Mind the Gap: Confidence Discrepancy Can Guide Federated Semi-Supervised Learning Across Pseudo-Mismatch

Supplementary Material

A. Pseudo-Code of SAGE

The pseudo-code of SAGE is shown in Algorithm 1.

Algorithm 1: **S**emi-supervised **A**ggregation for **G**lobally-Enhanced **E**nsemble (SAGE)

Input: Set of clients $\mathcal{C}$; number of online clients in each round M; number of communication rounds T; number of local training epochs E; weak augmentation $\alpha(\cdot)$; strong augmentation $\mathcal{A}(\cdot)$; confidence threshold τ; learning rate γ; unsupervised loss weight μ_u; dynamic correction coefficient $\lambda(\cdot)$; sensitivity hyper-parameter κ

1 **ServerExecutes:**
2 Randomly initialize global model parameters θ_g;
3 **for** $t = 0$ **to** $T - 1$ **do**
4 Randomly select online clients $\mathcal{C}_M \subseteq \mathcal{C}$;
5 **foreach** *client* $C_m \in \mathcal{C}_M$ ***in parallel*** **do**
6 $\theta_{l,m} \leftarrow$ **ClientUpdate**(θ_g)
7 **end**
8 $|D| = \sum_{C_m \in \mathcal{C}_\mathcal{M}} (|\mathcal{D}_m^s| + |\mathcal{D}_m^u|)$;
9 $\theta_g \leftarrow \frac{1}{|D|} \cdot \sum_{C_m \in \mathcal{C}_\mathcal{M}} ((|\mathcal{D}_m^s| + |\mathcal{D}_m^u|) \cdot \theta_{l,m})$;
10 **end**
11 **return** θ_g^T
12 **ClientUpdate**(θ_g):
13 $\theta_l \leftarrow \theta_g$;
14 **for** $e = 0$ **to** $E - 1$ **do**
15 **foreach** $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}^s, \mathbf{u} \in \mathcal{D}^u$ **do**
16 $\mathcal{L}_s \leftarrow \mathcal{L}_{CE}(p_l(y|\mathbf{x}, \mathbf{y}))$;
17 $p_l \leftarrow f_l(\alpha(\mathbf{u}))$;
18 $p_g \leftarrow f_g(\alpha(\mathbf{u}))$;
19 Calculate $\hat{y}$ by CPG in Eq. (1);
20 **if** $\max(p_l) \geq \tau$ **then**
21 $\Delta C = |\max(p_l) - \max(p_g)|$;
22 $\lambda \leftarrow \exp(-\kappa \cdot \Delta C)$;
23 $\delta_l \leftarrow$ one-hot$(\arg\max(p_l))$;
24 $\delta_g \leftarrow$ one-hot$(\arg\max(p_g))$;
25 Calculate $\hat{y}$ by CDSC in Eq. (6);
26 **end**
27 $L_u \leftarrow \mathrm{KL}(p_l(\mathcal{A}(\mathbf{u})) \parallel \hat{y}(\mathbf{u}))$;
28 $\theta_l \leftarrow \theta_l - \gamma \nabla_\theta (L_s + \mu_u \cdot L_u)$;
29 **end**
30 **end**
31 **return** θ_l, $\mathcal{D}^s$, $\mathcal{D}^u$

In the local training process of SAGE, standard supervised training is initially performed on labeled data (line 16) to compute L_s. Next, CPG assigns initial pseudo-labels $\hat{y}$ using Eq. (1) (lines 16 to 19), thereby enhancing the utilization of unlabeled data. Subsequently, the confidence discrepancy ΔC between the local and global models is calculated, and the pseudo-labels are dynamically refined by computing the correction coefficient λ (lines 20 to 25) using CDSC. Finally, the KL divergence between the corrected pseudo-labels and the strongly augmented predictions of the local model is calculated as the unsupervised loss L_u. Upon completing local training, clients upload the updated local models and dataset sizes to the server for standard federated aggregation (lines 4 to 9).

B. Additional Analysis of Preliminary Study

In Section 4.1, we identified an intriguing phenomenon: as data heterogeneity increases, the confidence discrepancy between local and global models progressively grows. The predictions of the local model become more aggressive, whereas those of the global model grow increasingly conservative, as described in Observation 1 and 2. In this section, we perform a more comprehensive observation and analysis of this phenomenon. First, we provide additional observations in Appendix B.1. Next, in Appendix B.2, we derive the underlying causes of this phenomenon and present a analytical process centered on Remark 1 and 2. Finally, in Appendix B.3, we design experiments to validate our analytical conclusions.

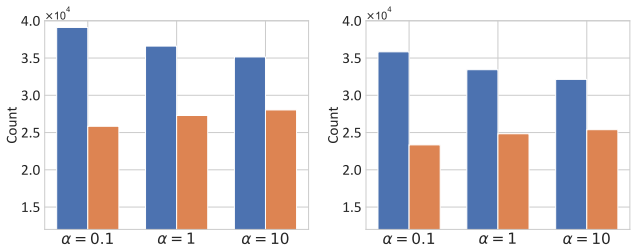
B.1. Additional Exploratory Experiments

To more comprehensively illustrate Observation 1 and 2, we follow the experimental setup of Fig. 2(a) and adjust the threshold values for displaying confidence distributions. As shown in Fig. 9, we observe similar patterns as in Fig. 2(a) of the main text: as data heterogeneity increases, the confidence of the local model tends to fall into high-confidence regions, while the global model shows the opposite trend.

Additionally, to expand on the comparison of pseudo-label counts between local and global models in Fig. 2(b), we conducted further experiments across different heterogeneity settings. As shown in Fig. 11, at varying levels of heterogeneity, the local model consistently maintains a high utilization rate of unlabeled data in the early training stages.

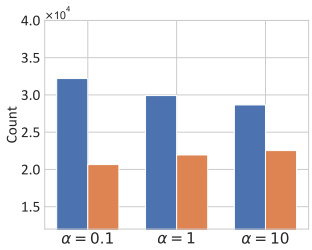
B.2. Analysis of Local-Global Discrepancies

In Section 4.1, we observed that as heterogeneity intensifies, the pseudo-labeling tendencies of the local and global

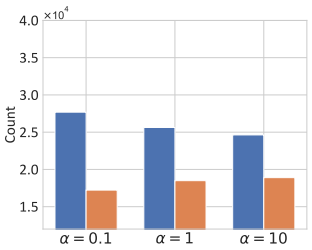


(a) Confidence > 0.95.

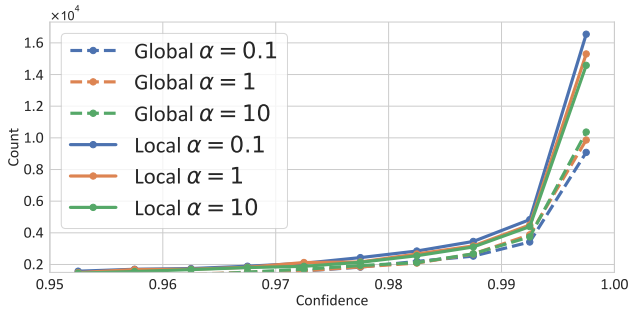
(b) Confidence > 0.96.



(c) Confidence > 0.97.

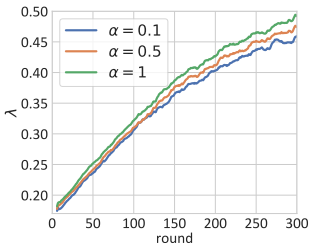


(d) Confidence > 0.98.

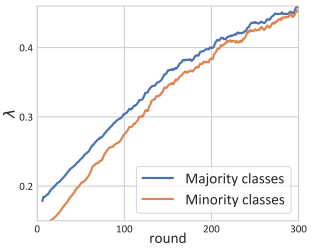


(e) Line Chart of Confidence Distribution.

Figure 9. Pseudo-label distribution of local and global models at different confidence distribution thresholds. Each subfigure represents a different threshold level, and the line chart shows the overall confidence distribution.



(a) λ under different data heterogeneity.

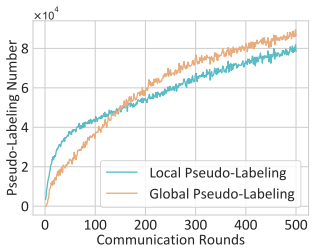


(b) λ under different data distribution.

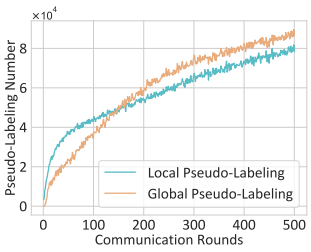
Figure 10. Ablation of λ on CIFAR-100.

models change in markedly different ways. These specific phenomena are detailed in Observations 1 and 2. In this section, we analyze the underlying reasons.

Local model. For the local model, we define the entropy of the local unsupervised data distribution as $H(Q^u(y))$,



(a) $\alpha = 1$.



(b) $\alpha = 10$.

Figure 11. The number of pseudo labels for local and global models under the additional heterogeneity setting.

Table 5. Ablation studies on soft label.

Method	Label	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$
FixMatch-LPL	Hard	49.32	49.67	49.55
	Soft	31.96	33.17	32.61
FixMatch -GPL	Hard	48.96	51.80	52.19
	Soft	48.68	50.77	48.64
SAGE	Hard	**54.18**	**55.82**	**56.06**
	Soft	53.05	54.53	55.90

aiming to explore the relationship between the entropy of the local data distribution $H(Q^u(y))$ and the entropy of model predictions $H(p(y|x, \mathcal{D}^u))$. For $p(y|\mathcal{D}^u)$, during local training, since $N^u \gg N^s$, as the local training time t increases, the local model adjusts $p(y|\mathcal{D}^u)$ based on the pseudo-labels $\hat{y}_l^i$ of the unlabeled sample $\mathbf{u}$:

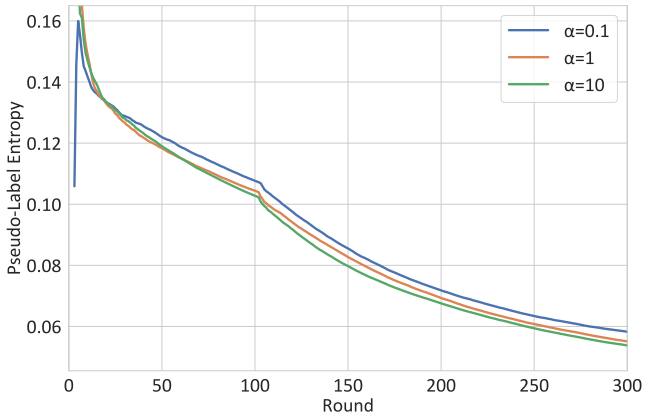
$$p^{(t+1)}(y|\mathcal{D}^u) = \gamma \cdot \left(p(\hat{y}_l^i = y|x, \mathcal{D}^u) - p^{(t)}(y|\mathcal{D}^u)\right) + p^{(t)}(y|\mathcal{D}^u). \tag{10}$$

As time t progresses, the prior distribution $p(y|\mathcal{D}^u)$ gradually couples with the true local unsupervised distribution $Q^u(y)$, this indicates a correlation between $H(p(y|\mathcal{D}^u))$ and $H(Q^u(y))$. For $p(y|x, \mathcal{D}^u)$, we expand it using Bayes' theorem as follows:

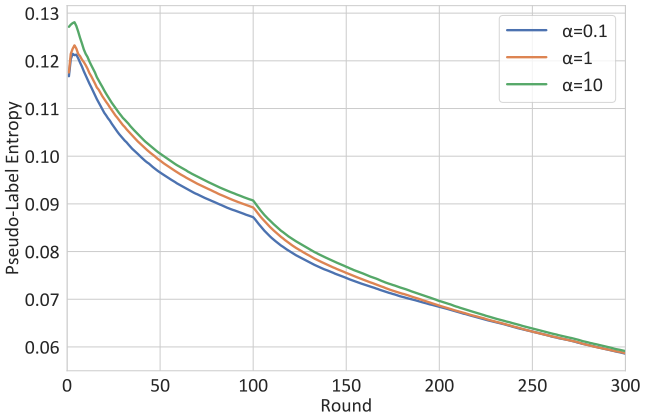
$$p(y|x, \mathcal{D}^u) = \frac{p(x|y, \mathcal{D}^u) \cdot p(y|\mathcal{D}^u)}{p(x|\mathcal{D}^u)}, \tag{11}$$

here, $p(y|\mathcal{D}^u)$ denotes the prior distribution of classes, $p(x|y, \mathcal{D}^u)$ is the feature distribution, and $p(x|\mathcal{D}^u)$ is the marginal distribution.The entropy $H(p(y|x, \mathcal{D}^u))$, when expanded according to Bayes' theorem, can be expressed as:

$$\begin{aligned} H(p(y|x, \mathcal{D}^u)) &= -\sum_y p(y|x, \mathcal{D}^u) \log p(y|x, \mathcal{D}^u) \\ &= -\sum_y \left(\frac{p(x|y, \mathcal{D}^u) \cdot p(y|\mathcal{D}^u)}{p(x|\mathcal{D}^u)}\right) \\ &\quad \cdot \log \left(\frac{p(x|y, \mathcal{D}^u) \cdot p(y|\mathcal{D}^u)}{p(x|\mathcal{D}^u)}\right). \end{aligned} \tag{12}$$

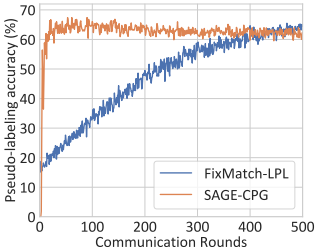


(a) Pseudo-label entropy of the global model under different heterogeneity.

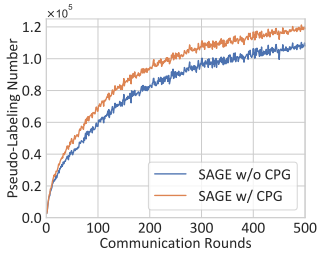


(b) Pseudo-label entropy of the local model under different heterogeneity.

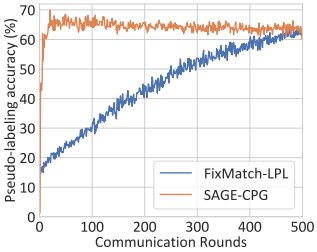
Figure 12. Changes in the pseudo-label confidence entropy of the global and local model as heterogeneity increases. Experiments show that as heterogeneity increases, global pseudo-label entropy will gradually increase, while local pseudo-label entropy will gradually decrease.



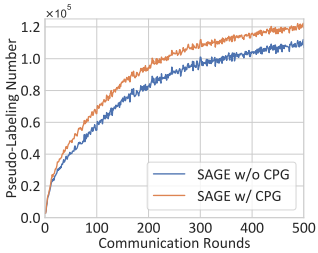
(a) Pseudo-labeling accuracy with $\alpha = 0.5$.



(b) Comparison of the number of pseudo-labels with $\alpha = 0.5$.



(c) Pseudo-labeling accuracy with $\alpha = 1$.



(d) Comparison of the number of pseudo-labels with $\alpha = 1$.

Figure 13. Additional ablation of CPG on CIFAR-100.

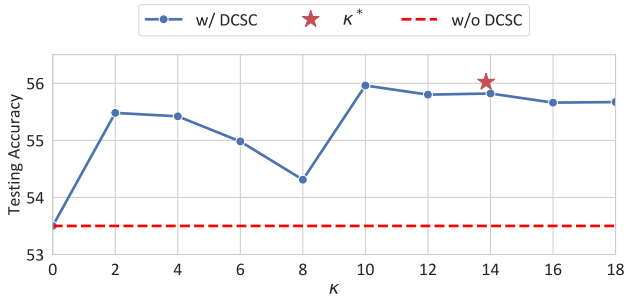
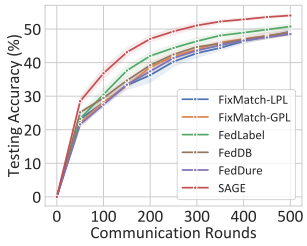
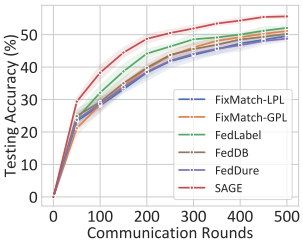


Figure 14. Ablation study on κ.



(a) Convergence curves of SAGE and other baseline methods with $\alpha = 0.1$.



(b) Convergence curves of SAGE and other baseline methods with $\alpha = 0.5$.

Figure 15. Additional convergence curves under different heterogeneities.

Consider the term associated with the prior distribution $p(u|\mathcal{D}^u)$:

$$H(p(y|x,\mathcal{D}^u)) = -\sum_y \frac{p(x|y,\mathcal{D}^u)\cdot p(y|\mathcal{D}^u)}{p(x|\mathcal{D}^u)} \log p(y|\mathcal{D}^u) - \sum_y \frac{p(x|y,\mathcal{D}^u)\cdot p(y|\mathcal{D}^u)}{p(x|\mathcal{D}^u)} \log p(x|y,\mathcal{D}^u), \tag{13}$$

the first term represents the entropy of the model's prior distribution:

$$H(p(y|\mathcal{D}^u)) = -\sum_y p(y|\mathcal{D}^u) \log p(y|\mathcal{D}^u). \tag{14}$$

The second term encapsulates a component that quantifies the feature distribution:

$$\mathrm{KL}(p(x|y,\mathcal{D}^u) \parallel p(x|\mathcal{D}^u)) = \sum_y p(y|\mathcal{D}^u) \log \frac{p(x|y,\mathcal{D}^u)}{p(x|\mathcal{D}^u)}. \tag{15}$$

Finally, the entropy of the predictive distribution $H(p(y|x,\mathcal{D}^u))$ can be written as follows:

$$H(p(y|x,\mathcal{D}^u)) = H(p(y|\mathcal{D}^u)) + \underbrace{\mathrm{KL}\left(p\left(x|y,\mathcal{D}^u\right) \middle\| p\left(x|\mathcal{D}^u\right)\right)}_{\text{Contribution of features}}, \tag{16}$$

This indicates that $H(p(y|x,\mathcal{D}^u))$ can be decomposed into the entropy of the prior distribution $H(p(y|\mathcal{D}^u))$ and a KL-divergence term contributed by the feature distribution. Under the heterogeneous setting, the local model struggles to establish robust feature discrimination across clients in the early stages of training, limiting the influence of the feature distribution on the predictive distribution. This implies that $H(p(y|x,\mathcal{D}^u))$ is mainly influenced by $H(p(y|\mathcal{D}^u))$, i.e., $H(p(y|x,\mathcal{D}^u)) \sim H(p(y|\mathcal{D}^u))$. Therefore, we conclude that $H(p(y|x,\mathcal{D}^u))$ is influenced by $H(p(y|\mathcal{D}^u))$ and correlates with $H(Q^u(y))$. As the degree of heterogeneity increases, $H(Q^u(y))$ decreases, consequently affecting $H(p(y|x,\mathcal{D}^u))$ and causing it to decrease accordingly.

Global model. The global model updates by aggregating parameters from multiple local models, it aims to learn a "compromise" global distribution that balances all client-side local distributions. The global model's confidence predictions are not directly influenced by the local class distribution of any specific client. However, As the degree of non-IIDness increases, the differences between local class distributions become more pronounced. The global model cannot simultaneously satisfy the extreme requirements of each local data distribution, so it makes high-confidence predictions only for samples with greater consistency across clients:

$$p(y|x,\theta_g) \approx \frac{1}{|\mathcal{C}_M|} \sum_{m=1}^{|\mathcal{C}_M|} p(y|x,\theta_{l,m}). \tag{17}$$

As a result, the global model's confidence predictions increasingly focus on classes with higher consistency across clients, demonstrating more conservative prediction behavior.

B.3. Experimental Support for Analysis Results

To support the analytical conclusions in Appendix B.2 and Remark 1 and 2 in Section 4.1, we conducted further exploratory experiments on CIFAR-100, analyzing how the entropy of pseudo-label confidence for the local and global models changes with heterogeneity. As shown in Fig. 12(a), when data heterogeneity intensifies, the entropy of the global model's pseudo-label confidence tends to increase, indicating greater uncertainty. This causes the global model's pseudo-labeling strategy to become more conservative. Conversely, in Fig. 12(b), the entropy of the local model's pseudo-label confidence tends to decrease as data heterogeneity increases, especially in the early stages of training when the local model has not yet developed robust feature differentiation capabilities. This suggests that the local model's predictions become overly reliant on the local imbalanced distribution, leading to overfitting and overly confident predictions.

C. Additional Ablation Study

In this section, we conduct further studies on the CPG and CDSC modules of SAGE, building on the ablation experiments in the main manuscript to demonstrate the effectiveness of these components.

C.1. Corrected Soft Label or Direct Soft Label?

The corrected soft labels produced by SAGE can mitigate the harmful effects of incorrect predictions. Additionally, we investigate whether directly using the model's predicted soft labels could achieve a similar effect. As shown in Tab. 5, directly using soft labels results in decreased performance, even worse than directly using hard labels. This is because directly using model predictions as soft labels suppresses all classes except the predicted one, thereby failing to mitigate the harm of incorrect pseudo-labels and potentially introducing extra noise. In contrast, the soft labels generated by SAGE ensure that prediction signals from both models are preserved, thereby enhancing their consensus.

C.2. Ablation Study on the correction coefficient λ

We define the dynamic correction coefficient λ to regulate the contribution of local and global pseudo-labels. We conduct an in-depth study of λ on CIFAR-100, as shown in Fig. 10: (1) According to Fig. 10(a), λ increases as heterogeneity intensifies, indicating that SAGE effectively detects the increase in heterogeneity and subsequently relies more on the global model. (2) According to Fig. 10(b), λ for local minority classes is smaller than that for local majority classes, suggesting that local minority classes tend to rely more on the predictions of the global model. (3) As training progresses, λ increases, and the gap between majority and minority narrows, suggesting an increase in the consensus between the models, consistent with the conclusion in Fig. 8.

C.3. Additional Ablation Study on CPG

In Fig. 7 of Section 5.5, we conducted the effectiveness analysis of CPG under the setting of $\alpha = 0.1$, confirming that CPG can significantly improve the quantity and quality of

Table 6. Comparison of convergence rates between SAGE and other baseline methods with $\alpha = 0.1$.

Acc.	30%		40%		45%		50%	
Method	Round ↓	Speedup ↑	Round ↓	Speedup ↑	Round ↓	Speedup ↑	Round ↓	Speedup↑
FixMatch-LPL	119	×1.00	242	×1.00	360	×1.00	562	×1.00
FixMatch-GPL	114	×1.04	226	×1.07	322	×1.12	524	×1.07
FedLabel	94	×1.27	175	×1.38	259	×1.39	429	×1.31
FedDB	103	×1.16	206	×1.17	321	×1.12	None	None
FedDure	114	×1.04	234	×1.03	341	×1.06	542	×1.04
SAGE	**60**	**×1.98**	**124**	**×1.95**	**174**	**×2.07**	**267**	**×2.10**

Table 7. Comparison of convergence rates between SAGE and other baseline methods with $\alpha = 0.5$.

Acc.	30%		40%		45%		50%	
Method	Round ↓	Speedup ↑	Round ↓	Speedup ↑	Round ↓	Speedup ↑	Round ↓	Speedup↑
FixMatch-LPL	121	×1.00	221	×1.00	334	×1.00	546	×1.00
FixMatch-GPL	113	×1.07	210	×1.05	274	×1.22	419	×1.30
FedLabel	83	×1.46	160	×1.38	222	×1.50	366	×1.49
FedDB	94	×1.29	205	×1.08	282	×1.18	492	×1.11
FedDure	110	×1.10	222	×1.00	315	×1.06	552	×0.99
SAGE	**55**	**×2.20**	**105**	**×2.10**	**159**	**×2.10**	**241**	**×2.27**

pseudo-labels. In this section, we conducted additional experiments under different heterogeneity settings to verify the robustness of CPG. As shown in Fig. 13, under the settings of $\alpha = \{0.5, 1\}$, CPG is still able to generate high-accuracy pseudo-labels in the early stages of training, supplementing the local model's pseudo-label predictions for local minority classes and further enhancing the utilization of unlabeled data.

C.4. Ablation Study on the Sensitivity Coefficient κ

In the implementation of CDSC, κ in Eq. (3) adjusts the sensitivity of the correction coefficient $\lambda(\mathbf{u})$ to the confidence discrepancy $\Delta C(\mathbf{u})$. On CIFAR-100, we divided clients with $\alpha = 1$ and varied κ in increments of 2 to study the robustness of SAGE with respect to κ. The results shown in Fig. 14 indicate that CDSC remains effective regardless of the value of κ. As κ increases, SAGE performance stabilizes, indicating low sensitivity to the hyperparameter κ.

In our experimental setup, we chose the value of κ heuristically: we referenced the confidence interval of pseudo-labels in FixMatch, $I_\tau = [0.95, 1]$, aiming for $\lambda(\cdot)$ to allocate equal weight to the local and global models when the confidence discrepancy reaches the interval length $|I_\tau| = 0.05$. Thus,

$$\exp(-\kappa^* \cdot |I_\tau|) = 0.5. \tag{18}$$

Solving this equation, we find $\kappa^* \approx 13.86$. In our experimental setups, κ^* yielded the best results.

D. Additional Comparison with Baselines

To demonstrate the effectiveness of SAGE, we present a comparison between SAGE and baseline methods with a 10% labeling ratio in Section 4 of the main manuscript. In this supplementary material, we further illustrate the robustness of SAGE with less or more labeled data by comparing SAGE with baseline methods at 20% labeling ratio. Additionally, to verify that SAGE consistently improves convergence rate, we compare the convergence of SAGE and baseline methods under varying degrees of heterogeneity.

D.1. Convergence Rate

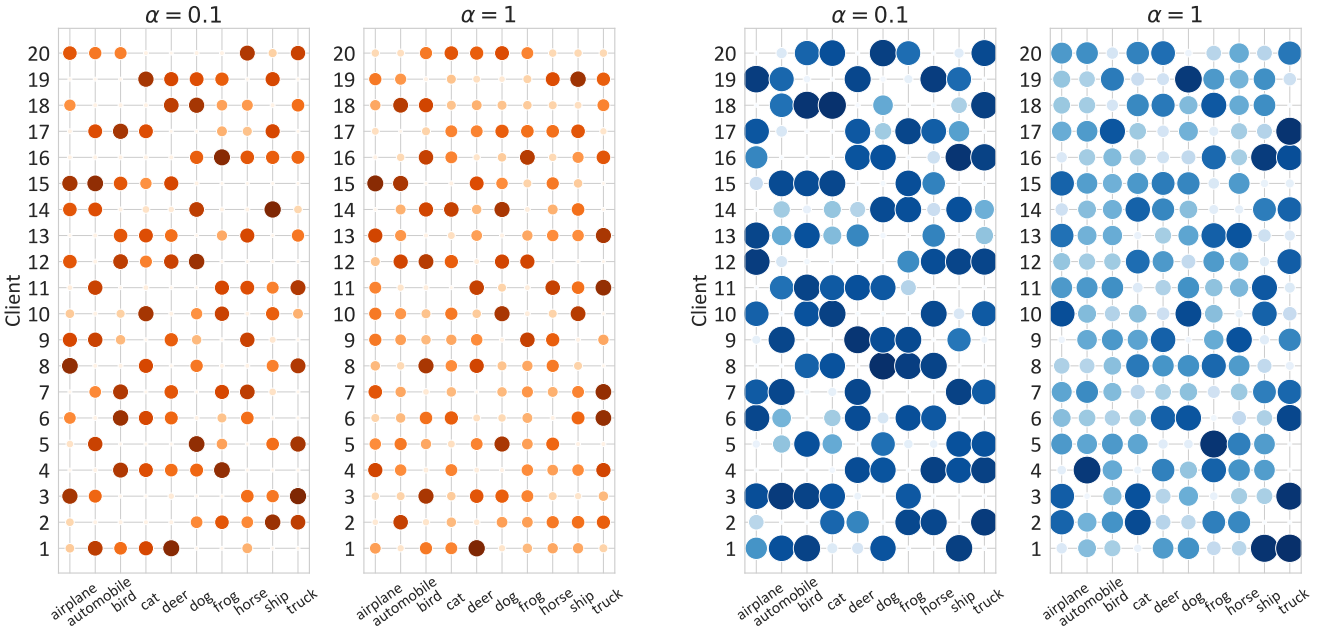
In Section 5.3 of the main manuscript, we conducted experiments under the $\alpha = 1$ setting, where the SAGE method significantly improved model convergence speed and test accuracy on the CIFAR-100 dataset. Here, we provide a detailed comparison of SAGE and baseline performance under different heterogeneity settings. As shown in Fig. 15, Tab. 6 and Tab. 7, SAGE still achieves substantial acceleration in early convergence speed under the settings of $\alpha = \{0.1, 0.5\}$.

D.2. Labeling Ratio

Tab. 8 present SAGE performance compared to baseline methods at 20% labeling ratios, respectively. SAGE consistently achieves the best performance across different labeling ratios.

Table 8. Experimental results on CIFAR-10, CIFAR-100, SVHN and CINIC-10 under 20% label. Bold text indicates the best result, while underlined text indicates the second-best result. The last row presents the improvement of SAGE over existing methods.

Methods	CIFAR-10			CIFAR-100			SVHN			CINIC-10		
	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha 1$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha 1$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha 1$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha 1$
SL methods												
FedAvg	86.37	87.06	87.97	45.72	46.57	47.55	88.37	89.05	89.97	66.24	68.29	69.21
FedProx	86.78	88.11	88.44	45.96	47.33	47.89	87.99	88.56	91.10	65.53	69.57	69.91
FedAvg-SL	90.46	91.24	91.32	67.98	68.83	69.10	94.11	94.41	94.40	77.82	80.42	81.29
SSL methods												
FixMatch-LPL	87.22	89.61	89.23	56.80	57.35	57.59	93.66	94.11	94.21	72.51	75.14	76.03
FixMatch-GPL	88.55	89.69	89.83	57.02	57.85	57.85	93.89	94.12	94.17	76.14	77.35	77.82
FedProx+FixMatch	87.47	89.46	89.56	57.44	57.91	57.87	93.60	93.93	94.05	72.36	75.15	76.06
FedAvg+FlexMatch	76.36	78.66	78.76	58.24	58.44	58.79	56.94	58.58	62.19	73.32	75.75	75.95
FSSL methods												
FedMatch	82.44	84.13	85.21	45.07	47.29	48.40	93.01	93.58	93.76	66.94	68.60	72.34
FedLabel	87.37	88.86	88.93	58.63	58.98	59.23	93.44	94.38	94.59	60.13	67.30	72.22
FedLoke	84.57	85.26	86.98	53.87	53.67	54.56	93.26	93.45	93.57	70.63	71.61	71.78
FedDure	88.56	89.63	89.95	56.14	57.23	57.89	93.81	94.42	94.37	76.21	77.13	77.75
FedDB	85.19	86.36	86.65	52.81	54.62	55.48	93.22	93.50	94.27	74.18	75.00	75.65
SAGE (ours)	**89.87**	**90.53**	**90.54**	**60.86**	**61.49**	**62.01**	**94.31**	**94.56**	**94.68**	**77.51**	**78.23**	**78.77**
	↑ 1.31	↑ 0.84	↑ 0.59	↑ 2.23	↑ 2.51	↑ 2.78	↑ 0.42	↑ 0.14	↑ 0.09	↑ 1.30	↑ 0.88	↑ 0.95



(a) Labeled Distribution

(b) Unlabeled Distribution

Figure 16. Distribution of labeled and unlabeled data across clients under different heterogeneity levels, using CIFAR-10 with 10% labeling as an example. The size of each bubble represents the count of data points of class y on client k.

E. Class Distribution Mismatch

In this work, our experiments follow the Class Distribution Mismatch setting, where both labeled and unlabeled data within each client adhere to different heterogeneous distributions. Using CIFAR-10 as an example, Fig. 16 shows the visualized data distribution across 20 clients.