DoMA: Geometry-Aware Aggregation for Stable Continual Federated Fine-Tuning on the Heterogeneous Edge

Anonymous authors
Paper under double-blind review

Abstract

Federated fine-tuning (FFT) adapts large models to decentralized, privacy-sensitive data through parameter-efficient updates. Existing methods predominantly consider heterogeneous client updates in Euclidean space, implicitly treating independently optimized directions as linearly compatible. Under non-IID and continually evolving data, however, these directions encode distinct client- and task-specific semantics; Euclidean operation can attenuate, cancel, or distort informative components. Repeated across adaptation stages, this cross-client directional mismatch accumulates into global trajectory drift, forcing a stability–plasticity trade-off between acquiring new knowledge and preserving earlier capabilities. To address this, we propose **DoMA** (*Drift-Controlled On-Manifold Aggregation*), which jointly addresses these spatial and temporal sources of interference. DoMA decouples update magnitude from direction, computes intrinsic Riemannian Fréchet means on a product-sphere manifold, and adaptively regulates inter-stage displacement. Extensive experiments across task-isolated language adaptation, sequential task-category learning, and multimodal continual transfer show that DoMA consistently improves both adaptation and retention, outperforming existing FFT methods by up to 8.72% in task performance and 8.38% in new-task plasticity, while lowering average forgetting by up to 5.99% compared to benchmarks. Code is available at `https://anonymous.4open.science/r/DoMA-453D/`.

1 Introduction

Deploying large language models (LLMs) in privacy-sensitive and resource-heterogeneous environments requires adapting them to decentralized data without transferring raw client samples. Parameter-efficient fine-tuning (PEFT), particularly low-rank adaptation (LoRA) Hu et al. (2022), makes local adaptation feasible by updating only a small fraction of model parameters, while federated fine-tuning (FFT) coordinates these updates across clients Zhang et al. (2024a); Wang et al. (2024). Real deployments, however, are rarely stationary: client populations, data distributions, and task requirements evolve continually, requiring the global model to acquire new capabilities while preserving previously learned ones Yu et al. (2024). This makes continual FFT fundamentally more challenging than either centralized PEFT or single-stage federated adaptation.

Continual FFT is vulnerable to update interference. Each client follows an adaptation trajectory determined by its local data, task semantics, and computational budget. Under non-IID data, these independently learned updates can disagree substantially, and repeated merging propagates this disagreement across both clients and adaptation stages Xu et al. (2024); Gao et al. (2022). Figure 1 illustrates the resulting instability. Held-out perplexity provides a task-agnostic measure of whether adaptation preserves language-model competence. Whereas centralized fine-tuning remains comparatively stable, federated adaptation exhibits persistent perplexity growth as tuning pro-

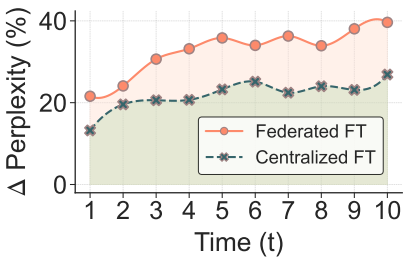


Figure 1: **Instability under continual fine-tuning.** Federated adaptation accumulates greater held-out language-model drift than centralized fine-tuning.

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.

ceeds. This widening gap shows that decentralized aggregation progressively destabilizes the global model, motivating an aggregation mechanism that can preserve both task adaptation and general language-model behavior.

We trace this instability to a coupled spatial–temporal limitation in existing FFT methods. Conventional aggregators combine heterogeneous low-rank updates through Euclidean operations, implicitly assuming that independently optimized client adaptations are mutually compatible. Under non-IID data, however, their directions and scales encode distinct client- and task-specific semantics, such that linear merging can attenuate, cancel, or distort informative components. DoRA Liu et al. (2024) improves local fine-tuning by decoupling weight magnitude from direction, but remains a local reparameterization rather than a federated coordination mechanism. When optimized independently across clients, DoRA provides neither a principled global consensus for heterogeneous directions nor a mechanism for preventing their disagreement from propagating across stages. Naively applying it to FFT therefore retains cross-client interference, which becomes embedded in successive global updates and accumulates into trajectory drift. Continual FFT must instead jointly achieve *stability*, preserving earlier capabilities, and *plasticity*,acquiring incoming tasks without improving one by suppressing the other.

We introduce DoMA (*Drift-Controlled On-Manifold Aggregation*), a unified framework that coordinates both the spatial structure of heterogeneous client updates and their temporal evolution. DoMA separates each local update into magnitude and directional components, computes an intrinsic Riemannian Fréchet mean of the directions on a product-sphere manifold, and coordinates their magnitudes separately. This formulation preserves directional consensus without forcing geometrically incompatible updates through a shared linear average. DoMA further introduces stage-adaptive drift control to regulate trajectory displacement as continual adaptation proceeds, limiting destabilizing changes without restricting subsequent task acquisition. The two mechanisms address complementary failure modes: on-manifold aggregation resolves cross-client geometric misalignment, while drift control prevents the remaining disagreement from accumulating across stages. Our main contributions are summarized as:

- **We propose DoMA**, a drift-controlled on-manifold aggregation framework for continual FFT that jointly coordinates heterogeneous client updates across both geometric and temporal dimensions.
- **DoMA improves federated adaptation performance** across task-isolated language adaptation, sequential task-category learning, heterogeneous model scales, and multimodal continual transfer, improving final average accuracy and new-task plasticity by up to 8.72% and 8.38%, respectively.
- **DoMA enhances post-adaptation stability** by preserving broad language-model competence under evolving client data, consistently reducing held-out perplexity drift and accumulated optimization loss across homogeneous and heterogeneous settings.
- **DoMA mitigates catastrophic forgetting without suppressing task acquisition.** Under sequential task-category adaptation, it lowers average forgetting by up to 5.99 percentage points relative to the strongest-plasticity baseline while improving final multi-task performance and backward transfer.
- **We provide a geometric perspective on continual FFT** that connects cross-client directional misalignment with cross-stage trajectory drift. Matched factorial analyses verify that intrinsic aggregation and adaptive drift control are critical to stable adaptation.

2 PRELIMINARIES

Low-rank adaptation. Given a pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA Hu et al. (2022) constrains its update to $\Delta W = BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and $r \ll \min(d, k)$. The adapted weight is

$$W' = W_0 + \Delta W = W_0 + BA, \tag{1}$$

with only (B, A) being trainable. This low-dimensional parameterization reduces both optimization and communication costs.

Weight-decomposed low-rank adaptation. DoRA Liu et al. (2024) improves local adaptation by decoupling the magnitude and direction of a weight matrix:

$$W_0 = m \odot \frac{V}{\|V\|_c}, \qquad m = \|W_0\|_c, \tag{2}$$

where $\|\cdot\|_c$ denotes column-wise ℓ_2 norms. During adaptation, the magnitude becomes trainable and the directional parameter is updated as $V' = V + BA$, yielding

$$W' = m' \odot \frac{V'}{\|V'\|_c}. \tag{3}$$

This decomposition improves the expressiveness of an individual adapter, but does not specify how independently optimized client components should be coordinated or how their disagreement should be controlled across continual federated stages.

Federated fine-tuning. In FFT, a server coordinates K clients with private datasets. At round t, client k computes a local update $\Delta W_k^{(t)}$ from the global model $W^{(t)}$, and the server performs

$$W^{(t+1)} = W^{(t)} + \sum_{k=1}^{K} p_k \Delta W_k^{(t)}, \tag{4}$$

where p_k is the aggregation weight. With LoRA or DoRA, clients exchange parameter-efficient updates rather than full model parameters. The key challenge is therefore not only representing each local update, but coordinating heterogeneous updates without accumulating interference.

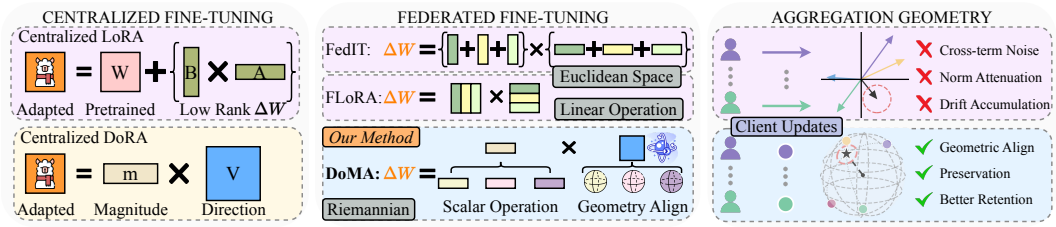


Figure 2: **Centralized and federated fine-tuning.** Existing FFT methods use Euclidean aggregation, whereas DoMA coordinates heterogeneous updates through on-manifold aggregation and adaptive drift control.

Limitations of Euclidean FFT aggregation. Existing methods encounter two distinct limitations. First, factor-wise aggregation may not reconstruct the intended global update. In FedIT Zhang et al. (2024a), client k produces $\Delta W_k = B_k A_k$, while the desired FedAvg McMahan et al. (2023) update is

$$\Delta W^\star = \sum_{k=1}^{K} p_k B_k A_k. \tag{5}$$

Let $\bar{B} = \sum_k p_k B_k$ and $\bar{A} = \sum_k p_k A_k$ denote the independently averaged factors. Their product becomes

$$\Delta W^{\text{FedIT}} = \bar{B}\bar{A} = \sum_k p_k^2 B_k A_k + \sum_{i \neq j} p_i p_j B_i A_j. \tag{6}$$

The cross-client terms are absent from $\Delta W^\star$. Moreover, because $BA = (BQ)(Q^{-1}A)$ for any invertible Q, averaging client factors is not invariant to equivalent representations of the same update.

FLoRA Wang et al. (2024) resolves this algebraic inconsistency by stacking client factors so that their product reconstructs the weighted Euclidean update exactly. Exact reconstruction, however, does not ensure effective coordination. Let $\langle \cdot, \cdot \rangle_F$ denote the Frobenius inner product. The norm of the aggregated update satisfies

$$\|\Delta W^\star\|_F^2 = \sum_k p_k^2 \|\Delta W_k\|_F^2 + 2 \sum_{i<j} p_i p_j \langle \Delta W_i, \Delta W_j \rangle_F. \tag{7}$$

Under non-IID data, weakly aligned updates provide little mutual reinforcement, while opposing updates partially cancel and attenuate the global adaptation signal. In continual FFT, each aggregated update also initializes the next stage, allowing residual directional mismatch to propagate through subsequent client trajectories and accumulate into global drift. Effective continual FFT therefore requires both structure-aware coordination of heterogeneous updates and explicit regulation of their temporal accumulation.

3 DoMA: Drift-Controlled On-Manifold Aggregation

We propose DoMA (*Drift-Controlled On-Manifold Aggregation*), a continual FFT framework that jointly coordinates heterogeneous client updates across space and time. Its central premise is that stable continual adaptation requires more than an expressive local adapter: the server must construct a coherent global update from geometrically heterogeneous client information and prevent residual disagreement from accumulating across stages. DoMA addresses these two requirements through three coupled components: (i) a geometry-bearing decomposition of the global update, (ii) local adaptation with stage-adaptive drift control, and (iii) intrinsic aggregation on a product-sphere manifold. Figure 3 summarizes the framework.

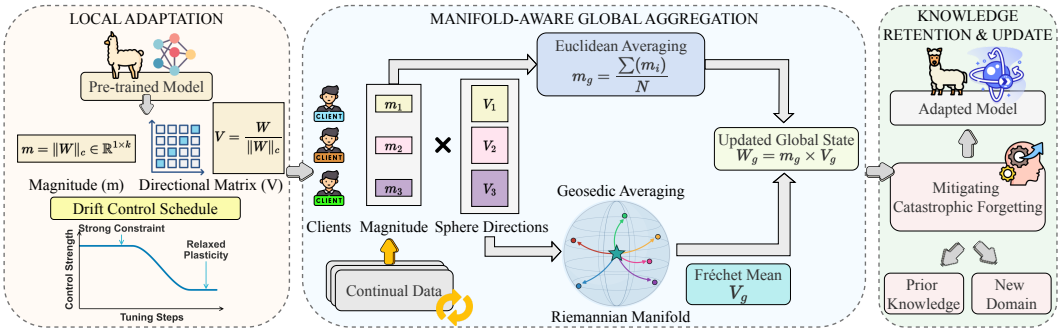


Figure 3: **Overview of DoMA.** DoMA separates update magnitude from direction, regulates local trajectory displacement, and constructs global directional consensus through intrinsic on-manifold aggregation.

3.1 Geometry-Bearing Update Decomposition

Given a weight matrix W, we represent its column-wise magnitude and direction as

$$W = m \odot \hat{V}, \qquad m = \|W\|_c, \qquad \hat{V} = \mathrm{Normalize}(V), \tag{8}$$

where $\|\cdot\|_c$ denotes column-wise ℓ_2 norms and $\mathrm{Normalize}(\cdot)$ normalizes each column. This decomposition separates update strength from update orientation. Although motivated by the magnitude–direction parameterization used for local adaptation Liu et al. (2024), DoMA assigns it a fundamentally different role: it becomes the interface through which independently optimized clients form a global geometric consensus.

For $V \in \mathbb{R}^{d\times k}$, every column of $\hat{V}$ has unit norm. We therefore model the directional state on the product manifold

$$\mathcal{M} = (\mathbb{S}^{d-1})^k, \qquad \mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : \|v\|_2 = 1\}. \tag{9}$$

For $\hat{U}, \hat{V} \in \mathcal{M}$, their product-manifold distance is

$$d^2_{\mathcal{M}}(\hat{U}, \hat{V}) = \sum_{j=1}^{k} \arccos^2(\mathrm{clip}(\langle \hat{u}_j, \hat{v}_j \rangle, -1, 1)) \,. \tag{10}$$

This distance measures column-wise angular disagreement rather than coordinate-wise Euclidean deviation, providing a geometry-consistent notion of proximity between heterogeneous client directions.

Given directional states $\{\hat{V}_i\}_{i\in\mathcal{S}_t} \subset \mathcal{M}$ from clients participating at stage t, their weighted Riemannian Fréchet mean is

$$\arg\min_{\hat{V}\in\mathcal{M}} \sum_{i\in\mathcal{S}_t} p_i \, d^2_{\mathcal{M}}(\hat{V}, \hat{V}_i), \tag{11}$$

where p_i is the aggregation weight. Unlike coordinate-wise averaging, this objective directly minimizes intrinsic directional disagreement. Under the bounded-dispersion condition described in Appendix D, the mean is unique and defines a stable global consensus.

3.2 Local Adaptation with Stage-Adaptive Drift Control

At stage t, the server distributes the current global state $\{m^{(t)}, \hat{V}^{(t)}\}$ together with a lightweight drift policy $\Pi^{(t)}$. Client i initializes its trainable PEFT parameters $\theta_i = \{m_i, A_i, B_i\}$ from the received state and records their stage-wise anchors $\theta_{p,i}^{(t,0)}$, $p \in \{m, A, B\}$. Local adaptation minimizes

$$\mathcal{L}_i(\theta_i) = \mathcal{L}_{\mathrm{LM}}(\theta_i; D_i) + \lambda_d(\ell, t) \sum_{p \in \{m,A,B\}} \left\| \theta_{p,i}^{(t,\ell)} - \theta_{p,i}^{(t,0)} \right\|_2^2, \tag{12}$$

where ℓ denotes the local optimization step. The second term limits abrupt departures from the current global state, preventing client-specific deviations from dominating the next aggregation.

A fixed constraint can preserve the current model only by restricting its ability to absorb incoming information. DoMA instead adapts the regularization strength to both local progress and continual stage:

$$\lambda_d(\ell, t) = \lambda_0 \exp\left(-\alpha_\ell \bar{\ell} - \alpha_t \bar{t}\right), \qquad \bar{\ell} = \frac{\ell - 1}{L - 1}, \quad \bar{t} = \frac{t - 1}{T - 1}. \tag{13}$$

The controller applies stronger anchoring when optimization is most susceptible to disruptive displacement and progressively restores adaptation flexibility as learning proceeds. It therefore improves stability without obtaining low drift through under-adaptation.

3.3 Intrinsic Global Aggregation

After local training, DoMA coordinates magnitudes and directions according to their respective structures. For participating clients $\mathcal{S}_t$, magnitudes are aggregated in Euclidean space:

$$m^{(t+1)} = \sum_{i \in \mathcal{S}_t} p_i m_i^{(t)}, \qquad p_i = \frac{|D_i|}{\sum_{j \in \mathcal{S}_t} |D_j|}. \tag{14}$$

Each client induces a directional state from the current global directional reference $V^{(t)}$ and its low-rank adaptation:

$$V_i^{(t)} = V^{(t)} + \Delta_i^{(t)}, \qquad \Delta_i^{(t)} = \frac{\alpha}{r_i} B_i^{(t)} A_i^{(t)}, \qquad \hat{V}_i^{(t)} = \mathrm{Normalize}(V_i^{(t)}), \tag{15}$$

where r_i may vary across clients. The server then computes

$$\hat{V}^{(t+1)} = \arg\min_{\hat{V} \in \mathcal{M}} \sum_{i \in \mathcal{S}_t} p_i \, d_{\mathcal{M}}^2\left(\hat{V}, \hat{V}_i^{(t)}\right). \tag{16}$$

We solve Equation equation 16 using Riemannian descent. Starting from the normalized Euclidean mean, iteration q computes the intrinsic consensus direction

$$G^{[q]} = \sum_{i \in \mathcal{S}_t} p_i \, \log_{\hat{V}^{[q]}}\left(\hat{V}_i^{(t)}\right), \qquad \hat{V}^{[q+1]} = \exp_{\hat{V}^{[q]}}\left(\eta_q G^{[q]}\right), \tag{17}$$

where $\log$ and $\exp$ are applied column-wise on the product-sphere manifold. The resulting global model is

$$W^{(t+1)} = m^{(t+1)} \odot \hat{V}^{(t+1)}. \tag{18}$$

The global directional state is cached and propagated to the next adaptation stage, avoiding the need to identify a single global low-rank factorization.

DoMA thus establishes two complementary forms of coordination: intrinsic aggregation removes cross-client geometric mismatch at each stage, while adaptive drift control limits the temporal propagation of the remaining disagreement. Their combination enables the global model to acquire new information without destabilizing previously learned capabilities. Complete pseudocode and theoretical analysis are provided in Appendices C and D.

4 Experiments

4.1 Experimental Setup

Benchmarks and models. We evaluate DoMA on eight commonsense reasoning (CR) benchmarks Liu et al. (2024) and seven natural language understanding (NLU) benchmarks, including tasks

from GLUE Wang et al. (2018). We use Qwen3-1.7B and Qwen3-8B Yang et al. (2025) for CR and TinyLlama Zhang et al. (2024b) for NLU. Clients hold non-IID data and heterogeneous adaptation budgets, and participating clients are sampled at each stage from a population of 100 clients. We compare against representative homogeneous- and heterogeneous-rank federated fine-tuning methods. Complete dataset, baseline, and implementation details are provided in Appendix B.1.

Continual protocols and metrics. We consider two complementary protocols. The *task-isolated protocol* continually adapts each benchmark from the pretrained model and reports downstream accuracy and ΔPPL on held-out WikiText-2 Merity et al. (2016), measuring task performance and language-model stability. The *sequential task-category protocol* adapts one model across tasks without resetting. Over five task orders, we report average accuracy (AA), new-task plasticity, and average forgetting (AF) to evaluate cross-task acquisition and retention. Metric definitions and results are provided in Appendix B.3. Code is available at `https://anonymous.4open.science/r/DoMA-453D/`.

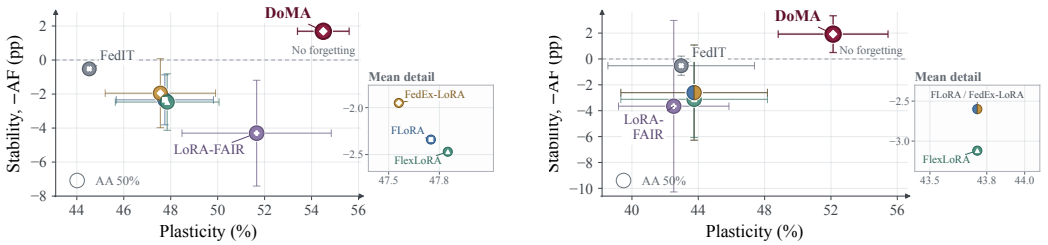
4.2 Plasticity-Qualified Sequential Task-Category Adaptation

The task-isolated results evaluate continual data arrival within a fixed task category. We now consider a stricter setting in which the task category itself changes over time. A single global model is sequentially adapted across BoolQ, PIQA, SIQA, and HellaSwag without resetting. Results are averaged over five task-order trajectories. Unlike task-isolated evaluation, this protocol directly exposes cross-task interference and distinguishes genuine retention from insufficient acquisition of incoming tasks.

Low forgetting can conceal under-adaptation. Forgetting alone cannot distinguish effective retention from insufficient acquisition of incoming tasks. On Qwen3-1.7B, FedIT obtains a seemingly favorable AF of 0.52%, but reaches only 44.53% plasticity and 44.14% final AA. Baselines that acquire incoming tasks more effectively generally incur greater interference with earlier categories, demonstrating why forgetting must be interpreted jointly with new-task acquisition and final multi-task performance.

Table 1: **Plasticity-qualified sequential task-category adaptation.** Mean ± standard deviation over five task-order trajectories. Higher AA and plasticity and lower AF are better; negative AF denotes positive backward improvement. Best means are bolded.

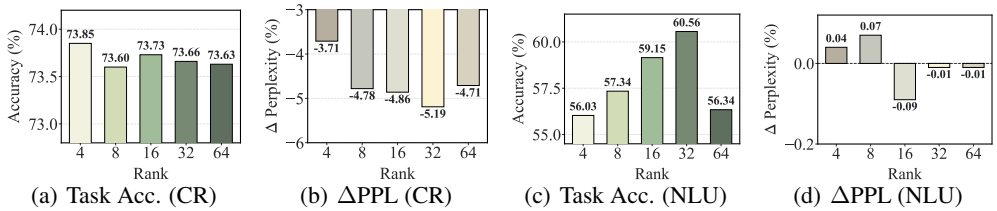
Qwen3-1.7B				Qwen3-8B			
Method	AA ↑	Plas. ↑	AF ↓	Method	AA ↑	Plas. ↑	AF ↓
FedIT	44.14 ± 0.00	44.53 ± 0.00	0.52 ± 0.00	FedIT	44.53 ± 2.21	42.97 ± 4.42	0.52 ± 0.74
FLoRA	46.09 ± 0.83	47.75 ± 2.07	2.34 ± 1.47	FLoRA	47.27 ± 3.87	43.75 ± 4.42	2.60 ± 3.68
FedEx-LoRA	46.09 ± 0.83	47.56 ± 2.35	1.95 ± 2.03	FedEx-LoRA	47.27 ± 3.87	43.75 ± 4.42	2.60 ± 3.68
FlexLoRA	46.09 ± 0.83	47.85 ± 2.21	2.47 ± 1.66	FlexLoRA	46.88 ± 3.31	43.75 ± 4.42	3.12 ± 2.95
LoRA-FAIR	49.22 ± 0.55	51.66 ± 3.18	4.30 ± 3.11	LoRA-FAIR	47.66 ± 3.31	42.52 ± 3.33	3.65 ± 6.63
DoMA	**56.57** ± 0.97	**54.50** ± 1.10	**-1.69** ± 0.18	**DoMA**	**56.38** ± 3.87	**52.13** ± 3.31	**-1.92** ± 1.42



(a) Qwen3-1.7B (b) Qwen3-8B

Figure 4: **Stability–plasticity frontiers.** Results on Qwen3-1.7B (left) and Qwen3-8B (right). Upper right is better; marker area denotes final average accuracy (AA).

DoMA shifts the stability–plasticity frontier. As shown in Table 1 and Figure 4, DoMA improves both adaptation and retention. On Qwen3-1.7B, it reaches 56.57% final AA and 54.50% plasticity, exceeding the strongest baselines by 7.35% and 2.84%, respectively, while reducing AF to -1.69%.



(a) Task Acc. (CR) (b) ΔPPL (CR) (c) Task Acc. (NLU) (d) ΔPPL (NLU)

Figure 5: **Robustness across adapter ranks.** Task accuracy and auxiliary held-out language-model drift under varying LoRA ranks on CR (SIQA) and NLU (WNLI).

On Qwen3-8B, DoMA improves AA and plasticity by 8.72% and 8.38% and attains an AF of -1.92%. It is the only evaluated method with negative mean AF at both model scales, showing that its stability does not arise from suppressing adaptation. Instead, DoMA shifts the frontier by jointly improving new-task acquisition, final multi-task performance, and backward transfer. Complete trajectory-level results are reported in Appendix B.3, while a matched factorial analysis confirms the complementary roles of Riemannian aggregation and drift regulation in Appendix B.3.2.

4.3 Task-Isolated Continual Federated Adaptation

Downstream adaptation performance. Table 2 summarizes the task-isolated results, with complete per-task comparisons reported in Appendix B.2. DoMA achieves the highest average federated accuracy in every model–heterogeneity block. On CR, it improves over the strongest prior method by 1.48 and 2.01 points for Qwen3-1.7B under homogeneous and heterogeneous ranks, and by 1.40 and 1.94 points for Qwen3-8B. The advantage is especially pronounced under direct rank alignment: DoMA exceeds zero-padding by 22.87 points on Qwen3-1.7B with heterogeneous ranks. The trend extends to NLU, where DoMA reaches 71.65% and 71.50% average accuracy, improving over the strongest baselines by 4.09 and 2.87 points. These consistent gains show that on-manifold coordination consolidates heterogeneous client information into a stronger downstream model.

Table 2: **Task-isolated continual FFT.** Average task accuracy and relative perplexity change on held-out text. Higher accuracy and lower ΔPPL are better. Best and second-best results are bolded and underlined. Complete results are in Appendix B.2.

Task / Model	Metric	Homogeneous						Heterogeneous					
		FedIT	FLoRA	FedEx	Flex	L-FAIR	**DoMA**	ZP	FLoRA	FedEx	Flex	L-FAIR	**DoMA**
CR / Qwen3-1.7B	Acc. ↑	66.54	<u>70.46</u>	67.73	70.17	69.73	**71.94**	49.52	<u>70.38</u>	68.70	70.36	69.62	**72.39**
	ΔPPL ↓	21.15	59.30	18.52	17.10	<u>13.85</u>	**11.81**	16.24	21.38	16.14	<u>15.18</u>	15.55	**10.08**
CR / Qwen3-8B	Acc. ↑	<u>85.03</u>	81.93	80.93	82.10	81.55	**86.43**	82.32	72.03	<u>85.22</u>	83.26	84.26	**87.16**
	ΔPPL ↓	−5.99	<u>−10.27</u>	−10.17	−9.43	−9.64	**-15.30**	−1.86	−7.76	<u>−10.11</u>	−9.54	−9.35	**-15.25**
NLU / TinyLlama	Acc. ↑	65.93	<u>67.56</u>	65.88	66.66	67.31	**71.65**	64.87	<u>68.63</u>	67.10	67.66	67.89	**71.50**
	ΔPPL ↓	4.23	6.59	0.77	0.55	<u>−0.02</u>	**-1.26**	<u>−0.98</u>	3.53	1.42	0.46	−0.45	**-3.75**

Language-model stability. Held-out ΔPPL complements downstream accuracy by measuring how well adaptation preserves general language-model behavior. DoMA achieves the lowest average ΔPPL in all six model–heterogeneity blocks, improving over the strongest prior result by 2.04–5.14 points on CR and by 1.24 and 2.77 points on NLU. The effect is most pronounced on Qwen3-8B, where DoMA reaches -15.30% and -15.25% under homogeneous and heterogeneous ranks, respectively. Thus, its accuracy gains are accompanied by greater, rather than weaker, language-model stability. Retention of previously acquired task capabilities under sequential task-category adaptation is evaluated using standard continual-learning metrics in Section 4.2.

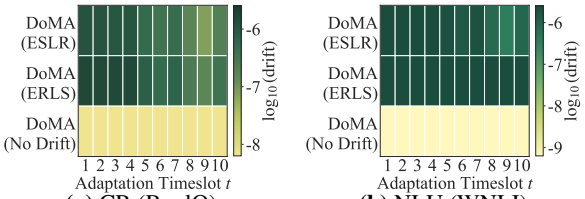
Robustness across adapter ranks. Figure 5 evaluates sensitivity to the available LoRA rank. DoMA maintains stable CR accuracy over $r \in \{4, \ldots, 64\}$ while producing consistently low ΔPPL. On NLU, accuracy initially improves with rank and peaks near $r = 32$, followed by a mild decline at larger ranks, suggesting that additional adaptation capacity eventually yields diminishing returns. Across both task families, the relatively stable perplexity curves indicate that DoMA remains robust as client adaptation capacity varies.

4.4 Multimodal Continual Transfer

We further evaluate DoMA on multi-stage vision–language adaptation with Qwen2-VL-2B-Instruct across multiple task orders. Table 3 shows that DoMA extends its stability–plasticity advantage beyond text, achieving the best AA, plasticity, retention, and AF. Relative to the strongest baseline for each metric, DoMA improves AA and plasticity by 4.5% and 5.4%, respectively, while reducing forgetting by 13.1%. The simultaneous gains reflect more effective cross-task knowledge consolidation. These results demonstrate stronger multimodal acquisition without compromising previously learned visual capabilities. Additional vision–language results are provided in Appendix F.

Table 3: **Results on multi-modal continual adaptation across task-order trajectories.**

Method	AA ↑	AF ↓	Ret. ↑	Plas. ↑
FedIT	69.1	6.1	78.7	70.9
FLoRA	68.4	7.0	79.2	71.9
FlexLoRA	68.8	8.6	78.6	72.5
LoRA-FAIR	68.2	7.5	80.1	72.5
DoMA	**72.2**	**5.3**	**80.7**	**76.4**

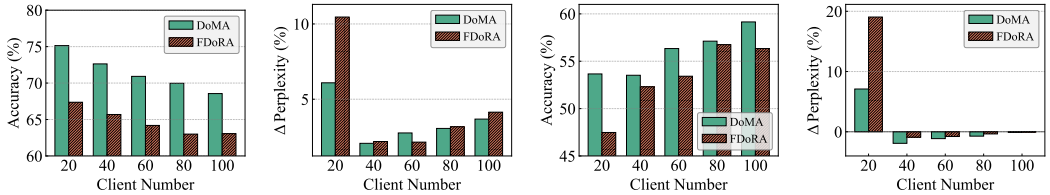


(a) CR (BoolQ) **(b)** NLU (WNLI)

Figure 6: **Drift-control trajectories.** Log-scale drift variations on DoMA variants.

4.5 Ablation Studies

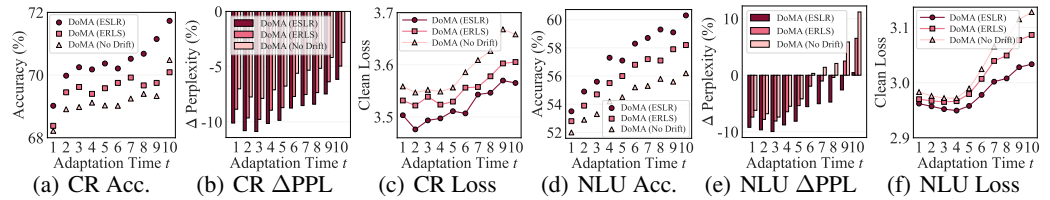
Effect of aggregation geometry. Figure 7 isolates the effect of manifold-aware directional aggregation by comparing DoMA with its controlled Euclidean counterpart, FDoRA, as the number of heterogeneous clients increases from 20 to 100. Across CR (BoolQ) and NLU (WNLI), DoMA consistently maintains higher task accuracy and lower held-out perplexity drift, improving accuracy by up to 7.8 points on BoolQ and 6.2 points on WNLI while reducing ΔPPL by up to 4.4 points. These results show that respecting the geometry of directional updates improves both knowledge consolidation and language-model stability under increasing client heterogeneity; the matched sequential-task analysis in Appendix B.3.2 further confirms this advantage using standard continual-learning metrics.



(a) Task Acc. (CR) (b) ΔPPL (CR) (c) Task Acc. (NLU) (d) ΔPPL (NLU)

Figure 7: **Effect of aggregation geometry.** Comparison of DoMA and its Euclidean counterpart under increasing client heterogeneity on CR (BoolQ) and NLU (WNLI). DoMA maintains higher task accuracy and lower auxiliary language-model drift.

Effect of adaptive drift control. We compare early-strong–late-relaxed scheduling (ESLR) with late-strong–early-relaxed scheduling (ERLS) and a variant without drift control. Across BoolQ and WNLI, ESLR yields higher accuracy, lower ΔPPL, and less accumulated loss (Figure 8), while limiting large early displacements and gradually restoring adaptation flexibility (Figure 6). These results show that adaptive drift control stabilizes early optimization without suppressing subsequent task acquisition. The sequential-task factorial analysis in Appendix B.3.2 further demonstrates that drift regulation complements, rather than replaces, Riemannian aggregation.



(a) CR Acc. (b) CR ΔPPL (c) CR Loss (d) NLU Acc. (e) NLU ΔPPL (f) NLU Loss

Figure 8: **Effect of adaptive drift control.** ESLR provides a more favorable adaptation trajectory on CR (BoolQ) and NLU (WNLI), combining higher task accuracy with lower auxiliary language-model drift and reduced loss accumulation.

5 DISCUSSION

Why DoMA stabilizes continual FFT. At stage t, let $\mu_t = \sum_{i \in \mathcal{S}_t} p_i \delta_{\hat{V}_i^{(t)}}$ denote the empirical distribution of client directions. DoMA sets $\hat{V}^{(t+1)} = \mathrm{bar}(\mu_t)$ through the Riemannian Fréchet barycenter and measures inter-stage drift intrinsically as

$$\mathcal{D}_t = d_{\mathcal{M}}\big(\hat{V}^{(t+1)}, \hat{V}^{(t)}\big). \tag{19}$$

Proposition 5.1 (Geometric–temporal stability)**.** *Under the bounded-dispersion and local-regularity conditions in Appendix E, DoMA satisfies*

$$\mathcal{D}_t \leq \frac{C_\varepsilon L_{\mathrm{dir}} \Gamma_t}{2\lambda_t}, \tag{20}$$

where C_ε characterizes barycenter sensitivity, L_{dir} the local direction map, Γ_t the effective optimization residual, and λ_t the drift-control strength.

Equation equation 20 highlights the complementary roles of DoMA: on-manifold aggregation reduces cross-client directional dispersion, while drift control bounds global displacement. Scheduling λ_t tightly constrains early updates and gradually restores flexibility, limiting representational overwriting without suppressing new-task acquisition. Proofs and the cumulative bound appear in Appendix E. Riemannian operations are server-side and add no client parameters or adapter payload; communication costs are reported in Appendix E.4.

6 RELATED WORK

Parameter-Efficient Fine-Tuning. PEFT adapts large models by optimizing a small parameter subset. LoRA Hu et al. (2022) learns low-rank weight updates, while DoRA Liu et al. (2024) improves local adaptation by decoupling weight magnitude and direction. Other approaches include quantized fine-tuning Dettmers et al. (2023), prefix-tuning Li & Liang (2021), and adapters Houlsby et al. (2019). Although effective for centralized adaptation, these methods do not specify how independently learned updates should be coordinated across heterogeneous clients or continual stages.

Federated Fine-Tuning. FFT combines local parameter-efficient updates through server-side aggregation. FedIT Zhang et al. (2024a) federates LoRA through factor averaging, whereas FLoRA Wang et al. (2024) supports heterogeneous ranks through exact update reconstruction. FlexLoRA Bai et al. (2024), FedEx-LoRA Singhal et al. (2024), and LoRA-FAIR Bian et al. (2025) improve heterogeneous adapter aggregation. However, these methods remain based on Euclidean coordination and do not preserve directional structure or regulate how aggregation errors propagate across stages.

Continual Federated Adaptation. Sequential domain shifts and client heterogeneity can cause federated aggregation to overwrite previously acquired capabilities Luo et al. (2025); Li et al. (2024). Federated continual-learning methods mitigate such interference through regularization, replay, or optimization control Yoon et al. (2021); Zhu et al. (2021), but are not designed for the geometry and heterogeneous ranks of parameter-efficient updates. DoMA instead couples intrinsic aggregation with adaptive trajectory regulation, jointly addressing cross-client misalignment and cross-stage drift in continual FFT.

7 CONCLUSION

We presented DOMA (*Drift-Controlled On-Manifold Aggregation*), a geometric framework for stable continual federated fine-tuning. DoMA addresses two coupled sources of interference: cross-client directional misalignment induced by Euclidean aggregation and its accumulation into cross-stage trajectory drift. By separating update magnitude from direction, computing intrinsic directional consensus on a product-sphere manifold, and adaptively regulating inter-stage displacement, DoMA preserves prior capabilities without suppressing adaptation to incoming tasks. Across task-isolated language adaptation, sequential task-category learning, and multimodal continual transfer, DoMA consistently improves task acquisition, final performance, and retention under heterogeneous data and adapter ranks. Our theoretical and factorial analyses further establish the complementary roles of intrinsic aggregation and drift control. These results position geometric–temporal update coordination as a principled foundation for scalable and reliable continual FFT.

Ethics Statement. This work uses public benchmarks and pretrained models and involves no human-subject research or collection of personal data. Although federated adaptation reduces the need to centralize client data, it does not eliminate risks related to privacy leakage, model bias, or harmful downstream use. Practical deployments should therefore include privacy safeguards, access controls, and safety and fairness evaluations. We adhere to the ICLR Code of Ethics.

Reproducibility Statement. The main text specifies the proposed algorithm, evaluation protocols, and metrics. Appendix B provides dataset, baseline, implementation, and hyperparameter details, while Appendix E states the theoretical assumptions and complete proofs. Anonymous source code is available at `https://anonymous.4open.science/r/DoMA-453D/`.

REFERENCES

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. *Advances in Neural Information Processing Systems*, 37:14457–14483, 2024.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. Lora-fair: Federated lora fine-tuning with aggregation and initialization refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3737–3746, 2025.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. Heterogeneous lora for federated fine-tuning of on-device foundation models. *arXiv preprint arXiv:2401.06432*, 2024.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction, 2022. URL `https://arxiv.org/abs/2203.11751`.

Zhihan Guo, Yifei Zhang, Zhuo Zhang, Zenglin Xu, and Irwin King. Fedhlt: Efficient federated low-rank adaption with hierarchical language tree for multilingual modeling. In *Companion Proceedings of the ACM Web Conference 2024*, pp. 1558–1567, 2024.

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.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting catastrophic forgetting in large language model tuning. *arXiv preprint arXiv:2406.04836*, 2024.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

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.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2023. URL `https://arxiv.org/abs/1602.05629`.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. Fedex-lora: Exact aggregation for federated and efficient fine-tuning of foundation models. *arXiv preprint arXiv:2410.09432*, 2024.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *Advances in Neural Information Processing Systems*, 37:22513–22533, 2024.

Binqian Xu, Xiangbo Shu, Haiyang Mei, Zechen Bai, Basura Fernando, Mike Zheng Shou, and Jinhui Tang. Dofit: Domain-aware federated instruction tuning with alleviated catastrophic forgetting. *Advances in Neural Information Processing Systems*, 37:85846–85866, 2024.

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.

Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In *International conference on machine learning*, pp. 12073–12086. PMLR, 2021.

Hao Yu, Xin Yang, Xin Gao, Yan Kang, Hao Wang, Junbo Zhang, and Tianrui Li. Personalized federated continual learning via multi-granularity prompt. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4023–4034, 2024.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. Towards building the federatedgpt: Federated instruction tuning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6915–6919. IEEE, 2024a.

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

Zhuangdi Zhu, Junyuan Hong, and Jiayu Zhou. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, pp. 12878–12889. PMLR, 2021.

A Summary of Notations

We summarize the key notations used throughout the paper in Table 4 for reference.

B Hyperparameter and Experimental Details

We summarize the detailed benchmark configurations and hyperparameter settings used throughout the evaluation. For commonsense reasoning (CR), we adopt eight benchmarks Liu et al. (2024), including BoolQ, PIQA, SIQA, HellaSwag, WinoGrande, ARC-e, ARC-c, and OBQA, covering logical inference, physical reasoning, social reasoning, and multi-step commonsense understanding. For natural language understanding (NLU), we use seven GLUE benchmark subsets Wang et al. (2018), including SST-2, SST-5, MRPC, CoLA, QNLI, RTE, and WNLI.

We evaluate downstream adaptation quality using task-specific accuracy and assess generalization and stability via perplexity (PPL) measured on held-out WikiText-2 Merity et al. (2016). Lower PPL indicates better preservation of pretrained language modeling capability and reduced catastrophic forgetting during continual adaptation. Unless otherwise specified, all experiments use AdamW optimization with a learning rate of 1×10^{-4}. The default LoRA rank is set to $r = 16$ in homogeneous settings, while heterogeneous settings adopt mixed ranks sampled from $[4, 64]$. DoRA decomposition

Table 4: Summary of Notations

Notation	Description
N	Number of clients.
i	Client index, $i \in \{1, \ldots, N\}$.
$t,\ T$	Round index and total rounds.
$D_i,\ \lvert D_i \rvert$	Client i's local data and its size.
p_i	Aggregation weight (e.g., $\lvert D_i \rvert / \sum_j \lvert D_j \rvert$).
$W^{(t)}$	Global model (or an adapted layer) at round t.
W_0	Frozen base weight of an adapted layer.
$\lVert \cdot \rVert_c$	Column-wise ℓ_2 norm.
$\mathrm{Normalize}(\cdot)$	Column-wise normalization to the unit sphere.
$m^{(t)}$	Decomposed magnitude (per-column scale).
$\hat{V}^{(t)}$	Decomposed direction on the unit sphere.
$r,\ \alpha$	LoRA rank and scaling factor.
$A_i^{(t)}, B_i^{(t)}$	Client i's LoRA matrices.
$\Delta_i^{(t)}$	Low-rank update, $\Delta_i^{(t)} = \frac{\alpha}{r} B_i^{(t)} A_i^{(t)}$.
$\Pi^{(t)} = \{\lambda_d, s(\ell), \rho\}$	Drift-control policy broadcast by the server.
λ_d	Drift regularization strength.
$s(\ell),\ \ell$	Drift schedule and local step index.
ρ	Warmup fraction in $s(\ell)$.
$\mathcal{L}_{\mathrm{LM}}$	Local language-model fine-tuning loss.
$\mathcal{L}_{\mathrm{total}}$	Drift-regularized local objective.
$\mathcal{M}$	Product-sphere manifold for $\hat{V}$.
$\log_{\hat{V}}(\cdot)$	Log map on $\mathcal{M}$.
$\exp_{\hat{V}}(\cdot)$	Exp map on $\mathcal{M}$.
η_R	Riemannian step size.
$K,\ \epsilon$	Max manifold iterations and tolerance.
PPL	Perplexity (lower is better; stability proxy).
CF%	Forgetting degree via relative PPL increase.

is applied to attention projection layers, including q, k, v, and output projections. Drift scheduling follows the proposed early-strong–late-relaxed strategy.

All experiments were conducted on a distributed GPU cluster consisting of four servers equipped with 8× NVIDIA RTX 3090 GPUs and one server with 4× NVIDIA L40S GPUs. Additional scalability analysis under partial participation is provided in Appendix G. Detailed hyperparameter settings are summarized in Table 5.

Table 5: Default hyperparameters used in our evaluations. These settings are shared across models and tasks unless otherwise stated.

Group	Category	Symbol	Value (default)
Optimization	Optimizer	–	AdamW
	Learning rate	η	3×10^{-4}
	Batch size	B	128
	Micro-batch	b	16
	Warmup steps	–	100
	Dropout	–	0.05
	Sequence length	L	512
Riemannian GD	Riemann step size	η_R	0.1
	Max Riemann iters	K	20
	Tolerance	ϵ	10^{-6}
Drift Control	Drift strength	λ_d	$[10^{-3} \to 10^{-4}]$
	Warmup fraction	ρ	0.2
	Scale schedule	$\alpha^{(t)}$	stage-wise

B.1 Detailed Experimental Setup

Datasets. The commonsense reasoning suite contains BoolQ, PIQA, SIQA, HellaSwag, WinoGrande, ARC-Easy, ARC-Challenge, and OpenBookQA, covering linguistic, physical, social, and scientific reasoning. The natural language understanding suite contains SST-2, SST-5, MRPC, CoLA, QNLI, RTE, and WNLI, spanning sentiment analysis, linguistic acceptability, paraphrase detection, and textual inference. WikiText-2 Merity et al. (2016) is used only to measure held-out language-model perplexity and is not used for downstream task training.

Models and client population. We use Qwen3-1.7B and Qwen3-8B Yang et al. (2025) for commonsense reasoning and TinyLlama Zhang et al. (2024b) for natural language understanding. Clients contain non-IID local data and may use different adapter ranks and training budgets. At every adaptation stage, participating clients are sampled from a population of 100 heterogeneous clients. All methods within the same model and protocol use identical data partitions, client samples, local-training budgets, and evaluation sets.

Task-isolated continual adaptation. Each downstream benchmark is initialized independently from the same pretrained model. Client data for that fixed task category then arrive continually across adaptation stages. This protocol isolates the effects of evolving non-IID client data, client-scale heterogeneity, and heterogeneous adapter ranks without introducing interference between task categories.

We report downstream task accuracy together with relative perplexity change (ΔPPL) on held-out WikiText-2. Accuracy measures acquisition of the target task, whereas ΔPPL measures the change in general language-model behavior induced by adaptation. Lower values indicate less perplexity growth, and negative values indicate that held-out perplexity improves after adaptation. Because this protocol contains only one task category at a time, ΔPPL is treated as an auxiliary stability indicator rather than a direct continual-learning forgetting metric.

Sequential task-category adaptation. A single global model is sequentially adapted across BoolQ, PIQA, SIQA, and HellaSwag without resetting between categories. We construct five task-order trajectories and evaluate the current global model on all categories observed so far after every adaptation stage. This produces the task-by-stage performance matrix required to measure new-task acquisition, backward transfer, and forgetting.

For Qwen3-1.7B, each client uses one local epoch with at most 256 training examples, and evaluation is capped at 128 examples per task. For Qwen3-8B, heterogeneous adapter ranks are drawn from $\{16, 8, 4\}$, with 24 training examples per client and an evaluation cap of 32 examples per task. Because the two model scales use different compute and evaluation budgets, their absolute results are interpreted independently. Within each model block, however, all methods use the same protocol and resources.

We report final average accuracy (AA), new-task plasticity, and average forgetting (AF) in the main paper. The appendix additionally reports average incremental accuracy ($\mathrm{ACC}_{\mathrm{inc}}$), backward transfer (BWT), retention, worst-case forgetting (WCF), and area under the retention curve (AURC). Formal definitions are given in Appendix B.3.1.

Baselines. For task-isolated evaluation, centralized LoRA and DoRA Liu et al. (2024) provide non-federated reference points. Among federated methods, FedIT Zhang et al. (2024a) applies FedAvg McMahan et al. (2023) to LoRA parameters, while FLoRA Wang et al. (2024) supports clients with heterogeneous adapter ranks. Zero-padding Cho et al. (2024) aligns heterogeneous adapters by padding them to a common rank, and standalone adaptation Wang et al. (2024); Guo et al. (2024) trains each client independently without server aggregation.

We further compare with FlexLoRA Bai et al. (2024), FedEx-LoRA Singhal et al. (2024), and LoRA-FAIR Bian et al. (2025). These methods represent complementary approaches to heterogeneous federated adapter aggregation, including weight synthesis, exact update reconstruction, and bias-aware initialization. All external baselines are evaluated under the same client sampling, data, and adaptation budgets as DoMA.

B.2 Complete Task-Isolated Adaptation Results

Tables 6 and 7 report the complete per-task results underlying the main-text summary. We include all centralized references and federated baselines under homogeneous and heterogeneous adapter ranks. Accuracy evaluates downstream task acquisition, while relative ΔPPL on held-out text measures preservation of general language-model behavior.

Table 6: **Performance on Commonsense Reasoning.** Task accuracy (Acc.) and relative perplexity change (ΔPPL) on eight CR benchmarks under homo/heter settings for Qwen3-1.7B and Qwen3-8B. Higher accuracy and lower ΔPPL are better. All results are reported in percentage (%).

Task Category		Commonsense Reasoning																
Settings		Qwen3-1.7B									Qwen3-8B							
Scene	Method	BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA	Avg.	BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA	Avg.
		Task Accuracy (Acc.) ↑																	
Centralized	LoRA	66.70	80.47	73.75	51.42	70.68	87.02	75.12	80.00	73.15	80.37	82.08	78.36	90.20	82.08	96.35	89.96	88.47	85.98
	DoRA	67.37	81.01	74.51	51.20	71.59	86.67	75.59	80.80	73.59	86.51	83.79	79.32	91.83	85.71	96.45	89.96	89.33	87.96
Homo	**DoMA**	**72.69**	**80.74**	**74.32**	**40.02**	**63.84**	86.07	**76.62**	**81.22**	**71.94**	**87.16**	**88.74**	**77.94**	**89.95**	**76.40**	96.49	**87.96**	86.80	**86.43**
	FedIT	69.85	76.28	59.31	36.59	51.78	85.94	72.78	79.80	66.54	86.48	87.27	76.97	87.89	72.29	96.49	86.83	86.00	85.03
	FLoRA	69.58	78.67	72.98	37.74	61.96	**86.24**	76.28	80.20	70.46	80.79	86.94	74.36	69.17	73.56	96.49	87.29	86.80	81.93
	FlexLoRA	70.03	74.70	73.20	38.45	62.41	86.18	75.94	80.45	70.17	81.31	87.14	74.87	68.44	75.30	96.49	86.42	86.80	82.10
	FedEx-LoRA	69.69	74.59	65.82	37.08	55.34	85.80	73.61	79.92	67.73	82.59	86.33	70.84	68.67	71.03	96.49	85.63	85.83	80.93
	LoRA-FAIR	70.06	77.20	71.50	37.56	60.88	85.73	75.21	79.71	69.73	81.64	87.22	73.26	69.04	73.41	96.49	86.11	85.21	81.55
	Standalone	65.80	74.90	67.60	36.66	61.56	83.67	70.48	76.30	67.12	76.49	78.26	70.83	77.34	70.54	88.46	79.52	75.47	77.11
Heter	**DoMA**	**73.73**	**81.07**	**74.10**	**40.49**	**67.72**	**86.24**	**74.57**	**81.20**	**72.39**	**87.16**	**88.63**	**77.89**	**90.24**	**82.24**	**96.49**	87.63	**87.00**	**87.16**
	Zero-padding	68.44	67.57	33.16	27.16	51.57	67.63	41.64	39.00	49.52	85.96	82.10	75.37	79.11	67.01	96.32	87.29	85.40	82.32
	FLoRA	69.72	78.73	73.54	39.78	62.93	85.62	73.49	79.20	70.38	80.43	86.83	75.03	25.06	82.01	96.07	**88.63**	42.20	72.03
	FlexLoRA	70.39	76.91	73.62	39.22	63.40	85.91	73.70	79.70	70.36	82.38	87.92	75.91	76.34	75.08	96.25	86.08	86.10	83.26
	FedEx-LoRA	70.36	77.63	68.42	37.80	58.10	85.48	72.84	78.95	68.70	86.32	87.93	76.82	88.46	74.10	96.18	86.35	85.62	85.22
	LoRA-FAIR	69.33	80.61	70.26	38.14	61.22	85.10	73.11	79.18	69.62	84.31	87.76	75.20	84.98	74.18	95.77	85.34	86.50	84.26
	Standalone	66.99	72.72	66.22	36.12	56.62	82.22	67.66	71.00	64.94	76.18	73.93	65.38	76.51	N/A	89.18	79.03	73.43	76.23
		Δ Perplexity (PPL) ↓																	
Homo	**DoMA**	**-4.32**	1.32	**12.37**	**11.32**	**34.28**	**3.50**	**15.04**	20.99	**11.81**	**-18.84**	**-17.48**	**-15.99**	**-19.89**	**-15.49**	**-11.93**	**-7.69**	**-15.07**	**-15.30**
	FedIT	6.57	5.17	40.12	13.07	55.05	7.13	17.12	24.95	21.15	-12.28	-10.94	-9.88	-10.32	2.13	-5.64	2.52	-3.51	-5.99
	FLoRA	26.25	6.98	76.98	67.33	213.92	9.66	47.84	25.44	59.30	-11.15	-13.22	-12.59	-8.31	-8.97	-9.56	-7.90	-10.46	-10.27
	FlexLoRA	3.88	**-1.23**	28.64	18.21	41.77	6.45	20.12	18.92	17.10	-10.42	-12.68	-11.31	-7.94	-7.58	-8.92	-6.84	-9.75	-9.43
	FedEx-LoRA	3.38	-0.28	31.15	20.07	45.90	6.98	21.44	19.55	18.52	-11.86	-13.05	-12.18	-9.42	-8.15	-9.36	-7.12	-10.21	-10.17
	LoRA-FAIR	-4.04	**-6.96**	25.43	15.87	38.66	5.91	18.75	**17.21**	13.85	-10.95	-12.47	-11.74	-8.68	-7.91	-8.83	-6.95	-9.62	-9.64
Heter	**DoMA**	**-18.15**	**-8.13**	**12.65**	**18.53**	**41.04**	**-3.51**	**12.67**	25.51	**10.08**	**-18.78**	**-17.39**	**-15.97**	**-19.77**	**-15.45**	-11.86	**-7.71**	**-15.05**	**-15.25**
	Zero-padding	-3.25	-3.47	15.37	24.70	51.49	-2.99	24.24	**23.82**	16.24	-2.72	-1.43	-1.35	-2.47	-1.73	-3.41	-0.57	-1.21	-1.86
	FLoRA	-5.57	-1.41	16.06	23.76	58.95	4.21	48.75	26.25	21.38	-11.36	-12.17	12.83	-8.93	-8.39	**-14.83**	-7.10	-12.13	-7.76
	FlexLoRA	-4.18	-3.37	14.72	21.90	46.38	1.85	19.55	24.60	15.18	-10.08	-11.74	-10.92	-8.36	-7.84	-10.57	-6.62	-10.15	-9.54
	FedEx-LoRA	-2.47	-2.92	15.81	22.30	48.05	2.44	20.83	25.12	16.14	-10.96	-12.36	-11.48	-9.05	-8.28	-11.24	-6.88	-10.63	-10.11
	LoRA-FAIR	-3.18	5.03	13.94	19.76	44.20	1.97	18.67	24.01	15.55	-9.84	-11.92	-10.66	-8.12	-7.63	-10.38	-6.31	-9.92	-9.35

Table 7: **Performance on Natural Language Understanding.** Task accuracy and relative perplexity change on seven NLU benchmarks for TinyLlama across settings. All results are in percentage (%).

Task Category		Natural Language Processing															
Settings		Accuracy (Acc.) ↑								Δ Perplexity (PPL) ↓							
Scene	Method	SST-2	SST-5	MRPC	CoLA	QNLI	RTE	WNLI	Avg.	SST-2	SST-5	MRPC	CoLA	QNLI	RTE	WNLI	Avg.
Homo	**DoMA**	**94.11**	**54.23**	**69.61**	**69.13**	**88.32**	**66.98**	**59.19**	**71.65**	**-8.29**	**-7.15**	**-4.82**	**-1.78**	**7.26**	**5.06**	0.93	**-1.26**
	FedIT	92.58	45.05	68.63	56.66	85.42	63.87	49.30	65.93	3.64	4.15	3.30	1.02	11.17	5.64	**0.72**	4.23
	FLoRA	92.46	48.77	68.87	65.98	84.03	62.09	50.70	67.56	4.36	4.26	4.00	15.37	11.61	5.77	0.73	6.59
	FlexLoRA	93.10	43.32	68.95	64.80	84.91	63.55	45.98	66.66	-5.65	-3.01	-2.88	2.10	9.45	4.88	-3.01	0.55
	FedEx-LoRA	92.72	42.96	68.55	63.90	84.50	62.88	43.66	65.88	-4.74	-2.83	-2.65	2.55	9.90	5.01	-2.83	0.77
	LoRA-FAIR	92.90	49.13	69.02	66.10	85.20	63.12	46.72	67.31	-6.01	-6.31	-3.10	1.88	8.72	4.56	0.11	-0.02
Heter	**DoMA**	**95.72**	**55.70**	**68.38**	**69.13**	**87.16**	**65.26**	**59.15**	**71.50**	**-9.16**	-6.17	**-20.27**	**-1.12**	5.55	**5.02**	**-0.09**	**-3.75**
	Zero-padding	85.12	39.46	67.40	68.46	83.52	61.82	49.30	64.87	-3.07	-3.40	-4.88	-0.65	**0.06**	5.14	-0.06	-0.98
	FLoRA	94.55	52.75	67.14	67.56	84.83	62.50	51.11	68.63	-4.60	-1.92	-12.40	18.92	15.34	9.39	-0.02	3.53
	FlexLoRA	94.92	47.86	67.88	67.10	85.40	63.12	44.32	67.66	-3.53	**-6.32**	-10.55	4.80	12.88	7.22	-1.28	0.46
	FedEx-LoRA	94.31	47.41	67.45	66.90	84.90	62.72	45.98	67.10	-2.15	-4.27	-9.88	5.25	13.44	7.66	-0.14	1.42
	LoRA-FAIR	94.80	48.86	67.92	67.85	85.60	63.50	43.66	67.89	-6.09	-5.91	-11.20	3.77	11.66	6.80	-2.20	-0.45

B.3 Complete Results for Plasticity-Qualified Continual Adaptation

This section complements the main continual-adaptation results with formal metric definitions, the complete metric suite, and a controlled factorial analysis of aggregation geometry and drift regulation.

B.3.1 Metrics and Complete Results

Metric definitions. Let $R_{t,j}^{(o)}$ denote the accuracy on task j after adaptation stage t under task-order trajectory o. For a trajectory containing T tasks, we define final average accuracy, plasticity, and average forgetting as

$$\mathrm{AA}^{(o)} = \frac{1}{T}\sum_{j=1}^{T} R_{T,j}^{(o)}, \mathrm{Plas}^{(o)} = \frac{1}{T}\sum_{t=1}^{T} R_{t,t}^{(o)}, \mathrm{AF}^{(o)} = \frac{1}{T-1}\sum_{j=1}^{T-1}\left(\max_{t\in\{j,\ldots,T-1\}} R_{t,j}^{(o)} - R_{T,j}^{(o)}\right). \tag{21}$$

Negative AF indicates positive backward improvement, i.e., subsequent adaptation improves earlier-task performance beyond its best pre-final value.

We additionally report average incremental accuracy and backward transfer:

$$\mathrm{ACC}_{\mathrm{inc}}^{(o)} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{t}\sum_{j=1}^{t} R_{t,j}^{(o)}, \mathrm{BWT}^{(o)} = \frac{1}{T-1}\sum_{j=1}^{T-1}\left(R_{T,j}^{(o)} - R_{j,j}^{(o)}\right). \tag{22}$$

Positive BWT indicates that learning subsequent tasks improves performance on previously encountered tasks.

Retention and worst-case forgetting are defined as

$$\mathrm{Ret}^{(o)} = \frac{1}{T}\sum_{j=1}^{T}\frac{R_{T,j}^{(o)}}{\max_{t\in\{j,\ldots,T\}} R_{t,j}^{(o)}}, \mathrm{WCF}^{(o)} = \max_{j<T}\left(\max_{t\in\{j,\ldots,T\}} R_{t,j}^{(o)} - R_{T,j}^{(o)}\right). \tag{23}$$

AURC summarizes retention over the complete adaptation trajectory; values above one indicate positive retention relative to initial task acquisition. Each metric is first computed for each task-order trajectory and then averaged over the five trajectories.

Complete results. Table 8 reports the complete metric suite. These metrics must be interpreted jointly: low forgetting can reflect either effective retention or insufficient acquisition of new tasks. AA, $\mathrm{ACC}_{\mathrm{inc}}$, and plasticity distinguish genuine stability after meaningful adaptation from this under-adaptation regime.

Table 8: **Complete plasticity-qualified continual adaptation results.** Mean ± standard deviation over five task-order trajectories. All values except AURC are percentages. Higher is better except for AF and WCF. Best means are bolded within each model block.

	Adaptation quality			Stability and retention				
Method	**AA↑**	**$\mathrm{ACC}_{\mathrm{inc}}$↑**	**Plas.↑**	**AF↓**	**BWT↑**	**Ret.↑**	**WCF↓**	**AURC↑**
Qwen3-1.7B								
FedIT	44.14 ± 0.00	57.78 ± 0.00	44.53 ± 0.00	0.52 ± 0.00	−0.52 ± 0.00	**98.82 ± 0.00**	**0.78 ± 0.00**	0.9905 ± 0.00
FLoRA	46.09 ± 0.83	58.37 ± 1.04	47.75 ± 2.07	2.34 ± 1.47	−2.21 ± 1.66	94.63 ± 2.67	8.20 ± 3.87	0.9744 ± 0.02
FedEx-LoRA	46.09 ± 0.83	58.19 ± 1.29	47.56 ± 2.35	1.95 ± 2.03	−1.95 ± 2.03	95.09 ± 3.32	7.42 ± 4.97	0.9778 ± 0.02
FlexLoRA	46.09 ± 0.83	58.41 ± 1.10	47.85 ± 2.21	2.47 ± 1.66	−2.34 ± 1.84	94.45 ± 2.93	8.59 ± 4.42	0.9726 ± 0.02
LoRA-FAIR	49.22 ± 0.55	59.16 ± 1.45	51.66 ± 3.18	4.30 ± 3.11	−3.26 ± 3.50	93.01 ± 4.07	12.50 ± 8.84	0.9406 ± 0.05
DoMA	**56.57 ± 0.97**	**61.97 ± 0.96**	**54.50 ± 1.10**	**-1.69 ± 0.18**	**1.95 ± 0.18**	92.45 ± 2.39	14.06 ± 5.52	**1.1084 ± 0.04**
Qwen3-8B								
FedIT	44.53 ± 2.21	53.04 ± 2.49	42.97 ± 4.42	0.52 ± 0.74	2.08 ± 2.95	93.80 ± 1.57	4.69 ± 2.21	1.0767 ± 0.00
FLoRA	47.27 ± 3.87	55.86 ± 2.16	43.75 ± 4.42	2.60 ± 3.68	4.69 ± 0.74	93.98 ± 4.23	7.81 ± 2.21	1.1073 ± 0.02
FedEx-LoRA	47.27 ± 3.87	55.16 ± 2.16	43.75 ± 4.42	2.60 ± 3.68	4.69 ± 0.74	93.98 ± 4.23	7.81 ± 2.21	1.0840 ± 0.09
FlexLoRA	46.88 ± 3.31	56.97 ± 2.03	43.75 ± 4.42	3.12 ± 2.95	4.17 ± 1.47	93.98 ± 4.23	7.81 ± 2.21	1.1103 ± 0.02
LoRA-FAIR	47.66 ± 3.31	57.55 ± 1.10	42.52 ± 3.33	3.65 ± 6.63	5.21 ± 0.00	92.07 ± 6.93	9.38 ± 4.42	1.0062 ± 0.07
DoMA	**56.38 ± 3.87**	**58.50 ± 1.43**	**52.13 ± 3.31**	**-1.92 ± 1.42**	**5.73 ± 0.74**	**97.80 ± 5.21**	**3.81 ± 6.63**	**1.1373 ± 0.01**

Protocol. All methods use the same sequential-task protocol and client pool within each model block. Participating clients are sampled at each stage from a population of 100 heterogeneous clients.

Results across model scales. On Qwen3-1.7B, DoMA achieves the highest AA, $\mathrm{ACC}_{\mathrm{inc}}$, plasticity, BWT, and AURC, together with the only negative mean AF. It improves final AA and plasticity over the strongest baselines by 6.35 and 2.44 percentage points, respectively. Although FedIT attains higher retention and lower WCF, these values coincide with the lowest AA and plasticity. This contrast illustrates why forgetting must be interpreted conditional on meaningful task acquisition.

The separation becomes stronger on Qwen3-8B. DoMA improves AA and plasticity over the next-best baselines by 8.72 and 8.38 percentage points while attaining negative AF (-1.92%). It also achieves the highest $\mathrm{ACC}_{\mathrm{inc}}$, BWT, retention, and AURC and the lowest WCF. DoMA therefore improves both adaptation and retention rather than obtaining stability through conservative updates.

Across both model scales, DoMA consistently occupies the high-plasticity, low-forgetting regime. This behavior is consistent with its geometric design: manifold-aware aggregation preserves the directional structure of heterogeneous client updates, while adaptive drift control regulates their displacement across task transitions.

Table 9: **Factorial analysis of aggregation geometry and drift control.** All configurations follow the same matched sequential-task protocol. All values except AURC are percentages. Higher is better except for AF and WCF; best means are bolded within each model block.

Configuration			Adaptation quality			Stability and retention				
Geometry	**Drift**	**Method**	**AA↑**	**ACC$_{inc}$↑**	**Plas.↑**	**AF↓**	**BWT↑**	**Ret.↑**	**WCF↓**	**AURC↑**
Qwen3-1.7B										
Euclidean	Off	FDoRA	55.37 ± 0.41	60.59 ± 1.30	54.69 ± 1.38	−0.52 ± 0.74	0.91 ± 1.29	89.70 ± 1.50	20.70 ± 3.87	**1.1319 ± 0.0082**
Euclidean	ESLR	FDoRA+ESLR	55.76 ± 0.97	60.59 ± 0.79	54.37 ± 0.69	−0.39 ± 0.55	0.52 ± 0.37	90.38 ± 1.39	19.14 ± 3.87	1.1055 ± 0.0392
Riemannian	Off	DoMA w/o drift	55.27 ± 0.83	60.69 ± 1.59	**55.18 ± 2.35**	0.26 ± 1.47	0.13 ± 2.03	89.64 ± 1.59	21.09 ± 4.42	1.1094 ± 0.0276
Riemannian	ESLR	**DoMA**	**56.57 ± 0.97**	**61.97 ± 0.96**	54.50 ± 1.10	**-1.69 ± 0.18**	**1.95 ± 0.18**	**92.45 ± 2.39**	**14.06 ± 5.52**	1.1084 ± 0.0424
Qwen3-8B										
Euclidean	Off	FDoRA	48.05 ± 3.87	57.19 ± 1.52	43.75 ± 4.42	−0.52 ± 2.21	3.73 ± 0.74	97.79 ± 0.98	4.69 ± 2.21	1.1073 ± 0.0201
Euclidean	ESLR	FDoRA+ESLR	48.83 ± 1.66	**58.50 ± 1.43**	46.09 ± 3.31	0.00 ± 0.00	3.65 ± 2.21	97.07 ± 2.00	6.25 ± 4.42	1.0840 ± 0.0873
Riemannian	Off	DoMA w/o drift	52.05 ± 3.87	53.32 ± 1.70	47.75 ± 4.42	−0.52 ± 2.21	3.73 ± 0.74	97.79 ± 0.98	4.69 ± 2.21	1.1103 ± 0.0160
Riemannian	ESLR	**DoMA**	**56.38 ± 3.87**	**58.50 ± 1.06**	**52.13 ± 3.31**	**-1.92 ± 1.42**	**5.73 ± 0.74**	**97.80 ± 5.21**	**3.81 ± 6.63**	**1.1373 ± 0.0057**

B.3.2 Disentangling Aggregation Geometry and Drift Control

To identify the source of this stability–plasticity improvement, we construct a matched 2×2 comparison that independently varies the aggregation geometry (Euclidean or Riemannian) and the ESLR drift schedule (disabled or enabled). The Euclidean variants are controlled counterparts of DoMA that replace manifold aggregation with Euclidean averaging while retaining the remaining adaptation protocol. All configurations within each model block use identical data, task orders, client sampling, and optimization budgets.

Complementary roles of geometry and drift regulation. Table 9 shows that the complete DoMA configuration achieves the strongest joint adaptation–retention performance at both model scales. On Qwen3-1.7B, it attains the best final and incremental accuracy, AF, BWT, retention, and WCF. On Qwen3-8B, DoMA attains the highest or tied mean for every reported metric. Single-component variants occasionally favor an individual metric, but neither achieves the same overall balance.

The controlled comparisons further separate the two effects. Under the same ESLR schedule, Riemannian aggregation improves both final accuracy and forgetting over Euclidean aggregation at both model scales, with a larger separation on Qwen3-8B. Within the Riemannian variants, enabling ESLR further improves final performance and limits average and worst-case forgetting. Geometry-aware aggregation and drift regulation therefore address complementary sources of sequential interference: the former preserves the directional structure of heterogeneous updates, whereas the latter controls their displacement across task transitions. Together, they produce the most favorable stability–plasticity trade-off.

C DoMA Framework

We present DoMA in a general multi-round form with T adaptation rounds. At initialization, the server decomposes the global weight into a magnitude-direction representation on the unit sphere. At the beginning of each round, the server broadcasts the global representation together with a *drift-control policy* $\Pi^{(t)}$ (Module 1), which acts as a system-level stability prior for local optimization. Clients then perform local adaptation with drift control (Algorithm 1) and upload PEFT components. Finally, the server aggregates magnitudes in Euclidean space and directions via an approximate Fréchet mean on the product-sphere manifold using iterative Riemannian gradient descent (Algorithm 2).

Algorithm 1 describes the client-side local adaptation (Module 2). Upon receiving the broadcast state, each client initializes its PEFT parameters and records drift anchors. Local optimization follows a system-specified drift-regularized objective governed by $\Pi^{(t)}$, where the schedule $s(\ell)$ is strong in early steps and decays later (Module 1). The client then uploads the PEFT components to the server.

Algorithm 2 describes the server-side manifold-structured aggregation (Module 3). The server aggregates magnitudes by (optionally weighted) Euclidean averaging. For directions, it first constructs each client direction as the DoRA-normalized fraction on the unit sphere, and then computes an approximate Fréchet mean via iterative Riemannian gradient descent using logarithmic and

Algorithm 1 Client Local Adaptation at Stage t

Input: client data $D_i^{(t)}$, broadcast state $\{m^{(t)}, \hat{V}^{(t)}\}$, drift policy $\Pi^{(t)}$, local step budget L
Output: client PEFT components $\{m_i^{(t)}, A_i^{(t)}, B_i^{(t)}\}$
Initialize local parameters $\theta_i^{(t)} \leftarrow \{m^{(t)}, A_i^{(t)}, B_i^{(t)}\}$
Record received parameters as anchors $\theta_{p,i}^{(0,t)}$ for $p \in \{m, A, B\}$
for $\ell = 1$ **to** L **do**
    Update $\theta_i^{(t)}$ by minimizing the drift-regularized local objective
    Update $\theta_i^{(t)}$ by minimizing Eq. equation 12 under the drift policy $\Pi^{(t)}$
end for
Upload $\{m_i^{(t)}, A_i^{(t)}, B_i^{(t)}\}$ to the server

Algorithm 2 Continual Federated Adaptation with Manifold-aware Aggregation

Input: initial global state $\{m^{(1)}, W_0\}$, stages T, clients $\{1, \ldots, N\}$, aggregation weights $\{p_i\}$
Output: global states $\{ m^{(t+1)}, \hat{V}^{(t+1)} \}_{t=1}^{T}$
for $t = 1$ **to** T **do**
    Broadcast $\{m^{(t)}, \hat{V}^{(t)}, \Pi^{(t)}\}$ to all clients
    Clients perform local adaptation in parallel using Algorithm 1
    Aggregate magnitudes by weighted averaging with Eq. equation 14
    Construct client directions and normalize them onto the unit sphere with Eq. equation 15
    Compute the aggregated direction via an approximate Fréchet mean
    Launching Riemannian gradient descent with Eq. equation 17 for global aggregated direction
    Set $\{m^{(t+1)}, \hat{V}^{(t+1)}\}$ as the global state for the next stage
end for

exponential maps. Finally, the server obtains $\{m^{(t+1)}, \hat{V}^{(t+1)}\}$ and may optionally merge them to form $W^{(t+1)}$. In implementation, $\hat{V}^{(t+1)}$ can be cached and directly used in the forward pass to reduce overhead.

D Theoretical Analysis

D.1 Properties of Fréchet Mean on the Riemann Unit Sphere

Theorem D.1 (Uniqueness of Fréchet Mean on Sphere)**.** *On the unit sphere $\mathbb{S}^{d-1}$, the Fréchet mean is unique if all points v_i lie within a geodesic ball of radius less than $\pi/2$. For points distributed more widely, multiple local minima may exist, but gradient-based methods converge to a valid barycenter.*

Theorem D.1 addresses the uniqueness properties of the Fréchet mean on the unit sphere $\mathbb{S}^{d-1}$, a fundamental result in Riemannian geometry with applications in manifold-based aggregation methods like those in DoMA. In general, the Fréchet mean provides a geometrically intrinsic notion of averaging points on curved spaces, ensuring that the aggregate respects the manifold's structure. Its significance lies in guaranteeing stable and unique solutions for aggregation in scenarios where data points (such as normalized direction vectors) are sufficiently concentrated, preventing ambiguity in federated updates and enhancing convergence in optimization algorithms. For broader distributions, while uniqueness may not hold, local minima still serve as valid barycenters, making gradient-based approximations practical for high-dimensional problems in fine-tuning.

Proof. Consider the unit sphere $\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d \mid \|v\|_2 = 1\}$, equipped with the Riemannian metric induced by the Euclidean inner product in the ambient space $\mathbb{R}^d$. The geodesic distance $d_g(u, v)$ between points $u, v \in \mathbb{S}^{d-1}$ is

$$d_g(u, v) = \arccos(\langle u, v \rangle), \tag{24}$$

where $\langle \cdot, \cdot \rangle$ is the standard dot product.

The Fréchet mean $\bar{v}$ of points $\{v_i\}_{i=1}^N \subset \mathbb{S}^{d-1}$ with weights $\{p_i\}_{i=1}^N$ (where $p_i \geq 0$ and $\sum p_i = 1$) minimizes the variance functional

$$f(v) = \sum_{i=1}^{N} p_i d_g^2(v, v_i) = \sum_{i=1}^{N} p_i \arccos^2(\langle v, v_i \rangle). \tag{25}$$

Assume the points $\{v_i\}$ lie within a geodesic ball $B_r(p) = \{q \in \mathbb{S}^{d-1} \mid d_g(p, q) < r\}$ of radius $r < \pi/2$ centered at some $p \in \mathbb{S}^{d-1}$. Such balls are strongly convex subsets of $\mathbb{S}^{d-1}$, meaning that for any two points in $B_r(p)$, the unique minimizing geodesic connecting them lies entirely within $B_r(p)$.

The functional $f(v)$ is geodesically convex on $\mathbb{S}^{d-1}$, and strictly convex within $B_{\pi/2}(p)$ due to the positive sectional curvature $K = 1$ of the sphere. Strict convexity implies that the Hessian $\mathrm{Hess}\, f(v)$ is positive definite for $v \in B_r(p)$, ensuring at most one critical point.

To show existence and uniqueness, note that $f(v)$ is continuous and $\mathbb{S}^{d-1}$ is compact, so a minimizer exists. The critical points satisfy $\nabla f(v) = 0$, where

$$\nabla f(v) = -2 \sum_{i=1}^{N} p_i \frac{\log_v(v_i)}{\sin(d_g(v, v_i))} \cdot \cos(d_g(v, v_i)), \tag{26}$$

but more directly, at equilibrium, $\sum_{i=1}^N p_i \log_v(v_i) = 0$.

Given strict convexity, there is exactly one critical point, which is the global minimum. Thus, the Fréchet mean is unique when points are confined to a ball of radius $< \pi/2$.

For wider distributions (diameter $\geq \pi/2$), the convexity fails, and multiple local minima may exist (e.g., antipodal points yield ambiguous means). However, gradient-based methods, such as Riemannian gradient descent, converge to a local minimum, which acts as a valid barycenter for practical purposes. □

D.2 Riemannian Gradient of the Fréchet Objective

We consider directional representations constrained on the unit sphere. Let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}$ denote the $(d-1)$-dimensional unit sphere endowed with the standard Riemannian metric. For any $x, y \in \mathbb{S}^{d-1}$, the geodesic distance is defined as

$$d(x, y) = \arccos(\langle x, y \rangle). \tag{27}$$

Definition D.2 (Weighted Fréchet Objective on the Sphere)**.** Given a set of directional vectors $\{\hat{v}_i^{(t)}\}_{i=1}^N \subset \mathbb{S}^{d-1}$ and aggregation weights $\{p_i\}_{i=1}^N$ satisfying $p_i \geq 0$ and $\sum_{i=1}^N p_i = 1$, the weighted Fréchet objective is defined as

$$\mathcal{F}(x) = \frac{1}{2} \sum_{i=1}^{N} p_i \, d\Big(x, \hat{v}_i^{(t)}\Big)^2, \qquad x \in \mathbb{S}^{d-1}. \tag{28}$$

Any minimizer $x^\star \in \mathbb{S}^{d-1}$ of $\mathcal{F}(x)$ is referred to as the Fréchet mean.

The Fréchet objective admits a simple and interpretable gradient form on the sphere.

Lemma D.3 (Riemannian Gradient Representation)**.** *Assume $\hat{v}_i^{(t)} \neq -x$ for all i. The Riemannian gradient of $\mathcal{F}(x)$ at $x \in \mathbb{S}^{d-1}$ is given by*

$$\nabla \mathcal{F}(x) = -\sum_{i=1}^{N} p_i \, \log_x\Big(\hat{v}_i^{(t)}\Big), \tag{29}$$

where $\log_x(\cdot)$ denotes the Riemannian logarithmic map from $\mathbb{S}^{d-1}$ to the tangent space $T_x\mathbb{S}^{d-1}$.

Proof. On the unit sphere, the Riemannian gradient of $\frac{1}{2}d(x, y)^2$ with respect to x is given by $-\log_x(y)$ whenever $y \neq -x$. By linearity of differentiation, summing over all weighted terms yields the stated expression. □

Lemma D.3 shows that the aggregation direction update in Algorithm 2 corresponds to a Riemannian gradient descent step on the Fréchet objective.

D.3 Uniqueness and Local Convexity of the Fréchet Mean

The Fréchet objective on the sphere is generally non-convex. However, under mild concentration conditions, it admits a unique minimizer and exhibits favorable local geometry.

Assumption D.4 (Directional Concentration)**.** There exists a reference direction $c \in \mathbb{S}^{d-1}$ and a radius $\delta < \frac{\pi}{2}$ such that

$$d\Big(\hat{v}_i^{(t)}, c\Big) \le \delta, \qquad \forall i \in \{1, \dots, N\}. \tag{30}$$

Theorem D.5 (Existence and Uniqueness of the Fréchet Mean)**.** *Under Assumption D.4, the Fréchet objective $\mathcal{F}(x)$ admits a unique minimizer $x^\star \in \mathbb{S}^{d-1}$. Moreover, $x^\star$ lies within the same geodesic ball of radius δ.*

Proof. On the unit sphere, any geodesic ball of radius strictly smaller than $\frac{\pi}{2}$ is geodesically convex. Within such a region, the squared geodesic distance $d(x, y)^2$ is strictly convex in x for fixed y. Since $\mathcal{F}(x)$ is a weighted sum of strictly convex functions over a geodesically convex domain, it is itself strictly convex on this domain and therefore admits a unique minimizer. □

Assumption D.4 is mild in federated fine-tuning, as directional updates originate from a shared base model and typically exhibit limited angular dispersion. Under this condition, Theorem D.5 ensures a unique and stable manifold-aware aggregation, which is essential for continual federated adaptation by avoiding ambiguity and oscillation in global update directions over time.

D.4 Convergence Analysis

Convergence of Continual DoMA Adaptation. We provide a theoretical analysis to show that DoMA admits a stable and convergent behavior under continual adaptation. In this setting, the system evolves over adaptation stages $t \in \{1, 2, \dots\}$, where at each stage clients perform a single round of local fine-tuning with newly arrived data and the server aggregates the updates using manifold-aware aggregation with adaptive drift control. We show that under mild regularity conditions, the global model sequence induced by DoMA converges as t increases, with the drift magnitude vanishing asymptotically.

Let $\theta^{(t)} \in \Theta$ denote the global model parameters after server aggregation at adaptation stage t. The sequence $\{\theta^{(t)}\}_{t \ge 1}$ is generated by iteratively performing local fine-tuning on newly arrived client data with DoMA aggregation.

We provide a more formal convergence analysis to characterize the asymptotic behavior of DoMA under continual adaptation. Specifically, we establish that the global parameter sequence induced by manifold-aware aggregation with adaptive drift control converges to a stable limit, and that the associated update magnitude vanishes over time. We further clarify the relationship between drift decay and parameter convergence.

Let $\theta^{(t)} \in \Theta$ denote the global model parameters after server aggregation at adaptation stage t. At each stage, clients perform a single local update using newly arrived data, and the server aggregates the resulting updates via the DoMA operator defined in eq. (x). This induces a stochastic iterative process of the form

$$\theta^{(t+1)} = \theta^{(t)} + \alpha_t \, \mathcal{A}_{\mathrm{DoMA}}\Big(\{\Delta_k^{(t)}\}_{k=1}^K\Big), \tag{31}$$

where $\Delta_k^{(t)}$ denotes the local update at client k, $\mathcal{A}_{\mathrm{DoMA}}(\cdot)$ is the manifold-aware aggregation operator, and $\alpha_t \in (0, 1]$ is the adaptive drift control coefficient at stage t.

We first introduce standard regularity assumptions on the underlying optimization process.

Assumption D.6 (Smoothness and Controlled Stochasticity)**.** The stage-wise objective $\mathcal{L}^{(t)}(\theta)$ is L-smooth for all t, i.e.,

$$\|\nabla \mathcal{L}^{(t)}(\theta_1) - \nabla \mathcal{L}^{(t)}(\theta_2)\| \le L\|\theta_1 - \theta_2\|. \tag{32}$$

Moreover, the stochastic gradient induced by local updates satisfies

$$\mathbb{E}\Big[\Delta_k^{(t)} \mid \theta^{(t)}\Big] = -\eta_t \nabla \mathcal{L}^{(t)}(\theta^{(t)}), \tag{33}$$

with bounded second moment

$$\mathbb{E}\Big[\|\Delta_k^{(t)}\|^2 \mid \theta^{(t)}\Big] \le G^2, \tag{34}$$

for some constant $G > 0$.

DoMA's manifold-aware aggregation plays a crucial role in stabilizing the update trajectory. By decomposing local updates into magnitude and direction components and performing aggregation on the unit Riemann sphere, DoMA avoids angular distortion and enforces geometric consistency across heterogeneous clients.

Lemma D.7 (Controlled Increment under Manifold-aware Aggregation)**.** *Under Assumption D.6, there exists a constant $C > 0$ such that the global update satisfies*

$$\mathbb{E}\Big[\|\theta^{(t+1)} - \theta^{(t)}\| \;\Big|\; \theta^{(t)}\Big] \le C\,\alpha_t. \tag{35}$$

Lemma D.7 shows that the magnitude of each global update is explicitly regulated by the drift control coefficient α_t. This property does not hold for Euclidean aggregation, where heterogeneous directions can amplify update norms even when individual local steps are small.

We next formalize a summability condition that is essential for convergence in continual adaptation.

Assumption D.8 (Summable Drift Control Schedule)**.** The adaptive drift control schedule satisfies

$$\sum_{t=1}^{\infty} \alpha_t < \infty. \tag{36}$$

This condition naturally holds for the ESLR strategy of DoMA, where strong early constraints are gradually relaxed over time.

Lemma D.9 (Finite Total Variation of the Parameter Trajectory)**.** *Under Assumptions D.6 and D.8, the global parameter sequence has finite total variation:*

$$\sum_{t=1}^{\infty} \|\theta^{(t+1)} - \theta^{(t)}\| < \infty. \tag{37}$$

Lemma D.9 implies that the cumulative displacement of the global model is bounded, even though adaptation proceeds indefinitely over time. This property is critical in continual settings, where no explicit convergence is enforced at individual stages. We then derive the main convergence.

Theorem D.10 (Convergence of Continual DoMA Adaptation)**.** *Under Assumptions D.6 and D.8, the global model sequence $\{\theta^{(t)}\}_{t\ge1}$ generated by eq. equation 31 converges almost surely to a limiting point $\theta^\star \in \Theta$, i.e.,*

$$\lim_{t\to\infty} \theta^{(t)} = \theta^\star. \tag{38}$$

Theorem D.10 establishes that DoMA admits a well-defined asymptotic limit under continual adaptation, despite the absence of multi-round local optimization or explicit stationarity at each stage. We finally clarify the relationship between convergence and drift decay.

Corollary D.11 (Vanishing Drift Magnitude)**.** *Let $\delta^{(t)} = \|\theta^{(t+1)} - \theta^{(t)}\|$ denote the mean drift magnitude at stage t. Under the conditions of Theorem D.10, we have*

$$\lim_{t\to\infty} \delta^{(t)} = 0. \tag{39}$$

Corollary D.11 shows that drift decay is a necessary consequence of convergence, and thus serves as a measurable indicator of long-term stability. This theoretical result aligns with the empirical observations in Figure 6, where DoMA with ESLR exhibits consistently decreasing drift magnitude over adaptation stages, while weaker or absent drift control leads to sustained or growing instability.

Convergence of Riemannian Gradient Descent. We analyze the convergence behavior of the iterative procedure used to compute the Fréchet mean in Algorithm 2. Recall that the update rule takes the form:

$$x^{(k+1)} = \exp_{x^{(k)}}\Big(-\eta_R\,\nabla\mathcal{F}\Big(x^{(k)}\Big)\Big), \tag{40}$$

where $\mathcal{F}$ denotes the Fréchet objective defined in Definition 2.1 and $\eta_R > 0$ is the Riemannian step size.

We first establish a descent property of the Fréchet objective along the Riemannian gradient direction.

Lemma D.12 (Descent Property)**.** *Under Assumption D.4, there exists a constant $L > 0$ such that the Riemannian gradient of $\mathcal{F}$ is L-Lipschitz continuous within the geodesic ball defined therein. For any step size $\eta_R \in (0, 1/L]$, the iterates satisfy*

$$\mathcal{F}\Big(x^{(k+1)}\Big) \leq \mathcal{F}\Big(x^{(k)}\Big) - \frac{\eta_R}{2}\left\|\nabla\mathcal{F}\Big(x^{(k)}\Big)\right\|^2. \tag{41}$$

Proof. Within a geodesically convex neighborhood of radius $\delta < \frac{\pi}{2}$, the Fréchet objective $\mathcal{F}$ is smooth and its Riemannian gradient is Lipschitz continuous. The standard smoothness inequality on Riemannian manifolds implies that for any $u \in T_{x^{(k)}}\mathbb{S}^{d-1}$,

$$\mathcal{F}(\exp_{x^{(k)}}(u)) \leq \mathcal{F}\Big(x^{(k)}\Big) + \Big\langle \nabla\mathcal{F}\Big(x^{(k)}\Big), u\Big\rangle + \frac{L}{2}\|u\|^2. \tag{42}$$

Substituting $u = -\eta_R\nabla\mathcal{F}(x^{(k)})$ and using $\eta_R \leq 1/L$ yields the desired result. □

Lemma D.12 guarantees that the Fréchet objective decreases monotonically along the iteration sequence. We now establish convergence to a stationary point.

Theorem D.13 (Convergence to the Fréchet Mean)**.** *Suppose Assumption D.4 holds and the initialization $x^{(0)}$ lies within the corresponding geodesic ball. For any fixed step size $\eta_R \in (0, 1/L]$, the Riemannian gradient descent sequence*

$$x^{(k+1)} = \exp_{x^{(k)}}\Big(-\eta_R\,\nabla\mathcal{F}\Big(x^{(k)}\Big)\Big) \tag{43}$$

generates a non-increasing and convergent sequence of objective values $\mathcal{F}(x^{(k)})$, satisfies

$$\lim_{k\to\infty}\left\|\nabla\mathcal{F}\Big(x^{(k)}\Big)\right\| = 0, \tag{44}$$

and converges to the unique Fréchet mean $x^\star$.

Proof. By Lemma D.12, summing the descent inequality over k yields

$$\sum_{k=0}^{\infty}\left\|\nabla\mathcal{F}\Big(x^{(k)}\Big)\right\|^2 \leq \frac{2}{\eta_R}\left(\mathcal{F}\Big(x^{(0)}\Big) - \inf_x \mathcal{F}(x)\right) < \infty. \tag{45}$$

Hence, $\|\nabla\mathcal{F}(x^{(k)})\| \to 0$. Under Assumption D.4, Theorem D.5 guarantees that $\mathcal{F}$ admits a unique minimizer $x^\star$. Therefore, any accumulation point of $\{x^{(k)}\}$ must coincide with $x^\star$, implying convergence of the entire sequence. □

Remark D.14. The above analysis establishes local convergence of the Riemannian gradient descent procedure used in Algorithm 2. In practice, the directional updates are tightly concentrated due to their shared initialization from a common base model, making the stated assumptions well satisfied.

E Mechanistic Stability and Communication Analysis

E.1 Setup and Regularity Conditions

At stage t, participating clients are initialized from the same global state $\theta_i^{(t,0)}$, whose directional component is $\hat{V}^{(t)}$. After local adaptation, client i returns $\theta_i^{(t,L)}$ and induces $\hat{V}_i^{(t)} \in \mathcal{M}$. Define

$$r_t = \left(\sum_{i\in\mathcal{S}_t} p_i \left\|\theta_i^{(t,L)} - \theta_i^{(t,0)}\right\|_2^2\right)^{1/2} \tag{46}$$

as the weighted local displacement. The directional distributions before and after local adaptation are

$$\mu_t^{(0)} = \delta_{\hat{V}^{(t)}}, \qquad \mu_t = \sum_{i\in\mathcal{S}_t} p_i\delta_{\hat{V}_i^{(t)}}. \tag{47}$$

Hence $\mathrm{bar}(\mu_t^{(0)}) = \hat{V}^{(t)}$ and $\mathrm{bar}(\mu_t) = \hat{V}^{(t+1)}$.

Assumption E.1 (Local regularity)**.** At each stage t:

1. The anchor and client directions lie in a common geodesically convex ball $\mathcal{B}(c_t, \varepsilon)$, $\varepsilon < \pi/2$, where the Fréchet objective is μ_ε-strongly geodesically convex and its logarithmic map is L_ε-Lipschitz. We define $C_\varepsilon = L_\varepsilon/\mu_\varepsilon$.
2. The map Φ_{dir} from PEFT parameters to normalized directions is locally L_{dir}-Lipschitz:
$$d_{\mathcal{M}}\big(\Phi_{\mathrm{dir}}(\theta), \Phi_{\mathrm{dir}}(\theta')\big) \le L_{\mathrm{dir}}\|\theta - \theta'\|_2. \tag{48}$$
3. At the output of local training,
$$\Big\|\nabla\mathcal{L}_{\mathrm{LM},i}\big(\theta_i^{(t,L)}\big) + 2\lambda_t\big(\theta_i^{(t,L)} - \theta_i^{(t,0)}\big)\Big\|_2 \le \xi_{i,t}, \tag{49}$$
where $\lambda_t = \lambda_d(L,t) > 0$ and $\|\nabla\mathcal{L}_{\mathrm{LM},i}\|_2 \le G_{i,t}$ in the local region.

Define the effective stage-wise residual as
$$\Gamma_t = \left(\sum_{i\in\mathcal{S}_t} p_i(G_{i,t} + \xi_{i,t})^2\right)^{1/2}. \tag{50}$$

E.2 Local and Geometric Stability

Lemma E.2 (Drift-controlled local displacement)**.** *Under Assumption E.1,*
$$r_t \le \frac{\Gamma_t}{2\lambda_t}. \tag{51}$$

Proof. Equation equation 49 and the triangle inequality give
$$2\lambda_t\left\|\theta_i^{(t,L)} - \theta_i^{(t,0)}\right\|_2 \le G_{i,t} + \xi_{i,t}. \tag{52}$$
Squaring, weighting by p_i, summing over clients, and taking the square root proves the result. □

Lemma E.3 (Intrinsic barycenter sensitivity)**.** *Under Assumption E.1,*
$$d_{\mathcal{M}}\left(\mathrm{bar}(\mu_t), \mathrm{bar}(\mu_t^{(0)})\right) \le C_\varepsilon L_{\mathrm{dir}} r_t. \tag{53}$$

Proof. Pairing every adapted direction with its common stage-wise anchor and applying the Lipschitz direction map yields
$$W_2(\mu_t, \mu_t^{(0)}) \le L_{\mathrm{dir}} r_t, \tag{54}$$
where W_2 is the 2-Wasserstein distance induced by $d_{\mathcal{M}}$. Strong geodesic convexity and Lipschitz continuity of the logarithmic map imply
$$d_{\mathcal{M}}\left(\mathrm{bar}(\mu_t), \mathrm{bar}(\mu_t^{(0)})\right) \le C_\varepsilon W_2(\mu_t, \mu_t^{(0)}). \tag{55}$$
Combining the inequalities proves the lemma. □

E.3 Proof and Cumulative Trajectory Bound

Proof of Proposition 5.1. Because $\mathrm{bar}(\mu_t) = \hat{V}^{(t+1)}$ and $\mathrm{bar}(\mu_t^{(0)}) = \hat{V}^{(t)}$, Lemma E.3 gives $\mathcal{D}_t \le C_\varepsilon L_{\mathrm{dir}} r_t$. Applying Lemma E.2 yields
$$\mathcal{D}_t \le \frac{C_\varepsilon L_{\mathrm{dir}}\Gamma_t}{2\lambda_t}, \tag{56}$$
which proves Equation equation 20. □

Corollary E.4 (Controlled continual trajectory)**.** *Suppose the reconstruction map $\mathcal{R}(m, \hat{V}) = m \odot \hat{V}$ is locally Lipschitz:*
$$\left\|\mathcal{R}(m,\hat{V}) - \mathcal{R}(m',\hat{V}')\right\|_F \le L_m\|m - m'\|_2 + L_V d_{\mathcal{M}}(\hat{V}, \hat{V}'). \tag{57}$$
Then, with $\Delta_m^{(t)} = \|m^{(t+1)} - m^{(t)}\|_2$,
$$\|W^{(T)} - W^{(0)}\|_F \le \sum_{t=0}^{T-1}\left(L_m\Delta_m^{(t)} + \frac{L_V C_\varepsilon L_{\mathrm{dir}}\Gamma_t}{2\lambda_t}\right). \tag{58}$$

Table 10: **Parameter and communication efficiency.** Param. denotes the trainable-parameter ratio relative to the backbone; Comm. denotes aggregate adapter-level communication.

Setting	Method	GPT2-large		Qwen3-8B	
		Param. (%)	Comm. (MB)	Param. (%)	Comm. (MB)
Centralized	Full FT	100.00	0.00	100.00	0.00
	LoRA	1.52	0.00	0.56	0.00
	DoRA	1.57	0.00	0.58	0.00
Federated	FedIT	15.24	450.00	5.64	1720.00
	FLoRA	15.24	450.00	5.64	1720.00
	FDoRA	15.66	462.30	5.79	1767.50
	DoMA	**15.66**	**462.30**	**5.79**	**1767.50**

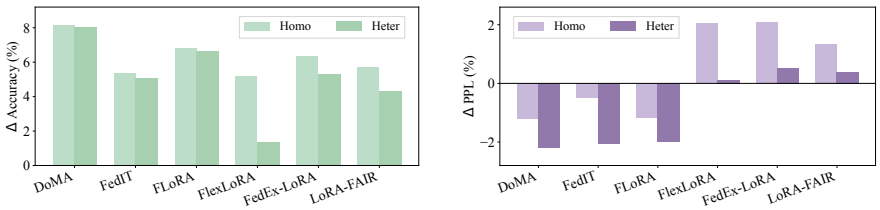


Figure 9: **Vision-Language Instruction Tuning (Classification).** Performance under Homo and Heter rank settings. **Left:** Δ Accuracy (%), where higher is better. **Right:** Δ PPL (%), where lower (more negative) indicates better retention of pretrained knowledge. DoMA consistently achieves the highest accuracy gains while maintaining stable or reduced perplexity, demonstrating robustness under multimodal and heterogeneous settings.

Proof. Apply Equation equation 57 at each stage, substitute Proposition 5.1, and sum using the triangle inequality. □

The result separates the two sources of stability. On-manifold aggregation controls the sensitivity of the global consensus through C_ε, whereas adaptive drift regularization controls local movement through λ_t. The scheduled decrease of λ_t therefore expands the admissible trajectory progressively, preserving early stability while restoring capacity for later task acquisition.

E.4 Parameter and Communication Efficiency

The Riemannian barycenter iterations are executed entirely on the server. Clients optimize and transmit the same magnitude and low-rank directional components as FDoRA, so DoMA adds neither client-trainable parameters nor adapter-level payload dimensions. Table 10 reports costs under the evaluation protocol, where participating clients are sampled from a population of 100 heterogeneous clients. Centralized entries report parameters per model; federated entries aggregate client payloads under the evaluated sampling configuration.

F Evaluation on Vision-Language Instruction Tuning

To evaluate the generality of DoMA across task scenarios, we further consider vision-language instruction tuning tasks, following a classification setting based on a pretrained vision-language model of Qwen2-VL-2B-Instruct. In this setting, each sample consists of an image paired with a textual instruction, and the model is required to generate the corresponding class label, introducing multimodal alignment and increased task complexity (as shown in Figure 10). We follow the heterogeneous FFT in the main text, where each client performs local adaptation with its own visual instruction data, and evaluate both homogeneous and heterogeneous settings. We report the change in task accuracy (ΔAccuracy) relative to the pretrained model, together with the relative change in perplexity (ΔPPL) on held-out Wikitext-2 to assess model stability.

As shown in Figure 9, DoMA consistently achieves the highest accuracy gains across both settings, outperforming FLoRA by around 20% relatively and remaining superior to recent benchmarks of FlexLoRA, FedEx-LoRA, and LoRA-FAIR. Under heterogeneous ranks, DoMA maintains stable

Example: Multimodal Instruction under Data Heterogeneity

Input Prompt

Identify the disease type shown in the image.
Return only the class ID.
Options: [0: Angular Leaf Spot, 1: Bean Rust, 2: Healthy]
Answer:

Input Image



(Image from the Beans dataset showing leaf symptoms)

Model Output

0 (Angular Leaf Spot)

Ground Truth

0 (Angular Leaf Spot)

Figure 10: **Example of Vision-Language Instruction of the Beans Task.**

performance with only minor degradation, while several baselines exhibit noticeable accuracy drops, particularly under severe rank mismatch. In addition, DoMA consistently yields a reduction in perplexity (PPL), indicating an improved, more stable retention of pretrained knowledge, whereas multiple baselines incur positive PPL change, suggesting increased risk of knowledge forgetting. These results demonstrate that DoMA generalizes effectively to multimodal instruction-following tasks and remains robust under increased heterogeneity in both data modality and client capacity.

G Scalability Analysis

DoMA is designed to operate under large-scale federated settings with partial participation. In our experiments, we choose a client population of $N = 100$ with uniform sampling without replacement at each adaptation stage, which follows standard FFT practice and allows the system to scale to large client populations without increasing per-round communication cost.

Meanwhile, according to the overhead analysis in Section 5, the communication overhead of DoMA scales linearly with the number of participating clients K, rather than the total client population N. Since DoMA operates on parameter-efficient updates and does not introduce additional parameters beyond existing PEFT-based methods, its communication cost remains comparable to methods such as FDoRA. From a computational perspective in Appendix Section D.4, the aggregation step involves computing a Riemannian Fréchet mean over the K client updates, which scales linearly with K and converges efficiently in practice within a small number of iterations. While our empirical evaluation focuses on moderate participation sizes, the framework naturally extends to larger-scale deployments with hundreds or thousands of clients through standard client sampling strategies.

H LIMITATIONS

While DoMA demonstrates strong stability and performance under continual FFT, our current design assumes that directional updates at each stage admit a reasonably concentrated geometric structure, under which Riemannian aggregation is well behaved. In scenarios with extremely sparse participation or highly adversarial client updates, additional robustness mechanisms or adaptive manifold constraints may be beneficial. Further exploring adaptive strategies for selecting aggregation mechanisms across different tasks and client populations is a promising direction for future work.

I BROADER IMPACTS

DoMA aims to improve the stability and efficiency of continual federated fine-tuning for large language models under heterogeneity. Potential positive impacts include enabling more reliable decentralized adaptation without centralized data collection, thereby improving client-wise flexibility, data preservation, and reducing computing cost in practical deployments.

At the same time, the proposed framework may also introduce potential risks if deployed irresponsibly. More stable federated adaptation could be misused to personalize harmful or biased models across decentralized clients continually. We therefore encourage future work on integrating safety monitoring, robustness mechanisms, and fairness-aware adaptation into continual federated fine-tuning systems.

J AI USE STATEMENT.

This paper was edited for grammar and language clarity using GPT-5.6-Sol. We reviewed and verified all AI-assisted text, languages, and take full responsibility for the final content of this work.