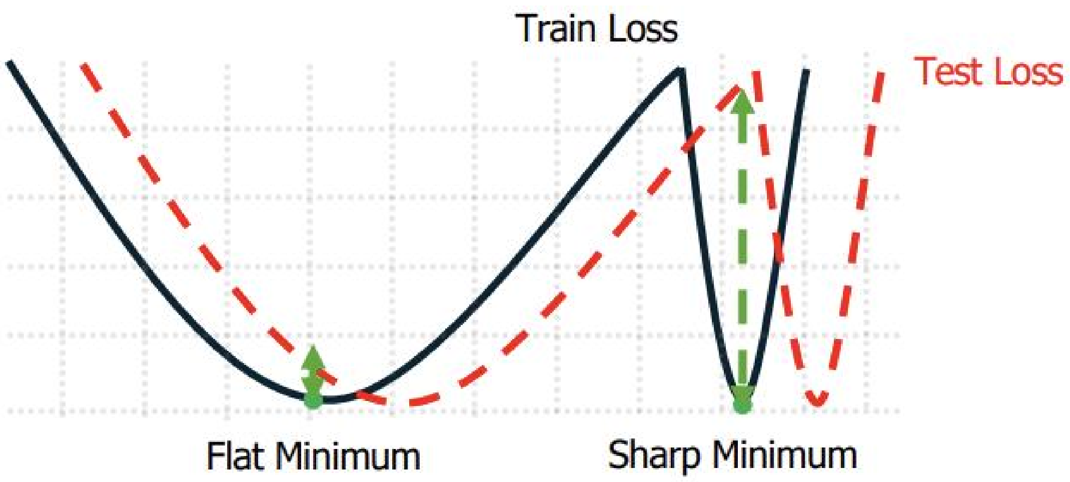
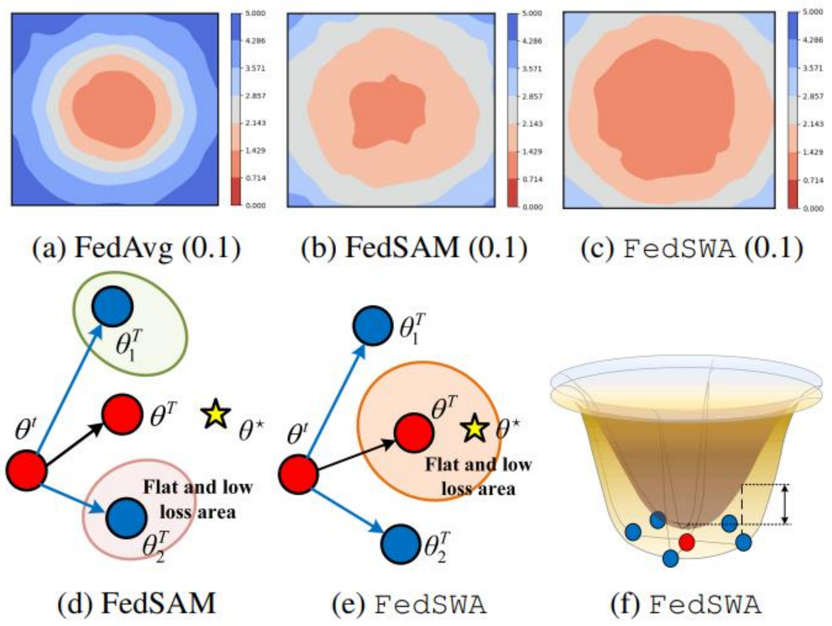
Improving Generalization in Federated Learning with Highly Heterogeneous Data via Momentum-Based Stochastic Controlled Weight Averaging (ICML'2025)

In order to improve the generalization ability of federated learning in data heterogeneity scenarios, the FedSWA algorithm is proposed based on the SWA optimizer. Compared with the FedSAM proposed by ICML2023, our algorithm is more inclined to find the global flat minimum value.

研究工作的创新性构思

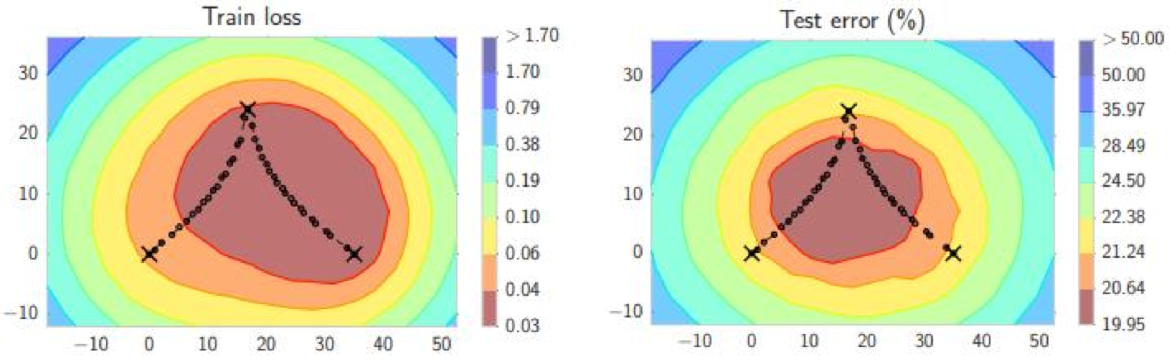


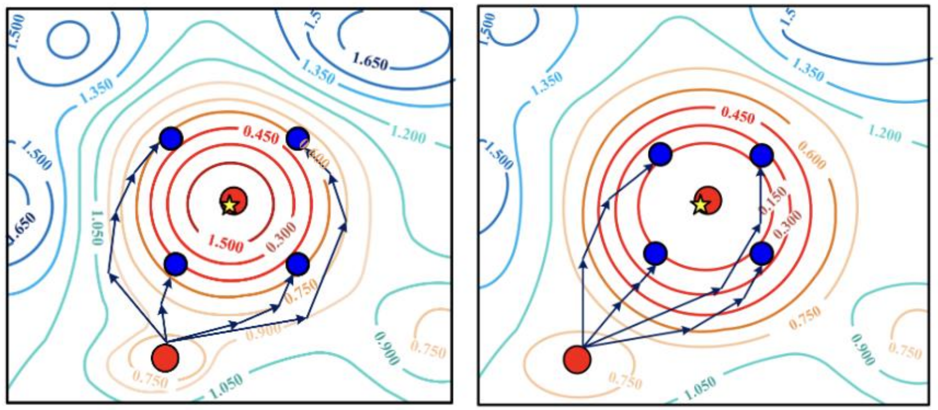
Flat minimum value has better generalization ability



The FedSWA algorithm is superior to FedSAM and flatter

SWA is inspired by a practical observation: at the end of each learning rate cycle, the model often reaches a local minimum near the edge of the loss surface, where the loss is relatively low. By averaging several of these points, we can likely obtain a more general and even lower-loss solution.





(a) FedSWA (b) FedMoSWA

Algorithm 1 Stochastic Weight Averaging

Require:
weights $\hat{w}$, LR bounds α_1, α_2,
cycle length c (for constant learning rate $c = 1$), number of iterations n

Ensure: w_{SWA}

$w \leftarrow \hat{w}$ {Initialize weights with $\hat{w}$}

$w_{\text{SWA}} \leftarrow w$

for $i \leftarrow 1, 2, \ldots, n$ **do**

$\alpha \leftarrow \alpha(i)$ {Calculate LR for the iteration}

$w \leftarrow w - \alpha \nabla \mathcal{L}_i(w)$ {Stochastic gradient update}

if $\mathrm{mod}(i, c) = 0$ **then**

$n_{\text{models}} \leftarrow i/c$ {Number of models}

$w_{\text{SWA}} \leftarrow \frac{w_{\text{SWA}} \cdot n_{\text{models}} + w}{n_{\text{models}} + 1}$ {Update average}

end if

end for

{Compute BatchNorm statistics for w_{SWA} weights}

Theoretical Results: We analyze the generalization ability and convergence rate of the FedSAM algorithm from ICML and our proposed FedSWA algorithm using uniform stability theory. The results show that the proposed FedSWA algorithm outperforms FedSAM.

Algorithm	Generalization error	Optimization error
FedSAM (Qu et al., 2022)	$\mathcal{O}\left(\frac{L}{mn\beta}e^{1+\frac{1}{T}}(\bar{c}L+\bar{c}\sigma_g+\bar{c}\sigma)\right)$	$\mathcal{O}\left(\frac{\beta F}{\sqrt{TKs}}+\frac{\sqrt{K}{\sigma_g}^2}{\sqrt{Ts}}+\frac{L^2\sigma^2}{T^{3/2}K}+\frac{L^2}{T^2}\right)$
MoFedSAM (Qu et al., 2022)	$\mathcal{O}\left(\frac{L}{mn\beta}e^{1+\frac{1}{T}}(\bar{c}L+\bar{c}\sigma_g+\bar{c}\sigma)\right)$	$\mathcal{O}\left(\frac{\beta LF}{\sqrt{TKs}}+\frac{\beta\sqrt{K}{\sigma_g}^2}{\sqrt{Ts}}+\frac{L^2\sigma^2}{T^{3/2}K}+\frac{\sqrt{K}L^2}{T^{3/2}\sqrt{s}}\right)$
FedSWA (ours)	$\mathcal{O}\left(\frac{L}{mn\beta}e^{1+\frac{1}{T}}(\tilde{c}L+\tilde{c}\sigma_g+\tilde{c}\sigma)\right)$	$\mathcal{O}\left(\frac{\beta(\sigma+\sqrt{K}\sigma_g)\sqrt{F}}{\sqrt{TKs}}+\frac{F^{2/3}(\beta{\sigma_g}^2)^{1/3}}{T^{2/3}}+\frac{\beta F}{T}\right)$
FedMoSWA (ours)	$\mathcal{O}\left(\frac{L}{mn\beta}e^{1+\frac{1}{T}}(\tilde{c}L+\sigma_g+\tilde{c}\sigma)\right)$	$\mathcal{O}\left(\frac{\sigma\sqrt{F}}{\sqrt{TKs}}\left(\sqrt{1+\frac{s}{\alpha^2}}\right)+\frac{\beta F}{T}\left(\frac{m}{s}\right)^{\frac{2}{3}}\right)$

Theorem 5.1 (Generalization Error). *Assuming all clients participate in each round with Option I:*

***Strongly convex**: Under Assumptions 1-5, suppose loss $\ell(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})$ is μ-strongly convex. By setting $\eta_k^t \le \frac{1}{\beta KT}$, $\tilde{b} = 1+\left(\frac{\mu}{(\beta+\mu)K}\right)^{K-1}\frac{1}{T}$, the generalization error satisfies:*

$$\texttt{FedSWA}:\ \varepsilon_{gen} \le \frac{2L}{mn\beta}e^{1-\frac{\mu}{(\beta+\mu)T}}\left(\tilde{b}L+\tilde{b}\sigma_g+\tilde{b}\sigma\right)$$

$$\texttt{FedMoSWA}:\ \varepsilon_{gen} \le \frac{2L}{mn\beta}e^{1-\frac{\mu}{(\beta+\mu)T}}\left(\tilde{b}L+\sigma_g+\tilde{b}\sigma\right)$$

***Non-convex**: Under Assumptions 2-5, assume $\ell(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})$ is β-smooth. Together with $\eta_k^t \le \frac{1}{\beta KT}$, $\tilde{c} = 1+\left(2+\frac{1}{KT}\right)^{K-1}\frac{1}{T}$, the generalization error satisfies:*

$$\texttt{FedSWA}:\ \varepsilon_{gen} \le \frac{2L}{mn\beta}e^{\frac{1}{T}+1}\left(\tilde{c}L+\tilde{c}\sigma_g+\tilde{c}\sigma\right).$$

$$\texttt{FedMoSWA}:\ \varepsilon_{gen} \le \frac{2L}{mn\beta}e^{\frac{1}{T}+1}\left(\tilde{c}L+\sigma_g+\tilde{c}\sigma\right).$$

Theorem 5.2 (Optimization Error of FedMoSWA). *For β-smooth functions $\{F_i\}$, which satisfy Assumptions 6-9, and are the same as in the SCAFFOLD (Karimireddy et al., 2020) algorithm (see the Appendix for details), the output of FedMoSWA has expected error smaller than ϵ.*

***Strongly convex**: $\eta_k^t \le \min\left(\frac{1}{\beta K\alpha},\frac{s}{\mu mK\alpha}\right), T \ge \max\left(\frac{\beta}{\mu},\frac{m}{s}\right)$ then*

$$\mathcal{O}\left(\frac{\sigma^2}{\mu TKs}\left(1+\frac{s}{\alpha^2}\right)+\frac{m\mu}{s}D^2\exp\left(-\left\{\frac{s}{m}+\frac{\mu}{\beta}\right\}T\right)\right)$$

***Non-convex**: $\eta_k^t \le \frac{1}{K\alpha\beta}\left(\frac{s}{m}\right)^{\frac{2}{3}}$, and $T \ge 1$, then*

$$\mathcal{O}\left(\frac{\sigma\sqrt{F}}{\sqrt{TKs}}\left(\sqrt{1+\frac{m}{\alpha^2}}\right)+\frac{\beta F}{T}\left(\frac{m}{s}\right)^{\frac{2}{3}}\right).$$

Here $D^2 := \left\|\boldsymbol{\theta}^0-\boldsymbol{\theta}^\star\right\|^2+\frac{1}{2m\beta^2}\sum_{i=1}^m\left\|\boldsymbol{c}_i^0-\nabla F_i(\boldsymbol{\theta}^\star)\right\|^2$, $F := F(\boldsymbol{\theta}^0)-F(\boldsymbol{\theta}^\star)$.

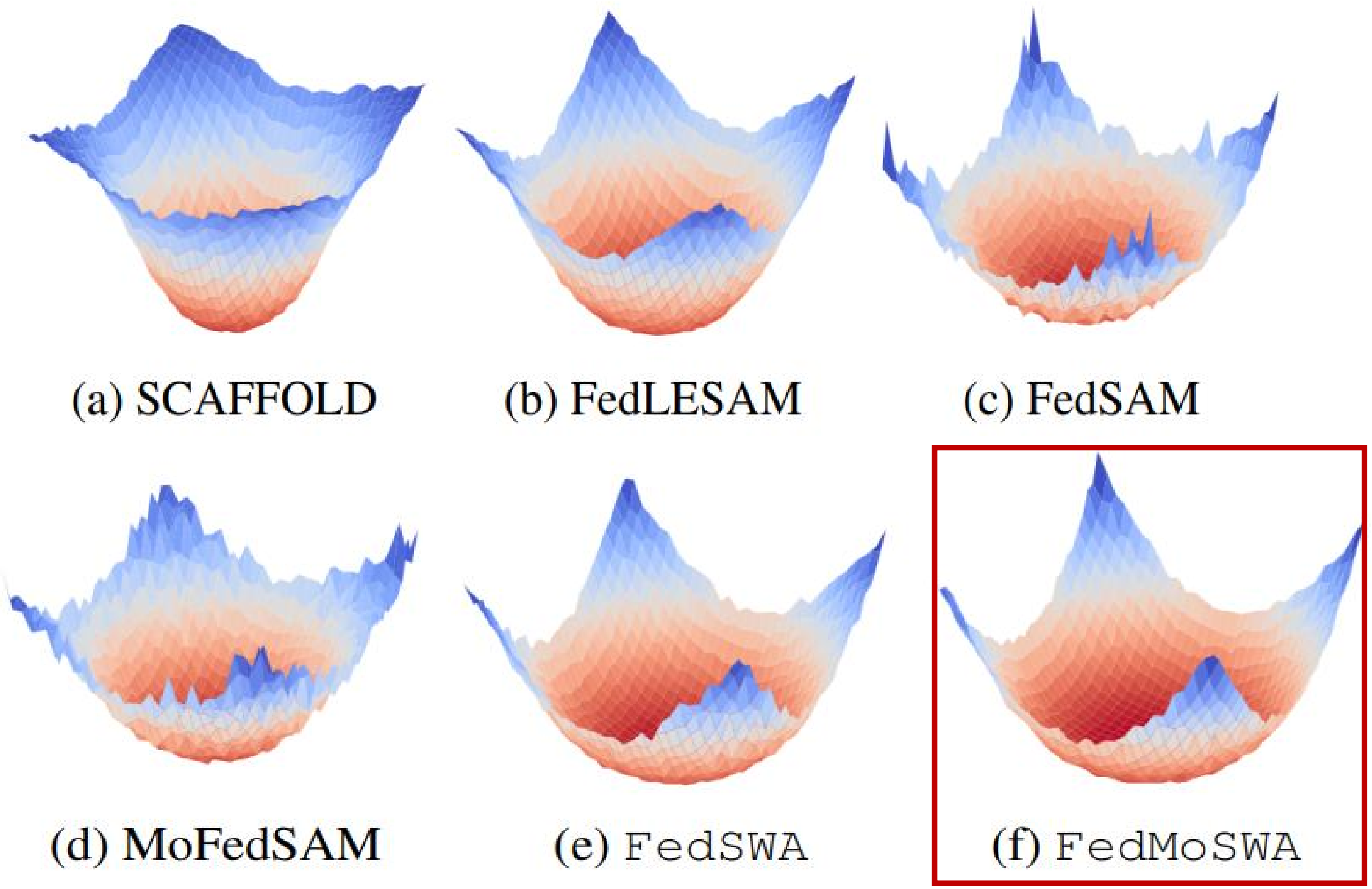
Algorithm 1 FedSWA, FedMoSWA algorithm.

Input: λ, ρ, initial server model $\boldsymbol{\theta}_0$, number of clients N, number of communication rounds T, number of local iterations K, local learning rate η_l.
for $t = 0, ..., T$ **do**
  Communicate $(\boldsymbol{\theta}_{t-1})$ to selected clients $i \in [s]$.
  Communicate $(\boldsymbol{\theta}_{t-1}, \boldsymbol{m})$ to selected clients $i \in [s]$.
  for $i = 1, \dots, s$ clients in parallel **do**
    for $k = 0, \dots, K$ local update **do**
      Compute mini-batch gradient $g_i\left(\boldsymbol{\theta}_{i,k}^t\right)$.
      $\eta_k^t = \eta_l\left(1-\frac{k}{K}\right)+\frac{k}{K}\rho\eta_l$.
      $\boldsymbol{\theta}_{i,k+1}^t \leftarrow \boldsymbol{\theta}_{i,k}^{(t)} - \eta_k^t\left(g_i\left(\boldsymbol{\theta}_{i,k}^t\right)\right)$.
      $\boldsymbol{\theta}_{i,k+1}^t \leftarrow \boldsymbol{\theta}_{i,k}^{(t)} - \eta_k^t\left(g_i\left(\boldsymbol{\theta}_{i,k}^t\right) - \boldsymbol{c_i} + \boldsymbol{m}\right)$.
    end for
    Communicate $\left(\boldsymbol{\theta}_{i,K}^t\right)$ to server.
    $\boldsymbol{c}_i^+ \leftarrow$ (i) $g_i(\boldsymbol{x})$ or (ii) $\boldsymbol{c}_i - \boldsymbol{m} + \frac{1}{\sum_k \eta_k^t}\left(\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_{i,k}^t\right)$.
    Communicate $\left(\boldsymbol{\theta}_{i,K}^t, \boldsymbol{c}_i^+ - \boldsymbol{m}\right)$ to server, $\boldsymbol{c_i} \leftarrow \boldsymbol{c_i}^+$.
  end for
  $\boldsymbol{m} \leftarrow \boldsymbol{m} + \gamma\frac{1}{s}\sum_{i\in[s]}\Delta\boldsymbol{c}_i$.
  $\boldsymbol{v}_t = \frac{1}{s}\sum_{i=1}^s \boldsymbol{\theta}_{i,K}^t$, $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \alpha\left(\boldsymbol{v}_t - \boldsymbol{\theta}_{t-1}\right)$.
end for

Improved accuracy by 6.3% on the CIFAR100 dataset. Improved accuracy by 1.8% on the ImageNet dataset.

Table 3: Comparison of each algorithm on the CIFAR100 and Tiny ImageNet datasets with different data heterogeneity.

Method	CIFAR100 (ResNet-18)						Tiny ImageNet (ViT-Base)					
	Dirichlet-0.1		Dirichlet-0.3		Dirichlet-0.6		Dirichlet-0.1		Dirichlet-0.3		Dirichlet-0.6	
	Acc.(%) 1000R	Rounds 55%	Acc.(%) 1000R	Rounds 55%	Acc.(%) 1000R	Rounds 55%	Acc.(%) 400R	Rounds 70%	Acc.(%) 400R	Rounds 70%	Acc.(%) 400R	Rounds 70%
FedAvg	$45.8_{\pm 0.3}$	1000+	$52.5_{\pm 0.3}$	1000+	$54.2_{\pm 0.2}$	1000+	$70.9_{\pm 0.1}$	258	$71.8_{\pm 0.1}$	223	$72.8_{\pm 0.1}$	208
FedDyn	$45.8_{\pm 0.2}$	1000+	$45.9_{\pm 0.3}$	1000+	$46.5_{\pm 0.2}$	1000+	$67.5_{\pm 0.3}$	400+	$68.2_{\pm 0.3}$	400+	$69.3_{\pm 0.3}$	400+
SCAFFOLD	$44.3_{\pm 0.3}$	1000+	$50.3_{\pm 0.3}$	1000+	$52.3_{\pm 0.2}$	1000+	$71.6_{\pm 0.1}$	202	$72.5_{\pm 0.1}$	192	$73.1_{\pm 0.2}$	169
FedSAM	$40.1_{\pm 0.4}$	1000+	$49.0_{\pm 0.3}$	1000+	$51.9_{\pm 0.5}$	1000+	$71.4_{\pm 0.2}$	212	$72.2_{\pm 0.2}$	194	$72.9_{\pm 0.4}$	180
MoFedSAM	$51.5_{\pm 0.2}$	1000+	$57.5_{\pm 0.2}$	770	$60.1_{\pm 0.1}$	603	$71.6_{\pm 0.4}$	229	$72.4_{\pm 0.3}$	214	$72.5_{\pm 0.4}$	209
FedLESAM	$48.7_{\pm 0.2}$	1000+	$53.3_{\pm 0.4}$	1000+	$52.1_{\pm 0.1}$	1000+	$71.9_{\pm 0.3}$	210	$72.1_{\pm 0.2}$	188	$72.5_{\pm 0.3}$	182
FedASAM	$47.7_{\pm 0.3}$	1000+	$46.6_{\pm 0.2}$	1000+	$49.8_{\pm 0.1}$	1000+	$69.2_{\pm 0.3}$	400+	$71.3_{\pm 0.2}$	234	$72.1_{\pm 0.3}$	196
FedACG	$52.2_{\pm 0.4}$	1000+	$57.7_{\pm 0.2}$	717	$61.7_{\pm 0.4}$	518	$66.2_{\pm 0.2}$	400+	$68.5_{\pm 0.1}$	400+	$70.2_{\pm 0.3}$	386
FedSWA	$50.3_{\pm 0.3}$	1000+	$55.5_{\pm 0.4}$	889	$59.8_{\pm 0.3}$	574	$71.9_{\pm 0.3}$	199	$72.6_{\pm 0.2}$	179	$73.2_{\pm 0.2}$	168
FedMoSWA	**$61.9_{\pm 0.5}$**	**577**	**$66.2_{\pm 0.4}$**	**468**	**$67.9_{\pm 0.4}$**	**330**	**$73.8_{\pm 0.3}$**	**161**	**$74.4_{\pm 0.3}$**	**152**	**$74.7_{\pm 0.1}$**	**144**



Through the visualization of model loss landscapes, it can also be observed that our algorithm has flatter loss landscapes and lower loss values.

- *Junkang Liu*, *Fanhua Shang, Yuanyuan Liu, Hongying Liu.* **Improving Generalization in Federated Learning with Heterogeneous Data via Momentum-Based Stochastic Controlled Weight Averaging.** *ICML 2025*.