BI-LEVEL HETEROGENEOUS LEARNING FOR TIME SERIES FOUNDATION MODELS: A FEDERATED LEARNING APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

Heterogeneity in time series data is far more pronounced than in vision or language, manifesting at two coupled levels: *inter-domain* discrepancies in sensing modalities, sampling rates, and physical processes, and *intra-domain* variability such as latent sub-domains and temporal concept drift. We empirically show that this bi-level heterogeneity is the root cause of two failures of mixed-batch centralized pretraining of time series foundation models (TSFMs), namely representation collapse and gradient conflict, which existing federated learning (FL) methods only partially address since they assume each client is internally homogeneous. We propose FedTRL, a heterogeneity-aware FL pretraining method whose two components target the two levels respectively: a local objective that combines diffusion-based reconstruction with sub-domain adversarial regularization and prototype alignment to reduce intra-domain conflict without disturbing temporal modeling, and a domain-aware aggregation (DaG) on the server that re-weights clients by prototype discriminability and global semantic alignment to control inter-domain interference. Extensive experiments across point/probabilistic forecasting show FedTRL matches or surpasses centralized and federated TSFM baselines and degrades least under domain imbalance, drift, and unseen-domain transfer, turning bi-level heterogeneity from an obstacle into a controllable training signal.

1 INTRODUCTION

Time series forecasting plays a critical role in decision-making domains such as weather (Bi et al., 2023), energy (Wu et al., 2022), and urban computing (Zhang et al., 2024b). Foundation models have shifted research from task-specific architectures toward general-purpose forecasting models trained on diverse datasets (Goswami et al., 2024; Shi et al., 2024). However, existing time series foundation models (TSFMs) are typically built under centralized paradigms (Shi et al., 2024; Das et al., 2024; Woo et al., 2024), where heterogeneous datasets are merged into mixed batches for pretraining. As shown in **Fig. 1**, this mixing is far from benign: relative to federated alternatives (Chen et al., 2025), it produces both *representation collapse* (smallest inter-domain centroid distance, largest within-domain variance) and *strongest gradient conflicts* (lowest pairwise cosine similarity, highest gradient-norm variance). Beyond model quality, temporal data are inherently fragmented across organizations, making centralized collection impractical and aggravating cross-domain heterogeneity.

Concretely, these limitations manifest as a form of bi-level heterogeneity. The first level is inter-domain heterogeneity (Chen et al., 2025), where datasets from different sources exhibit covariate shifts and divergent temporal patterns due to variations in sensing modalities, sampling rates, and environments. This misalignment often causes centralized or mixed-batch training to overfit domain-specific signals and fail to capture globally consistent dynamics. The second level is intra-domain conflict, which arises within a single dataset or domain (Wang et al., 2024b). Temporal concept drift from reconfiguration or task evolution, together with latent sub-domains such as geographically distributed devices, leads to semantically similar but contextually divergent fragments. Unlike inter-domain heterogeneity, this conflict cannot be absorbed by client-level partitioning, since the conflicting fragments coexist within each client. For example, a weather agency acting as one client holds records from stations whose diurnal cycles differ in amplitude and phase and whose noise characteristics shift after recalibration (intra-domain), while a power operator whose load series

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.

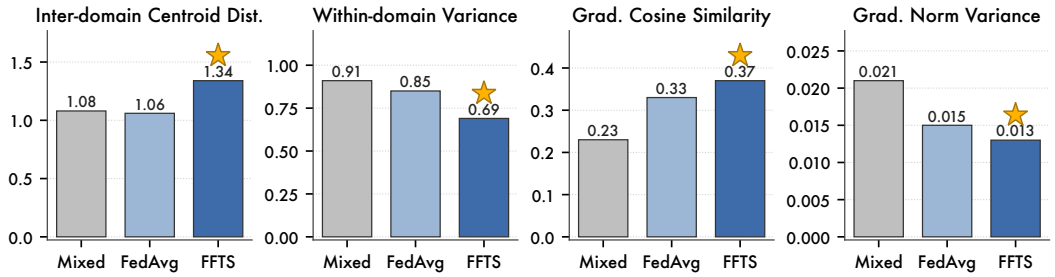


Figure 1: Bi-level heterogeneity. For centralized mixing (*Mixed*), FedAvg, and FFTS, inter-domain centroid distance and gradient cosine similarity stay low (↑ better), while within-domain variance and gradient-norm variance stay high (↓ better). Time-MoE-300B sub-domains, details in App. E.9.

follow sharp weekday peaks at a different sampling rate differs from it at the inter-domain level. Naive training over such mixtures produces incoherent embeddings that are neither specialized nor aligned within sub-domains. The bi-level heterogeneity undermines the learning of domain-invariant and temporally coherent patterns, thereby limiting the generalization of TSFMs.

A more flexible learning paradigm is required to handle bi-level heterogeneity in TSFMs training, one that can integrate knowledge across diverse domains while limiting cross-domain interference. Federated learning (FL) (McMahan et al., 2017) naturally offers such a strategy: by enabling decentralized training over distributed datasets, it allows domains to collaborate without centralizing raw data. Yet, existing FL methods primarily address inter-domain heterogeneity through personalization (Tan et al., 2022) or domain-aware adaptations (Yang et al., 2023), while typically assuming that each client's local data is homogeneous and stationary. While such assumptions may hold in static modalities like CV and NLP, they break down in time series domains where a single client can contain multiple latent sub-domains or undergo temporal concept drift. These intra-domain conflicts blur local representations, leaving them neither domain-specialized nor globally aligned. Recent work (Chen et al., 2025) explores task-agnostic FL for time series, but it largely targets inter-domain misalignment while overlooking intra-domain conflicts and evolving semantics. The open question is how to design a FL paradigm that can simultaneously resolve bi-level heterogeneity, and in doing so, provide a scalable pathway for building robust TSFMs.

To this end, we propose FedTRL, a heterogeneity-aware FL method for TSFM pretraining whose two components target the two levels of heterogeneity respectively. Locally, each client combines diffusion-based reconstruction with sub-domain adversarial regularization and prototype alignment. The adversarial signal acts only on the prototype-classification path so that temporal dynamics in the encoder are left intact, while only the sub-domain classifiability of prototypes is suppressed. Globally, a domain-aware aggregation (DaG) re-weights clients on the server by prototype discriminability and global semantic alignment, controlling inter-domain interference without diluting clients that carry clear, transferable semantic signatures. As shown in **Fig. 2**, FedTRL closes/surpasses centralized TSFMs on zero-shot point (TSLib (Wu et al., 2022)) and probabilistic forecasting (GIFT-Eval (Aksu et al., 2024)), and its advantage amplifies under domain imbalance, drift, and unseen-domain transfer. Our contributions:

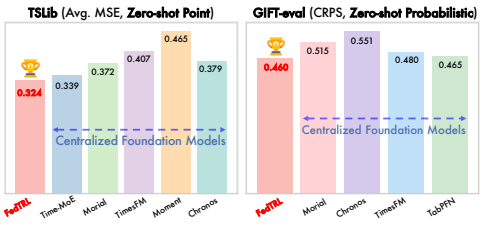


Figure 2: FedTRL outperforms centralized TSFMs in *zero-shot point* and *probabilistic forecasting*. Details are provided in Sec. 4

- We empirically isolate two failure modes of mixed-batch centralized TSFM pretraining, namely representation collapse and gradient conflict, and tie them to bi-level heterogeneity, motivating the treatment of TSFM pretraining as a heterogeneity-aware problem.
- We propose a fine-grained joint optimization and aggregation that combines sub-domain adversarial regularization with prototype alignment and domain-aware aggregation, enforcing semantically consistent representations while adaptively weighting client updates.
- Extensive experiments across representation learning, full/zero-shot point forecasting, and zero-shot probabilistic forecasting benchmarks show that FedTRL effectively scales FL to train TSFMs, achieving performance competitive with or superior to centralized approaches.

2 Related Work

Unsupervised Time Series Representation Learning. Unsupervised representation learning extracts general temporal patterns from unlabeled sequences (Zhang et al., 2024a), forming the foundation of TSFMs (*see Appendix A*), in contrast to task-specific supervised methods (Wu et al., 2022). Existing works mainly fall into masked reconstruction and contrastive learning. Masked reconstruction predicts missing inputs, with point-wise strategies such as TST (Zerveas et al., 2021), TimeMAE (Cheng et al., 2023), and SimMTM (Dong et al., 2023), and patch-wise variants like PatchTST (Nie et al., 2022) and CrossTimeNet (Cheng et al., 2025) that better capture local structures. Contrastive learning instead distinguishes positives from negatives, exemplified by TS-TCC (Eldele et al., 2021), which exploits temporal smoothness, and CoST (Woo et al., 2022), which leverages time–frequency augmentations. While effective on single datasets, these approaches face two issues when scaled to TSFMs: *(i) dependence on dataset-specific fine-tuning* and *(ii) conflicting embeddings across heterogeneous domains*. This highlights the need for more flexible training paradigms that unify representation learning across domains, where FL (McMahan et al., 2017) offers a strategy for collaborative pretraining on decentralized heterogeneous data.

Federated Learning in Time Series. Recent studies have extended FL to time series, exploring personalized cross-domain adaptation (Liu et al., 2024a; Soi et al., 2024) and task-agnostic modeling across domains (Chen et al., 2025; 2023; Sun et al., 2024). FeDaL (Chen et al., 2026) debiases dataset-level heterogeneity locally and corrects the server aggregate, but keeps aggregation weights proportional to client sample size. However, these methods often assume that local data within each client is homogeneous and domain-stable, an assumption rarely satisfied in practice. Most of them primarily target inter-domain heterogeneity across clients while overlooking intra-domain variability within individual clients. In reality, a single client may contain multiple latent sub-domains with divergent patterns, and cross-client distributions often remain misaligned despite similar marginal statistics. Such bi-level heterogeneity disrupts representation learning, yielding incoherent local embeddings and misaligned global models that struggle to generalize across domains and tasks. We propose FedTRL, a heterogeneity-aware FL pretraining method that tackles both inter-domain and intra-domain heterogeneity, enabling robust and scalable *pretraining of TSFMs from scratch*.

3 Methodology

Problem Definition. We consider an FL system with a server and K clients (domains), where each client k holds a local dataset $\mathcal{D}_k$ that may include data from one/multiple sub-domains. These datasets are heterogeneous due to differences in sampling rates, temporal dynamics, and domain-specific patterns. The objective is to collaboratively learn a unified representation that generalizes across domains and supports diverse forecasting tasks without supervision. The global objective is:

$$\min_{\theta} F(\theta) := \sum\nolimits_{k=1}^{K} \frac{n_k}{n} F_k(\theta_k; \mathcal{D}_k), \tag{1}$$

where F_k is the local objective on client (domain) k, n_k is its sample size, and $n = \sum_{k=1}^{K} n_k$. However, this standard objective fails to yield generalizable patterns because of two factors. (i) *Inter-domain heterogeneity*, where distinct distributions across domains lead to divergent local optima θ_k^*, making naive aggregation ineffective. (ii) *Intra-domain conflicts*, where a single domain hosts multiple latent sub-domains, and contextual shifts create inconsistent or hybrid embeddings.

Overview. We propose FedTRL (**Fig. 3**), a heterogeneity-aware FL method that targets the two levels of heterogeneity with complementary signals. FedTRL integrates: **(i)** a local objective that combines diffusion-based reconstruction with prototype-level sub-domain adversarial regularization and prototype alignment to control intra-domain conflict while leaving the encoder's temporal modeling capacity intact, and **(ii)** a domain-aware aggregation (DaG) on the server that re-weights clients by prototype *discriminability* and global semantic alignment, so that clients with cleaner, more reliable prototypes contribute more to the global encoder.

Definitions. We define: (i) *Sub-domain invariance* of a representation $\mathbf{z} = E(\mathbf{x})$ is defined as the smallness of the $\mathcal{H}$-divergence between latent distributions across sub-domains, $d_{\mathcal{H}}(\mathcal{D}_i^Z, \mathcal{D}_j^Z) =$

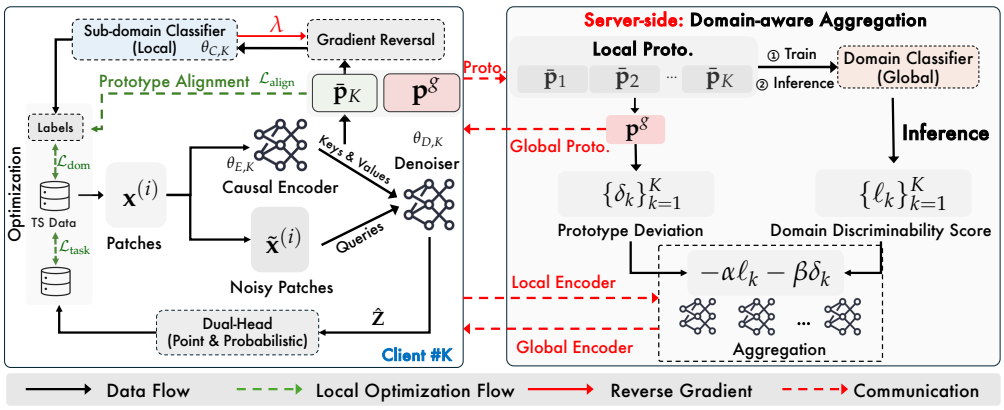


Figure 3: Structure of FedTRL (single client example). Each client performs unsupervised diffusion-based reconstruction and uploads only the encoder θ_E and prototypes $\bar{\mathbf{p}}$ per round, after which the server applies domain-aware aggregation (DaG) to produce a unified encoder.

$2\sup_{h\in\mathcal{H}}\left|\Pr_{\mathbf{z}\sim\mathcal{D}_i^z}[h(\mathbf{z})=1]-\Pr_{\mathbf{z}\sim\mathcal{D}_j^z}[h(\mathbf{z})=1]\right|$ for a hypothesis class $\mathcal{H}$ of binary discriminators. This is a constraint on the *classifiability* of representations in $\mathcal{H}$ rather than a constraint on downstream temporal predictions. (ii) *Globally consistent dynamics* are defined as the 2-Wasserstein barycenter of local prototype distributions, i.e., the distribution $\mathcal{D}^g$ minimizing $\sum_k w_k W_2^2(\mathcal{D}_k^p, \mathcal{D}^g)$ for non-negative weights w_k summing to one. Formal statements and proofs are deferred to App. D.

3.1 Local Updating on Clients

To address heterogeneity at different levels, each client optimizes a unified objective: (i) reconstruction for temporal semantics and (ii) adversarial regularization for mitigating intra-domain conflicts, while (iii) prototype alignment is designed to promote inter-domain consistency. We first decompose a multivariate time series $\mathbf{X}\in\mathbb{R}^{T\times C}$ of length T into univariate series $\mathbf{x}\in\mathbb{R}^T$ along channels. Each univariate series is first stationarized using a non-parametric two-stage instance normalization (Liu et al., 2022) to mitigate temporal distribution shifts and anomalous ranges, thereby improving robustness across latent sub-domains. We then divide each sequence into non-overlapping patches of length P, yielding $N=\lfloor T/P\rfloor$ patches $\mathcal{P}=\{\mathbf{x}^{(i)}\}_{i=1}^N$. This preserves local temporal structure while reducing computation for long sequences. Given these patches, we adopt diffusion reconstruction (Wang et al., 2025a) to perform unsupervised representation learning. Specifically, we apply a forward diffusion process $\mathcal{D}_{\text{fwd}}$ to obtain its noisy counterpart:

$$\tilde{\mathbf{x}}^{(i)}=\sqrt{\bar{\alpha}_t}\,\mathbf{x}^{(i)}+\sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon},\boldsymbol{\epsilon}\sim\mathcal{N}(0,I),\ t\sim\mathcal{U}(0,T_d) \tag{2}$$

Unlike contrastive (Woo et al., 2022) or masked reconstruction methods (Chen et al., 2025) that implicitly assume data homogeneity, our approach prioritizes intrinsic temporal dynamics over domain-specific artifacts by jointly modeling point-, patch-, and sequence-level patterns. All clean patches are embedded and stacked into a token sequence: $\mathbf{E}\in\mathbb{R}^{N\times d}$, which is processed by a causal Transformer encoder θ_E: $\mathbf{H}=\theta_E(\mathbf{E})\in\mathbb{R}^{N\times d}$. In parallel, noisy patches $\tilde{\mathbf{x}}^{(i)}$ are embedded as query tokens $\tilde{\mathbf{E}}\in\mathbb{R}^{N\times d}$. A denoising decoder θ_D reconstructs latent signals via cross-attention, with queries $\tilde{\mathbf{E}}\mathbf{W}_q$ attending to encoder outputs $\mathbf{H}\mathbf{W}_k$, $\mathbf{H}\mathbf{W}_v$: $\hat{\mathbf{Z}}=\theta_D(\tilde{\mathbf{E}}\mathbf{W}_q,\ \mathbf{H}\mathbf{W}_k,\ \mathbf{H}\mathbf{W}_v)$.

Sub-domain Adversarial and Prototype-Aligned Regularization. To mitigate conflicts among semantically related but contextually divergent sub-domains, we adopt sub-domain adversarial learning (Ganin et al., 2016). A sub-domain classifier D_{θ_C} is trained on each client to predict the sequence-level sub-domain label y^{dom}, where each time series (and all patches derived from it) shares the same sub-domain category. Labels come from metadata already present in the data (e.g., sub-dataset, station, or sensor identifiers), and k-means pseudo-labels serve as a substitute when metadata is missing (App. B.3). Given encoder outputs $\mathbf{H}$, we summarize the client's representation distribution by mean pooling: $\bar{\mathbf{p}}=\text{Mean}(\mathbf{H})$. Here $\bar{\mathbf{p}}$ serves as a *distribution-level centroid* for cross-client semantic alignment, while per-step temporal dynamics are encoded in $\mathbf{H}$ by the causal Transformer θ_E and remain independent of how the prototype is summarized. The classifier objective

is

$$\mathcal{L}_{\text{dom}} = \frac{1}{N}\sum_{i=1}^{N} \text{CE}\big(D_{\theta_C}(\text{GRL}(\bar{\mathbf{p}}^{(i)})),\ y_i^{\text{dom}}\big), \tag{3}$$

where GRL (Gradient Reverse Layer) acts as identity in the forward pass and scales gradients by $-\lambda$ in the backward pass. The GRL gradient is routed through the prototype-classification path only, so the diffusion reconstruction loss and the per-step prediction error are not affected by $\mathcal{L}_{\text{dom}}$, and the regularizer acts on the sub-domain classifiability of the prototype rather than on temporal patterns at the patch level. The coefficient λ controls the strength of this prototype-level regularization. Beyond local adversarial regularization, we facilitate cross-domain knowledge transfer via lightweight prototype-based alignment. Rather than constraining full model parameters, which are entangled with domain-specific dynamics, each client aligns its local prototype $\bar{\mathbf{p}}$ with the global prototype $\mathbf{p}^g$ as:

$$\mathcal{L}_{\text{align}} = ||\bar{\mathbf{p}} - \mathbf{p}^g||_2^2. \tag{4}$$

This pulls local representation distributions toward a shared centroid while preserving per-domain flexibility, yielding a semantically consistent prototype space that supports stable aggregation.

Point and Probabilistic Forecasting Support. Each latent token $\hat{\mathbf{z}}^{(i)} \in \hat{\mathbf{Z}}$ is projected back to the patch space through an MLP, yielding reconstructed patches $\hat{\mathbf{x}}^{(i)}$. On top of these shared representations, we support both point and probabilistic forecasting within a unified dual-head architecture. The deterministic head outputs point predictions $\hat{y}_i$, optimized by mean square error (MSE), while the probabilistic head parameterizes a Student-t distribution with mean μ_i, scale σ_i and degrees of freedom ν (fixed $\nu = 5$ empirically) encouraging the model to capture heavy-tailed uncertainties via negative log-likelihood (NLL). Formally, the joint training objective is:

$$\mathcal{L}_{\text{task}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \beta \cdot \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}\log\big(\pi(\nu-2)\sigma_i^2\big) + \frac{\nu+1}{2}\log\left(1 + \frac{(y_i-\mu_i)^2}{(\nu-2)\sigma_i^2}\right)\right], \tag{5}$$

where y_i is the ground-truth value of the forecasting target derived from the original time series $\mathbf{x}$, and β balances the deterministic and probabilistic objectives. To stabilize optimization under federated heterogeneity, we adopt a warm-up schedule: only the deterministic head is trained in the initial R_{warm} rounds, after which the probabilistic head is activated by annealing $\beta = 0$ to 1. This training design ensures accurate point forecasts first, followed by well-calibrated probabilistic predictions.

Local Optimization Objective. Each client jointly optimizes the three components introduced above. The reconstruction objective ensures temporal fidelity, the adversarial loss suppresses subdomain-specific bias, and the prototype alignment term promotes semantic consistency across clients. The overall local objective can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{dom}} \cdot \mathcal{L}_{\text{dom}} + \lambda_{\text{align}} \cdot \mathcal{L}_{\text{align}}, \tag{6}$$

where λ_{dom} and λ_{align} are balancing coefficients controlling the contribution of subdomain adversarial regularization and prototype alignment, respectively. Only the encoder θ_E and the local prototypes $\bar{\mathbf{p}}$ are shared with the server, as the goal is to collaboratively learn a domain-generalizable encoder. The decoder θ_D and sub-domain classifier θ_C are retained locally as auxiliary components to facilitate denoising and mitigate domain-specific bias. Their exclusion from aggregation prevents local semantics from interfering with the global encoder, enhancing generalization across domains.

3.2 Domain-aware Aggregation on Server

Naive FL aggregation (McMahan et al., 2017) treats all clients uniformly and dilutes the signal of clients whose representations are clean and informative when others are noisy or drifted. We propose Domain-aware Aggregation (DaG), which combines two complementary signals, prototype *discriminability* and prototype *alignment*, to weight the encoder update. The server receives local encoders $\{\theta_{E,k}\}_{k=1}^K$ and prototypes $\{\bar{\mathbf{p}}_k\}_{k=1}^K$, and updates a global prototype $\mathbf{p}^g$ via FedAvg. We then train a global domain classifier $f_d : \mathbb{R}^d \to \mathbb{R}^K$ over the local prototypes with client indices as labels and fix it for inference. For each $\bar{\mathbf{p}}_k$, we define the **Domain Discriminability Score**:

$$\ell_k = \mathcal{L}_{\text{CE}}\big(f_d(\bar{\mathbf{p}}_k),\ k\big), \tag{7}$$

which quantifies how confidently $\bar{\mathbf{p}}_k$ is recognized as belonging to its source client. A *low* ℓ_k indicates that $\bar{\mathbf{p}}_k$ has a clear semantic signature (a clean, internally-consistent summary of client

k's representation distribution) and is therefore a reliable contributor to global aggregation, while a *high* ℓ_k indicates an ambiguous prototype whose encoder update is less trustworthy. ℓ_k is thus a reliability signal rather than a transferability claim, and it is never used alone. We complement this with a continuous prototype-deviation term $\delta_k = \|\bar{\mathbf{p}}_k - \mathbf{p}^g\|_2^2$ (Eq. 4) that measures global semantic drift. The two signals are combined into a single aggregation score as follows:

$$s_k = -\alpha\,\ell_k - \beta\,\delta_k, \qquad \alpha, \beta > 0, \tag{8}$$

so that clients with clearer prototype signatures (low ℓ_k) and tighter alignment to the global state (low δ_k) receive higher softmax weight, while ambiguous or drifted clients are diluted. Final aggregation is obtained via softmax weighting, which can be formulated as:

$$\theta_E^g = \sum\nolimits_{k=1}^{K} \mathrm{Softmax}(s_k) \cdot \theta_{E,k}. \tag{9}$$

By weighting clients according to prototype discriminability and global alignment, DaG emphasizes reliable, globally consistent representations while downweighting ambiguous or drifted clients, improving cross-domain generalization. The resulting weights uniquely minimize an entropy-regularized reliability objective, preserve a positive weight for every client, and admit a convergence guarantee for the induced time-varying objective (App. D). The global prototype $\mathbf{p}^g$ and updated model θ_E^g are then broadcast for the next round. The full workflow is in **Algorithm 1**, with code available at Link.

4 Main Results

We evaluate FedTRL across forecasting settings (**Appendix C**). First, we compare it with unsupervised representation learning methods on in-domain forecasting (Sec. 4.1). We assess its scaling to large TSFMs (Sec. 4.2), with full-shot adaptation and zero-shot generalization (Sec. 4.2.1, Sec. 4.2.2) on standard benchmarks and real-world weather datasets. Finally, we conduct ablation, hyperparameter, and scalability studies (Sec. 4.3). FedTRL shows FL pretraining scales to large TSFMs, achieving competitive point/probabilistic forecasting performance while often surpassing centralized models.

4.1 In-domain Point Forecasting

Setup. We conduct experiments on eight benchmark datasets from TSLib (Wu et al., 2022). Each dataset is treated as an independent client during federated training (without probabilistic head). For fine-tuning, we adapt the pretrained model to the corresponding target datasets. The look-back window is fixed at 512, and forecasting performance is evaluated across four horizons {96,192,336,720} using MSE and MAE. Model architectures and training configurations are summarized in **Table 10**.

Table 1: In-domain forecasting results (averaged over horizons $\{96, 192, 336, 720\}$). **Bold**: best, Underline: second best. †: FedAvg aggregation protocol. Full results in **Table 13**.

Methods	Ours		Federated Self-supervised†								Federated FMs				Centralized	
	FedTRL		SimMTM		PatchTST		TimeMAE		CoST		FFTS		FedAvg		All Mixed	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	**0.448**	**0.472**	0.495	0.512	0.478	0.495	0.502	0.518	0.521	0.540	0.463	0.485	0.476	0.495	0.458	0.488
ETTh2	**0.382**	**0.414**	0.442	0.468	0.410	0.432	0.455	0.474	0.471	0.490	0.395	0.419	0.410	0.433	0.399	0.424
ETTm1	0.375	0.402	0.415	0.439	0.386	0.408	0.421	0.444	0.436	0.455	0.380	0.406	0.399	0.406	**0.372**	**0.398**
ETTm2	**0.278**	**0.330**	0.324	0.352	0.296	0.340	0.330	0.360	0.345	0.369	0.290	0.338	0.299	0.341	0.286	0.336
Electricity	**0.181**	**0.272**	0.216	0.306	0.192	0.288	0.225	0.315	0.242	0.329	0.187	0.280	0.198	0.293	0.189	0.285
Traffic	0.426	0.288	0.472	0.315	0.438	0.295	0.485	0.322	0.498	0.335	**0.420**	**0.285**	0.434	0.294	0.428	0.289
Weather	**0.241**	**0.282**	0.285	0.310	0.260	0.295	0.295	0.312	0.305	0.322	0.252	0.290	0.264	0.299	0.254	0.292
Exchange	**0.375**	**0.423**	0.460	0.480	0.405	0.445	0.468	0.490	0.472	0.492	0.390	0.440	0.403	0.444	0.390	0.438
1st Count	**12**		0		0		0		0		2		0		2	

Results. We compare FedTRL with three categories of baselines: (i) Masked reconstruction-based methods, including SimMTM (Dong et al., 2023) and PatchTST (Nie et al., 2022), (ii) Contrastive Learning-based methods, including TS-TCC (Eldele et al., 2021) and CoST (Woo et al., 2022), and (iii) FL methods, including FFTS (Chen et al., 2025) and FedAvg (McMahan et al., 2017). Each client perform TimeDART (Wang et al., 2025a) in FedAvg. Across eight datasets, FedTRL achieves consistent improvements over both federated and centralized baselines. It not only surpasses existing federated pretraining approaches by clear margins, but also outperforms centralized training on several datasets, showing that properly aligned and aggregated FL can exceed centralized solutions. Its notable gains over FedAvg highlight robustness under heterogeneous clients.

4.2 SCALING TO TIME SERIES FOUNDATION MODELS

FedTRL scales to large TSFMs by federated pretraining across heterogeneous domains. We pretrain on Time-MoE-300B (Shi et al., 2024), which contains over 300 billion time points from nine domains, treating each domain as an independent client. All evaluation datasets are excluded from pretraining to prevent data leakage. Model and training configurations are provided in **Table 10**.

4.2.1 FULL/ZERO-SHOT POINT FORECASTING

We evaluate FedTRL on widely used forecasting benchmarks, using TSLib (Wu et al., 2022) for point forecasting and excluding ECL to avoid data leakage. Beyond these standard benchmarks, we introduce RW-Bench, a collection of real-world weather datasets from 15 geographically diverse regions. RW-Bench preserves noisy observations rather than relying on curated academic data, and covers 10 variables at hourly resolution, providing a challenging testbed for zero-shot forecasting under realistic conditions. Dataset statistics are provided in **Appendix C.5**.

Table 2: Full-shot point forecasting, averaged over horizons $\{96, 192, 336, 720\}$. FedTRL is adapted to each target dataset for only one epoch. **Bold**: the best, Underline: the second best.

Models	**FedTRL**		TimeMixer		TimeXer		iTransformer		PatchTST		TimesNet		DLinear		Autoformer		Stationary		LightTS	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	**0.372**	**0.399**	0.448	0.443	0.491	0.489	0.474	0.476	0.413	0.430	0.458	0.450	0.422	0.437	0.496	0.487	0.570	0.537	0.491	0.479
ETTh2	**0.330**	0.383	0.364	0.394	0.356	0.398	0.372	0.411	**0.330**	**0.379**	0.414	0.427	0.431	0.446	0.450	0.459	0.526	0.516	0.602	0.543
ETTm1	**0.316**	**0.369**	0.381	0.395	0.381	0.402	0.368	0.396	0.351	0.380	0.400	0.406	0.357	0.378	0.588	0.517	0.481	0.456	0.435	0.437
ETTm2	0.257	**0.313**	0.275	0.323	0.275	0.329	0.271	0.331	**0.255**	0.315	0.291	0.333	0.267	0.333	0.327	0.371	0.306	0.347	0.409	0.436
Electricity	**0.143**	**0.247**	0.175	0.272	0.171	0.273	0.161	0.256	0.161	0.252	0.192	0.295	0.188	0.295	0.227	0.338	0.193	0.296	0.229	0.329
Traffic	**0.355**	**0.256**	0.405	0.284	0.419	0.302	0.385	0.275	0.390	0.263	0.620	0.336	0.433	0.295	0.628	0.379	0.624	0.340	0.622	0.392
Weather	**0.214**	**0.252**	0.241	0.272	0.227	0.268	0.235	0.274	0.255	0.264	0.259	0.287	0.248	0.300	0.338	0.382	0.288	0.314	0.261	0.312
Exchange	**0.372**	**0.390**	0.428	0.445	0.474	0.466	0.442	0.457	0.400	0.431	0.674	0.591	0.556	0.542	0.762	0.686	0.954	0.669	0.704	0.616
1st Count	**12**		0		0		0		3		0		0		0		0		0	

Full-shot Adaption. We compare FedTRL against several advanced deep time series forecasting models, including TimeMixer (Wang et al., 2024a), TimeXer (Wang et al., 2024c), PatchTST (Nie et al., 2022), TimesNet (Wu et al., 2022), DLinear (Zeng et al., 2023), iTransformer (Liu et al., 2023), Autoformer (Wu et al., 2021), Non-Stationary Transformer (Liu et al., 2022), and LightTS (Zhang et al., 2022), as shown in **Table 2**. FedTRL-trained model outperforms all baselines, showing notable advantages in both complex and simple domains despite requiring only 1 adaptation epoch.

Zero-shot Inference. We evaluate the FedTRL-trained TSFM on TSLib and RW-Bench following (Shi et al., 2024). As shown in **Table 3**, FedTRL achieves state-of-the-art zero-shot performance across 20 datasets, with particularly strong gains on real-world weather data characterized by greater temporal variability and distribution shifts. FedTRL outperforms all three Time-MoE scales despite using fewer parameters than Time-MoE$_{\text{ultra}}$, and also surpasses TimesFM and MOMENT, which require fine-tuning for variable-length adaptation (**Table 18**). These results demonstrate the effectiveness of federated pretraining for zero-shot TSFM generalization. Broader comparisons with LightGTS (Wang et al., 2025b), VisionTS (Chen et al., 2024), Sundial (Liu et al., 2025b), Chronos-2 (Ansari et al., 2025), Moirai-MoE (Liu et al., 2025a), and FeDaL (Chen et al., 2026) are reported in Table 16, with fine-tuned MOMENT, Timer, SymTime, and GTM included in App. E.6.

Table 3: Zero-shot point forecasting, averaged over horizons $\{96, 192, 336, 720\}$ with look-backs $\{512, 1024, 2048, 3072\}$. **Bold**: the best, Underline: the second best, "–": not evaluable. TSLib results from (Shi et al., 2024), full RW-Bench results in **Table 15**.

Models	**FedTRL**		**Time-MoE$_b$**		**Time-MoE$_l$**		**Time-MoE$_u$**		**Moirai$_s$**		**Moirai$_b$**		**Moirai$_l$**		**TimesFM**		**Moment**		**Chronos$_s$**		**Chronos$_b$**		**Chronos$_l$**	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.399	0.420	0.400	0.424	**0.394**	**0.419**	0.412	0.426	0.428	0.427	0.417	**0.419**	0.480	0.439	0.473	0.443	0.683	0.566	0.545	0.472	0.591	0.468	0.588	0.466
ETTh2	**0.350**	0.380	0.366	0.404	0.405	0.415	0.371	0.399	0.361	0.384	0.362	0.382	0.367	**0.377**	0.392	0.406	0.361	0.409	0.424	0.430	0.405	0.410	0.455	0.427
ETTm1	**0.349**	**0.383**	0.394	0.415	0.376	0.405	0.356	0.391	0.436	0.410	0.406	0.385	0.422	0.391	0.433	0.418	0.670	0.536	0.640	0.499	0.645	0.500	0.555	0.465
ETTm2	**0.282**	**0.330**	0.317	0.365	0.316	0.361	0.288	0.344	0.307	0.347	0.311	0.337	0.329	0.343	0.328	0.346	0.316	0.365	0.349	0.380	0.310	0.350	0.295	0.338
Weather	**0.238**	0.279	0.265	0.297	0.270	0.300	0.270	0.300	0.275	0.286	0.287	0.281	0.264	**0.273**	–	–	0.294	0.326	0.300	0.318	0.292	0.315	0.279	0.306
RW-Bench	**0.599**	**0.519**	0.621	0.536	0.608	0.525	–	–	1.328	0.627	1.029	0.598	1.244	0.606	0.991	0.747	1.091	0.756	1.097	0.683	1.156	0.697	1.117	0.683
1st Count	**8**		1		2		1		0		1		2		0		1		0		1		0	

4.2.2 ZERO-SHOT PROBABILISTIC FORECASTING

Beyond point forecasting, probabilistic forecasting is important to real-world applications with uncertainty. We experiment on GIFT-Eval (Aksu et al., 2024) and FEV leaderboard (Ansari et al.,

2024), following their official evaluation suite and assessing point (MASE) and probabilistic (CPRS and WQL) metrics. Note that all evaluated datasets are excluded from the pre-training dataset.

GIFT-eval. Results are shown in **Table 4**. This benchmark covers 23 datasets and 15 baselines, including statistical models, task-specific methods (DeepAR (Salinas et al., 2020), TiDE (Das et al., 2023a), N-BEATS (Oreshkin et al., 2020), PatchTST (Nie et al., 2022), iTransformer (Liu et al., 2023)), time series FMs (TimesFM (Das et al., 2024), TabPFN (Hoo et al., 2024), Chronos (Ansari et al., 2024)), and federated FMs (FedAvg (McMahan et al., 2017), FFTS (Chen et al., 2025)). Across all unseen datasets, FedTRL achieves the best MASE/CRPS, and surpasses most pretrained foundation models, highlighting its effectiveness for large-scale cross-domain pretraining.

Table 4: Zero-shot results on GIFT-eval with 100 generated series (Woo et al., 2024). Lower is better, and rank is over 97 configurations. Baselines from (Aksu et al., 2024).

Type	Statistical Methods				Supervised Task-Specific Models					Zero-Shot Models					Zero-Shot Models (FL)		
Model	Naïve	Seasonal Naïve	Auto ARIMA	Auto Theta	DeepAR	TiDE	N-BEATS	PTST.	iTrans.	TimesFM	TabPFN	Chronos	Chronos-2	Moirai	FedAvg	FFTS	**FedTRL (Ours)**
MASE	1.260	1.000	0.964	0.978	1.206	0.980	0.842	0.762	0.802	0.680	0.748	0.786	0.677	0.809	0.872	0.766	**0.675**
CRPS	1.383	1.000	0.770	1.051	0.721	0.652	0.689	0.496	0.524	0.465	0.480	0.551	—	0.515	0.700	0.521	**0.460**
Rank	28.072	26.175	21.515	24.031	18.938	18.557	21.381	10.052	11.320	**8.237**	8.268	14.309	—	10.175	18.231	9.787	8.676

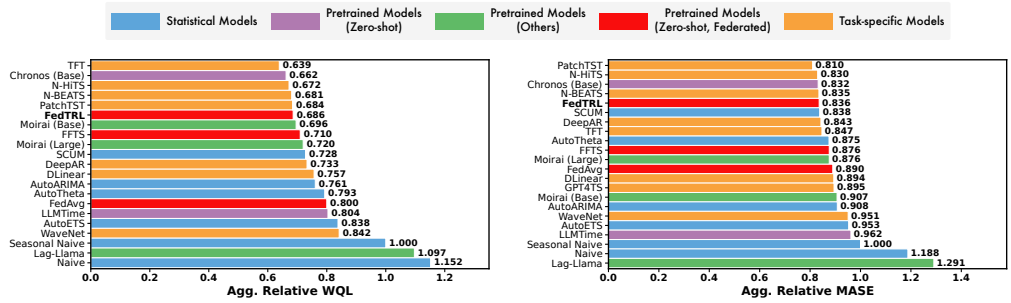


Figure 4: Results on FEV leaderboard (lower is better). (Others): pretrained models whose pretraining data overlap with the benchmark. FedTRL uses 20 generated series (Ansari et al., 2024).

FEV Leaderboard. We evaluate FedTRL on the FEV leaderboard (Ansari et al., 2024), covering 27 datasets for probabilistic forecasting. As shown in **Fig. 4**, FedTRL achieves strong zero-shot performance, outperforming most statistical and task-specific models trained with in-domain supervision, ranking second among pretrained FMs (ahead of Chronos) and first among federated FMs (ahead of FFTS and FedAvg). This performance is supported by diffusion-based reconstruction that captures fine-grained temporal dynamics at both patch and point levels, together with a warm-started dual-head design that extends accurate point forecasting to uncertainty modeling. Combined with DaG, these components make federated pretraining effective for both point and probabilistic forecasting.

4.3 Framework Analysis

We first confirm that FedTRL reverses both failure modes in **Fig. 1**, with the highest domain purity, the lowest within-domain variance, and the least gradient conflict (App. E.9), and then analyze its components, generality to other reconstruction methods, scalability, and hyperparameter sensitivity.

Table 5: Ablation study on point forecasting tasks.

Tasks		In-domain		Full-shot		Zero-shot	
	Variants	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	**FedTRL**	**0.448**	**0.472**	**0.372**	**0.399**	**0.399**	**0.420**
	w/o GRL	0.464	0.486	0.377	0.400	0.407	0.435
	w/o Proto	0.458	0.482	0.382	0.410	0.417	0.432
	w/o DaG	0.472	0.494	0.394	0.421	0.436	0.459
ETTm1	**FedTRL**	**0.375**	**0.402**	**0.316**	**0.369**	**0.350**	**0.380**
	w/o GRL	0.388	0.414	0.324	0.372	0.359	0.384
	w/o Proto	0.382	0.408	0.320	0.372	0.361	0.384
	w/o DaG	0.400	0.420	0.333	0.364	0.380	0.399
Weather	**FedTRL**	**0.241**	**0.282**	**0.214**	**0.252**	**0.238**	**0.279**
	w/o GRL	0.260	0.295	0.217	0.257	0.240	0.287
	w/o Proto	0.252	0.290	0.222	0.259	0.251	0.282
	w/o DaG	0.254	0.292	0.239	0.275	0.289	0.292

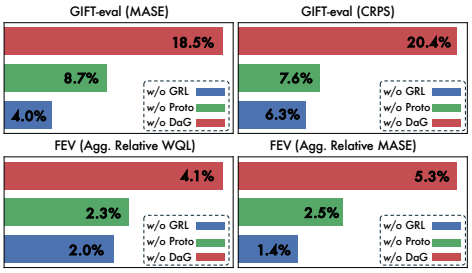


Figure 5: Relative performance drop of ablations on zero-shot probabilistic forecasting (GIFT-eval, FEV).

Ablation Study. We ablate three key components: GRL-based adversarial regularization, prototype alignment, and DaG, replacing DaG with FedAvg. As shown in **Table 5** and **Fig. 5**, removing any component consistently degrades both point and probabilistic forecasting. GRL contributes most to probabilistic forecasting through subdomain-invariant representations, while prototype alignment is particularly important for bridging client-level representation gaps in zero-shot transfer. Additional ablations replacing diffusion with masked or contrastive objectives and removing either DaG term independently show similar performance drops, confirming that each design choice matters (Table 26).

Can FedTRL Improve Masked Reconstruction Methods? To evaluate the generality of our proposed FedTRL, we integrate its key components into other masked reconstruction-based representation methods, including SimMTM and PatchTST. As shown in **Table 6**, incorporating FedTRL consistently improves performance across four long-term forecasting datasets. These demonstrate that FedTRL can seamlessly enhance different reconstruction-based representation methods, highlighting its effectiveness and broad applicability.

Table 6: Results on ETT, Weather and Traffic (averaged over horizons). Improvement (%) over baseline.

Model	ETTh1		ETTm1		Weather		Traffic	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
SimMTM	0.495	0.512	0.415	0.439	0.285	0.310	0.472	0.315
+ FedTRL	0.476	0.501	0.401	0.430	0.263	0.301	0.459	0.305
Improvement (%)	**3.84**	**2.15**	**3.37**	**2.05**	**7.88**	**2.48**	**2.82**	**3.13**
PatchTST	0.478	0.495	0.386	0.408	0.260	0.295	0.438	0.295
+ FedTRL	0.461	0.487	0.378	0.404	0.252	0.288	0.430	0.291
Improvement (%)	**3.56**	**1.62**	**2.07**	**0.98**	**3.08**	**2.37**	**1.83**	**1.36**

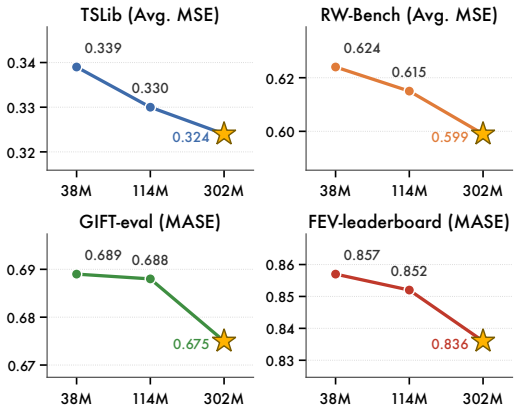


Figure 6: Model scalability across zero-shot point and probabilistic forecasting benchmarks.

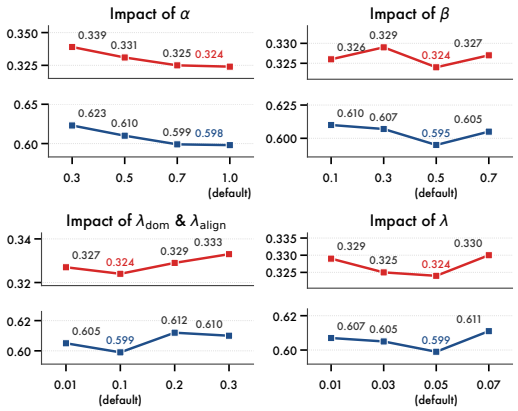


Figure 7: Hyperparameter sensitivity (Avg. MSE). **Red**: TSLib, **Blue**: RW-Bench.

Model Scalability. Scalability is a key property for large time series foundation models. We evaluate the proposed FedTRL across different backbone sizes (from 38M to 302M, more structure details are provided in **Table 18**). Results (**Fig. 6**) consistently show that FedTRL maintains strong performance improvements on both point and probabilistic forecasting benchmarks as the backbone grows, demonstrating its robustness as a scalable training paradigm for models of varying capacity.

Hyperparameter Sensitivity. We evaluate hyperparameter sensitivity (α, β, λ_{dom}&λ_{align}, λ) in **Fig. 7**, with consistent trends across zero-shot point forecasting. First, reducing α gradually degrades performance, highlighting the importance of strong domain discriminability. Second, β exhibits a U-shaped trend: too small under-utilizes alignment, too large over-regularizes, confirming that moderate values balance local and global consistency. Third, λ_{dom}, λ_{align} are stable within $[0.1, 0.3]$, but very small values harm invariance. Finally, λ also follows a concave trend, where weak reversal fails to suppress domain bias and strong reversal over-penalizes representations. Overall, FedTRL is robust across these settings, with default parameters balancing discriminability and invariance.

5 Conclusion

We propose FedTRL, a heterogeneity-aware federated pretraining framework for TSFMs that addresses inter- and intra-domain heterogeneity through prototype-level adversarial regularization and domain-aware aggregation. Extensive experiments on point and probabilistic forecasting show that FedTRL matches or outperforms centralized and federated baselines, with larger gains under domain imbalance, drift, and unseen-domain transfer. These results demonstrate that bi-level heterogeneity can be explicitly modeled to improve federated TSFM pretraining.

AI USE STATEMENT

We used large language models (LLMs) solely for language polishing of the manuscript, such as improving grammar, wording, and readability. LLMs were not used for any other purpose, including research ideation, method design, experiments, data analysis, or the writing of technical content. All text was reviewed by the authors, who take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

Source code is available at the anonymous link `https://anonymous.4open.science/r/FedTRL-Review-FDED`. The training procedure is given in Alg. 1, and the classifiers and subdomain label protocol in App. B.2–B.3. Architectures and hyperparameters are listed in Tables 10 and 18. Datasets, baselines, evaluation protocols, RW-Bench statistics, and the decontamination protocol are detailed in App. C. All proofs, with their assumptions, are in App. D.

REFERENCES

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

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

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, 2025. URL `https://arxiv.org/abs/2510.15821`.

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The uea multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. *arXiv preprint arXiv:2408.17253*, 2024.

Shengchao Chen, Guodong Long, Tao Shen, and Jing Jiang. Prompt federated learning for weather forecasting: Toward foundation models on meteorological data. *arXiv preprint arXiv:2301.09152*, 2023.

Shengchao Chen, Guodong Long, Jing Jiang, and Chengqi Zhang. Federated foundation models on heterogeneous time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15839–15847, 2025.

Shengchao Chen, Guodong Long, Michael Blumenstein, and Jing Jiang. Fedal: Federated dataset learning for general time series foundation models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=HK6t5x5gJq`.

Mingyue Cheng, Qi Liu, Zhiding Liu, Hao Zhang, Rujiao Zhang, and Enhong Chen. Timemae: Self-supervised representations of time series with decoupled masked autoencoders. *arXiv preprint arXiv:2303.00320*, 2023.

Mingyue Cheng, Xiaoyu Tao, Qi Liu, Hao Zhang, Yiheng Chen, and Defu Lian. Cross-domain pre-training with language models for transferable time series representations. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 175–183, 2025.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pp. 2089–2099. PMLR, 2021.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35: 16344–16359, 2022.

Abhimanyu Das, Weihao Kong, Andrew Leach, Shaan Mathur, Rajat Sen, and Rose Yu. Long-term forecasting with tide: Time-series dense encoder. *arXiv preprint arXiv:2304.08424*, 2023a.

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

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.

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

Samuel Dooley, Gurnoor Singh Khurana, Chirag Mohapatra, Siddartha V Naidu, and Colin White. Forecastpfn: Synthetically-trained zero-shot forecasting. *Advances in Neural Information Processing Systems*, 36, 2024.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

En Fu and Yanyan Hu. Frequency-masked embedding inference: a non-contrastive approach for time series representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16639–16647, 2025.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Azul Garza and Max Mergenthaler-Canseco. Timegpt-1. *arXiv preprint arXiv:2310.03589*, 2023.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. Moment: A family of open time-series foundation models. *arXiv preprint arXiv:2402.03885*, 2024.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36, 2024.

Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtárik. Lower bounds and optimal algorithms for personalized federated learning. *Advances in Neural Information Processing Systems*, 33:2304–2315, 2020.

Cheng He, Xu Huang, Gangwei Jiang, Zhaoyi Li, Defu Lian, Hong Xie, Enhong Chen, Patrick PC Lee, et al. Gtm: A general time-series model for enhanced representation learning of time-series data. In *International Conference on Learning Representations*, volume 2026, pp. 7280–7309, 2026.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. The tabular foundation model tabpfn outperforms specialized time series forecasting models based on simple features. In *NeurIPS Workshop on Time Series in the Age of Large Models*, 2024.

Xiang Lan, Hanshu Yan, Shenda Hong, and Mengling Feng. Towards enhancing time series contrastive learning: A dynamic bad pair mining approach. In *International conference on learning representations*, volume 2024, pp. 10359–10379, 2024.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*, pp. 6357–6368. PMLR, 2021a.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *arXiv preprint arXiv:2102.07623*, 2021b.

Qingxiang Liu, Xu Liu, Chenghao Liu, Qingsong Wen, and Yuxuan Liang. Time-ffm: Towards lm-empowered federated foundation model for time series forecasting. *arXiv preprint arXiv:2405.14252*, 2024a.

Xu Liu, Juncheng Liu, Gerald Woo, Taha Aksu, Yuxuan Liang, Roger Zimmermann, Chenghao Liu, Junnan Li, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Moirai-MoE: Empowering time series foundation models with sparse mixture of experts. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38940–38962. PMLR, 13–19 Jul 2025a. URL https://proceedings.mlr.press/v267/liu25an.html.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in Neural Information Processing Systems*, 35: 9881–9893, 2022.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625*, 2023.

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

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. *arXiv preprint arXiv:2502.00816*, 2025b.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

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

Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-beats: Neural basis expansion analysis for interpretable time series forecasting, 2020. URL https://arxiv.org/abs/1905.10437.

Junwoo Park, Daehoon Gwak, Jaegul Choo, and Edward Choi. Self-supervised contrastive learning for long-term forecasting. In *International Conference on Learning Representations*, volume 2024, pp. 19115–19137, 2024.

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

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3): 1181–1191, 2020.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-moe: Billion-scale time series foundation models with mixture of experts. *arXiv preprint arXiv:2409.16040*, 2024.

Zhi Wen Soi, Chenrui Fan, Aditya Shankar, Abele Mălan, and Lydia Y Chen. Federated time series generation on feature and temporally misaligned data. *arXiv preprint arXiv:2410.21072*, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Hao Sun, Xiaoli Tang, Chengyi Yang, Zhenpeng Yu, Xiuli Wang, Qijie Ding, Zengxiang Li, and Han Yu. Hifi-gas: Hierarchical federated learning incentive mechanism enhanced gas usage estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22824–22832, 2024.

Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE transactions on neural networks and learning systems*, 34(12):9587–9603, 2022.

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

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

Daoyu Wang, Mingyue Cheng, Zhiding Liu, and Qi Liu. Timedart: A diffusion autoregressive transformer for self-supervised time series representation, 2025a. URL https://arxiv.org/abs/2410.05711.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. *arXiv preprint arXiv:2405.14616*, 2024a.

Wenxuan Wang, Kai Wu, Dan Wang, Xiaoyu Zhang, et al. Synthetic series-symbol data generation for time series foundation models. *Advances in Neural Information Processing Systems*, 38: 10491–10554, 2026.

Yihang Wang, Yuying Qiu, Peng Chen, Kai Zhao, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. LightGTS: A lightweight general time series forecasting model. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025b.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. *arXiv preprint arXiv:2405.19363*, 2024b.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Advances in Neural Information Processing Systems*, 37:469–498, 2024c.

Gerald Woo, Chenghao Liu, Doyen Sahoo, Akshat Kumar, and Steven Hoi. Cost: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. *arXiv preprint arXiv:2202.01575*, 2022.

Gerald Woo, Chenghao Liu, Akshat Kumar, and Doyen Sahoo. Pushing the limits of pre-training for time series forecasting in the cloudops domain. *arXiv preprint arXiv:2310.05063*, 2023.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers, 2024. URL https://arxiv.org/abs/2402.02592.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

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

Maxwell Xu, Alexander Moreno, Hui Wei, Benjamin M Marlin, and James Rehg. Rebar: Retrieval-based reconstruction for time-series contrastive learning. In *International Conference on Learning Representations*, volume 2024, pp. 15440–15464, 2024.

Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Petzold, William Yang Wang, and Wei Cheng. A survey on detection of llms-generated content. *arXiv preprint arXiv:2310.15654*, 2023.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 11121–11128, 2023.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2114–2124, 2021.

Kexin Zhang, Qingsong Wen, Chaoli Zhang, Rongyao Cai, Ming Jin, Yong Liu, James Y Zhang, Yuxuan Liang, Guansong Pang, Dongjin Song, et al. Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. *IEEE transactions on pattern analysis and machine intelligence*, 46(10):6775–6794, 2024a.

Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020.

Tianping Zhang, Yizhuo Zhang, Wei Cao, Jiang Bian, Xiaohan Yi, Shun Zheng, and Jian Li. Less is more: Fast multivariate time series forecasting with light sampling-oriented mlp structures. *arXiv preprint arXiv:2207.01186*, 2022.

Ziyi Zhang, Shaogang Ren, Xiaoning Qian, and Nick Duffield. Towards invariant time series forecasting in smart cities. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 1344–1350, 2024b.

Roadmap. This appendix provides supplementary materials omitted from the main text, together with extended experimental results, theoretical analyses, and additional discussions. Specifically:

1. **Appendix A** presents extended related work, covering large-scale *time series foundation models* (Appendix A.1) and *heterogeneous federated learning* (Appendix A.2).
2. **Appendix B** provides supplementary methodological details, including the formulation of the *Gradient Reversal Layer* (Appendix B.1), the design of *sub-domain and global domain classifiers* (Appendix B.2), the *sub-domain label acquisition* protocol (Appendix B.3), the complete *workflow of FedTRL* (Appendix B.4), and the *client–domain assignment strategy* (Appendix B.5).
3. **Appendix D** presents formal definitions and theoretical analyses of key concepts in FedTRL, together with proofs establishing domain invariance, domain awareness (including a uniform lower bound on the aggregation weights), convergence of DaG-weighted federated optimization, and global dynamics.
4. **Appendix C** details the experimental settings, datasets, baseline implementations, and the *pre-training/evaluation decontamination protocol* (Appendix C.7) used across all forecasting tasks.
5. **Appendix E** reports additional experimental results and in-depth analyses, including model scale comparisons (Appendix E.1), evaluations of federated baselines under heterogeneity (Appendix E.2), extensive forecasting results (Appendices E.3–E.5), a full-shot comparison with fine-tuned TSFMs (Appendix E.6), scalability studies (Appendix E.7), *computational and communication efficiency* (Appendix E.8), and further analyses of generalization, representation and gradient behavior, robustness, extended ablations, component coupling, and statistical reliability (Appendix E.9).
6. **Appendix F** provides showcases during pretraining and forecasting.
7. **Appendix G** discusses limitations and broader impacts of FedTRL.

A More Related Work

A.1 Time Series Foundation Models

Pre-trained models that scale in capacity and data can evolve into foundation models capable of transferring across diverse domains. Moving beyond dataset-specific architectures, recent efforts have focused on building large time series foundation models (TSFMs) via cross-domain pretraining for zero-shot inference. Early examples include ForecastFPN (Dooley et al., 2024) using synthetic data, CloudOps (Woo et al., 2023) with masked modeling for domain-specific prediction, and TimeGPT-1 (Garza & Mergenthaler-Canseco, 2023) as the first commercial API. More recent advances such as MOMENT (Goswami et al., 2024), Moirai (Woo et al., 2024), TimeFM (Das et al., 2024), Chronos (Ansari et al., 2024), Timer (Liu et al., 2024b), and Time-MoE (Shi et al., 2024) scale pretraining to ultra-large datasets, achieving strong cross-task transfer. However, most TSFMs are trained under centralized paradigms (Shi et al., 2024; Das et al., 2024; Woo et al., 2024), where heterogeneous datasets are simply merged into mixed batches. This practice amplifies gradient conflicts and obscures domain-specific structure, leading to suboptimal representations. Temporal data are also fragmented across organizations and applications, making centralized collection impractical. These challenges call for flexible training strategies that explicitly address bi-level heterogeneity, spanning both inter-domain misalignment and intra-domain variability. Our work therefore introduces an FL framework for bi-level heterogeneous learning, enabling scalable TSFM pretraining.

A.2 Heterogeneous Federated Learning

FL (McMahan et al., 2017) enables decentralized model training across distributed clients without sharing raw data. A key challenge is statistical heterogeneity, which degrades performance due to distributional shifts across clients. Early approaches reduce client–server divergence through alignment or regularization, but generally assume mild heterogeneity and shared feature spaces, an assumption that breaks down in diverse real-world domains. Personalized Federated Learning (PFL) tailors models to individual clients using strategies such as regularization-based decomposition(Hanzely

et al., 2020; Li et al., 2021a), partial model sharing (Li et al., 2021b; Collins et al., 2021), adaptive aggregation (Zhang et al., 2020), or meta-learning (Fallah et al., 2020). While effective in vision and language tasks, these methods rely on transferable low-level features (Chen et al., 2025), which are less applicable to time series due to deeper variability in resolution, semantics, and physical context. More importantly, PFL prioritizes personalization over generalization, making it ill-suited for TSFM pretraining where cross-domain transferability is crucial. Instead of producing personalized models, our goal is to train a unified TSFM that explicitly resolves bi-level heterogeneity, capturing both inter-domain discrepancies and intra-domain conflicts during pretraining.

B Supplementary Materials

B.1 Gradient Reversal Layer

The Gradient Reversal Layer (GRL) (Ganin et al., 2016) is a standard technique in domain-adversarial learning. It acts as an identity function during the forward pass but reverses and scales gradients during backpropagation. Formally, given an input feature $\mathbf{h}$, GRL defines:

$$\mathrm{GRL}(\mathbf{h}) = \mathbf{h}, \quad \frac{\partial\, \mathrm{GRL}(\mathbf{h})}{\partial \mathbf{h}} = -\lambda I, \tag{10}$$

where $\lambda \geq 0$ is a tunable coefficient and I denotes the identity matrix. This operation allows the sub-domain classifier to minimize its standard cross-entropy loss, while the encoder is simultaneously optimized to maximize the same objective, thereby learning domain-invariant representations. The coefficient λ controls the strength of adversarial regularization, balancing domain invariance against the preservation of task-relevant temporal dynamics. Note that for training the local sub-domain classifier, we assign each sub-domain an index as its category label y^{dom}, while for the global domain classifier, we use the client index as the category label.

B.2 Sub-domain and Domain Classifier

The local sub-domain classifier promotes intra-domain invariance by mitigating fragmentation within each client, whereas the global domain classifier evaluates inter-domain separability to guide adaptive aggregation. Together, they target both levels of heterogeneity and support robust, scalable pretraining of TSFMs. The specific details of these classifiers are described below.

Local Sub-domain Classifier On each client, a sub-domain classifier D_{θ_C} is trained on $\bar{\mathbf{p}}$ with sub-domain index y^{dom} as supervision. To encourage sub-domain-invariant representations, we apply a GRL to the encoder pathway and optimize the sub-domain classifier. The classifier learns to predict sub-domains while the encoder learns to remove sub-domain cues (GRL scales gradients by $-\lambda$). This min–max game reduces the mutual information between embeddings and sub-domain labels, mitigating intra-domain drift without collapsing temporal content.

Global Domain Classifier At the server level, we must evaluate how domain-specific each client's representation remains to guide aggregation. For this purpose, we train a global domain classifier $f_d : \mathbb{R}^d \to \mathbb{R}^K$, where K is the number of clients (domains). Using client indices as supervision, f_d is trained on client prototypes $\{\bar{\mathbf{p}}_k\}$. After training, f_d is fixed and used for inference. For each prototype $\bar{\mathbf{p}}_k$, the **Domain Discriminability Score** is defined as $\ell_k = \mathcal{L}_{\mathrm{CE}}(f_d(\bar{\mathbf{p}}_k), k)$. Low ℓ_k indicates that the prototype has a clear semantic signature and is therefore a reliable contributor to aggregation, while high ℓ_k indicates an ambiguous prototype whose update should be down-weighted. These scores, combined with alignment to the global prototype, determine the DaG weights.

B.3 Sub-domain Label Acquisition

The sub-domain labels y^{dom} used by the local classifier come from metadata that is already available in the datasets or client records and require no semantic annotation: sub-dataset identifiers within each Time-MoE domain, station identifiers for weather clients, and sensor or meter identifiers where present. The number of sub-domains M_k therefore varies per client, from 1 to 15 in our pretraining corpus. When such metadata is missing, we cluster the local mean-pooled representations with k-means and select the number of clusters by the silhouette score. A single label per client is

the conservative fallback, under which GRL is inactive. **Table 7(a)** shows that FedTRL degrades gracefully when 20% or 50% of the labels are randomly flipped, that k-means pseudo-labels nearly match metadata labels, and that discarding sub-domain information entirely (single label) causes the largest drop, confirming that the intra-client level matters while curated annotation is not required.

B.4 WORKFLOW OF FEDTRL

The workflow of our proposed FedTRL is shown in **Alg. 1**. At each round, the server broadcasts the current global encoder and global prototype. Each selected client performs diffusion-based reconstruction to learn temporal features, applies a local sub-domain classifier with a GRL to promote domain invariance, and aligns its prototype to the previous global prototype. The client then uploads only its encoder and prototype. The server forms a new global prototype via weighted averaging, trains a global domain classifier on client prototypes, and computes for each client a domain discriminability score (ℓ_k) and a prototype deviation (δ_k). These are combined into aggregation logits $a_k = -\alpha\ell_k - \beta\delta_k$, softmax-normalized to weights w_k, which are used to aggregate encoders into the updated global model. The decoder and sub-domain classifier remain local throughout, preventing client-specific bias from contaminating the global encoder.

Algorithm 1: Workflow of FedTRL

Input: Initial encoder $\theta^{(0)}$; total rounds T; local epochs E; learning rate η; weighting factors $\lambda_{\text{domain}}, \lambda_{\text{align}}$

Output: Global encoder $\theta_E^{g,T}$ as the time series foundation model

for $t = 1, \ldots, T$ **do**

  Server selects a full client set $\mathcal{S}^t \subseteq \{1, \ldots, K\}$ of clients;

  foreach *client* $k \in \mathcal{S}^t$ ***in parallel*** **do**

    Download the last global encoder θ_E^{t-1} and prototype $\mathbf{p}^{t-1}$;

    for $e = 1$ **to** E **do**

      Sample mini-batch $(\mathbf{X}_k, \mathbf{Y}_k) \sim \mathcal{D}_k$;

      Generate noised patches and perform denoising reconstruction;

      Compute reconstruction loss: $\mathcal{L}_{\text{task}}$;

      Compute domain loss via local classifier and GRL: $\mathcal{L}_{\text{dom}}$;

      Compute prototype deviation: $\mathcal{L}_{\text{align}}^t = \|\bar{\mathbf{p}}_k^t - \mathbf{p}^{g,t-1}\|^2$;

      Combine local optimization objective: $\mathcal{L}_k^t = (\mathcal{L}_{\text{task},k}^t + \lambda_{\text{dom}}\mathcal{L}_{\text{dom},k}^t + \lambda_{\text{align}}\mathcal{L}_{\text{align},k}^t)$;

      Update local encoder via: $\theta_{E,k}^t \leftarrow \theta_{E,k}^{t-1} - \eta\nabla_{\theta_{E,k}^t}\mathcal{L}_k^t$;

      (**in parallel**) Updated local decoder via: $\theta_{D,k}^t \leftarrow \theta_{D,k}^{t-1} - \eta\nabla_{\theta_{D,k}^t}\mathcal{L}_k^t$;

      (**in parallel**) Updated local classifier via: $\theta_{C,k}^t \leftarrow \theta_{C,k}^{t-1} - \eta\nabla_{\theta_{C,k}^t}\mathcal{L}_k^t$;

    Upload only $\theta_{E,k}^t$, $\bar{\mathbf{p}}_k^t$ (local prototype) to server for aggregation;

  /* Server-side aggregation */

  Aggregate prototypes: $\mathbf{p}^{g,t} =: \sum_k w_k \bar{\mathbf{p}}_k^t$;

  Update and evaluate the global domain classifier on all $\mathbf{p}_k^t$, obtain domain discriminability scores ℓ_k^t;

  Compute prototype deviations: $\delta_k = \|\bar{\mathbf{p}}_k^t - \mathbf{p}^{g,t}\|^2$;

  Compute aggregation logits: $a_k = -\alpha\,\ell_k^t - \beta\,\delta_k^t$, then weights $w_k = \mathrm{Softmax}(a_k)$;

  Aggregate encoder: $\theta_E^t = \sum_k w_k \theta_{E,k}^t$;

  Broadcast new $\theta_E^{g,t}$, $\mathbf{p}^{g,t}$ to all clients;

return *Time series foundation model* θ_E^T

B.5 CLIENT–DOMAIN ASSIGNMENT STRATEGY

Throughout our experiments, we follow the standard FL-for-time-series convention (Chen et al., 2025) of treating each dataset (or each pre-training domain in Time-MoE-300B) as a single client. This is a deliberate design choice for reproducibility and direct comparability with prior FL-for-time-series work, rather than a strict requirement of FedTRL.

Two natural alternatives exist. (i) *Similar-domain merging*: clients holding closely related domains (e.g., two energy datasets with comparable sampling rates and physical processes) can be combined to increase per-client sample size and reduce the overall number of communication rounds. (ii) *Sub-domain splitting*: a single dataset that already exhibits strong internal heterogeneity (e.g., a multi-site weather corpus spanning many geographically distinct stations) can be partitioned into multiple clients to make the local adversarial regularizer more effective on each finer partition.

Both alternatives interact with FedTRL through a single mechanism: the prototype discriminability score ℓ_k and prototype deviation δ_k that determine the DaG aggregation weights w_k. When two genuinely-similar domains are merged, the combined prototype tends to be more clearly classified (lower ℓ_k) and is therefore upweighted by DaG. When an internally heterogeneous client is left intact, its prototype is more ambiguous (higher ℓ_k) and DaG correspondingly down-weights it relative to clients with cleaner signatures. In this sense, the client–domain assignment is not an isolated hyperparameter that must be tuned in isolation: its effect is observable through prototype statistics, and DaG's reliability-based reweighting absorbs a substantial fraction of suboptimal assignments. We therefore expose the assignment policy as a deployment-side knob rather than as a method-side requirement, and adopt the one-dataset-per-client convention purely for evaluation consistency with previous federated TSFM work.

Empirical Sensitivity to the Partition. We re-partition the Time-MoE pretraining corpus in four ways and re-train FedTRL (**Table 7(b)**). Coarse (5 clients) or fine (18 clients) assignments incur only minor degradation, consistent with the absorbing effect of DaG described above. A random partition removes the inter-client structure, yet FedTRL still outperforms centralized mixing and remains well ahead of FedAvg, because GRL keeps acting inside each client. We further compare FedTRL with a reduction to partitioned training with weighted aggregation, keeping the client partition and local training but computing prototype weights once and holding them fixed (**Table 7(b)**). One-shot prototype weighting recovers part of the gap to FedTRL, and restoring the local regularizers and per-round reweighting accounts for the rest, so FedTRL is not a fixed-weight scheme.

Table 7: Robustness to sub-domain labels and client partitions (MSE averaged over horizons, TSLib avg. over five datasets). (a) Label source and noise. (b) Partition and aggregation weighting, zero-shot.

(a) Sub-domain labels

Labels	**Weather** (in-dom.)	**RW-Bench** (zero-shot)
Metadata (default)	**0.241**	**0.599**
20% labels flipped	0.243	0.602
50% labels flipped	0.248	0.610
k-means pseudo-labels	0.244	0.603
Single label (no GRL)	0.260	0.617

(b) Partition and aggregation

Setting	**TSLib avg.**	**RW-Bench**
FedTRL (9 domain clients)	**0.324**	**0.599**
coarse (5 clients)	0.329	0.607
fine (18 clients)	0.326	0.602
random partition	0.341	0.612
Partitioned, fixed weights	0.337	0.631
+ GRL & prototype align.	0.330	0.616
All Mixed (centralized)	0.345	0.614
FedAvg (9 domain clients)	0.348	0.729

C Experiments Details

C.1 In-domain Point Forecasting

This task evaluates self-supervised time series representation learning. We introduce baselines across three categories: (i) masked reconstruction methods, including SimMTM (Dong et al., 2023) and PatchTST (Nie et al., 2022), (ii) contrastive learning methods, including TS-TCC (Eldele et al., 2021) and CoST (Woo et al., 2022), and (iii) federated learning methods, including FFTS (Chen et al., 2025) and FedAvg (McMahan et al., 2017), where each client adopts TimeDART (Wang et al., 2025a). Introducation of these baselines are summarized as follows:

- **Masked Reconstruction Methods:**

  - **SimMTM** (Dong et al., 2023): It recovers masked time series by learning point-wise similarities between masked and unmasked sequences through a manifold learning approach.
  - **PatchTST** (Nie et al., 2022): It introduces a patching mechanism and channel-independent strategy to extract local semantic information and capture long-range dependencies effectively.
  - **TimeMAE** (Cheng et al., 2023): It learns decoupled masked autoencoder representations over sub-series windows, reconstructing masked tokens and predicting their discrete codewords.
- **Contrastive Learning Methods:**
  - **TS-TCC** (Eldele et al., 2021): It employs temporal and contextual contrastive modules to learn robust representations by ensuring consistency across different augmented views.
  - **CoST** (Woo et al., 2022): It learns disentangled seasonal-trend representations by applying contrastive losses in both the time and frequency domains.
- **Federated Learning Methods:**
  - **FedAvg** (McMahan et al., 2017): The standard federated learning algorithm that aggregates global models by performing a weighted average of local model parameters.
  - **FFTS** (Chen et al., 2025): A federated learning framework specifically designed for time series that addresses data heterogeneity across distributed clients to achieve a unified representation space.
- **Centralized Counterpart**
  - **Centralized**: A baseline where all datasets are pooled into a single server for training, serving as an empirical performance upper bound for the federated scenarios.

For fair comparison, all unsupervised baselines (except the FL ones) are implemented under the FedAvg protocol. In addition, a centralized baseline is constructed by mixing all datasets and training with the channel-independent (Nie et al., 2022). All methods adopt a consistent Transformer backbone. The results are shown in **Table 1**. We additionally evaluate four recent self-supervised objectives under the identical in-domain federated protocol with FedAvg aggregation and matched backbone capacity: REBAR (Xu et al., 2024) (retrieval-based reconstruction for contrastive positive-pair identification), AutoCon (Park et al., 2024) (autocorrelation-based contrastive learning for long-term forecasting), FEI (Fu & Hu, 2025) (frequency-masked embedding inference), and DBPM (Lan et al., 2024) (dynamic bad-pair mining, a plug-in that we apply to CoST). Results are reported in **Table 12**.

C.2 Full-shot Point Forecasting

This task evaluates long-term point forecasting in a fully supervised setting, where models are trained or adapted using the entire training dataset. We benchmark FedTRL against advanced deep forecasting models, including TimeMixer (Wang et al., 2024a), TimeXer (Wang et al., 2024c), PatchTST (Nie et al., 2022), TimesNet (Wu et al., 2022), DLinear (Zeng et al., 2023), iTransformer (Liu et al., 2023), Autoformer (Wu et al., 2021), Non-Stationary Transformer (Liu et al., 2022), and LightTS (Zhang et al., 2022). Brief introductions of these baselines are summarized as follows:

- **TimeMixer** (Wang et al., 2024a): It analyzes temporal variations through a decomposable multiscale mixing architecture, effectively integrating both fine-grained and coarse-grained information.
- **TimeXer** (Wang et al., 2024c): It reconciles endogenous and exogenous variables by introducing a novel embedding layer and capturing dependencies at both patch and variate levels.
- **PatchTST** (Nie et al., 2022): It leverages a patching mechanism and channel-independent strategy to enhance the extraction of local semantic patterns and the modeling of long-range dependencies.
- **TimesNet** (Wu et al., 2022): It transforms 1D time series into 2D tensors based on multi-periodicity to model intra-period and inter-period variations using 2D convolutional kernels.

- **DLinear** (Zeng et al., 2023): A simple yet effective decomposition-based MLP model that separately maps trend and seasonal components through linear layers.
- **iTransformer** (Liu et al., 2023): It inverts the Transformer architecture by treating each variate's entire series as a token, allowing attention to capture multivariate correlations and FFNs to model temporal patterns.
- **Autoformer** (Wu et al., 2021): It integrates a deep decomposition architecture with an auto-correlation mechanism to replace standard self-attention for efficient long-term forecasting.
- **Non-Stationary Transformer** (Liu et al., 2022): It addresses distribution shifts in real-world data using Series Stationarization and De-stationary Attention to restore original non-stationary information.
- **LightTS** (Zhang et al., 2022): A lightweight, sampling-oriented MLP-based framework that employs interval sampling and multi-scale structures for fast and robust multivariate forecasting.

For FedTRL, the model is first pretrained on Time-MoE-300B and then adapted to each target dataset with only a single epoch for fast transfer. The implementation is based on TSLib (Wu et al., 2022). Results are reported in **Table 2**. Fine-tuned centralized TSFMs are compared in Appendix E.6.

C.3 Zero-shot Point Forecasting

This task evaluates zero-shot long-term point forecasting, where models are pretrained on large time series datasets. We benchmark FedTRL against advanced foundation models, including Time-MoE (Shi et al., 2024), TimesFM (Das et al., 2024), Moirai (Woo et al., 2024), Chronos (Ansari et al., 2024), and Moment (Goswami et al., 2024). Evaluation is conducted on two real-world benchmarks: TSLib (four ETT-series and Weather) and RW-Bench (15 datasets from diverse regions, with details in Section C.5). Following the protocol of (Shi et al., 2024), the look-back window is set to $\{512, 1024, 2048, 3072\}$ with corresponding prediction horizons $\{96, 192, 336, 720\}$. Among these baselines, TimesFM and Moment require fine-tuning for adaptation. The results on TSLib are officially from (Shi et al., 2024). Time-MoE$_{\text{ultra}}$ is excluded since it is not publicly available. The results are shown in **Table 3** and full results in **Table 14**.

Inference Protocol on Unseen Domains. For an unseen domain, no new client is registered and no additional federated updates are performed. In the zero-shot setting, the final global encoder is frozen and applied directly with instance normalization. In the full-shot setting, the same encoder is fine-tuned locally on the target-domain training data for one epoch, without federation.

C.4 Zero-shot Probabilistic Forecasting

This task evaluates zero-shot probabilistic forecasting. We evaluate FedTRL on two widely used benchmarks. GIFT-Eval (Aksu et al., 2024) comprehensively assesses performance across 23 datasets, covering 144K time series and 177M data points, with a total of 97 forecasting configurations. We follow the official evaluation protocol provided by Salesforce and report aggregated results in **Table 4**. In addition, we evaluate both accuracy and inference efficiency on the FEV leaderboard(Ansari et al., 2024), maintained by AutoGluon, which includes 27 datasets for standardized zero-shot testing. Aggregated results are reported in **Fig. 4**.

C.5 RW-Bench for Zero-shot Point Forecasting Evaluation

To assess the real-world applicability of pretrained FMs, we introduce the Real-world Weather Benchmark (RW-Bench), constructed from 15 ground-based weather observation stations collected in collaboration with our industry partners. Unlike curated academic datasets, RW-Bench preserves the complexity of raw operational data: no outliers are removed, and missing values are simply zero-filled. This design ensures that models are tested under conditions closer to real deployment scenarios. RW-Bench spans January 2022 to August 2025, providing timely and authentic observational records that complement the synthetic reanalysis data (e.g., ERA5) used in large-scale pretraining corpora such as Time-MoE-300B. *Importantly, the spatiotemporal non-overlap between RW-Bench and pretraining datasets enables a fair zero-shot evaluation of model generalization.* Each dataset contains hourly

multivariate time series covering ten meteorological variables: air pressure, air temperature, relative humidity, precipitation, wind direction, wind speed, maximum wind direction, maximum wind speed, maximum temperature, and minimum temperature. RW-Bench presents a challenging testbed for zero-shot forecasting. Dataset statistics for TSLib and RW-Bench are provided in **Table 8** and **Table 9**, while sample visualizations appear in **Fig. 8**.

Table 8: Statistics of TSLib (Wu et al., 2022). Channels: number of variables. Size: (training, validation, testing). †SE (spectral entropy), ACF (autocorrelation-based), and Proxy (error-based) measure forecastability, where higher means easier to predict.

Dataset	Domain	# Channels	Frequency	Size	Forecast Length	SE†	ACF†	Proxy†
ETTh1	Power	7	Hourly	(8545, 2881, 2881)	{96, 192, 336, 720}	0.523	20.372	0.857
ETTh2	Power	7	Hourly	(8545, 2881, 2881)	{96, 192, 336, 720}	0.652	35.940	0.943
ETTm1	Power	7	15 Minute	(34465, 11521, 11521)	{96, 192, 336, 720}	0.580	24.335	0.949
ETTm2	Power	7	15 Minute	(34465, 11521, 11521)	{96, 192, 336, 720}	0.696	40.276	0.978
Electricity	Energy	321	Hourly	(17805, 2537, 5166)	{96, 192, 336, 720}	0.706	12.137	0.814
Traffic	Transportation	862	Hourly	(11673, 1661, 3413)	{96, 192, 336, 720}	0.567	6.977	0.681
Weather	Weather	21	10 Minute	(36792, 5271, 10540)	{96, 192, 336, 720}	0.549	25.818	0.842
Exchange	Finance	8	1 Day	(4704, 665, 1422)	{96, 192, 336, 720}	0.793	46.054	0.998

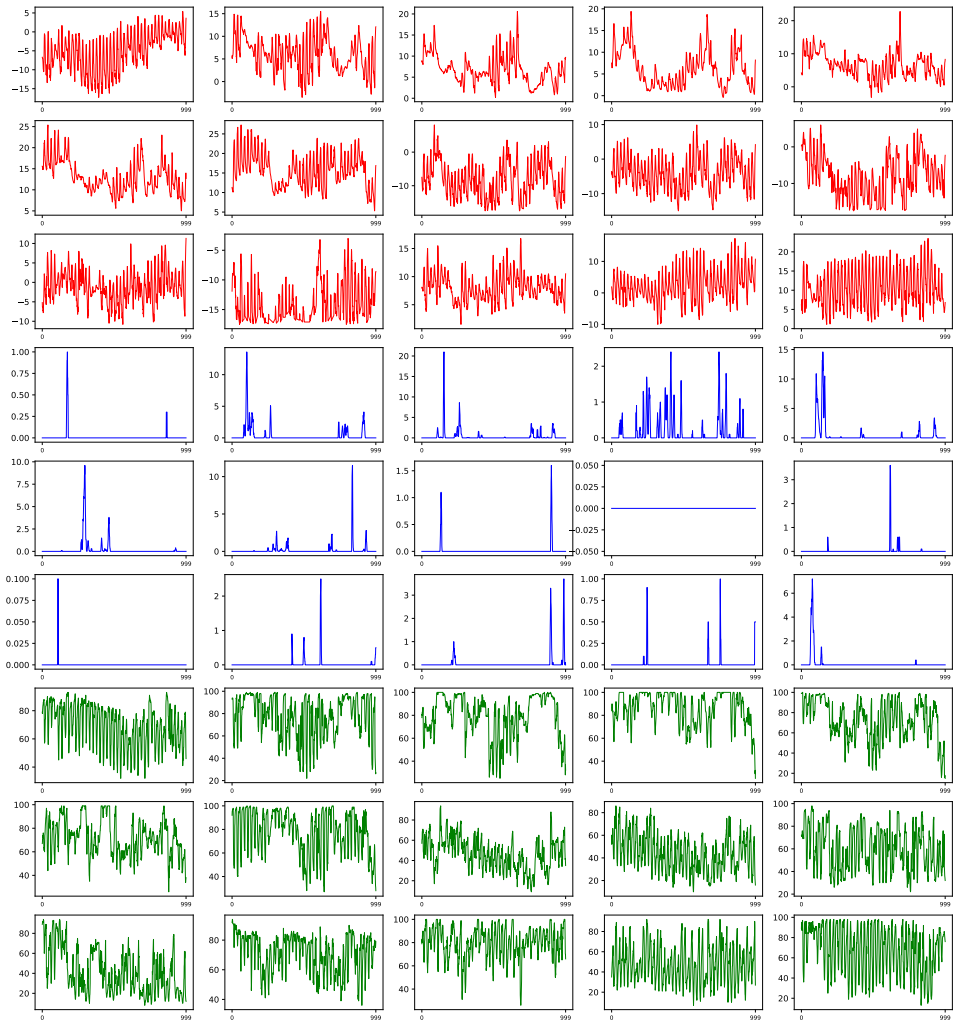


Figure 8: Visualization of RW-Bench samples: red denotes temperature, blue denotes precipitation, and green denotes humidity. Variables are arranged following the order in Table 9. RW-Bench will be open-sourced once the required permissions are obtained.

Table 9: Statistics of RW-Bench. Channels: number of variables. Size: (training, validation, testing). †SE (spectral entropy), ACF (autocorrelation-based), and Proxy (error-based) measure forecastability, where higher means easier to predict.

Dataset	Region	# Channels	Frequency	Size	Forecast Length	SE†	ACF†	Proxy†
RW-1	Yuzhno-Sakhalinsk	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.438	17.089	0.682
RW-2	Tokyo	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.438	17.090	0.682
RW-3	Beijing	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.431	16.660	0.711
RW-4	Mumbai/Santacruz	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.398	16.474	0.706
RW-5	Ho Chi Minh City	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.462	17.292	0.768
RW-6	Cairo	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.456	16.912	0.734
RW-7	Nairobi Dagoretti	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.425	16.217	0.656
RW-8	Nauru	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.452	11.586	0.705
RW-9	Honolulu	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.435	11.696	0.734
RW-10	Los Angeles	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.425	15.879	0.726
RW-11	Edmonton	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.437	18.865	0.805
RW-12	Oxford	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.455	18.871	0.816
RW-13	Zurich	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.463	17.805	0.801
RW-14	McMurdo	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.431	16.438	0.720
RW-15	Novolazarevskaya	10	Hourly	January 2022 - August 2025	{96, 192, 336, 720}	0.426	15.508	0.699

C.6 Model Architecture and Training Configuration

We adopt a vanilla Transformer (Vaswani et al., 2017) as the backbone. The local encoder is implemented as a causal Transformer encoder. For in-domain forecasting, we employ a standard multi-head attention mechanism with positional embeddings. For large-scale foundation model pretraining, we replace this with FlashAttention (Dao et al., 2022) and RoPE (Su et al., 2024) to improve efficiency, consistent with advanced foundation model practices. Details in **Table 10**.

Table 10: Detailed model architecture and training configuration.

Category	**Configuration (In-domain Forecasting)**	**Configuration (Large Foundation Model)**
Optimizer	Adam	Adam
Batch Size	128	2048
Local Epochs	10	100
Global Rounds	200	10,000
Learning Rate (Local Updating)	0.0001	0.0005
Input Sequence Length	512	3072
Learning Scheduler	StepLR	StepLR
Diffusion Step	1250	2500
Computation Devices	4 × Nvidia RTX A5500-24GB GPUs	16× Nvidia A100-80G GPUs
Attention	Vanilia	FlashAttention with RoPE
Encoder Layer	4	24
(Denoising) Decoder Layers	2	6
Feedforward Dimension	512	4096
Model Dimension	512	1024
Number of Heads	4	8
Dropout	0.4	0.4
Activation Function	GELU	GELU
Patch Length	16	16
Stride	16	16
Domain Classifier Structure	2 × MLP with ReLU	4 × MLP with ReLU
Training Epoch on Server	5	20
Learning Rate	0.0005	0.0001
Domain Discriminability Weight α	1.0	1.0
Semantic Alignment Weight β	0.5	0.5
Domain-adversarial Weight λ_{dom}	0.1	0.1
Prototype Alignment Weight λ_{align}	0.1	0.1
Warmup Coefficients R_{warm}	Not applicable	4,000

C.7 Pre-training/Evaluation Decontamination Protocol

For every evaluation split used in this paper, we compute dataset-level identifiers from a combination of source name, observation period, sampling frequency, and variable schema, audit the Time-MoE-300B pre-training shards for matching *instances*, and remove any matches from pre-training before

the corresponding model is trained. This instance-level protocol follows the convention adopted by prior large-scale TSFM work (Shi et al., 2024; Woo et al., 2024; Ansari et al., 2024).

D Theoretical Analysis and Proofs

In this section, we provide formal definitions and proof-level justifications for the concepts of **Domain-Invariant Representations**, **Domain Awareness**, and **Global Dynamics**. We distinguish local sub-domain invariance from server-side prototype discriminability: the local GRL objective suppresses sub-domain classifiability *within* a client, whereas the server-side score $\ell_k = \mathcal{L}_{\mathrm{CE}}(f_d(\bar{\mathbf{p}}_k), k)$ is a *Domain Discriminability Score*, where low values indicate clear, reliable prototypes and high values indicate ambiguous prototypes. We formulate the learning problem within domain adaptation and distributed optimization. Let $\mathcal{X} \subseteq \mathbb{R}^{T\times C}$ be the input space and $\mathcal{Z} \subseteq \mathbb{R}^{N\times d}$ be the representation space. We consider K source domains (clients), where client k is characterized by a distribution $\mathcal{D}_k$ over $\mathcal{X} \times \mathcal{Y}$ and may contain M_k sub-domains $m = 1, \ldots, M_k$ with latent distributions $\mathcal{D}^Z_{k,m}$ induced by the encoder $E_{\theta_E} : \mathcal{X} \to \mathcal{Z}$, and we write $\bar{\mathcal{D}}^Z_k = M_k^{-1} \sum_{m=1}^{M_k} \mathcal{D}^Z_{k,m}$ for the uniform latent mixture over its sub-domains. Aggregation logits and softmax weights are denoted a_k and w_k.

D.1 Domain-Invariant Representations

Definition D.1 (**$\mathcal{H}$-Divergence for Domain Invariance**). Let $\mathcal{H}$ be a hypothesis class of binary classifiers (domain discriminators). The $\mathcal{H}$-divergence between two distributions $\mathcal{D}^Z_i$ and $\mathcal{D}^Z_j$ in the latent space induced by E is defined as:

$$d_{\mathcal{H}}(\mathcal{D}^Z_i, \mathcal{D}^Z_j) = 2 \sup_{h\in\mathcal{H}} \left| \Pr_{\mathbf{z}\sim\mathcal{D}^Z_i}[h(\mathbf{z}) = 1] - \Pr_{\mathbf{z}\sim\mathcal{D}^Z_j}[h(\mathbf{z}) = 1] \right|. \tag{11}$$

A representation is *domain-invariant* if $d_{\mathcal{H}}(\mathcal{D}^Z_i, \mathcal{D}^Z_j)$ is minimized for all pairs i, j.

Theorem D.2 (**Adversarial Reduction of Empirical Sub-domain Separability**). *For a fixed client k, assume balanced sampling over its sub-domains and let $\bar{\mathcal{D}}^Z_k$ denote the uniform latent mixture over those sub-domains. If the local discriminator class $\mathcal{H}$ is used as an empirical surrogate for one-vs-mixture sub-domain tests and is optimized near its empirical optimum for each fixed encoder, then the GRL update on θ_E reduces an empirical discriminator-based proxy for sub-domain separability. This is a proxy for reducing empirical $\mathcal{H}$-divergence among local sub-domains, and is* not *a direct minimization of the population $\mathcal{H}$-divergence. In Ben-David-style domain-adaptation bounds, such divergence terms can tighten a target-risk bound only together with the source risk and the ideal joint risk term $\lambda^* = \min_{h\in\mathcal{H}}\big(\epsilon_i(h) + \epsilon_j(h)\big)$.*

Proof. For fixed encoder parameters θ_E, the local classifier D_{θ_C} is trained to minimize the sub-domain cross-entropy loss $\mathcal{L}_{\mathrm{dom}}$. Cross-entropy is used here as a surrogate for empirical domain-classification error, and is not itself the $\mathcal{H}$-divergence. In the balanced binary case, the empirical $\mathcal{H}$-divergence proxy between two latent distributions P and Q is related to the best empirical domain-classification error by

$$\widehat{d}_{\mathcal{H}}(P, Q) = 2\big(1 - 2\,\widehat{\epsilon}^*_{\mathcal{H}}(P, Q)\big), \tag{12}$$

where $\widehat{\epsilon}^*_{\mathcal{H}}(P, Q)$ is the minimum balanced empirical 0–1 error over classifiers in $\mathcal{H}$. Thus, making the best discriminator less accurate decreases this empirical separability proxy.

With a GRL, the classifier-side update minimizes $\mathcal{L}_{\mathrm{dom}}$, while the encoder receives the reversed gradient. For the adversarial component alone, this corresponds to the idealized game

$$\theta^*_E \in \arg\max_{\theta_E} \min_{\theta_C} \mathcal{L}_{\mathrm{dom}}(\theta_E, \theta_C), \tag{13}$$

up to the GRL scale factor. Therefore, under the stated surrogate and near-optimal-discriminator assumptions, the encoder is encouraged to make local sub-domain prototypes harder to classify, reducing an empirical proxy for sub-domain separability. The standard domain-adaptation generalization bound additionally contains source-task error and the ideal joint risk λ^*, so this adversarial term should be interpreted as a regularizer that can support generalization rather than as a standalone unseen-domain guarantee. □

D.2 Domain Awareness via Entropy-Regularized Aggregation

Here we justify the softmax weighting mechanism used in DaG.

Definition D.3 (Prototype Aggregation Penalty). We define the prototype aggregation penalty for client k as

$$C_k = \alpha\,\ell_k + \beta\,\delta_k, \tag{14}$$

where ℓ_k is the Domain Discriminability Score and $\delta_k = \|\bar{\mathbf{p}}_k - \mathbf{p}^g\|_2^2$ is the squared deviation from the global prototype. Under the main-text interpretation, low ℓ_k indicates a clear, reliable prototype, while high ℓ_k indicates an ambiguous prototype. Thus, high C_k means that the client prototype is unreliable, drifted, or both.

Theorem D.4 (Optimality of Softmax Aggregation). *Assume $K < \infty$, $C_k \in \mathbb{R}$ for all k, and $\tau > 0$. The weights*

$$w_k = \frac{\exp(-C_k/\tau)}{\sum_{j=1}^{K} \exp(-C_j/\tau)} \tag{15}$$

are the unique minimizer of the entropy-regularized aggregation-penalty objective

$$\min_{\mathbf{w}\in\Delta^{K-1}} \sum_{k=1}^{K} w_k\,C_k - \tau\,H(\mathbf{w}). \tag{16}$$

For $\tau = 1$, this recovers the FedTRL weighting rule $w_k = \mathrm{Softmax}(-\alpha\ell_k - \beta\delta_k)$.

Proof. Let $\mathbf{w} = [w_1, \ldots, w_K] \in \Delta^{K-1}$ and use the convention $0\log 0 = 0$. Since $\tau > 0$, the objective

$$\mathcal{J}(\mathbf{w}) = \sum_{k=1}^{K} w_k\,C_k + \tau \sum_{k=1}^{K} w_k \log w_k \tag{17}$$

is strictly convex on the simplex whenever the costs are finite. Its minimizer is interior, so the Lagrangian first-order conditions apply:

$$\mathcal{L}(\mathbf{w}, \nu) = \sum_{k=1}^{K} w_k\,C_k + \tau \sum_{k=1}^{K} w_k \log w_k + \nu\Big(\sum_{k=1}^{K} w_k - 1\Big). \tag{18}$$

Setting $\partial\mathcal{L}/\partial w_k = 0$ gives

$$C_k + \tau(1 + \log w_k) + \nu = 0, \tag{19}$$

and therefore

$$w_k = \exp(-C_k/\tau)\,\exp(-\nu/\tau - 1). \tag{20}$$

The simplex constraint determines the normalizer, yielding

$$w_k = \frac{\exp(-C_k/\tau)}{\sum_{j=1}^{K} \exp(-C_j/\tau)}. \tag{21}$$

Substituting $C_k = \alpha\ell_k + \beta\delta_k$ and setting $\tau = 1$ gives $w_k = \mathrm{Softmax}(-\alpha\ell_k - \beta\delta_k)$. Thus the rule down-weights ambiguous or drifted prototypes (high C_k) while entropy regularization prevents exact single-client collapse for any fixed $\tau > 0$. □

No Aggregation Collapse. Theorem D.4 gives the round-wise optimum; the next result shows that this optimum never concentrates on a single client.

Corollary D.5 (Uniform Weight Floor). *Assume $0 \le \ell_k \le \ell_{\max}$ and $0 \le \delta_k \le \delta_{\max}$ for all k, and let $C_{\max} = \max_k C_k$, $C_{\min} = \min_k C_k$. Then the weights of Theorem D.4 satisfy*

$$w_k \;\ge\; w_{\min} := \frac{1}{K}\exp\Big(-\frac{C_{\max} - C_{\min}}{\tau}\Big) \;\ge\; \frac{1}{K}\exp\Big(-\frac{\alpha\ell_{\max} + \beta\delta_{\max}}{\tau}\Big) > 0, \tag{22}$$

and consequently $\max_k w_k \le 1 - (K-1)w_{\min} < 1$ for $K \ge 2$.

Proof. Since $C_j \geq C_{\min}$ for all j, $\sum_j \exp(-C_j/\tau) \leq K \exp(-C_{\min}/\tau)$, hence $w_k \geq \exp(-(C_k - C_{\min})/\tau)/K \geq \exp(-(C_{\max} - C_{\min})/\tau)/K$. Because cross-entropy and squared distances are nonnegative, $C_{\min} \geq 0$ and $C_{\max} \leq \alpha\ell_{\max} + \beta\delta_{\max}$, which gives the second inequality. Finally, $\max_k w_k = 1 - \sum_{j \neq k} w_j \leq 1 - (K-1)w_{\min}$. □

Lemma D.6 (Stability of Softmax Weights). *For* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^K$, $\|\operatorname{Softmax}(\mathbf{a}) - \operatorname{Softmax}(\mathbf{b})\|_1 \leq \operatorname{osc}(\mathbf{a} - \mathbf{b}) \leq 2\|\mathbf{a} - \mathbf{b}\|_\infty$, *where* $\operatorname{osc}(\mathbf{v}) = \max_k v_k - \min_k v_k$.

Proof. Let $\mathbf{v} = \mathbf{a} - \mathbf{b}$ and $\mathbf{w}(s) = \operatorname{Softmax}(\mathbf{b} + s\mathbf{v})$ for $s \in [0,1]$. The softmax Jacobian gives $\frac{d}{ds} w_k(s) = w_k(s)\big(v_k - \bar{v}(s)\big)$ with $\bar{v}(s) = \sum_j w_j(s) v_j \in [\min_j v_j, \max_j v_j]$. Therefore $\|\frac{d}{ds}\mathbf{w}(s)\|_1 = \sum_k w_k(s)|v_k - \bar{v}(s)| \leq \operatorname{osc}(\mathbf{v})$, and integrating over $s \in [0,1]$ yields the claim. □

Convergence with Time-varying DaG Weights. We now analyze the encoder dynamics induced by DaG. Because the weights are recomputed every round, the natural target is the time-varying objective $F_t(\theta) = \sum_k w_k^t F_k(\theta)$, rather than a single static objective. We analyze an encoder-level surrogate in which client k runs R local SGD steps on a fixed local objective F_k: $\theta_k^{t,0} = \theta^t$, $\theta_k^{t,r+1} = \theta_k^{t,r} - \eta\, g_k^{t,r}$, and $\theta^{t+1} = \sum_k w_k^t \theta_k^{t,R}$.

Assumption D.7. (i) Each F_k is L-smooth and $F^{\inf} := \min_k \inf_\theta F_k(\theta) > -\infty$. (ii) $g_k^{t,r} = \nabla F_k(\theta_k^{t,r}) + \varepsilon_k^{t,r}$ with $\mathbb{E}[\varepsilon_k^{t,r} \mid \mathcal{G}_{t,r}] = 0$ and $\mathbb{E}[\|\varepsilon_k^{t,r}\|^2 \mid \mathcal{G}_{t,r}] \leq \sigma^2$, where $\mathcal{G}_{t,r}$ contains all randomness before local step r of round t. (iii) $\|\nabla F_k(\theta) - \nabla F_j(\theta)\| \leq \zeta$ for all k, j, θ. (iv) $\sup_\theta \big(\max_k F_k(\theta) - \min_k F_k(\theta)\big) \leq B$. (v) All clients participate and $\mathbf{w}^t \in \Delta^{K-1}$ is $\mathcal{G}_{t,0}$-measurable, i.e., the weights used in round t are computed from the prototypes of round $t-1$.

Theorem D.8 (Convergence of DaG-weighted Federated Optimization). *Under Assumption D.7, if* $\eta LR \leq 1/8$, *then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F_t(\theta^t)\|^2 \leq \frac{8\Delta_0}{\eta RT} + \frac{4B\,V_T}{\eta RT} + 8L\eta\sigma^2 + 16\eta^2 L^2 R\sigma^2 + 48\eta^2 L^2 R^2 \zeta^2, \tag{23}$$

where $\Delta_0 = \max_k F_k(\theta^0) - F^{\inf}$ *and* $V_T = \sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{w}^{t+1} - \mathbf{w}^t\|_1$. *With* $\eta = 1/(8LR\sqrt{T})$,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F_t(\theta^t)\|^2 \leq \frac{64L\Delta_0 + 32LB\,V_T + \sigma^2/R}{\sqrt{T}} + \frac{3\zeta^2}{4T} + \frac{\sigma^2}{4RT}, \tag{24}$$

which vanishes whenever $V_T = o(\sqrt{T})$. *By Lemma D.6 with* $\tau = 1$, $V_T \leq 2\sum_t \mathbb{E}\big(\alpha\|\boldsymbol{\ell}^{t+1} - \boldsymbol{\ell}^t\|_\infty + \beta\|\boldsymbol{\delta}^{t+1} - \boldsymbol{\delta}^t\|_\infty\big)$, *so the weight-variation term is controlled by how much the reliability scores move.*

Proof. Write $G_t^2 = \|\nabla F_t(\theta^t)\|^2$ and $S_t = \sum_k w_k^t \sum_{r=0}^{R-1} \|\theta_k^{t,r} - \theta^t\|^2$. All expectations below are conditional on $\mathcal{G}_{t,0}$ unless stated otherwise.

Step 1 (client drift). Fix k and let $u_r = \mathbb{E}\|\theta_k^{t,r} - \theta^t\|^2$, $U = \sum_{r=0}^{R-1} u_r$. Since $\theta_k^{t,r} - \theta^t = -\eta\sum_{s<r}(\nabla F_k(\theta_k^{t,s}) + \varepsilon_k^{t,s})$, the Cauchy–Schwarz inequality and the vanishing of cross terms $\mathbb{E}\langle \varepsilon_k^{t,s}, \varepsilon_k^{t,s'}\rangle$ for $s < s'$ (condition on $\mathcal{G}_{t,s'}$) give $u_r \leq 2\eta^2 r \sum_{s<r} \mathbb{E}\|\nabla F_k(\theta_k^{t,s})\|^2 + 2\eta^2 r\sigma^2$. Decomposing $\nabla F_k(\theta_k^{t,s})$ into $[\nabla F_k(\theta_k^{t,s}) - \nabla F_k(\theta^t)] + [\nabla F_k(\theta^t) - \nabla F_t(\theta^t)] + \nabla F_t(\theta^t)$ and using smoothness and $\|\nabla F_k - \sum_j w_j^t \nabla F_j\| \leq \sum_j w_j^t \|\nabla F_k - \nabla F_j\| \leq \zeta$, we get $\mathbb{E}\|\nabla F_k(\theta_k^{t,s})\|^2 \leq 3L^2 u_s + 3\zeta^2 + 3G_t^2$. Summing over $r < R$ with $\sum_{r<R} r \leq R^2/2$ yields $U \leq 3\eta^2 L^2 R^2 U + 3\eta^2 R^3(\zeta^2 + G_t^2) + \eta^2 R^2 \sigma^2$. As $3\eta^2 L^2 R^2 \leq 3/64 \leq 1/2$, $U \leq 6\eta^2 R^3(\zeta^2 + G_t^2) + 2\eta^2 R^2\sigma^2$, and averaging over k with the $\mathcal{G}_{t,0}$-measurable weights gives $\mathbb{E}S_t \leq 6\eta^2 R^3(\zeta^2 + G_t^2) + 2\eta^2 R^2\sigma^2$.

Step 2 (one-round descent). F_t is L-smooth as a convex combination of L-smooth functions. Let $\mathbf{h} = \sum_k w_k^t \sum_r \nabla F_k(\theta_k^{t,r})$ and $\boldsymbol{\varepsilon} = \sum_k w_k^t \sum_r \varepsilon_k^{t,r}$, so $\theta^{t+1} - \theta^t = -\eta(\mathbf{h} + \boldsymbol{\varepsilon})$ and

$$F_t(\theta^{t+1}) \leq F_t(\theta^t) - \eta\langle \nabla F_t(\theta^t), \mathbf{h} + \boldsymbol{\varepsilon}\rangle + \tfrac{L\eta^2}{2}\|\mathbf{h} + \boldsymbol{\varepsilon}\|^2. \tag{25}$$

(a) $\mathbb{E}\langle \nabla F_t(\theta^t), \varepsilon_k^{t,r}\rangle = 0$ by conditioning on $\mathcal{G}_{t,r}$, since θ^t and $\mathbf{w}^t$ are $\mathcal{G}_{t,0}$-measurable. (b) Write $\mathbf{h} = R\nabla F_t(\theta^t) + \mathbf{e}$; Jensen's inequality over $\mathbf{w}^t$, Cauchy–Schwarz over r and smoothness give $\|\mathbf{e}\|^2 \leq$

RL^2S_t, so by Young's inequality $-\eta\langle\nabla F_t(\theta^t), \mathbf{h}\rangle \le -\frac{\eta R}{2}G_t^2 + \frac{\eta L^2}{2}S_t$. (c) $\|\mathbf{h}+\boldsymbol{\varepsilon}\|^2 \le 2\|\mathbf{h}\|^2 + 2\|\boldsymbol{\varepsilon}\|^2$, $\|\mathbf{h}\|^2 \le 2R^2G_t^2 + 2RL^2S_t$, and $\mathbb{E}\|\boldsymbol{\varepsilon}\|^2 = \sum_r \mathbb{E}\|\sum_k w_k^t \varepsilon_k^{t,r}\|^2 \le R\sigma^2$ by orthogonality across steps and Jensen's inequality within a step. Collecting terms and using $\eta LR \le 1/8$ (so $2L\eta^2R^2 \le \eta R/4$ and $2L^3\eta^2 R \le \eta L^2/4$),

$$\mathbb{E}F_t(\theta^{t+1}) \le F_t(\theta^t) - \tfrac{\eta R}{4}G_t^2 + \eta L^2\,\mathbb{E}S_t + L\eta^2 R\sigma^2. \tag{26}$$

Inserting Step 1 and $\eta L^2 \cdot 6\eta^2R^3 = \eta R \cdot 6(\eta LR)^2 \le \frac{6}{64}\eta R$ gives

$$\mathbb{E}F_t(\theta^{t+1}) \le F_t(\theta^t) - \tfrac{\eta R}{8}G_t^2 + 6\eta^3L^2R^3\zeta^2 + 2\eta^3L^2R^2\sigma^2 + L\eta^2R\sigma^2. \tag{27}$$

Step 3 (change of objective). Since $\sum_k(w_k^{t+1} - w_k^t) = 0$, for any θ and any constant c, $F_{t+1}(\theta) - F_t(\theta) = \sum_k(w_k^{t+1} - w_k^t)(F_k(\theta) - c)$. Taking $c = \frac{1}{2}(\max_k F_k(\theta) + \min_k F_k(\theta))$ gives $|F_k(\theta) - c| \le B/2$ by Assumption D.7(iv), hence $F_{t+1}(\theta^{t+1}) \le F_t(\theta^{t+1}) + \frac{B}{2}\|\mathbf{w}^{t+1} - \mathbf{w}^t\|_1$.

Step 4 (telescoping). Taking total expectations, combining Steps 2–3 and summing over $t = 0, \dots, T-1$,

$$\tfrac{\eta R}{8}\sum_{t=0}^{T-1}\mathbb{E}G_t^2 \le F_0(\theta^0) - \mathbb{E}F_T(\theta^T) + \tfrac{B}{2}V_T + T\big(6\eta^3L^2R^3\zeta^2 + 2\eta^3L^2R^2\sigma^2 + L\eta^2R\sigma^2\big). \tag{28}$$

Using $F_0(\theta^0) \le \max_k F_k(\theta^0)$ and $F_T \ge F^{\inf}$ and multiplying by $8/(\eta RT)$ proves the first bound. Substituting $\eta = 1/(8LR\sqrt{T})$, which satisfies $\eta LR \le 1/8$ for all $T \ge 1$, proves the second. □

Remark D.9 (**Scope**). Theorem D.8 is a statement about the aggregation dynamics of DaG given a local SGD oracle. It uses weights computed from the previous round's prototypes (Assumption D.7(v)); in Alg. 1 the weights are computed from the current round, which makes them correlated with that round's gradient noise, and the one-round-lagged variant is the version covered here. The surrogate F_k also treats the local auxiliary modules and the alignment target $\mathbf{p}^g$ as fixed within the analysis and does not cover the saddle-point dynamics of the GRL term. Combined with Corollary D.5, every F_t keeps weight at least $w_{\min}$ on each client, so F_t never degenerates to a single client.

D.3 Global Dynamics as Wasserstein Barycenter

Definition D.10 (Global Prototype Barycenter). Given local prototypes $\{\bar{\mathbf{p}}_k\}_{k=1}^K \subset \mathbb{R}^d$ and nonnegative weights $\{w_k\}_{k=1}^K$ with $\sum_{k=1}^K w_k = 1$, we interpret each prototype as a Dirac measure $\delta_{\bar{\mathbf{p}}_k}$. The global prototype barycenter is the Dirac measure $\delta_{\mathbf{p}^g}$ minimizing the weighted sum of squared 2-Wasserstein distances to these local prototype measures.

Theorem D.11 (Weighted Prototype Mean as a Point-Mass Barycenter). *Under the point-mass prototype interpretation above, the weighted prototype update*

$$\mathbf{p}^g = \sum_{k=1}^K w_k\,\bar{\mathbf{p}}_k \tag{29}$$

is the unique Fréchet mean (equivalently, the unique 2-Wasserstein barycenter) of the local prototype measures $\{\delta_{\bar{\mathbf{p}}_k}\}_{k=1}^K$.

Proof. For Dirac measures in $\mathbb{R}^d$,

$$W_2^2(\delta_{\bar{\mathbf{p}}_k}, \delta_{\mathbf{p}^g}) = \|\bar{\mathbf{p}}_k - \mathbf{p}^g\|_2^2. \tag{30}$$

Therefore the barycenter objective reduces to the finite-dimensional Euclidean problem

$$\min_{\mathbf{p}^g}\ \mathcal{J}(\mathbf{p}^g) = \min_{\mathbf{p}^g}\sum_{k=1}^K w_k\,\|\bar{\mathbf{p}}_k - \mathbf{p}^g\|_2^2. \tag{31}$$

Because $w_k \ge 0$ and $\sum_k w_k = 1$, the Hessian is $2I$, so $\mathcal{J}$ is strictly convex with a unique minimizer. Setting the gradient to zero gives

$$\nabla_{\mathbf{p}^g}\mathcal{J} = 2\sum_{k=1}^K w_k(\mathbf{p}^g - \bar{\mathbf{p}}_k) = 0, \tag{32}$$

which yields $\mathbf{p}^g = \sum_k w_k \bar{\mathbf{p}}_k$. This establishes the barycentric interpretation of the weighted prototype update. The alignment loss in Eq. 4 encourages local prototypes to remain close to the broadcast global prototype, but does not by itself prove convergence of the full nonconvex training dynamics. If Gaussian prototype distributions are considered instead of point masses, the weighted Euclidean average characterizes the barycenter *mean*, while the full Gaussian Wasserstein barycenter additionally depends on covariance and reduces to the same prototype-only expression only when covariance is fixed, common, or intentionally ignored (Villani et al., 2008). □

E Discussion and Additional Results

E.1 Model Scale Discussion

Table 11 compares recent time series foundation models across architecture, scale, tokenization, and training setups. Centralized approaches such as Time-MoE, Moirai, and Chronos demonstrate strong zero-shot performance (**Tables 3** and **4**, **Figs. 4**), validating the effectiveness of large-scale pretraining on aggregated data. Unlike these centralized pipelines, FedTRL is pretrained in a fully federated manner, showing that models with hundreds of millions of parameters can be effectively trained without centralizing data. While comparable in scale to centralized FMs, FedTRL yields flexible yet domain-invariant representations. Beyond matching performance, FedTRL exemplifies a paradigm shift in FM training, demonstrating that FL can serve not only as a privacy-preserving alternative but also as a viable strategy for building large-scale models across distributed domains.

Table 11: Comparison of TSFMs. *Architecture*: Transformer variant. *Model Size*: parameter counts. *Pre-training Scale*: number of pretraining time points. *Token Level*: token granularity. *Tokenization*: embedded values. *Context Length*: maximum input length. *Probabilistic*: whether multiple possible predictions are generated. *Training*: centralized or federated.

Method	**FedTRL (Ours)**	Time-MoE (2024)	Moirai (2024)	MOMENT (2024)	LLMTime (2024)	Chronos (2024)	Lag-Llama (2023)	TimesFM (2023b)
Architecture	Encoder	Decoder	Encoder	Encoder	Decoder	EncDec	Decoder	Decoder
Model Size	302M	113M 453M 2.4B	14M 91M 311M	40M 125M 385M	–	46M 200M 710M	200M	17M 70M 200M
Pre-training Scale	300B	300B	231B	1.13B	–	84B	0.36B	100B
Token Level	Patch	Point	Patch	Patch	Point	Point	Point	Patch
Tokenization	Continuous	Continuous	Continuous	Continuous	Discrete	Discrete	Continuous	Continuous
Context Length	$\leq$3072	$\leq$4096	$\leq$5000	= 512	-	$\leq$512	$\leq$1024	$\leq$512
Probabilistic	True	False	True	False	True	True	True	False
Training	**Federated**	Centralized	Centralized	Centralized	Centralized	Centralized	Centralized	Centralized

Table 12: In-domain MSE (averaged over horizons) of FL aggregation baselines and recent self-supervised objectives. The latter use FedAvg under the same protocol and backbone capacity, with DBPM applied to CoST. **Bold**: the best.

		FL aggregation baselines						Recent self-supervised objectives (FedAvg)			
Dataset	FedTRL	FFTS	FedAvg	FedProx	FedPer	FedRep	pFedMe	CoST+DBPM	REBAR	AutoCon	FEI
ETT-h1	**0.448**	0.463	0.476	0.477	0.492	0.482	0.499	0.472	0.464	0.469	0.461
ETT-m1	**0.375**	0.380	0.399	0.395	0.405	0.401	0.414	0.392	0.390	0.383	0.386
Weather	**0.241**	0.252	0.264	0.262	0.281	0.280	0.287	0.267	0.259	0.262	0.255

E.2 Federated Baselines under Heterogeneity

A wide range of FL algorithms have been developed to address data heterogeneity (Tan et al., 2022), with strong results in CV and NLP. However, our experiments reveal that these methods transfer poorly to time series representation learning. Using the setup in Section 4.1, we evaluate representative approaches (FedProx (Li et al., 2020), FedPer (Arivazhagan et al., 2019), FedRep (Collins et al., 2021), and pFedMe (T Dinh et al., 2020)) on in-domain point forecasting, aggregating final-round local models for fair comparison. As shown in **Table 12**, only FedProx slightly improves over vanilla FedAvg, while others perform worse. We attribute this to fundamental differences between discrete classification benchmarks and continuous time series forecasting. Classification tasks have

low-dimensional outputs where personalization strategies are effective, whereas forecasting requires modeling high-dimensional, continuous targets with complex temporal dependencies. Algorithms designed for discrete label spaces struggle to capture such dynamics, leading to poor generalization. In contrast, FedTRL directly addresses both inter- and intra-client heterogeneity through domain-invariant local objectives and domain-aware aggregation, yielding more robust TSFMs. The recent self-supervised objectives in the right block of **Table 12** improve on the earlier self-supervised baselines of **Table 1**, but none of them addresses intra-client sub-domain conflict or reliability-aware aggregation, and a consistent gap to FedTRL remains.

E.3 In-domain Forecasting

The full results of in-domain forecasting are shown in **Table 13**, where our proposed FedTRL consistently achieves the best performance across nearly all datasets and horizons. Compared to federated baselines, it shows clear error reductions and greater stability, while also rivaling or surpassing centralized training in several cases. These results show that FedTRL not only mitigates heterogeneity more effectively than existing FL methods but also delivers representations strong enough to match centralized pretraining, validating its robustness across forecasting scenarios.

Table 13: Full in-domain forecasting results over horizons $\{96, 192, 336, 720\}$. **Bold**: the best, Underline: the second best. †: FedAvg (McMahan et al., 2017) aggregation.

Methods		Ours		Federated Unsupervised†								Federated FMs				Centralized	
		FedTRL		SimMTM		PatchTST		TimeMAE		CoST		FFTS		FedAvg		All Mixed	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.367**	**0.441**	0.399	0.477	0.391	0.462	0.410	0.485	0.414	0.498	0.387	0.456	0.393	0.462	0.372	0.454
	192	**0.433**	**0.464**	0.466	0.503	0.449	0.482	0.476	0.507	0.500	0.531	0.444	0.481	0.454	0.490	0.437	0.478
	336	**0.467**	**0.482**	0.519	0.523	0.504	0.508	0.531	0.530	0.550	0.555	0.479	0.491	0.495	0.502	0.485	0.500
	720	**0.525**	**0.501**	0.595	0.546	0.569	0.528	0.591	0.549	0.620	0.576	0.542	0.511	0.562	0.527	0.539	0.520
ETTh2	96	**0.317**	**0.389**	0.358	0.434	0.327	0.403	0.369	0.441	0.372	0.451	0.327	0.393	0.338	0.405	0.325	0.393
	192	**0.371**	**0.411**	0.416	0.460	0.397	0.427	0.439	0.468	0.443	0.481	0.383	0.417	0.400	0.433	0.381	0.419
	336	**0.396**	**0.419**	0.469	0.480	0.440	0.442	0.479	0.482	0.499	0.501	0.409	0.424	0.421	0.436	0.418	0.433
	720	**0.445**	**0.437**	0.526	0.497	0.476	0.456	0.532	0.505	0.570	0.526	0.462	0.442	0.481	0.458	0.472	0.452
ETTm1	96	0.312	0.378	0.335	0.410	0.314	0.380	0.337	0.413	0.358	0.426	0.309	0.378	0.321	0.380	**0.309**	**0.374**
	192	**0.351**	0.395	0.396	0.433	0.363	0.402	0.405	0.439	0.408	0.447	0.369	0.403	0.385	0.404	0.352	**0.392**
	336	0.397	0.409	0.443	0.447	0.411	0.418	0.443	0.450	0.453	0.461	0.399	0.413	0.419	0.416	**0.382**	**0.403**
	720	**0.440**	0.426	0.486	0.466	0.456	0.432	0.499	0.474	0.525	0.487	0.442	0.430	0.471	0.424	0.445	**0.423**
ETTm2	96	0.234	0.313	0.261	0.328	0.240	0.320	0.269	0.337	0.278	0.345	0.236	0.318	0.239	0.320	**0.232**	**0.313**
	192	**0.269**	**0.325**	0.313	0.347	0.285	0.336	0.310	0.353	0.330	0.363	0.280	0.333	0.287	0.337	0.272	0.331
	336	**0.292**	**0.338**	0.340	0.359	0.307	0.343	0.350	0.365	0.365	0.375	0.301	0.344	0.312	0.348	0.302	0.343
	720	**0.317**	**0.344**	0.382	0.374	0.353	0.361	0.390	0.384	0.407	0.393	0.343	0.356	0.360	0.359	0.339	0.356
Electricity	96	**0.151**	**0.259**	0.173	0.287	0.155	0.271	0.179	0.293	0.198	0.309	0.153	0.264	0.158	0.271	0.153	0.265
	192	**0.173**	**0.269**	0.209	0.302	0.185	0.282	0.213	0.309	0.229	0.323	0.178	0.275	0.188	0.283	0.182	0.283
	336	**0.189**	**0.274**	0.230	0.313	0.202	0.295	0.240	0.323	0.255	0.336	0.195	0.283	0.209	0.296	0.200	0.291
	720	**0.211**	**0.287**	0.251	0.322	0.225	0.304	0.268	0.335	0.286	0.349	0.221	0.298	0.237	0.322	0.221	0.301
Traffic	96	0.355	0.272	0.384	0.293	0.354	0.275	0.395	0.300	0.394	0.310	**0.339**	**0.265**	0.343	0.269	0.352	0.272
	192	0.408	0.283	0.453	0.310	0.422	0.291	0.462	0.316	0.478	0.331	**0.395**	**0.281**	0.405	0.288	0.414	0.285
	336	0.451	0.295	0.505	0.324	0.461	0.300	0.511	0.329	0.533	0.346	**0.444**	**0.294**	0.460	0.304	0.452	0.296
	720	**0.491**	0.303	0.547	0.334	0.516	0.314	0.572	0.343	0.588	0.353	0.502	**0.301**	0.528	0.316	0.495	0.304
Weather	96	**0.201**	**0.266**	0.229	0.288	0.208	0.277	0.237	0.289	0.248	0.303	0.204	0.271	0.228	0.285	0.210	0.277
	192	**0.231**	**0.279**	0.276	0.308	0.252	0.293	0.280	0.310	0.290	0.318	0.244	0.288	0.252	0.295	0.239	0.286
	336	**0.254**	**0.286**	0.302	0.319	0.277	0.301	0.310	0.317	0.316	0.326	0.262	0.296	0.270	0.304	0.266	0.297
	720	**0.277**	**0.297**	0.333	0.325	0.303	0.309	0.353	0.332	0.365	0.342	0.298	0.305	0.306	0.312	0.301	0.308
Exchange	96	**0.311**	**0.397**	0.376	0.449	0.326	0.414	0.374	0.452	0.385	0.459	0.319	0.410	0.332	0.414	0.322	0.410
	192	**0.358**	**0.419**	0.442	0.474	0.388	0.439	0.449	0.485	0.446	0.483	0.367	0.432	0.380	0.436	0.367	0.430
	336	**0.393**	**0.429**	0.477	0.485	0.431	0.458	0.493	0.499	0.502	0.503	0.409	0.446	0.430	0.452	0.413	0.448
	720	**0.438**	**0.446**	0.545	0.512	0.475	0.470	0.556	0.523	0.555	0.523	0.465	0.472	0.470	0.476	0.458	0.464
1st Count		**49**		0		0		0		0		7		0		8	

E.4 Zero-shot Point Forecasting

We evaluate our FedTRL-trained model on two zero-shot point forecasting benchmarks: TSLib (5 datasets) and RW-Bench (15 datasets), following the evaluation protocol of (Shi et al., 2024). The full results are reported in Table 14 (TSLib) and Table 15 (RW-Bench). Across both benchmarks and

multiple prediction horizons, FedTRL consistently achieves state-of-the-art performance, surpassing centralized foundation models trained on aggregated data. These results further demonstrate the effectiveness of FedTRL.

Table 14: Full zero-shot forecasting results (lower is better). TimesFM is not evaluated on Weather (−), which is in its pretraining data. **Bold**: the best, Underline: the second best.

Models		Ours		Pretrained Time Series Foundation Models (Zero-shot)																					
		FedTRL		Time-MoE$_{base}$		Time-MoE$_{large}$		Time-MoE$_{ultra}$		Moirai$_{small}$		Moirai$_{base}$		Moirai$_{large}$		TimesFM		Moment		Chronos$_{small}$		Chronos$_{base}$		Chronos$_{large}$	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	96	**0.346**	0.381	0.357	0.381	0.350	0.382	0.349	**0.379**	0.401	0.402	0.376	0.392	0.349	**0.379**	0.414	0.404	0.688	0.557	0.466	0.409	0.440	0.393	0.441	0.390
	192	0.386	0.407	**0.384**	**0.404**	0.388	0.412	0.395	0.413	0.388	0.412	0.412	0.413	0.434	0.415	0.465	0.434	0.688	0.560	0.530	0.450	0.492	0.426	0.502	0.424
	336	0.420	0.443	**0.411**	0.434	**0.411**	0.430	0.447	0.453	0.433	**0.428**	0.433	**0.428**	0.495	0.445	0.503	0.456	0.675	0.563	0.570	0.486	0.550	0.462	0.576	0.467
	720	0.444	0.449	0.449	0.477	**0.427**	0.455	0.457	0.462	0.439	0.454	0.447	**0.444**	0.611	0.510	0.511	0.481	0.683	0.585	0.615	0.543	0.882	0.591	0.835	0.583
	Avg.	0.399	0.420	0.400	0.424	**0.394**	**0.419**	0.412	0.426	0.428	0.427	0.417	**0.419**	0.480	0.439	0.473	0.443	0.683	0.566	0.545	0.472	0.591	0.468	0.588	0.466
ETTh2	96	0.300	0.352	0.305	0.359	0.302	0.354	**0.292**	0.352	0.297	0.336	0.294	**0.330**	0.296	**0.330**	0.315	0.349	0.342	0.396	0.307	0.356	0.308	0.343	0.320	0.345
	192	**0.335**	**0.369**	0.351	0.386	0.364	0.385	0.347	0.379	0.368	0.381	0.365	0.375	0.361	0.371	0.388	0.395	0.354	0.402	0.376	0.401	0.384	0.392	0.406	0.399
	336	0.371	0.392	0.391	0.418	0.417	0.425	0.406	0.419	0.370	0.393	0.376	**0.390**	0.390	**0.390**	0.422	0.427	**0.356**	0.407	0.408	0.431	0.429	0.430	0.492	0.453
	720	**0.394**	**0.407**	0.419	0.454	0.537	0.496	0.439	0.447	0.411	0.426	0.416	0.433	0.423	0.418	0.443	0.454	0.395	0.434	0.604	0.533	0.501	0.477	0.603	0.511
	Avg.	**0.350**	0.380	0.366	0.404	0.405	0.415	0.371	0.399	0.361	0.384	0.362	0.382	0.367	**0.377**	0.392	0.406	0.361	0.409	0.424	0.430	0.405	0.410	0.455	0.427
ETTm1	96	0.284	0.342	0.338	0.368	0.309	0.557	**0.281**	**0.341**	0.418	0.392	0.363	0.356	0.380	0.361	0.361	0.370	0.654	0.527	0.511	0.423	0.454	0.408	0.457	0.403
	192	0.319	0.366	0.353	0.388	0.346	0.381	**0.305**	**0.358**	0.431	0.405	0.388	0.375	0.412	0.383	0.414	0.405	0.662	0.532	0.618	0.485	0.567	0.477	0.530	0.450
	336	**0.358**	0.405	0.381	0.413	0.373	0.408	0.369	0.395	0.433	0.412	0.416	**0.392**	0.436	0.400	0.445	0.429	0.672	0.537	0.683	0.524	0.662	0.525	0.577	0.481
	720	**0.435**	0.419	0.504	0.493	0.475	0.477	0.469	0.472	0.462	0.432	0.460	**0.418**	0.462	0.420	0.512	0.471	0.692	0.551	0.748	0.566	0.900	0.591	0.660	0.526
	Avg.	**0.349**	**0.383**	0.394	0.415	0.376	0.405	0.356	0.391	0.436	0.410	0.406	0.385	0.422	0.391	0.433	0.418	0.670	0.536	0.640	0.499	0.645	0.500	0.555	0.465
ETTm2	96	0.205	0.279	0.201	0.291	**0.197**	0.286	0.198	0.288	0.214	0.288	0.205	0.273	0.211	0.274	0.202	**0.270**	0.260	0.335	0.209	0.291	0.199	0.274	**0.197**	0.271
	192	0.245	0.321	0.258	0.334	0.250	0.322	**0.235**	**0.312**	0.284	0.332	0.275	0.316	0.281	0.318	0.289	0.321	0.289	0.350	0.280	0.341	0.261	0.322	0.254	0.314
	336	0.311	**0.343**	0.324	0.373	0.337	0.375	**0.293**	0.348	0.331	0.362	0.329	0.350	0.341	0.355	0.360	0.366	0.324	0.369	0.354	0.390	0.326	0.366	0.313	0.353
	720	**0.367**	**0.377**	0.488	0.464	0.480	0.461	0.427	0.423	0.402	0.408	0.437	0.411	0.485	0.428	0.462	0.430	0.394	0.409	0.553	0.499	0.455	0.439	0.416	0.415
	Avg.	**0.282**	**0.330**	0.317	0.365	0.316	0.361	0.288	0.344	0.307	0.347	0.311	0.337	0.329	0.343	0.328	0.346	0.316	0.365	0.349	0.380	0.310	0.350	0.295	0.338
Weather	96	**0.149**	**0.210**	0.160	0.214	0.159	0.213	0.157	0.211	0.198	0.222	0.220	0.217	0.199	0.211	–	–	0.243	0.255	0.211	0.243	0.203	0.238	0.194	0.235
	192	**0.188**	**0.240**	0.210	0.260	0.215	0.266	0.208	0.256	0.247	0.265	0.271	0.259	0.246	0.251	–	–	0.278	0.329	0.263	0.294	0.256	0.290	0.249	0.285
	336	**0.249**	**0.282**	0.274	0.309	0.291	0.322	0.255	0.290	0.283	0.303	0.286	0.297	0.274	0.291	–	–	0.306	0.346	0.321	0.339	0.314	0.336	0.302	0.327
	720	0.366	0.384	0.418	0.405	0.415	0.400	0.405	0.397	0.373	0.354	0.373	0.354	**0.337**	**0.340**	–	–	0.350	0.374	0.404	0.397	0.397	0.396	0.372	0.378
	Avg.	**0.238**	0.279	0.265	0.297	0.270	0.300	0.256	0.288	0.275	0.286	0.287	0.281	0.264	**0.273**	–	–	0.294	0.326	0.300	0.318	0.292	0.315	0.279	0.306
1st Count		22		3		5		9		1		7		7		1		1		0		0		1	

Comparison with Most Recent TSFMs. We additionally compare FedTRL against LightGTS (Wang et al., 2025b), VisionTS (Chen et al., 2024), Sundial 94B/230B/1032B (Liu et al., 2025b), Chronos-2 (Ansari et al., 2025), Moirai-MoE (Liu et al., 2025a), and the federated FeDaL (Chen et al., 2026) under the same zero-shot protocol as Table 3. As shown in Table 16, FedTRL remains competitive with the strongest centralized models, closely followed by LightGTS on a few datasets and Sundial$_{1032B}$ on long-horizon settings, while Chronos-2 leads on ETTm2 and Weather. FedTRL is best on ETTh1 and RW-Bench and outperforms the federated FeDaL on all six benchmarks. These results are obtained without centralizing raw data.

E.5 Zero-shot Probabilistic Forecasting

We evaluate our FedTRL-trained model on GIFT-Eval (Aksu et al., 2024), a benchmark designed to comprehensively assess forecasting across diverse time series. It contains 23 datasets with 144,000 series and 177 million data points, covering 97 forecasting configurations. Following the official evaluation suite, we report aggregated results in Table 4. We further assess zero-shot forecasting and inference efficiency on FEV leaderboard (Ansari et al., 2024), established by AutoGluon, which includes 27 datasets. Aggregated forecasting metrics are presented in Fig. 4.

E.6 Full-shot Comparison with Fine-tuned TSFMs

TimesFM and Moment in **Table 3** are already run with the fine-tuning their protocols require. **Table 17** further compares FedTRL with MOMENT (Goswami et al., 2024), Timer (Liu et al., 2024b), SymTime (Wang et al., 2026), and GTM (He et al., 2026) in the full-shot setting under their original configurations, where FedTRL leads every model on all three datasets. The controlled centralized comparison is All Mixed in **Table 1**, which pools the same data and trains the same backbone under the same protocol. The last row of **Table 17** re-runs the one-epoch adaptation of the single pretrained checkpoint under three seeds, isolating the randomness introduced after pretraining. The advantage on ETTh1, ETTm1, and Weather is substantially larger than the observed spread, while on ETTh2 and ETTm2 the difference to the best supervised model is comparable to it.

Table 15: Full results of real-world weather benchmark (RW-bench).

Models		Ours		Pretrained Time Series Foundation Models (Zero-shot)																			
		FedTRL		Time-MoE$_{base}$		Time-MoE$_{large}$		Moirai$_{small}$		Moirai$_{base}$		Moirai$_{large}$		TimesFM		Moment		Chronos$_{small}$		Chronos$_{base}$		Chronos$_{large}$	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
RW1	96	0.508	**0.453**	0.510	0.460	**0.505**	0.455	0.885	0.547	0.879	0.529	0.986	0.530	0.883	0.703	0.674	0.594	0.850	0.581	0.879	0.589	0.873	0.583
	192	**0.528**	**0.473**	0.540	0.492	0.532	0.485	0.950	0.569	0.896	0.546	1.011	0.548	0.903	0.711	0.717	0.600	0.917	0.620	0.979	0.634	0.950	0.623
	336	**0.546**	**0.488**	0.566	0.519	0.556	0.505	1.148	0.588	0.813	0.552	0.868	0.560	0.889	0.708	0.913	0.700	0.957	0.649	1.047	0.667	0.973	0.646
	720	**0.552**	**0.497**	0.643	0.580	0.568	0.519	1.869	0.667	0.958	0.602	1.259	0.625	0.877	0.705	1.740	1.019	1.089	0.708	1.142	0.715	1.012	0.687
	Avg	**0.534**	**0.477**	0.565	0.513	0.540	0.491	1.213	0.593	0.887	0.557	1.031	0.566	0.888	0.707	1.011	0.728	0.953	0.639	1.012	0.651	0.952	0.635
RW2	96	0.433	**0.420**	**0.431**	0.426	0.436	0.424	1.541	0.526	1.314	0.483	1.877	0.491	0.846	0.702	0.677	0.567	0.778	0.543	0.779	0.546	0.811	0.553
	192	**0.460**	**0.445**	**0.460**	0.455	0.470	0.456	1.407	0.560	1.021	0.504	2.416	0.529	0.841	0.700	0.788	0.623	0.855	0.587	0.874	0.595	0.895	0.598
	336	**0.481**	**0.459**	0.497	0.487	0.497	0.477	1.666	0.596	1.168	0.529	1.910	0.553	0.842	0.700	0.881	0.653	0.943	0.625	0.942	0.629	0.960	0.630
	720	**0.505**	**0.481**	0.575	0.545	0.539	0.505	2.236	0.660	2.600	0.603	4.302	0.637	0.847	0.702	1.933	1.044	1.146	0.691	1.054	0.681	1.049	0.677
	Avg	**0.470**	**0.451**	0.491	0.479	0.485	0.465	1.713	0.585	1.526	0.530	2.626	0.552	0.844	0.701	1.070	0.721	0.931	0.612	0.912	0.613	0.929	0.614
RW3	96	0.580	0.485	**0.566**	**0.474**	0.579	0.485	1.461	0.616	1.254	0.576	1.511	0.561	0.952	0.728	0.726	0.598	0.977	0.614	0.980	0.625	0.943	0.600
	192	**0.594**	0.501	**0.594**	**0.498**	0.597	0.508	1.224	0.610	0.991	0.577	1.450	0.571	0.963	0.730	0.808	0.629	1.032	0.649	1.058	0.668	1.029	0.640
	336	**0.605**	**0.511**	0.620	0.522	0.608	0.519	1.578	0.629	0.900	0.583	1.053	0.587	0.973	0.735	0.908	0.674	1.084	0.675	1.121	0.696	1.054	0.658
	720	**0.617**	**0.529**	0.681	0.576	0.631	0.544	2.740	0.709	1.233	0.650	1.755	0.653	0.953	0.732	2.197	1.137	1.181	0.725	1.218	0.741	1.165	0.729
	Avg	**0.599**	**0.506**	0.615	0.517	0.604	0.514	1.751	0.641	1.094	0.597	1.442	0.593	0.960	0.731	1.159	0.760	1.068	0.666	1.094	0.682	1.048	0.657
RW4	96	**0.501**	0.484	0.512	0.486	0.502	**0.483**	1.634	0.603	1.313	0.562	1.808	0.571	0.895	0.728	0.707	0.588	0.891	0.629	0.870	0.628	0.889	0.630
	192	**0.517**	**0.501**	0.535	0.510	0.522	0.505	1.299	0.606	1.104	0.571	1.651	0.592	0.886	0.726	0.776	0.619	0.951	0.663	0.959	0.671	0.961	0.669
	336	**0.534**	**0.512**	0.557	0.534	0.544	0.522	1.457	0.629	0.931	0.575	1.174	0.610	0.893	0.727	0.862	0.667	1.004	0.690	1.025	0.701	1.013	0.695
	720	**0.558**	**0.530**	0.616	0.582	0.582	0.547	2.364	0.703	1.576	0.644	2.106	0.683	0.888	0.726	2.205	1.126	1.169	0.747	1.170	0.757	1.089	0.740
	Avg	**0.527**	**0.507**	0.555	0.528	0.538	0.514	1.688	0.635	1.231	0.588	1.685	0.614	0.890	0.727	1.138	0.750	1.004	0.682	1.006	0.689	0.988	0.684
RW5	96	**0.489**	0.469	0.491	0.470	0.498	**0.468**	1.002	0.571	0.911	0.541	0.941	0.533	0.900	0.732	0.707	0.600	0.897	0.622	0.911	0.626	0.924	0.627
	192	**0.509**	**0.490**	0.513	0.494	0.514	0.491	1.049	0.588	0.815	0.547	0.863	0.551	0.912	0.735	0.759	0.617	0.973	0.661	0.998	0.672	0.991	0.668
	336	**0.517**	**0.499**	0.538	0.516	0.529	0.507	1.503	0.614	0.805	0.555	1.040	0.577	0.905	0.732	1.013	0.727	1.039	0.695	1.071	0.706	1.053	0.696
	720	**0.526**	**0.510**	0.613	0.570	0.541	0.524	2.697	0.687	1.226	0.616	1.825	0.653	0.910	0.735	1.741	1.008	1.155	0.748	1.199	0.755	1.166	0.738
	Avg	**0.510**	**0.492**	0.539	0.512	0.520	0.497	1.563	0.615	0.939	0.565	1.167	0.578	0.907	0.733	1.055	0.738	1.016	0.682	1.045	0.690	1.033	0.683
RW6	96	0.786	**0.499**	0.796	0.516	0.785	**0.499**	1.442	0.596	1.169	0.561	1.555	0.566	1.290	0.778	**0.692**	0.597	1.365	0.656	1.388	0.660	1.349	0.660
	192	0.808	**0.517**	0.808	0.531	0.809	0.521	1.228	0.597	1.021	0.565	1.706	0.586	1.299	0.779	**0.754**	0.627	1.462	0.692	1.533	0.704	1.463	0.699
	336	**0.821**	**0.525**	0.835	0.549	0.828	0.537	1.733	0.623	1.126	0.585	1.456	0.610	1.301	0.780	0.860	0.667	1.524	0.720	1.630	0.737	1.552	0.727
	720	**0.831**	**0.539**	0.897	0.599	0.837	0.552	2.410	0.694	1.877	0.657	2.881	0.690	1.290	0.780	1.703	0.982	1.627	0.772	1.752	0.786	1.670	0.766
	Avg	**0.812**	**0.520**	0.834	0.549	0.815	0.527	1.703	0.628	1.298	0.592	1.900	0.613	1.295	0.779	1.002	0.718	1.495	0.710	1.576	0.722	1.508	0.713
RW7	96	**0.479**	**0.478**	0.492	0.481	0.493	0.479	0.703	0.550	0.695	0.541	0.660	0.531	0.887	0.726	0.728	0.615	0.879	0.625	0.897	0.626	0.888	0.626
	192	**0.503**	0.504	0.514	**0.503**	0.514	0.504	0.712	0.561	0.686	0.549	0.655	0.553	0.891	0.727	0.817	0.634	0.939	0.662	0.981	0.670	0.953	0.665
	336	**0.518**	**0.520**	0.536	0.525	0.527	0.522	0.791	0.582	0.696	0.567	0.661	0.566	0.885	0.725	0.899	0.696	0.981	0.687	1.028	0.697	1.001	0.691
	720	**0.526**	**0.532**	0.582	0.564	0.537	0.533	1.462	0.643	0.733	0.610	0.868	0.615	0.880	0.724	1.961	1.071	1.081	0.731	1.094	0.735	1.021	0.722
	Avg	**0.506**	**0.508**	0.531	0.518	0.518	0.509	0.917	0.584	0.703	0.567	0.711	0.566	0.886	0.725	1.101	0.754	0.970	0.676	1.000	0.682	0.966	0.676
RW8	96	0.520	0.468	0.525	0.471	0.527	0.466	0.853	0.552	0.828	0.525	0.881	0.531	1.035	0.748	0.786	0.643	0.858	0.598	0.929	0.622	0.910	0.615
	192	0.539	0.492	**0.538**	**0.489**	0.542	**0.489**	0.769	0.563	0.773	0.539	0.736	0.546	1.019	0.746	0.920	0.701	0.925	0.638	1.058	0.679	1.012	0.665
	336	**0.545**	0.503	0.557	0.509	0.547	**0.502**	0.800	0.581	0.805	0.557	0.718	0.559	1.018	0.748	1.026	0.752	0.986	0.668	1.170	0.723	1.119	0.702
	720	**0.560**	**0.516**	0.621	0.556	0.573	0.527	1.054	0.631	0.818	0.589	0.937	0.614	1.018	0.749	2.017	1.111	1.147	0.718	1.356	0.789	1.298	0.777
	Avg	**0.541**	**0.495**	0.560	0.506	0.547	0.496	0.869	0.582	0.806	0.553	0.818	0.563	1.022	0.748	1.187	0.802	0.979	0.655	1.128	0.703	1.085	0.690
RW9	96	**0.576**	0.479	**0.576**	**0.476**	0.585	0.482	1.143	0.578	1.005	0.557	1.304	0.559	1.012	0.720	0.791	0.645	0.939	0.605	1.062	0.636	1.041	0.630
	192	0.622	0.513	**0.613**	**0.508**	0.629	0.517	1.124	0.607	0.996	0.582	1.066	0.586	1.025	0.725	0.913	0.685	1.039	0.654	1.228	0.703	1.178	0.695
	336	**0.644**	**0.528**	0.648	0.532	0.648	0.529	1.121	0.623	0.966	0.596	1.061	0.618	1.036	0.728	1.091	0.791	1.106	0.686	1.357	0.748	1.321	0.738
	720	**0.658**	**0.544**	0.707	0.573	0.669	0.552	1.313	0.661	1.029	0.634	1.557	0.704	1.037	0.731	2.049	1.133	1.221	0.734	1.518	0.802	1.500	0.778
	Avg	**0.625**	**0.516**	0.636	0.522	0.633	0.520	1.175	0.617	0.999	0.592	1.247	0.617	1.028	0.726	1.211	0.813	1.076	0.670	1.291	0.722	1.260	0.710
RW10	96	**0.566**	**0.505**	0.585	0.518	0.579	**0.505**	0.923	0.607	0.926	0.591	0.836	0.578	0.989	0.769	0.752	0.626	1.014	0.657	1.073	0.661	1.016	0.651
	192	**0.599**	**0.537**	0.612	0.546	0.610	**0.537**	0.935	0.629	0.949	0.617	0.874	0.615	0.986	0.768	0.777	0.638	1.087	0.703	1.144	0.708	1.098	0.698
	336	**0.616**	**0.553**	0.626	0.566	0.621	0.554	0.907	0.637	0.972	0.633	0.864	0.634	0.976	0.768	0.852	0.678	1.117	0.728	1.173	0.730	1.120	0.720
	720	**0.618**	**0.564**	0.655	0.600	0.629	**0.564**	1.119	0.682	1.063	0.680	1.070	0.690	0.968	0.767	1.872	1.069	1.206	0.774	1.260	0.774	1.186	0.755
	Avg	**0.600**	**0.540**	0.620	0.558	0.609	**0.540**	0.971	0.639	0.977	0.630	0.911	0.629	0.980	0.768	1.063	0.753	1.106	0.716	1.162	0.718	1.105	0.706
RW11	96	**0.529**	**0.487**	0.556	0.498	0.546	0.495	0.975	0.601	0.929	0.593	0.967	0.583	0.901	0.740	0.707	0.623	0.906	0.601	0.956	0.608	0.906	0.596
	192	**0.586**	0.539	0.588	**0.535**	0.596	0.542	1.062	0.642	0.999	0.630	0.936	0.619	0.898	0.738	0.726	0.625	0.999	0.656	1.108	0.670	1.014	0.651
	336	0.628	0.567	**0.614**	**0.559**	0.639	0.571	0.916	0.635	0.900	0.643	0.846	0.635	0.906	0.740	0.985	0.765	1.089	0.690	1.146	0.698	1.070	0.679
	720	**0.655**	0.586	0.673	0.600	**0.655**	**0.585**	0.931	0.669	0.915	0.675	0.920	0.675	0.908	0.739	1.685	1.023	1.264	0.738	1.267	0.746	1.328	0.712
	Avg	**0.599**	**0.545**	0.608	0.548	0.609	0.548	0.971	0.637	0.936	0.635	0.917	0.628	0.903	0.739	1.026	0.759	1.065	0.671	1.119	0.681	1.080	0.659
RW12	96	**0.708**	**0.554**	0.751	0.569	0.726	0.556	1.124	0.661	1.168	0.657	1.134	0.641	1.140	0.798	0.782	0.654	1.196	0.693	1.227	0.684	1.191	0.682
	192	**0.711**	**0.605**	0.791	**0.605**	0.781	0.606	1.285	0.707	1.206	0.690	1.183	0.688	1.149	0.802	0.836	0.650	1.331	0.754	1.415	0.755	1.342	0.746
	336	**0.811**	**0.629**	0.821	0.631	0.827	0.637	1.544	0.720	1.252	0.720	1.236	0.722	1.146	0.800	1.011	0.755	1.423	0.792	1.502	0.793	1.454	0.783
	720	**0.838**	**0.650**	0.880	0.668	0.858	0.657	2.417	0.780	1.342	0.768	1.529	0.790	1.158	0.804	1.910	1.092	1.589	0.842	1.662	0.846	1.572	0.827
	Avg	**0.782**	**0.610**	0.811	0.618	0.798	0.614	1.592	0.717	1.242	0.709	1.270	0.710	1.148	0.801	1.135	0.788	1.385	0.770	1.451	0.770	1.390	0.759
RW13	96	**0.584**	**0.493**	0.601	0.509	0.597	0.499	0.969	0.593	0.832	0.571	0.845	0.570	1.047	0.773	0.736	0.623	1.057	0.657	1.081	0.650	1.075	0.650
	192	**0.619**	**0.522**	0.643	0.541	0.622	0.523	1.055	0.616	0.908	0.601	0.944	0.609	1.045	0.770	0.781	0.636	1.149	0.706	1.202	0.701	1.151	0.693
	336	**0.644**	**0.539**	0.668	0.564	0.648	0.542	0.986	0.632	0.901	0.617	0.914	0.627	1.048	0.771	0.901	0.688	1.232	0.740	1.248	0.732	1.195	0.723
	720	**0.657**	**0.556**	0.725	0.613	0.660	**0.556**	1.290	0.689	0.921	0.654	1.021	0.666	1.035	0.771	1.925	1.065	1.329	0.793	1.299	0.776	1.243	0.755
	Avg	**0.626**	**0.528**	0.659	0.557	0.632	0.530	1.075	0.633	0.891	0.611	0.931	0.618	1.044	0.771	1.086	0.753	1.192	0.724	1.208	0.715	1.166	0.705
RW14	96	**0.648**	**0.520**	0.666	0.532	0.653	0.525	1.144	0.620	1.044	0.602	1.111	0.597	1.107	0.782	0.743	0.630	1.118	0.667	1.166	0.674	1.129	0.669
	192	**0.682**	**0.550**	0.696	0.560	0.688	0.557	1.211	0.646	0.989	0.615	1.026	0.623	1.098	0.777	0.819	0.649	1.221	0.713	1.281	0.727	1.256	0.721
	336	**0.714**	**0.573**	0.722	0.581	0.720	0.580	1.358	0.674	1.029	0.637	1.080	0.648	1.109	0.780	0.883	0.688	1.277	0.742	1.401	0.766	1.314	0.752
	720	**0.724**	**0.587**	0.789	0.628	0.734	0.592	1.793	0.732	1.096	0.689	1.219	0.697	1.099	0.778	1.891	1.067	1.359	0.786	1.459	0.805	1.389	0.783
	Avg	**0.692**	**0.557**	0.718	0.575	0.699	0.563	1.376	0.668	1.039	0.636	1.109	0.641	1.103	0.780	1.084	0.759	1.244	0.727	1.327	0.743	1.272	0.731
RW15	96	0.508	0.484	**0.505**	**0.483**	0.512	0.488	0.891	0.565	0.797	0.559	0.765	0.540	0.972	0.760	0.740	0.623	0.843	0.587	0.861	0.600	0.836	0.588
	192	0.559	0.524	**0.547**	**0.519**	0.559	0.529	1.051	0.613	0.837	0.599	0.758	0.584	0.962	0.758	0.751	0.632	0.932	0.635	0.980	0.659	0.933	0.639
	336	**0.583**	**0.543**	0.595	0.555	0.594	0.555	1.329	0.638	0.852	0.615	0.886	0.616	0.973	0.762	0.852	0.689	0.999	0.664	1.045	0.693	0.998	0.670
	720	**0.616**	**0.568**	0.672	0.608	0.641	0.585	2.119	0.704	0.967	0.663	1.145	0.682	0.976	0.762	1.787	1.044	1.120	0.715	1.169	0.745	1.101	0.740
	Avg	**0.566**	**0.530**	0.580	0.541	0.576	0.539	1.348	0.630	0.863	0.609	0.888	0.606	0.971	0.760	1.033	0.747	0.974	0.650	1.014	0.675	0.967	0.659
1st Count		**124**		21		12		0		0		0		0		2		0		0		0	

Table 16: Zero-shot comparison with recent TSFMs (MSE averaged over horizons $\{96, 192, 336, 720\}$ with look-backs $\{512, 1024, 2048, 3072\}$). All models use the same inputs, and probabilistic models report the median forecast. **Bold**: the best per row. "–": not reported. 1^{st} Count covers the five TSLib datasets.

Dataset/Model	FedTRL (Ours)	LightGTS	VisionTS	Sundial$_{94B}$	Sundial$_{230B}$	Sundial$_{1032B}$	Chronos-2	Moirai-MoE	FeDaL
ETTh1	**0.399**	0.401	0.412	0.402	0.403	0.411	0.428	0.411	0.407
ETTh2	0.350	0.362	0.365	0.377	0.364	**0.333**	0.358	0.359	0.361
ETTm1	0.349	0.345	0.391	0.367	0.352	**0.336**	0.397	0.357	0.360
ETTm2	0.282	**0.249**	0.296	0.280	0.273	0.258	0.264	0.286	0.292
Weather	0.238	**0.219**	0.303	0.254	0.252	0.234	0.234	0.249	0.255
RW-Bench	**0.599**	–	–	–	–	–	0.612	0.703	0.638
1^{st} Count	1	**2**	0	0	0	**2**	0	0	0

Table 17: Full-shot comparison with fine-tuned centralized TSFMs and seed variability of FedTRL (MSE averaged over horizons). Timer uses look-back 672, others 512. Best supervised: best model in Table 2. Last row: one-epoch adaptation of the same checkpoint over three seeds (mean $\pm$ std). "–": not evaluated.

Model	ETTh1	ETTh2	ETTm1	ETTm2	Weather
MOMENT (Goswami et al., 2024)	0.418	–	0.344	–	0.228
Timer (Liu et al., 2024b)	0.418	–	0.352	–	0.232
SymTime (Wang et al., 2026)	0.424	–	0.376	–	0.249
GTM (He et al., 2026)	0.441	–	0.357	–	0.238
Best supervised	0.413	**0.330**	0.351	**0.255**	0.227
FedTRL	**0.372**	–	**0.316**	–	**0.214**
FedTRL (three seeds)	0.374$\pm$0.0044	0.331$\pm$0.0021	0.315$\pm$0.0026	0.258$\pm$0.0023	0.215$\pm$0.0015

E.7 Scalability

We evaluate the scalability of FedTRL by training models of different sizes, with configurations summarized in Table 18 and results reported in Fig. 6. We further assess scalability with respect to dataset size by controlling the amount of data used for pretraining across {90B, 120B, 180B, 300B, 450B, 540B}, where the 450B and 540B configurations are augmented using ERA5 daily, weekly, and monthly data. We evaluate the resulting TSFMs on full-shot, zero-shot, and zero-shot probabilistic forecasting tasks. As shown in **Table 19**, performance consistently improves with larger pretraining datasets, demonstrating the positive scaling behavior of FedTRL.

Table 18: Model configurations of scalability exploration, where gray shading is the default setting.

Patch Size (P)	Context Length (T)	Prediction Length (F)	Layers (L)	Dimension (D, D_{ff})	Heads H	Total Parameters #Count
16	3072	$\{16, 720\}$	12	$(512, 2048)$	8	38M
16	3072	$\{16, 720\}$	16	$(768, 3072)$	12	114M
16	3072	$\{16, 720\}$	24	$(1024, 4096)$	16	302M

E.8 Computational and Communication Efficiency

We summarize FedTRL's computational and communication overhead relative to standard local objectives and federated aggregation, with numbers measured under the configurations of Table 10.

Per-step Computation. Compared to PatchTST or SimMTM as the local self-supervised objective under FedAvg, FedTRL's diffusion-based reconstruction together with the local sub-domain adversarial regularizer adds approximately $1.4\times$ training time per local step (e.g., $0.42\,\text{s} \rightarrow 0.59\,\text{s}$ on the in-domain pretraining setting) and $1.2\times$ peak GPU memory. Under large-scale TSFM pretraining on a single A100-80G, the per-step time grows from $\sim 0.31\,\text{s}$ to $\sim 0.43\,\text{s}$, with memory usage rising by $\sim 1.19\times$. The overhead is dominated by the diffusion forward/denoising pass, while the GRL and

Table 19: Full/Zero-shot point forecasting (MSE report) and zero-shot probabilistic forecasting (Avg. MASE report) results across different pretraining datasets scales.

Task / Report	Datasets / Metric	Pretraining Datasets Scale (Billion)					
		90B	120B	180B	300B (Original)	450B	540B
Full-shot Forecasting							
MSE Report	ETTh1	0.401	0.387	0.372	0.371	0.363	**0.363**
	ETTm1	0.343	0.336	0.328	0.316	0.312	**0.310**
	Weather	0.232	0.231	0.227	0.214	0.204	**0.202**
Zero-shot Point Forecasting							
MSE Report	ETTh1	0.428	0.432	0.411	0.399	0.400	**0.393**
	ETTm1	0.402	0.371	0.376	0.350	0.344	**0.341**
	Weather	0.275	0.273	0.269	0.238	0.214	**0.208**
	RW-Bench	0.689	0.647	0.641	0.599	0.581	**0.575**
Zero-shot Probabilistic							
Ave. MASE Report	GIFT-eval	0.770	0.744	0.721	0.675	0.677	**0.669**
	FEV Leaderboard	0.945	0.873	0.847	0.836	0.833	**0.829**

prototype-alignment terms are essentially free in compute. **Table 20(a)** extends this comparison to a contrastive objective and reports accuracy alongside cost: diffusion costs the most per pretraining step and returns the lowest error.

Deployment Cost. The denoising decoder is used only during pretraining and discarded afterward, so the choice of local objective does not change inference cost. **Table 20(b)** compares deployment cost and accuracy with recent foundation models and with supervised models trained per dataset. Among models that need no target-domain training, FedTRL attains the lowest average MSE, ahead of Chronos-2 and both Time-MoE scales, and it is faster and uses less memory than either Time-MoE variant while holding 2.7 times the parameters of the base one, since FedTRL carries 16 time points per token whereas Time-MoE tokenizes at point level. Supervised models trained per dataset are substantially cheaper, but they require a training split for every target domain, which is unavailable in the leave-one-domain-out setting and on RW-Bench. Where such a split exists, one epoch of local adaptation places FedTRL ahead of every supervised model in the table.

Table 20: Training and deployment cost. (a) Per-step pretraining time, peak memory, and average in-domain MSE on the eight datasets of Table 1, under a shared backbone and federated protocol. (b) Deployment cost (one NVIDIA L4, look-back 512, horizon 96) and MSE averaged over the five TSLib datasets and all horizons. Bold: lowest MSE per target-domain training regime.

(a) Pretraining cost

Local objective	Per-step time (s)	Peak memory (GB)	Avg. in-domain MSE
CoST (contrastive)	0.38	1.3	0.411
TimeMAE (masked)	0.37	0.6	0.398
PatchTST (masked)	0.41	0.5	0.358
SimMTM (masked)	0.42	1.7	0.389
FedTRL (diffusion)	0.59	2.0	**0.338**

(b) Deployment cost

Model	Params	Target-domain training	B=1 (ms)	B=32 (ms)	Peak mem. at B=32 (GB)	Avg. MSE
DLinear	0.1M	per dataset	0.12	0.13	0.01	0.345
iTransformer	6.6M	per dataset	0.88	1.42	0.04	0.344
PatchTST	9.5M	per dataset	0.99	3.09	0.10	**0.321**
Chronos-2	119M	none	12.8	48.9	0.50	0.336
Time-MoE base	113M	none	42.0	559.9	1.90	0.348
Time-MoE large	453M	none	59.4	1305.0	4.25	0.352
FedTRL (zero-shot)	302M	none	17.6	58.4	1.29	**0.324**
FedTRL (one epoch)	302M	one epoch	17.6	58.4	1.29	**0.298**

Communication. Each round transmits only the encoder θ_E and a single prototype vector $\bar{\mathbf{p}}_k$ from each client, while the decoder θ_D and sub-domain classifier θ_C never leave the client. The per-round prototype payload is below 0.01 MB and is therefore negligible relative to the encoder, so the total per-round communication cost is effectively identical to FedAvg.

Server-side Aggregation. The DaG step trains a lightweight global domain classifier on the K client prototypes and computes K scalar discriminability scores, with cost scaling linearly in K and bounded by a few seconds per round in our setting, independent of the pre-training data scale. The whole DaG block remains negligible when amortized over the per-round client compute.

E.9 Additional Discussions

This subsection provides additional analyses of our proposed FedTRL, including its generalization to unseen domains, its representation- and gradient-level behavior in handling heterogeneity, and its robustness under different simulated heterogeneity settings.

Generalization to Unseen Domain. To further test generalization, we conducted a leave-one-domain-out study (**Tables 21–22** below), where an entire domain was removed during pretraining. Even under this stricter setting, FedTRL consistently outperforms centralized mixed-domain pretraining and other FL baselines on the unseen domain. Performance naturally decreases when excluding related domains (an expected phenomenon for all foundation models) but FedTRL remains the most robust under such conditions.

Table 21: In-domain MSE (averaged over horizons) when the corresponding domain (energy or traffic) is excluded from pretraining.

Method	ETTh1	ETTm1	Traffic	Weather
FedAvg	0.511	0.433	0.488	0.295
FFTS	0.492	0.421	0.479	0.290
Centralized All Mixed	0.498	0.427	0.483	0.290
FedTRL (Original)	0.448	0.375	0.426	0.241
FedTRL (Ours)	**0.472**	**0.411**	**0.477**	**0.282**

Table 22: Zero-shot MSE (averaged over horizons) when the corresponding domain (energy or traffic) is excluded from pretraining.

Method	ETTh1	ETTm1	Weather
FedAvg	0.438	0.384	0.292
FFTS	0.427	0.380	0.282
Centralized All Mixed	0.431	0.382	0.288
FedTRL (Original)	0.399	0.349	0.238
FedTRL	0.418 (↓ 4.76%)	0.366 (↓ 4.87%)	0.269 (↓ 13.03%)

Representation- and Gradient-Level Analysis. **Table 23** shows that centralized mixed-domain pretraining yields the smallest inter-domain centroid distance and the largest within-domain variance, indicating collapsed and weakly structured representations. In contrast, FedTRL maintains clear domain separation while producing the most compact within-domain clusters. The same diagnostics computed on public centralized checkpoints (MOMENT, Timer, and Time-MoE) follow the pattern of All Mixed, so the phenomenon extends beyond our own backbone. FedTRL also attains the highest domain-level clustering purity while driving a within-client sub-domain probe close to chance, which is the intended bi-level structure: domains separate, while sub-domains inside a client mix. **Table 24** further shows that All-Mixed exhibits the strongest gradient conflict (lowest cosine similarity and highest gradient-norm variance), whereas FedTRL substantially alleviates these conflicts. The conflict grows with the amount of mixing: when the number of domains mixed during centralized pretraining increases from 2 to 4 to 9, with the total number of training sequences and update steps fixed, the mean gradient cosine similarity falls monotonically from 0.41 to 0.23, while that of FedTRL declines

only from 0.49 to 0.43. To measure collapse directly, **Fig. 9** reports the singular-value spectrum of the representation covariance over nine domains, computed with a re-implementation on a shared 128-dimensional backbone together with public TSFM checkpoints. Normalized by the embedding dimension, FedTRL retains the largest effective rank (0.55, versus 0.42–0.44 for All Mixed, FedAvg, and FFTS and 0.12–0.46 for MOMENT, Timer, and Time-MoE), so its representations spread over more directions rather than collapsing onto a few dominant ones.

Table 23: Representation analysis over the nine Time-MoE domains: inter-domain centroid distance, within-domain variance, domain-level clustering purity, and within-client sub-domain probe accuracy (chance ≈34%). The upper block uses public centralized TSFM checkpoints. "–": not computed.

Method	Inter-domain Centroid Distance (↑)	Within-domain Variance (↓)	Domain Purity (↑)	Sub-domain Probe Acc. (↓)
MOMENT (Goswami et al., 2024)	1.04	0.92	–	–
Timer (Liu et al., 2024b)	1.07	0.88	–	–
Time-MoE (base) (Shi et al., 2024)	1.12	0.81	–	–
All Mixed (Centralized)	1.08	0.91	0.55	76%
FedAvg	1.06	0.85	0.58	78%
FFTS	1.34	0.69	0.68	69%
FedTRL (Ours)	**1.27**	**0.53**	**0.79**	**42%**

Table 24: Gradient conflict across domains: average pairwise cosine similarity of domain-wise gradients and variance of gradient norms.

Method	Avg. Gradient Cosine Similarity (↑)	Gradient Norm Variance (↓)
All Mixed (Centralized)	0.23	0.021
FedAvg	0.33	0.015
FFTS	0.37	0.013
FedTRL (Ours)	**0.43**	**0.010**

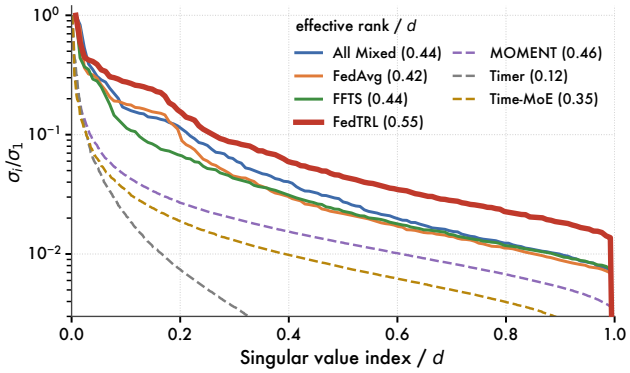


Figure 9: Singular-value spectrum of the representation covariance over nine domains, with the index normalized by the embedding dimension d. Legend: effective rank$/d$ (higher means less collapse).

Robustness Under Increasing Heterogeneity. By gradually increasing domain imbalance (from Balanced to Mildly Imbalanced to Real Skew) by controlling the data scale of each domain, we observe consistent performance degradation for All-Mixed, FedAvg, and FFTS across in-domain, full-shot, zero-shot, and probabilistic forecasting (as shown in **Table 25**). In contrast, FedTRL remains markedly more stable in all settings, indicating that severe heterogeneity harms federated training and that FedTRL effectively mitigates this effect.

Extended Ablations. **Table 26(a)** replaces the diffusion objective with masked reconstruction or contrastive learning while keeping the encoder, GRL, prototype alignment, and DaG unchanged; diffusion improves consistently in both in-domain and zero-shot settings. **Table 26(b)** shows that the gain of diffusion over masked reconstruction widens as the number of latent sub-domains per client grows and nearly disappears once each client is homogeneous, so the objective acts on the intra-domain level rather than being a generically stronger pretraining task. The diffusion noise level

Table 25: Forecasting results under different inter-domain imbalance levels.

Task Type / Dataset (Metric)	Method	Inter-domain Imbalance Level		
		Balanced (Simulated)	Mildly Imbalanced (Simulated)	Real Skew (Original Pretraining)
In-domain Forecasting (MSE on Weather)	All Mixed (Centralized)	0.235	0.248	0.254
	FedAvg	0.231	0.248	0.264
	FFTS	0.224	0.235	0.252
	FedTRL (Ours)	**0.216**	**0.230**	**0.241**
Full-shot Forecasting (MSE on Weather)	All Mixed (Centralized)	0.227	0.235	0.238
	FedAvg	0.228	0.239	0.240
	FFTS	0.212	0.219	0.227
	FedTRL (Ours)	**0.204**	**0.211**	**0.214**
Zero-shot Forecasting (MSE on RW-Bench)	All Mixed (Centralized)	0.610	0.608	0.614
	FedAvg	0.708	0.710	0.729
	FFTS	0.605	0.614	0.614
	FedTRL (Ours)	**0.587**	**0.592**	**0.599**
Zero-shot Probabilistic (MASE on GIFT-eval / FEV)	All Mixed (Centralized)	0.812 / 0.867	0.820 / 0.875	0.823 / 0.894
	FedAvg	0.835 / 0.877	0.849 / 0.879	0.872 / 0.890
	FFTS	0.752 / 0.852	0.754 / 0.855	0.766 / 0.876
	FedTRL (Ours)	**0.670 / 0.826**	**0.669 / 0.832**	**0.675 / 0.836**

$t \sim \mathcal{U}(0, T_d)$ is drawn per sample, so the encoder recovers temporal dynamics under a continuum of corruption strengths rather than one fixed masking ratio, which favors dynamics shared across the sub-domains coexisting inside a client. **Table 26(c)** ablates the two DaG terms separately: removing either term degrades performance, removing both causes the largest drop, and the discriminability term ℓ_k contributes more, consistent with the sensitivity trends in **Fig. 7**.

Table 26: Extended ablations (MSE averaged over horizons). (a) Replacing the diffusion objective, with all other components unchanged. (b) Zero-shot Weather MSE versus sub-domains per client, with total training sequences fixed. (c) Removing one DaG term ($\alpha = 0$ or $\beta = 0$), zero-shot.

(a) Local self-supervised objective

Local objective	In-domain			Zero-shot		
	ETTh1	ETTm1	Weather	ETTh1	ETTm1	Weather
Diffusion (ours)	**0.448**	**0.375**	**0.241**	**0.399**	**0.350**	**0.238**
Masked reconstruction	0.462	0.384	0.253	0.412	0.362	0.249
Contrastive	0.471	0.391	0.259	0.422	0.369	0.256

(b) Sub-domains per client

Sub-domains	Diffusion	Masked	Gain
1 (homogeneous)	0.244	0.246	0.8%
2	0.240	0.245	2.0%
4	0.242	0.248	2.4%
Default (2–15)	**0.238**	0.249	**4.4%**

(c) DaG terms

Variant	ETTh1	ETTm1	Weather	RW-Bench
Full DaG	**0.399**	**0.350**	**0.238**	**0.599**
w/o ℓ_k ($\alpha = 0$)	0.421	0.368	0.266	0.647
w/o δ_k ($\beta = 0$)	0.408	0.361	0.251	0.631
w/o DaG (FedAvg)	0.436	0.380	0.289	0.668

Coupling between Local and Server Components. The components of FedTRL are coupled through the prototype $\bar{\mathbf{p}}_k$ that each client uploads: diffusion reconstruction learns the representation the prototype summarizes, GRL removes sub-domain-specific variation from it, prototype alignment places client prototypes in a comparable space, and DaG weights aggregation by their discriminability and deviation. **Table 27** measures two links in this chain. Removing GRL raises the within-client sub-domain probe accuracy from 42% to 74% and lowers the accuracy of the server-side domain classifier on prototypes from 94.2% to 87.6%, while raising the mean and variance of ℓ_k across clients. Replacing diffusion with masked reconstruction also raises the mean ℓ_k. GRL thus gives up sub-domain separability, which aggregation does not need, in exchange for cleaner client-level signatures, which DaG relies on. The two classifiers act at different granularities and do not compete.

Statistical Reliability. **Table 28** counts, for each evaluation grid, how often FedTRL is ahead of its closest competitor across all dataset, horizon, and metric cells. Since cells within a dataset are correlated, we also aggregate each dataset and each RW-Bench station into one evaluation unit; FedTRL leads on 27 of the 28 units, which a two-sided sign test places at $p < 10^{-6}$. Pretraining

Table 27: Effect of the local components on the prototype signal used by DaG. Sub-domain probe: within-client probe accuracy (chance ≈34%). Global domain acc.: accuracy of f_d on uploaded prototypes. ℓ_k statistics are computed across clients. "–": not computed.

Variant	Sub-domain Probe (↓)	Global Domain Acc. (↑)	Mean ℓ_k (↓)	Var. ℓ_k (↓)
FedTRL	**42%**	**94.2%**	**0.31**	**0.010**
w/o GRL	74%	87.6%	0.52	0.028
Diffusion → masked reconstruction	–	–	0.44	–

is also repeated: the four client partitions in **Table 7(b)** are separately pretrained models that differ in data order and client sampling, and each stays ahead of centralized mixing and FedAvg on both benchmarks. Seed variability of the target adaptation is reported in **Table 17**.

Table 28: Cell-level win–loss of FedTRL against its closest competitor (smallest margin) on each grid, where a cell is one dataset, horizon, and metric. Ties are excluded.

Grid	Cells	Closest competitor	FedTRL win–loss
In-domain representation comparison (Table 13)	64	All Mixed	54–8
TSLib zero-shot (Table 14)	40	Time-MoE$_{\text{ultra}}$	28–11
RW-Bench zero-shot (Table 15)	120	Time-MoE$_{\text{large}}$	102–9

F SHOWCASES

Reconstruction from Noise in Pretraining. To illustrate the effectiveness of our diffusion-based reconstruction strategy, we showcase examples of noisy inputs, reconstructions, and original sequences. As shown in **Fig. 10**, the model successfully restores a wide range of temporal dynamics, including non-stationary series with high-frequency fluctuations, periodic components, approximate stationarity, and nonlinear trends. These cases highlight the ability of FedTRL to preserve fine-grained temporal structures and denoise corrupted inputs, even under complex and unstable conditions, thereby ensuring robust representation learning during pretraining.

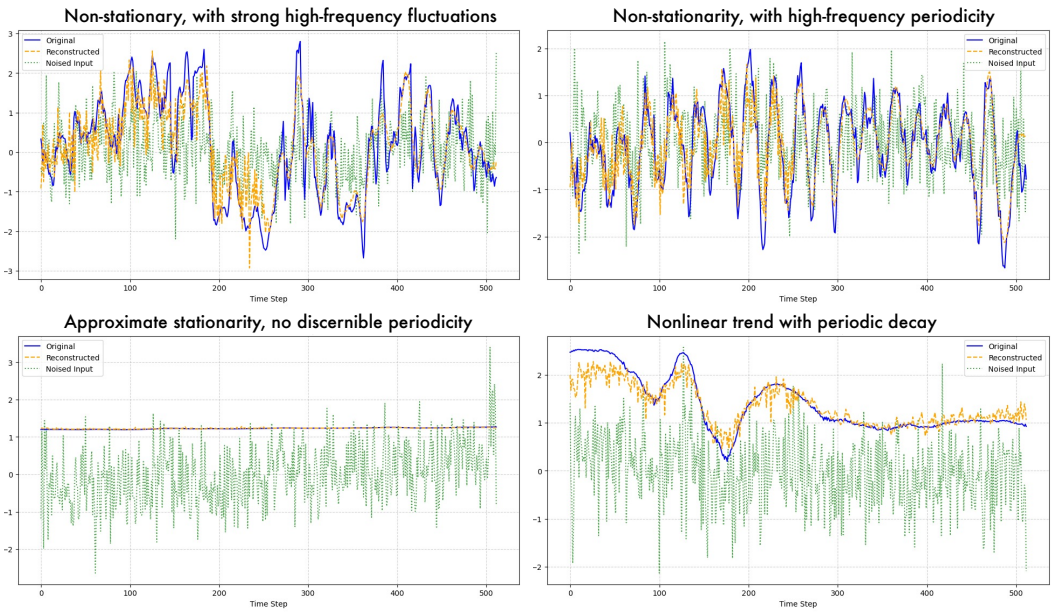


Figure 10: Showcase of diffusion-based reconstruction during pretraining.

Forecasting. **Fig. 11- 12** present zero-shot forecasting showcases on all the datasets from our RW-Bench (point forecasting) and FEV-leaderboard (Ansari et al., 2024) (probabilistic forecasting).

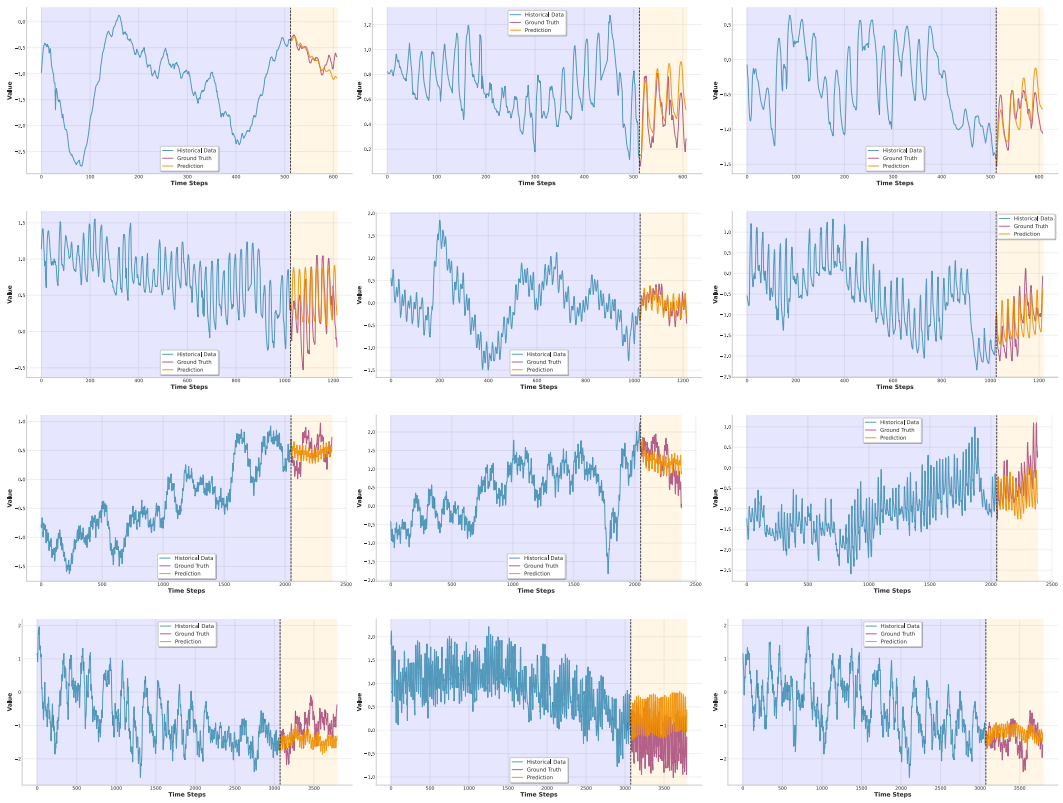


Figure 11: Zero-shot point forecasting results on our RW-Bench. From top to bottom, the prediction horizons are 96, 192, 336, and 720, corresponding to lookback window lengths of 512, 1024, 2048, and 3072, respectively. The visual samples are randomly selected from RW-Bench.

G Limitations and Broader Impacts

Limitations. Like other foundation-model pre-training pipelines, the overall training cost of FedTRL grows with model and data scale, so the absolute computational footprint is non-trivial even though the per-step overhead introduced by FedTRL on top of standard local objectives is modest. As the number of participating clients grows, the server-side aggregation step also incurs additional coordination cost (linear in client count), which, while small relative to the local pre-training compute, may become non-negligible in very large-scale deployments. Our claims are scoped to forecasting. As a first indication beyond it, a linear probe on three UEA classification datasets (Bagnall et al., 2018) improves from 68.9% with the FFTS encoder to 71.4% with the FedTRL encoder, while imputation and anomaly detection remain unevaluated. Finally, our convergence analysis (Theorem D.8) covers the DaG aggregation dynamics with one-round-lagged weights and an encoder-level surrogate objective, and does not cover the saddle-point dynamics of the GRL term.

Broader Impacts. FedTRL enables cross-domain TSFM pre-training in settings where data are fragmented or sensitive, such as climate observation, energy systems, public health, and industrial monitoring, lowering institutional barriers in domains under data-sovereignty constraints. By handling bi-level heterogeneity, it also reduces the risk that a few easily-aggregated domains dominate the representation space, favoring fairness across data sources. As with other forecasting models, predictions may be misused for surveillance or market manipulation, and biases in participating datasets can propagate downstream. The federated design provides a structural mitigation since raw data never leaves its source, while standard practices of dataset transparency, bias auditing under domain shift, and human-in-the-loop verification remain advisable for high-stakes use.

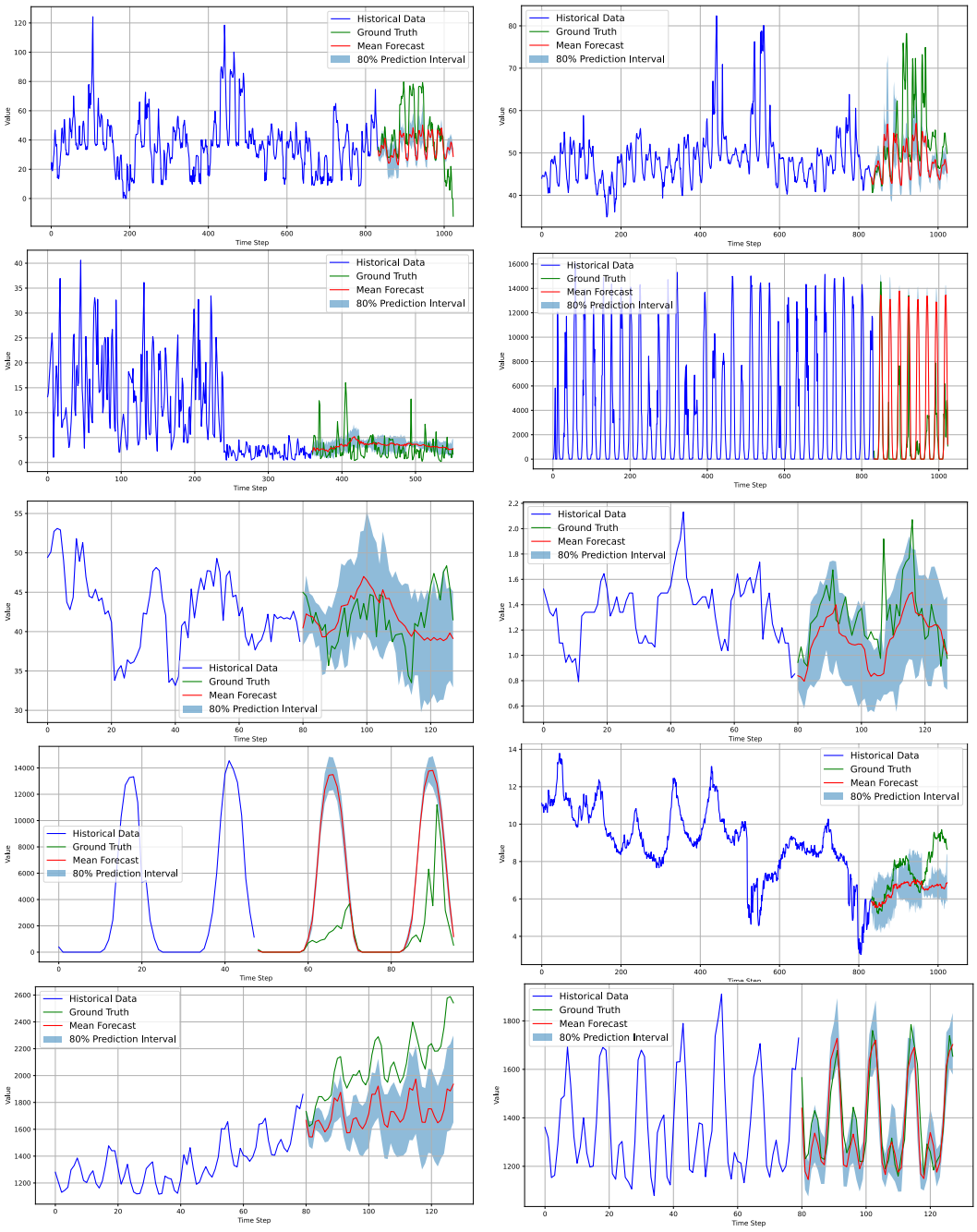


Figure 12: Zero-shot probabilistic forecasting results on FEV-leaderboard datasets, with visual samples randomly drawn from the benchmark.