Handling Spatial-Temporal Data Heterogeneity for Federated Continual Learning via Tail Anchor

Hao Yu
Southwestern University of Finance and Economics
yuhao2033@163.com

Xin Yang*
Southwestern University of Finance and Economics
yangxin@swufe.edu.cn

Le Zhang
University of Electronic Science and Technology of China
lezhang@uestc.edu.cn

Hanlin Gu
WeBank
allengu@webank.com

Tianrui Li
Southwest Jiaotong University
trli@swjtu.edu.cn

Lixin Fan
WeBank
lixinfan@webank.com

Qiang Yang
Hong Kong University of Science and Technology
qyang@cse.ust.hk

CVPR 2025

Background

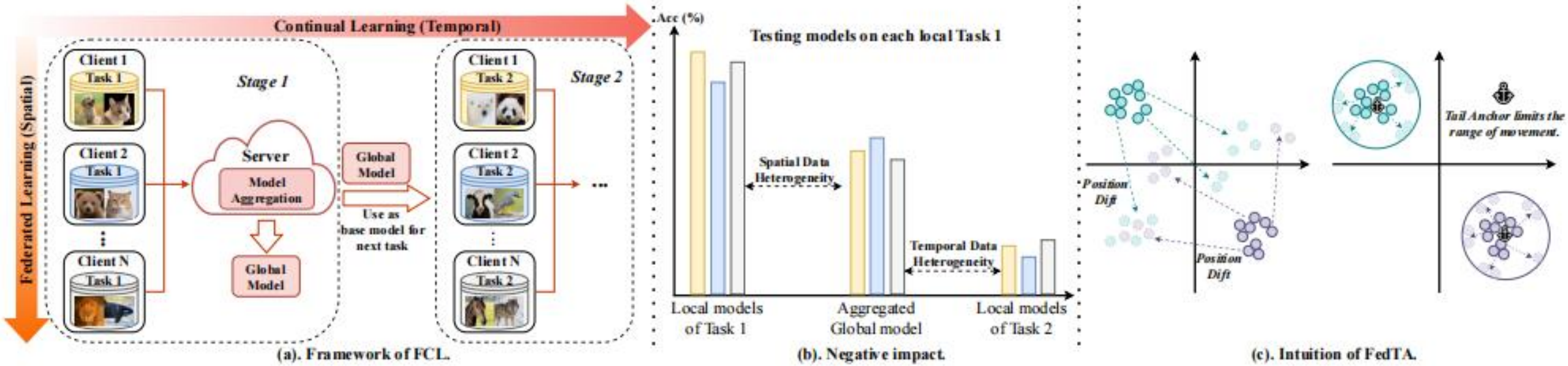


Figure 1: Illustration of FCL, the negative impact of spatial-temporal data heterogeneity and the intuition of Tail Anchor.

Traditional FL：training data of all clients will remain unchanged（unrealistic static assumption）
Federated Continual Learning (FCL)：allowing clients to continually accumulate knowledge from task sequences
Challenge：*spatial-temporal data heterogeneity*

空间异质性 (Spatial)	不同客户端数据分布不同（Non-IID）	各客户端数据分布不均，私有类+公共类混合
时间异质性 (Temporal)	同一客户端不同任务数据分布不同	每个客户端有任务序列 $\mathcal{T}_i = \{T_i^1, T_i^2, \dots\}$

Negative Impact

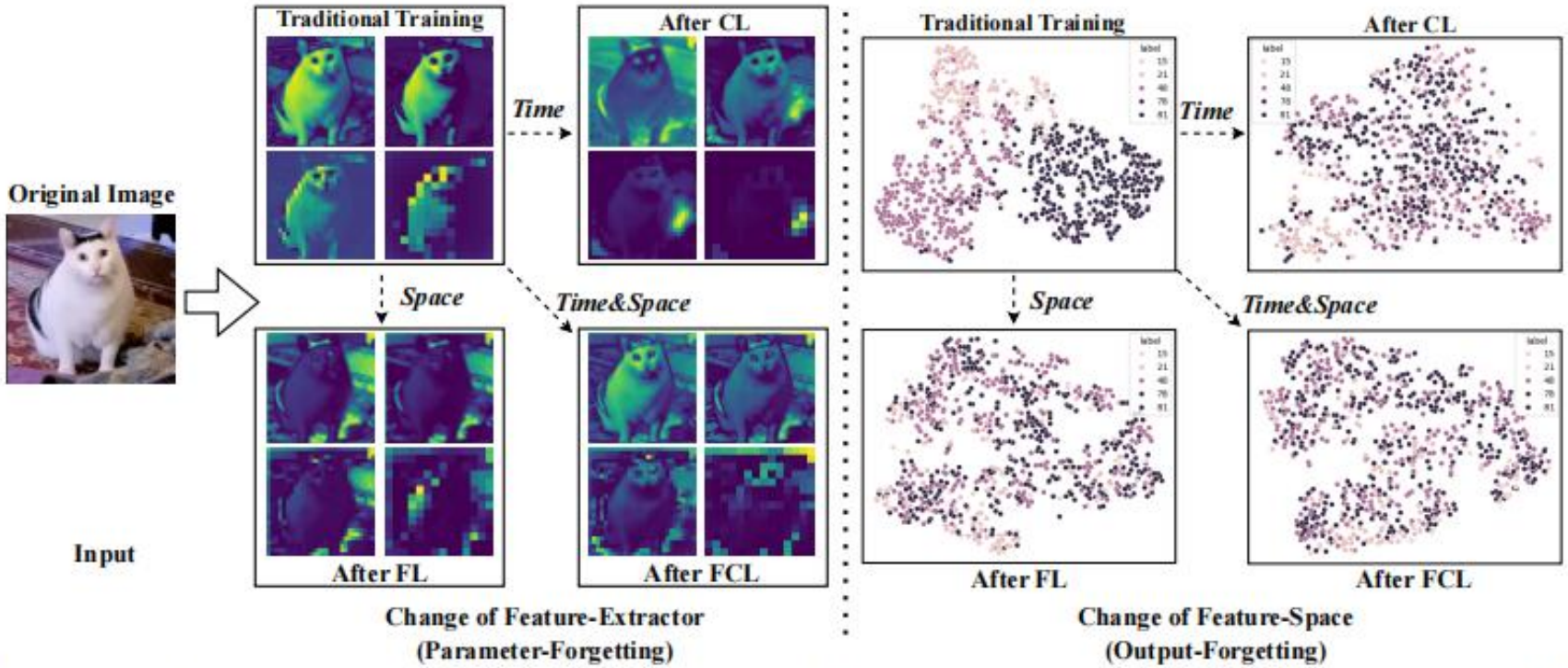
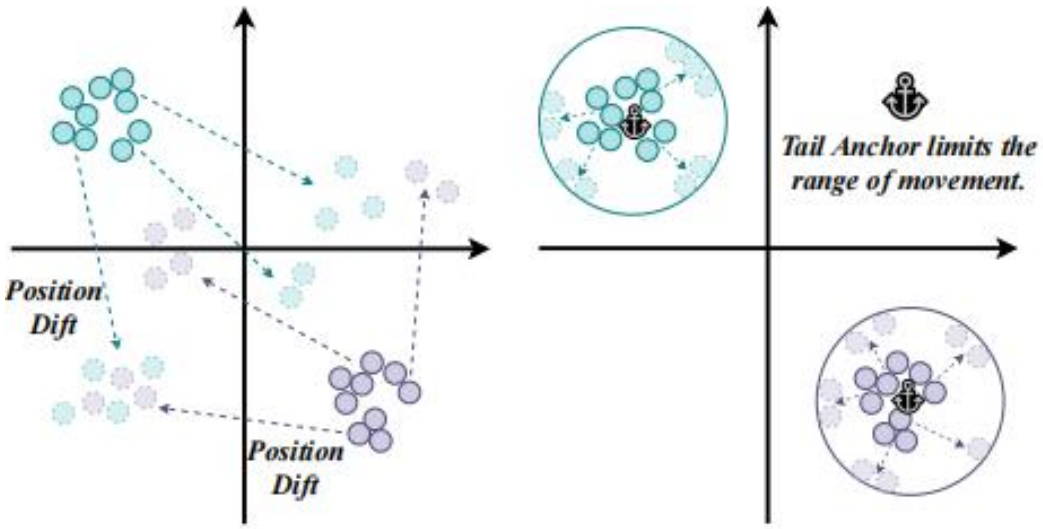


Figure 2: Illustrations of the negative impact of spatial-temporal data heterogeneity on the feature extractor and feature space. [Left Side] illustrates the variation of significant features extracted by the feature extractor for the same input sample, where brighter colors indicate more important features. As spatial-temporal changes occur, the extracted features gradually shift away from "cat", even extracting features near the image edges. [Right Side] depicts the changes in the positions of the features in the feature space after undergoing spatial-temporal transformations for the same batch.

参数遗忘 (Parameter-Forgetting)：特征提取器对同一输入提取的特征内容发生改变，不再提取与类别相关的语义特征（特征提取器参数被新任务数据覆盖）

输出遗忘 (Output-Forgetting)：特征在空间中的位置发生偏移，边界模糊（分类头持续更新导致特征位置漂移）

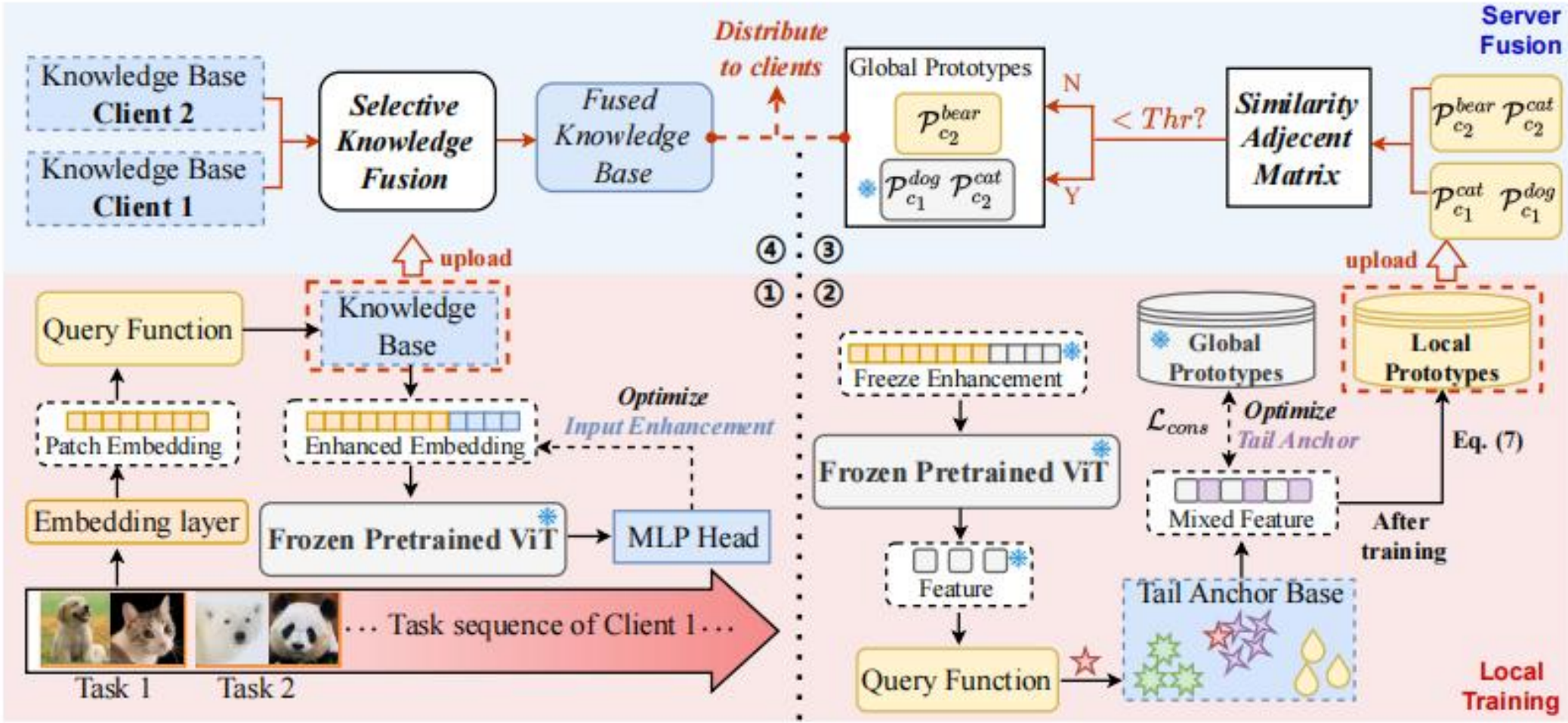
Motivation



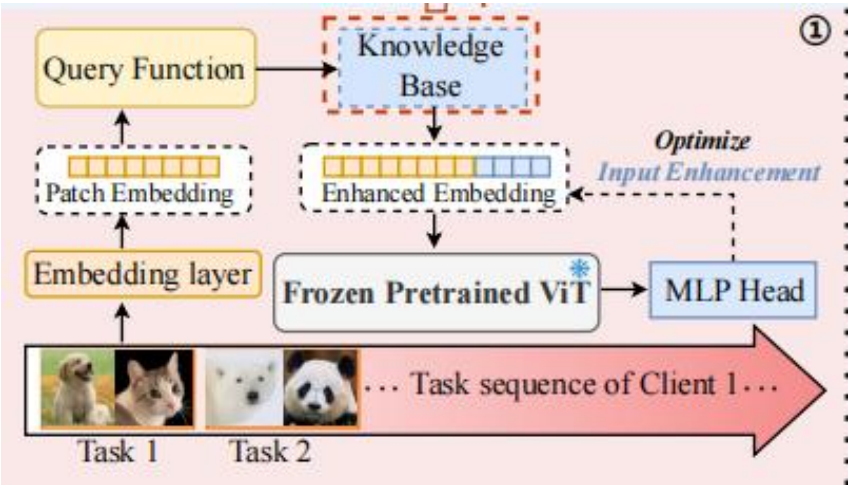
(c). Intuition of FedTA.

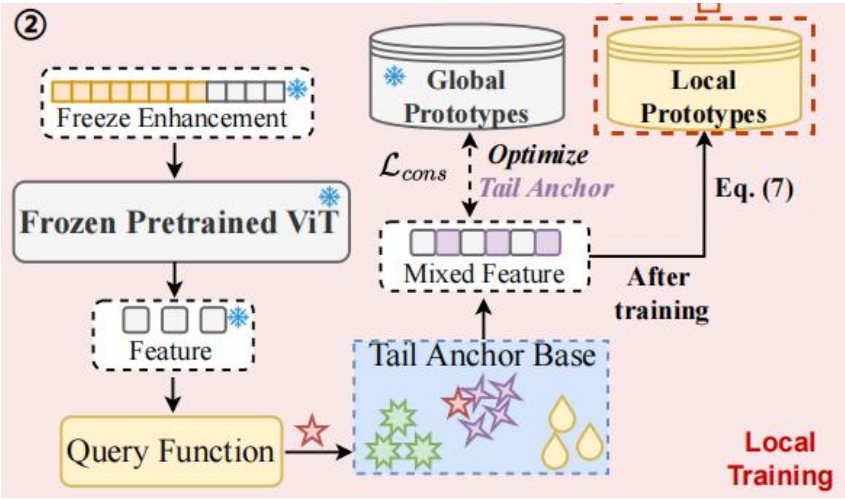
- **Feature Consistency:** Feature extractor parameters change across time and clients, causing parameter-forgetting. -> Freeze the pre-trained ViT-B/16 feature extractor shared by all clients.
- **Position Fixing:** Classification head continuously fits new tasks, causing output features to gradually drift in space. -> Learnable **Tail Anchors** mix with frozen features, aligned with server-selected and fixed global prototypes.
- **Common Future Space:** data distribution differences across clients. -> server establishes unique global prototypes via Best Global Prototype Selection(BGPS), contrastive loss forces all client features toward the same prototypes.

Method


Server Fusion
Knowledge Base Client 2
Knowledge Base Client 1
Selective Knowledge Fusion
Fused Knowledge Base
Distribute to clients
Global Prototypes
N
< Thr?
Y
Similarity Adjecent Matrix
upload
upload
Query Function
Knowledge Base
Patch Embedding
Enhanced Embedding
Optimize Input Enhancement
Embedding layer
Frozen Pretrained ViT
MLP Head
Task 1
Task 2
... Task sequence of Client 1 ...
Freeze Enhancement
Frozen Pretrained ViT
Feature
Query Function
Global Prototypes
Local Prototypes
Optimize Tail Anchor
Mixed Feature
Tail Anchor Base
Eq. (7)
After training
Local Training

Method





Input Enhancement

$$\mathcal{KB}_i = \{IE_1, IE_2, \ldots, IE_M\},$$

$$E' = [IE_{s_1}, \ldots, IE_{s_N}; E], 1 \leq N \leq M,$$

$$\min \mathcal{L}(\mathcal{V}_e^i(E'), y) + \lambda_1 \sum_{\mathcal{K}_s^{ie}} \mathrm{dis}(K_{in}^{ie}, Ks_i{}^{ie}),$$

Optimization tail anchor

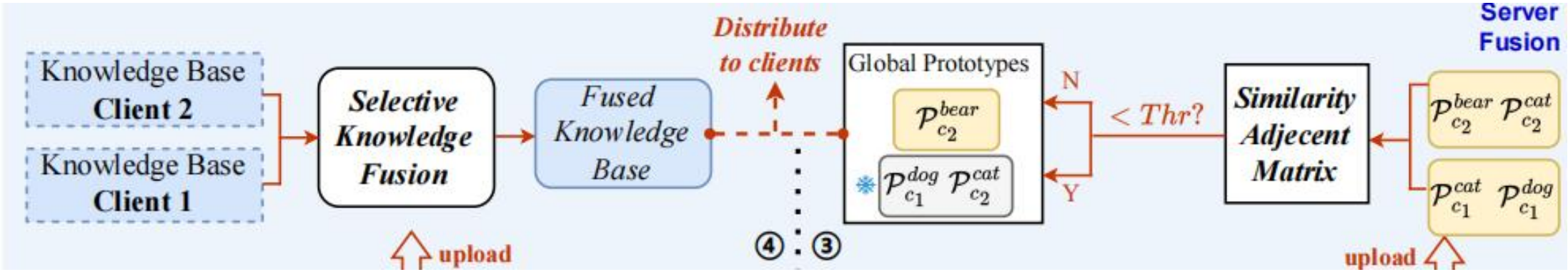
$$\mathcal{TA} = \{(K_1^{ta}, TA_1), (K_2^{ta}, TA_2), \ldots, (K_m^{ta}, TA_m)\}$$

$$K_s^{ta} = \arg\min \mathrm{dis}(\mathcal{F}_{out}, K_i^{ta})$$

$$\mathcal{L}_{cons}(\mathcal{F}_{TA}) = -\log \frac{\exp(\mathcal{F}_{TA} \cdot \mathcal{G}^y/\tau)}{\sum_{y_a \in \mathcal{Y}^t} \exp(\mathcal{F}_{TA} \cdot \mathcal{G}^{y_a}/\tau)},$$

$$\mathcal{L}_{ta} = \mathcal{L}_{CE}(\mathcal{F}_{TA}) + \lambda_2 \mathcal{L}_{cons}(\mathcal{F}_{TA}) + \lambda_3 \mathrm{dis}(\mathcal{F}_{TA}, K_s^{ta}),$$

Method



$$\mathcal{L}_{KD} = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} \underset{x_s \in \mathcal{D}_s}{\mathrm{MSE}}(\mathcal{V}(E_i'), \mathcal{V}(E_j')).$$

$$\mathcal{M}_{ij} = \mathrm{dis}(P_G^i, P_G^j), 0 < i \leq j \leq |P_G|.$$

$$\mathcal{G}^y = P_g^s = \underset{y_{low} \leq i \leq y_{high}}{\mathrm{argmin}} \ \bar{\mathcal{M}}_i = \frac{1}{|P_G|} \sum_{j=1}^{|P_G|} A_{ij},$$

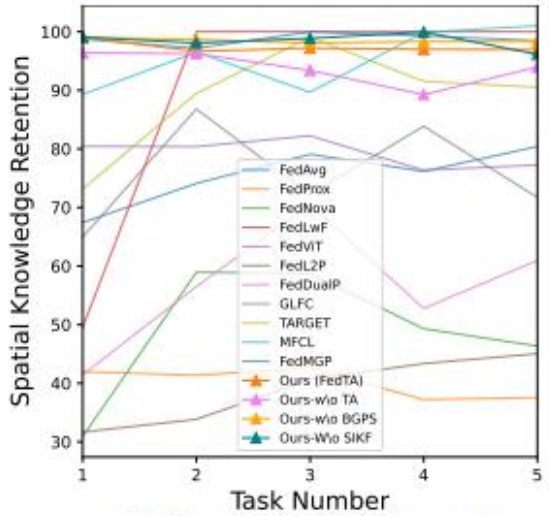
最优全局锚点： 与其他原型平均相似度最低
（即最具判别性，更不容易被误认为其他类别）

Experiments

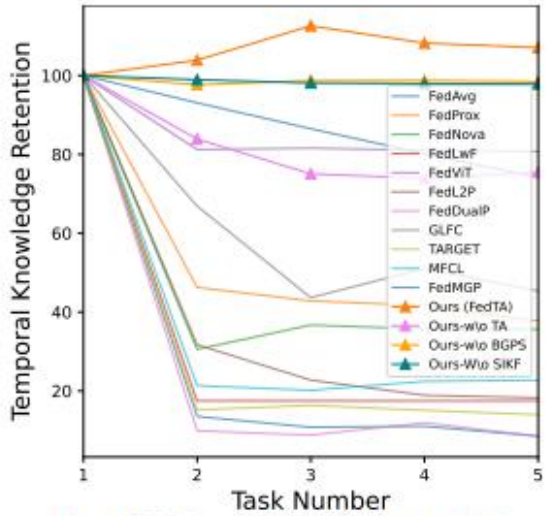
Table 1: Accuracy of the aggregated global model on local test sets with 5 class-incremental tasks.

Algorithm	Type	CIFAR-100 Task ID					Imagenet-R Task ID				
		1	2	3	4	5	1	2	3	4	5
FedAvg [26]	FL	43.9	50.6	57.3	55.5	61.2	37.7	35.4	35.5	35.8	36.7
FedProx [17]		23.7	22.8	26.0	22.1	23.6	20.2	19.7	19.7	18.9	17.8
FedNova [31]		13.7	18.8	20.1	16.1	15.1	2.0	4.7	4.6	7.8	7.7
FedLwF [23]	FL+CL	36.9	12.5	17.1	13.6	9.7	5.9	7.0	2.6	4.0	3.8
FedViT [4]		70.2	70.0	71.4	66.0	67.3	68.2	59.8	57.3	59.8	57.9
FedL2P [32]		28.4	29.9	29.3	25.4	25.7	27.1	27.6	24.8	25.1	26.5
FedDualP [33]		31.7	42.8	52.8	39.0	46.3	23.5	26.6	26.4	30.2	32.0
GLFC [3]	FCL	82.0	63.1	73.4	64.2	64.8	61.9	67.1	67.0	71.7	57.2
TARGET [41]		54.0	41.4	32.2	13.9	15.9	39.9	15.0	16.0	17.5	16.1
MFCL [1]		46.7	16.2	10.6	14.6	13.5	28.8	14.5	16.2	13.3	13.8
FedMGP [40]		90.2	85.3	90.7	89.2	82.2	77.3	76.8	78.0	75.6	75.4
Ours (FedTA)	FCL	**96.1**	**94.0**	**94.6**	**94.4**	**89.4**	**81.5**	78.8	79.2	80.6	**85.0**
Ours-w/o TA		78.7	75.5	73.4	69.3	72.3	79.6	72.3	72.3	74.1	63.2
Ours-w/o SIKF		90.7	88.8	91.1	91.4	89.1	80.0	**80.5**	**81.1**	**82.9**	81.7
Ours-w/o BGPS		90.8	92.5	88.4	91.4	88.6	80.5	78.7	78.3	81.7	79.1

Experiments

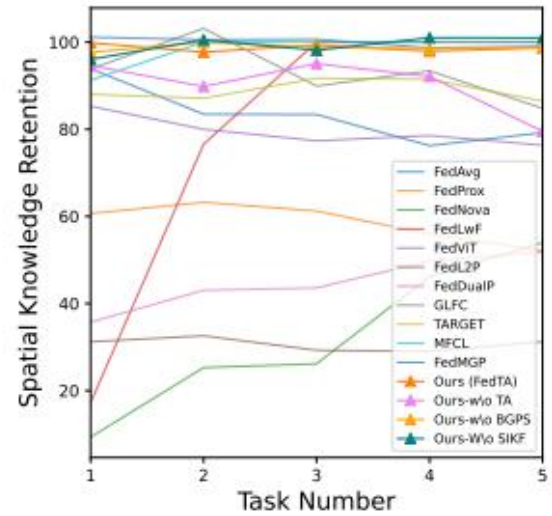


(a) KR_s on CIFAR-100.

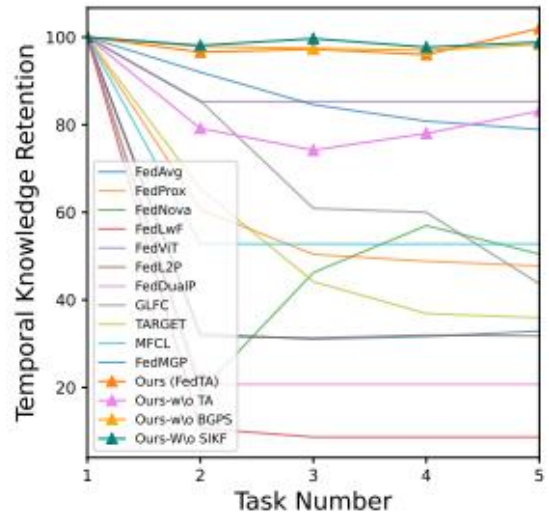


(b) KR_t on CIFAR-100.

$$KR_t = \frac{1}{a}\sum_{i=1}^{a}\frac{Acc(\theta_i^r; T_i^0)}{Acc(\theta_i^0; T_i^0)},$$



(c) KR_s on ImageNet-R.



(d) KR_t on ImageNet-R.

$$KR_s = \frac{1}{a}\sum_{i=1}^{a}\frac{Acc(\theta_g^r; T_i^r)}{Acc(\theta_i^r; T_i^r)},$$

Experiments

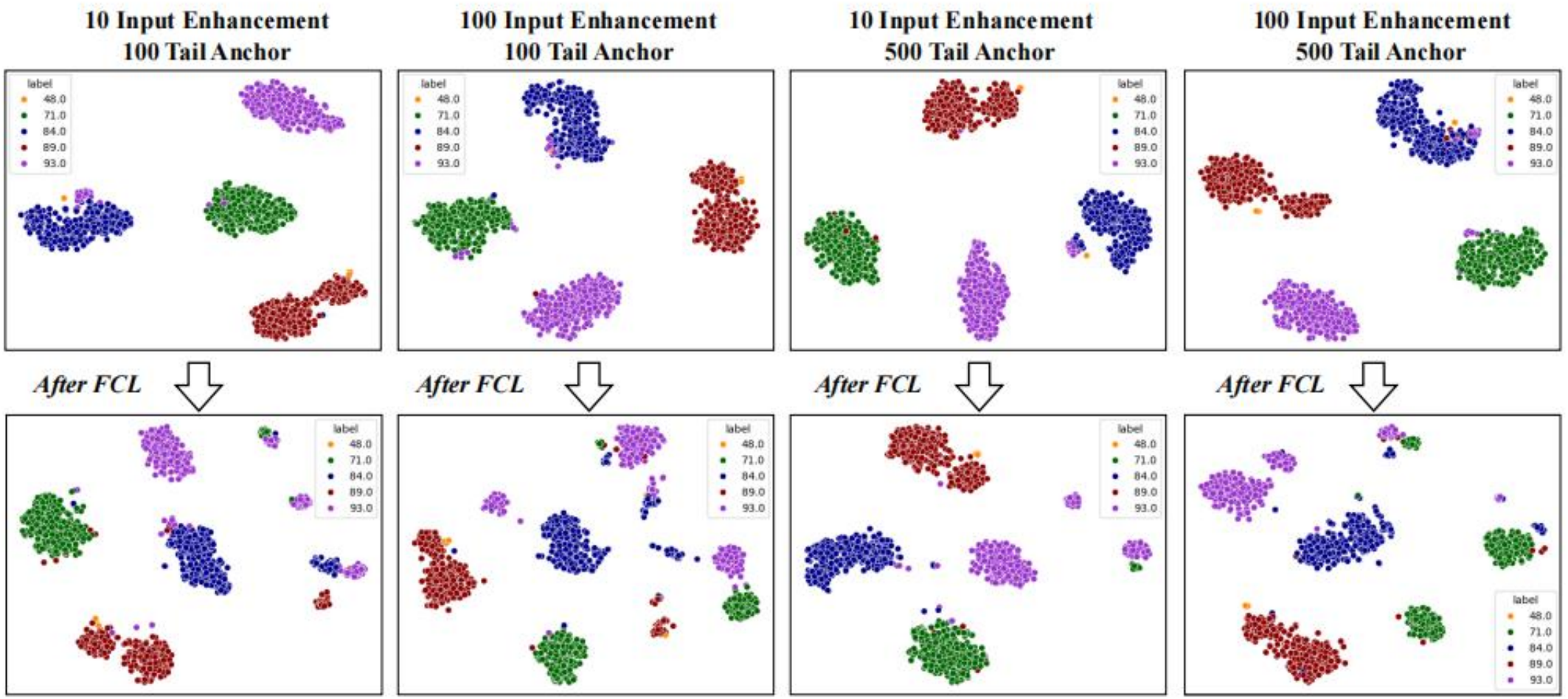


Figure 5: T-SNE for position changes of features corresponding to the same samples after FCL.

南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Personalized Federated Continual Learning via Multi-granularity Prompt

Hao Yu
School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics
Chengdu, China
yuhao2033@163.com

Xin Yang*
School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics
Chengdu, China
yangxin@swufe.edu.cn

Xin Gao
School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics
Chengdu, China
xingaocs@hotmail.com

Yan Kang
Webank
Shenzhen, China
kangyan2003@gmail.com

Hao Wang
College of Computer Science, Sichuan University
Chengdu, China
cshaowang@gmail.com

Junbo Zhang
JD Intelligent Cities Research
JD iCity, JD Technology
Beijing, China
msjunbozhang@outlook.com

Tianrui Li
School of Computing and Artificial Intelligence, Southwest Jiaotong University
Chengdu, China
trli@swjtu.edu.cn

KDD 2024/TPAMI 2026

Background

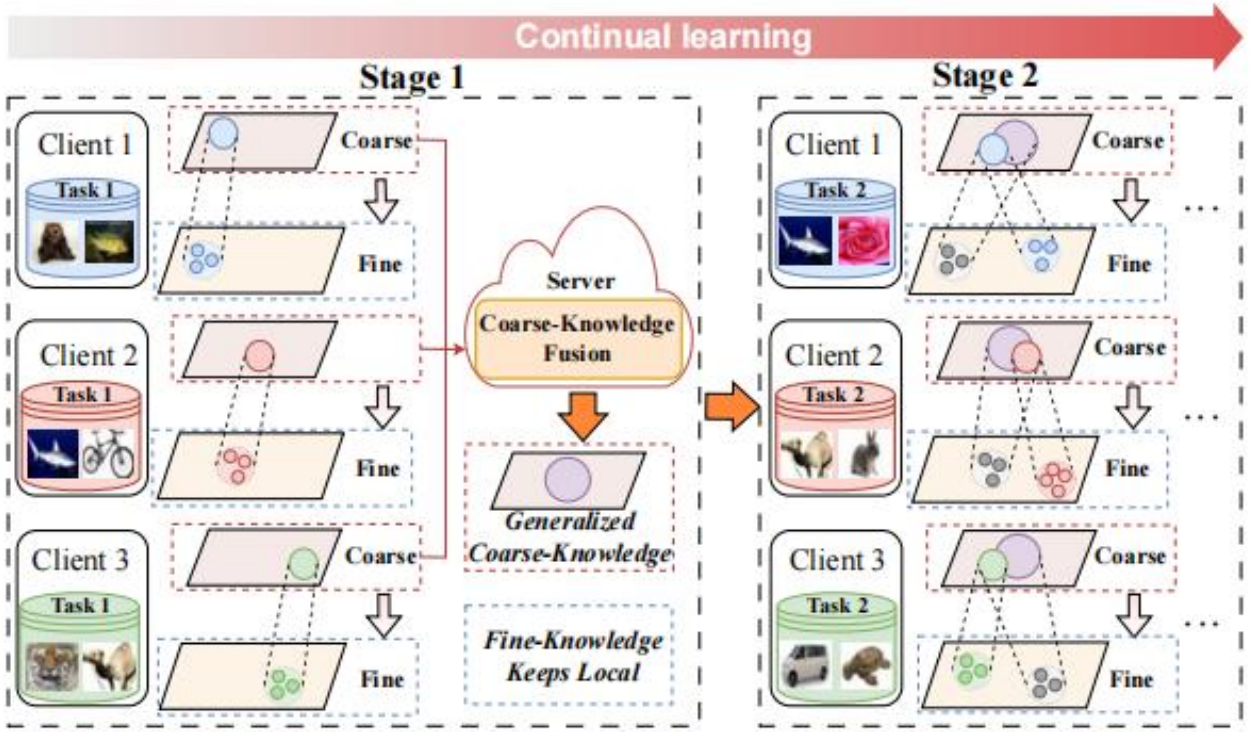
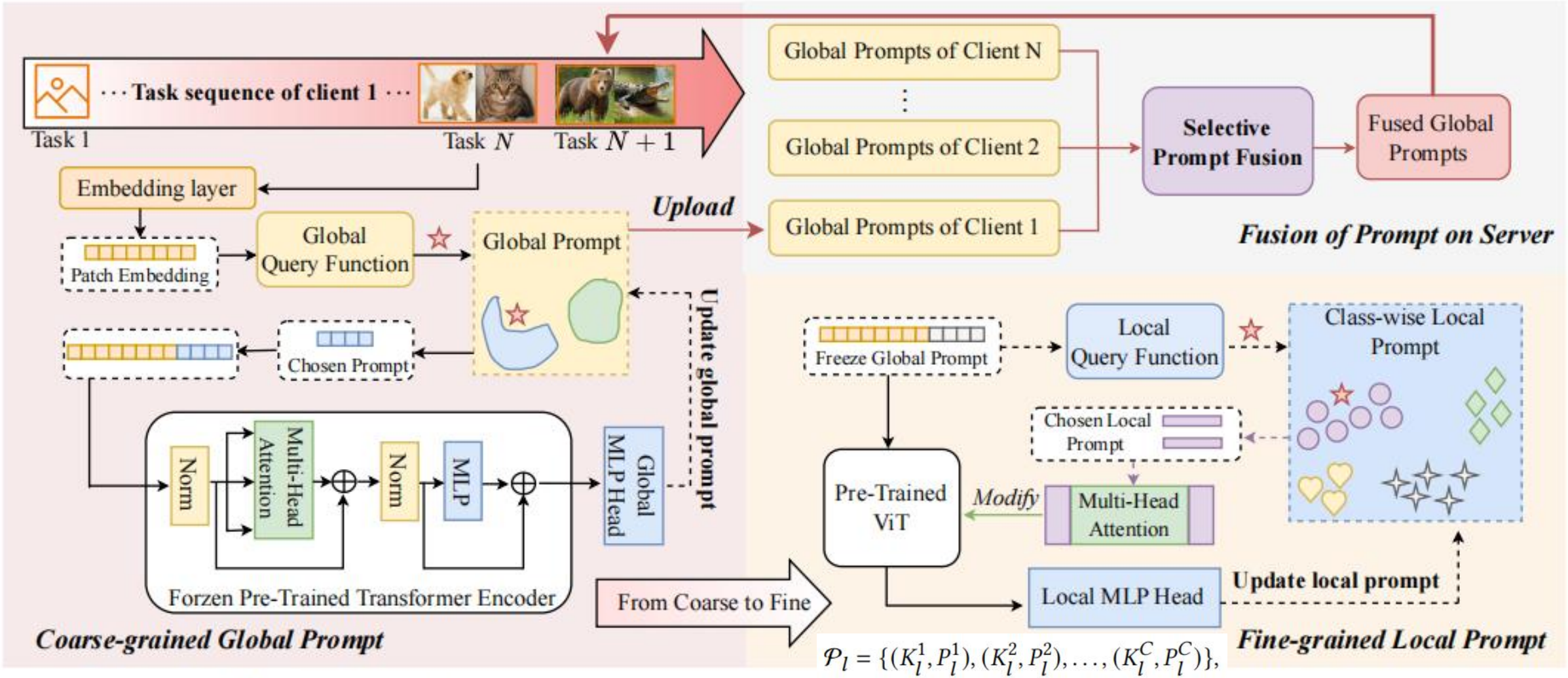


Figure 1: An illustration of constructing a multi-granularity knowledge space in PFCL. By dividing local knowledge into coarse-grained knowledge and fine-grained knowledge, better aggregation of common representations can be achieved on the server side. On the local side, fine-grained knowledge is used to personalize the generalized representation. The two different levels of knowledge can accumulate over time.

Method

$$\mathcal{L}_{CE} = \underset{x_s \in \mathcal{D}_f}{\text{MSE}}\left(\mathcal{V}(E_i'), \mathcal{V}(E_j')\right)$$



$$\mathcal{P}_g = \{P_g^1, P_g^2, \ldots, P_g^M\}, \qquad E' = \left[P_g^{s_1}, \ldots, P_g^{s_k}; E\right], 1 \le N \le M,$$

$$\mathcal{K}_g^s = \underset{\mathcal{K}_g}{\text{argmin}} \sum_{i=1}^{N} \text{dis}(K_g^{in}, K_g^i), \qquad \min_{H_g^i, \mathcal{P}_g, \mathcal{K}_g} \mathcal{L}(\mathcal{V}_g^i(E'), y) + \lambda_1 \sum_{\mathcal{K}_g^s} \text{dis}(K_g^{in}, K_g^{s_i}),$$

$$\mathcal{P}_l = \{(K_l^1, P_l^1), (K_l^2, P_l^2), \ldots, (K_l^C, P_l^C)\},$$

$$\text{MSA}(h_q, h_k, h_v) = \text{Concat}\left(\text{h}_1, \ldots, \text{h}_z\right) W^O,$$

$$\text{MSA}' = \text{MSA}(h_Q, [p_V; h_K], [p_V; h_V]).$$

$$\min_{H_l^i, \mathcal{P}_l, \mathcal{K}_l} \mathcal{L}(\mathcal{V}_l^i(E'), y) + \lambda_2 \sum_{\mathcal{K}_l^s} \text{dis}(K_l^{in}, K_l^{s_i}).$$

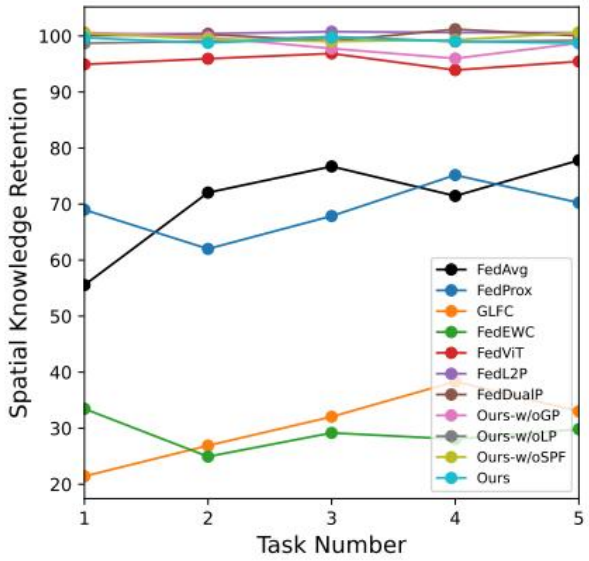
Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

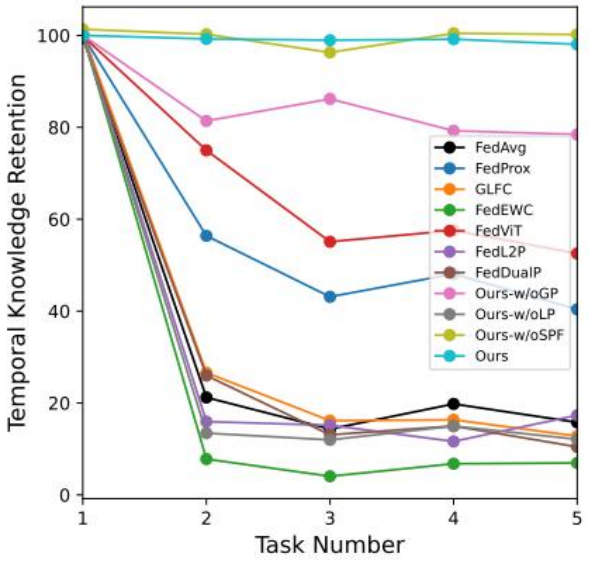
Table 1: Accuracy of the aggregated global model on local test sets CIFAR-100 with 5 class-incremental tasks each client.

Algorithm	Backbone	Asynchronous Task ID						Synchronous Task ID					
		1	2	3	4	5	Average	1	2	3	4	5	Average
FedAvg[31]	ResNet-18	47.39	62.63	67.25	62.69	68.72	61.74	67.77	77.72	76.65	74.59	82.00	75.74
FedProx[23]		68.26	56.94	65.20	63.82	67.48	64.34	43.07	23.09	51.41	45.01	50.62	42.64
FedEWC		27.77	20.66	25.70	24.11	26.41	24.93	49.94	71.00	70.22	70.09	77.89	67.83
GLFC[7]		14.22	18.84	23.93	26.70	22.52	21.24	5.22	8.93	24.61	35.43	42.33	23.30
FedViT	ViT	83.02	82.39	83.24	80.34	83.32	82.46	70.30	71.05	69.10	64.36	71.01	69.16
FedL2P		89.63	89.68	90.45	90.02	90.75	90.11	80.22	82.81	81.61	80.68	84.14	81.89
FedDualP		82.09	81.17	80.05	80.52	81.48	81.06	63.84	65.62	63.16	61.10	63.28	63.40
Ours(FedMGP)	ViT	**90.26**	**90.14**	**91.29**	**90.30**	**90.83**	**90.56**	**82.23**	**84.14**	**82.01**	**82.47**	**86.44**	**83.46**
Ours-w/oLP		88.35	89.19	89.91	89.16	90.20	89.36	79.80	82.04	79.71	80.20	83.56	81.06
Ours-w/oGP		86.93	88.52	82.85	84.11	87.29	85.94	78.73	78.92	77.74	75.16	79.10	77.93

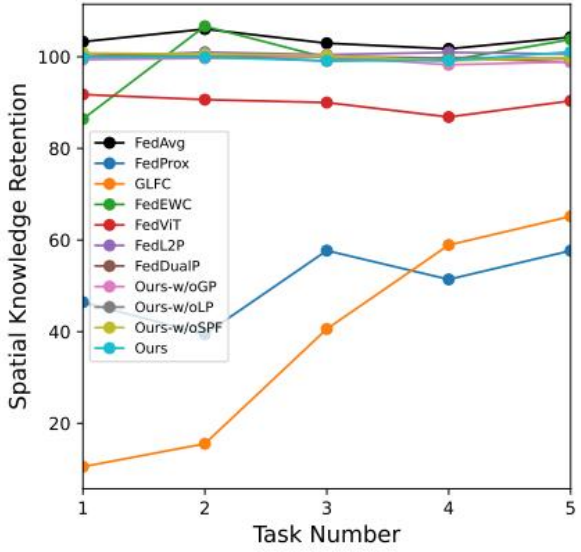
Experiments



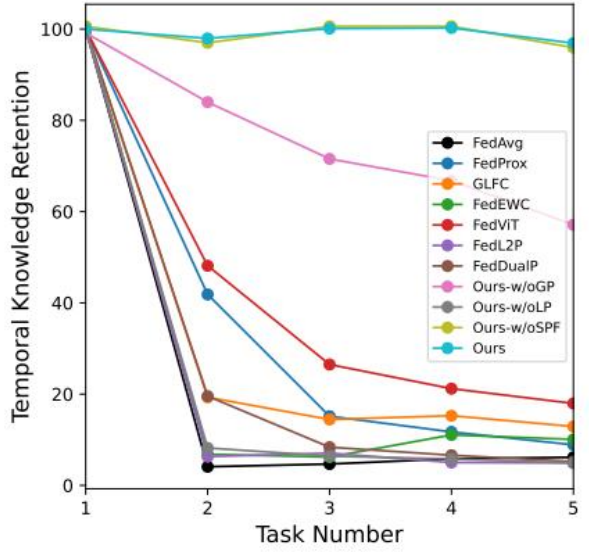
(a) Asynchronous KR_s



(b) Asynchronous KR_t



(c) Synchronous KR_s



(d) Synchronous KR_t

Method

(ICCV 2023, TARGET: Federated Class-Continual Learning via Exemplar-Free Distillation)

$$\mathcal{L}_G^{ce} = CE(\theta_{k-1}(\hat{x}), \hat{y}).$$

$$\mathcal{L}_G^{div} = -\omega KL(\theta_{k-1}(\hat{x}), \theta_S(\hat{x})), \text{ and}$$
$$\omega = \mathbb{1}(\arg\max \theta_{k-1}(\hat{x}) \neq \arg\max \theta_S(\hat{x})),$$

$$\mathcal{L}_G^{bn} = \sum_l (||\mu_l(\hat{x}) - \mu_l|| + ||\sigma_l^2(\hat{x}) - \sigma_l^2||),$$

$$\mathcal{L}_G = \mathcal{L}_G^{ce} + \lambda_1 \mathcal{L}_G^{div} + \lambda_2 \mathcal{L}_G^{bn}$$

$$\mathcal{L}_S = KL(\theta_{k-1}(\hat{x}), \theta_S(\hat{x}))$$

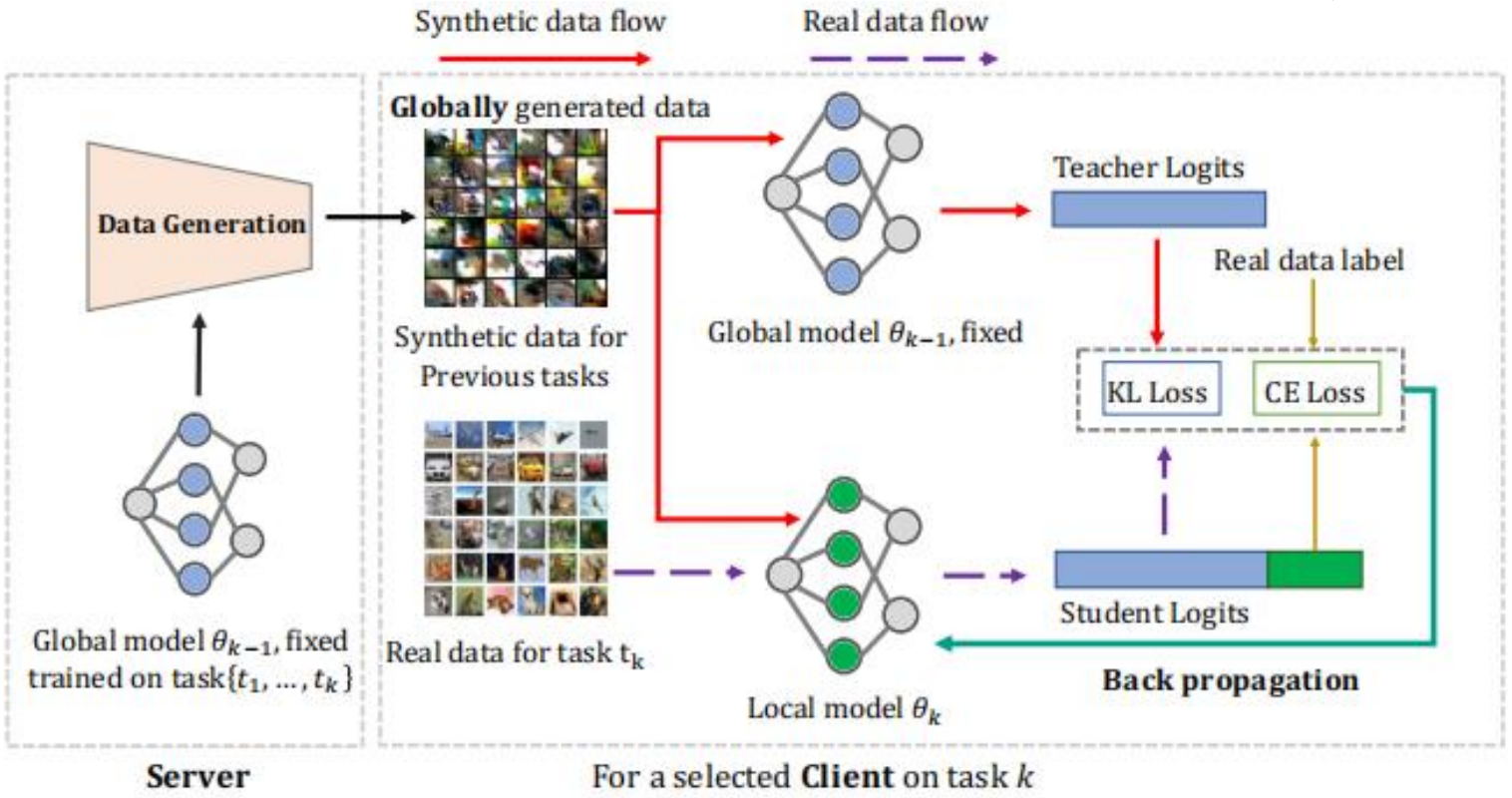


Figure 3: Pipeline of TARGET. We utilize the global model (trained on task $k-1$) to synthesize the data with a global distribution, subsequently employing this synthesized data for training the k-th task.

Experiments

Table 2: The Average Accuracy (%) and Forgetting for all learned tasks on CIFAR-100 for various numbers of tasks (5, 10) under both IID and non-IID settings. Results are reported as an average of 3 runs. 'Acc' refers to average accuracy, and '$\mathcal{F}$' represents the forgetting measure utilized in Equation 2. The best results are in bold.

Data partition	IID				NIID (1)				NIID (0.5)			
Tasks	T=5		T=10		T=5		T=10		T=5		T=10	
Method	Acc(↑)	$\mathcal{F}$(↓)	Acc(↑)	$\mathcal{F}$(↓)	Acc(↑)	$\mathcal{F}$(↓)	Acc(↑)	$\mathcal{F}$(↓)	Acc(↑)	$\mathcal{F}$(↓)	Acc(↑)	$\mathcal{F}$(↓)
Finetune	16.12	0.78	7.83	0.75	16.33	0.77	8.45	0.74	15.49	0.74	7.64	0.71
FedEWC	16.51	0.71	8.01	0.65	16.06	0.68	8.84	0.62	16.86	0.66	8.04	0.65
FedWeIT	28.45	0.52	20.39	0.43	28.56	0.49	19.68	0.45	24.57	0.54	15.45	0.48
FedLwF	30.61	0.45	23.27	0.37	30.94	0.42	21.16	0.41	27.59	0.44	17.98	0.45
Ours	**36.31**	**0.22**	**24.76**	**0.26**	**34.89**	**0.24**	**22.85**	**0.26**	**33.33**	**0.27**	**20.71**	**0.29**

Method

(NIPS 2023, A Data-Free Approach to Mitigate Catastrophic Forgetting in Federated Class Incremental Learning for Vision Tasks)

$$\min_{\mathcal{G}} \mathcal{L}_{CE} + w_{div}\mathcal{L}_{div} + w_{BN}\mathcal{L}_{BN} + w_{pr}\mathcal{L}_{pr}$$

$$\mathcal{L}_{pr} = ||\tilde{x} - Smooth(\tilde{x})||_2^2.$$

$$\min_{\mathcal{F}_t} \mathcal{L}_{CE}^t + w_{FT}\mathcal{L}_{FT}^t + w_{KD}\mathcal{L}_{KD}^t$$

$$\mathcal{L}_{KD}^t = ||\mathcal{W}(\mathcal{F}_t^{1:L-1}([x,\tilde{x}])) - \mathcal{W}(\mathcal{F}_{t-1}^{1:L-1}([x,\tilde{x}]))||_2^2$$

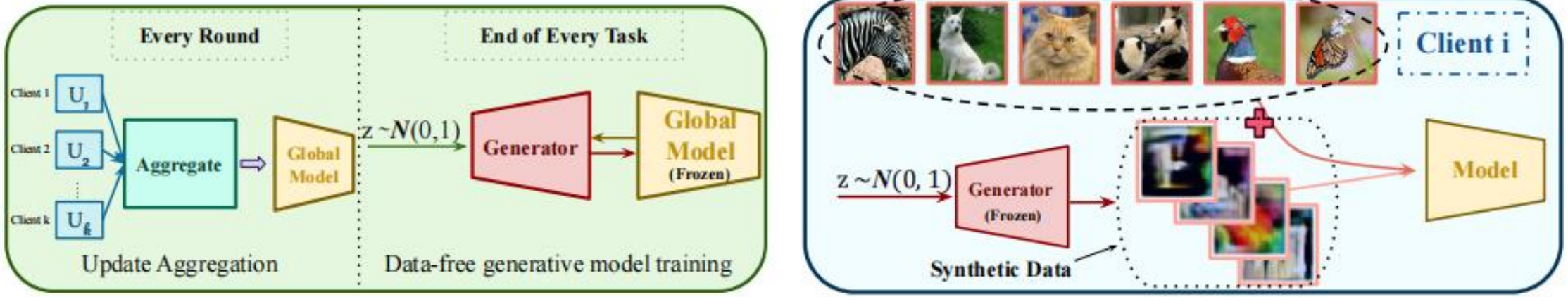
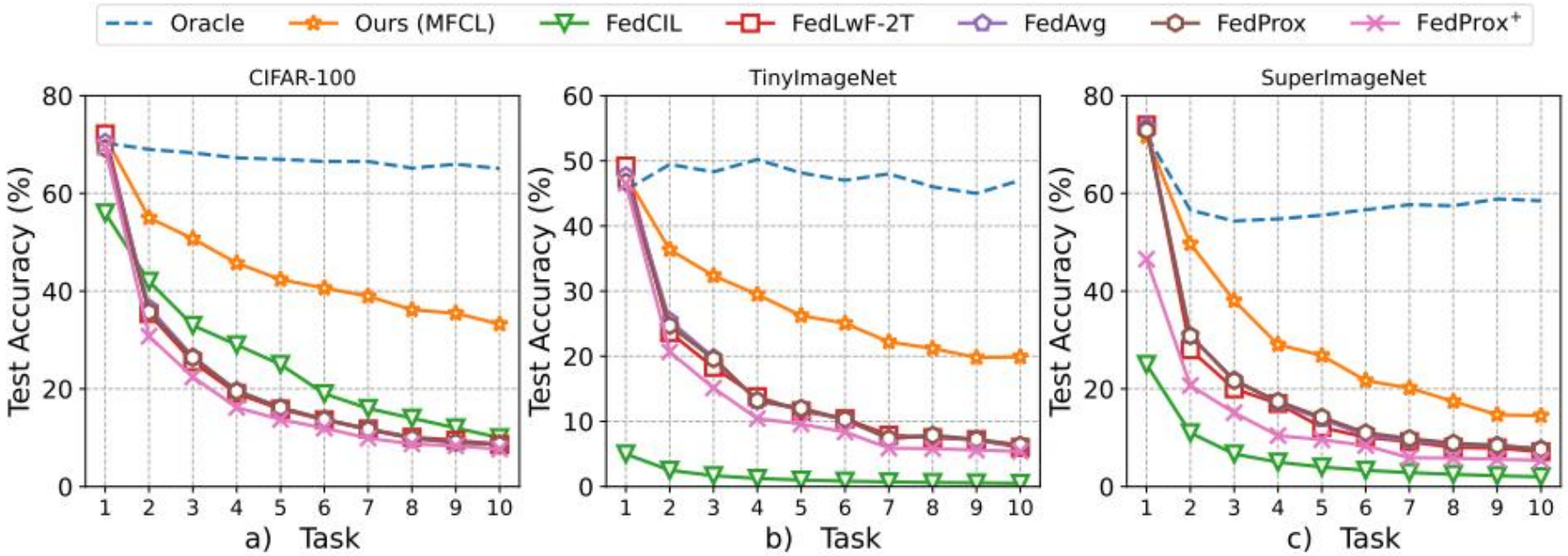


Figure 3: Overview of MFCL. **Left.** The server aggregates the updates every round and trains a generator using data-free methods at the end of each task. **Right.** Clients train their models locally using their local data and synthetic images of past tasks from the generator.

Experiments

Figure 4: Building SuperImageNet by regrouping ImageNet dataset. Labels in Blue are the original labels, and in Red are the labels in SuperImageNet.



Thanks