A. Scalability discussion

We compare client and server computation complexity (Client/Server Comp.) and communication costs for upload and download (UL/DL Comm. cost).

Method	Client Comp.	Server Comp.	Comm. cost
FlexLoRA	$O(rd^2)$	$O(Kdr^2 + Kd^2 + d^3)$	UL:$2rd$ DL:$2rd$
FLoRA	$O(rd^2)$	$O(Krd^2)$	UL:$2rd$ DL:d^2
Te-LoRA (Ours)	$O(rd)$	$O(K^2rd + Kr^2d + K^2r^2d) + O(Kr^2)$	UL:rd DL:rd

Table A. Computation complexity

Since (rank) $r \ll d$, our Te-LoRA greatly reduces client computation and communication cost. With $K = 10$ (clients) and $d = 4096$ (LoRA matrix $\in \mathbb{R}^{d\times r}$), it is also efficient on the server side for small (< 100) and medium ($100 - 1000$) scales, but loses advantage when $K > d$ in large-scale settings.

B. More baselines

For the heterogeneous LoRA scenario, most existing methods have been thoroughly compared in the original paper, as this is a key issue addressed in our work. To enrich results, we include additional baselines (Table B). Te-LoRA outperforms FFA-LoRA, which trains only the B matrix, and LoRA-A^2, which uses score-based rank selection with alternating freezing, in both homogeneous and heterogeneous settings.

Method		MMLU		MT-Bench
Heter / Homo	Wizard	Dolly	Alpaca-GPT4	Wizard
FFA-LoRA	21.11	25.83	43.62	3.13
LoRA-A^2	23.52 / 22.92	27.93 / 27.91	45.50 / 45.20	3.26 / 3.21
Te-LoRA (Ours)	**23.71** / **23.35**	**28.37** / **28.44**	**46.16** / **45.86**	**3.33** / **3.31**

Table B. More baselines

C. Convergence analysis

Theoretical Assumptions

Lipschitz continuity: Assume that the model's loss function is L-Lipschitz continuous with respect to the parameters θ and is bounded, such that the change in loss due to small perturbations of the parameters is controlled. Here, L is the Lipschitz constant, and $|\theta| \le R$, where R is the radius of the parameter space.

Bounded alignment and tensor errors: Assume that the alignment error ψ (from PAA) and tensor error τ (from T2M) are bounded within a constant range.

Convergence Theorem

Under the aforementioned assumptions, let the sample size per client be N, the number of clients be K, and the total dimension of the LoRA parameters be $\mathcal{P}$. After applying PAA+T2M aggregation, the generalization error (or expected risk difference) of the model satisfies the following:

$$\mathcal{E}(\hat{\theta}) = O\Big(L(\psi+\tau) + \sqrt{\frac{\mathcal{P}\ln\frac{R}{(\psi+\tau)}}{|K|\;N}}\Big) \quad (1)$$

Thus, with high probability, the generalization error comprises two components: the approximation error $O(\psi+\tau)$ induced by alignment/tensor errors, and the statistical error term $O\big(\sqrt{\frac{\mathcal{P}\ln(\frac{R}{\psi+\tau})}{|K|N}}\big)$. This result preserves the dependency structure of the generalization error concerning the sample size, number of clients, parameter dimension, and errors.

Key Points of Deduction

Local perturbation error: From the Lipschitz property, we know that if the parameter vectors differ by $\Delta\theta$, then the change in loss is at most $O(L|\Delta\theta|)$. Therefore, when the alignment error and tensor error are combined into $|\Delta\theta| = O(\psi+\tau)$, the resulting model error is $O(L(\psi+\tau))$.

Coverage and statistical error: Assume that the parameter space can be regarded as a k-dimensional sphere with radius R and ε-net coverage number $|N| = O\big(\frac{R}{\varepsilon}\big)^{\mathcal{P}}$. Combining Hoeffding's concentration inequality, parallel estimation for all θ yields a generalisation error term of $O\big(\sqrt{\frac{\mathcal{P}\ln(\frac{R}{\varepsilon})}{|K|N}}\big)$.

Sample Complexity

Let the generalization error target be ϵ (ignoring the alignment error term), which must satisfy $\sqrt{\frac{\mathcal{P}\ln(\frac{R}{\psi+\tau})}{|K|N}} \approx O(\epsilon)$. The sample size required for a single client is $N = O\Big(\frac{\mathcal{P}}{|K|\epsilon^2}\ln\frac{R}{\psi+\tau}\Big)$.