Supplementary Material

A Details of Non-iid Settings.

In quantity-based label imbalance, we randomly assign α different label IDs to each client at each stage. Then, for each labeled sample, we randomly and equally distribute it among the clients associated with that label. In distribution-based label imbalance, each client receives a portion of the samples for each label based on the Dirichlet distribution. Formally, we sample P_k from $Dir_N(\beta)$ and allocate a proportion $P_{k,j}$ of class k instances to the client j. Fig. 1 shows these two partitioning strategies.

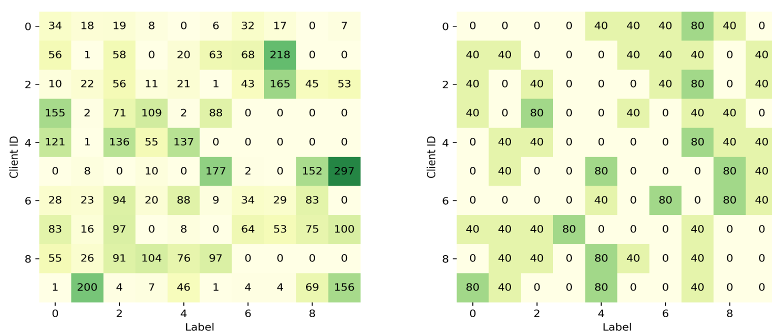

Fig. 1: An example of distribution-based label imbalance partition and quantity-based label imbalance partition on CIFAR-100 (10 classes) with $\beta = 0.5$ (left) and $\alpha = 6$ (right).

B Additional Experiments.

B.1 Impact of LoRA embedded in different blocks

In our method, we embed the LoRA module in the first block of the model. Here we test the results of the LoRA module embedding in each ViT block and compute the relative accuracy (e.g., $\Delta A_N^i = A_N^i - A_N^0|_{i=1,\dots,12}$) of embedding in each block versus embedding in the first block. As can be seen in Fig. 2, embedding LoRA in the first layer consistently outperforms embedding it in any other layers. Therefore, in our method, we fix LoRA to be embedded specifically in the first block.

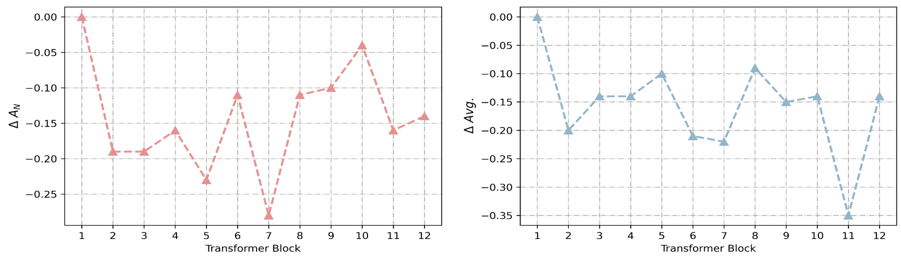

Fig. 2: Results of LoRA embedded in different blocks. We visualize the relative accuracy using embedding into the first block as a baseline.

B.2 Similarity analysis of LoRA parameters

In order to better demonstrate the role of orthogonal regularization, we compute the average cosine similarity of LoRA parameters between different stages, and the results are shown in Fig. 3. It can be seen that under the effect of orthogonal regularization, the cosine similarity between LoRAs at different stages is relatively low, indicating that the parameter space is closer to orthogonality, and therefore mitigates catastrophic forgetting.

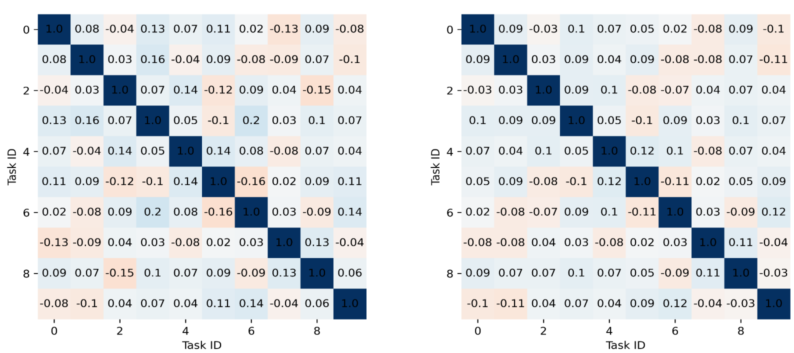

Fig. 3: LoRA cosine similarity visualization. Left: without l_{ort}; Right: with L_{ort}.

B.3 Large-scale experiments

We also test the performance of the model on large-scale datasets, specifically, we randomly select 200 classes from Imagenet-1k as a new dataset and use self-supervised pre-trained weights. As can be seen in Tab. 1, in quantity-based label imbalance, our model still maintains good performance on large-scale datasets. In addition, we also test the performance of the model on longer incremental

Table 1: Experiments on large scale dataset (Imagenet-200).

	10 Tasks			**20 Tasks**		
Methods	A_N	Avg.	F_N	A_N	Avg.	F_N
L2P+FL	33.5	53.2	12.5	15.6	39.8	16.2
Ours	**80.1**	**83.8**	**4.0**	**79.5**	**83.4**	**4.6**

phases (20 tasks), and our method effectively mitigates catastrophic forgetting in the long-phase incremental task compared to L2P+FL.