Appendix

This appendix contains the following sections:

- **Section A: Dataset**
- **Section B: Cross-dataset Testing**
- **Section C: Optimization of Scene-Similarity Adaptive Local Aggregation**
- **Section D: Exploring Different Similarity Measures in SSALA**
- **Section E: Limitation and Failed Cases**
- **Section F: Detailed quantitative results**
- **Section G: Comparative Analysis of SSALA and SOTA Federated Aggregation Algorithms**

Figure 1: Client hardware devices used in our experiments.

A. Dataset

Fig. 1 shows the edge devices we use in the federated learning setting.

ShanghaiTech consists of 437 fixed-angle street surveillance videos across 13 different scenes (307 normal and 130 anomalous videos). We adopt the standard WSVAD setting, which incorporates a subset of anomalous test videos into the training set, ensuring that both training and testing sets cover all scenes.

UBnormal is a large-scale open-set dataset, comprising 543 videos across 29 scenes synthesized using Cinema4D software. The key challenge of UBnormal lies in the disjoint types of anomalies between the training and testing sets, enabling a rigorous evaluation of detection methods in real-world scenarios.

B. Cross-dataset Testing

We tested the proposed method on two datasets, UCSD Ped2 and Avenue, which both capture single-scene surveillance scenarios but differ significantly in resolution and frame rate. Specifically, the UCSD Ped2 dataset consists of 16 training videos showcasing normal pedestrian activity and 12 test videos containing abnormal events. Examples of anomalies include bikers, skaters, and vehicles appearing in a pedestrian-only area. Each video has a resolution of 240×360 pixels and a frame rate of 10 fps. The Avenue dataset consists of 16 training videos containing normal activities and 21 test videos that include various anomalous events. These anomalies involve actions such as people running, throwing objects, or walking in the wrong direction. Each video is recorded at a resolution of 360×640 pixels with a frame rate of 25 fps.

These two datasets are commonly used to evaluate semi-supervised and self-supervised methods, with the training set containing only normal videos. To align with the weakly supervised setting, we re-divided the training and test sets of these datasets, and the details of the splitting method can be found in the supplementary materials.

To assess the performance of our method on fully heterogeneous, single-scene datasets, we assign each dataset to a distinct client in accordance with the FL setup. For a fair comparison, we re-implemented AR-Net and RTFM within the FL framework using publicly available source code and the recommended parameters. These adaptations are referred to as Fed-AR-Net and Fed-RTFM, respectively.

The experimental results, presented in Table 1, demonstrate that our method achieves a balanced improvement in both AUC and FAR compared to Fed-RTFM and Fed-AR-Net. Notably, our method achieves the best FAR, highlighting its robustness in fully heterogeneous single-scene settings. This result highlights the effectiveness and adaptability of our approach, making it a compelling choice for FL-WSVAD under such challenging scenarios. It is important to note that the inputs to Fed-RTFM and Fed-AR-Net are vector features, so the tensor features from VideoMAE cannot be used as input. For a fair comparison, we also used I3D features as the input for the first branch (CAAD) and disabled CSAD.

C. Optimization of Scene-Similarity Adaptive Local Aggregation

The parameter optimization of SSALA are as follows:

$$\begin{cases} W_i^t \leftarrow W_i^t - \eta \nabla_{W_i^t} \mathcal{L}(\hat{\Theta}_i^t, D_i^{s,t}; \Theta^{t-1}) \\ A_i^t \leftarrow A_i^t - \eta \nabla_{A_i^t} \mathcal{L}(\hat{\Theta}_i^t, D_i^{s,t}; \Theta^{t-1}) \end{cases} \tag{1}$$

The gradients of the loss function with respect to A_i and W_i are computed as:

$$\begin{cases} \nabla_{A_i^t} \mathcal{L}_i = \left(W_i^t \odot \left(\theta^{t-1} - \theta_i^{t-1}\right)\right) \odot \nabla_{\theta_i^t} \mathcal{L}_i \\ \nabla_{W_i^t} \mathcal{L}_i^t = \left(\theta^{t-1} - \theta_i^{t-1}\right) \odot \nabla_{\theta_i^t} \mathcal{L}_i \end{cases} \tag{2}$$

At each iteration, the adaptive weights A_i and W_i are adjusted to better capture the personalized information for each client:

Method	Feature	PED2		Avenue	
		AUC	FAR	AUC	FAR
Fed-RTFM (Fedavg)	I3D	93.74%	3.98%	96.26%	2.78%
Fed-AR-Net (Fedavg)	I3D	87.62%	13.21%	92.18%	19.82%
Ours (CAAD+SSALA)	I3D	95.05%	1.56%	95.38%	2.80%
Ours (Fedavg)	MAE	98.21%	0.01%	96.67%	0.12%
Ours (SSALA)	MAE	**99.82%**	**0.00%**	**98.94%**	**0.00%**

Table 1: Performance comparison of FL on Ped2 and Avenue datasets with different resolutions and frame rates.

$$A_i^t = A_i^t - \mu \nabla_{A_i^t} \tag{3}$$

$$A_i^t = A_i^t - \mu \frac{\partial L_i}{\partial \theta_i} \odot \theta_i^{t-1} \tag{4}$$

$$W_i^t = W_i^t - \eta \nabla_{W_i^t} \tag{5}$$

$$W_i^t = W_i^t - \eta \frac{\partial L_i}{\partial \theta_i} \odot \left(\theta^{t-1} - \theta_i^{t-1}\right) \tag{6}$$

where μ and η are hyperparameters.

D. Exploring Different Similarity Measures in SSALA

We have evaluated various similarity measures, including SSIM, Normalized Cross-Correlation (NCC), and Feature Similarity Index (FSIM), to assess the performance of our model. As shown in Table 2, SSIM achieves the highest average AUC (97.86%) and the lowest FAR (0.03%), demonstrating its superior performance compared to both NCC and FSIM. Below, we provide a brief introduction to the two metrics.

NCC is a measure based on pixel intensity correlation, suitable for template matching and image registration tasks, especially when the images are well-aligned. It evaluates similarity by calculating the normalized dot product between images. FSIM evaluates image similarity by considering both low-level and high-level features, such as edges, textures, and phase consistency. It strikes a balance between structural and feature similarity, offering stronger robustness than NCC. FSIM is also better at adapting to image distortions compared to SSIM, and performs exceptionally well in tasks requiring more detailed perceptual quality evaluation.

As shown in Fig. 2 and Fig. 3, we also provide scene similarity analysis for both the ShanghaiTech and UBnormal datasets, using the Structural Similarity Index (SSIM) as the evaluation metric. These figures demonstrate the degree of visual consistency across different scenes, offering valuable insights into the variability and shared characteristics within each dataset.

E. Limitation and Failed cases

Illumination variations and occlusions are well-recognized challenges in computer vision, with significant progress made in recent research. However, these challenges are particularly pronounced in the WSVAD task. The absence of fine-grained annotations makes it difficult to construct an end-to-end network. Consequently, we rely on pre-trained feature extractors to capture features from video snippets. Unfortunately, these pre-trained extractors, which are trained on source domains unrelated to the VAD task, may fail to extract effective representations of anomalous objects. As a result, reasoning based on these transferred features can lead to higher FAR when inferring occluded objects.

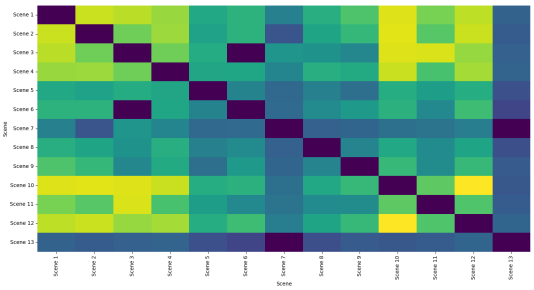

Figure 2: Scene Similarity in the ShanghaiTech Dataset Measured by SSIM.

As demonstrated in Figs. 4 and 5, the majority of existing datasets label occluded objects as normal, which reduces the difficulty for semi-supervised and weakly supervised methods. For instance, in Fig. 4, the bicycle is classified as anomalous, but during the fully occluded phase, its ground truth label is marked as normal. Similarly, in Fig. 5, the scooter is classified as anomalous, though the anomalous target is very small and close to the normal target. During the fully occluded phase, its ground truth label is also marked as normal. It is evident that small objects present a significant challenge in WSVAD.

To address these issues, we plan to incorporate human skeleton information and methods for day-night domain adaptation in future work to enhance model robustness under varying illumination conditions.

The failure cases of our method, as depicted in Fig. 6, highlight several critical challenges. The primary causes of detection failure stem from small-scale targets and discrepancies between visual and semantic interpretations. In Fig. 1(a), the anomalous target (an electric scooter) moves longitudinally towards the camera from a distance. Initially, the target is too small to be accurately detected or described, which is a common challenge across various visual tasks.

Metric	Metric	1	2	3	4	5	6	7	8	9	10	11	12	13	Average
SSIM	**AUC (%)**	97.63	99.22	88.52	96.28	98.51	99.55	97.54	97.31	100.00	99.69	100.00	98.94	100.00	**97.86**
	FARN (%)	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.03**
NCC	**AUC (%)**	96.09	98.54	86.32	91.65	98.51	76.35	97.45	94.15	100.00	98.46	100.00	98.54	100.00	**94.46**
	FARN (%)	0.18	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.04**
FSIM	**AUC (%)**	95.96	96.47	85.93	95.28	98.51	76.35	97.45	97.14	100.00	97.83	100.00	98.28	100.00	**95.13**
	FARN (%)	0.09	0.00	12.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.13**

Table 2: AUC and FAR Performance Across Similarity Metrics on the ShanghaiTech Dataset. SSIM denotes Structural Similarity Index, NCC denotes Normalized Cross-Correlation, and FSIM denotes Feature Similarity Index.

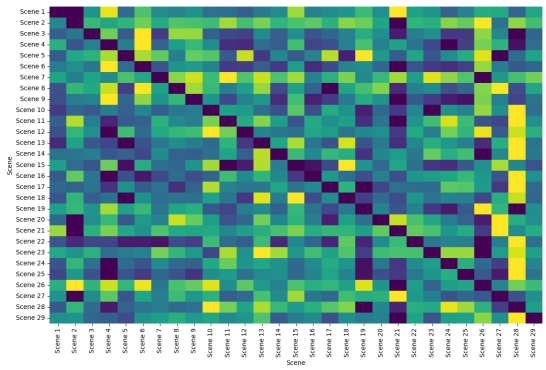

Figure 3: Scene Similarity in the UBnormal Dataset Measured by SSIM.

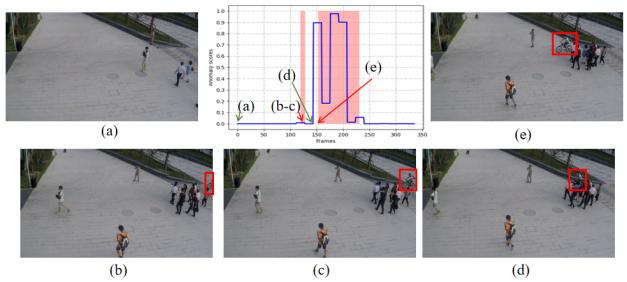

Figure 4: Ground truth vs. predictions: Occlusion scenario in video 01_0052 from the ShanghaiTech dataset.

Furthermore, as shown in Fig. 1(b) , while the model successfully detects instances like a pedestrian being knocked down, it struggles to capture the entirety of the anomalous event as specified in the semantic annotation. This issue arises from the inherent subjectivity in anomaly labeling, where the human semantic understanding of events may not align with the visual interpretation by deep networks, particularly in weakly-supervised settings.

To mitigate these challenges, we can utilize multi-scale feature representations to improve the detection of small-scale targets. Additionally, implementing a more objective ground-truth annotation method for anomalous events would enable a more accurate evaluation of weakly-supervised models' performance.

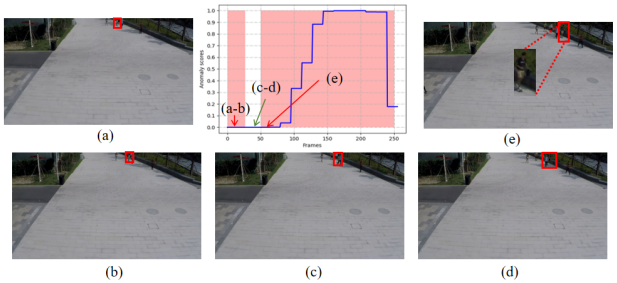

Figure 5: Ground truth vs. predictions: Occlusion scenario in video 01_0076 from the ShanghaiTech dataset.

F. Detailed quantitative results

We evaluated the performance of CAAD and CSAD individually across different scenarios in the ShanghaiTech dataset. The results are shown in Table 3.

Impact of federated learning algorithms. In Table 4, We evaluate existing FL algorithms for aggregating global parameters in SSALA, including FedAvg, FedMedian, FedAdam, FedYogi, and FedProx, on detection performance. The findings indicate that the FedAvg algorithm we used yields the best average performance, although other algorithms may outperform it in specific scenarios.

Performance Analysis Across Different Scenarios. We provide a detailed analysis of the performance of each client in different scenarios of ShanghaiTech dataset. As shown in Table 5, when considering tasks in different scenarios, we observed that most clients achieved the high AUC at the frame level, probably attributable to anomalies induced by typical anomalous behaviors, such as running. In contrast, the AUC for Clients 3 and 4 was comparatively lower, possibly due to anomalies occurring at greater distances from the camera or being influenced by perspective distortion. Table 6 shows the performance of each scene in UBnormal dataset.

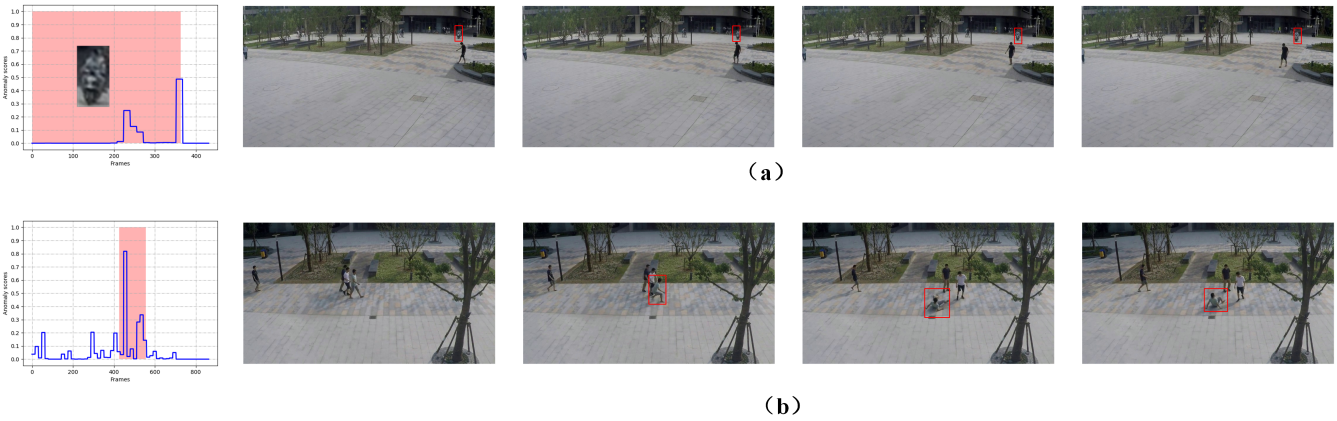

Figure 6: Typical failure cases on ShanghaiTech dataset.

Table 3: Performance Metrics for AUC and FAR across Two Detectors on ShanghaiTech.

Metric	Module	1	2	3	4	5	6	7	8	9	10	11	12	13	Average
AUC (%)	CAAD	96.39	96.62	87.90	94.36	98.29	99.26	97.26	95.15	100.00	99.55	100.00	98.76	100.00	96.86
	CSAD	97.01	95.42	87.46	93.59	98.79	98.86	98.02	95.38	100.00	98.37	100.00	98.17	100.00	96.97
FARN (%)	CAAD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	CSAD	0.37	0.00	0.11	0.00	0.00	0.05	0.00	0.00	0.00	0.31	0.00	0.00	0.00	0.12

Table 4: Performance Metrics for Different Aggregation Algorithms.

Algorithm	1	2	3	4	5	6	7	8	9	10	11	12	13	Average
FedAvg	97.63	99.22	88.52	96.28	98.51	99.55	97.54	97.31	100.00	99.69	100.00	98.94	100.00	**97.86**
FedTrimmedAvg	97.86	95.88	84.53	95.46	98.59	99.52	95.82	97.27	100.00	94.05	100.00	98.57	100.00	97.35
FedAdam	94.85	98.51	81.95	89.24	98.27	99.17	95.68	92.88	100.00	86.90	100.00	98.16	100.00	95.11
FedYogi	95.17	96.15	83.21	94.19	97.92	98.09	97.77	94.59	100.00	77.20	100.00	97.33	100.00	95.25
FedAvgM	97.23	96.37	85.45	91.11	98.98	98.13	97.20	94.22	100.00	92.96	100.00	98.67	100.00	96.46
FedProx	97.21	92.72	86.09	95.38	98.58	98.81	96.99	96.51	100.00	98.60	100.00	98.79	100.00	97.12

Table 5: Detailed Performance on ShanghaiTech dataset.

Metric	1	2	3	4	5	6	7	8	9	10	11	12	13	Average
AUC (%)	97.63	99.22	88.52	96.28	98.51	99.55	97.54	97.31	100.00	99.69	100.00	98.94	100.00	97.86
FAR_N (%)	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03

G. Comparative Analysis of SSALA and SOTA Federated Aggregation Algorithms

Fig. 8 presents a 2D visualization of the local learning trajectory in a pathological heterogeneous setting on CIFAR-10. The figure shows the normalized learning path of a model across different rounds of federated learning, projected onto a 2D plane using PCA. Specifically, it displays the local model projections for client ID 8 from round 130 to round 200. The green dots indicate the model projections from rounds 130 to 160, during which only the FedAvg model was applied without using the SSALA local initialization algorithm. In contrast, the red dots represent model projections from rounds 160 to 200, where the SSALA local initialization algorithm was activated. This suggests that SSALA is more effective at guiding the local model's updates, leading to smoother convergence and more efficient learning by capturing and utilizing relevant information from the global model.

Following the evaluation protocol used in FedALA, Table 7 presents the test accuracy in both pathological and practical heterogeneous settings, highlighting the performance of various FL methods, including FedAvg, FedProx, and SSALA, across datasets like MNIST, Cifar10, Cifar100, and AG News. In the pathological heterogeneous setting, where clients receive disjoint subsets of data (e.g., 2 out of 10 classes for MNIST and Cifar10), the results showcase the challenge of extreme data heterogeneity. In the practical heterogeneous setting, controlled by a Dirichlet distribution (β = 0.1) to simulate more realistic uneven data distributions among clients, SSALA consistently outperforms other

Table 6: Detailed Performance on UBnormal dataset.

	1	2	3	4	5	6
AUC (%)	78.46	88.69	81.78	91.90	85.89	63.34
FARN (%)	0.00	0.00	0.00	0.00	0.00	0.00
	7	**8**	**9**	**10**	**11**	**12**
AUC (%)	66.42	73.56	78.76	69.22	62.10	92.67
FARN (%)	0.00	0.00	0.00	0.00	0.00	0.00
	13	**14**	**15**	**16**	**17**	**18**
AUC (%)	77.08	77.73	87.34	64.82	76.24	87.12
FARN (%)	0.00	0.00	0.00	0.00	0.00	0.00
	19	**20**	**21**	**22**	**23**	**24**
AUC (%)	37.55	78.90	87.18	63.33	72.63	72.98
FARN (%)	0.00	0.00	0.00	0.00	0.00	0.00
	25	**26**	**27**	**28**	**29**	**Average**
AUC (%)	82.44	93.68	91.62	60.69	68.53	76.51
FARN (%)	0.00	0.00	0.00	0.00	0.00	0.00

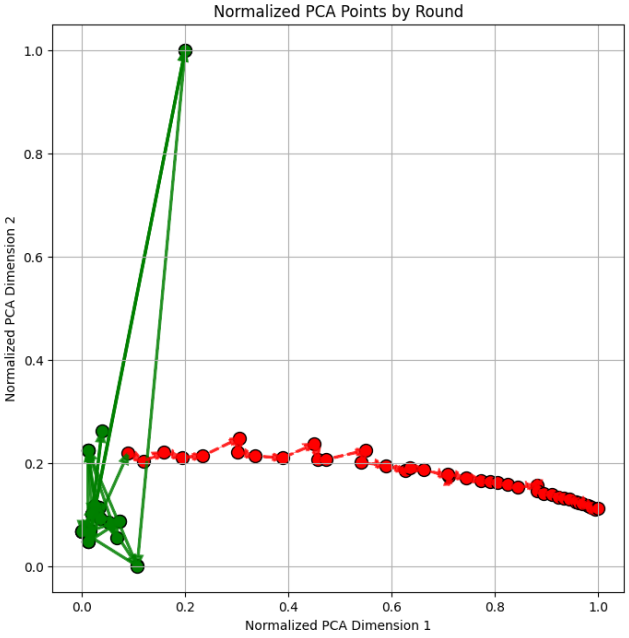

Figure 7: PCA Projection of Learning Trajectories.

methods. SSALA demonstrates particularly strong performance in highly heterogeneous environments, achieving superior accuracy. These results underline SSALA's effectiveness in addressing statistical heterogeneity in federated learning scenarios.

Table 8 presents the test accuracy and improvements brought by the SSALA module on the Tiny-ImageNet, MNIST, and Cifar100 datasets, under various heterogeneity and scalability settings. The table compares the performance of SSALA against state-of-the-art FL methods, including FedAvg, FedProx, and various personalized FL approaches. It highlights that SSALA consistently achieves superior accuracy, with notable improvements, particularly in highly heterogeneous settings (Dirichlet 0.01) and with larger numbers of clients. These results demonstrate the effectiveness of ALA in precisely capturing the desired information from global models, yielding better personalization for local models. Furthermore, the scalability experiments show that SSALA maintains high accuracy with minimal degradation, even as the number of clients increases to 100, showcasing its robustness and applicability across various federated learning environments.

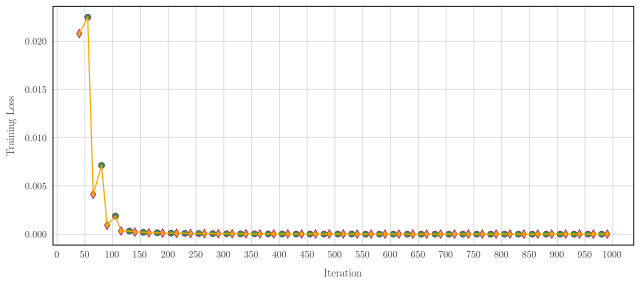

Figure 8: The training loss of the global objective in SSALA on Cifar10 in the pathological setting.

As shown in Fig. 8, we record the average loss of the local models after training (orange diamonds) and the average loss before local training following local initialization (green dots) at each iteration. For clarity, we display a point every 30 rounds. When the number of iterations exceeds 500, both the loss function values represented by the orange diamonds and green dots fall below 1×10^{-3}. Moreover, the loss values corresponding to the orange diamonds and green dots are nearly equal, indicating that SSALA has converged.

Settings	Pathological heterogeneous setting			Practical heterogeneous setting				
Methods	MNIST	Cifar10	Cifar100	Cifar10	Cifar100	TINY	TINY*	AG News
FedAvg	97.93±0.05	55.09±0.83	25.98±0.13	59.16±0.47	31.89±0.47	19.46±0.20	19.45±0.13	79.57±0.17
FedProx	98.01±0.09	55.06±0.75	25.94±0.16	59.21±0.40	31.99±0.41	19.37±0.22	19.27±0.23	79.35±0.23
FedAvg-C	99.79±0.00	91.38±0.03	66.17±0.03	90.34±0.01	51.80±0.02	30.67±0.08	36.94±0.10	95.89±0.25
FedProx-C	99.80±0.04	91.97±0.04	66.07±0.08	90.33±0.01	51.84±0.07	30.77±0.13	38.78±0.52	96.10±0.22
Per-FedAvg	99.63±0.02	89.63±0.23	56.80±0.26	87.74±0.19	44.28±0.33	25.07±0.07	21.81±0.54	93.27±0.25
FedRep	99.77±0.03	91.93±0.14	67.56±0.31	90.40±0.24	52.39±0.35	37.27±0.20	39.95±0.61	96.29±0.11
pFedMe	99.75±0.02	90.11±0.10	58.20±0.14	88.09±0.32	47.34±0.46	26.93±0.19	33.44±0.33	91.41±0.22
Ditto	99.81±0.00	91.49±0.06	67.23±0.07	90.35±0.04	52.87±0.64	32.15±0.04	35.92±0.43	95.45±0.17
FedAMP	99.76±0.02	90.79±0.16	64.34±0.37	88.70±0.18	47.69±0.49	27.83±0.11	29.11±0.15	94.18±0.09
FedPHP	99.73±0.00	90.01±0.00	63.09±0.04	88.92±0.02	50.52±0.16	28.17±3.26	29.90±0.51	94.38±0.12
FedFomo	99.83±0.00	90.71±0.07	62.49±0.22	88.06±0.02	45.39±0.45	26.33±0.22	26.84±0.11	95.84±0.15
APPLE	99.75±0.01	90.97±0.05	65.80±0.08	89.37±0.11	53.22±0.20	35.04±0.47	39.93±0.52	95.63±0.21
PartialFed	99.86±0.01	89.60±0.13	61.39±0.12	87.38±0.08	48.81±0.20	35.26±0.18	37.50±0.16	85.20±0.16
FedALA	99.88±0.02	91.83±0.02	67.30±0.06	90.78±0.03	55.27±0.04	40.59±0.04	42.11±0.05	96.10±0.08
Our	**99.91±0.01**	**92.20±0.07**	**67.85±0.04**	**91.08±0.02**	**55.51±0.05**	**40.68±0.05**	**44.38±0.04**	**96.33±0.09**

Table 7: The test accuracy (%) in the pathological heterogeneous setting and practical heterogeneous setting.

	Heterogeneity		Scalability		Applicability of module			
Datasets	Tiny-ImageNet	MNIST	Cifar100		Tiny-ImageNet		Cifar100	
Methods	$Dir(0.01)$	$Dir(0.1)$	100 clients	100 clients	Acc.	Imps.	Acc.	Imps.
FedAvg	15.70±0.46	98.81±0.01	31.95±0.37	39.51±1.22	40.68±0.05	21.22	55.51±0.05	23.62
FedProx	15.66±0.36	98.82±0.01	31.97±0.24	31.97±0.24	41.31±0.07	21.94	56.36±0.55	24.37
FedAvg-C	49.88±0.11	99.65±0.00	47.90±0.12	47.94±0.25	—	—	—	—
FedProx-C	49.84±0.02	15.70±0.46	48.02±0.02	48.11±0.19	—	—	—	—
Per-FedAvg	39.39±0.30	98.90±0.05	36.07±0.24	47.96±0.8	39.02±0.16	13.95	55.67±0.44	11.36
FedRep	55.43±0.15	99.48±0.02	44.61±0.20	41.48±0.05	—	—	—	—
pFedMe	41.45±0.14	99.52±0.02	46.45±0.18	43.27±0.46	31.04±0.17	3.74	47.55±0.41	0.21
Ditto	50.62±0.02	99.64±0.00	52.89±0.22	48.94±0.04	43.28±0.12	11.13	56.41±0.11	3.54
FedAMP	48.42±0.06	99.47±0.02	40.43±0.17	—	27.88±0.17	0.05	47.81±0.23	0.12
FedPHP	48.63±0.02	99.58±0.00	49.70±0.31	49.99±0.73	31.21±0.54	3.04	53.60±0.16	3.08
FedFomo	46.36±0.54	99.33±0.04	38.91±0.08	37.70±0.10	—	—	—	—
APPLE	48.04±0.10	15.70±0.46	52.81±0.29	—	—	—	—	—
PartialFed	49.38±0.02	99.67±0.01	39.31±0.01	36.49±0.07	35.36±0.05	0.10	49.06±0.04	0.25
FedALA	57.03±0.03	**99.71±0.00**	**54.68±0.57**	**54.81±0.03**	—	—	—	—
Our	**57.35±0.04**	99.70±0.01	52.01±0.33	53.01±0.04	—	—	—	—

Table 8: The test accuracy (%) (and improvement (%)) on Tiny-ImageNet, MNIST, and Cifar100.