A. Experimental Results

A.1. Detailed Results

We provide results of all tasks in Table 3, which echoes observations in Sec 4.2.

Table 3. Comparisons between our proposed method and several alternative methods mentioned in Section 3.2. As highlighted in block font, our method successfully tackles the challenge of "task heterogeneity" and benefits performance by unifying more data.

Dataset	Method	Performance				
		Seg.(mIoU)	**Sal.**(mIoU)	**Edge**(odsF)	**Norm.**(rmse)	**H.Parts**(mIoU)
Pascal-Context (5)	Single-Task FL	0.2449	0.5619	0.5032	**0.2023**	0.3244
	Plain Many-Task FL	0.2731	0.5638	0.5104	0.2284	0.3312
	FedProx	0.2745	0.5661	0.5432	0.2112	0.3623
	Ours (DG w/ sc agg.)	**0.2762**	**0.5774**	**0.5812**	0.2041	**0.3747**
Pascal-Context & NYUD (9)	Single-Task FL	0.2449	0.5619	0.4995	0.2023	0.3244
	Plain Many-Task FL	0.2707	0.5714	0.5215	0.2354	0.3109
	FedProx	0.2611	0.5745	0.5327	0.2132	0.3634
	Ours (DG w/ sc agg.)	**0.2938**	**0.5814**	**0.5898**	**0.2014**	**0.3782**

A.2. Grouping Fairness

We explore the fairness of grouping, which is to check whether some clients are never chosen by others. The grouping is unfair if a client is barely aggregated by others. In that case, its global model will become different with others which probably lead to different training behavior in the next training steps, thus letting the client less likely to be chosen in the next round, too. Fortunately, as shown in Figure 7, we can observe fair grouping results and all clients are actively included, eliminating our concerns. Specifically, for each client, we plot the times they are chosen in each communication round.

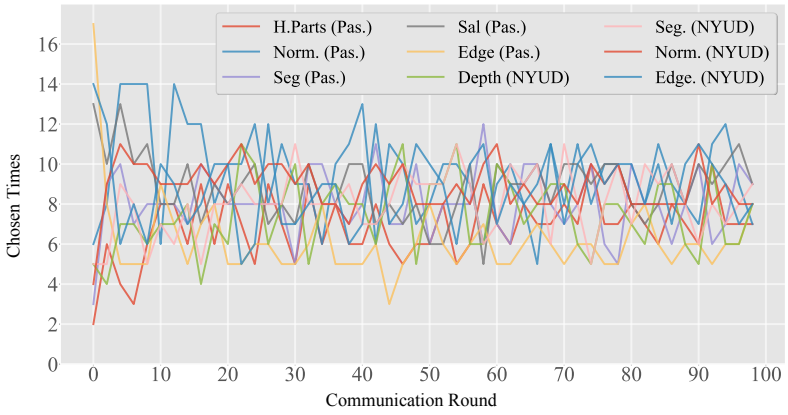

Figure 7. For every client, we plot the times it is chosen for aggregation in each communication round. If client i includes client j in its aggregation step, client j is *chosen* by client i. Note that we only provide one client's result for each task.