MANIFOLD-AWARE TEMPORAL DOMAIN GENERALIZATION FOR LARGE LANGUAGE MODELS

Yiheng Yao[1] **Zekun Cai**[2,4*] **Xinyuan Song**[3] **Hiroki Hill Kobayashi**[1]
Xuan Song[2] **Ryosuke Shibasaki**[4] **Liang Zhao**[3]
[1]The University of Tokyo, Tokyo, Japan [2]Jilin University, Changchun, China
[3]Emory University, Atlanta, GA, USA [4]LocationMind, Tokyo, Japan
yihengyao@g.ecc.u-tokyo.ac.jp, {caizekun,songxuan}@jlu.edu.cn
{xinyuan.song,liang.zhao}@emory.edu, shiba@locationmind.com
kobayashi@ds.itc.u-tokyo.ac.jp

ABSTRACT

Temporal distribution shifts are pervasive in real-world deployments of Large Language Models (LLMs), where data evolves continuously over time. While Temporal Domain Generalization (TDG) seeks to model such structured evolution, existing approaches characterize model adaptation in the full parameter space. This formulation becomes computationally infeasible for modern LLMs. This paper introduces a geometric reformulation of TDG under parameter-efficient fine-tuning. We establish that the low-dimensional temporal structure underlying model evolution can be preserved under parameter-efficient reparameterization, enabling temporal modeling without operating in the ambient parameter space. Building on this principle, we propose Manifold-aware Temporal LoRA (MaT-LoRA), which constrains temporal updates to a shared low-dimensional manifold within a low-rank adaptation subspace, and models its evolution through a structured temporal core. This reparameterization dramatically reduces temporal modeling complexity while retaining expressive power. Extensive experiments on synthetic and real-world datasets, including scientific documents, news publishers, and review ratings, demonstrate that MaT-LoRA achieves superior temporal generalization performance with practical scalability for LLMs.

1 INTRODUCTION

Domain generalization (DG) (Tremblay et al., 2018; Huang et al., 2020; Du et al., 2020; Qiao et al., 2020; Wang et al., 2022) aims to learn models that generalize to unseen domains without access to target-domain data, and has become a central problem for reliable deployment of machine learning systems. In many real-world scenarios, however, domains are not isolated but are interconnected through underlying physical dynamics or evolving structures. Temporal domain generalization (TDG) (Li et al., 2020; Nasery et al., 2021; Qin et al., 2022; Bai et al., 2023; Yong et al., 2023; Zeng et al., 2024; Cai et al., 2024; Yu et al., 2025) represents a prominent and practically important instance of DG, where domains are indexed by time and assume data distributions evolve sequentially.

TDG formalizes generalization as trajectories of model parameters in a high-dimensional space, aiming to synchronize model evolution with distribution dynamics to enable generalization to future domains. State-of-the-art approaches characterize parameter evolution using mechanisms such as LSTM architectures Bai et al. (2023), continuous-time neural dynamics Cai et al. (2024), and spectral analyses of parameter trajectories Yu et al. (2025). Despite their methodological differences, these approaches aim to uncover low-dimensional structure within full-parameter trajectories. However, learning such a structure requires operating over the entire high-dimensional parameter space, leading to parameter overhead that scales with model size. As shown in Fig. 1, this overhead becomes prohibitively large for billion-parameter models, revealing the scalability limitation of TDG.

It has become increasingly evident that large language models (LLMs) deployed in real-world environments are also subject to distribution shifts analogous to those studied in TDG. Empirical

* Corresponding author

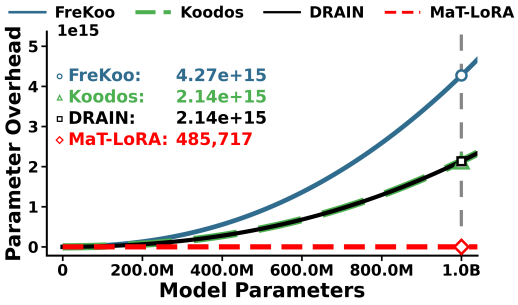

Figure 1: Illustration of the scalability bottleneck of TDG(Yu et al., 2025; Cai et al., 2024; Bai et al., 2023). Full-parameter TDG scales prohibitively with model size, whereas MaT-LoRA maintains overhead nearly constant and yields over $10^{10}\times$ reduction at the 1B model scale.

analyses reveal that model performance degrades as the evaluation time moves further away from the training period Luu et al. (2022), and simply scaling up model size does not eliminate this temporal generalization gap Lazaridou et al. (2021). Importantly, recent studies suggest that temporal variation in fine-tuned LLMs follows structured trajectories in weight space rather than random drift. Fine-tuned LLMs from neighboring time periods exhibit closely aligned update directions, and interpolating between them often yields effective models for unseen times with minimal additional training Nylund et al. (2024); Dziadzio et al. (2025). These empirical findings suggest that temporal model evolution in large-scale LLMs is also governed by low-dimensional dynamics, motivating the extension of TDG principles from moderate-scale models to modern large-scale LLMs.

Extending TDG to large-scale LLMs is non-trivial due to several fundamental challenges. **1) Mismatching of the parameter space for generalization.** Classical TDG characterizes temporal evolution as trajectories in the full parameter space, assuming that model parameters can be freely optimized over time. In contrast, LLMs are typically fine-tuned via parameter-efficient fine-tuning, which restricts learning to parameter increments around a pretrained initialization. Consequently, temporal trajectories are expressed in a different parameter space. **2) Expressing temporal dynamics under constrained updates.** The widely adopted LoRA-based fine-tune method Hu et al. (2022) for LLMs is confined to a low-rank subspace of the parameter space. Under such structural restriction, it is unclear whether the intrinsic low-dimensional temporal dynamics identified in TDG can remain fully representable, and whether the associated geometric structure can be preserved.

To bridge this gap, we develop a Manifold-aware Temporal LoRA (MaT-LoRA). We reinterpret generalization as occurring in a translated increment space anchored at a pretrained initialization. We show that if full-parameter optima lie on a low-dimensional manifold, then the corresponding parameter increments inherit the same intrinsic geometric structure via diffeomorphic translation. Building on this observation, we further identify that under low-rank adaptation, temporal updates should not be treated as independent rank-constrained solutions, but as points residing on a shared low-dimensional manifold embedded within the rank-restricted space. This perspective shifts the objective from fitting disconnected per-domain adapters to identifying and modeling the underlying manifold structure governing temporal evolution. MaT-LoRA confines temporal evolution independent of the LLMs size, but with the intrinsic dimensionality of the shared update subspace. As illustrated in Fig. **??**, MaT-LoRA reduces the effective temporal modeling complexity by more than ten orders of magnitude for TDG on billion-scale models. This makes TDG practically feasible for modern LLMs. We validate this framework through extensive experiments on both synthetic benchmarks and three real-world datasets, results demonstrating that MaT-LoRA consistently outperforms existing baselines while maintaining computational efficiency suitable for large-scale LLMs models.

2 Related Work

Domain Generalization (DG) and Temporal Domain Generalization (TDG) Domain Generalization (DG) aims to generalize to unseen target domains, a paradigm that Temporal Domain Generalization (TDG) extends by explicitly modeling time-varying data distributions. Following the taxonomy established by Wu et al. (2025), existing TDG methodologies are organized into three primary categories: (1) *Data Distribution*, delineates the nature of temporal shifts into Covariate Shift Kim et al. (2021); Kulinski et al. (2020); Zhang et al. (2025) and Concept Drift Wen et al. (2023); Zhang et al. (2022); (2) *Representation Learning*, encompasses Decoupling-based Zhou et al. (2022), Invariant-based Hu et al. (2023), and Adaptive Mechanism-based frameworks Wen et al. (2023), alongside the integration of Large Time Series Models Jin et al. (2024); and (3) *Out-Of-Distribution*

(OOD) Evaluation, stratifies assessment protocols into Extrinsic Kim et al. (2021); Zhang et al. (2022); Sun et al. (2025) and Intrinsic Xie et al. (2023); Cai et al. (2024; 2025) paradigms.

Parameter-Efficient Fine-Tuning (PEFT) Parameter-Efficient Fine-Tuning (PEFT) offers a resource-efficient paradigm for adapting large-scale models to diverse downstream tasks. Following the taxonomy proposed by Han et al. (2024), PEFT strategies can be categorized into four distinct groups: *additive PEFT*, augments the model architecture by injecting additional trainable modules or parameters(Houlsby et al. (2019); Mahabadi et al. (2021)); *selective PEFT*, designates only a specific subset of existing parameters to be updated during fine-tuning(Liu et al. (2022); Sung et al. (2021)); *reparameterized PEFT*, optimizes a low-dimensional reparameterization of the model weights during training that is subsequently merged for inference(Hu et al. (2022); Wu et al. (2024); and *hybrid PEFT* (Mao et al. (2022); Zhou et al. (2024)).

Current methodologies present a fundamental dichotomy. Traditional TDG methods rigorously model temporal dynamics but scale poorly with increasing parameter dimensionality. Conversely, PEFT strategies excel at adapting large-scale models yet typically treat adaptation as a static process, failing to capture the continuous evolution inherent in distributional shifts. The isolation of these paradigms highlights a critical research gap: the lack of a unified framework that efficiently reconciles temporal dynamic modeling with the computational constraints of LLMs.

3 PROBLEM DEFINITION

Temporal Domain Generalization (TDG) for large language models (LLMs) studies adaptation under continuous distribution shift over time. We consider a sequence of T source domains $\{\mathcal{D}_t\}_{t=1}^{T}$, where $\mathcal{D}_t = \{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^{N_t}$ and $x_i^{(t)}$, $y_i^{(t)}$, and N_t denote inputs, labels, and sample size at timestamp $t \in \mathbb{R}^1$. The shift is modeled as a temporal evolution of the conditional distribution $P_t(Y \mid X)$ over $\mathcal{X}_t \times \mathcal{Y}_t$, implying a time-varying input–output relationship.

Starting from a pretrained model W_{pre}, we represent adaptation at time t by an update ΔW_t, yielding the mapping $g_{W_{\text{pre}}+\Delta W_t} : \mathcal{X}_t \to \mathcal{Y}_t$. TDG aims to learn and extrapolate the trajectory $\{\Delta W_t\}_{t=1}^{T}$ to infer suitable weights for the future unseen domains $\{\mathcal{D}_{T+1}, \mathcal{D}_{T+2}, \ldots\}$.

4 METHODOLOGY

In this section, we develop MaT-LoRA(Manifold-aware Temporal LoRA), a framework for temporal domain generalization of LLMs. We first motivate a low-dimensional manifold structure for the trajectory of optimal updates under distribution shift. We then introduce a Manifold-Constrained Factorization that separates time-invariant spatial bases from a time-varying temporal core, retaining the expressivity of per-domain adapters while using constant memory. Finally, we describe general ways to parameterize the temporal core, from structured continuous dynamical systems to flexible arbitrary functions, to match different types of temporal evolution.

4.1 THE GEOMETRY OF PARAMETER INCREMENTS FOR TEMPORAL DOMAINS

In many temporal learning problems, distribution shift is driven by coherent physical, social, or semantic factors rather than arbitrary stochastic noise. We thus assume that the conditional distribution $P_t(Y \mid X)$ evolves over time according to an unknown but deterministic latent dynamics.

Prior work in TDG Cai et al. (2024; 2025) suggests that the sequence of full-parameter optima $\{W_t^*\}_{t=1}^{T}$ lies on a low-dimensional embedded manifold $\mathcal{M}$ in the ambient parameter space. We now formalize the corresponding statement in the parameter-increment space.

Definition 1 (Parameter increments)**.** *Fix a pretrained initialization $W_{\text{pre}} \in \mathbb{R}^p$. For each time step $t \in [T]$, define the optimal increment*

$$\Delta W_t^* := W_t^* - W_{\text{pre}}. \tag{1}$$

Define the map $\phi : \mathbb{R}^p \to \mathbb{R}^p$ by

$$\phi(W) := W - W_{\text{pre}}. \tag{2}$$

Lemma 1 (Embedded manifolds)**.** *Let $\mathcal{M} \subset \mathbb{R}^p$ be an embedded m-dimensional submanifold. Then $\mathcal{M}' := \phi(\mathcal{M}) = \mathcal{M} - W_{\text{pre}}$ is also an embedded m-dimensional submanifold of $\mathbb{R}^p$.*

Proof. The map $\phi(W) = W - W_{\text{pre}}$ is smooth and bijective with smooth inverse $\phi^{-1}(U) = U + W_{\text{pre}}$; hence ϕ is a diffeomorphism on $\mathbb{R}^p$. The image of an embedded submanifold under a diffeomorphism is an embedded submanifold of the same dimension. Therefore $\mathcal{M}' = \phi(\mathcal{M})$ is an embedded m-dimensional submanifold. □

Lemma 2 (Parameter-increment manifold)**.** *Assume $\{W_t^*\}_{t=1}^T \subset \mathcal{M}$ for some embedded m-dimensional submanifold $\mathcal{M} \subset \mathbb{R}^p$. Then $\{\Delta W_t^*\}_{t=1}^T \subset \mathcal{M}'$, where $\mathcal{M}' := \mathcal{M} - W_{\text{pre}}$ is also an embedded m-dimensional submanifold.*

Proof. By Definition 1, $\Delta W_t^* = \phi(W_t^*)$. Since $W_t^* \in \mathcal{M}$ for each t, it follows that $\Delta W_t^* \in \phi(\mathcal{M}) = \mathcal{M}'$. The manifold claim follows from Lemma 1. □

This result motivates working in the parameter-increment space: modeling temporal generalization via ΔW_t provides a principled reparameterization for large pretrained models under temporal domain shifts, and it supports extrapolation by learning trajectories on a low-dimensional manifold.

4.2 Manifold-Constrained Low-Rank Factorization

TDG aims to model the continuous evolution of the data-generating process. For smaller models, prior methods (e.g., Bai et al. (2023); Cai et al. (2024)) often project the full parameter trajectory onto a compact latent manifold to obtain a compressed representation of temporal dynamics. For LLMs, such global dimensionality reduction is computationally infeasible due to the scale of the weight matrices, and directly regressing full weights can overwrite pretrained generalization Luo et al. (2025); Li et al. (2024). We therefore shift the perspective from compressing the full parameter space to modeling temporal generalization within a constrained update subspace that matches the inherently low-rank structure of effective model variations.

4.2.1 Geometric Incoherence in Discrete Generalization

Low-Rank Adaptation (LoRA) Hu et al. (2022) is not only parameter efficient; it restricts adaptation to an intrinsic low-dimensional coordinate system. Specifically, LoRA constrains the update to $W_{\text{pre}} + BA$, where $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, so that $\Delta W = BA$ has rank at most r. This confines optimization to a much smaller subset of the ambient parameter space and makes temporal evolution of updates easier to model. Consequently, learning temporal distribution shift via parameter dynamics reduces to modeling a time-varying embedded submanifold within the rank-r update space.

In the standard approach, each domain t is fitted with an independent LoRA factorization (B_t, A_t) Hu et al. (2022). While this enables domain-specific adaptation, it ignores temporal structure. In over-parameterized regimes, the loss surface is highly non-convex and admits many distinct solutions with similarly low loss; empirically, different optima can be connected by low-loss curves (mode connectivity), indicating large near-optimal sets rather than isolated points Garipov et al. (2018); Draxler et al. (2018). As a result, independently optimizing $\{(B_t, A_t)\}_{t=1}^T$ can yield updates that are scattered even within the low-rank update space, which makes extrapolation to unseen domains $\{\mathcal{D}_{T+1}, \ldots\}$ ill-posed.

In contrast, if the full-parameter optima exhibit a manifold structure, then the same structure holds in the increment space by Lemmas 1–2. Specifically, assuming $\{W_t^*\}_{t=1}^T \subset \mathcal{M}$ for some embedded low-dimensional submanifold $\mathcal{M} \subset \mathbb{R}^p$, Lemma 1 implies that the translated set $\mathcal{M}' = \mathcal{M} - W_{\text{pre}}$ remains an embedded submanifold of the same intrinsic dimension. Lemma 2 then yields $\{\Delta W_t^*\}_{t=1}^T \subset \mathcal{M}'$. Therefore, optimal increments are not arbitrary points in the ambient rank-r space; they are constrained to lie on a compact low-dimensional manifold (up to translation). This shifts the modeling objective from fitting disconnected per-domain adapters to identifying and modeling the underlying manifold structure so that temporal extrapolation is well-defined.

4.2.2 Subspace-Shared Time-Varying Parameterization

Let $\{\Delta W_t = B_t A_t\}_{t=1}^T$ be a sequence of LoRA updates where $B_t \in \mathbb{R}^{d\times r}$ and $A_t \in \mathbb{R}^{r\times k}$. The global column subspace could be $\mathcal{C} = \text{span}(\bigcup_{t=1}^T \text{col}(B_t))$ and the global row subspace be $\mathcal{R} = \text{span}(\bigcup_{t=1}^T \text{row}(A_t))$. Then, there exist fixed basis matrices $B \in \mathbb{R}^{d\times r'}$ and $A \in \mathbb{R}^{r'\times k}$, where r' satisfies $r \geq r' \geq \max(\dim(\mathcal{C}), \dim(\mathcal{R}))$. Then, construct B such that its columns form a basis for the global column subspace $\mathcal{C}$. Since $\text{col}(B_t) \subseteq \mathcal{C} = \text{col}(B)$ for all t, each column of B_t can be

expressed as a linear combination of the columns of B. Thus, there exists a unique coefficient matrix $C_t \in \mathbb{R}^{r' \times r}$ such that:

$$B_t = BC_t \tag{3}$$

Similarly, construct A such that its rows form a basis for the global row subspace $\mathcal{R}$. Since $\text{row}(A_t) \subseteq \mathcal{R} = \text{row}(A)$, each row of A_t lies in the row space of A. Thus, there exists a unique coefficient matrix $D_t \in \mathbb{R}^{r \times r'}$ such that:

$$A_t = D_t A \tag{4}$$

Substituting these factorizations back into the original update equation:

$$\Delta W_t = B_t A_t = (BC_t)(D_t A) = B(C_t D_t)A \tag{5}$$

To enforce geometric consistency and enable extrapolation, by defining the core matrix as $F_t := C_t D_t \in \mathbb{R}^{r' \times r'}$, we introduce the Manifold-Constrained Factorization, where the time-varying update $\Delta W(t)$ could be decomposed into time-invariant ambient bases and a dynamic temporal core:

$$\Delta W(t) = B \cdot F_t \cdot A \tag{6}$$

B and A identify the embedded submanifold within the rank-r latent space, and F_t governs the manifold generalization dynamics. We demonstrate in Sec. A.2 that this structured form is not a restrictive approximation but rather a sufficient representation for any sequence of low-rank updates that share a common support. This confirms that the factorization captures the full expressive power of the independent adapters, provided the updates reside within a shared ambient subspace.

4.3 Instantiations of the Temporal Core

As established in Section 4.1, the trajectory of the optimal update ΔW_t is diffeomorphic to the continuous evolution of the data distribution $\mathcal{D}_t$. Under the shared-basis factorization $\Delta W_t = BF_tA$, all temporal variation is captured by the low-dimensional core F_t while B and A remain fixed. Because the induced map on the active subspace is a linear isomorphism under full-rank bases, this diffeomorphic structure is preserved in F_t, making the core dynamics topologically equivalent to the intrinsic data-stream dynamics. We therefore choose the parameterization of F_t to match the inductive bias of the underlying evolution, and consider three instantiations for different regimes:

- **Continuous Linear Dynamical Systems:** If the data evolution is hypothesized to follow a smooth, continuous flow governed by an Autonomous Linear Dynamical System, we propose Lin-dym MaT-LoRA, which models the trajectory analytically using Lie algebra. Assuming the rate of change is governed by a constant velocity field $\mathcal{W}$, the core evolves as:

  $$F_t = \exp(t\mathcal{W}) \cdot F_0 \tag{7}$$

  where $\mathcal{W} \in \mathbb{R}^{r' \times r'}$ is a learnable coefficient matrix and $\exp(\cdot)$ denotes the matrix exponential. This analytic form enforces a strong inductive bias for structural continuity.

- **Sequential Markovian Evolution:** When the data evolution possesses the Markov property, a recursive modeling approach is required. We propose Markv MaT-LoRA, and parameterize F_t using a Recurrent Neural Network (e.g., RNN or LSTM) to capture autoregressive dependencies:

  $$\text{vec}(F_t) = \text{RNN}(h_{t-1}), \tag{8}$$

  This formulation allows the system to generate the current core F_t based on the trajectory of previous time steps, making it ideal for path-dependent shifts.

- **Arbitrary Non-Linear Dynamics:** For scenarios where the domain shift exhibits complex, high-dimensional non-linear patterns without clear sequential dependency, we leverage the universal approximation capability of Multi-Layer Perceptrons (MLPs), named Non-lin MaT-LoRA. In this setting, the core matrix is modeled as a direct function of the timestamp t:

  $$\text{vec}(F_t) = \text{MLP}(t) \tag{9}$$

  This form fits a global function to the temporal manifold, suitable for deterministic environments where the parameter state is strictly dependent on t.

5 EXPERIMENTS

This section empirically evaluates *MaT-LoRA* through comprehensive quantitative and qualitative analyses. We conducted experiments on four representative LLMs: DistilBERT(Sanh et al. (2019)), Qwen3-0.6B(Yang et al. (2025)), TinyLLaMA-1.1B(Zhang et al. (2024)), and LLaMA3-8B(Dubey et al. (2024)). By leveraging both synthetic and real-world datasets to ensure robust assessment, our evaluation aims to: (1) assess the temporal generalizability of *MaT-LoRA* within LLMs contexts; and (2) verify its fidelity in capturing underlying data dynamics.

5.1 SYNTHETIC DATASET

Synthetic datasets serve as an effective instrument for validation, avoiding the confounding factors of unknown generative laws and high-dimensional noise inherent in real-world scenarios. By restricting domain evolution to pure mathematical transformations, these datasets strictly conform to the TDG hypothesis of time-controlled dynamics. Validating on such data confirms MaT-LoRA's capability to capture dynamics within a regime that faithfully follows the field's most fundamental assumptions.

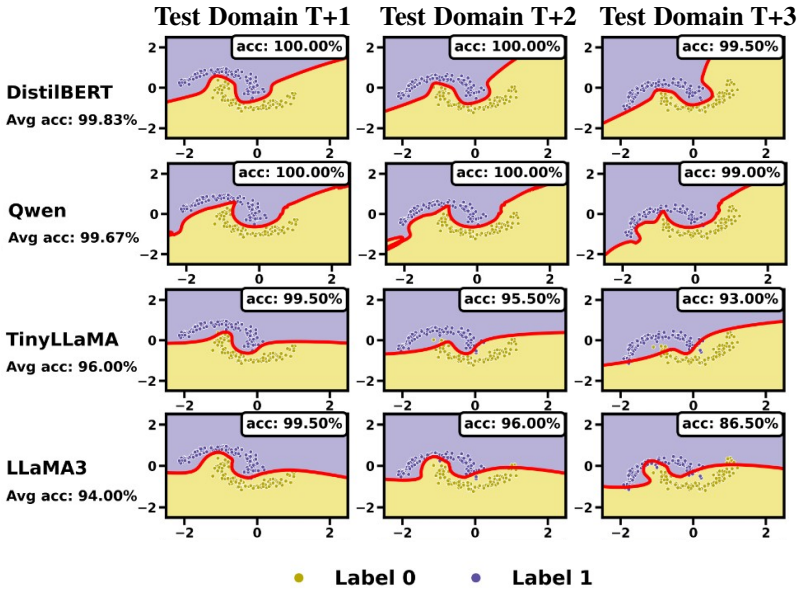

Figure 2: Visualization of Extrapolated Decision Boundaries on the Rotating 2-Moons Dataset across Four LLMs Backbones.

Experimental Setup: The **Rotating 2-Moons** dataset serves as a benchmark for controlled, continuous concept drift. In this setting, the classification boundary remains topologically invariant, while the data distribution undergoes a deterministic geometric shift. The sequence comprises 12 distinct domains, each containing 200 class-balanced instances generated via an 18° counterclockwise rotation per time step, perturbed by Gaussian noise ($\sigma = 0.10$). The first 9 domains constitute the training set, with the subsequent domains reserved for evaluation. To accommodate continuous 2D inputs, we project coordinates directly into the latent space via a linear layer, eschewing standard token and positional embeddings.

Qualitative Analysis of Decision Boundaries: Fig. 2 illustrates the decision boundaries derived from MaT-LoRA across four pre-trained LLMs: DistilBERT, Qwen, TinyLLaMA, and LLaMA-3. DistilBERT demonstrates superior generalization, with predictions that tightly track the ground-truth trajectory over multiple extrapolation steps, indicating effective capture of the underlying dynamics. Notably, despite the larger parameter counts of TinyLLaMA and LLaMA3 compared to Qwen, they exhibit clearer decision boundaries. While Qwen achieves superior test performance, visual inspection suggests it has captured spurious features, a phenomenon potentially attributable to its distinct Q, K normalization mechanism.

Overall, MaT-LoRA consistently enables pre-trained models—which inherently lack temporal extrapolation capabilities—to model data dynamics with high fidelity, demonstrating the robustness of our approach in enhancing generalization within the LLMs domain.

5.2 Real world Dataset

We utilize three existing temporal misalignment datasets(AIC, NewsCLS, and Yelp) introduced by Luu et al. (2022). AIC is a binary classification benchmark that requires identifying the publication venue (ICML vs. AAAI) based on a paper's title. NewsCLS is a three-class classification benchmark that predicts the publisher from news content. Yelp is a five-class classification benchmark that maps a user review to its corresponding numerical rating. Further datasets details are provided in C.1.

Table 1: Performance comparison on temporal domain datasets, reporting classification accuracy (%).

Method		AIC	NewsCLS	Yelp
Offline		84.00 ± 5.58	82.79 ± 0.97	69.59 ± 0.92
IncFinetune		90.58 ± 0.63	82.60 ± 1.69	70.46 ± 0.28
Last Domain		90.17 ± 0.63	80.55 ± 2.07	68.41 ± 0.42
MaT-LoRA	Lin-dym	91.17 ± 1.42	83.72 ± 4.89	72.04 ± 0.28
	Markv	$\mathbf{91.67} \pm 2.02$	$\mathbf{84.86} \pm 1.98$	$\mathbf{72.29} \pm 0.12$
	Non-lin	90.75 ± 1.75	84.70 ± 0.72	71.78 ± 0.24

Metrics. Accuracy (%) is used for classification tasks. All models were trained on training domains and then deployed on all unseen test domains. Each method's experiments were repeated three times, with mean results and standard deviations reported. Detailed parameter settings for each dataset are provided in Appendix C.2.

As presented in Table 1, all three variants of MaT-LoRA consistently outperform the baseline methods across real-world datasets. While approaches like Offline and LastDomain demonstrate limited efficacy, their failure to explicitly model temporal continuity results in significant performance gaps. In contrast, MaT-LoRA effectively captures the underlying evolution of data distributions. Notably, even the simplest linear variant performs exceptionally well, suggesting that real-world temporal variations and their manifestation in low-dimensional spaces may not be complex. Furthermore, our results demonstrate that all three datasets achieve state-of-the-art performance on RNN variants, indicating that distributional changes can be based on the evolution of previous states. Overall, these findings confirm MaT-LoRA as a robust benchmark for Temporal Domain Generalization in LLMs.

Table 2: Comparison of training and testing running time (seconds) on AIC.

Method	Train	Test
Offline	532 ± 153	2.11
Last domain	241 ± 31	2.73
IncFinetune	402 ± 77	2.46
MaT-LoRA	740 ± 18	2.99

Running time. MaT-LoRA targets TDG tasks within LLMs, where computational efficiency is paramount. To evaluate this, we conducted a comparative analysis against baselines on the AIC dataset over three independent runs, reporting the mean and standard deviation for training duration and the average inference latency. As shown in Table 2, MaT-LoRA incurs only a marginal increase in training time compared to baselines, a slight overhead attributed to optimizing parameter generalization on the submanifold. Crucially, inference latency remains on par with baselines, demonstrating that MaT-LoRA effectively mitigates domain shifts while maintaining high efficiency during both training and testing.

6 Conclusion

In this work, we presented MaT-LoRA, a framework for temporal domain generalization in LLMs that exploits the low-dimensional manifold structure of parameter updates. By introducing a Manifold-Constrained Factorization, our approach decouples adaptation into time-invariant spatial bases and a dynamic temporal core, effectively reconciling the expressivity of domain-specific adapters with

constant memory efficiency. Furthermore, the flexible parameterization of this core, ranging from continuous dynamical systems to arbitrary functions, enables robust generalization across diverse evolutionary patterns. Extensive experiments on both synthetic and real-world datasets validate the superior efficacy and computational efficiency of our proposed design.

7 ACKNOWLEDGMENT

This work was supported by the JST FOREST Program (Grant Number: JPMJFR211B); the grants of Jilin Provincial International Cooperation Key Laboratory for Super Smart City and Jilin Provincial Key Laboratory of Intelligent Policing.

REFERENCES

Guangji Bai, Chen Ling, and Liang Zhao. Temporal domain generalization with drift-aware dynamic neural networks. In *The Eleventh International Conference on Learning Representations*, 2023.

Zekun Cai, Guangji Bai, Renhe Jiang, Xuan Song, and Liang Zhao. Continuous temporal domain generalization. *Advances in Neural Information Processing Systems*, 37:127987–128014, 2024.

Zekun Cai, Yiheng Yao, Guangji Bai, Renhe Jiang, Xuan Song, Ryosuke Shibasaki, and Liang Zhao. Continuous domain generalization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pp. 519–528, 2003.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A. Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning (ICML)*, 2018.

Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. In *European Conference on Computer Vision*, pp. 200–216, 2020.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Sebastian Dziadzio, Vishaal Udandarao, Karsten Roth, Ameya Prabhu, Zeynep Akata, Samuel Albanie, and Matthias Bethge. How to merge your multimodal models over time? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20479–20491, 2025.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of DNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pp. 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Xuanming Hu, Wei Fan, Kun Yi, Pengfei Wang, Yuanbo Xu, Yanjie Fu, and Pengyang Wang. Boosting urban prediction via addressing spatial-temporal distribution shift. In *2023 IEEE International Conference on Data Mining (ICDM)*, pp. 160–169. IEEE, 2023.

Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *16th European Conference on Computer Vision, ECCV 2020*, pp. 124–140. Springer, 2020.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations*, 2021.

Sean Kulinski, Saurabh Bagchi, and David I Inouye. Feature shift detection: Localizing which features have shifted via conditional distribution tests. *Advances in neural information processing systems*, 33:19523–19533, 2020.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363, 2021.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Sequential learning for domain generalization. In *European conference on computer vision*, pp. 603–619. Springer, 2020.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4297–4308, 2024.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A Smith. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5944–5958, 2022.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 565–576, 2021.

Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Scott Yih, and Madian Khabsa. Unipelt: A unified framework for parameter-efficient language model tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6253–6264, 2022.

Anshul Nasery, Soumyadeep Thakur, Vihari Piratla, Abir De, and Sunita Sarawagi. Training for the future: A simple gradient interpolation loss to generalize along time. *Advances in Neural Information Processing Systems*, 34:19198–19209, 2021.

Kai Nylund, Suchin Gururangan, and Noah A Smith. Time is encoded in the weights of finetuned language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2571–2587, 2024.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*, 2002.

Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12556–12565, 2020.

Tiexin Qin, Shiqi Wang, and Haoliang Li. Generalizing to evolving domains with latent structure-aware sequential autoencoder. In *International Conference on Machine Learning*, pp. 18062–18082. PMLR, 2022.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Yanru Sun, Zongxia Xie, Emadeldeen Eldele, Dongyue Chen, Qinghua Hu, and Min Wu. Learning pattern-specific experts for time series forecasting under patch-level distribution shift. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yi-Lin Sung, Varun Nair, and Colin A Raffel. Training neural networks with fixed sparse masks. *Advances in Neural Information Processing Systems*, 34:24193–24205, 2021.

Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 969–977, 2018.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022.

Qingsong Wen, Weiqi Chen, Liang Sun, Zhang Zhang, Liang Wang, Rong Jin, Tieniu Tan, et al. Onenet: Enhancing time series forecasting models under concept drift by online ensembling. *Advances in Neural Information Processing Systems*, 36:69949–69980, 2023.

Taiqiang Wu, Jiahao Wang, Zhe Zhao, and Ngai Wong. Mixture-of-subspaces in low-rank adaptation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7880–7899, 2024.

Xin Wu, Fei Teng, Xingwang Li, Ji Zhang, Tianrui Li, and Qiang Duan. Out-of-distribution generalization in time series: A survey. *arXiv preprint arXiv:2503.13868*, 2025.

Mixue Xie, Shuang Li, Longhui Yuan, Chi Liu, and Zehui Dai. Evolving standardization for continual domain generalization over temporal drift. *Advances in Neural Information Processing Systems*, 36:21983–22002, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

LIN Yong, Fan Zhou, Lu Tan, Lintao Ma, Jianmeng Liu, HE Yansu, Yuan Yuan, Yu Liu, James Y Zhang, Yujiu Yang, et al. Continuous invariance learning. In *The Twelfth International Conference on Learning Representations*, 2023.

En Yu, Jie Lu, Xiaoyu Yang, Guangquan Zhang, and Zhen Fang. Learning robust spectral dynamics for temporal domain generalization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Qiuhao Zeng, Wei Wang, Fan Zhou, Gezheng Xu, Ruizhi Pu, Changjian Shui, Christian Gagné, Shichun Yang, Charles X Ling, and Boyu Wang. Generalizing across temporal domains with koopman operators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16651–16659, 2024.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

Weiyang Zhang, Xinyang Chen, Yu Sun, Weili Guan, and Liqiang Nie. Batch training for streaming time series: A transferable augmentation framework to combat distribution shifts. *Transactions on Machine Learning Research*, 2025.

Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, Zhou Qin, and Wenwu Zhu. Dynamic graph neural networks under spatio-temporal distribution shift. *Advances in neural information processing systems*, 35:6074–6089, 2022.

Han Zhou, Xingchen Wan, Ivan Vulić, and Anna Korhonen. Autopeft: Automatic configuration search for parameter-efficient fine-tuning. *Transactions of the Association for Computational Linguistics*, 12:525–542, 2024.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pp. 27268–27286. PMLR, 2022.

A Theoretical Analysis

A.1 Parameter Efficiency Analysis

Beyond geometric coherence, the proposed factorization is parameter-efficient. Consider LoRA updates of the form $\Delta W_t = B_t A_t$ with rank r, where $B_t \in \mathbb{R}^{d\times r}$ and $A_t \in \mathbb{R}^{r\times k}$. Maintaining an independent adapter per domain yields

$$\mathcal{O}(\text{Multi}) = T\,(dr + rk) = T(d+k)r. \tag{10}$$

A static single-adapter baseline uses

$$\mathcal{O}(\text{Single}) = dr + rk = (d+k)r. \tag{11}$$

Our manifold-constrained parameterization shares the large bases $B \in \mathbb{R}^{d\times r'}$ and $A \in \mathbb{R}^{r'\times k}$ across time and learns only a small time-varying core F_t (via a lightweight model with $|W_F|$ parameters), giving

$$\mathcal{O}(\text{Ours}) = (dr' + r'k) + |W_F| = (d+k)r' + |W_F|. \tag{12}$$

With $r' \le r \ll \min(d,k)$ and $|W_F|$ small, our total parameter cost is dominated by the shared bases and remains essentially independent of T. Consequently, compared to Discrete Multi-LoRA, MaT-LoRA reduces adapter storage by an approximate factor of T (up to the minor $|W_F|$ overhead), while still allowing domain-specific adaptation through the compact core F_t. This yields a scalable and VRAM-friendly alternative for long domain sequences, where keeping separate LoRA adapters quickly becomes prohibitive.

A.2 Stability and Justification of the Shared-Basis Form

To justify the shared-basis parameterization $\Delta W_t = BF_tA$, we also show that, under the training procedure (gradient descent on LoRA factors), the learned factor subspaces do not drift far over time. Concretely, starting from the initial factors (B_1, A_1), we prove that the components of B_t orthogonal to $\mathrm{col}(B_1)$ and of A_t orthogonal to $\mathrm{row}(A_1)$ remain small throughout the updates, up to a controlled leakage term. This establishes that using fixed shared subspaces for B and A is meaningful, and that temporal variation can be concentrated into a small core F_t.

Assumption 1 (Differentiability and GD dynamics)**.** *For each $t \in [T]$, the objective $\mathcal{J}_t : \mathbb{R}^{d\times r} \times \mathbb{R}^{r\times k} \to \mathbb{R}$ is differentiable. The LoRA factors (B_t, A_t) are updated by gradient descent*

$$B_{t+1} = B_t - \eta \nabla_B \mathcal{J}_t(B_t, A_t), \qquad A_{t+1} = A_t - \eta \nabla_A \mathcal{J}_t(B_t, A_t), \tag{13}$$

with step size $\eta > 0$.

Definition 2 (Reference shared subspaces)**.** *Let $B := B_1$ and $A := A_1$. Define the orthogonal projectors*

$$P_B := \mathrm{Proj}_{\mathrm{col}(B)} \in \mathbb{R}^{d\times d}, \qquad P_A := \mathrm{Proj}_{\mathrm{col}(A^\top)} \in \mathbb{R}^{k\times k}. \tag{14}$$

Equivalently, $\mathrm{Im}(P_A) = \mathrm{row}(A) \subset \mathbb{R}^k$.

Assumption 2 (Dissipativity of out-of-subspace components)**.** *There exist constants $\alpha_B, \alpha_A \ge 0$ and $\varepsilon_B, \varepsilon_A \ge 0$ such that for all $t \in [T]$,*

$$\big\langle (I-P_B)B_t, (I-P_B)\nabla_B \mathcal{J}_t(B_t,A_t)\big\rangle \ge \alpha_B \|(I-P_B)B_t\|_F^2 - \varepsilon_B, \tag{15}$$

$$\big\langle A_t(I-P_A), \nabla_A \mathcal{J}_t(B_t,A_t)(I-P_A)\big\rangle \ge \alpha_A \|A_t(I-P_A)\|_F^2 - \varepsilon_A, \tag{16}$$

where $\langle X, Y\rangle := \mathrm{tr}(X^\top Y)$.

Assumption 2 requires that the optimization dynamics do not amplify the components of B_t and A_t outside the selected subspaces. Instead, the gradient acts as a restoring force on $(I-P_B)B_t$ and $A_t(I-P_A)$, shrinking them whenever they become large, up to bounded slack terms $\varepsilon_B, \varepsilon_A$. Therefore, the evolution of the in-subspace components and the out-of-subspace components remains approximately decoupled, preserving the intended independence of the structured part from its complement throughout training.

Assumption 3 (Projected gradient growth outside the reference subspaces)**.** *There exist constants $\beta_B, \beta_A \ge 0$ and $\delta_B, \delta_A \ge 0$ such that for all $t \in [T]$,*

$$\|(I-P_B)\nabla_B \mathcal{J}_t(B_t,A_t)\|_F^2 \le \beta_B \|(I-P_B)B_t\|_F^2 + \delta_B, \tag{17}$$

$$\|\nabla_A \mathcal{J}_t(B_t,A_t)(I-P_A)\|_F^2 \le \beta_A \|A_t(I-P_A)\|_F^2 + \delta_A. \tag{18}$$

Assumption 3 states that the gradient energy outside the reference subspaces is controlled by the current out-of-subspace energy, up to a bounded residual term. Together with Assumption 2, this yields a stable discrete-time recursion under gradient descent: the out-of-subspace components are not only discouraged by the gradient direction, but their one-step fluctuation is also quantitatively bounded.

Theorem 3 (Stability of (B_t, A_t) subspaces under GD)**.** *Under Assumptions 1–2 and Assumption 3, define*

$$e_t^B := \|(I - P_B)B_t\|_F, \qquad e_t^A := \|A_t(I - P_A)\|_F, \qquad \Delta W_t := B_t A_t. \tag{19}$$

Suppose the step size $\eta > 0$ satisfies

$$\kappa_B := 1 - 2\eta\alpha_B + \eta^2\beta_B < 1, \qquad \kappa_A := 1 - 2\eta\alpha_A + \eta^2\beta_A < 1. \tag{20}$$

Then for all $t \in [T]$,

$$(e_{t+1}^B)^2 \le \kappa_B (e_t^B)^2 + 2\eta\varepsilon_B + \eta^2\delta_B, \tag{21}$$

$$(e_{t+1}^A)^2 \le \kappa_A (e_t^A)^2 + 2\eta\varepsilon_A + \eta^2\delta_A. \tag{22}$$

Consequently, for all $t \in [T]$,

$$(e_t^B)^2 \le \kappa_B^{t-1}(e_1^B)^2 + \frac{2\eta\varepsilon_B + \eta^2\delta_B}{1 - \kappa_B}\big(1 - \kappa_B^{t-1}\big), \tag{23}$$

$$(e_t^A)^2 \le \kappa_A^{t-1}(e_1^A)^2 + \frac{2\eta\varepsilon_A + \eta^2\delta_A}{1 - \kappa_A}\big(1 - \kappa_A^{t-1}\big). \tag{24}$$

Moreover, the induced update leakage satisfies

$$\|(I - P_B)\Delta W_t\|_F \le e_t^B \|A_t\|_2, \qquad \|\Delta W_t(I - P_A)\|_F \le \|B_t\|_2 e_t^A. \tag{25}$$

The detailed proof is provided in Section B. Theorem 3 shows that, under the dissipativity conditions in Assumption 2, the out-of-subspace energies $(e_t^B)^2$ and $(e_t^A)^2$ follow a contractive recursion with additive noise $\varepsilon_B, \varepsilon_A$. Hence, if the initialization is aligned ($e_1^B = e_1^A = 0$) and leakage is small, then $\mathrm{col}(B_t)$ stays close to $\mathrm{col}(B_1)$ and $\mathrm{row}(A_t)$ stays close to $\mathrm{row}(A_1)$ for all t, which in turn implies that the induced updates $\Delta W_t = B_t A_t$ largely remain supported on the shared subspaces, as quantified by (25). Therefore, a shared (B, A) with a time-varying core is a consistent structural restriction rather than an ad hoc constraint.

B Proof of Theorem 3

Proof. We prove the B-part; the A-part follows analogously.

Let $R_B := I - P_B$. By Assumption 2, P_B is an orthogonal projector, hence R_B is also an orthogonal projector:

$$R_B^\top = R_B, \qquad R_B^2 = R_B. \tag{26}$$

By Assumption 1,

$$B_{t+1} = B_t - \eta\nabla_B \mathcal{J}_t(B_t, A_t). \tag{27}$$

Left-multiplying by R_B, we obtain

$$R_B B_{t+1} = R_B B_t - \eta R_B \nabla_B \mathcal{J}_t(B_t, A_t). \tag{28}$$

Taking squared Frobenius norms and expanding,

$$\|R_B B_{t+1}\|_F^2 = \|R_B B_t\|_F^2 - 2\eta\big\langle R_B B_t,\ R_B\nabla_B\mathcal{J}_t(B_t, A_t)\big\rangle + \eta^2\|R_B\nabla_B\mathcal{J}_t(B_t, A_t)\|_F^2. \tag{29}$$

By Assumption 2, specifically (15),

$$\big\langle R_B B_t,\ R_B\nabla_B\mathcal{J}_t(B_t, A_t)\big\rangle \ge \alpha_B\|R_B B_t\|_F^2 - \varepsilon_B. \tag{30}$$

Substituting this into (29) yields

$$\begin{aligned}\|R_B B_{t+1}\|_F^2 &\le \|R_B B_t\|_F^2 - 2\eta\big(\alpha_B\|R_B B_t\|_F^2 - \varepsilon_B\big) + \eta^2\|R_B\nabla_B\mathcal{J}_t(B_t, A_t)\|_F^2 \\ &= (1 - 2\eta\alpha_B)\|R_B B_t\|_F^2 + 2\eta\varepsilon_B + \eta^2\|R_B\nabla_B\mathcal{J}_t(B_t, A_t)\|_F^2.\end{aligned}$$

Next, by Assumption 3, specifically (17),

$$\|R_B \nabla_B \mathcal{J}_t(B_t, A_t)\|_F^2 \le \beta_B \|R_B B_t\|_F^2 + \delta_B. \tag{31}$$

Substituting this bound into (31), we obtain

$$\begin{aligned}\|R_B B_{t+1}\|_F^2 &\le (1 - 2\eta\alpha_B)\|R_B B_t\|_F^2 + 2\eta\varepsilon_B + \eta^2\big(\beta_B \|R_B B_t\|_F^2 + \delta_B\big) \\ &= \big(1 - 2\eta\alpha_B + \eta^2\beta_B\big)\|R_B B_t\|_F^2 + 2\eta\varepsilon_B + \eta^2\delta_B.\end{aligned}$$

By the definition of κ_B in (20),

$$\|R_B B_{t+1}\|_F^2 \le \kappa_B \|R_B B_t\|_F^2 + 2\eta\varepsilon_B + \eta^2\delta_B. \tag{32}$$

Since $e_t^B = \|R_B B_t\|_F$, this proves

$$(e_{t+1}^B)^2 \le \kappa_B (e_t^B)^2 + 2\eta\varepsilon_B + \eta^2\delta_B, \tag{33}$$

which is (21).

Now define $x_t := (e_t^B)^2$ and $c_B := 2\eta\varepsilon_B + \eta^2\delta_B$. Then

$$x_{t+1} \le \kappa_B x_t + c_B. \tag{34}$$

Iterating the recursion gives

$$\begin{aligned}x_t &\le \kappa_B^{t-1} x_1 + c_B \sum_{j=0}^{t-2} \kappa_B^j \\ &= \kappa_B^{t-1} x_1 + \frac{c_B}{1-\kappa_B}\big(1 - \kappa_B^{t-1}\big),\end{aligned}$$

because $\kappa_B < 1$. Therefore,

$$(e_t^B)^2 \le \kappa_B^{t-1}(e_1^B)^2 + \frac{2\eta\varepsilon_B + \eta^2\delta_B}{1-\kappa_B}\big(1 - \kappa_B^{t-1}\big), \tag{35}$$

which proves (23).

The proof of (22) and (24) is identical with

$$R_A := I - P_A \tag{36}$$

acting on the right, using (16) and (18).

Finally, since

$$(I - P_B)\Delta W_t = (I - P_B) B_t A_t, \tag{37}$$

the submultiplicativity of the Frobenius and spectral norms gives

$$\|(I - P_B)\Delta W_t\|_F \le \|(I - P_B) B_t\|_F \|A_t\|_2 = e_t^B \|A_t\|_2. \tag{38}$$

Similarly,

$$\Delta W_t (I - P_A) = B_t A_t (I - P_A), \tag{39}$$

hence

$$\|\Delta W_t (I - P_A)\|_F \le \|B_t\|_2 \|A_t (I - P_A)\|_F = \|B_t\|_2 e_t^A. \tag{40}$$

This proves (25). □

C EXPERIMENT

C.1 DATASET

We utilize the three real-world benchmarks proposed by Luu et al. (2022) to evaluate performance under temporal shifts:

- *AI venue classification (AIC)*: We utilize the AIC dataset proposed by Luu et al. (2022), which constitutes a binary classification benchmark explicitly demonstrated to exhibit significant temporal distribution shifts. The task requires discriminating between papers published in AAAI and ICML across four sequential time intervals: 2009–2011, 2012–2014, 2015–2017, and 2018–2020. This setup serves as a challenging proxy for topic classification and author disambiguation under evolving data dynamics. A representative instance: *Input:* "PCA-SVM-Based Comprehensive Evaluation for Customer Relationship Management System of Power Supply Enterprise" *Output:* ICML (vs. AAAI)
- *News source classification (NewsCLS)*: Derived from the metadata of the Newsroom dataset, this benchmark constitutes a three-way classification task designed to identify publication sources(Fox News, The New York Times, and Washington Post) across four sequential time intervals: 2009–2010, 2011–2012, 2013–2014, and 2015–2016, the first three domains used for training and the last for testing. A representative instance: *Input:* A Muslim woman said Sunday that her viral article explaining why she voted for Donald Trump has angered her liberal pals as well as other Muslims. *Output:* FoxNews (vs NYTimes or WaPost)
- *Review rating classification (Yelp)*: This canonical sentiment analysis task involves predicting the numerical rating assigned by an author based on the review text Pang et al. (2002); Dave et al. (2003). Adopting the temporal partitioning strategy from Luu et al. (2022), the dataset spans the years 2013 through 2019. To enhance training efficiency, we subsample the first 10,000 instances from each year. The data corresponding to the initial four years (2013–2016) constitutes the training set, while the subsequent three years (2017–2019) are reserved for testing. *Input:* Best thai food in the area. Everything was authentic and delicious. Will definitely be back again and again. *Output:* 5.0

C.2 Experimental Setup

We employ a DistilBERT backbone (max_len=512, batch size=2, LoRA $r = 8$) and benchmark against three baseline strategies: *Offline* (training on the union of all domains), *LastDomain* (using only the most recent domain), and *IncFinetune* (sequential training). All baselines utilize a 90/10 train-validation split and an early stopping protocol with a patience of 10 and an evaluation interval of 50 steps. For *IncFinetune*, early stopping triggers the transition to the subsequent domain. In contrast, our proposed MaT-LoRA follows a fixed training schedule: 2 epochs for AIC and Yelp, and 10 epochs for NewsCLS to accommodate its larger scale.