Data-Centric Model Optimization

- **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 Optimization**
- **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 Optimization:
Balancing generalization and specialization 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

↓

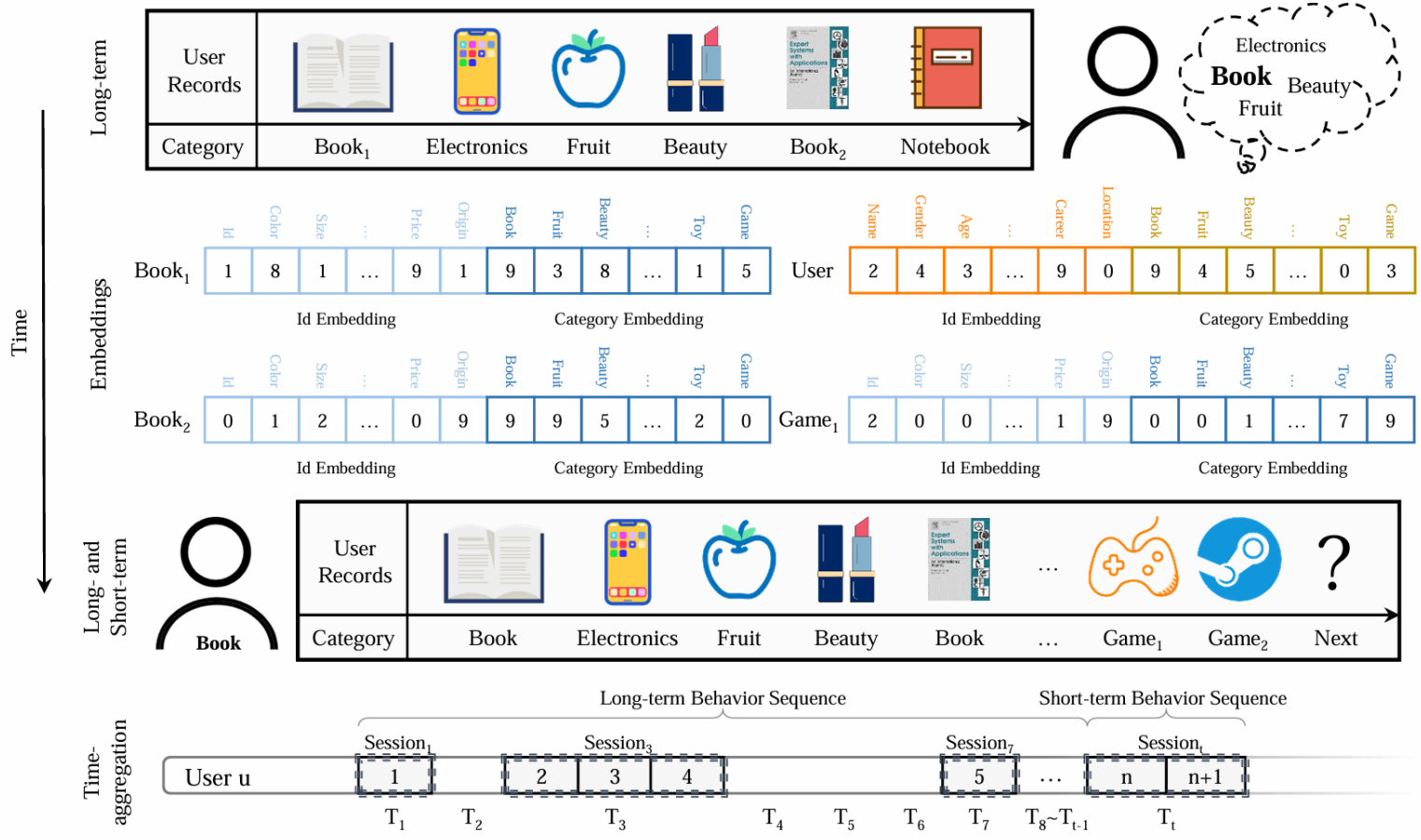
③**Synthetic dataset generation: Generating** vertical domain data based on few private data and large models

↓

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

Recommender System (RS)

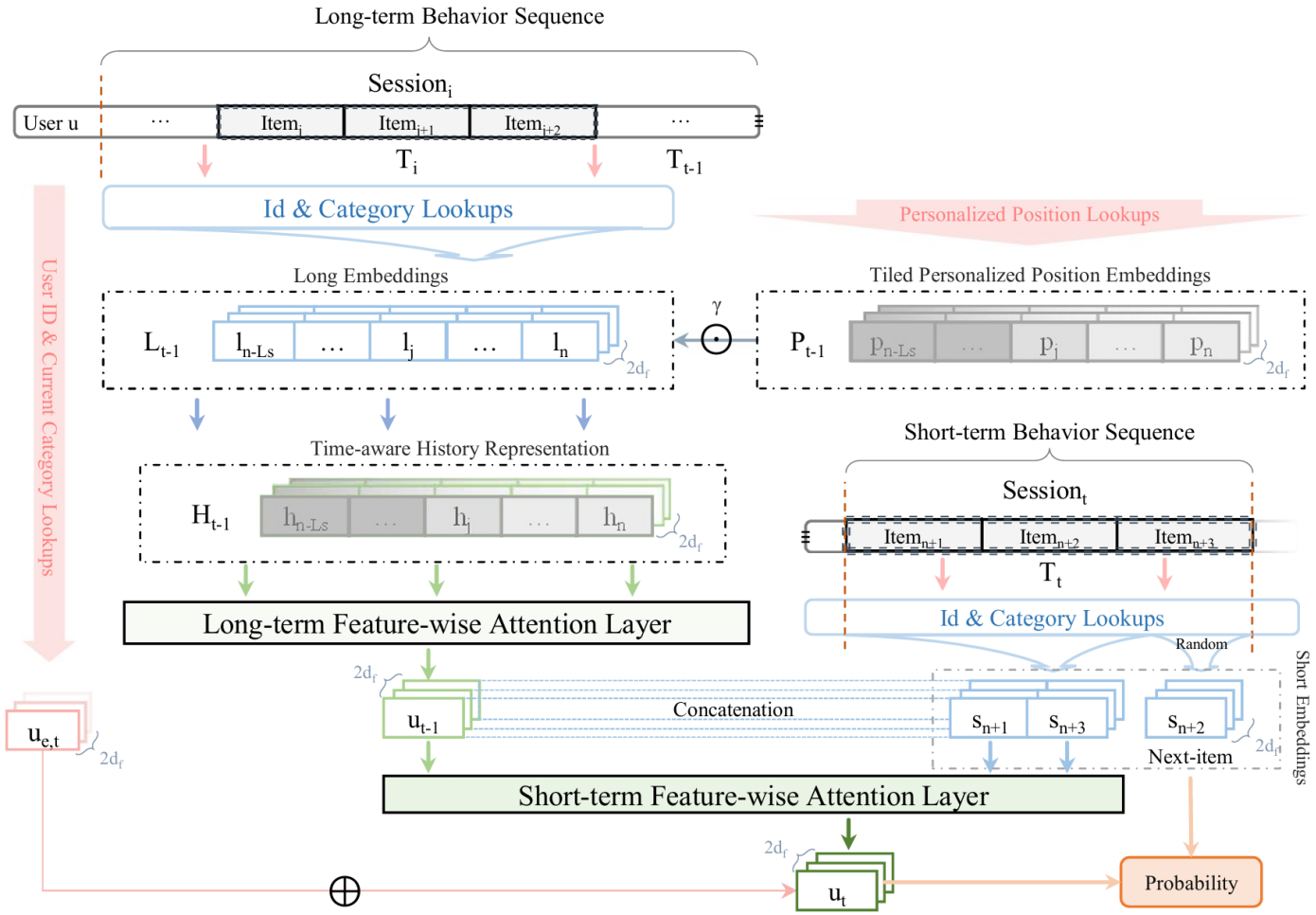
- **Problem:** Users have **personalized taste for time** in the behavior data besides others
- **Solution:** Time-aware Long- and Short-term Attention Network for Next-item Recommendation



Accepted by Neurocomputing 2021 **Outperform baselines by up to 20.79%** **Recently used by Amazon AWS**

[RS]: TLSAN

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



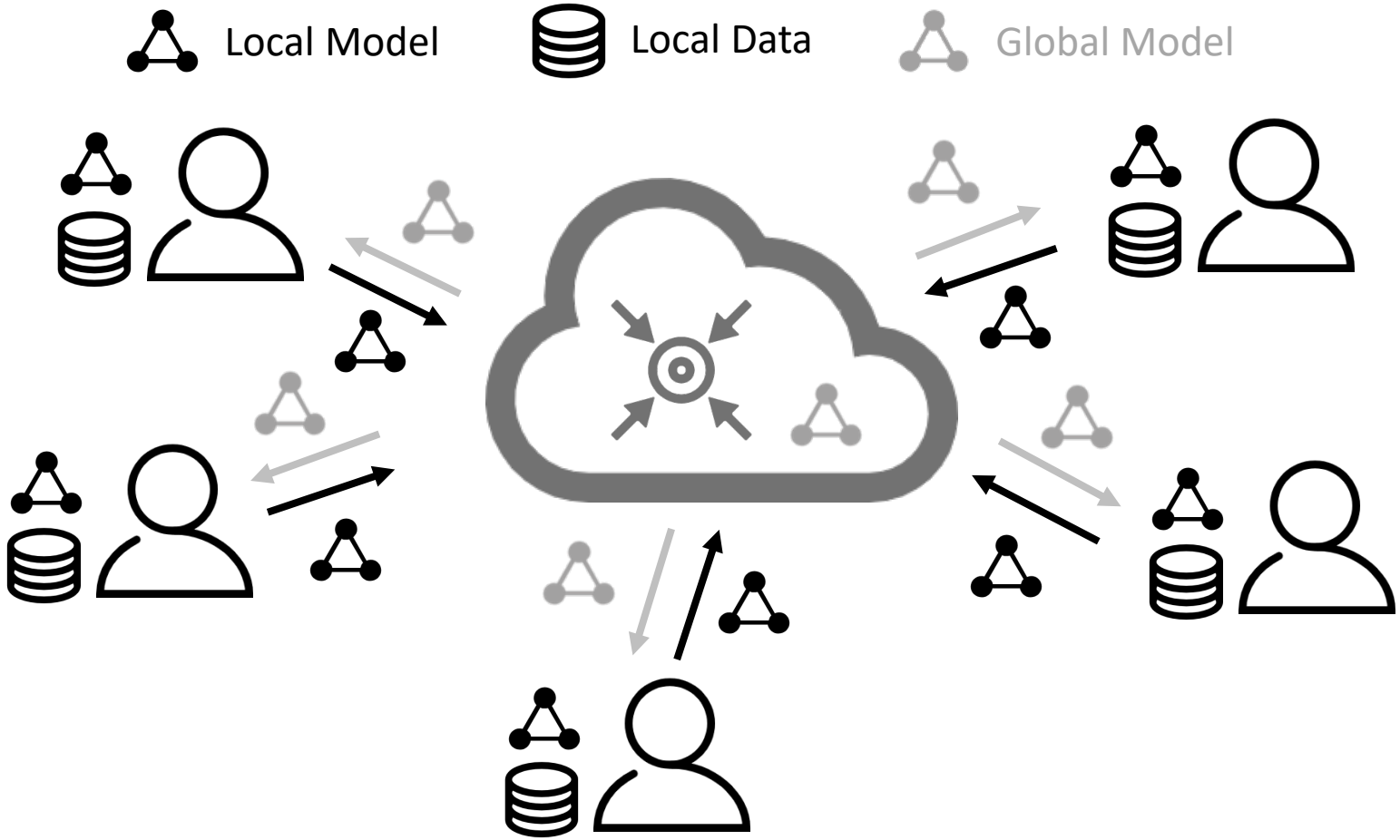
Accepted by Neurocomputing 2021

Outperform baselines by up to 20.79%

Recently used by Amazon AWS

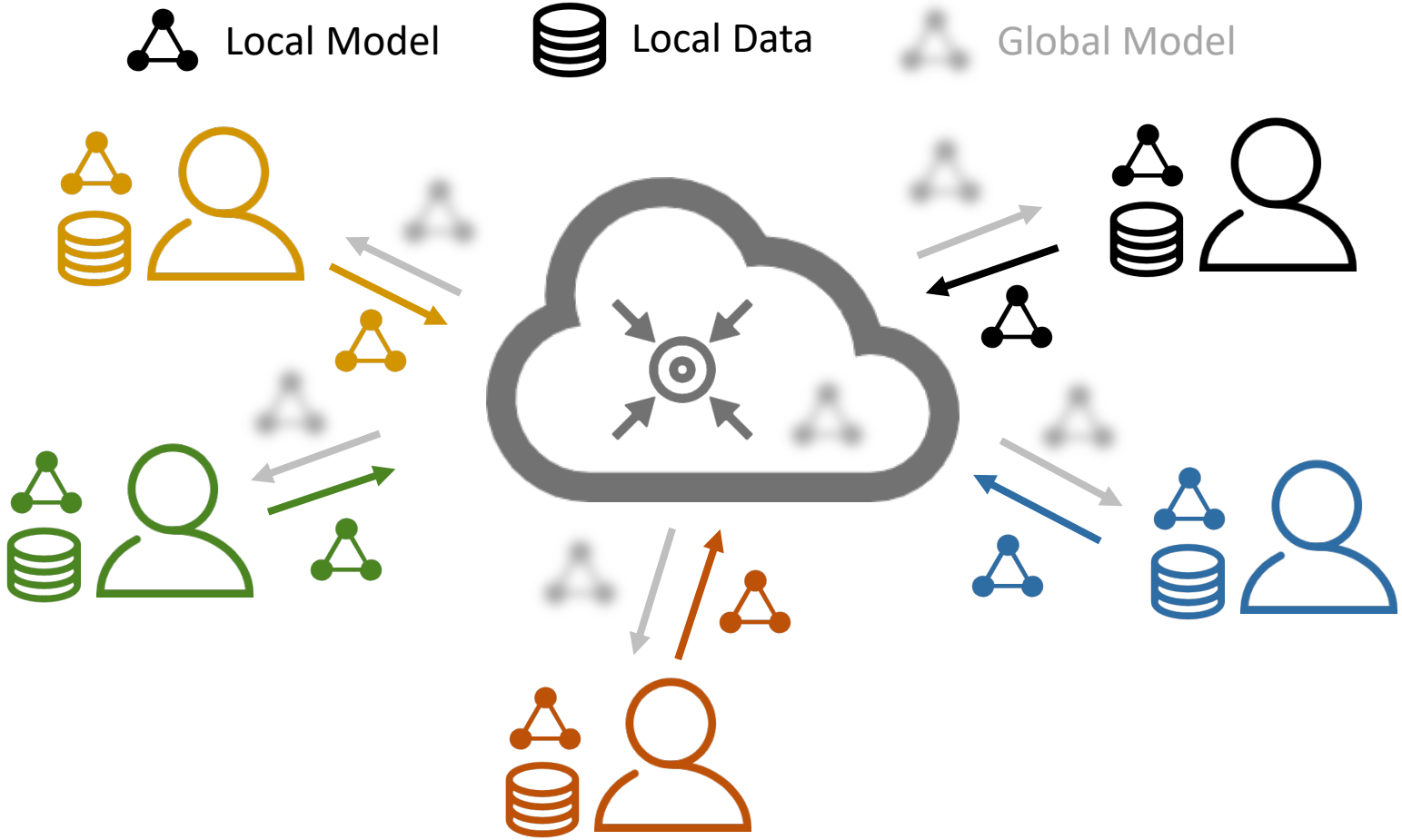
Federated Learning (FL)

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



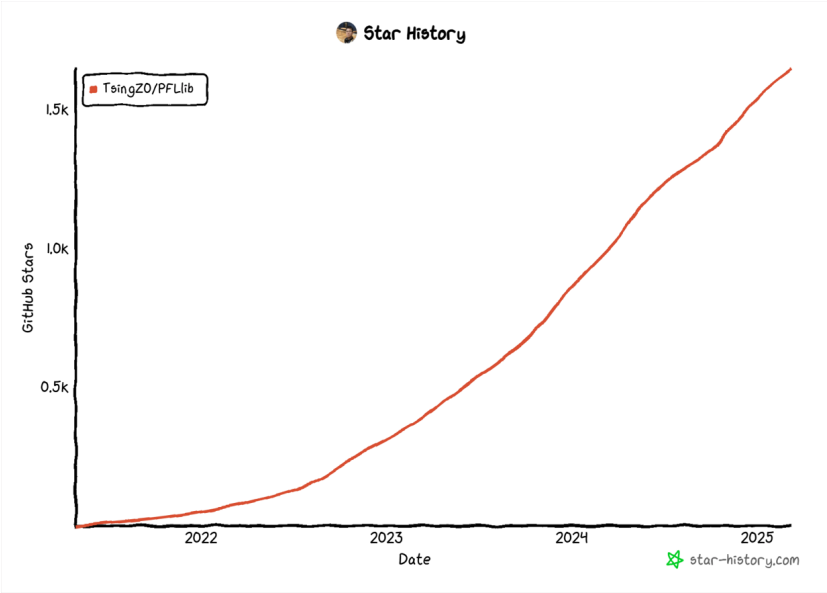
[FL]: Data Heterogeneity

- **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:**



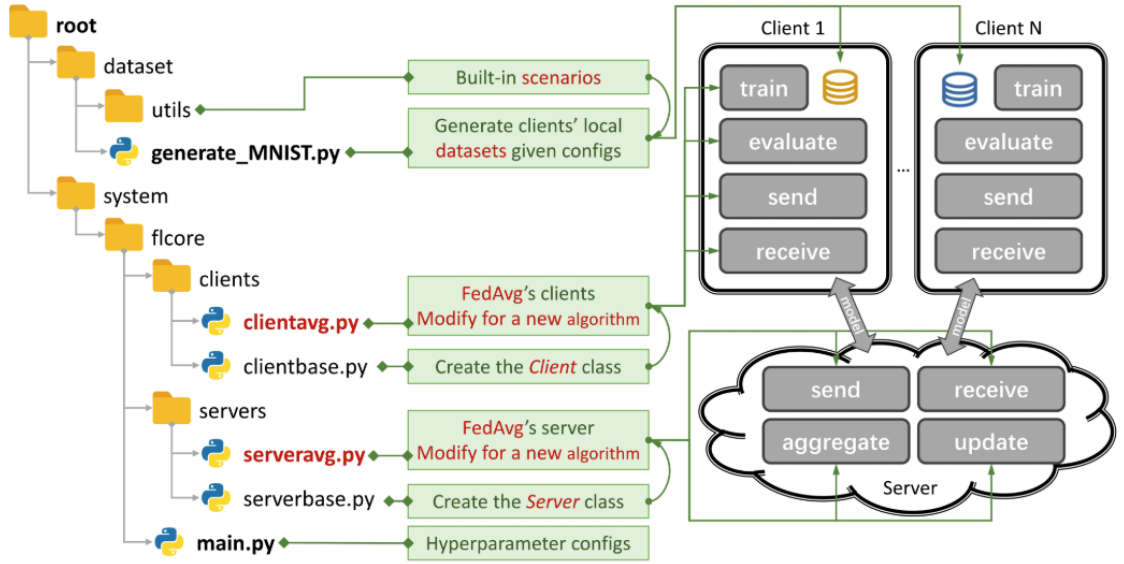
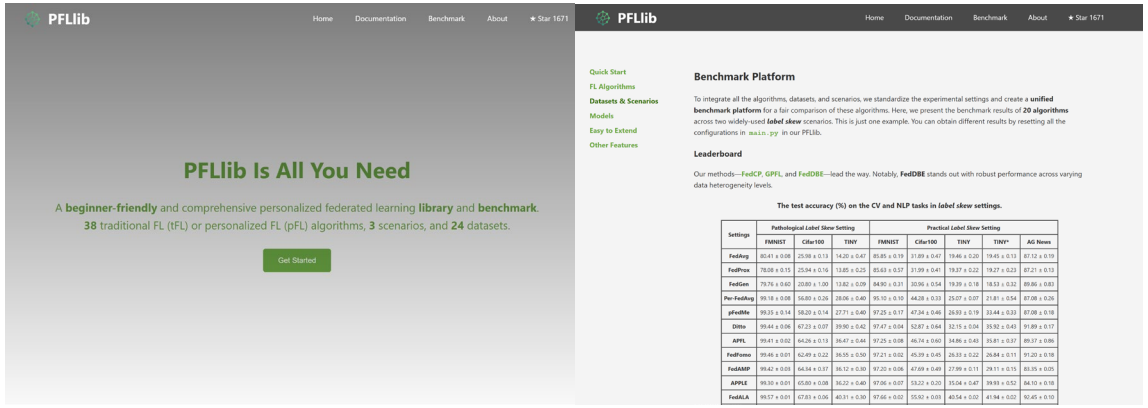


Figure 1: An Example for FedAvg. You can create a scenario using `generate_DATA.py` and run an algorithm using `main.py`, `clientNAME.py`, and `serverNAME.py`. For a new algorithm, you only need to add new features in `clientNAME.py` and `serverNAME.py`.

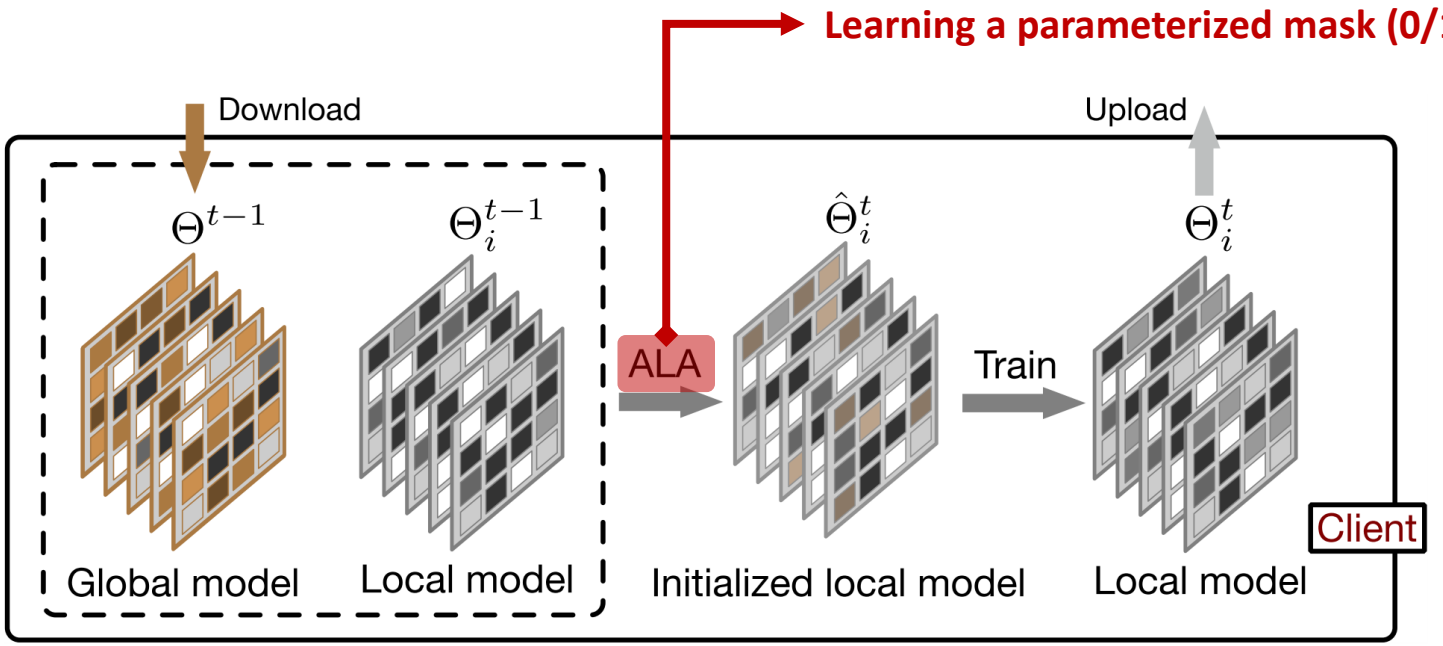


Accepted by JMLR 2025 (WAIC Outstanding Paper Top40)

https://github.com/TsingZ0/PFLlib and https://www.pfllib.com/

[FL]: FedALA

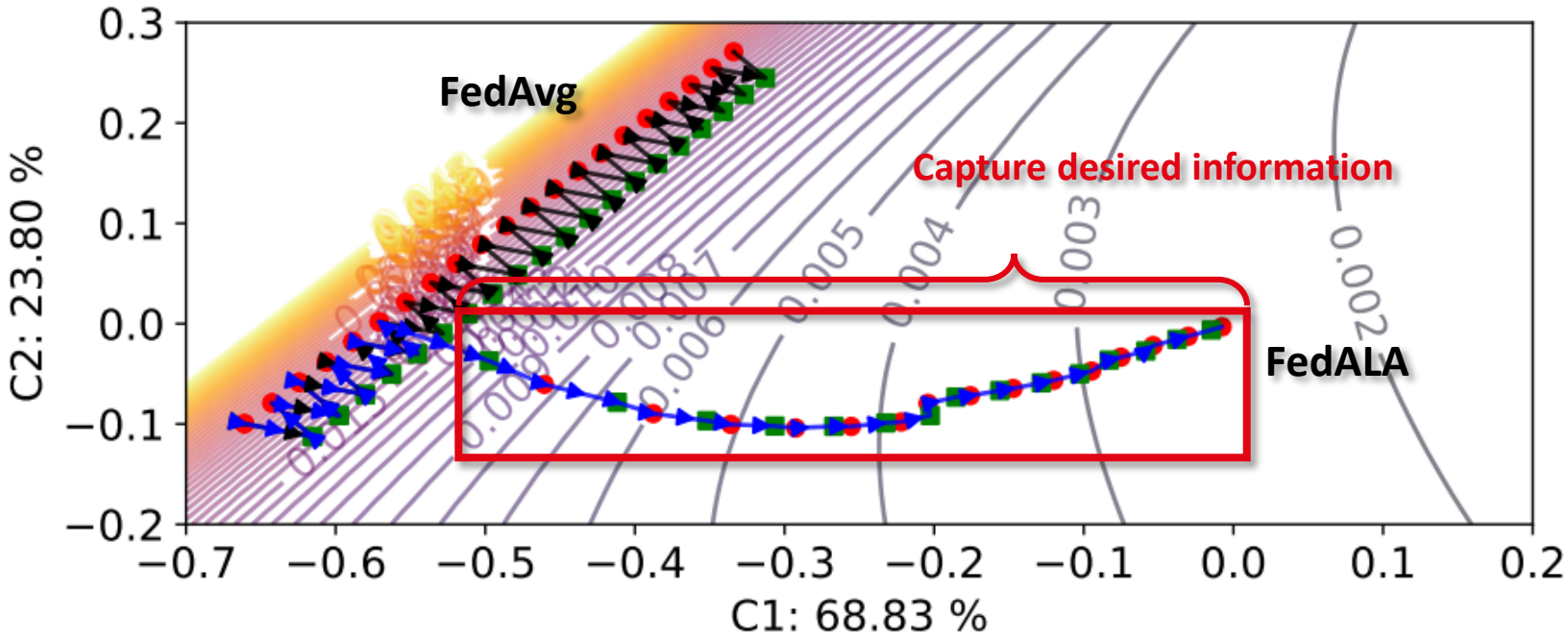
- Extract each client's **desired parameters** from the global model to facilitate local training
- **Adaptively aggregate** the parameters in the global and local model in each communication round



Workflow on the client in one iteration

[FL]: FedALA

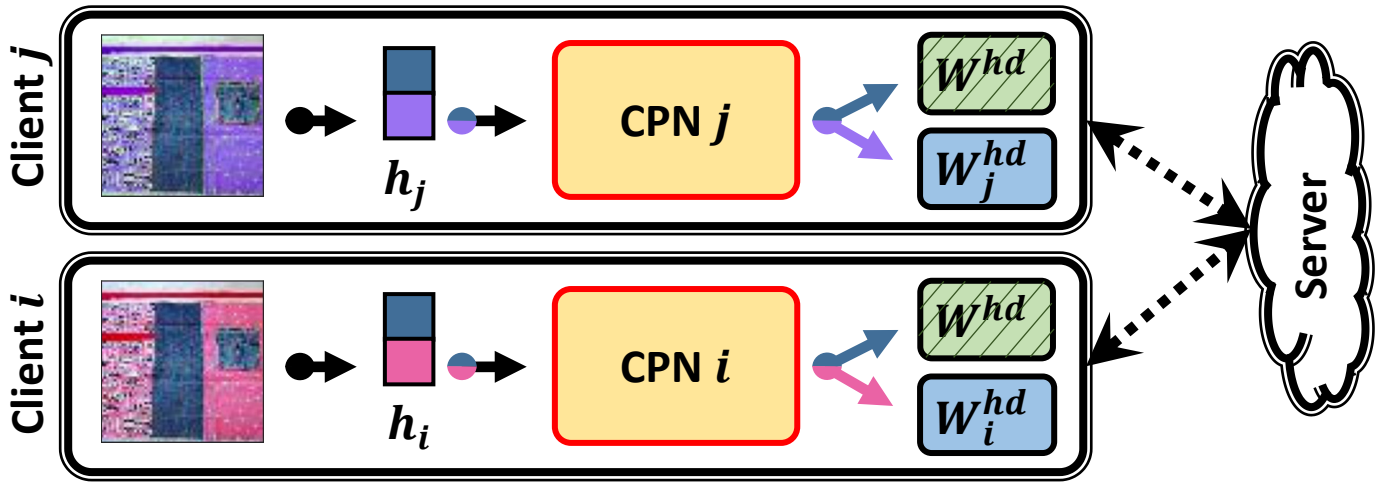
- Learning trajectory on one client: **FedAvg** vs. **FedALA**
- Activate **ALA** in the subsequent iterations



2D visualization of local learning trajectory

[FL]: FedCP

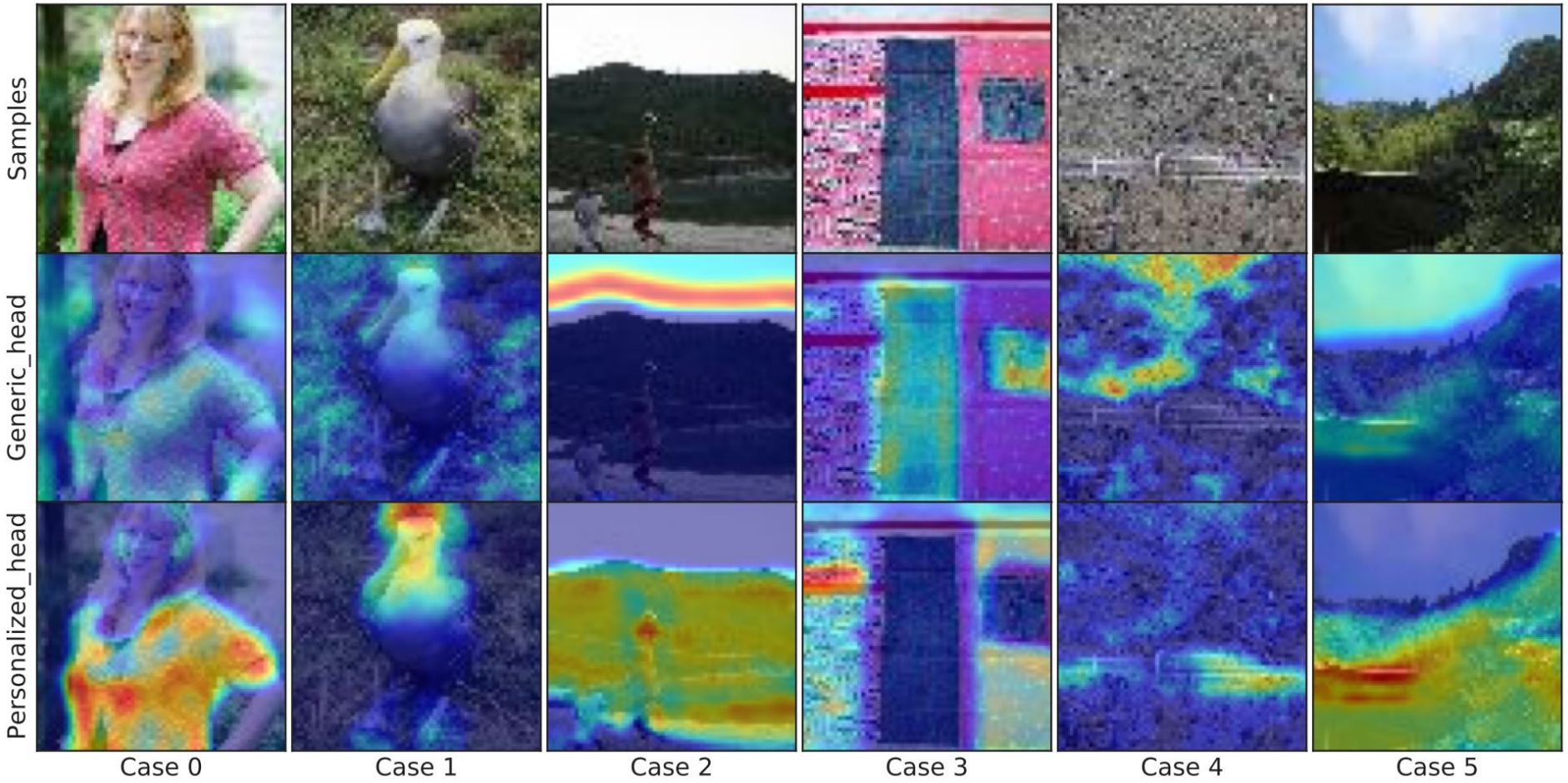
- We **separate feature information** via an auxiliary **Conditional Policy Network (CPN).**
  - Detach generic and personalized information
  - Sample-specific separation
  - Lightweight (e.g., 4.67% parameters of ResNet-18)



- 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

Accepted by KDD 2023

Outperform baselines by up to 6.69%

[FL]: GPFL

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

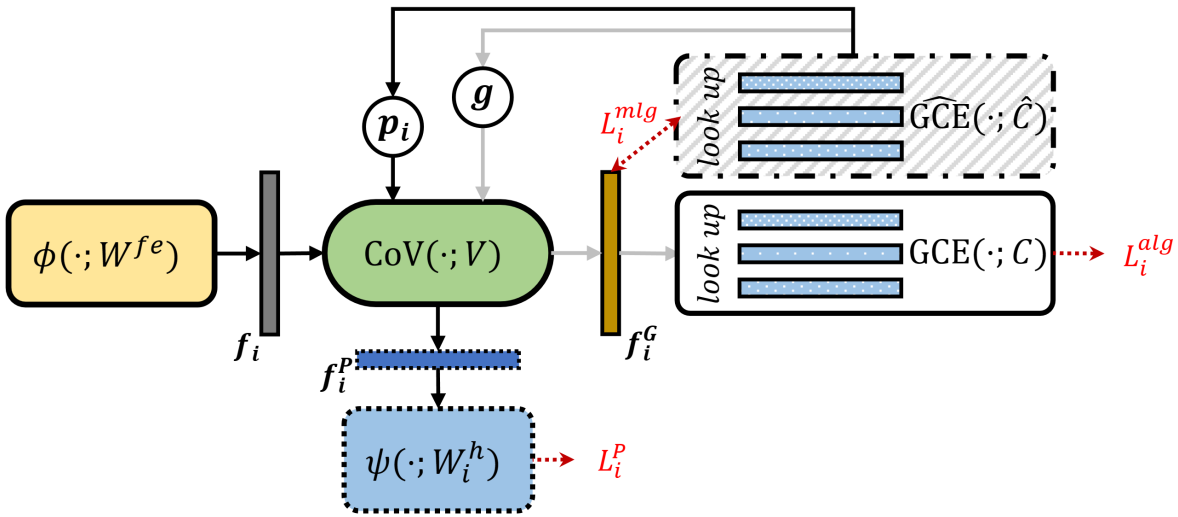
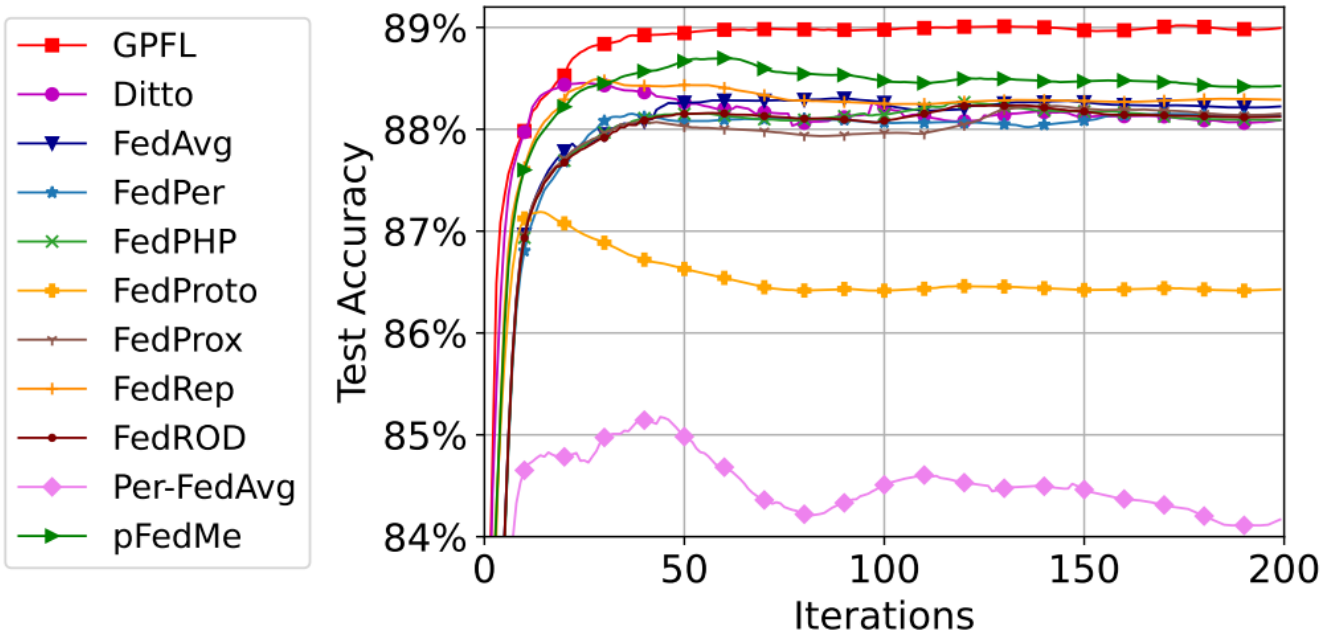


Illustration of client modules and data flow between them

[FL]: GPFL

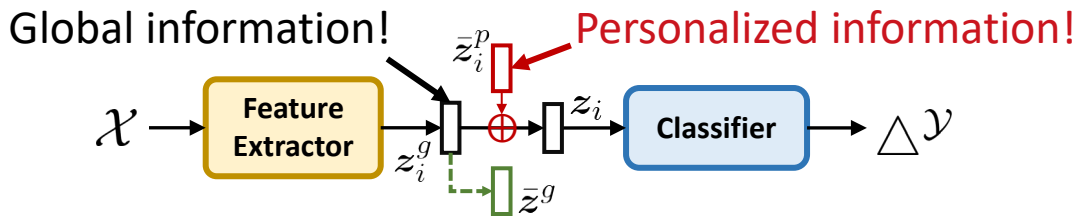
- Relieve the **widely existed** overfitting issue in pFL



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)

Accepted by NeurIPS 2023

Outperform baselines by up to 11.36%

[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 $1 - \delta$:*

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

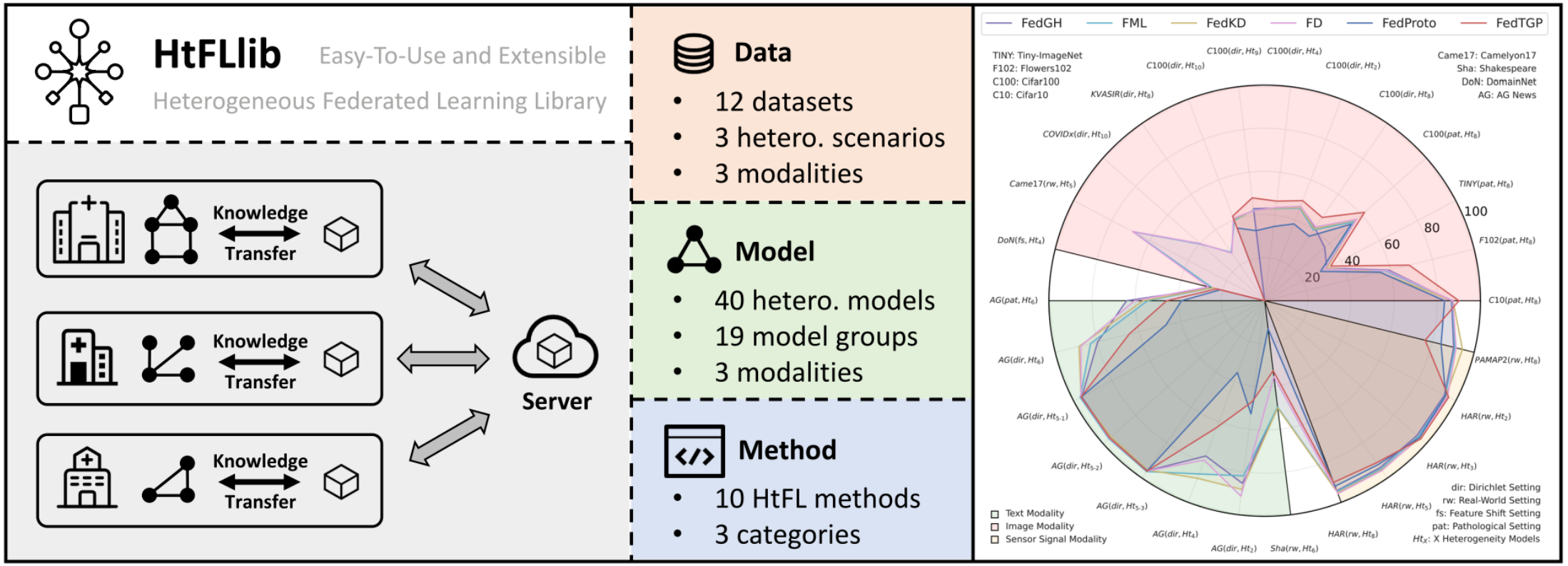
where $\mathcal{L}_{\hat{\mathcal{D}}_i}$ is the empirical loss on $\mathcal{D}_i$, e is the base of the natural logarithm, and $d_{\mathcal{H}}(\cdot, \cdot)$ is the $\mathcal{H}$-divergence between two distributions. $\lambda_i := \min_{h^g} \mathcal{L}_{\mathcal{D}}(h^g) + \mathcal{L}_{\mathcal{D}_i}(h^g)$, $\tilde{\mathcal{U}}_i^g \subseteq \mathcal{Z}$, $\tilde{\mathcal{U}}^g \subseteq \mathcal{Z}$, 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.

Accepted by NeurIPS 2023

Outperform baselines by up to 11.36%

[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.



Accepted by KDD 2025 (Best Paper Runner UP)

https://github.com/TsingZ0/HtFLlib

[FL]: HtFL on Real-World Devices

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

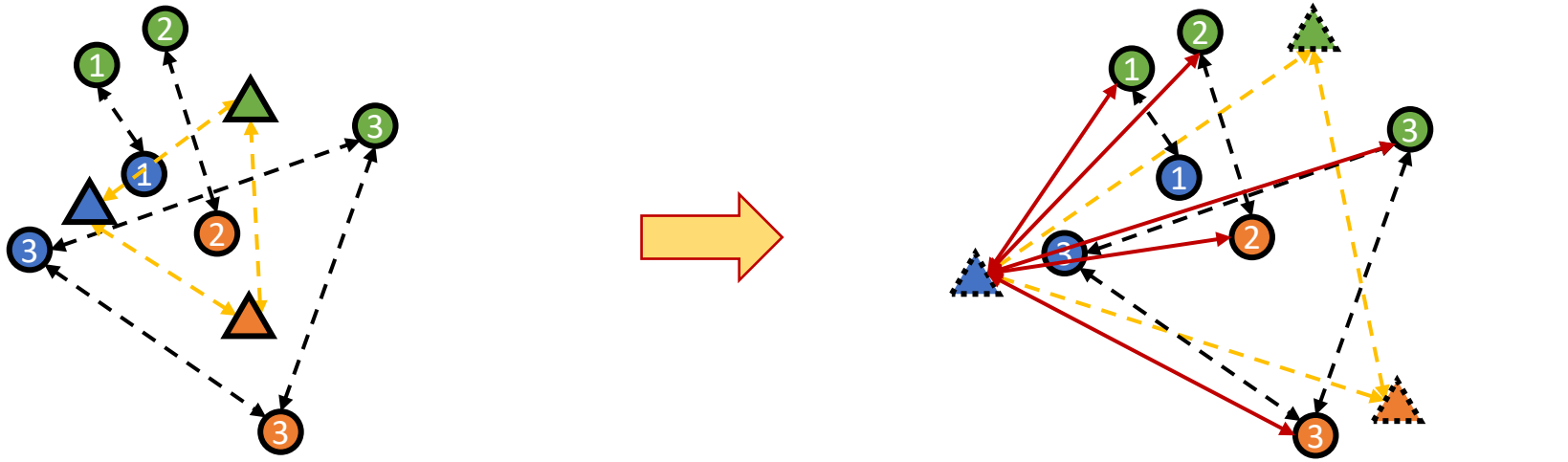
- 28 Single Board Computers (SBC)
  - Orange Pi, LattePanda, Nvidia Jetson
- 20 Smartphones
  - Samsung, Xiaomi, Google Pixel, Asus ROG, One Plus
- High Voltage Power Meter
- Wired and wireless networking
- Workstation - FL Server

Collaborating with KAUST

https://github.com/TsingZ0/HtFL-OnDevice

[FL]: FedTGP

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



$\blacktriangle = ① \cdot w_1 + ③ \cdot w_3$

▲ Trainable Global Prototype (TGP)

[FL]: FedTGP

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

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

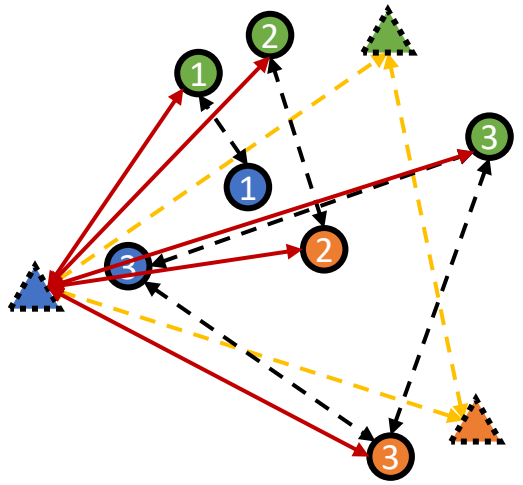
$$\mathcal{L}_P^c = \sum_{i \in \mathcal{I}^t} -\log \frac{e^{-(\phi(P_i^c, \hat{P}^c) + \delta(t))}}{e^{-(\phi(P_i^c, \hat{P}^c) + \delta(t))} + \sum_{c'} e^{-\phi(P_i^c, \hat{P}^{c'})}}$$

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

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

maximum cluster margin

τ is a margin threshold



$\hat{P}^c$: A TGP of class c

$\hat{\mathcal{P}}$: All TGP

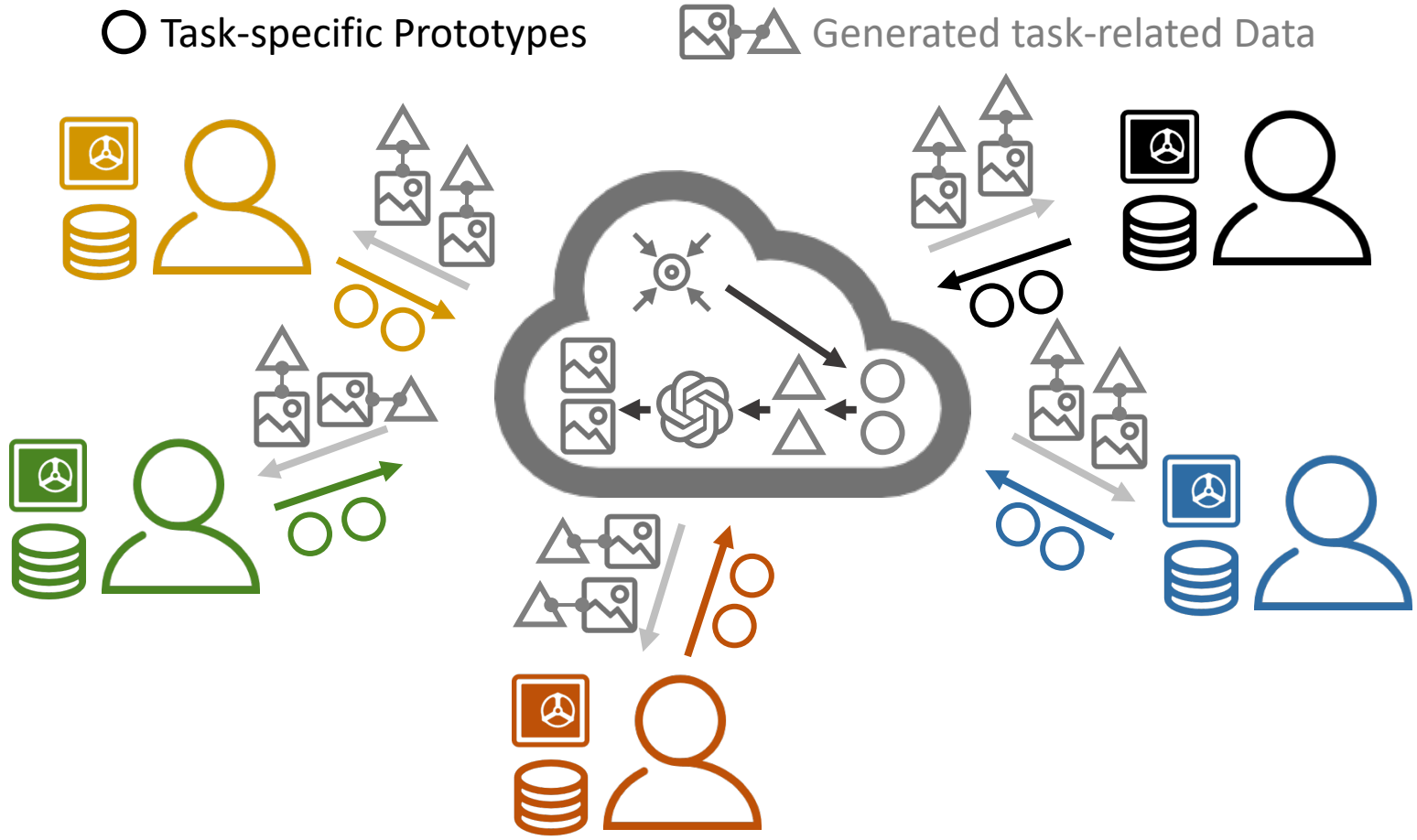
P_i^c: A prototype of class c from client i

Accepted by AAAI 2024

Outperform baselines by up to 9.08%, and achieve up to an 18.96% improvement over FedProto

[FL]: FedKTL

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



Accepted by CVPR 2024

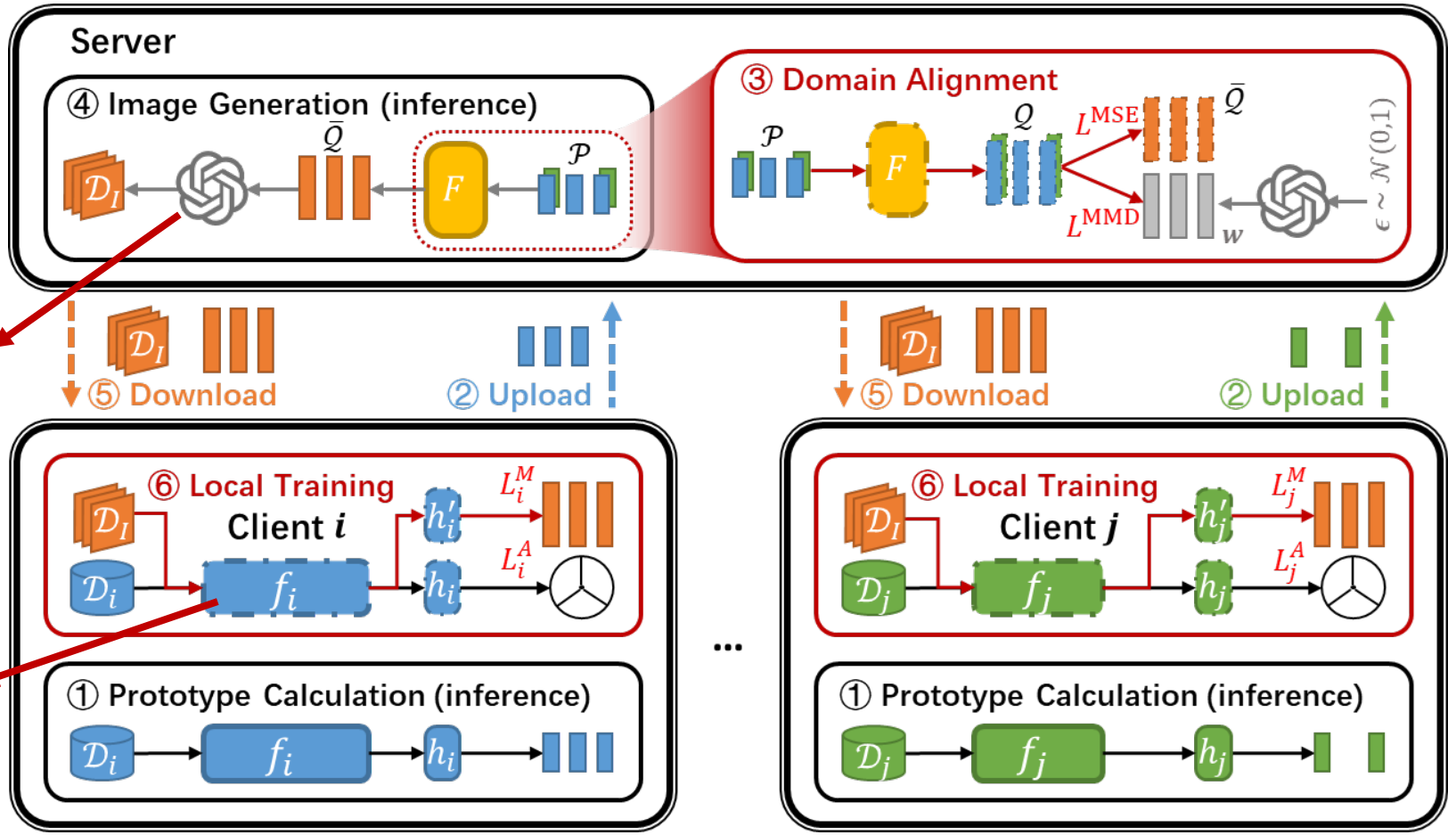
Outperform baselines by up to 7.31%

[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**

Diffusion models, etc.

CNNs, ResNets, ViTs, etc.



Accepted by CVPR 2024

Outperform baselines by up to 7.31%

[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

Accepted by CVPR 2024

Outperform baselines by up to 7.31%

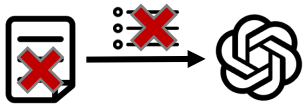
Synthetic Data Generation (SDG)

- 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

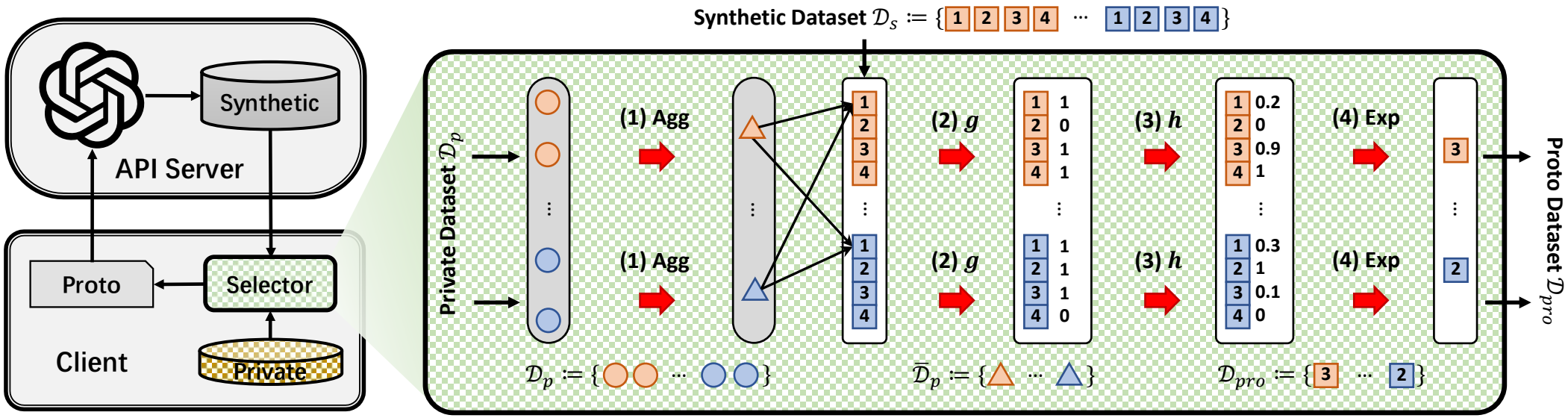


Few-shot ICL

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-l**: leather surface anomaly detection

Top-1 accuracy (%) on four specialized datasets

	COVIDx	Came17	KVASIR-f	MVAD-l
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**

Accepted by ICML 2025 (spotlight)

[SDG]: PCEvolve

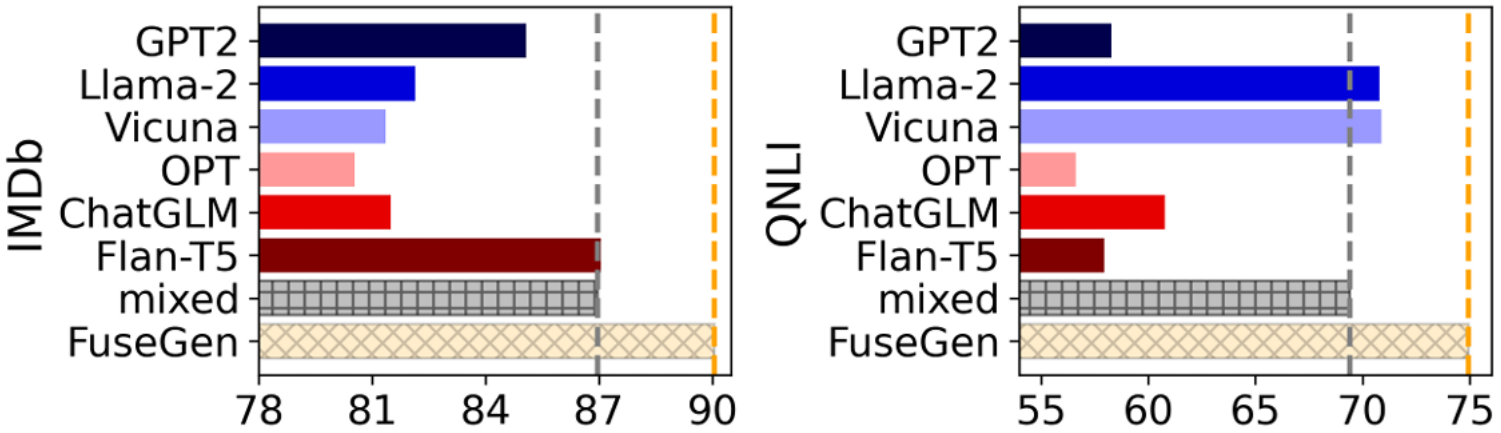
(a) Initial (b) PE (c) PCEvolve (d) Private

Generated leather surface images w.r.t. MVAD-l 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-l.

Accepted by ICML 2025 (spotlight)

[SDG]: FuseGen

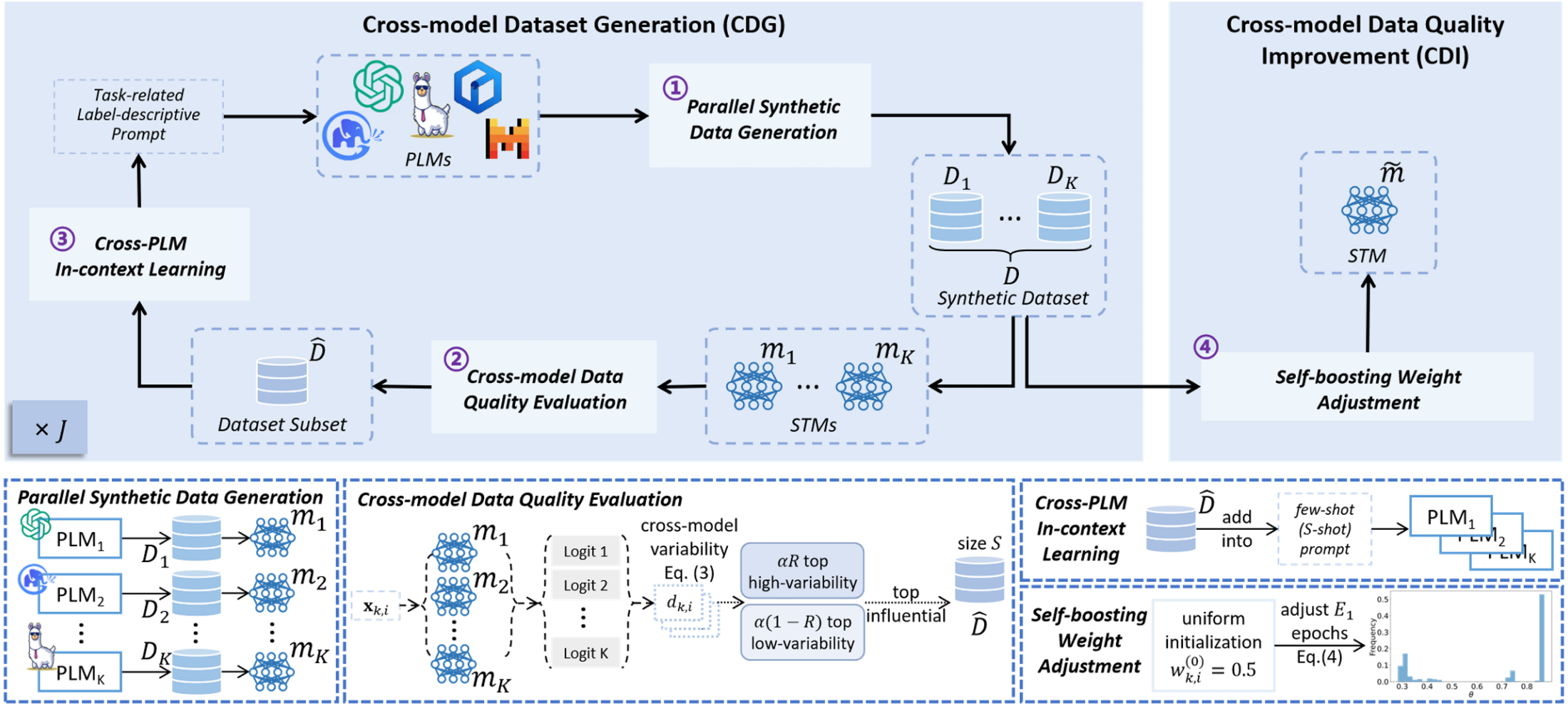
- **Problem:** Pre-trained Language Models (PLMs) have **different tastes** for specific domains
- **Solution:** We **merge models' outputs** to create **diverse datasets** through **evolution**



Accepted by EMNLP 2024

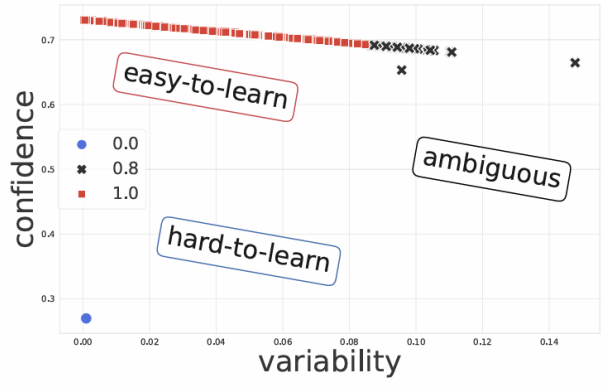
[SDG]: FuseGen

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

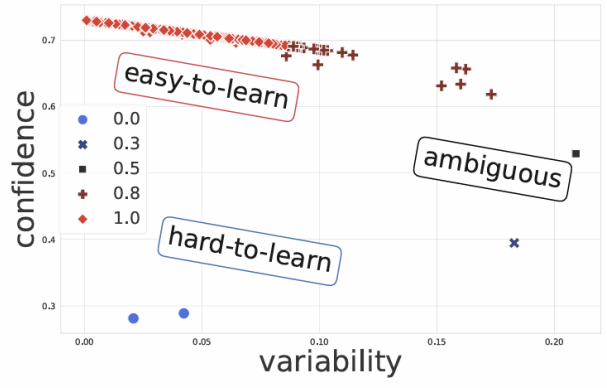


[SDG]: FuseGen

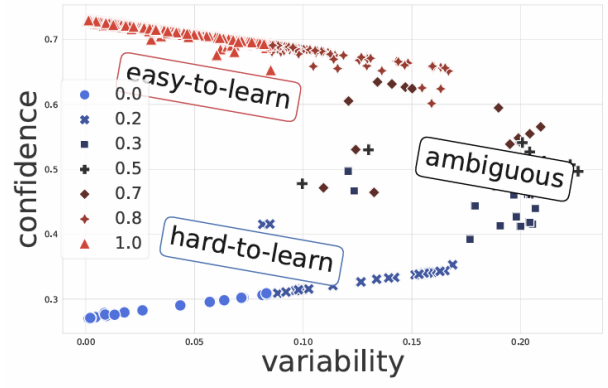
- Synthetic dataset cartography



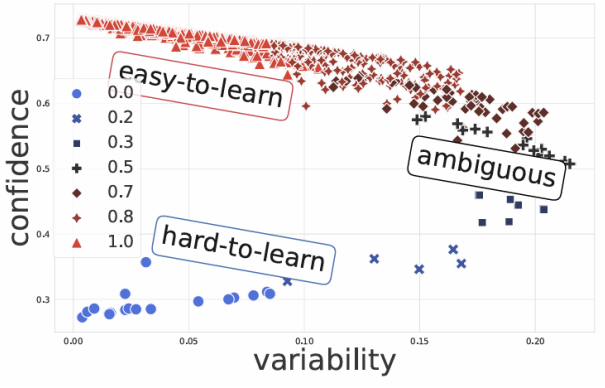
(a) Llama-2 ZeroGen $K = 1$ (84.23)



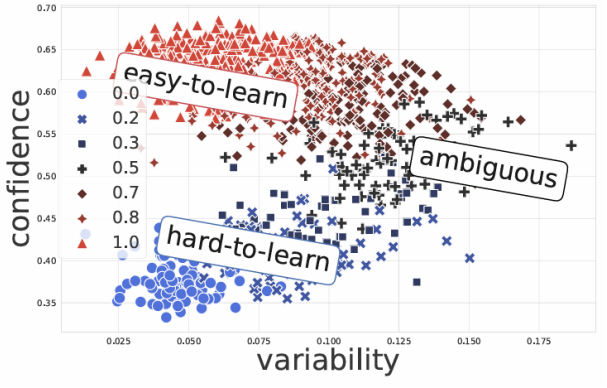
(b) Llama-2 ProGen $K = 1$ (84.24)



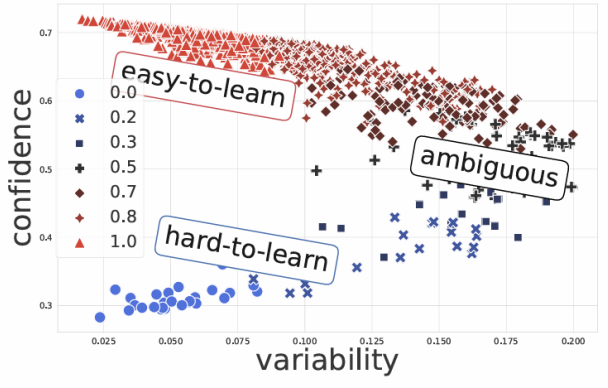
(c) Llama-2 Ours $K = 6$ (86.60)



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



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

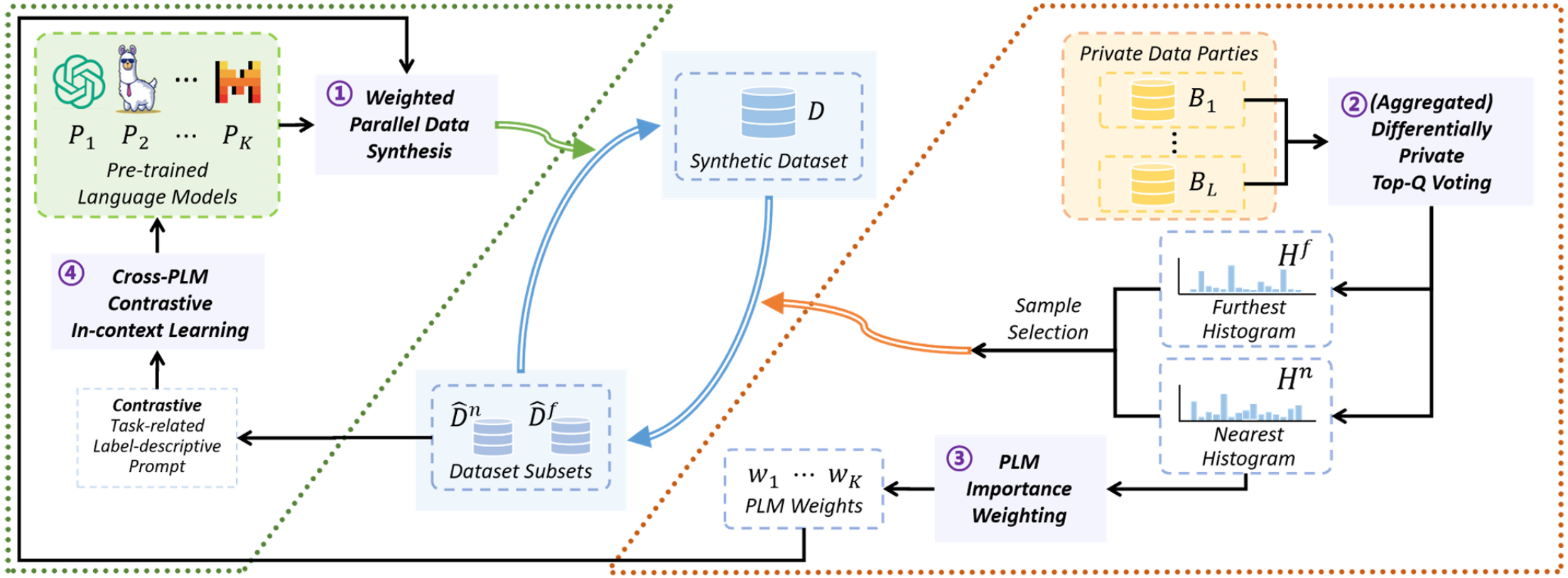


(f) Flan-T5 Ours $K = 6$ (88.73)

Accepted by EMNLP 2024

[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**



Accepted by ICML 2025

[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 it's 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 movie is an unwatchable mess, basically a rip-off of that scene in Dr. Strangelove where the U.S. Air Force gets into	negative

Summary

Data-Centric Model Optimization:
Balancing generalization and specialization 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!