SUPPLEMENTARY MATERIAL TO THE PAPER ENTITLED “CHASE: CLIENT HETEROGENEITY-AWARE DATA SELECTION FOR EFFECTIVE FEDERATED ACTIVE LEARNING”

6.1 Key Notations

The key notation in this paper is provided in Table 6.

TABLE 6: Key notations.

Symbol	Meaning
$\boldsymbol{x}_i, y_i$	The *i*-th sample and its corresponding truth label
$\hat{y}_i$	The predicted label of sample x_i
R, r	Overall communication round and the r-th round
E, e	Overall local training epoch and the e-th epoch
K, k	Overall size of clients and the client with serial number k
$\mathcal{B}$, $\mathcal{B}_k$	Overall fixed budget and the budget of client k
$\mathcal{D}_k^L$, $\mathcal{D}_k^U$, $\mathcal{D}_k^S$	Labeled, unlabeled, and selected sets at client k
N_k^L, N_k^U, N_k^S	*Size of* labeled, unlabeled, and selected sets at client k
$\tilde{\mathcal{D}}_k^U$	Unlabeled subset of client k by subset sampling
$\mathcal{D}_k^D$	Dormant set of client k
$\mathcal{D}_k^{U_0}$	Unlabeled data of client k with zero EVs
$\boldsymbol{\omega}^{r-1}$	The received global model of clients at round r
$\boldsymbol{\omega}_k^r$	The trained local model of client k after E epochs
$\tilde{\boldsymbol{\omega}}_k^e$	The partially trained local model of client k after e epochs
$\boldsymbol{V}_k^r(\boldsymbol{x}^i)$	The historical variations on sample $\boldsymbol{x}_i$ of client k at round r
$EV_k^r(\boldsymbol{x}^i)$	The epistemic variation on sample $\boldsymbol{x}_i$ of client k at round r

6.2 Details of the Experiment Settings

6.2.1 Datasets

We evaluate the efficacy of CHASe using well-known **image datasets** following studies [9], [29]:

- MNIST: This dataset for digit recognition includes 60,000 training images and 10,000 test images.
- EMNIST: As an extension, EMNIST contains 814,255 images (731,668 for training and 82,587 for testing), featuring numbers, upper and lower case letters across 62 categories.
- CIFAR-10: This set comprises 60,000 32x32x3 pixel color images within 10 classes, split into 50,000 for training and 10,000 for testing.
- CIFAR-100: Similar to CIFAR-10 but with 100 classes. Each class has 600 images, divided into 500 for training and 100 for testing.

CIFAR-100 and EMNIST are notable for their high number of classes, with EMNIST also being large-scale, providing varied environments for testing effectiveness and robustness.

To test CHASe’s generalizability, we also include the Shakespeare **text dataset** [53], which includes 4,226,15 samples across 80 categories, based on the complete works of William Shakespeare. Following Leaf’s guidelines [54], 90% of these samples are used for training, with the remaining 10% for testing in the next character prediction tasks.

6.2.2 Baselines

We evaluate the performance of our method against the following baselines:

- **FedRandom** that combines the FedAvg [6] approach with the random sample selection for annotation.
- **FedEntropy** that combines FedAvg with the entropy-based AL [12] that selects samples with high prediction entropy.
- **FedCoreset** that combines FedAvg with the Core-set [13] based AL.
- **FedLL4AL** that combines FedAvg with LL4AL [14]. As an adaption, the additional loss prediction employed in LL4AL is updated and aggregated also following FedAvg.
- **FAL(•)** that applies each classical AL method (i.e., Entropy, Core-set and LL4AL) to the FAL scheme [16] where data are being selected using the global model.

6.2.3 Metrics

We measure the **effectiveness** of all methods in terms of the accuracy on the testing set with a fixed number of communication rounds, aligning with prior AL works [14], [44]. We report the average accuracy measures on 3 runs with different seeds. On the other hand, we measure the **efficiency** of all methods in terms of both execution time and inference cost per FAL round. Particularly, the inference cost refers to the number of inferred unlabeled samples. We focus on reporting efficiency-related evaluation results in Section 4.3.2.

6.2.4 FL-related Setting

We assume $K = 10$ clients in MNIST, CIFAR-10 and CIFAR-100; and $K = 100$ in the larger EMNIST and Shakespeare datasets. We set the local epoch number $E = 10$ and communication round $R = 200$. Following a previous work [6] on simulating *various degrees* of Non-IID settings, we specify each client has $C_k = 5$, 2, and 10 classes for both labeled and unlabeled sets on CIFAR-10, MNIST and EMNIST, respectively. Following another study [55] that specifically simulates the Non-IID setting for CIFAR-100, we use the Dirichlet distribution $Dir(\alpha)$ to generate heterogeneous data partitions with the concentration parameter $\alpha = 1e - 2$. For Shakespeare, we follow previous studies [6], [54], creating a Non-IID dataset with 1,146 clients representing each speaking role in each play. However, we only include clients with more than 10,000 samples and randomly select 100 of these clients to participate in FAL.

6.2.5 AL-related Setting

To initiate the labeled set, we randomly select small data portions from the training set, specifically 4% for CIFAR-10 and Shakespeare, 10% for CIFAR-100, 1.33% for MNIST, and 2% for EMNIST. Regarding the annotation budget, we base our assumption on each client being honest and non-adversarial [56], thus not providing error labels. We examine two distinct client behavior scenarios:

- AbCo [Absolute Cooperation]: All clients are fully engaged in the annotation. The requirement per client per round is $\mathcal{B}_k = 10$ samples, e.g., given $K = 10$ clients, the server’s total budget per round is $\mathcal{B} = \mathcal{B}_k \times K = 100$.
- ReCo [Relative Cooperation]: Not all clients are equally cooperative. We use the Gaussian distribution to organize clients into passive, ordinary, and aggressive groups, and their ratio is 2:6:2. A passive/ordinary/aggressive client will annotate $\mathcal{B}_k = 5/7/10$ samples every 5/3/1 round(s).

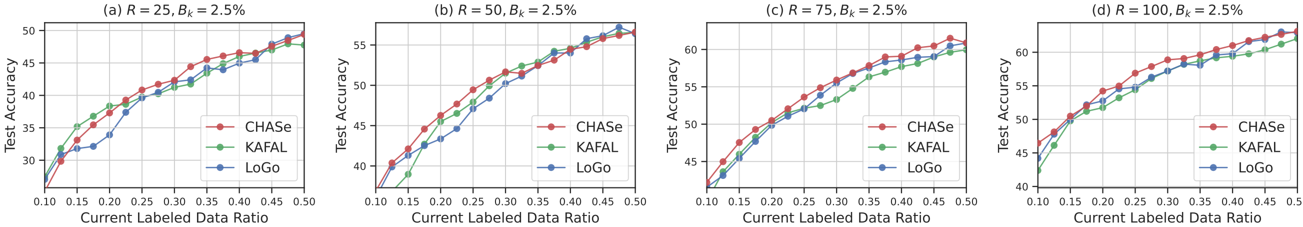


Fig. 17: Adaption to a continual FAL setting with a fixed annotation budget (simulated machine annotators). The results are smoothed using a moving average with a window size of 3 to reduce noise data.

6.2.6 *Other Settings*

Considering the limited storage space and computational resources of the clients, as in existing FL works, CNN (2 Conv + 2 FC) [57] is used for MNIST, EMNIST and CIFAR-10, and VGG-19 for CIFAR-100. We use the SGD optimization with a learning rate of 1e-3 for MNIST, EMNIST, and CIFAR-10, and 1e-2 for CIFAR100. For Shakespeare, following [6], we train two layers stacked character-level LSTM with a learning rate of 1.47 to predict the next character. For the alignment loss ℓ_{align}, we set $\tau = 0.5$ following the work [27], $\mu = 0.1$, 0.1, 1, 1, 0.3 for MNIST, EMNIST, CIFAR-10, CIFAR-100, and Shakespeare respectively. For FAmS, we set the subset size $N_s^U = 500$, the awaken threshold $\mathcal{B}' = 3 \times \mathcal{B}_k$ and the awaken ratio $\beta = 0.4$. The experiments on MNIST, CIFAR-10, and Shakespeare are performed on a single machine with Quadro RTX 6000 GPUs and PyTorch 1.6.0, while those on EMNIST and CIFAR-100 use Quadro RTX A5000 GPUs and PyTorch 1.8.0. The detailed hyperparameters for training are listed in Table 7.

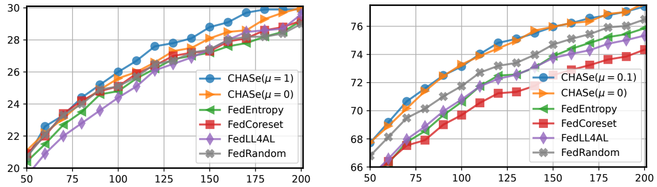
TABLE 7: The hyperparameters associated with datasets.

Hyperparameter	MNIST	EMNIST	CIFAR-10	CIFAR-100	Shakespeare
Learning Rate	0.001	0.001	0.001	0.01	1.47
Dropout Probability	0.5	0.5	0.5	0.0	0.2
Rounds (R)	200	200	200	200	200
Local Batch Size(B)	10	10	10	10	10
Local Epoch(E)	10	10	10	10	10
Size of Clients(K)	10	100	10	10	100
Subset Size(N_S^U)	500	500	500	500	500
Awaken Ratio (β)	0.4	0.4	0.4	0.4	0.4
Temperature (τ)	0.5	0.5	0.5	0.5	0.5
Weight in ℓ_{align} (μ)	0.1	0.1	1	1	0.3

6.3 Adaption CHASe to Continual FAL

Here, we explore the continual FAL with a fixed annotation budget. In this scenario, the annotation capacity of clients remains limited and constant, regardless of the number of training rounds. This reflects cases where clients rely on automated or resource-constrained methods (e.g., machine annotators), allowing them to label a fixed amount of data ($B_k = 2.5\%$) per sampling interval, irrespective of communication frequency.

Fig. 17 indicates that our method is always better or closer to KAFAL and LoGo. However, because the model itself has not learned enough knowledge in this scenario, the importance of unlabeled samples under the relatively large budget, therefore, the superiority of the strategy is easily covered under this continual FAL.



(a) CIFAR-100 / Dir(1e-2) / ReCo (b) EMNIST / C_k=10 / ReCo

Fig. 18: Test accuracy curves on CIFAR-100 and EMNIST.

6.4 Visualization of Abaltion Studay

To ensure a fair comparison, we conducted additional studies by removing the alignment loss and evaluating CHASe without it, as illustrated in Fig. 8(b) on Page 8 and Table 1 & Table 2 on Page 9 of the manuscript. Space constraints in the original manuscript limited our discussion to CIFAR-10, but here we provide further assurance by presenting additional ablation results for CIFAR-100 and EMNIST, shown in Figs. 18(a) and 18(b) of this letter. These experiments demonstrate that CHASe ($u = 0$) consistently outperforms other baselines, with decision boundary calibration further improving task performance under the same communication round.