When Data is Scarce, Learn to Adapt: Robust Federated Learning via Adversarial Meta-Optimization

Md Zarif Hossain Awal Ahmed Fime Ahmed Imteaj
Florida Atlantic University
{mdzarifhossa2025, afime2025, aimteaj}@fau.edu

https://speedlab-git.github.io/FAML-CVPR

A. Hyperparameter and FAML Components Ablations

A.1. Ablation on Robust Regularizer λ_{out}

Figure 3 illustrates the effect of the robust regularizer hyperparameter λ_{out} on clean and robust accuracy (RA) across SVHN [5], FMNIST [8], CIFAR10 and CIFAR100 [3] datasets. We observe that smaller values of λ_{out} yield higher clean accuracy but lower adversarial robustness across datasets. As λ_{out} increases, RA improves gradually, accompanied by a slight degradation in clean performance.

Table 1. Hyperparameters for different datasets.

Datasets	**Meta-LR** (η)	**LR** (μ)	**Robust Reg.** (λ_{out})
SVHN	0.01	0.01	0.6
FMNIST	0.007	0.01	0.6
CIFAR10	0.01	0.01	0.5
CIFAR100	0.03	0.01	0.6

However, when $\lambda_{\text{out}} > 0.7$, the clean accuracy across all datasets drops significantly. Specifically, the optimal trade-off between clean and robust performance is achieved at $\lambda_{\text{out}} = 0.6$ for SVHN, FMNIST, and CIFAR100, and $\lambda_{\text{out}} = 0.5$ for CIFAR10. The training hyperparameters (e.g., meta-learning rate η, learning rate μ) for each dataset are summarized in Table 1.

A.2. Ablation on Drift Regularizers

Figure 2 presents the ablation study on the impact of the drift regularizers, evaluated on the CIFAR10 dataset. The model drift, measured as the ℓ_2 distance between global and local parameters across communication rounds. In this ablation, we compare three variants of FAML: without any drift regularizer, with the clean drift regularizer, and with both clean and adversarial drift regularizers. The latter two variants employ the adaptive mask with drift regularizers.

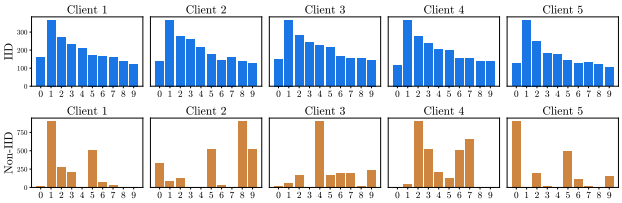


Figure 1. Data distribution across clients under Dirichlet partitioning ($\beta = 0.1$) on CIFAR10 dataset.

From Figure 2, we observe that without any drift regularization, the model exhibits high drift magnitude during early communication rounds and drift increases significantly at later stages. Incorporating the clean drift regularizer substantially mitigates this issue by aligning local predictions with the global model's clean predictions. However, a gradual rise in drift is observed in later stages as the adversarial training client introduces bias into the global update. The most stable training behavior is observed when both clean and adversarial regularizers are applied, resulting in the lowest drift magnitude throughout training. This indicates that adversarial regularization further suppresses the drift induced by adversarially trained clients, ensuring consistent and stable convergence across communication rounds.

A.3. Ablation on FAML components

For a more detailed analysis of our methodology, we examine the contribution of four key componenets in FAML: Clean Drift Regularizer (CDR), Robust Drift Regularizer (RDR), Adaptive Mask (AM), and Temporal Moving Average (TMA). To this end, we construct a initial baseline, denoted as $\mathcal{B}$, which only incorporates robust regularizer without any of the four componenets. We then incrementally introduce the key components to form three intermediate baselines, $\mathcal{B}_1$, $\mathcal{B}_2$, $\mathcal{B}_3$, allowing us to isolate their individual and combined impact on clean and robust performance. Table 2, summarizes the ablation study on four key

Table 2. Ablation study of the key components in our method on SVHN under PGD-20 attack. Columns denote the use of Clean Drift Regularizer **(CDR)**, Robust Drift Regularizer **(RDR)**, Adaptive Mask **(AM)**, and Temporal Moving Average (**TMA**). We report both clean and robust accuracy (%).

FAT Method	CDR	RDR	AM	TMA	Clean	Robust
CalFAT (baseline)	–	–	–	–	84.15	41.68
$\mathcal{B}$: Base (FAML w/o Reg. & mask)	✗	✗	✗	✗	79.42	37.50
$\mathcal{B}_1$: + Clean Reg. only	✓	✗	✗	✗	77.12	34.42
$\mathcal{B}_2$: + Clean + Robust Reg.	✓	✓	✗	✗	77.03	35.15
$\mathcal{B}_3$: + Clean + Robust Reg. + Mask	✓	✓	✓	✗	85.12	43.20
FAML (ours)	✓	✓	✓	✓	**85.30**	**45.92**

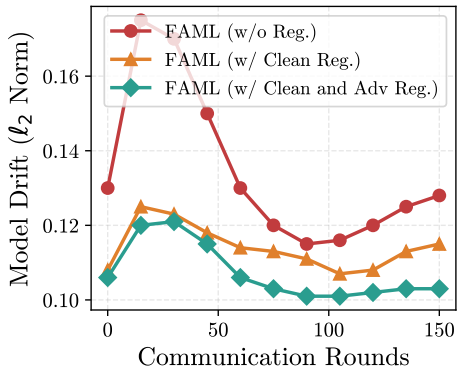


Figure 2. Ablation study of drift regularizers on CIFAR10. Model drift (ℓ_2 norm) is measured across communication rounds for FAML without drift regularizers, with clean drift regularaizer, and with both clean and adversarial drift regularaizer. Incorporating drift regularizers significantly reduces drift magnitude and stabilizes convergence.

components of FAML on the SVHN dataset under PGD-20 attack. We observe that without any drift regularizers and adaptive mask the baseline $\mathcal{B}$ exhibits notable performance degradation in both clean accuracy (79.42%) and robust accuracy (37.50%). This indicates that naively adopting meta-learning with only robust regularizer leads to unstable training due to client drift, resulting in poor performance.

Interestingly, with only clean drift regularizer in ($\mathcal{B}_2$) and with both drift regularizers in ($\mathcal{B}_3$), we observe further degradation in both clean and robust accuracy, even lower than the initial baseline $\mathcal{B}$. This occurs because, without adaptive mask, local predictions are aligned with confident but incorrect global model's predictions, which propagates unreliable decision boundaries and leads to performance deterioration. Notably, incorporating Adaptive Mask in ($\mathcal{B}_3$ yields a substantial improvement, with clean accuracy increasing to 85.12% and robust accuracy improving to 43.20%. This highlights the importance of adding adaptive mask, which adaptively filters out unreliable predictions from global model. As a result, the adaptive mask emerges as one of the most critical components of FAML. Finally, adding TMA to $\mathcal{B}_3$ results in a slight improvement and achieves the highest clean and robust accuracy, 85.30% and 45.92%, respectively.

B. Additional Details on Experimental Setup

B.1. Model Architectures

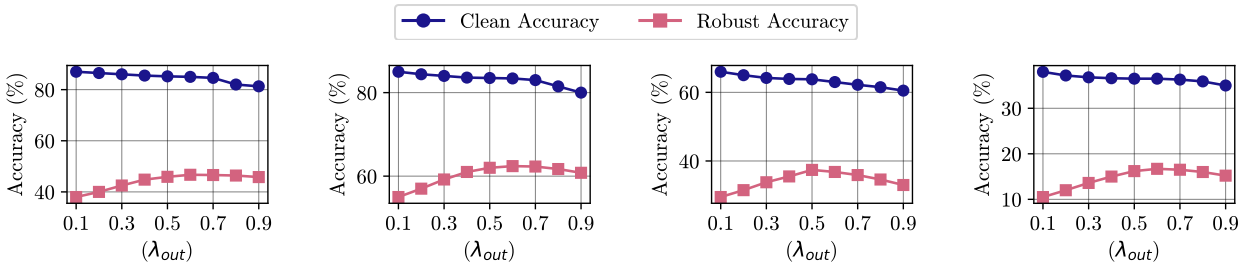
For SVHN and FMNIST, we use a simple CNN architecture consisting of three convolutional layers followed by three fully connected layers. For CIFAR10 and CIFAR100, we adopt a slightly deeper architecture (similar to DBFAT [12], CALFAT [1]) with four convolutional layers followed by three fully connected layers.

B.2. Data partition with Dirichlet Distribution

In our experiments, to simulate real-world statistical heterogeneity, we partition the training data among clients using a Dirichlet distribution. Specifically, we sample $p_l^k \sim \mathrm{Dir}(\beta)$ and allocate a proportion p_l^k of the data with label l to client k, where $\mathrm{Dir}(\beta)$ denotes the Dirichlet distribution with concentration parameter β. In Figure 1, we visualize the label distribution of 5 clients on the CIFAR10 dataset for $\beta = 10$, representing the IID scenario, and for $\beta = 0.1$, representing a highly skewed non-IID scenario. As shown in the figure, in the IID setting, the data of each client are relatively homogeneous, containing a nearly equal number of samples from each class. In contrast, under the non-IID setting, the label distribution becomes highly skewed. Some clients receive far fewer samples than others, and most of their data may come from only one class while several classes are completely missing. For example, in Figure 1, client 5 has no samples from classes 1, 4, and 8.

B.3. Data Scarcity Setting

In our experiments, we evaluate and compare FAML, which is trained using only 20% of the total training data. For CIFAR10, this corresponds to approximately $10,000$ samples out of $50,000$ training samples, which are then distributed among participating clients following the Dirichlet distribution settings. This setting allows us to evaluate the



(a) SVHN (b) FMNIST (c) CIFAR10 (d) CIFAR100

Figure 3. Ablation study of the robust regularizer λ_{out}. Clean and robust accuracy reported for different values of λ_{out} on SVHN, FMNIST, CIFAR10, and CIFAR100.

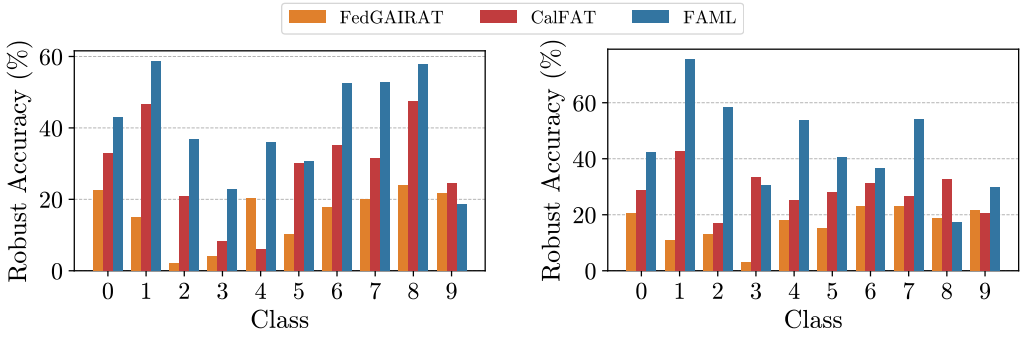


Figure 4. Per-class average robust accuracy (under PGD-20) of FAT methods on CIFAR10 **(left)** and SVHN **(right)** datasets.

Table 3. Robust accuracy (%) of different FAT methods under PGD-20 attacks on CIFAR10 and SVHN. Results are reported for ℓ_∞ perturbation budgets of $\epsilon \in \{8/255, 16/255, 32/255\}$.

Methods	**CIFAR10**			**SVHN**		
	$8/255$	$16/255$	$32/255$	$8/255$	$16/255$	$32/255$
NaiveFAT [13]	29.14	17.24	10.23	19.61	15.32	10.61
FedPGD [4]	26.59	15.13	8.54	19.33	15.12	11.87
FedTRADES [10]	27.75	16.48	8.91	36.92	28.35	20.11
FedMART [7]	18.49	10.61	7.32	19.71	15.54	11.34
FedGAIRAT [11]	29.66	20.41	13.42	38.30	30.47	24.78
FedRBN [2]	26.87	19.67	15.88	32.32	24.23	18.77
CalFAT [1]	31.12	27.45	18.65	41.68	36.33	29.65
FAML	**37.38**	**31.48**	**24.45**	**45.92**	**39.12**	**34.76**

model's performance in realistic low-data regimes. Importantly, throughout all experiments in both the main paper and the Appendix, the baseline FAT methods such as CalFAT, DBFAT, and FedGAIRAT are trained using the full 100% of the training data. In stark contrast, FAML achieves its performance while being trained on just 20% of the data.

B.4. Domain and Distribution Shift Evaluation Setup

To evaluate the generalization capability of FAML under domain and distribution shifts, we conduct cross-domain and out-of-distribution (OOD) experiments. In the cross-domain setting, the model is trained on SVHN, which contains real-world house number images, and evaluated on the unseen MNIST dataset consisting of handwritten digits.

For OOD evaluation, models trained on CIFAR10 are tested on STL10 and CIFAR10C. Evaluation on STL10 reflects a data distribution shift, where the training and testing data drawn from different data distribution (e.g., CIFAR10 from tiny images and STL10 from ImageNet) but share overlapping semantic labels. In contrast, CIFAR10C represents a natural distribution shift that includes various image corruptions such as blur and noise, simulating real-world variations caused by illumination, weather, and camera artifacts. Both of these settings correspond to covariate shift [6], where the training and test samples differ in visual characteristics while maintaining identical label spaces. In our CIFAR10C experiments, we apply six different corruption types: Gaussian Noise, Shot Noise, Speckle Noise, Impulse Noise, Defocus Blur, and Gaussian Blur. In the main paper, we report the average clean and robust accuracies across all corruption types. To provide qualitative insight into these variations, Figure 5 presents sample images from CIFAR10C under different corruption settings, while Figures 7 and 8 illustrate cross-domain and distributional differences between SVHN vs. MNIST and CIFAR10 vs. STL10, respectively. As shown in Figure 8, although CIFAR10 and STL10 share the same label categories, their images differ in background, camera angle, and resolution. Similarly, MNIST samples differ significantly from SVHN,

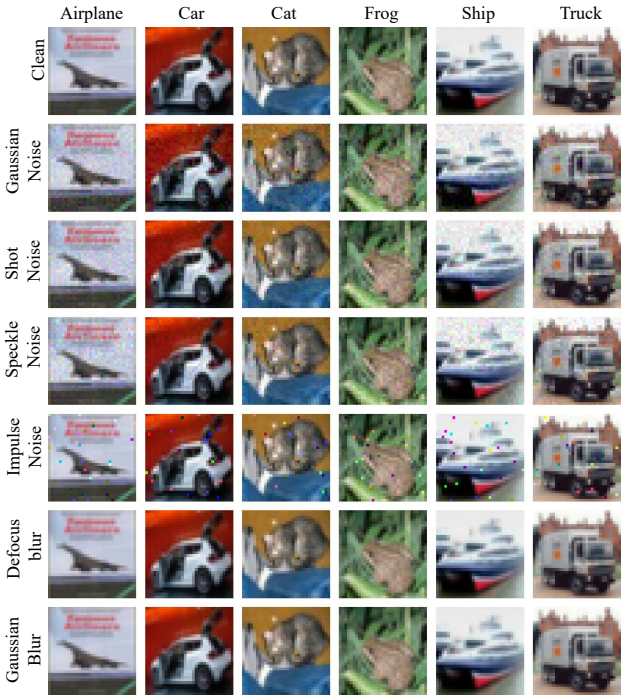
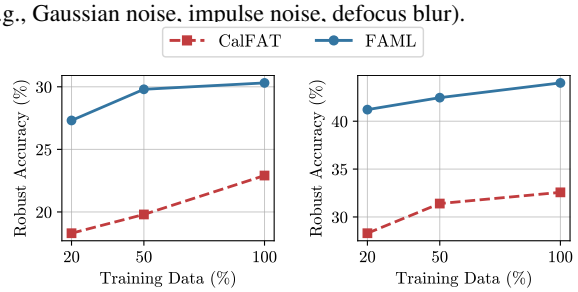


Figure 5. Samples from the CIFAR10C test set illustrating various corruption types. Each row corresponds to a corruption category (e.g., Gaussian noise, impulse noise, defocus blur).



(a) CIFAR10 (b) SVHN

Figure 6. Robust accuracy (under AutoAttack) of FAT methods on CIFAR10 and SVHN datasets under varying training data ratios ($\{20, 50, 100\}\%$).

as the former contains handwritten digits while the latter includes real-world street numbers.

C. Additional Results

C.1. Per-class performance of FAML

Adversarial training often amplifies disparities in performance across different classes [9], as the decision boundaries of AT-trained models tend to favor easier classes during classification. To further demonstrate the fairness and consistency of FAML, we compare its per-class average robust accuracy against the strongest baselines, CalFAT and FedGAIRAT, as shown in Figure 4. On CIFAR10 (Figure 4(left)), FAML consistently achieves higher robust accuracy across most classes compared to both baselines. Similarly, on SVHN (Figure 4(right)), FAML achieves the best per-class accuracy overall.

C.2. Robustness under varying perturbation strengths

Table 3 reports the robust accuracy of different FAT methods on CIFAR10 and SVHN under PGD-20 attacks with ℓ_∞ perturbation budgets of $\epsilon = {}^{8}/_{255}$, $\epsilon = {}^{16}/_{255}$, and $\epsilon = {}^{32}/_{255}$. A higher perturbation budget indicates a stronger and more challenging adversarial attack. As expected, the robust accuracy of all methods decreases as the perturbation strength increases. Notably, FAML consistently achieves the highest robustness across all budgets and datasets. On CIFAR10, FAML achieves 37.38% RA at $\epsilon = {}^{8}/_{255}$ and maintains 24.45% RA even under the strongest attack at $\epsilon = {}^{32}/_{255}$, outperforming the best baseline (CalFAT) by a notable margin. Similar trends are observed on SVHN, where FAML achieves RA of 45.92% at $\epsilon = {}^{8}/_{255}$ and 34.76% at $\epsilon = {}^{32}/_{255}$.

C.3. Robust performance under varying training data size for AutoAttack

To further validate the data efficiency of FAML, we extend our analysis to a stronger adversarial evaluation with AutoAttack across varying training data ratios $\{20, 50, 100\}\%$ on the CIFAR10 and SVHN datasets. Figure 6 presents a comparative analysis between FAML and CalFAT under varying training data ratios. Across both datasets, FAML consistently achieves significantly higher RA under AutoAttack, even when trained with only 20% of training data. On CIFAR10, FAML achieves robust accuracy above 27%, outperforming CalFAT by a large margin; CalFAT requires the full dataset (100%) to approach similar robustness. A similar trend is observed on SVHN, where FAML maintains robust accuracy above 40% and outperforms CalFAT. These results provide additional empirical validation of our main findings: FAML attains near-optimal robustness with substantially less training data, and its advantages become even more pronounced under strong adversarial attacks such as AutoAttack. Unlike CalFAT, which depends heavily on abundant data to achieve robustness, FAML demonstrates strong resilience and stable performance with limited data.

C.4. Additional results on scalability

In the main paper, we reported scalability results on SVHN and CIFAR10 across different numbers of participating clients, $K = \{10, 20, 50\}$ and measure both clean accuracy and robustness under PGD-20 and AutoAttack (AA). To complement those findings, Table 4 reports additional experiments on FMNIST and CIFAR100 datasets. On FMNIST, FAML consistently outperforms all baselines by a

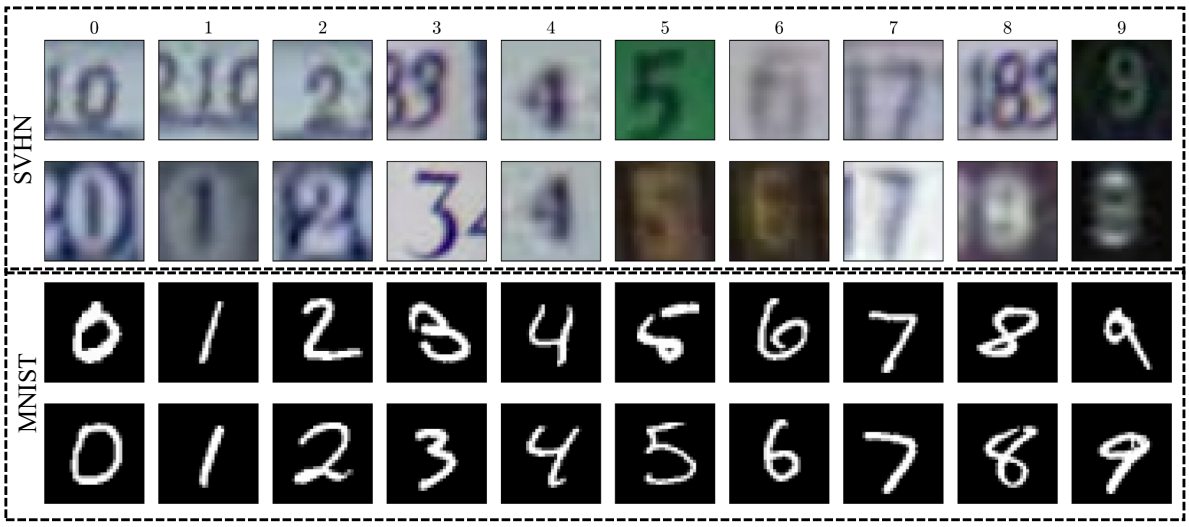


Figure 7. Examples of digit samples from the SVHN and MNIST datasets. SVHN contains real-world street-view digits with background clutter, while MNIST consists of handwritten digits on clean backgrounds, representing a cross-domain visual shift.

Figure 8. Visual comparison of image samples from the CIFAR10 and STL10 datasets. While both datasets contain several overlapping semantic categories, STL10 includes a monkey class that is absent from CIFAR10. The datasets also differ in image resolution and overall visual distribution due to variations in camera angle, lighting, and background context.

large margin across all client sizes. For $K = 10$, FAML achieves a clean accuracy of 83.45%, significantly surpassing the best baseline, CalFAT (74.78%). This strong performance extends to robust evaluations, where FAML achieves 66.86% (PGD) and 62.35% (AA), both being the highest among all methods. A similar trend is observed when the number of clients increases to $K = 20$ and $K = 50$, where FAML maintains superior clean accuracy of 82.32% and 85.84%, respectively, while simultaneously yielding the best PGD and AA robustness. On CIFAR100, a much more challenging dataset, the gap between FAML and other methods remains substantial. For $K = 10$, FAML attains the highest clean accuracy at 44.16%, along with robust accuracies of 16.55% (PGD) and 13.21% (AA), again outperforming all baselines. As the number of clients increases to $K = 20$ and $K = 50$, FAML maintains strong scalability and outperforms the best baseline by a large margin. Overall, when the number of clients increases, existing baselines struggle to maintain both clean and robust performance and their accuracy drops significantly, whereas FAML sustains high performance due to its strong generalization ability and fast adaptation capability.

Table 4. Clean and robust accuracies across different number of clients $K = \{10, 20, 50\}$ on FMNIST and CIFAR100 datasets. Robust accuracy is evaluated under PGD-20 and AA attacks. Best results are in bold.

Method	FMNIST									CIFAR100								
	$K=10$			$K=20$			$K=50$			$K=10$			$K=20$			$K=50$		
	Clean	PGD	AA	Clean	PGD	AA	Clean	PGD	AA	Clean	PGD	AA	Clean	PGD	AA	Clean	PGD	AA
NaiveFAT	54.35	37.28	26.82	52.37	34.98	25.81	48.23	32.20	23.76	17.16	11.77	8.47	15.02	10.03	7.05	13.46	9.12	6.73
FedPGD	55.10	34.11	26.67	53.09	32.51	24.54	48.90	29.99	22.65	21.25	13.16	10.29	20.05	8.28	8.63	17.25	8.10	7.78
FedTRADES	56.92	35.77	28.39	54.85	38.89	30.76	50.52	35.58	21.33	18.86	11.85	9.41	13.75	9.75	7.60	13.51	9.57	7.42
FedMART	57.23	42.59	33.27	55.15	36.91	28.71	50.80	33.96	26.26	28.77	10.41	9.72	28.12	9.82	8.53	27.05	8.14	7.12
FedGAIRAT	55.40	41.42	33.55	53.38	37.60	29.69	49.17	34.47	27.58	14.12	10.56	8.55	12.94	9.12	7.33	11.96	8.81	7.05
FedRBN	54.26	43.26	31.99	52.28	38.11	29.07	48.16	33.39	26.30	22.56	11.98	10.32	20.78	9.15	7.46	19.51	8.84	7.39
GEAR	59.02	28.71	18.81	56.87	25.81	17.28	52.38	21.52	13.53	24.71	12.02	7.88	22.50	10.22	6.70	22.10	9.85	6.20
CalFAT	74.78	45.93	43.80	54.78	37.51	34.80	51.83	35.41	27.52	37.35	13.04	10.30	22.64	10.52	8.89	18.48	9.37	7.70
FAML	**83.45**	**66.86**	**62.35**	**82.32**	**66.42**	**61.86**	**85.84**	**67.68**	**61.25**	**44.16**	**16.55**	**13.21**	**42.11**	**15.32**	**12.10**	**41.16**	**15.22**	**12.55**

Table 5. Clean and robust accuracies under various adversarial attacks (FGSM, BIM, CW, PGD, and AA) across different values of $\beta \in \{0.05, 0.3, 10\}$ on CIFAR10 and SVHN datasets. The lowest β represents a highly heterogeneous client scenario, while the highest corresponds to an IID setting. Best results are highlighted in **bold**.

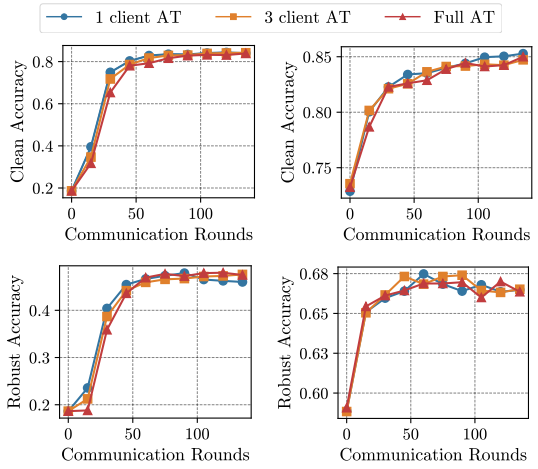
Method	CIFAR10																	
	$\beta = 0.05$						$\beta = 0.3$						$\beta = 10$ **(IID)**					
	Clean	FGSM	BIM	CW	PGD	AA	Clean	FGSM	BIM	CW	PGD	AA	Clean	FGSM	BIM	CW	PGD	AA
NaiveFAT	49.10	27.49	25.32	22.17	25.24	22.51	58.93	31.68	28.17	24.96	28.00	24.34	79.62	33.68	30.87	27.15	37.57	26.31
FedPGD	47.13	26.63	24.96	20.75	25.03	21.28	56.12	30.86	28.46	25.07	28.29	23.64	75.89	34.51	31.45	28.52	42.16	30.23
FedTRADES	40.24	26.02	25.06	22.48	24.99	20.16	54.26	30.83	29.39	24.74	29.26	23.87	74.29	35.13	32.14	29.77	44.35	31.12
FedMART	29.84	21.90	21.39	18.31	21.41	17.89	40.96	28.32	27.88	23.12	27.80	22.16	72.33	34.18	31.80	28.43	43.11	30.55
FedGAIRAT	50.41	28.89	26.30	22.66	26.34	23.81	60.63	33.31	30.12	25.50	29.67	24.75	73.43	35.29	32.00	29.33	43.49	31.12
FedRBN	39.35	25.92	24.40	21.55	24.77	19.47	53.54	29.88	28.76	24.11	28.63	23.14	72.44	34.11	31.37	28.71	41.67	29.95
CalFAT	61.00	32.40	29.75	23.55	29.50	25.66	69.95	34.25	30.80	27.76	30.96	26.84	74.23	37.68	33.27	30.88	**44.68**	30.34
FAML	**62.25**	**37.37**	**44.95**	**28.59**	**35.27**	**26.52**	**70.59**	**41.39**	**48.93**	**30.66**	**38.87**	**27.61**	**77.78**	**44.39**	**50.31**	**33.91**	44.60	**32.49**

Method	SVHN																	
	$\beta = 0.05$						$\beta = 0.3$						$\beta = 10$ **(IID)**					
	Clean	FGSM	BIM	CW	PGD	AA	Clean	FGSM	BIM	CW	PGD	AA	Clean	FGSM	BIM	CW	PGD	AA
NaiveFAT	18.80	18.90	18.90	18.90	19.00	14.20	20.70	20.80	20.80	20.80	21.00	15.60	22.50	22.90	23.00	22.90	23.10	16.70
FedPGD	18.70	18.50	18.60	18.90	18.70	13.10	20.40	20.60	20.70	21.00	21.10	15.00	22.20	22.60	22.70	22.80	22.90	16.50
FedMART	19.00	19.10	18.90	19.10	19.00	13.10	21.80	22.10	22.00	22.10	22.00	15.30	23.60	24.00	24.10	24.00	24.10	16.80
FedTRADES	54.80	35.40	33.70	30.00	35.00	30.30	59.30	39.10	37.20	32.80	38.80	33.60	63.80	41.90	40.30	34.80	41.70	35.70
FedGAIRAT	56.00	35.00	34.00	30.30	35.00	30.20	60.50	38.50	36.80	32.70	38.50	33.40	65.40	42.00	40.50	35.10	41.90	36.10
FedRBN	55.40	33.40	31.30	27.00	31.80	27.10	59.60	36.30	34.40	29.80	35.70	30.60	64.00	38.90	37.00	32.50	39.00	33.50
CalFAT	77.50	46.50	40.20	30.50	40.00	31.40	83.10	50.90	44.50	34.10	43.90	34.90	90.33	54.20	47.10	36.90	47.60	42.00
FAML	**81.00**	**52.00**	**63.00**	**35.50**	**44.90**	**40.00**	**87.10**	**58.00**	**68.90**	**39.80**	**50.40**	**44.80**	**91.20**	**60.60**	**72.90**	**42.90**	**54.10**	**48.10**

C.5. Robustness under varying label skewness

In the main paper, we reported the performance of FAT methods on CIFAR10 and SVHN under Dirichlet distributions $\beta = 0.05$ and $\beta = 10$, evaluated under BIM, PGD, CW and AA attacks. In Table 5, we extend this analysis by incorporating an additional Dirichlet distribution parameter, $\beta = 0.3$, which represents moderate distribution skew. We also include experimental results under the FGSM attack to provide a more comprehensive evaluation of robustness. This setup enables a thorough analysis of the robust-

ness and generalization capability of each method under realistic FL scenarios. Across all datasets and heterogeneity levels, the FAML framework consistently outperforms existing baselines. Under the most challenging non-IID configuration ($\beta = 0.05$), FAML achieves the highest clean accuracy and demonstrates significantly stronger resilience under AutoAttack. On CIFAR10, FAML surpasses FAT methods such as FedPGD, FedTRADES, and FedGAIRAT by a considerable margin, reflecting its ability to effectively learn robust representations even when client data distributions are highly skewed. When the heterogeneity is moderately reduced ($\beta = 0.3$), FAML continues to maintain superior performance, again offering the best clean and robust accuracy on both CIFAR10 and SVHN. In the IID setting ($\beta = 10$), all methods experience a performance improvement. Notably, FAML achieves the highest robustness under all adversarial attacks. These results demonstrate that FAML remains consistently effective across different levels of client heterogeneity.



(a) SVHN (b) FMNIST

Figure 9. Comparison of robust (PGD-20) and clean accuracy on SVHN and FMNIST with varying numbers of AT-performing clients in FAML.

C.6. Robust performance with single AT performing client on SVHN and FMNIST

Figure 9 presents the clean and robust (under PGD-20) performance of varying the number of AT-performing clients on SVHN and FMNIST. Similar to our ablations on CIFAR10 and CIFAR100 in the main paper, we observe that on both datasets, FAML achieves robust accuracy comparable full AT setting, even when only one client performs AT. Notably, on FMNIST, the single AT configuration results in higher clean accuracy, indicating that the global model benefits from learning a more generalized decision boundary guided predominantly by clean clients. These results further validates our claim that adversarial training on a single client is sufficient for achieving strong robustness in FAML while preserving desirable clean performance.

References

[1] Chen Chen, Yuchen Liu, Xingjun Ma, and Lingjuan Lyu. Calfat: Calibrated federated adversarial training with label skewness. *Advances in neural information processing systems*, 35:3569–3581, 2022. 2, 3

[2] Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. Federated robustness propagation: Sharing robustness in heterogeneous federated learning. *arXiv preprint arXiv:2106.10196*, 2021. 3

[3] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 3

[5] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, page 7. Granada, 2011. 1

[6] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551, 2020. 3

[7] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*, 2019. 3

[8] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 1

[9] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning*, pages 11492–11501. PMLR, 2021. 4

[10] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019. 3

[11] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. *arXiv preprint arXiv:2010.01736*, 2020. 3

[12] Jie Zhang, Bo Li, Chen Chen, Lingjuan Lyu, Shuang Wu, Shouhong Ding, and Chao Wu. Delving into the adversarial robustness of federated learning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11245–11253, 2023. 2

[13] Giulio Zizzo, Ambrish Rawat, Mathieu Sinn, and Beat Buesser. Fat: Federated adversarial training. *arXiv preprint arXiv:2012.01791*, 2020. 3