

Data-Centric Model Adaptation

- **Name:** Jianqing Zhang
- **2-year Ph.D.:** Shanghai Jiao Tong University
- **Visiting:** Hong Kong Polytechnic University (PolyU)
- **Collaborations:**
 - Qiang Yang, HKUST & PolyU, China
 - Yang Hua, Queen's University Belfast, UK
 - Yang Liu, PolyU, China
 - Marco Canini, KAUST, Saudi Arabia
 - Han Yu, NTU, Singapore
 - Yuyin Zhou, UC Santa Cruz, USA





Overview

- **Research interests:** **Data-Centric Model Adaptation**
- **Research fields:** *Code LLM, Synthetic Data Generation, Federated Learning, Recommender System*
- **Open-sourced projects** (initiator):
 - EvolveGen, PFLlib (**1,800+** stars, **300+** forks), HtFLlib, HtFLlib on Device, FL-IoT, etc.
- **Publications:** **9** first-author top-tier papers
 - AAAI'23 (oral), KDD'23, ICCV'23, NeurIPS'23, JMLR'25 (WAIC Outstanding Paper Top40)
 - AAAI'24, CVPR'24, KDD'25 (Best Paper Runner Up)
 - EMNLP'24, ICML'25, ICML'25 (spotlight)
- **Outstanding advantages:** **Passion, research + application, keen sense of fields**
- **Awards:** Youth Talent of China Association for Science and Technology (China Association for Artificial Intelligence, CAAI), Wenjun Wu Honorary Doctorate in AI, PhD National Scholarship, etc.
- **Applications:** ① Cross-hospital cancer recognition model, ② Cross-province intelligent 12345 hotline model, ③ HtFL testbed on real-world devices, ④ Led 9-member team in building a Federated ML platform for data centers
- **Impact:** **800+** citations, **30K+** views across major media, well-recognized by IEEE/ACM Fellows
- **Intern:** ByteDance AML, Tsinghua AIR, KAUST SANDs lab, Tencent CodeBuddy



Systematical Research Trace

Data-Centric Model Adaptation: Balancing generalization and specialization on new data from a data-driven perspective

① **Recommender:** **Exploiting** long- and short-term user behavior data



② **Federated learning:** **Exploiting** generalization and specialization in heterogeneous data across devices



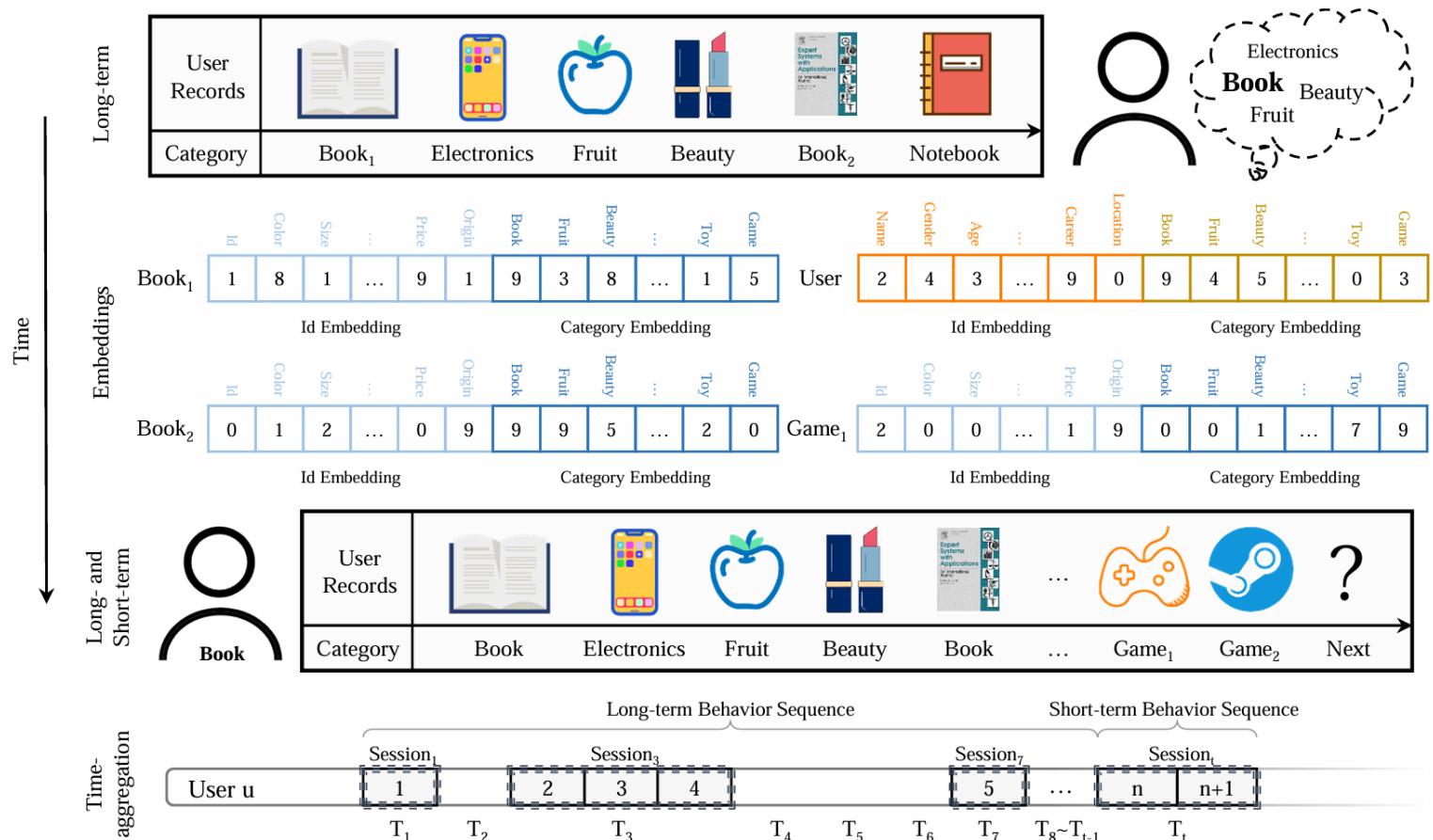
③ **Synthetic dataset generation:** **Generating** vertical domain data based on few private data and large models



④ **Code LLM reinforcement learning:** **Exploiting** self-generated data to reduce coding errors

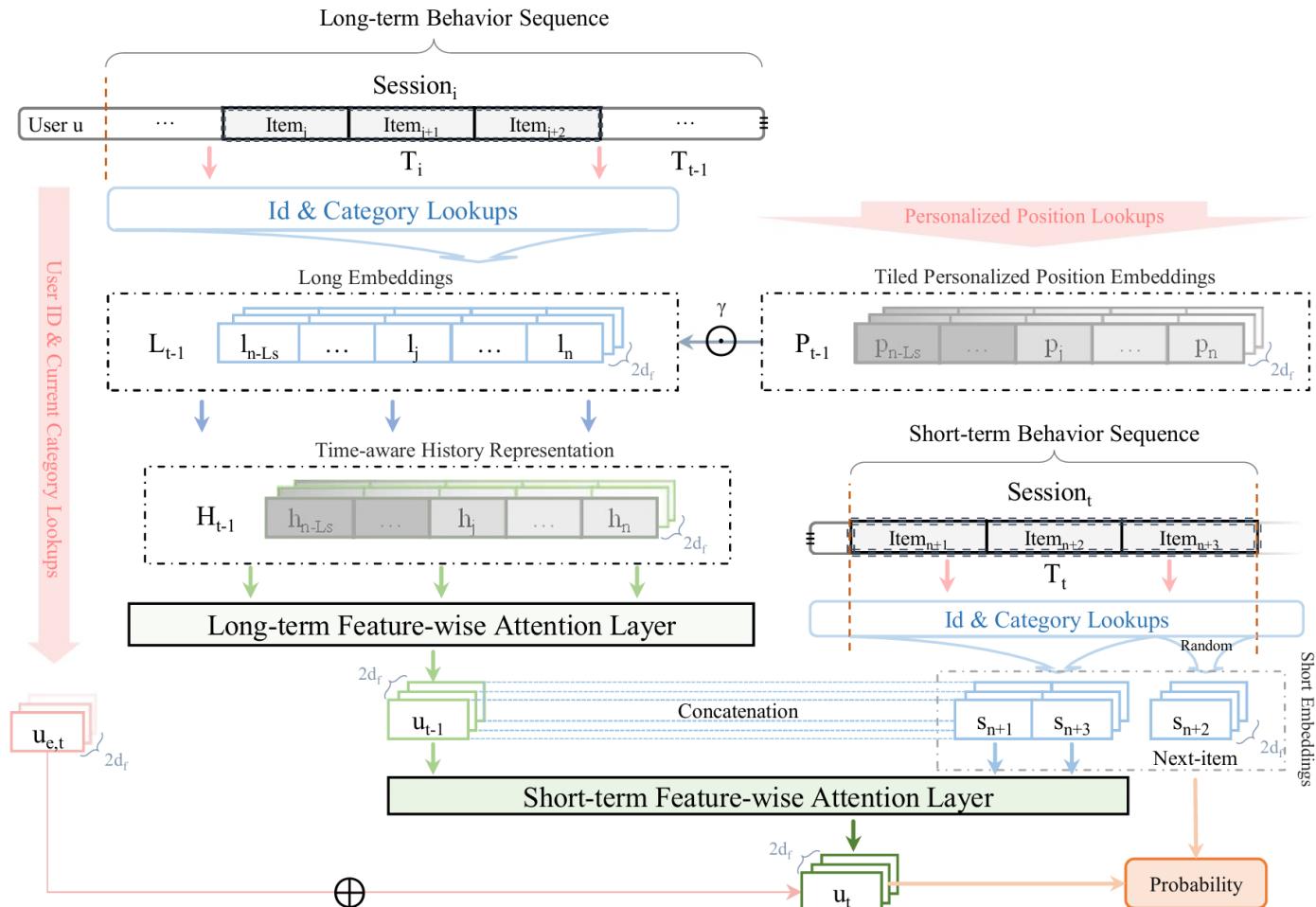
Recommender System (RS)

- **Problem:** Users have **personalized taste for time** in the behavior data besides others
- **Solution:** Time-aware Long- and Short-term Attention Network for Next-item Recommendation



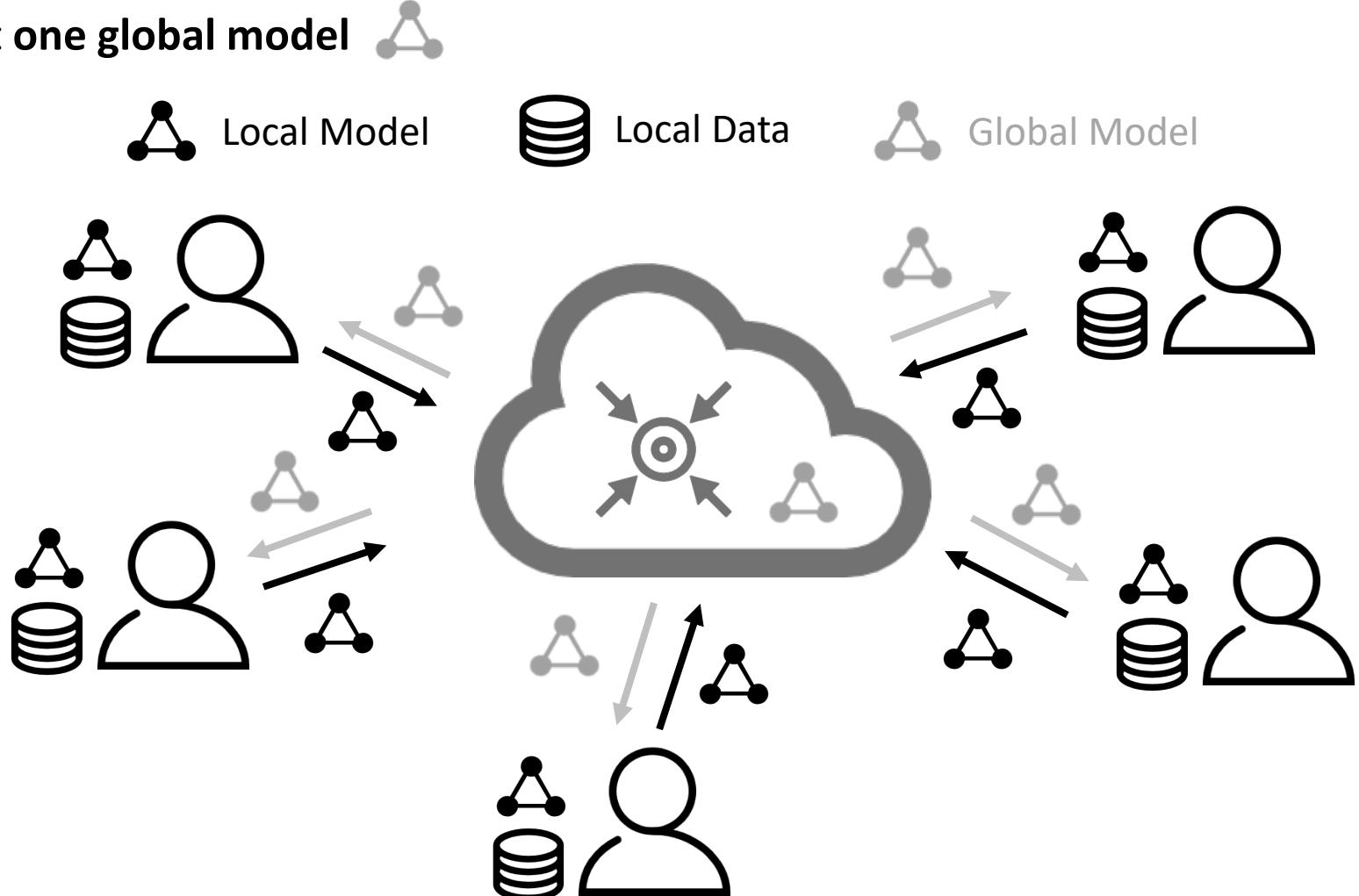
[RS]: TLSAN

- Capture personalized time-aggregation pattern in long-term behavior data via attention



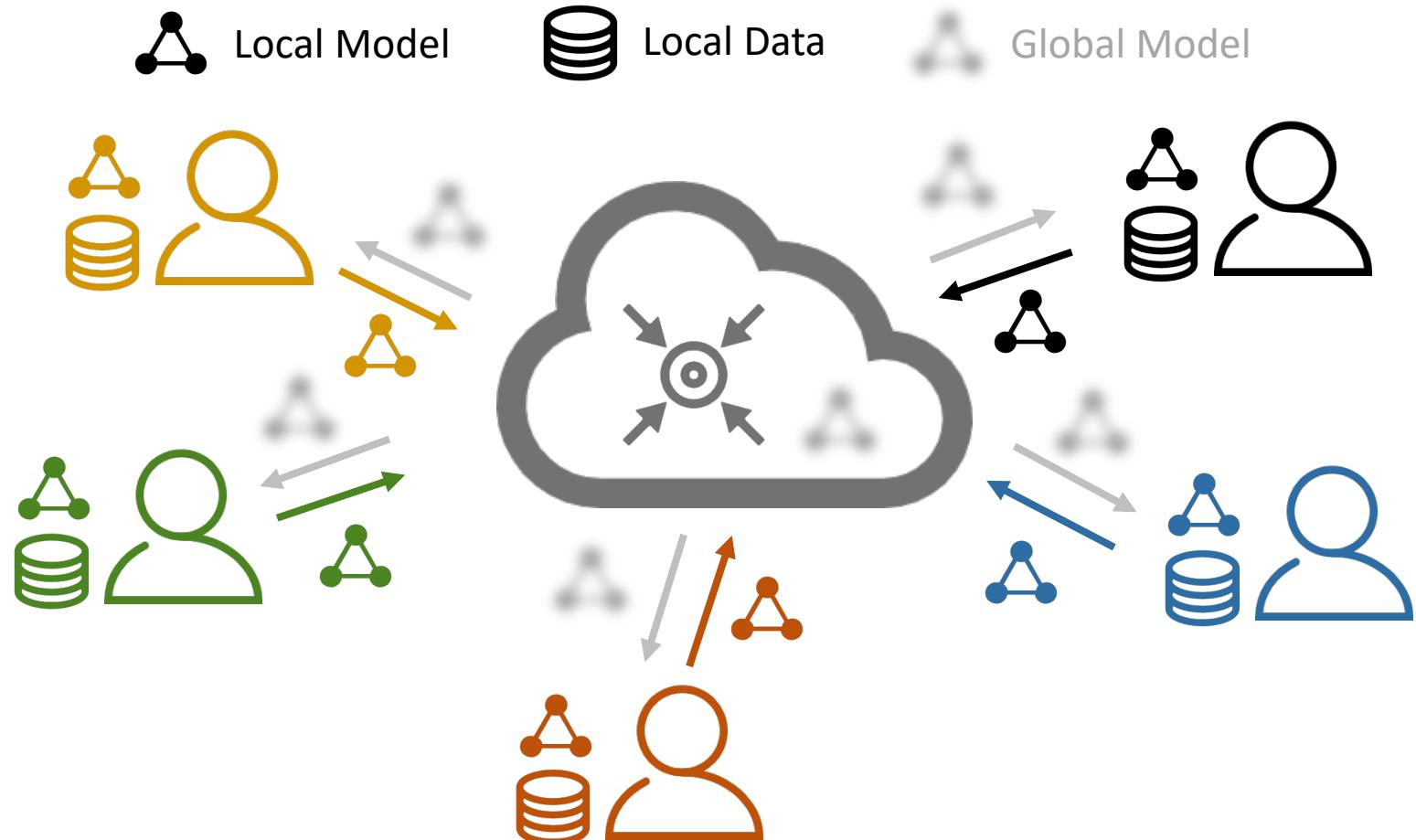
Federated Learning (FL)

- A **collaborative** and **privacy-preserving** technique for AI model training
- Finally output **one global model**



[FL]: Data Heterogeneity

- **Problem:** Different clients have different data distributions, resulting in a **poor global model**



[FL]: PFLlib: pFL algorithm library and benchmark

- Beginner-friendly
- 39 FL&pFL, 3 scenarios, 24 datasets
- Popular (1800+ stars)
- 500 clients: 5GB GPU memory
- Rapidly developing:

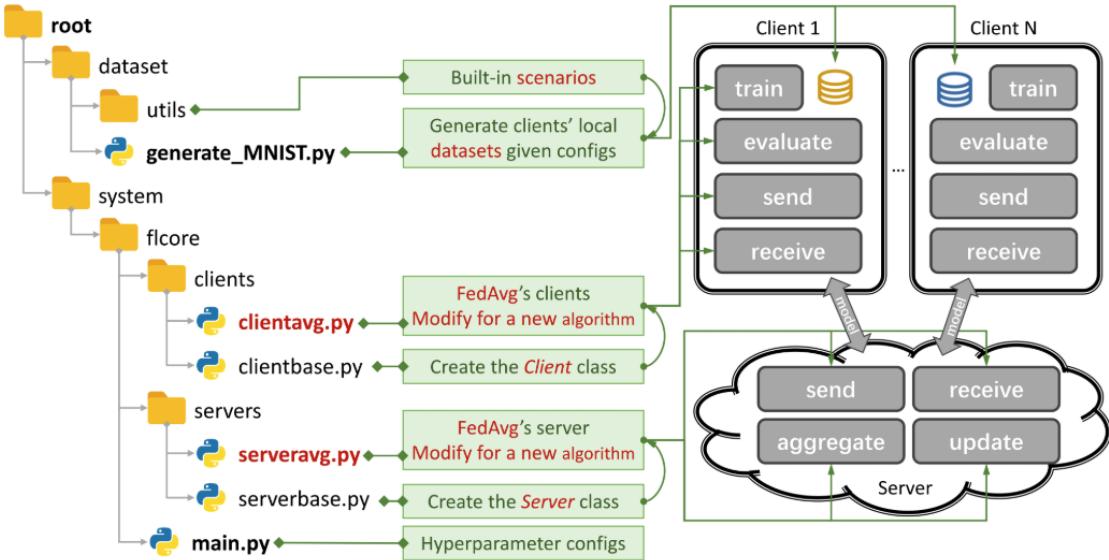
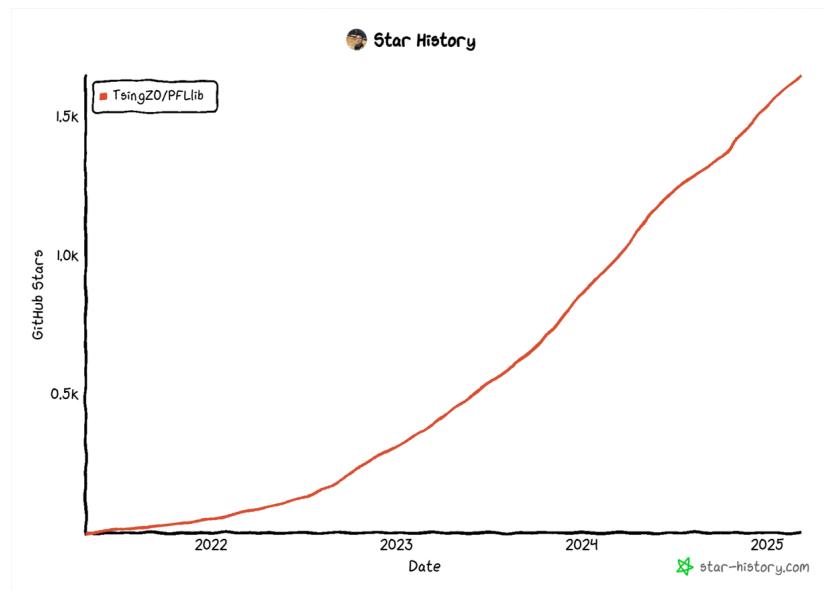


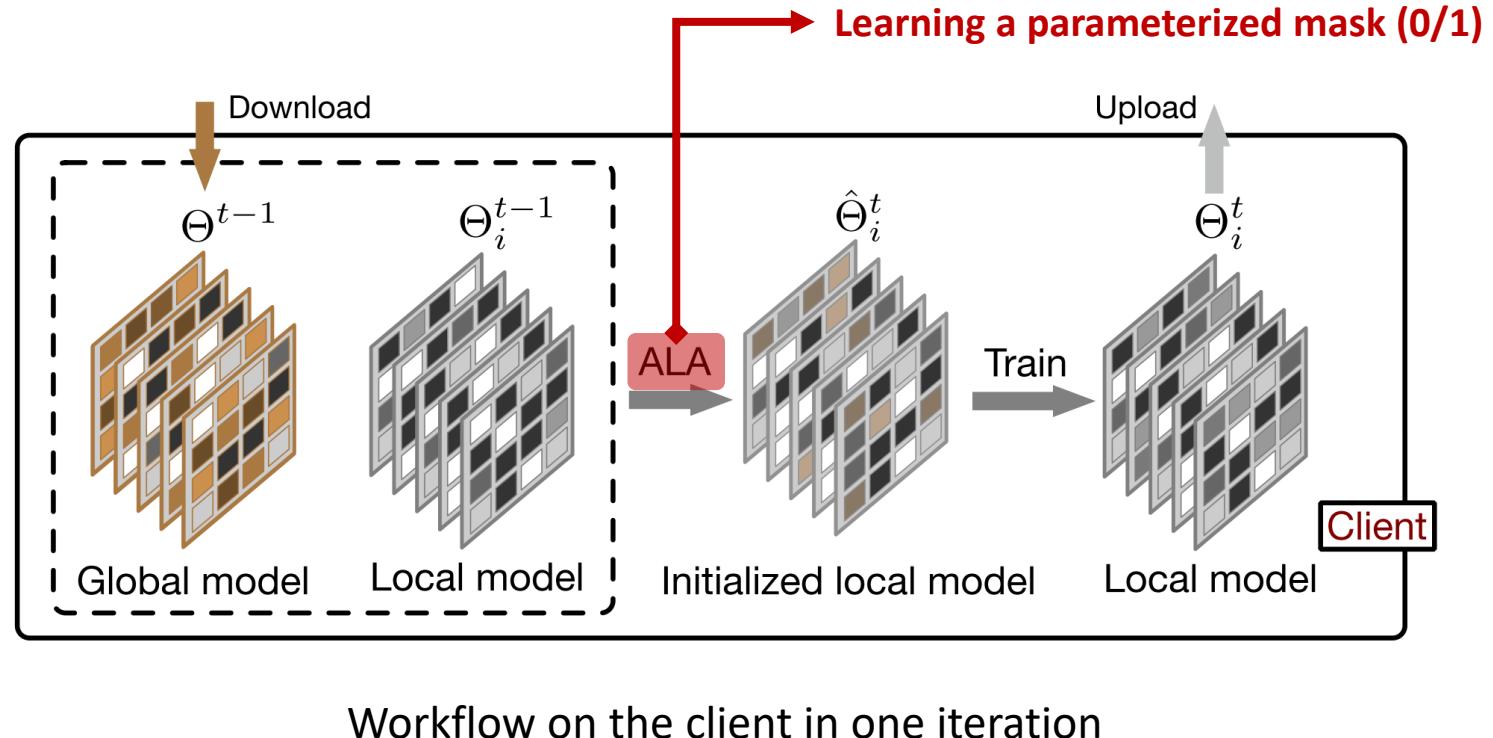
Figure 1: An Example for FedAvg. You can create a scenario using `generate_DATA.py` and run an algorithm using `main.py`, `clientNAME.py`, and `serverNAME.py`. For a new algorithm, you only need to add new features in `clientNAME.py` and `serverNAME.py`.

The screenshot shows the PFLlib website. The homepage features a 'Star History' chart and a 'PFLlib Is All You Need' section. The 'Benchmark Platform' section displays experimental results for various algorithms across different datasets and settings. The table below shows a portion of the benchmark data.

Setting	Pathological Label Skew Setting				Practical Label Skew Setting			
	MNIST	Cifar100	TINY	MNIST	Cifar100	TINY	MNIST	AG News
FedAvg	80.41 ± 0.06	47.23 ± 0.07	39.90 ± 0.62	97.45 ± 0.04	52.87 ± 0.04	32.15 ± 0.04	35.95 ± 0.43	98.89 ± 0.17
FedProx	78.05 ± 0.15	25.54 ± 0.16	14.20 ± 0.67	85.85 ± 0.19	31.89 ± 0.47	19.46 ± 0.20	19.85 ± 0.13	87.12 ± 0.19
FedDPS	79.75 ± 0.06	20.80 ± 1.00	13.85 ± 0.25	85.60 ± 0.57	31.89 ± 0.47	19.37 ± 0.22	19.87 ± 0.23	87.21 ± 0.13
FedFog	99.18 ± 0.54	56.80 ± 0.28	28.06 ± 0.60	95.15 ± 0.10	44.83 ± 0.07	25.07 ± 0.07	21.89 ± 0.54	98.86 ± 0.83
Per-FedAvg	99.13 ± 0.54	56.80 ± 0.28	28.06 ± 0.60	95.15 ± 0.10	44.83 ± 0.07	25.07 ± 0.07	21.89 ± 0.54	98.86 ± 0.83
pFedMe	99.13 ± 0.14	58.02 ± 0.14	27.71 ± 0.60	97.25 ± 0.17	47.34 ± 0.48	24.89 ± 0.19	33.44 ± 0.33	97.08 ± 0.18
Ditto	99.41 ± 0.06	47.23 ± 0.07	39.90 ± 0.62	97.45 ± 0.04	52.87 ± 0.04	32.15 ± 0.04	35.95 ± 0.43	98.89 ± 0.17
APFL	99.41 ± 0.03	54.26 ± 0.13	34.47 ± 0.64	97.25 ± 0.08	46.74 ± 0.08	34.86 ± 0.04	35.81 ± 0.37	98.97 ± 0.06
FedML	99.41 ± 0.03	52.69 ± 0.22	36.55 ± 0.50	97.25 ± 0.02	45.79 ± 0.48	24.83 ± 0.22	26.48 ± 0.11	92.02 ± 0.18
FedAMP	99.42 ± 0.03	56.84 ± 0.37	36.12 ± 0.30	97.25 ± 0.06	47.69 ± 0.48	27.99 ± 0.11	29.17 ± 0.15	88.35 ± 0.05
APPLE	99.13 ± 0.03	55.68 ± 0.08	36.22 ± 0.60	97.05 ± 0.07	53.22 ± 0.20	35.04 ± 0.47	39.93 ± 0.52	94.83 ± 0.18
FedAL	99.17 ± 0.03	47.81 ± 0.06	40.31 ± 0.30	97.66 ± 0.02	55.92 ± 0.07	41.94 ± 0.02	42.65 ± 0.10	

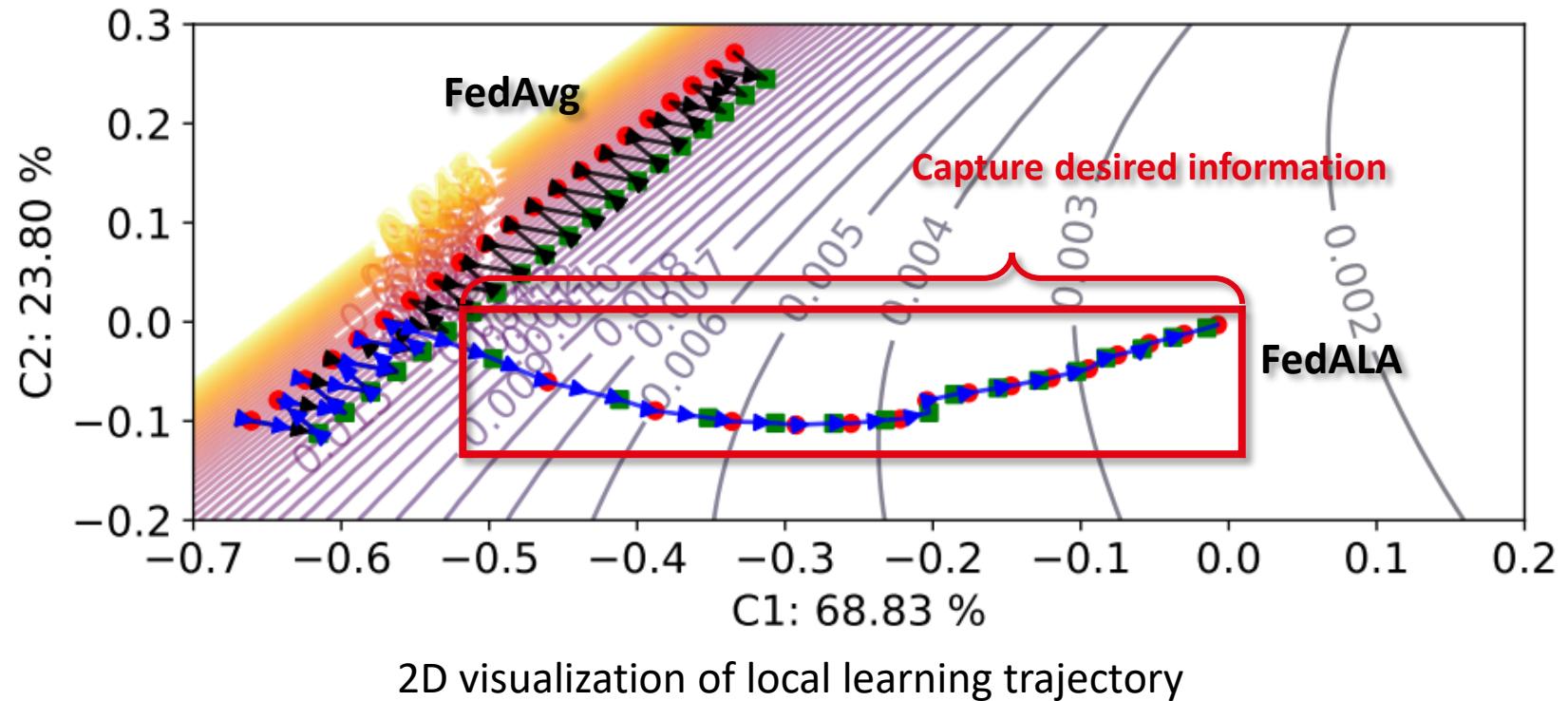
[FL]: FedALA

- Extract each client's **desired parameters** from the global model to facilitate local training
- **Adaptively aggregate** the parameters in the global and local model in each communication round



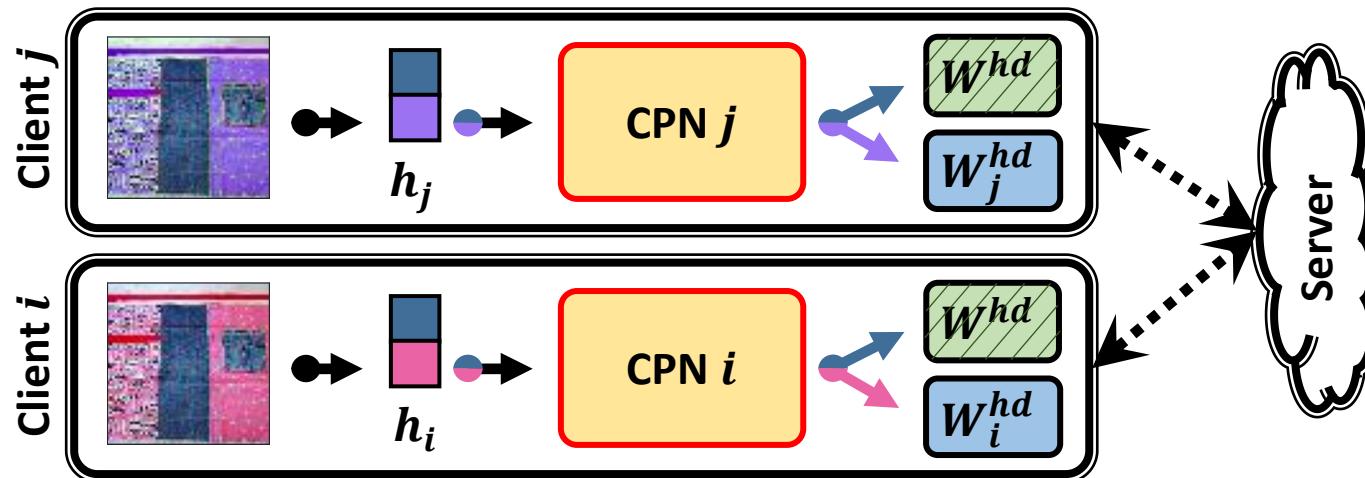
[FL]: FedALA

- Learning trajectory on one client: **FedAvg** vs. **FedALA**
- Activate ALA in the subsequent iterations



[FL]: FedCP

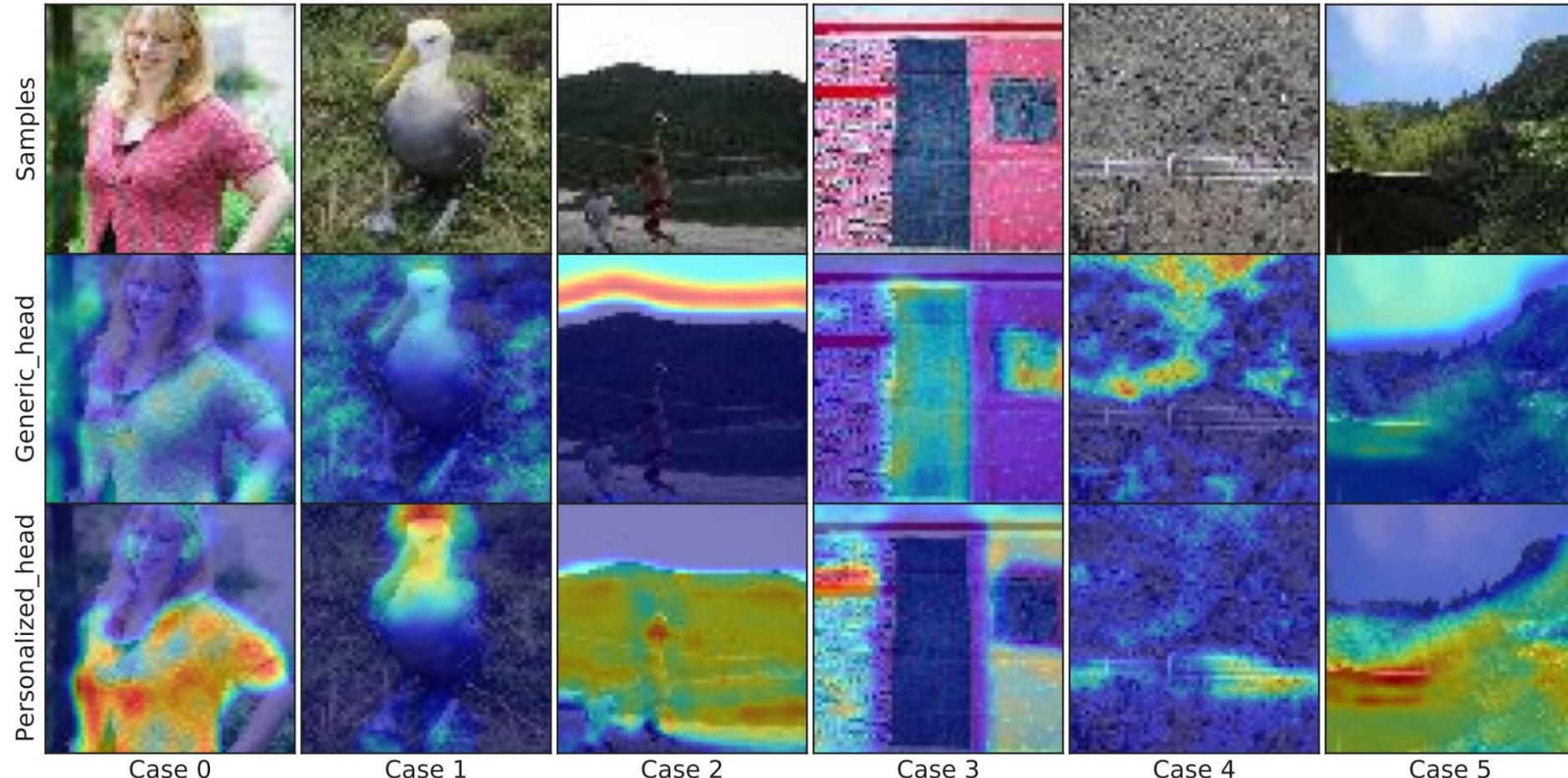
- We separate feature information via an auxiliary **Conditional Policy Network (CPN)**.
 - Detach generic and personalized information
 - Sample-specific separation
 - Lightweight (e.g., 4.67% parameters of ResNet-18)



- Then, we utilize **global and personalized information** via the corresponding heads.

[FL]: FedCP

- Effect of feature information separation



Six samples from the Tiny-ImageNet dataset

[FL]: GPFL

- GCE introduces **generic and personalized routes** like MoE
- CoV **eliminates the interaction** between global and personalized feature learning

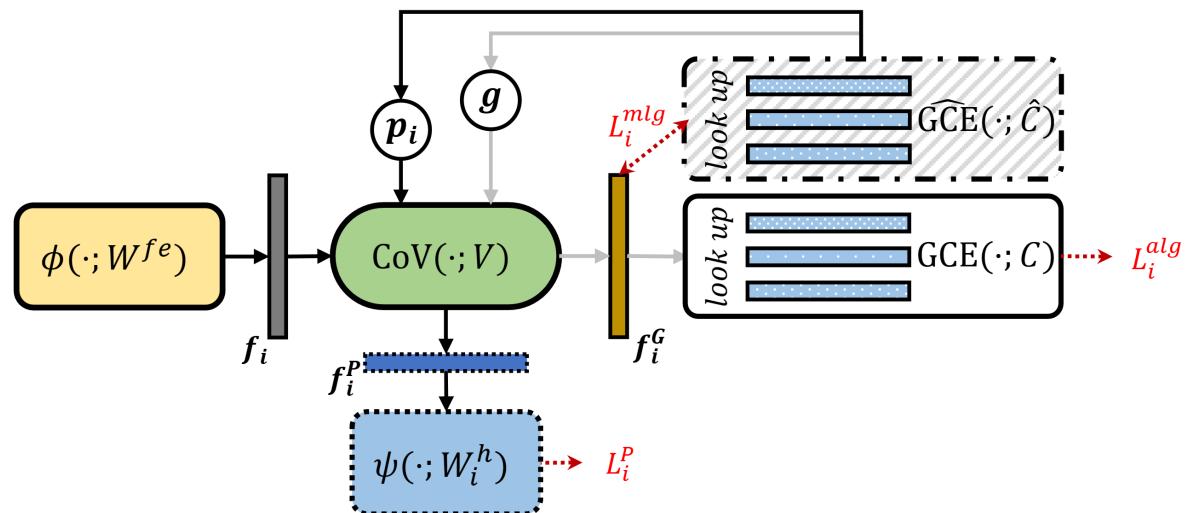
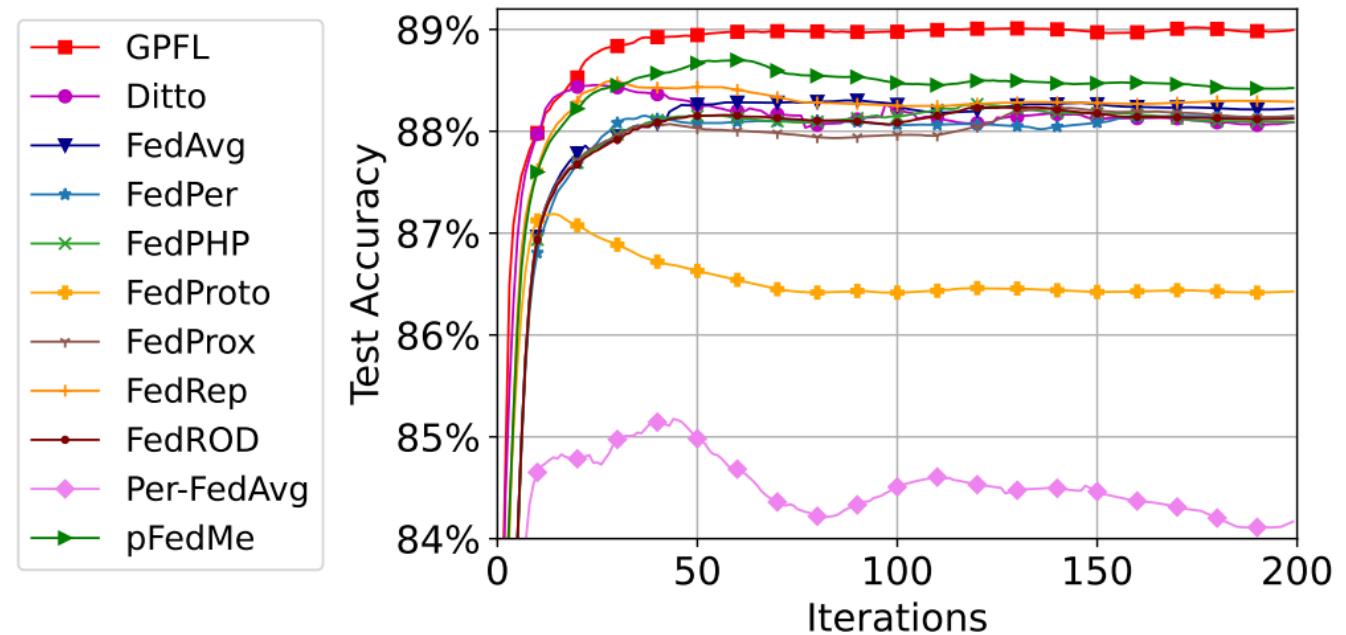


Illustration of client modules and data flow between them

[FL]: GPFL

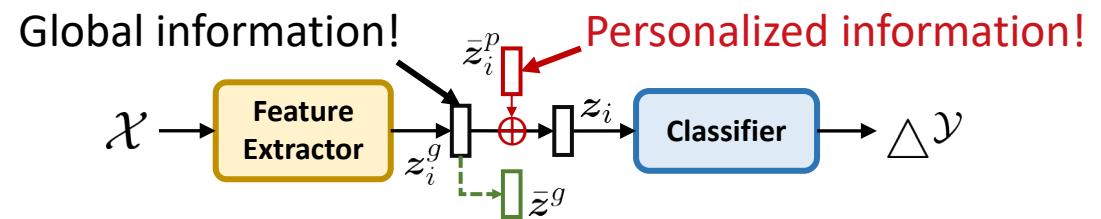
- Relieve the **widely existed** overfitting issue in pFL



Test accuracy curves in the feature shift setting

[FL]: DBE

- Eliminate domain bias by store **personalized information** in PRBM
- Enhance **information disentanglement** by guiding feature extractor with MR



Local model (with PRBM and MR)



[FL]: DBE

- Improve bi-directional knowledge transfer

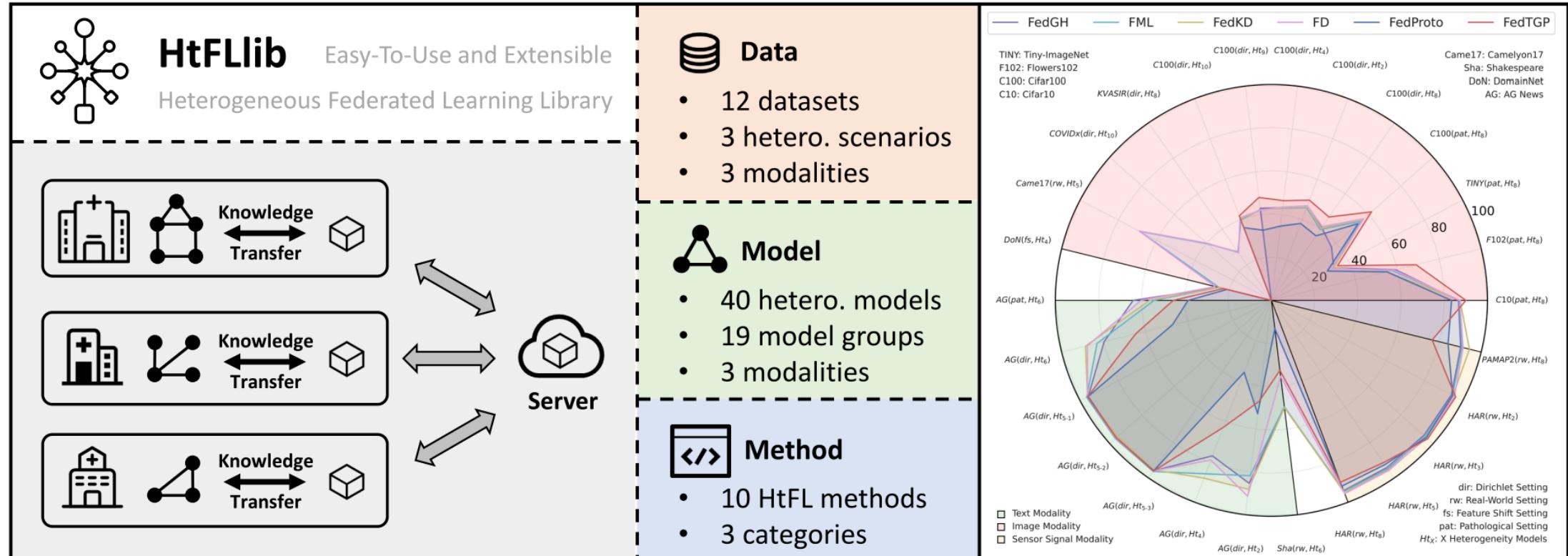
Corollary 1. Consider a local data domain \mathcal{D}_i and a virtual global data domain \mathcal{D} for client i and the server, respectively. Let $\mathcal{D}_i = \langle \mathcal{U}_i, c^* \rangle$ and $\mathcal{D} = \langle \mathcal{U}, c^* \rangle$, where $c^* : \mathcal{X} \mapsto \mathcal{Y}$ is a ground-truth labeling function. Let \mathcal{H} be a hypothesis space of VC dimension d and $h : \mathcal{Z} \mapsto \mathcal{Y}, \forall h \in \mathcal{H}$. When using DBE, given a feature extraction function $\mathcal{F}^g : \mathcal{X} \mapsto \mathcal{Z}$ that shared between \mathcal{D}_i and \mathcal{D} , a random labeled sample of size m generated by applying \mathcal{F}^g to a random sample from \mathcal{U}_i labeled according to c^* , then for every $h^g \in \mathcal{H}$, with probability at least $1 - \delta$:

$$\mathcal{L}_{\mathcal{D}}(h^g) \leq \mathcal{L}_{\hat{\mathcal{D}}_i}(h^g) + \sqrt{\frac{4}{m} \left(d \log \frac{2em}{d} + \log \frac{4}{\delta} \right)} + d_{\mathcal{H}}(\tilde{\mathcal{U}}_i^g, \tilde{\mathcal{U}}^g) + \lambda_i,$$

where $\mathcal{L}_{\hat{\mathcal{D}}_i}$ is the empirical loss on \mathcal{D}_i , e is the base of the natural logarithm, and $d_{\mathcal{H}}(\cdot, \cdot)$ is the \mathcal{H} -divergence between two distributions. $\lambda_i := \min_{h^g} \mathcal{L}_{\mathcal{D}}(h^g) + \mathcal{L}_{\mathcal{D}_i}(h^g)$, $\tilde{\mathcal{U}}_i^g \subseteq \mathcal{Z}$, $\tilde{\mathcal{U}}^g \subseteq \mathcal{Z}$, and $d_{\mathcal{H}}(\tilde{\mathcal{U}}_i^g, \tilde{\mathcal{U}}^g) \leq d_{\mathcal{H}}(\tilde{\mathcal{U}}_i, \tilde{\mathcal{U}})$. $\tilde{\mathcal{U}}_i^g$ and $\tilde{\mathcal{U}}^g$ are the induced distributions of \mathcal{U}_i and \mathcal{U} under \mathcal{F}^g , respectively. $\tilde{\mathcal{U}}_i$ and $\tilde{\mathcal{U}}$ are the induced distributions of \mathcal{U}_i and \mathcal{U} under \mathcal{F} , respectively. \mathcal{F} is the feature extraction function in the original FedAvg without DBE.

[FL]: HtFLlib: HtFL Library and Benchmark

- **Easy-to-use and extensible:** modify only two files to add a new algorithm
- **PFLlib compatible:** support all PFLlib's scenarios, datasets, tools, etc.
- **First & comprehensive:** 40 heterogeneous models, 3 modalities, 10 HtFL methods, etc.



[FL]: HtFL on Real-World Devices

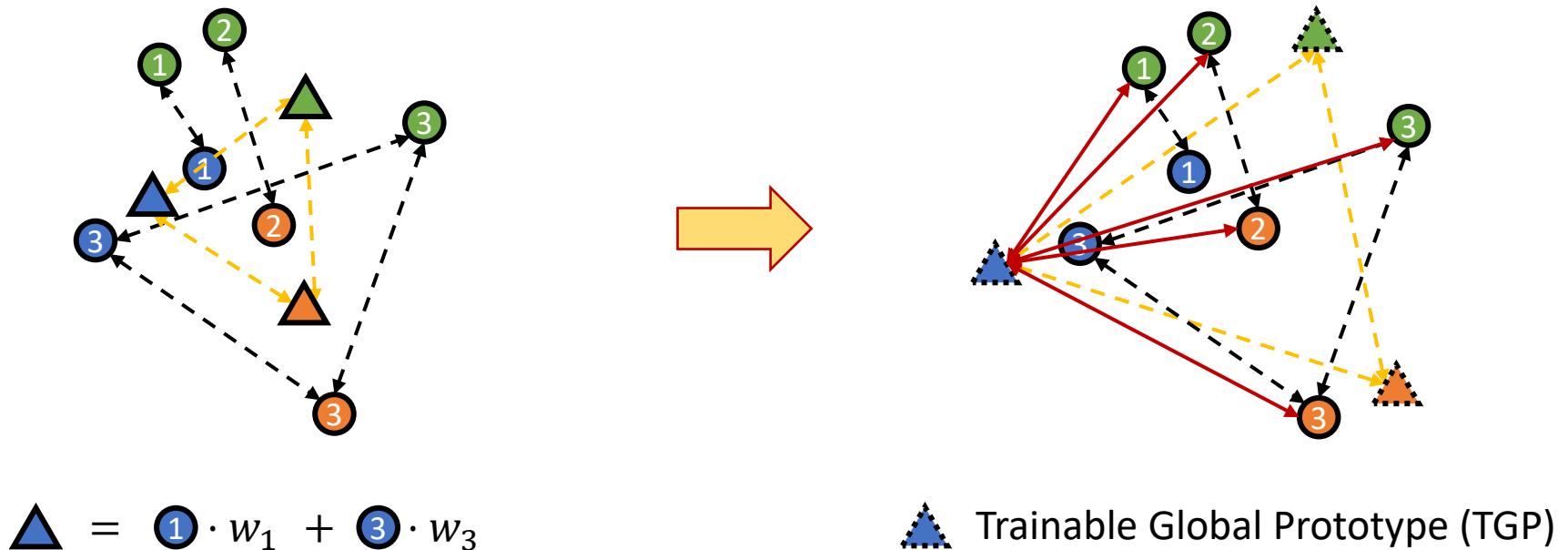
- **Real-world deployment, open-sourced**
 - + CoLEXT, + real-world datasets, + systematical metrics



- 28 Single Board Computers (SBC)
 - Orange Pi, LattePanda, Nvidia Jetson
- 20 Smartphones
 - Samsung, Xiaomi, Google Pixel, Asus ROG, One Plus
- High Voltage Power Meter
- Wired and wireless networking
- Workstation - FL Server

[FL]: FedTGP

- **Enlarge** the global prototype margin
- Ensure optimal feature quality across clients



[FL]: FedTGP

- Adaptive-margin-enhanced Contrastive Learning (ACL)
- ACL is **universal** and can be applied to other tasks

$$\min_{\hat{\mathcal{P}}} \sum_{c=1}^C \mathcal{L}_P^c,$$

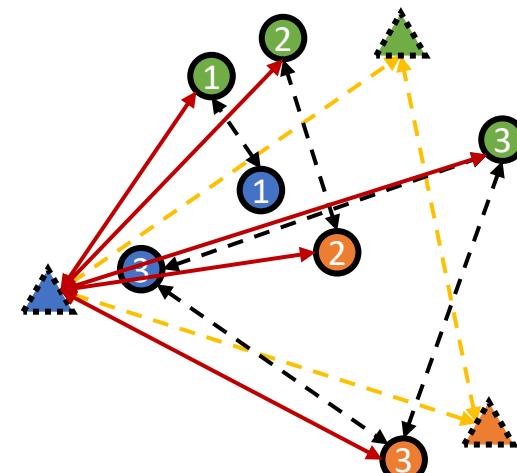
$$\mathcal{L}_P^c = \sum_{i \in \mathcal{I}^t} -\log \frac{e^{-(\phi(P_i^c, \hat{P}^c) + \delta(t))}}{e^{-(\phi(P_i^c, \hat{P}^c) + \delta(t))} + \sum_{c'} e^{-\phi(P_i^c, \hat{P}^{c'})}}$$

$$\delta(t) = \min(\max_{c \in [C], c' \in [C], c \neq c'} \phi(Q_t^c, Q_t^{c'}), \tau),$$

$$Q_t^c = \frac{1}{|\mathcal{P}_t^c|} \sum_{i \in \mathcal{I}^t} P_i^c, \forall c \in [C]$$

τ is a margin threshold

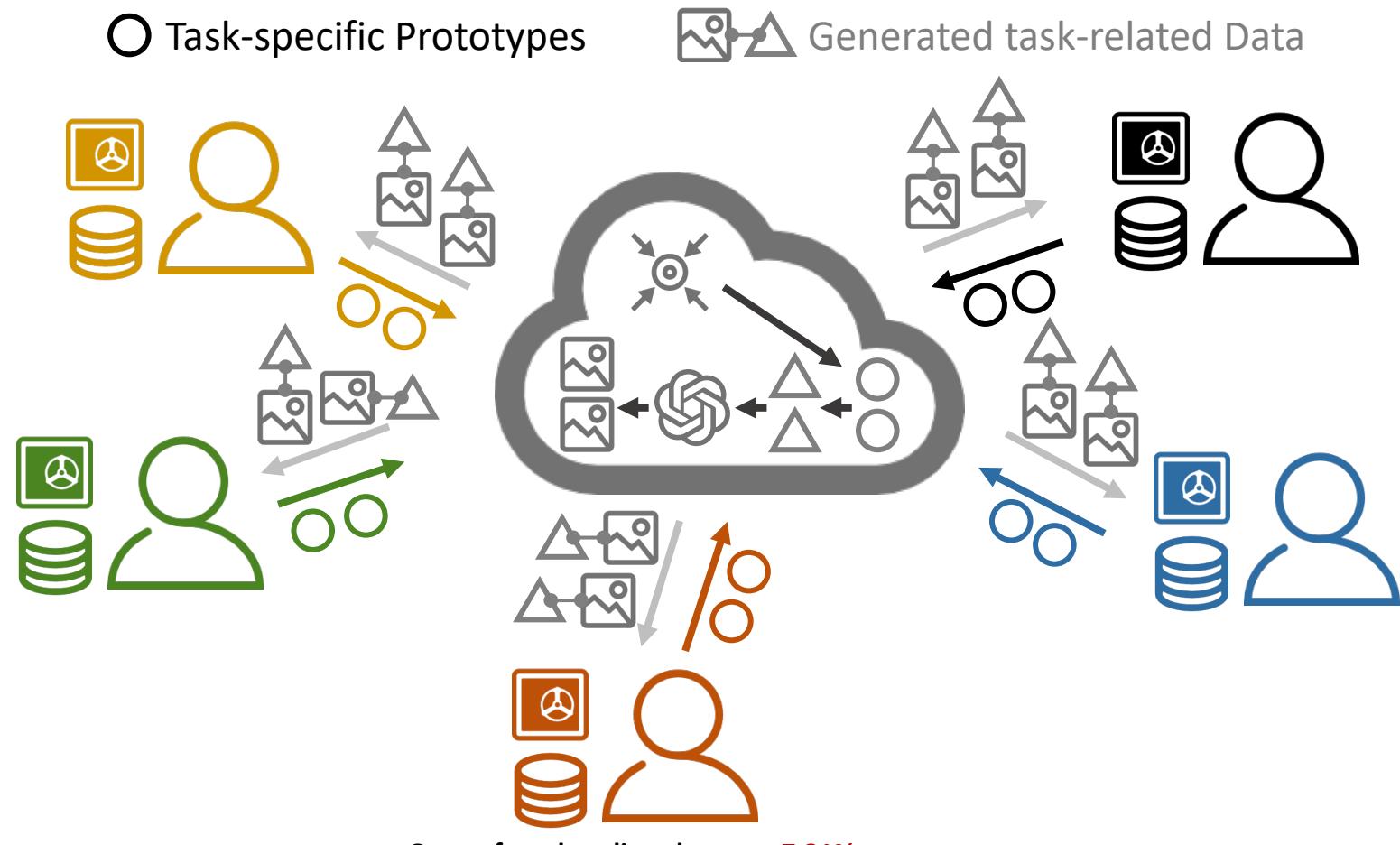
maximum cluster margin



- ▲ \hat{P}^c : A TGP of class c
- ▲ $\hat{\mathcal{P}}$: All TGP
- P_i^c : A prototype of class c from client i

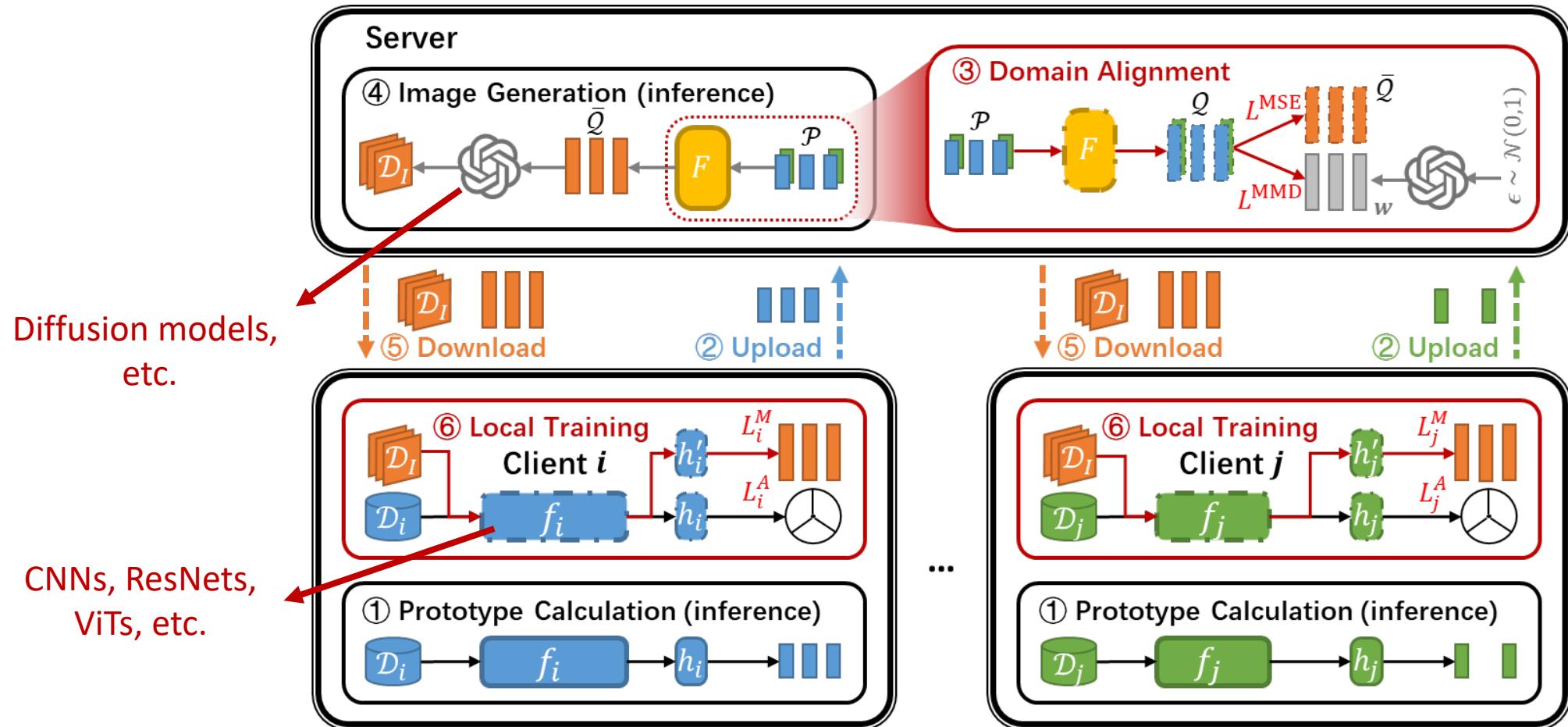
[FL]: FedKTL

- Transfer **common knowledge** from the pre-trained generators to clients
- Obtain **task-specific knowledge** from other clients



[FL]: FedKTL

- We need to Align small models' feature space with the generative model's for the transfer
- Transfer global knowledge using an additional supervised local task



[FL]: FedKTL

- FedKTL can **adapt to various generators** pre-trained on different datasets
- **No semantic limitations** on the generated images relative to the clients' data.



(a) Client #1



(b) AFHQv2



(c) Benches



(d) FFHQ-U



(e) WikiArt

Generators pre-trained on different image datasets

	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$
AFHQv2	26.82 ± 0.32	27.05 ± 0.26	26.32 ± 0.52
Bench	27.71 ± 0.25	28.36 ± 0.42	27.56 ± 0.50
FFHQ-U	27.28 ± 0.23	27.21 ± 0.35	26.59 ± 0.47
WikiArt	27.37 ± 0.51	27.48 ± 0.33	27.30 ± 0.15



Synthetic Data Generation (SDG)

- Given a **prompt**, with or without **data examples**,
- AI generates a dataset that **aligns with the user's request**.
 - Focusing on special domains (e.g., code, medicine, industry, etc.)



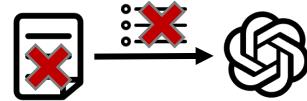
[SDG]: Existing general large models

- **Problem:** Existing general large models cannot fit specific domains:

- Fine-tuning
 - **Costly** for large model training, **data scarcity**
- Few-shot in-context learning (ICL)
 - **Privacy issue**, effortful **prompt engineering**
- Zero-shot ICL + selection
 - **Costly** for large amount data generating, effortful **prompt engineering**



Fine-tuning



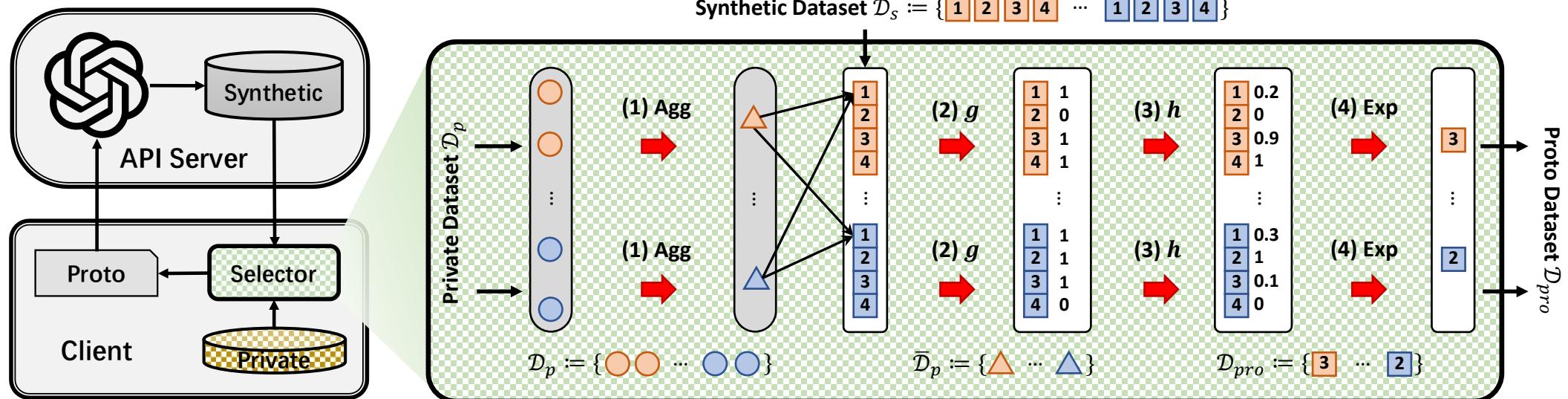
Few-shot ICL



Zero-shot ICL

[SDG]: PCEvolve

- **Solution:** You only need to provide a few samples — we'll **evolve** an entire dataset for you,
- While **protecting privacy** via differential privacy (DP) using Exponential mechanism





[SDG]: PCEvolve

- **COVIDx**: chest X-ray images for COVID-19
- **Came17**: tumor tissue patches from breast cancer metastases
- **KVASIR-f**: endoscopic images for gastrointestinal abnormal findings detection
- **MVAD-I**: leather surface anomaly detection

Top-1 accuracy (%) on four specialized datasets

	COVIDx	Came17	KVASIR-f	MVAD-I
Init	49.34	50.47	33.43	33.33
RF	50.01	54.82	34.66	48.17
GCap	50.86	55.77	32.66	27.33
B	50.42	54.41	32.57	43.21
LE	50.02	55.44	35.51	27.93
DPImg	49.14	61.06	33.35	37.03
PE	59.63	63.66	48.88	57.41
PE-EM	57.60	63.34	43.01	50.06
PCEvolve-GM	56.91	62.63	43.55	55.56
PCEvolve	64.04	69.10	50.95	59.26

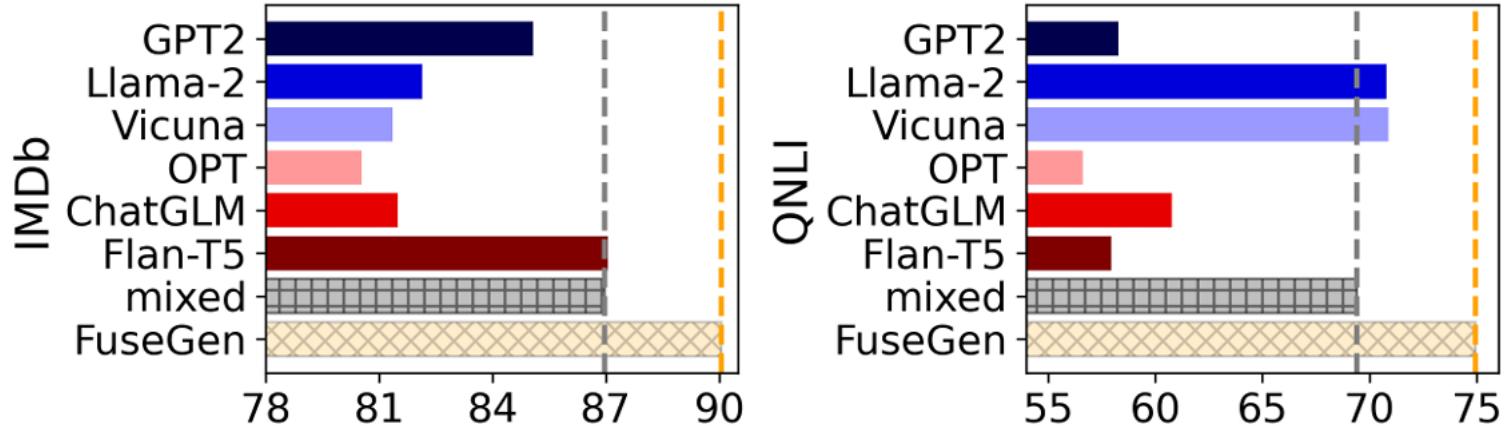
[SDG]: PCEvolve



Generated leather surface images w.r.t. MVAD-I for industry anomaly detection. The three rows show normal images, cut defects, and droplet defects. “Initial” denotes API-generated images using just the prompt. “Private” denotes the real images from MVAD-I.

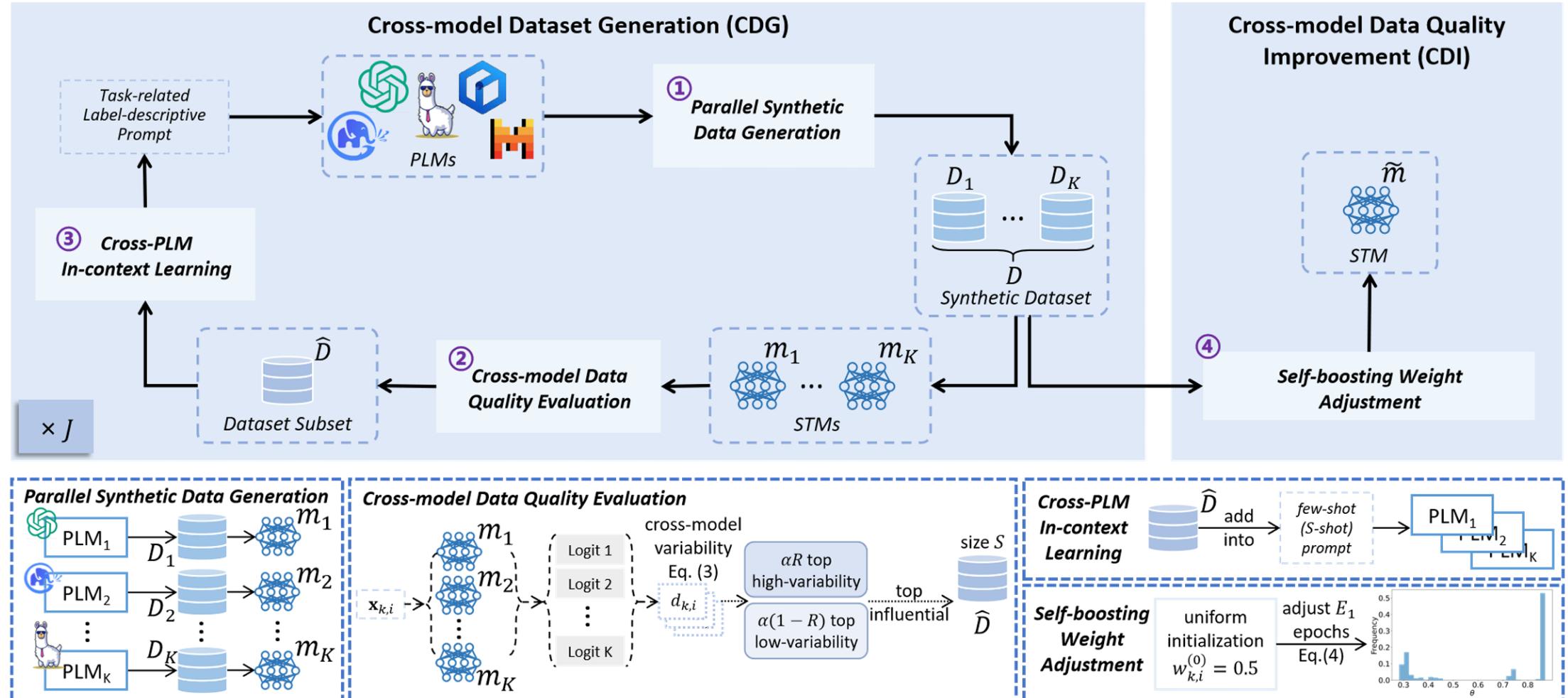
[SDG]: FuseGen

- **Problem:** Pre-trained Language Models (PLMs) have **different tastes** for specific domains
- **Solution:** We merge models' outputs to create **diverse datasets** through **evolution**



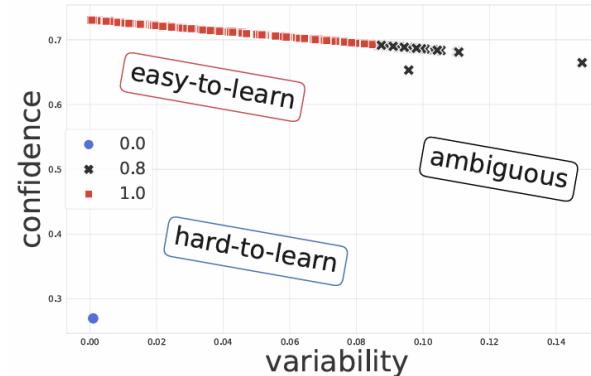
[SDG]: FuseGen

- We consider downstream models' feedback as reward signals for evolution

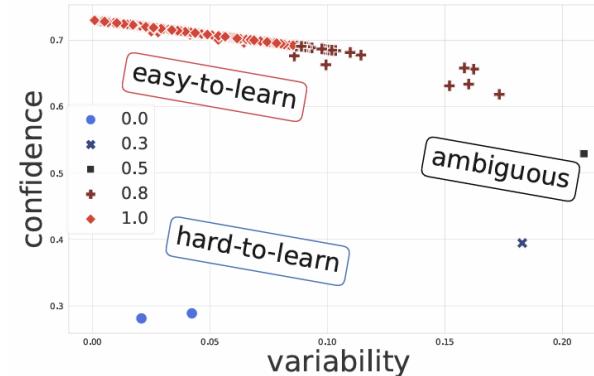


[SDG]: FuseGen

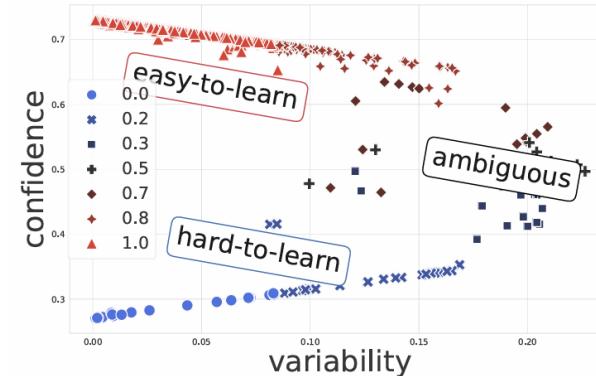
- Synthetic dataset cartography



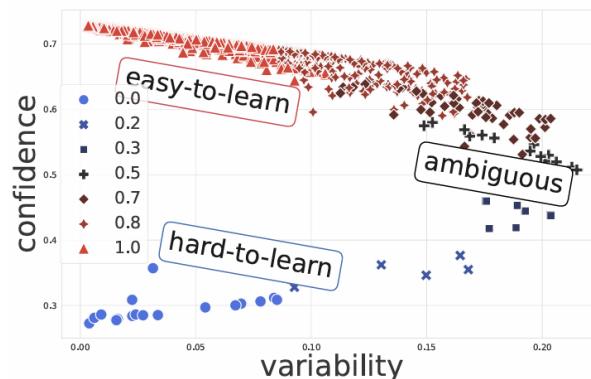
(a) Llama-2 ZeroGen $K = 1$ (84.23)



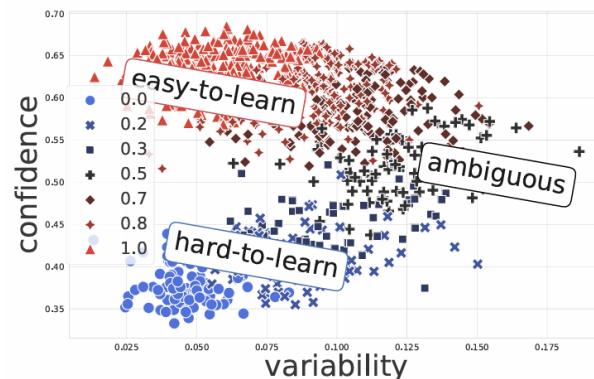
(b) Llama-2 ProGen $K = 1$ (84.24)



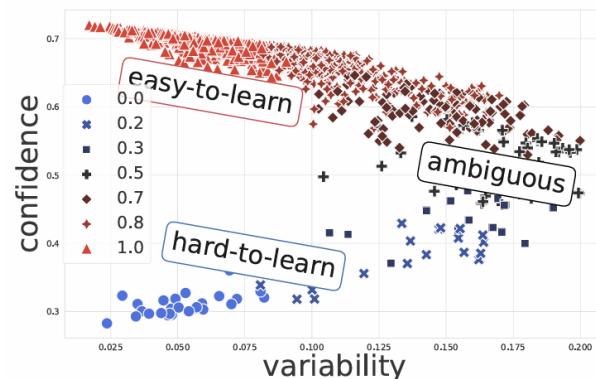
(c) Llama-2 Ours $K = 6$ (86.60)



(d) Flan-T5 ZeroGen $K = 1$ (88.18)



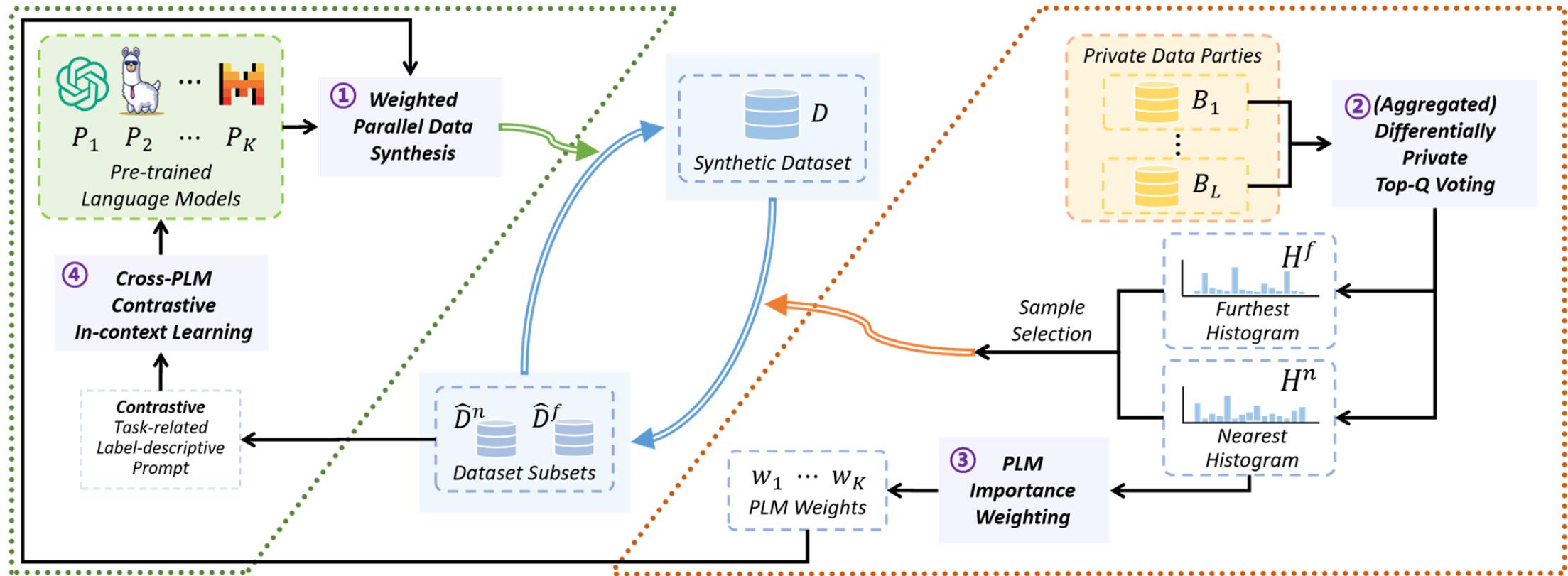
(e) Flan-T5 ProGen $K = 1$ (85.80)



(f) Flan-T5 Ours $K = 6$ (88.73)

[SDG]: WASP

- **Problem:** Using only positive examples for evolution lacks diversity for domain dataset generation
- **Solution:** Contrastive voting-based positive and negative sample selection for evolution





[SDG]: WASP

Model	High-quality Sample Text (Examples)	Label
GPT-2	"This movie isn't for everyone, but it's worth a look. But if you have to be in the dark to enjoy it, just buy the soundtrack."	positive
GPT-2	"is that it has some semi-interesting world-building, but the information it offers is grossly insufficient to allow one to progress in the story. The central characters have little to do with the world beyond playing around with their toys and talking to themselves."	negative
Llama-2	"Incredible film! I laughed, I cried, and I left the theater feeling inspired. Every performance was top-notch and the direction was flawless. A true masterpiece that will be remembered for years to come. 10/10."	positive
Llama-2	"This film was a colossal disappointment. I found myself glancing at my wrist every 10 minutes, eagerly anticipating its conclusion. The narrative was intricate and challenging to comprehend, and the personas were underdeveloped. I had no emotional investment in their plight. The performances were lacking, and the dialogue was trite and unconvincing. In summary, I would not advise viewing this motion picture to any individual."	negative
Vicuna	"A movie for people who want to see an action-packed adventure that will keep them on the edge of their seats."	positive
Vicuna	"Felt like a missed opportunity. Glacial pacing and shallow characters. The premise was intriguing but the execution fell flat. Avoid this one."	negative
OPT	"-This film is enjoyable to watch because of its well-developed storyline and most of all its storyline that i can relate to and would like to have. This film also has good acting performances and the songs help the film a lot. It's not a typical love story kind of a film, but it does have lots of interesting messages to give. Very well-done! Definitely recommend this film!"	positive
OPT	"Quote: The first sentence means nothing to me. Without context the first two sentences also mean nothing."	negative
ChatGLM3	"Attention getter, visually interesting and outstanding acting, the story of an American citizen that is in Mexico and gets involved in a murder is a good movie."	positive
ChatGLM3	"This model is an implementable model based on the off-the-shelf models in the ChatGLM family. All rights reserved. Information contained herein is confidential and proprietary to Shanghai Jiao Tong University."	positive

Summary



Data-Centric Model Adaptation: Balancing generalization and specialization on new data from a data-driven perspective

① **Recommender:** **Exploiting** long- and short-term user behavior data



② **Federated learning:** **Exploiting** generalization and specialization in heterogeneous data across devices



③ **Synthetic dataset generation:** **Generating** vertical domain data based on few private data and large models



④ **Code LLM reinforcement learning:** **Exploiting** self-generated data to reduce coding errors

Feel free to contact me!

Home page: <https://github.com/TsingZ0>



Thanks!