Summary of My Research

- **Name**: Jianqing Zhang
- **Ph.D.**: Shanghai Jiao Tong & Tsinghua University
- **Visiting**: Hong Kong Polytechnic University
- **Home Page**: github.com/TsingZ0
- **E-mail**: tsingz@sjtu.edu.cn
- **LinkedIn**: www.linkedin.com/in/tsingz/
- **X**: @TsingZ00

Overview

- **Research interests**: **Data-Centric Vertical Model Optimization**
- **Research fields**: *Code LLM, Synthetic Data Generation, Distributed Learning, Recommender System*
- **Outstanding advantages**: **Research capabilities** and **engineering experience**
- **Publications**: **9** first-author top-tier conference/journal papers
  - Stage ① AAAI'23 (**oral**), KDD'23, ICCV'23, NeurIPS'23, JMLR'25
  - Stage ② AAAI'24, CVPR'24, KDD'25
  - Stage ③ EMNLP'24, ICML'25, ICML'25 (**spotlight**)
- **Open-sourced projects** (initiator):
  - EvolveGen, PFLlib (**1800**+ stars, **300**+ forks), HtFLlib, HtFLlib on Device, FL-IoT, etc.
- **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
- **Projects**: ① Cross-hospital cancer recognition, ② Cross-province intelligent 12345 hotline model, ③ HtFL testbed on real-world devices, ④ Led 9-member team in building a distributed ML platform for data centers
- **Impact**: **700**+ citations, **30K**+ views across major media, well-recognized by IEEE/ACM Fellows
- **Intern**: ByteDance AML, Tsinghua AIR, KAUST SANDs lab, Tencent AI Code

总览

- **研究兴趣**: 以数据为中心的垂域模型优化
- **研究领域**: 代码大模型后训练、合成数据集生成、分布式机器学习、推荐系统
- **突出优势**: 科研能力 + 工程经验
- **论文发表**: **9** 篇一作顶会顶刊论文
  - 阶段① AAAI'23 (**oral**), KDD'23, ICCV'23, NeurIPS'23, JMLR'25
  - 阶段② AAAI'24, CVPR'24, KDD'25
  - 阶段③ EMNLP'24, ICML'25, ICML'25 (**spotlight**)
- **开源项目(发起人)**:
  - EvolveGen, PFLlib (**1800**+ stars, **300**+ forks), HtFLlib, HtFLlib on Device, FL-IoT, etc.
- **获奖**: 中国科协-博士青年人才托举 (中国人工智能学会, CAAI), 吴文俊人工智能荣誉博士, 博士生国家奖学金
- **落地项目**: ①与医院合作进行跨医院癌症相关研究、②为12345政务服务热线智能模型进行跨省份知识迁移、③在20+单片机上部署异构模型分布式训练、④带9人团队搭建分布式机器学习平台，交付给数据中心
- **影响力**: **700**+ 谷歌学术引用, **30K**+ 主流媒体曝光, 受到IEEE/ACM Fellows在CCF大会上对我工作的赞扬
- **实习交流**: 字节跳动 AML, 清华智能产业研究院, 阿卜杜拉国王科技大学 SANDs lab, 腾讯代码模型组

Systematical Research Trace

Data-Centric Vertical Model Optimization: Balancing generalization and specialization in vertical domains from a data-driven perspective

① **Recommender:** Hierarchical modeling of long- and short-term behaviors for next-item recommendation

↓

②**Distributed learning:** Generalization and specialization in heterogeneous data across distributed machines

↓

③**Synthetic dataset generation:** Evolution of large model–generated data in specialized domains

↓

④**Post-training of code LLM:** Dynamic sampling of code preference data in the code generation domain

系统性科研路径

以数据为中心的垂域模型优化：从数据出发，平衡垂直领域中模型的泛化和专业化能力

①推荐模型训练：下一项推荐领域，长期-短期行为数据的层次化建模

↓

②分布式训练：分布式设备上，异质数据中的共性和个性

↓

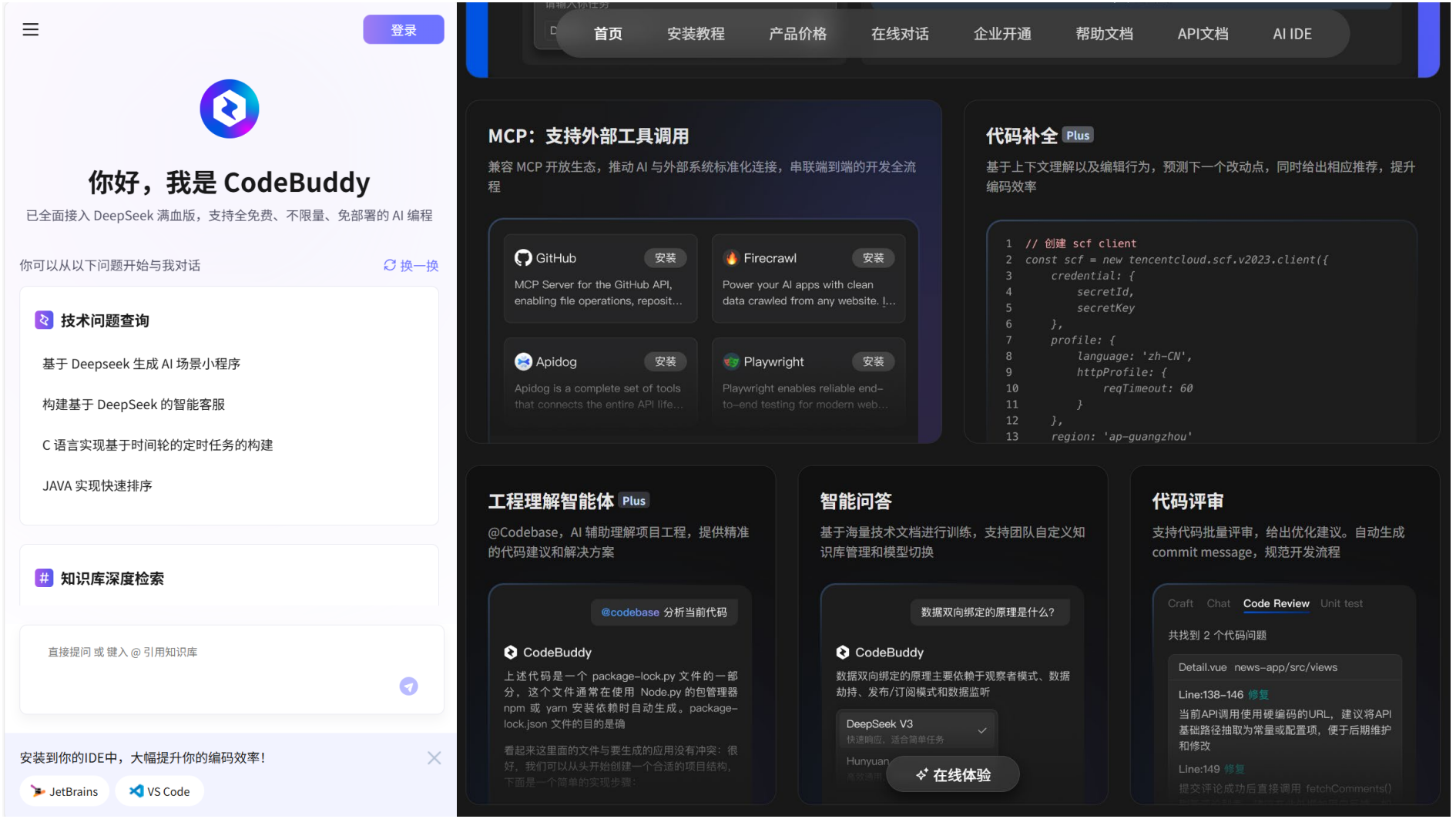
③合成数据集生成：特殊专业领域，大模型合成数据的迭代进化

↓

④代码大模型后训练：代码生成领域，大模型代码偏好数据的动态采样

Code LLM (Code)

- **CodeBuddy**, Cursor, Claude Code, GitHub Copilot, Trae, Lingma, CodeFuse, etc.



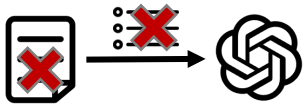
Synthetic Data Generation (SDG)

- 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 large models

- **Problem:** Existing large models cannot fit specifical 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

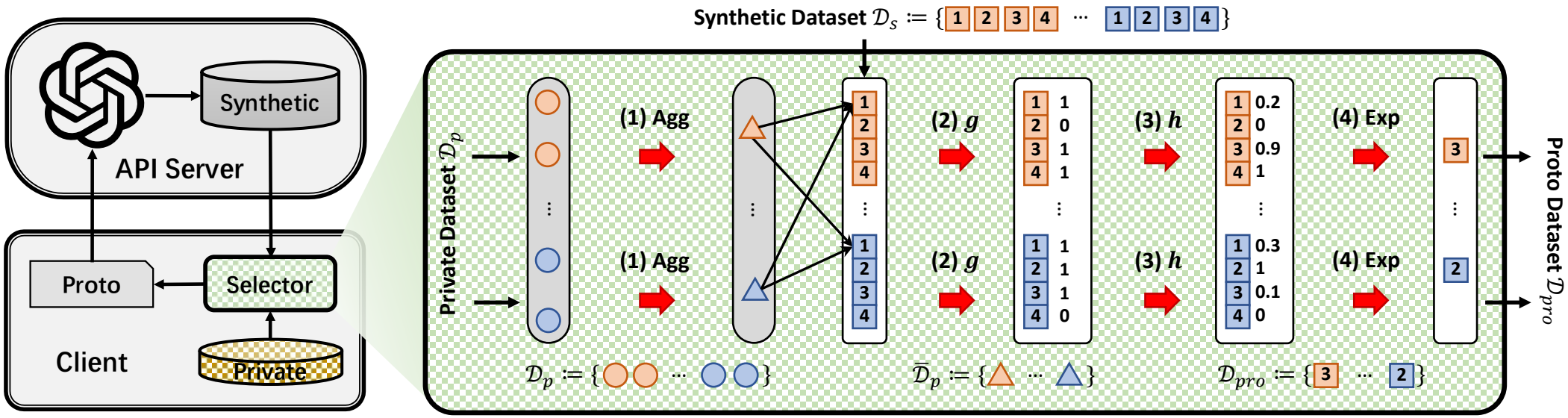


Few-shot ICL

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**



[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**

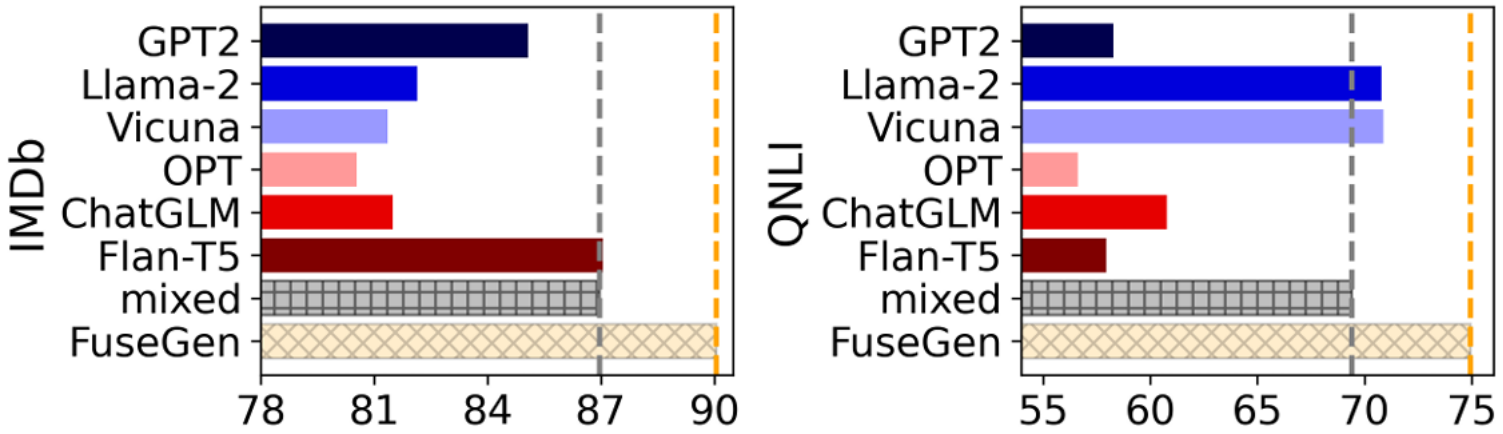
[SDG]: PCEvolve

(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.

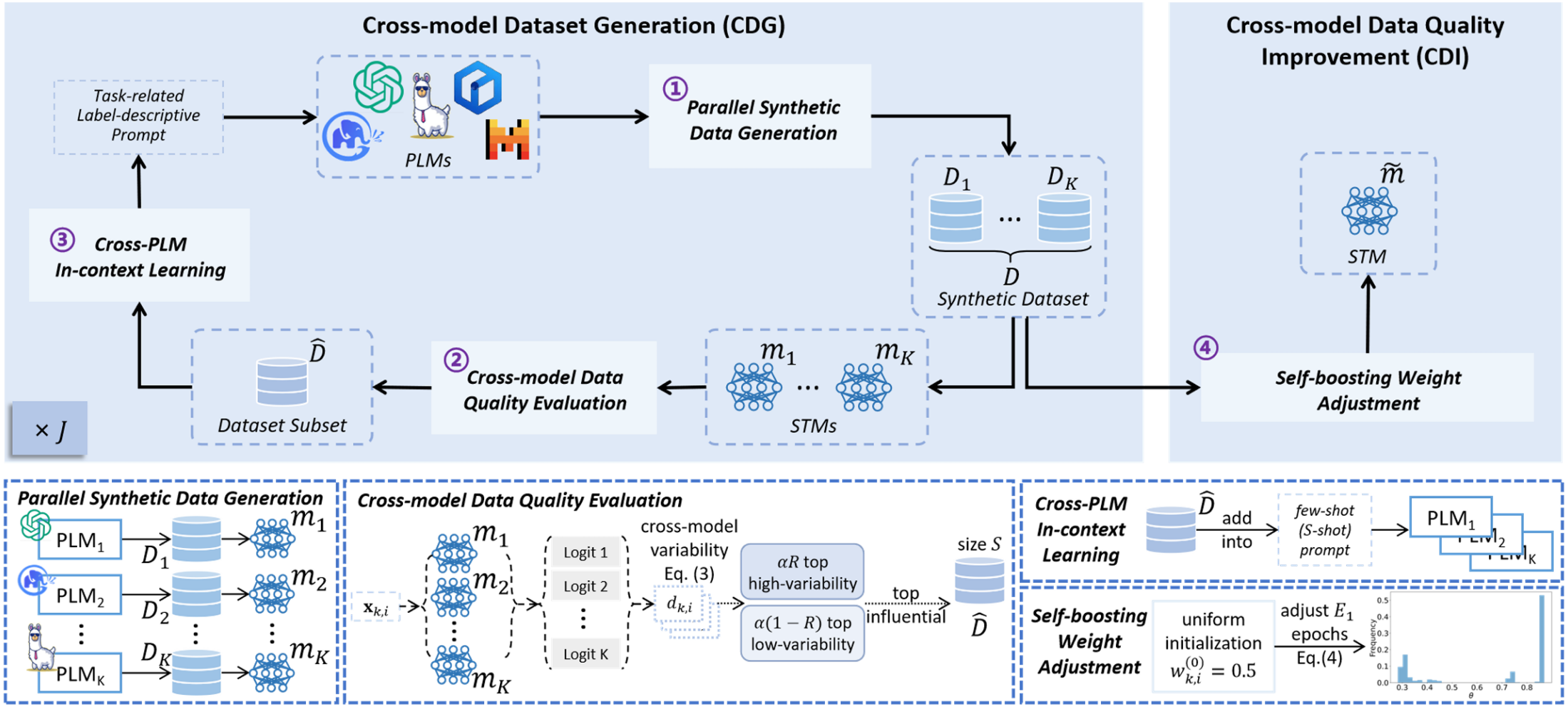
[SDG]: FuseGen

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



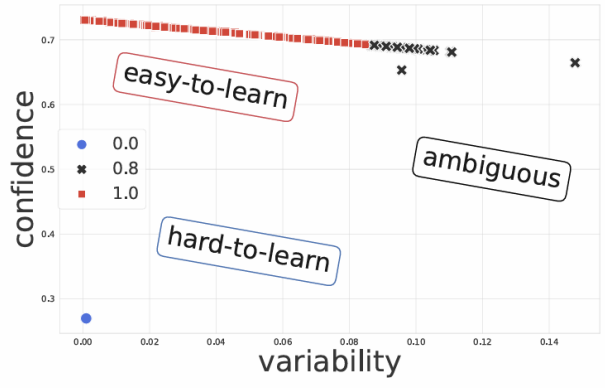
[SDG]: FuseGen

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

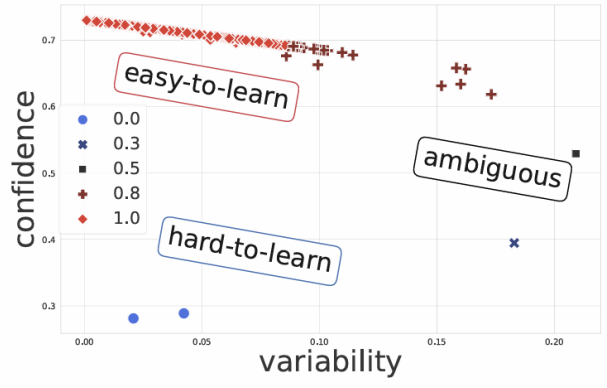


[SDG]: FuseGen

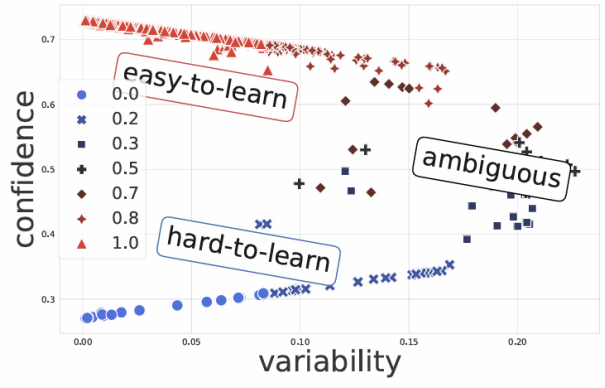
- Synthetic dataset cartography



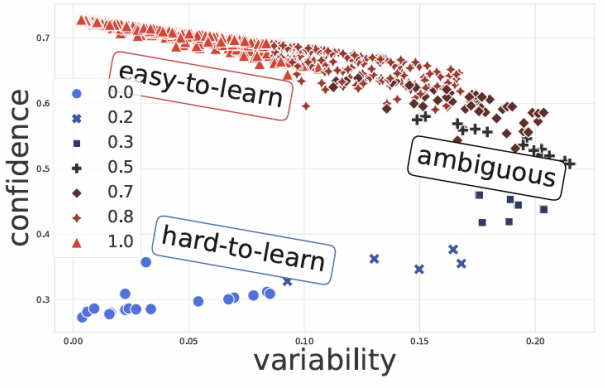
(a) Llama-2 ZeroGen $K = 1$ (84.23)



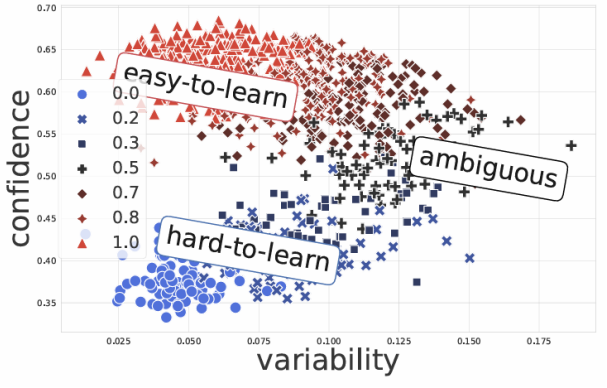
(b) Llama-2 ProGen $K = 1$ (84.24)



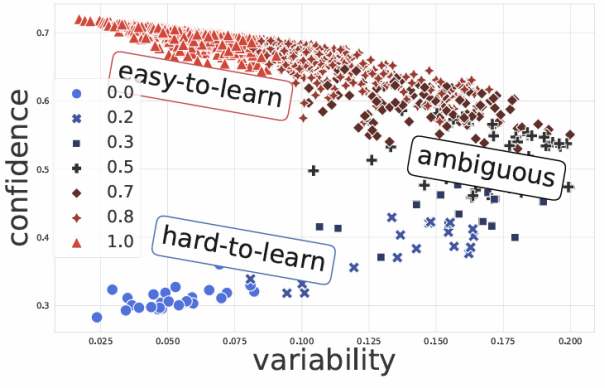
(c) Llama-2 Ours $K = 6$ (86.60)



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



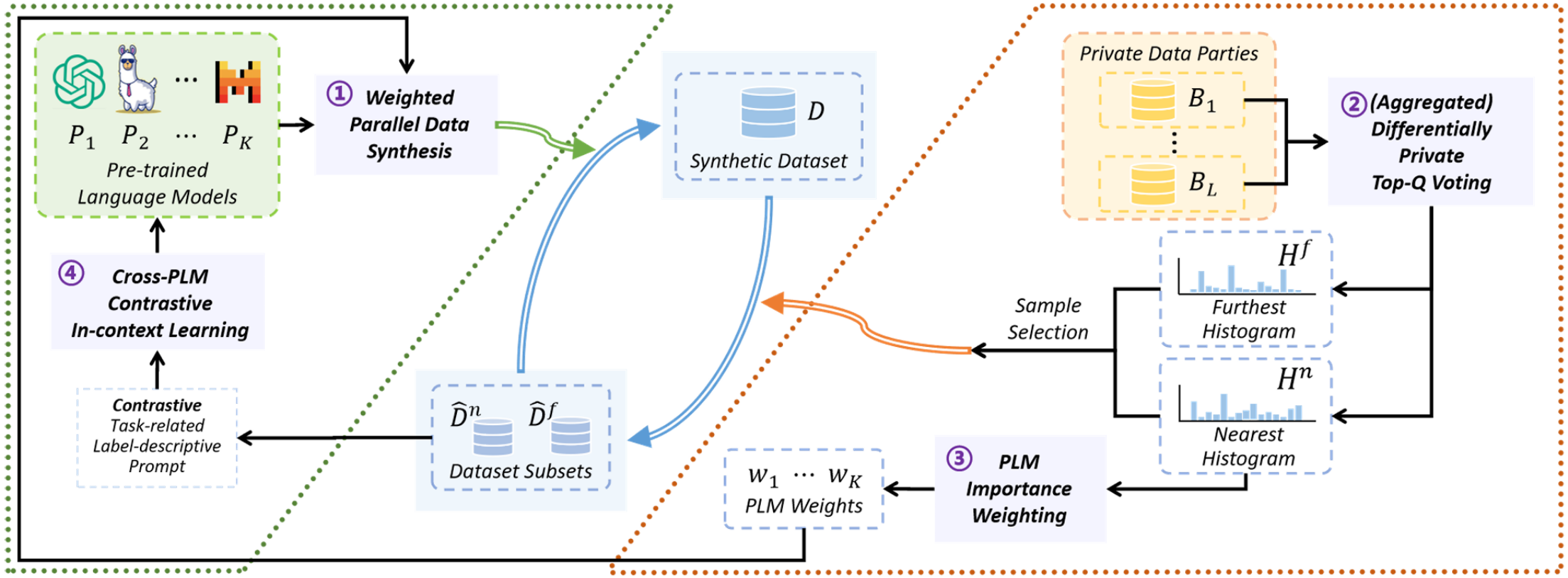
(e) Flan-T5 ProGen $K = 1$ (85.80)



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

[SDG]: WASP

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

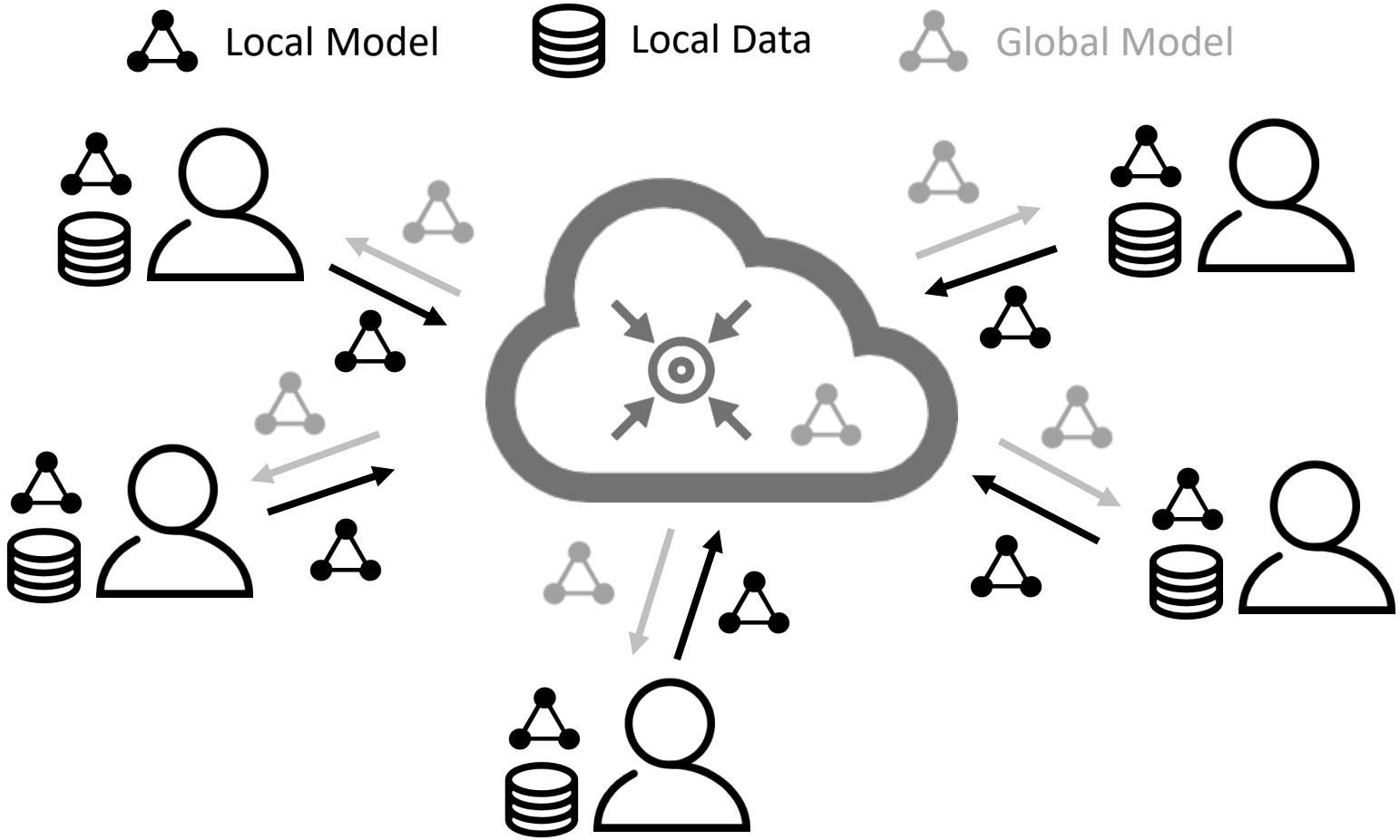


[SDG]: WASP

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 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

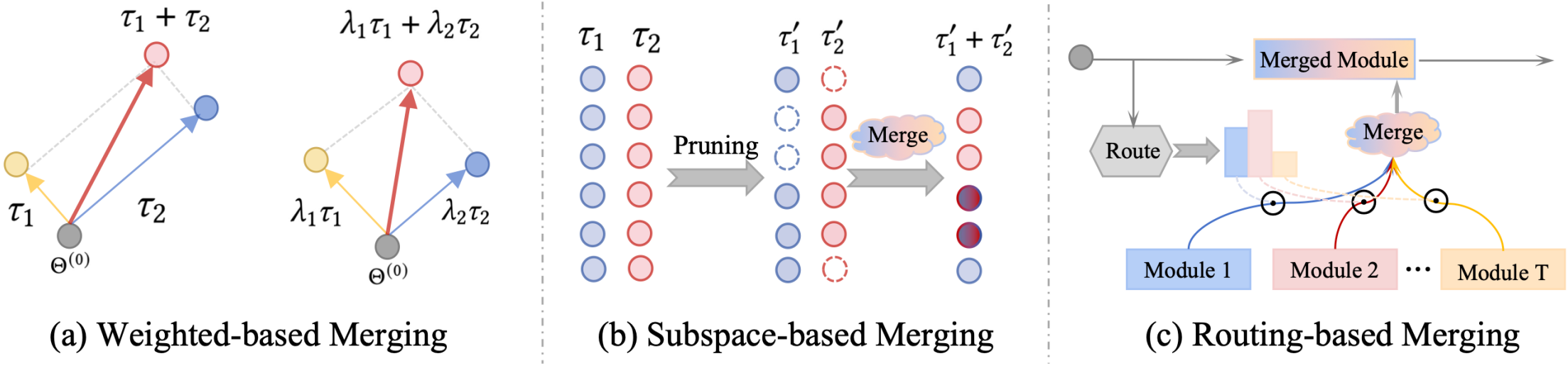
Distributed Learning (DL) (Model Merging)

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



Distributed Learning (DL) (Model Merging)

- I mainly focus on **model merging** in DL, which is also **popular** in **large model training** by
  - merging parameters, merging intermediate features, merge parameter-efficient LoRA modules, etc.
- **Multi-modal model**: obtain a single, effective, and parameter-efficient modality-agnostic model
- **RL**: DogeRM [1] merges the reward model with LLMs fine-tuned on different downstream domains to create domain-private reward models directly

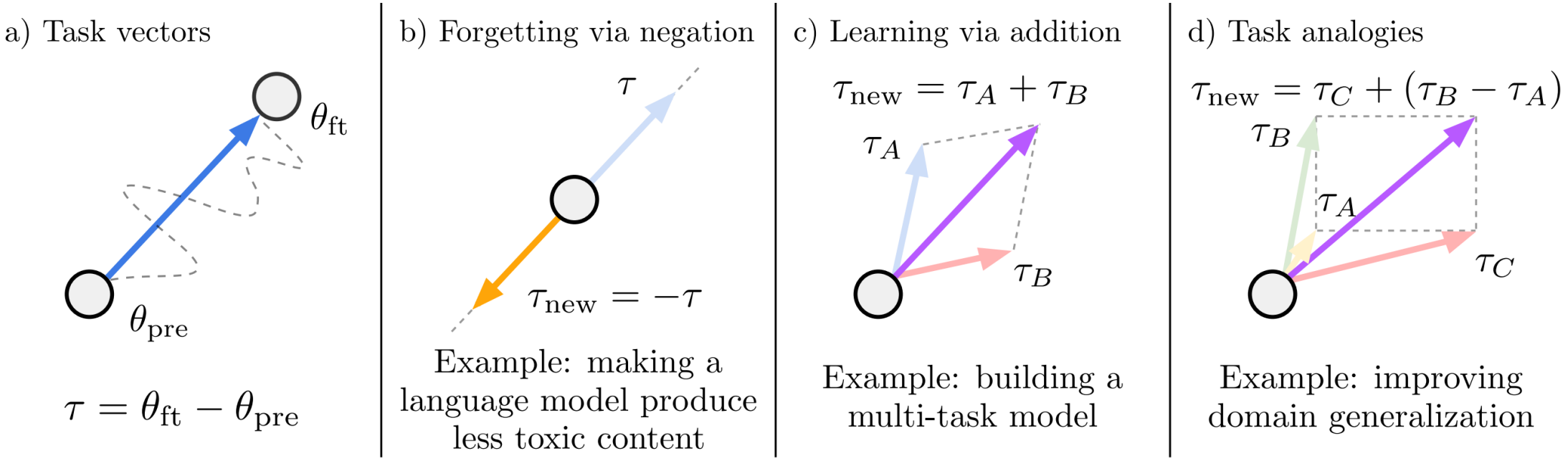


(a) Weighted-based Merging (b) Subspace-based Merging (c) Routing-based Merging

[1] Lin, Tzu-Han, et al. "DogeRM: Equipping Reward Models with Domain Knowledge through Model Merging." *EMNLP* 2024.

Distributed Learning (DL) (Model Merging)

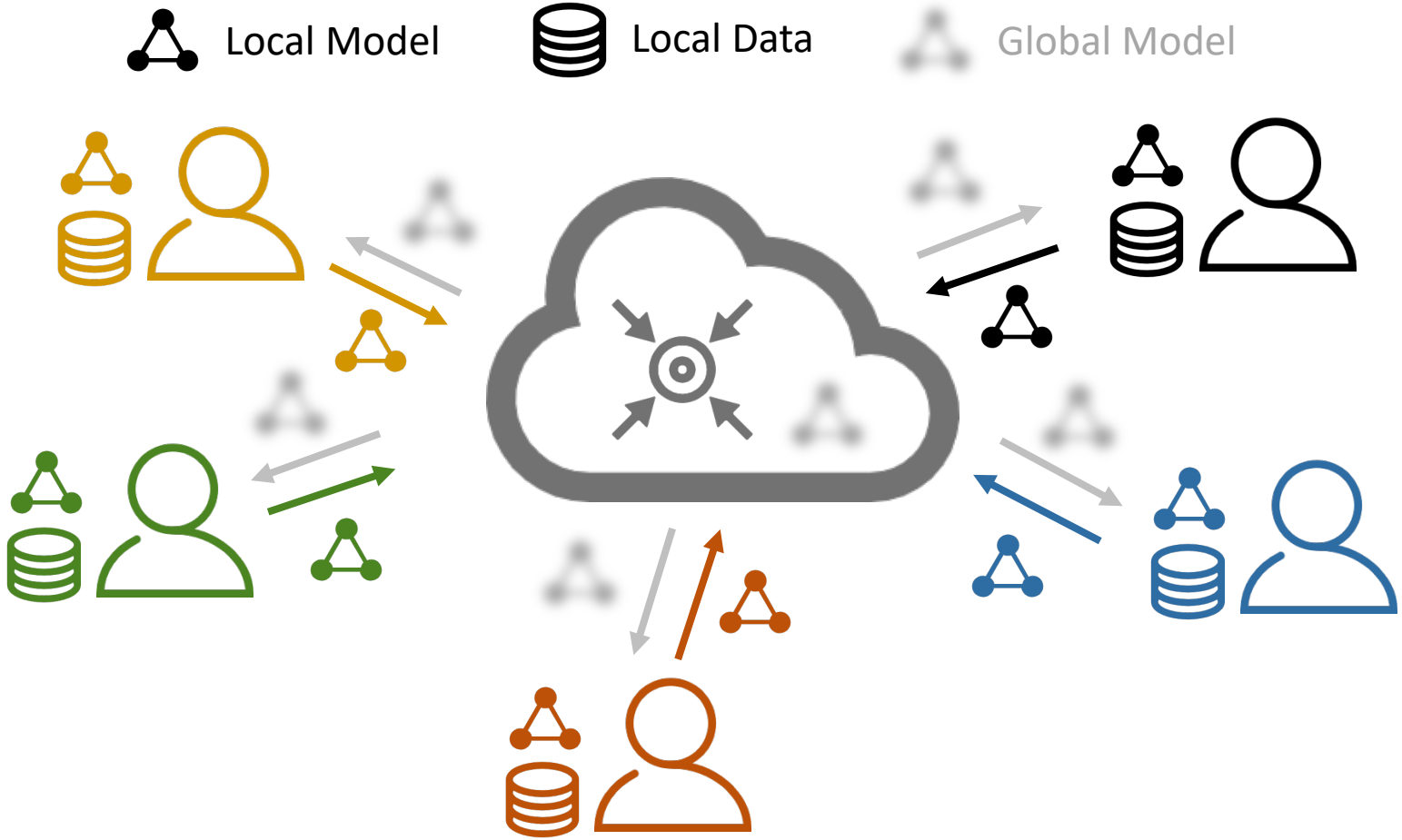
- I mainly focus on **model merging** in DL, which is also **popular** in **large model training** by
  - merging parameters, merging intermediate features, merge parameter-efficient LoRA modules, etc.
- **Model editing**: [1] shows that task vectors can be modified and combined together, and the behavior of the resulting model is steered accordingly. AlphaEdit [2] (ICLR'25 best paper)



[1] Ilharco, Gabriel, et al. "Editing models with task arithmetic." *ICLR* 2023.
[2] Fang, Junfeng, et al. "Alphaedit: Null-space constrained knowledge editing for language models." *ICLR* 2025.

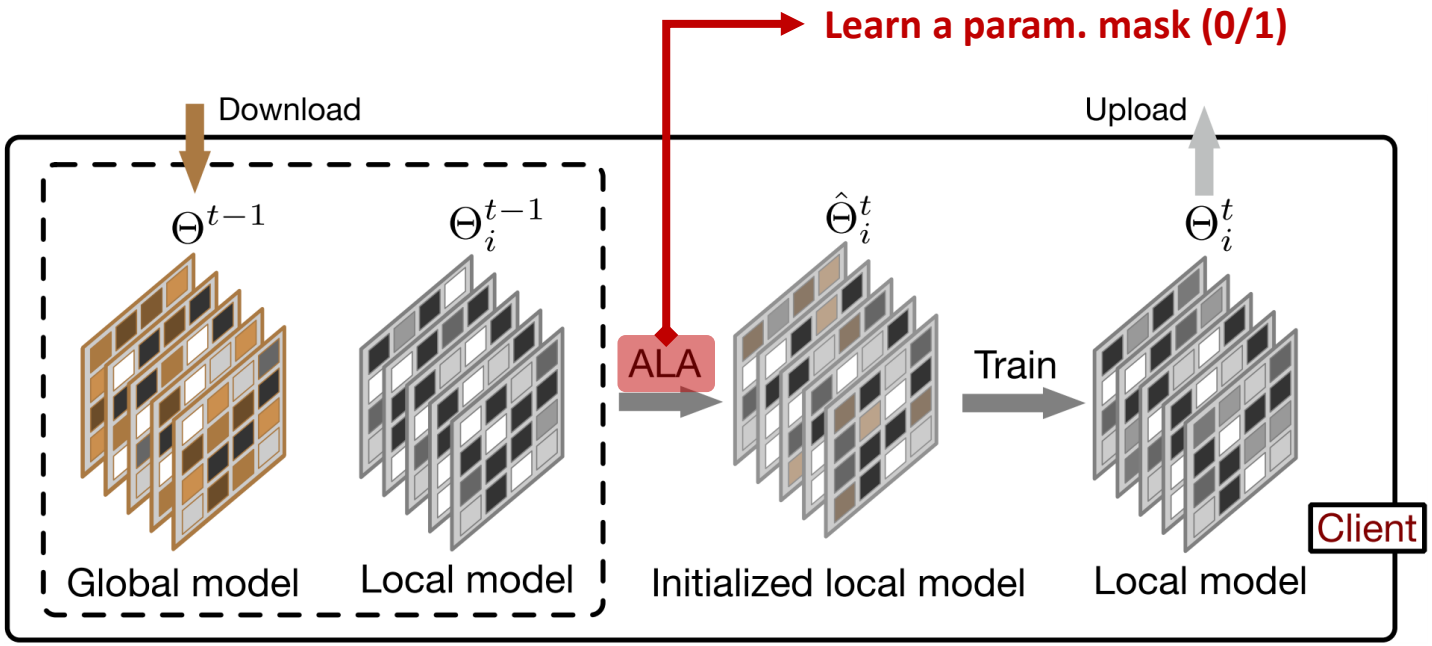
[DL]: Data Heterogeneity (Model Merging)

- **Problem:** Different tasks have **different** data distributions



[DL]: FedALA (Model Merging)

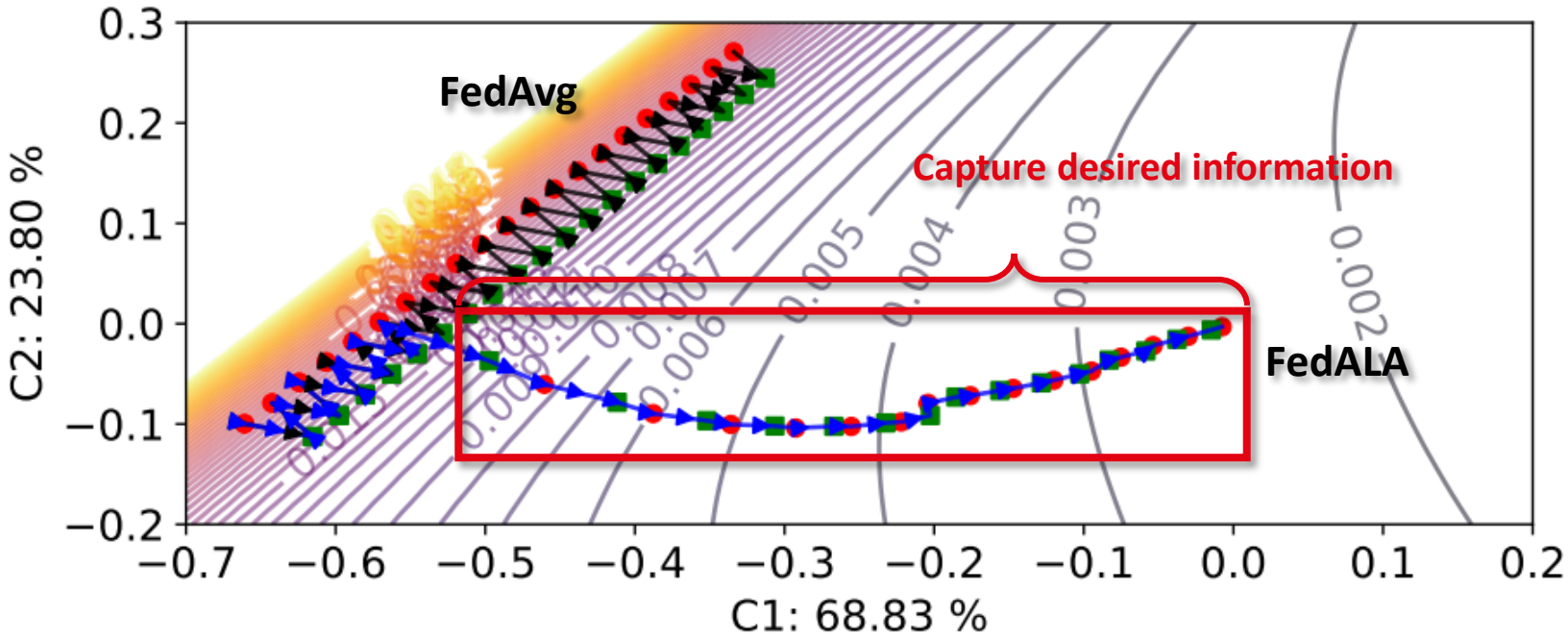
- Extract each client's **desired parameters** from the global model that facilitates local training
- **Adaptively aggregate** the parameters in the global and local model for initialization



Workflow on the client in one iteration

[DL]: FedALA (Model Merging)

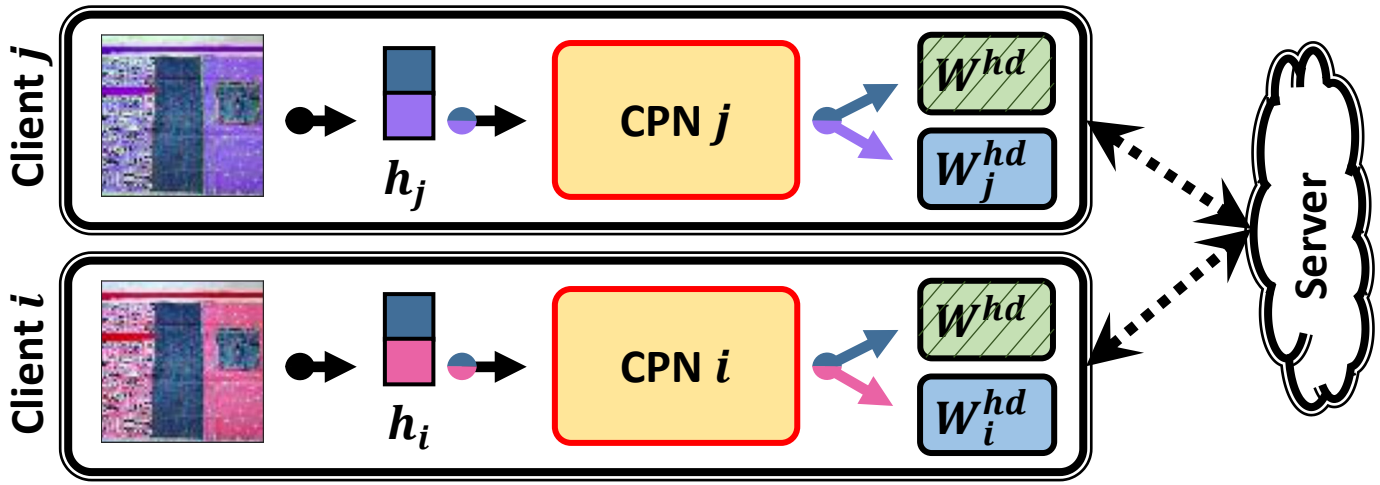
- Learning trajectory on one client: **FedAvg** vs. **FedALA**
- Activate **ALA** in the subsequent iterations



2D visualization of local learning trajectory

[DL]: FedCP (Model Routing)

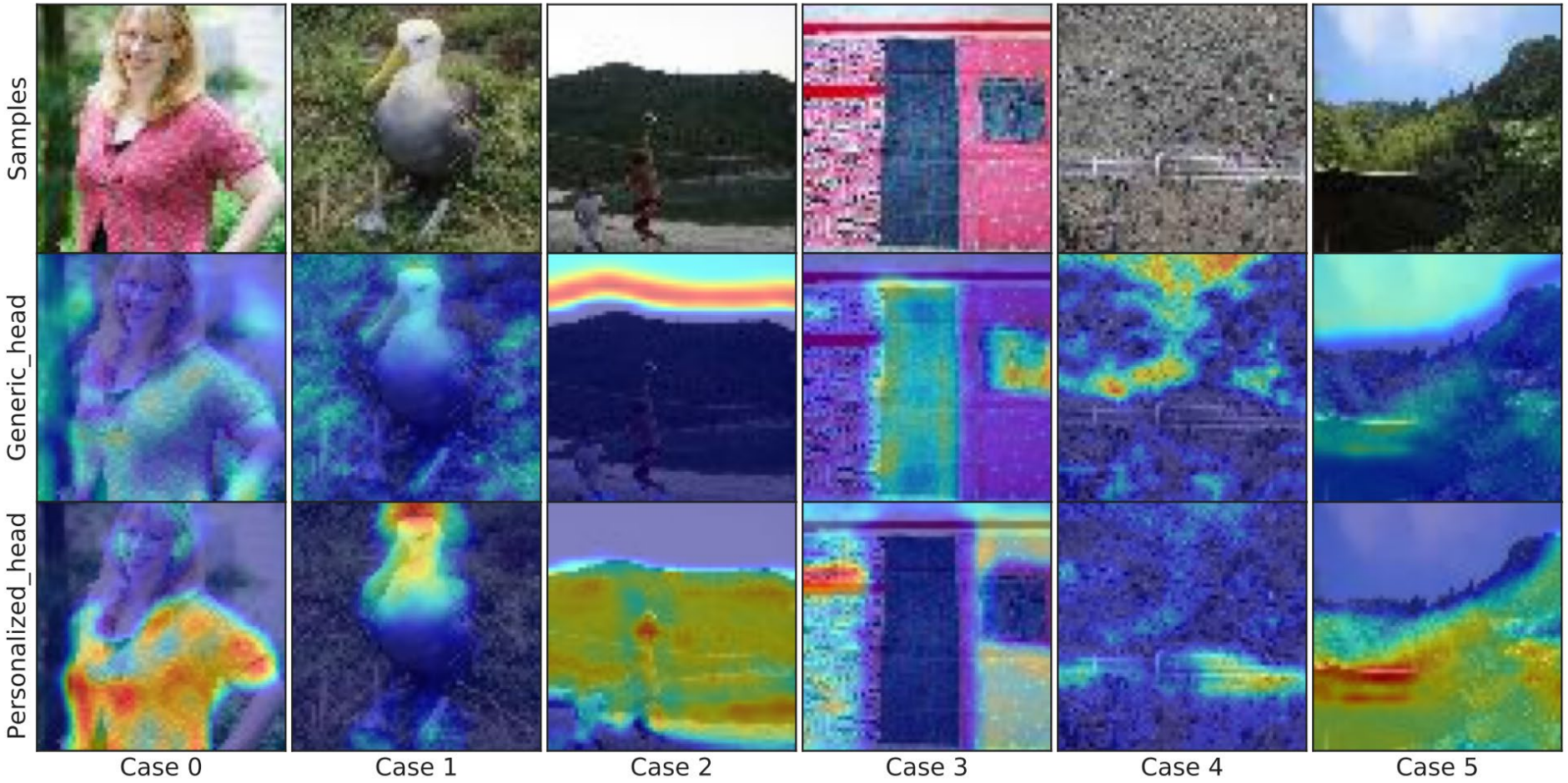
- We separate feature information via an auxiliary **Conditional Policy Network (CPN).**
  - Sample-specific separation
  - Lightweight (e.g., 4.67% parameters of ResNet-18)



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

[DL]: FedCP (Model Routing)

- **Separating Feature Information**



Six samples from the Tiny-ImageNet dataset

[DL]: GPFL (Model Routing)

- GCE introduces more global information **simultaneously** with local training
- CoV **eliminates the interaction** between global and personalized feature learning

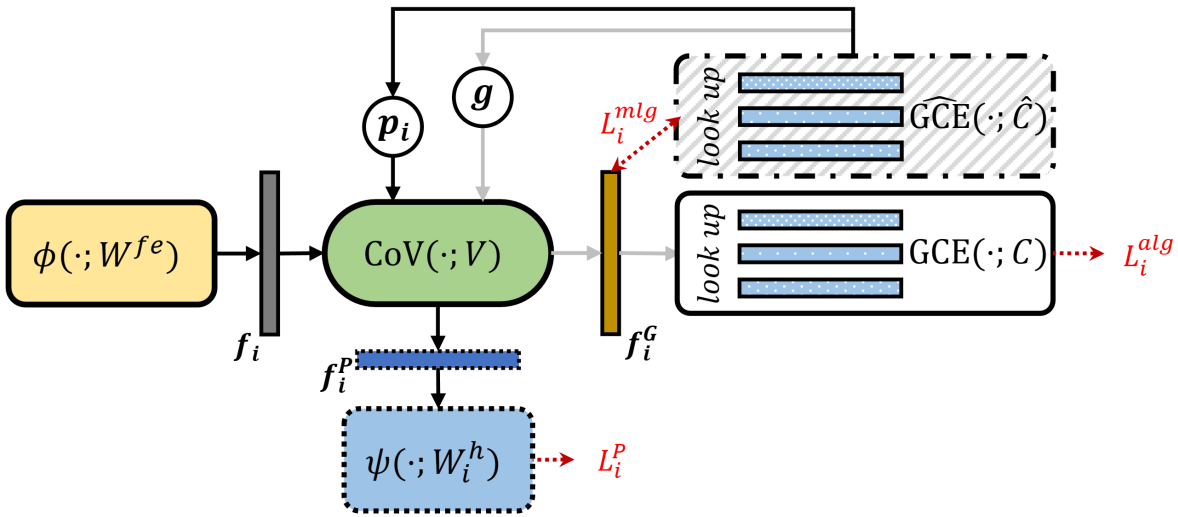
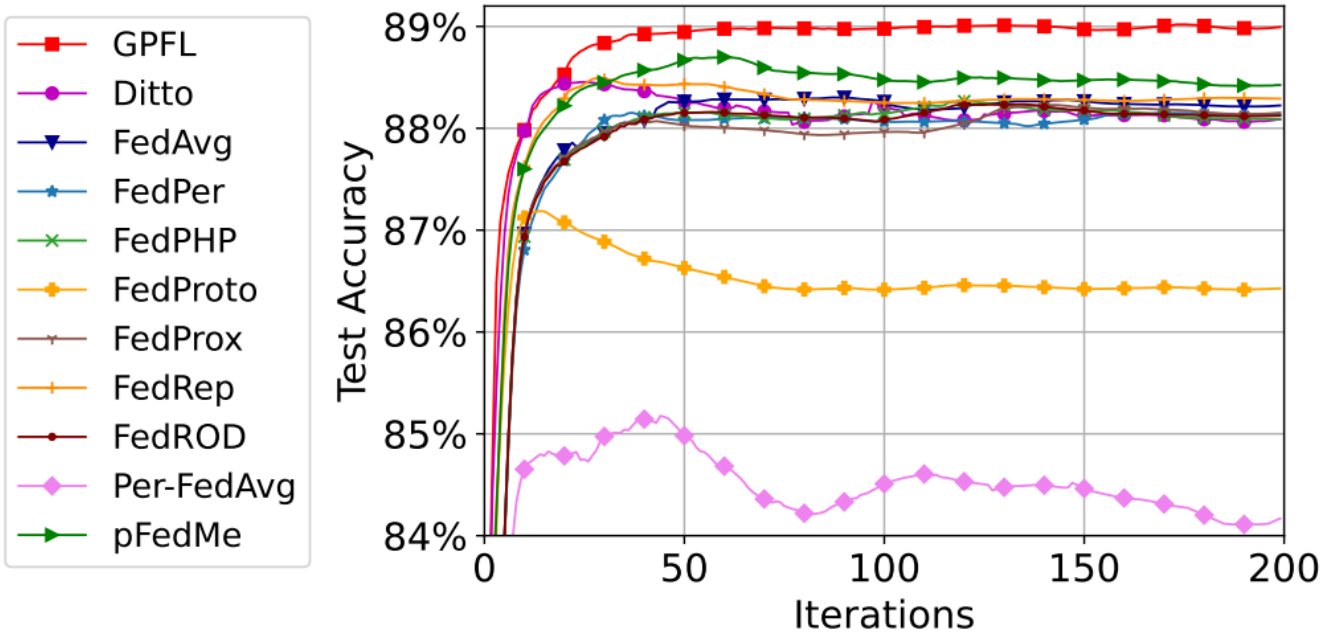


Illustration of client modules and data flow between them

[DL]: GPFL (Model Routing)

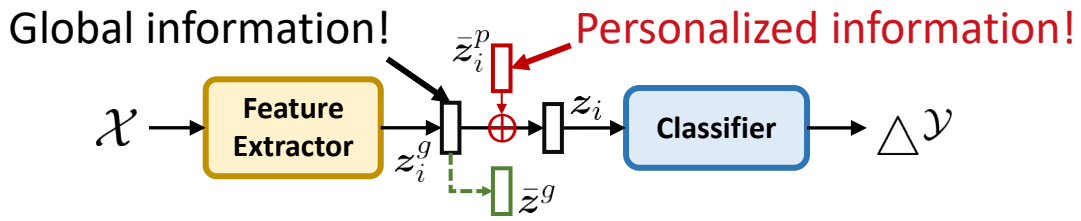
- Relieve the **widely existed** overfitting issue in pFL



Test accuracy curves in the feature shift setting

[DL]: DBE (Feature Decoupling)

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



Local model (with PRBM and MR)

[DL]: DBE (Feature Decoupling)

- 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.

[DL]: DL on Device

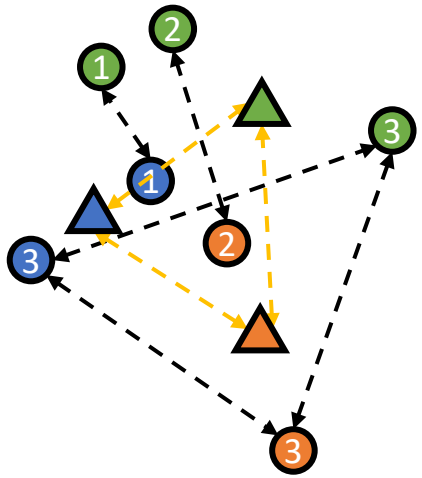
- **Real-world deployment**
  - **+ CoLExT, + real-world datasets, + systematical metrics**

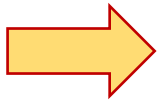
- 28 Single Board Computers (SBC)
  - Orange Pi, LattePanda, Nvidia Jetson
- 20 Smartphones
  - Samsung, Xiaomi, Google Pixel, Asus ROG, One Plus
- High Voltage Power Meter
- Wired and wireless networking
- Workstation - FL Server

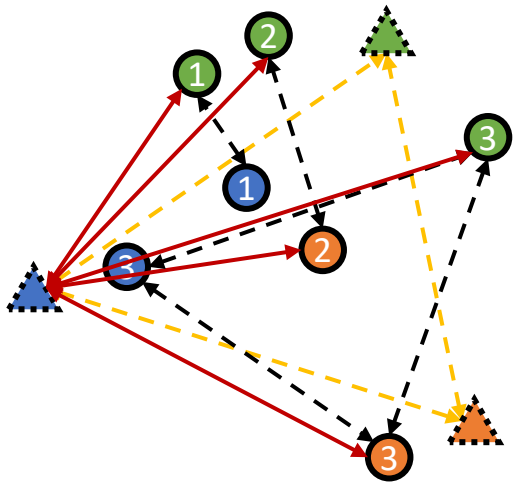
https://github.com/TsingZ0/HtFL-OnDevice

[DL]: FedTGP (Knowledge Distillation)

- Remove weighted-averaging
- Consider the uploaded client prototypes as data
- **Enlarge** the global prototype margin







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

Trainable Global Prototype (TGP)

[DL]: FedTGP (Knowledge Distillation)

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

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

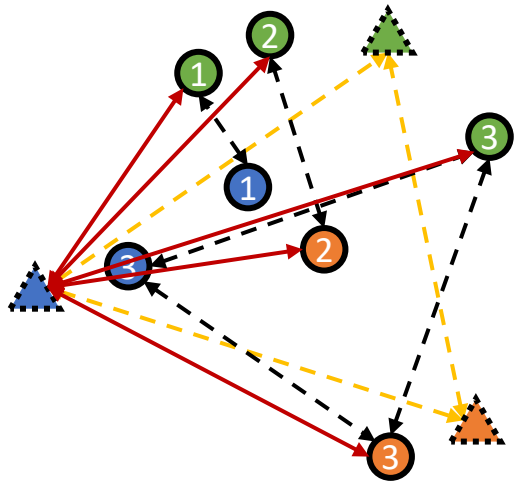
$$\mathcal{L}_P^c = \sum_{i \in \mathcal{I}^t} -\log \frac{e^{-(\phi(P_i^c, \hat{P}^c) + \delta(t))}}{e^{-(\phi(P_i^c, \hat{P}^c) + \delta(t))} + \sum_{c'} e^{-\phi(P_i^c, \hat{P}^{c'})}}$$

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

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

maximum cluster margin

τ is a margin threshold



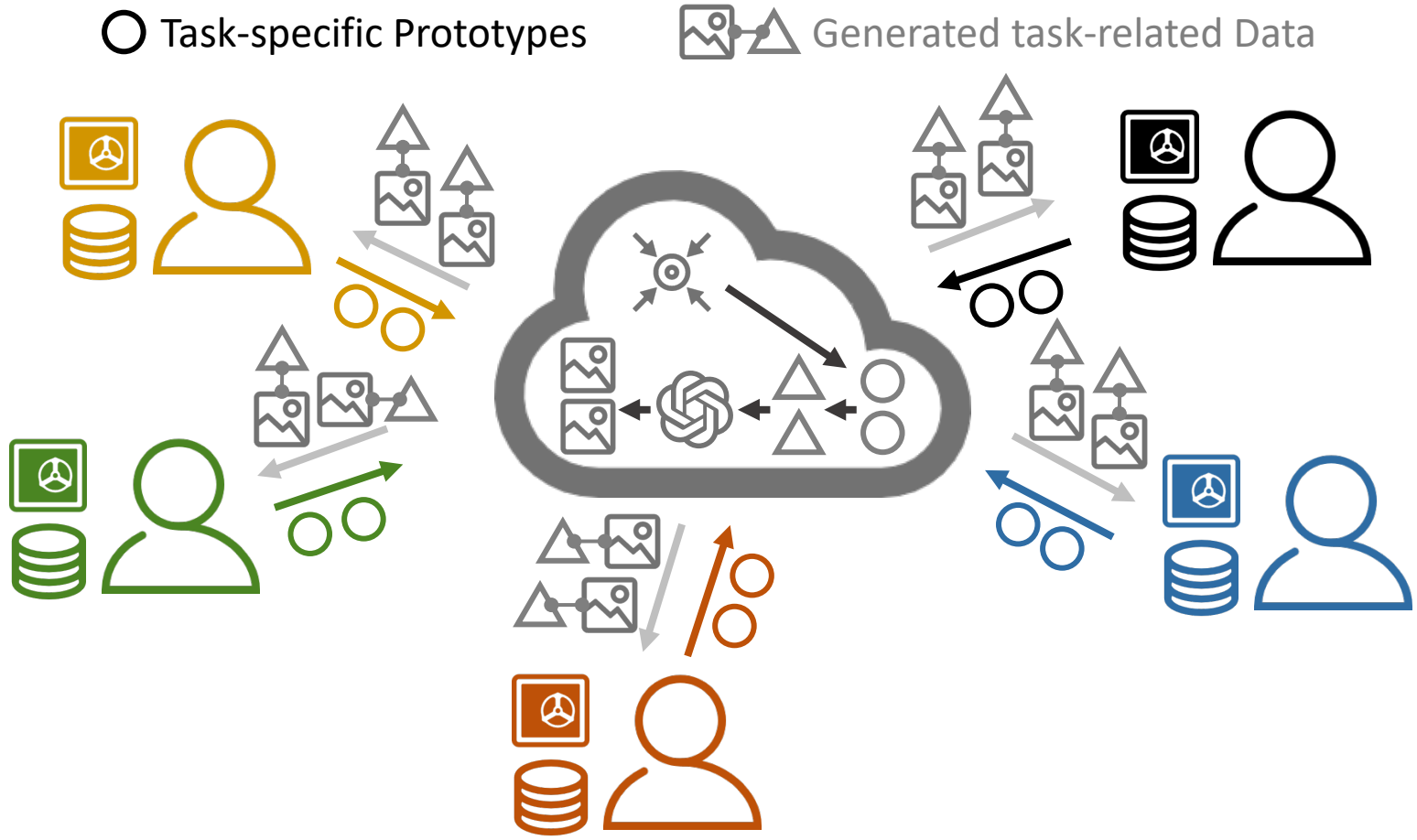
$\hat{P}^c$: A TGP of class c

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

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

[DL]: FedKTL (Knowledge Distillation)

- Transfer **common knowledge** from the generator to clients
- Obtain **task-specific knowledge** from other clients

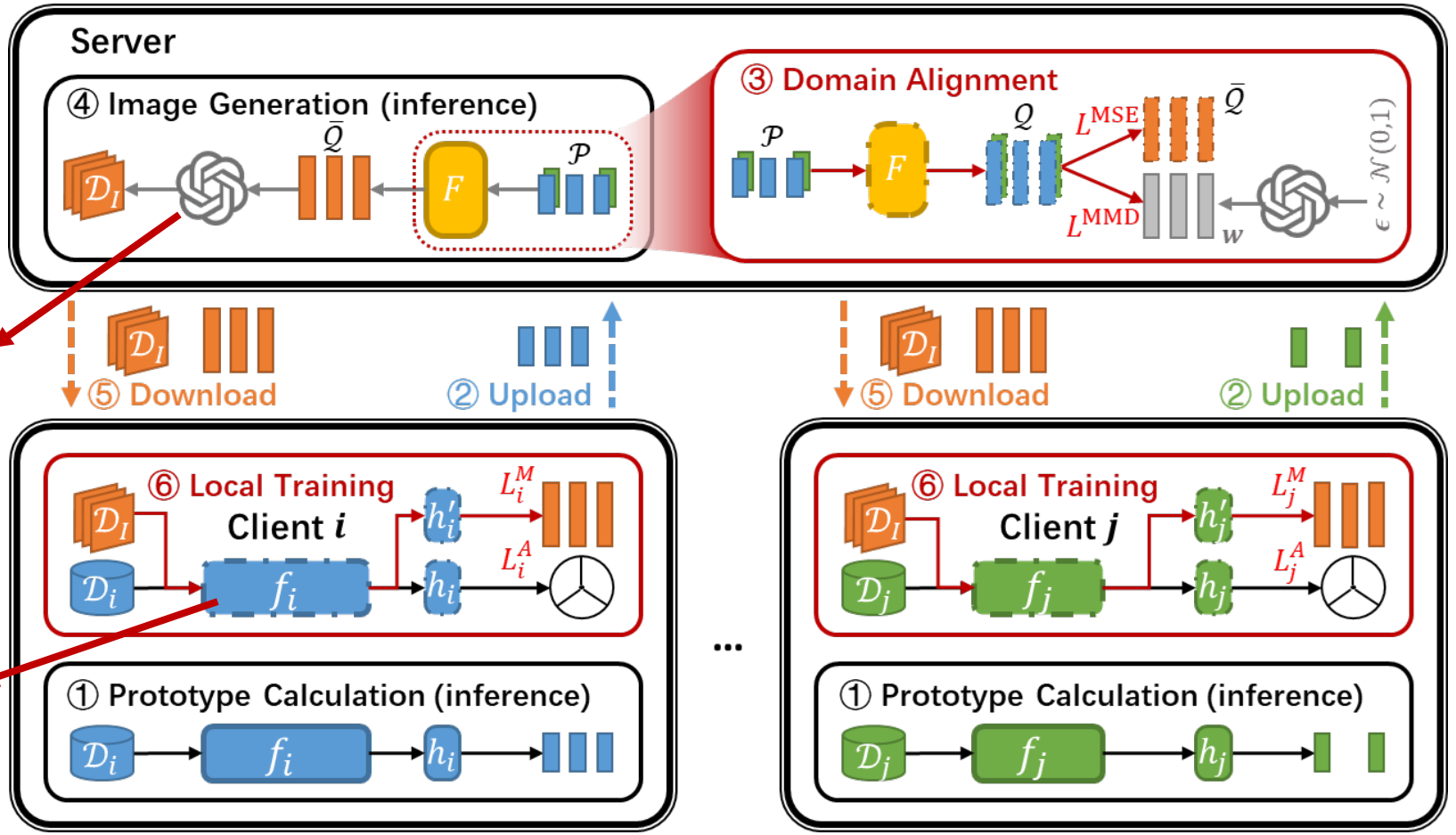


Accepted by CVPR 2024

Outperform baselines by up to 7.31%

[DL]: FedKTL (Knowledge Distillation)

- **Align** small models' feature space with the generative model's
- 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%

[DL]: FedKTL (Knowledge Distillation)

- FedKTL can **adapt to various generators** that were pre-trained using various datasets
- The **semantics of the generated images** can be different from clients' data

(a) Client #1 | (b) AFHQv2 (c) Benches (d) FFHQ-U (e) WikiArt

Generators pre-trained on different image datasets

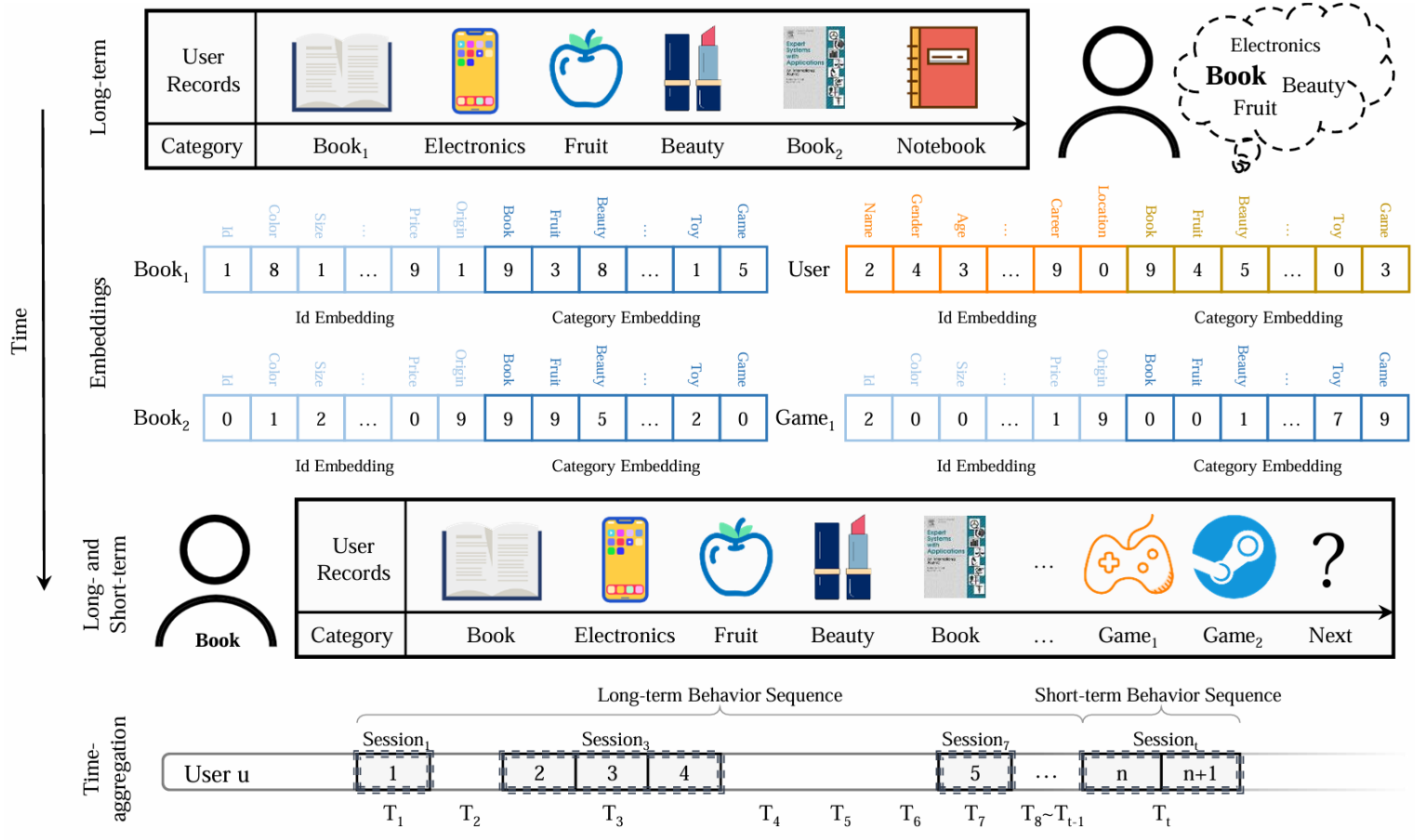
	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$
AFHQv2	26.82±0.32	**27.05±0.26**	26.32±0.52
Bench	27.71±0.25	**28.36±0.42**	27.56±0.50
FFHQ-U	**27.28±0.23**	27.21±0.35	26.59±0.47
WikiArt	27.37±0.51	**27.48±0.33**	27.30±0.15

Accepted by CVPR 2024

Outperform baselines by up to 7.31%

Recommender System (RS)

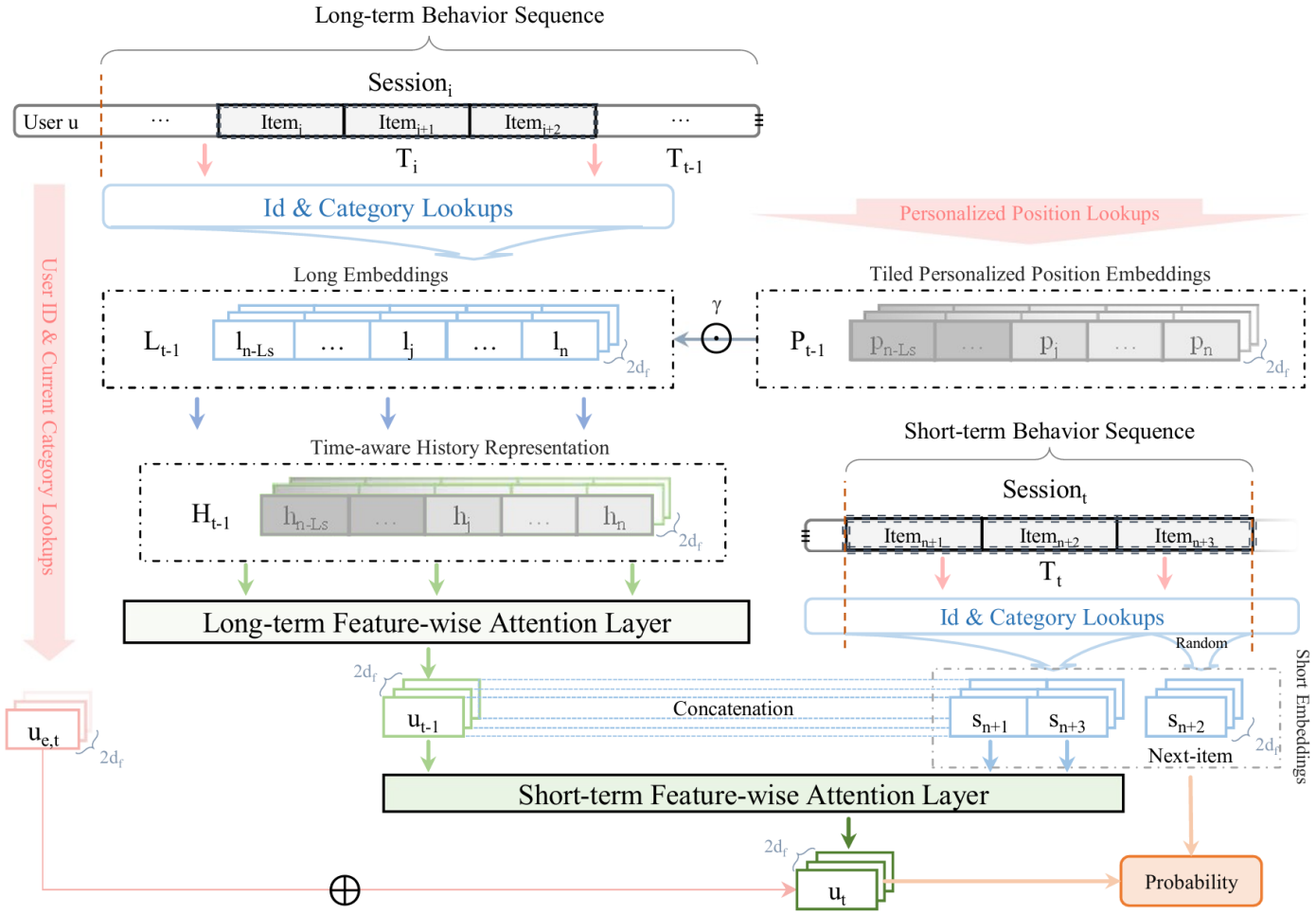
- **Problem:** Users have **personalized taste for time**
- **Solution:** Time-aware Long- and Short-term Attention Network for Next-item Recommendation



Outperform baselines by up to 20.79%

[RS]: TLSAN

- Capture **personalized time-aggregation pattern** in long-term attention



Outperform baselines by up to 20.79%

Feel free to contact me!

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

Thanks!