the selected teacher and student. Experiments on FEV-Bench and GIFT-Eval show that the full framework improves over GT-only training and standard response-based KD on the reported aggregate metrics. It also achieves pairwise win rates above 50% against each evaluated baseline on both benchmarks. These results suggest that selecting teacher responses by training-window quality and adapting their influence can improve the transfer of probabilistic forecasting knowledge to a compact student.

Several limitations remain. First, due to time constraints, our teacher pool does not include TimesFM-3.0, a recent state-of-the-art time-series foundation model. Evaluating it as an additional teacher may further clarify how our framework performs with stronger and more diverse teacher pools. Second, although our sample-level teacher selection chooses the teacher with the lowest forecasting loss on each training window, more theoretically grounded selection criteria may yield better or more principled choices. Developing such criteria is an important direction for future work. Third, generating teacher forecasts before student pretraining incurs additional upfront inference cost. Future work could investigate more efficient forecast generation and caching strategies. This cost is incurred during training preparation; the teachers are not required at inference time.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

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.

Alexander Alexandrov, Konstantinos Benidis, Michael Bohlke-Schneider, Valentin Flunkert, Jan Gasthaus, Tim Januschowski, Danielle C Maddix, Syama Rangapuram, David Salinas, Jasper Schulz, et al. Gluonts: Probabilistic and neural time series modeling in python. *Journal of Machine Learning Research*, 21(116):1–6, 2020.

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.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Luis M Candanedo, Véronique Feldheim, and Dominique Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and buildings*, 140:81–97, 2017.

Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017.

Eli Cortez, Anand Bonde, Alexandre Muzio, Mark Russinovich, Marcus Fontoura, and Ricardo Bianchini. Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms. In *Proceedings of the 26th symposium on operating systems principles*, pp. 153–167, 2017.

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

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.

Rupasree Dey, Abdul Matin, Nathan Orwick, Yao Zhang, Shrideep Pallickara, and Sangmi Lee Pallickara. When to trust, how to distill: Multi-foundation model guidance for lightweight, robust scientific time series forecasting. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 10807–10818, 2026.

Patrick Emami, Abhijeet Sahu, and Peter Graf. Buildingsbench: A large-scale dataset of 900k buildings and benchmark for short-term load forecasting. *Advances in Neural Information Processing Systems*, 36:19823–19857, 2023.

Cheng Feng, Long Huang, and Denis Krompass. General time transformer: an encoder-only foundation model for zero-shot multivariate time series forecasting. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 3757–3761, 2024.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I Webb, Rob J Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. *arXiv preprint arXiv:2105.06643*, 2021.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the association for computational linguistics: ACL 2023*, pp. 8003–8017, 2023.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Ruihan Jin, Pengpeng Shao, Zhengqi Wen, Jinyang Wu, Mingkuan Feng, Shuo Yang, Chu Yuan Zhang, and Jianhua Tao. Exploring knowledge purification in multi-teacher knowledge distillation for llms. In *International Conference on Learning Representations*, 2026.

Emaad Khwaja, Chris Lettieri, Gerald Woo, Eden Belouadah, Marc Cenac, Guillaume Jarry, Enguerrand Paquin, Xunyi Zhao, Viktoriya Zhukov, Othmane Abou-Amal, et al. Toto 2.0: Time series forecasting enters the scaling era. *arXiv preprint arXiv:2605.20119*, 2026.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Yuqi Li, Kuiye Ding, Chuanguang Yang, Szu-Yu Chen, and Yingli Tian. Distilling time series foundation models for efficient forecasting. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4631–4635. IEEE, 2026.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6555–6565, 2024.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, et al. Mopd: Multi-teacher on-policy distillation for capability integration in llm post-training. *arXiv preprint arXiv:2606.30406*, 2026.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1): 54–74, 2020.

Clayton Miller, Anjukan Kathirgamanathan, Bianca Picchetti, Pandarasamy Arjunan, June Young Park, Zoltan Nagy, Paul Raftery, Brodie W Hobson, Zixiao Shi, and Forrest Meggers. The building data genome project 2, energy meter data from the ashrae great energy predictor iii competition. *Scientific data*, 7(1):368, 2020.

John A Miller, Mohammed Aldosari, Farah Saeed, Nasid Habib Barna, Subas Rana, I Budak Arpinar, and Ninghao Liu. A survey of deep learning and foundation models for time series forecasting. *arXiv preprint arXiv:2401.13912*, 2024.

Tianyu Peng and Jiajun Zhang. Enhancing knowledge distillation of large language models through efficient multi-modal distribution alignment. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2025.

Patrick Podest, Marco Pichler, Elias Bürger, Levente Zólyomi, Bernhard Voggenberger, Wilhelm Berghammer, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex-2: Generalizing tirex to multivariate data and streaming. *arXiv preprint arXiv:2607.01204*, 2026.

Stephan Rasp, Peter D Dueben, Sebastian Scher, Jonathan A Weyn, Soukayna Mouatadid, and Nils Thuerey. Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.

Charles Reiss, John Wilkes, and Joseph L Hellerstein. Google cluster-usage traces: format+ schema. *Google Inc., White Paper*, 1(1-14):83, 2011.

Abdulwahed Salam and Abdelaaziz El Hibaoui. Power consumption of tetouan city. UCI Machine Learning Repository, 2018.

Muhamad Risqi U Saputra, Pedro PB De Gusmao, Yasin Almalioglu, Andrew Markham, and Niki Trigoni. Distilling knowledge from a deep pose regressor network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 263–272, 2019.

Iman Sharafaldin, Arash Habibi Lashkari, Ali A Ghorbani, et al. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, 1(2018):108–116, 2018.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting. *arXiv preprint arXiv:2509.26468*, 2025.

Makoto Shing, Kou Misaki, Han Bao, Sho Yokoi, and Takuya Akiba. Taid: Temporally adaptive interpolated distillation for efficient knowledge transfer in language models. In *International Conference on Learning Representations*, 2025.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7059–7073, 2023.

Qian Sun, Yong-Ming Tian, Jia-Wei Huang, Cheng Feng, and Shao-Qun Zhang. Rmisc: A large-scale real-world multivariate corpus for time series foundation models. *arXiv preprint arXiv:2607.06504*, 2026.

Yijun Tian, Yikun Han, Xiusi Chen, Wei Wang, and Nitesh V Chawla. Beyond answers: Transferring reasoning capabilities to smaller llms using multi-teacher knowledge distillation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 251–260, 2025.

Artur Trindade. ElectricityLoadDiagrams20112014, 2015.

Frauke Wiese, Ingmar Schlecht, Wolf-Dieter Bunke, Clemens Gerbaulet, Lion Hirth, Martin Jahn, Friedrich Kunz, Casimir Lorenz, Jonathan Mühlenpfordt, Juliane Reimann, et al. Open power system data–frictionless data for electricity system modelling. *Applied Energy*, 236:401–409, 2019.

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

Junran Wu, Ke Xu, Xueyuan Chen, Shangzhe Li, and Jichang Zhao. Price graphs: Utilizing the structural information of financial time series for stock prediction. *Information Sciences*, 588: 405–424, 2022.

Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4498–4502. IEEE, 2022.

A SRAW HYPERPARAMETER SETTINGS

Table 3: Hyperparameters of Sample-wise Relative Adaptive Weighting (SRAW).

Hyperparameter	Value
λ_{start}	0.40
λ_{end}	0.15
$\lambda_{\min}$	0.02
τ	0.01
β	0.50

We use a linearly decaying global distillation schedule together with the sample-wise adaptive weighting mechanism introduced in Section 4.3. The hyperparameters used in all experiments are summarized in Table 3.

B BENCHMARK DETAILS

B.1 GIFT-EVAL

GIFT-Eval is a comprehensive benchmark designed for evaluating general-purpose time-series forecasting models under zero-shot settings. It contains 23 datasets with more than 144,000 time series spanning seven application domains and a wide range of sampling frequencies, variate dimensions, and forecasting horizons. The benchmark includes both univariate and multivariate forecasting tasks and supports probabilistic evaluation, making it well suited for assessing the generalization ability of time-series foundation models across heterogeneous forecasting scenarios (Aksu et al., 2024).

B.2 FEV-BENCH

FEV-Bench is a large-scale benchmark designed to provide a realistic and statistically robust evaluation of time-series forecasting models. It comprises 100 forecasting tasks across seven domains, including 46 tasks with covariates, thereby covering a broad range of practical forecasting settings. In addition to standard forecasting tasks, FEV-Bench emphasizes reproducible evaluation and aggregates model performance using complementary measures such as win rates and skill scores with bootstrapped confidence intervals (Shchur et al., 2025).

C TRAINING DETAILS

C.1 DATASET DETAILS

Our pretraining corpus consists of a large and carefully curated collection of publicly available time-series datasets covering diverse application domains and temporal characteristics. We prioritize datasets with few or no native missing values and exclude sources with severe quality issues, such as obvious measurement artifacts or highly irregular fluctuations indicative of unreliable data. The resulting corpus spans a wide range of sequence lengths, sampling frequencies, and temporal patterns, including different degrees of trend, seasonality, and variability.

In addition to univariate time series, the corpus contains both real-world multivariate datasets and synthetically constructed multivariate data. The former preserves naturally occurring cross-variable dependencies, while the latter further increases the diversity of multivariate structures encountered during training. Overall, the corpus is designed to provide broad coverage across domains, temporal resolutions, sequence lengths, and variable configurations.The following tables show the composition of univariate and real multivariate datasets.

Table 4: Publicly available univariate datasets used in the pretraining corpus and their corresponding statistics and sources.

Dataset	Domain	# Series	Avg. Length	Source
Alibaba Cluster Trace 2018	cloud computing	467,272	1,629	(Woo et al., 2024)
Azure VM 2017	cloud computing	478,416	5,552	(Cortez et al., 2017)
BDG2	buildings/energy	985	9,337	(Miller et al., 2020)
Borg Cluster Data 2011	cloud computing	1,003,702	3,748	(Reiss et al., 2011)
Buildings-900K	buildings/energy	1,795,256	8,761	(Woo et al., 2024)
Electricity 15min	energy	370	113,341	(Trindade, 2015)
FordA	automotive/sensor	4,921	500	(Dau et al., 2019)
Hog	motion/sensor	152	2,403	(Woo et al., 2024)
InlineSkate	motion/sensor	650	1,882	(Dau et al., 2019)
KernelSynth-1M	synthetic	1,000,000	1024	(Ansari et al., 2024)
Largest 2017	transportation	8,196	105,120	(Woo et al., 2024)
London Smart Meters	energy	2,351	30,594	(Godahewa et al., 2021)
M4	mixed	100,000	252	(Makridakis et al., 2020)
Mexico City Bikes	transportation	494	78,313	(Ansari et al., 2024)
Monash Electricity	energy	642	13,230	(Godahewa et al., 2021)
Monash KDD Cup 2018	environment	270	10,897	(Godahewa et al., 2021)
Monash Pedestrian Counts	environment	66	47,459	(Godahewa et al., 2021)
Monash london smart meters	energy	5,560	29,951	(Godahewa et al., 2021)
Monash Temperature Rain	weather	32,072	725	(Godahewa et al., 2021)
Wind Farms	energy	674	4,434	(Godahewa et al., 2021)
Solar	energy	10,332	56,940	(Ansari et al., 2024)
Taxi	transportation	4,856	1,108	(Alexandrov et al., 2020)
TSMixup-10M	mixed	10,000,000	1,090	(Ansari et al., 2025)
Uber TLC	transportation	524	2,262	(Alexandrov et al., 2020)
WeatherBench	weather	675,840	122,445	(Rasp et al., 2020)
Wiki Daily 100K	web	100,000	2,741	(Ansari et al., 2024)

Table 5: Publicly available real multivariate datasets used in the pretraining corpus and their corresponding statistics and sources.

Dataset	Domain	# Series	Avg. Length	Source
AMarket China	Finance	3,600	2,060	(Wu et al., 2022)
Appliance Energy	Energy	26	19,735	(Candanedo et al., 2017)
Buildings Bench Com	Energy	692,725	8,759	(Emami et al., 2023)
CSE-CIC-IDS2018	Industry	4,981,515	250	(Sharafaldin et al., 2018)
ERA5 Hourly Single Levels	Environment	574	752,711	(Hersbach et al., 2020)
Froth Flotation	Industry	13	3,278	(Sun et al., 2026)

Continued on next page

Table 5 – continued from previous page

Dataset	Domain	# Series	Avg. Length	Source
Gas Sensor	Environment	296	387,202	(Sun et al., 2026)
Global Climate Change	Environment	2,578	2,185	(Sun et al., 2026)
OilWell	Energy	9,353	26,144	(Sun et al., 2026)
OPSD	Energy	1,585	104,185	(Wiese et al., 2019)
Tetuan Power Consumption	Energy	8	52,416	(Salam & El Hibaoui, 2018)
Tushare Index Daily	Finance	6,889	3,448	(Sun et al., 2026)
Tushare Stock	Finance	96,291	1,601	(Sun et al., 2026)
WeatherBench 5-625deg	Environment	4,994,748	8,766	(Rasp et al., 2020)

C.2 TRAINING DATA CONSTRUCTION

we obtained sampled training windows based on the datasets described at previous section.

Dataset-level balancing. To reduce the imbalance caused by the substantially different sizes of the source datasets, we first estimate the maximum number of valid training samples that can be extracted from each dataset. We then manually assign a sampling budget to each dataset using a reference dataset as the balancing point. Datasets with fewer available samples than the reference are sampled up to their maximum feasible capacity, whereas larger datasets are downsampled proportionally. This prevents a small number of very large datasets from dominating the pretraining distribution.

Within-dataset sampling. For each dataset, its sampling budget is distributed across individual time series according to sequence length, allowing longer series to contribute more training examples. For a series of length L, forecast origins are sampled from the central 90% of the sequence, i.e.,

$$t \in [0.05L,\ 0.95L].$$

When N origins are sampled from the same series, we enforce a minimum separation of

$$\Delta t_{\min} = \frac{0.9L}{2N},$$

to reduce excessive overlap among sampled windows.

Context and horizon construction. For each forecast origin t, the context contains all available historical observations preceding t, truncated to the most recent 2048 time steps:

$$C = \min(t, 2048).$$

For datasets with known sampling frequency, the forecasting horizon is determined using a predefined frequency-to-horizon mapping. For datasets without frequency metadata, we adopt a Chronos-2-inspired random horizon scheme. When $C = 2048$, we sample

$$H_{2048} = 16r, \qquad r \sim \mathrm{Uniform}\{1, \ldots, 8\}.$$

For shorter contexts, the horizon is scaled proportionally with the available context length:

$$H = \max\left(1, \left\lfloor H_{2048}\frac{C}{2048}\right\rfloor\right).$$

Overall, this hierarchical sampling procedure balances data sources of substantially different scales while preserving diversity in sequence length, temporal position, and forecasting horizon.

Table 6: Sampling proportions and capacity utilization of the largest training data sources in the final pretraining corpus.

Dataset	Training Sample Proportion	Capacity Utilization
Univariate		
WeatherBench	15.23%	5.05%
Buildings-900K	13.08%	21.62%
TSMixup-10M	12.93%	29.37%
Real-world Multivariate		
WeatherBench 5.625deg	23.46%	8.54%
OilWell	19.35%	100.00%
ERA5 Hourly Single Levels	11.72%	100.00%

C.3 Student model architecture

The student is the Chronos-2 Small model, initialized from scratch. Its architecture is summarized in Table 7.

Table 7: Chronos-2 Small student model architecture.

Component	Configuration
Architecture	Chronos-2 encoder
Number of parameters	27,934,624 (27.9M)
Model dimension (d_{model})	512
Feed-forward dimension (d_{ff})	2,048
Key/value dimension (d_{kv})	64
Encoder layers	6
Attention heads	8
Activation function	ReLU
Dropout rate	0.1
Layer normalization epsilon	10^{-6}
Input patch size / stride	16 / 16
Output patch size	16
Maximum context length	2,048
Quantile levels	$0.1, 0.2, \ldots, 0.9$
Regression token	Enabled
Arcsinh transformation	Enabled

C.4 Experimental hyperparameters

We train the student using the selected teacher predictions and ground-truth targets. Unless otherwise specified, the SRAW experiments use the teacher-guided cross-quantile Huber loss. Table 8 summarizes the training configuration.

Table 8: Training and distillation hyperparameters.

Hyperparameter	Value
Teacher models	Tirex-2, Toto-2, and Chronos-2
Teacher selection	Sample-wise best-teacher selection
KD loss	teacher-guided cross-quantile Huber loss
Ground-truth loss	Quantile pinball loss
Quantile levels	$0.1, 0.2, \ldots, 0.9$
Weighting method	Sample-wise Relative Adaptive Weighting (SRAW)
GPUs	4
Per-GPU batch size	1,024
Global batch size	4,096
Gradient accumulation steps	1
Maximum optimizer steps	100,000
Optimizer	Fused AdamW
Initial learning rate	1×10^{-3}
Minimum learning rate	0
Learning-rate schedule	Linear warmup followed by cosine decay
Warmup ratio	0.03
Weight decay	0.01
Maximum gradient norm	1.0
Training precision	BF16
TF32	Enabled
Random seed	123
Context length	2,048
Missingness augmentation	Disabled
Evaluation during training	None

D Baseline Implementation Details

All baselines use the same student architecture, training data, initialization protocol, optimizer, and training budget as our method. They differ in the distillation objective or teacher-weighting strategy. For all response-based baselines, we use the quantile grid $\mathcal{Q} = \{0.1, 0.2, \ldots, 0.9\}$ and the quantile loss $\mathcal{L}_Q$ defined in Section 3.

GT-only. The student is trained using only the ground-truth quantile forecasting loss:

$$\mathcal{L}_{\mathrm{GT-only}} = \mathcal{L}_{\mathrm{GT}}. \tag{20}$$

Response-based KD. We select the teacher with the lowest ground-truth quantile loss for each training window, as described in Section 4.1, and align the student and selected teacher at corresponding quantile levels:

$$\mathcal{L}_{\mathrm{RespKD}} = \mathcal{L}_{\mathrm{GT}} + \lambda_{\mathrm{KD}} \frac{1}{|\mathcal{V}_B|} \sum_{i \in \mathcal{V}_B} \mathcal{L}_Q(\widetilde{T}_i, \widetilde{S}_i), \tag{21}$$

where $\mathcal{V}_B$ is the set of training windows with valid teacher forecasts, and λ_{KD} controls the distillation strength.

Attentive Imitation Loss (AIL). AIL adjusts the strength of teacher imitation according to the teacher's forecasting error on each training window. For each window i, we select the teacher with the lowest mean pinball loss among the available teachers:

$$j_i^* = \arg\min_j \mathcal{L}_Q(Y_i, T_i^{(j)}), \tag{22}$$

where the selection loss is computed over finite target values, quantile levels, and forecast steps.

For the selected teacher, we compute a per-window error after normalizing the target and teacher forecasts using statistics from the context and applying the $\mathrm{arcsinh}$ transformation:

$$E_i^T = \frac{1}{|\Omega_i|} \sum_{(q,h)\in\Omega_i} 2\,\rho_q\Big(\widetilde{Y}_{i,h} - \widetilde{T}_{i,q,h}^{(j_i^*)}\Big), \tag{23}$$

where $\rho_q(u) = \max(qu, (q-1)u)$ is the quantile pinball loss and Ω_i contains valid target, quantile, and horizon entries. We estimate the minimum and maximum of this error over a calibration subset of the training data and define $\eta = E_{\max}^T - E_{\min}^T$. The teacher confidence is

$$\Phi_i = \mathrm{clip}\left(1 - \frac{E_i^T}{\eta},\, 0,\, 1\right). \tag{24}$$

The imitation term uses an aligned, horizon-adaptive Huber loss between the student and selected-teacher quantile forecasts:

$$\mathcal{L}_{\mathrm{Huber},i} = \frac{1}{|\Omega_i|} \sum_{(q,h)\in\Omega_i} \mathrm{Huber}_{\delta_{i,h}}\Big(\widetilde{S}_{i,q,h} - \widetilde{T}_{i,q,h}^{(j_i^*)}\Big), \qquad \delta_{i,h} = 0.1\left(\widetilde{T}_{i,0.9,h}^{(j_i^*)} - \widetilde{T}_{i,0.1,h}^{(j_i^*)}\right)_+ + \epsilon. \tag{25}$$

The resulting training objective is

$$\mathcal{L}_{\mathrm{AIL}}(s) = \mathcal{L}_Q(\widetilde{Y}, \widetilde{S}) + \lambda(s) \frac{1}{|\mathcal{R}_B|} \sum_{i\in\mathcal{R}_B} \Phi_i\, \mathcal{L}_{\mathrm{Huber},i}, \tag{26}$$

where $\lambda(s)$ is linearly scheduled from 0.4 to 0.15 over training.

Confidence-Aware Multi-Teacher KD (CA-MKD). CA-MKD assigns sample-wise weights to teachers based on their losses against the ground truth, then aggregates their predictions before distillation (Zhang et al., 2022). The original method uses the cross-entropy loss for classification; we replace it with the quantile loss for probabilistic forecasting. For teacher j and training window i, define

$$E_i^{(j)} = \mathcal{L}_Q(\widetilde{Y}_i, \widetilde{T}_i^{(j)}). \tag{27}$$

Following the confidence-weighting form of CA-MKD, we compute

$$w_i^{(j)} = \frac{1}{J-1}\left(1 - \frac{\exp(E_i^{(j)})}{\sum_{k=1}^J \exp(E_i^{(k)})}\right), \qquad \sum_{j=1}^J w_i^{(j)} = 1, \tag{28}$$

where J is the number of teachers. The aggregated quantile forecast is

$$\widetilde{T}_i^{\mathrm{agg}} = \sum_{j=1}^J w_i^{(j)} \widetilde{T}_i^{(j)}. \tag{29}$$

We then distill the aggregated response using the corresponding-quantile response loss:

$$\mathcal{L}_{\text{CA-MKD}} = \mathcal{L}_{\mathrm{GT}} + \lambda_{\mathrm{CA}} \frac{1}{|\mathcal{R}_B|} \sum_{i\in\mathcal{R}_B} \mathcal{L}_Q(\widetilde{T}_i^{\mathrm{agg}}, \widetilde{S}_i), \tag{30}$$

where λ_{CA} controls the distillation strength. Since the intermediate-feature component of the original CA-MKD is not used in this response-based adaptation, our baseline applies its confidence-aware weighting to teacher forecasts only.