浙江大学
ZHEJIANG UNIVERSITY

大小模型端云协同赋能人机交互

张圣宇 | 浙江大学

2025年4月

马斯克的大胆预言：碳基生命（也就是我们人类）只是硅基生命的启动程序。随着科技的不断发展，尤其是AI领域取得的突破，以人工智能为主的硅基生命形态将会在未来成为地球上的主宰生物。

DeepSeek & ChatGPT

- 2007年1月9日，乔布斯发布第一代iPhone苹果手机，把iPod、电话、移动互联网设备等进行有机整合，推动了移动互联网进入了黄金发展年代。

- 今天大模型给人类社会诸多生产、生活模式带来一次大变革。2023年2月，英伟达创始人兼CEO黄仁勋提出随着ChatGPT为代表的大模型出现，我们已经进入“人工智能的iPhone时刻（iPhone moment of AI）”，这一观点受到美国《财富》杂志、华尔街时报等媒体的广泛认可并转载。

- DeepSeek在模型算法和工程优化方面进行了系统级创新，在2048块英伟达H800 GPU（针对中国市场的低配版GPU）集群上完成训练，打破了大语言模型以大算力为核心的预期天花板，为在受限资源下探索通用人工智能开辟了新的道路。

GPT-4o

https://openai.com/index/hello-gpt-4o/

GPT-4o：By my eyes

https://openai.com/index/be-my-eyes/

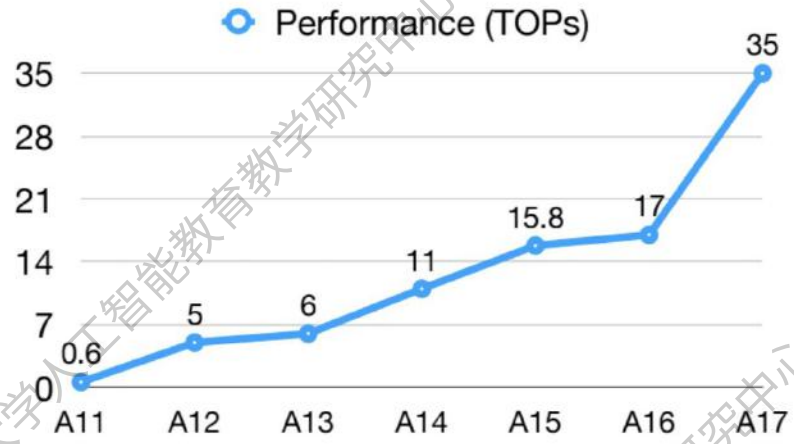
移动端智能：生活的方方面面

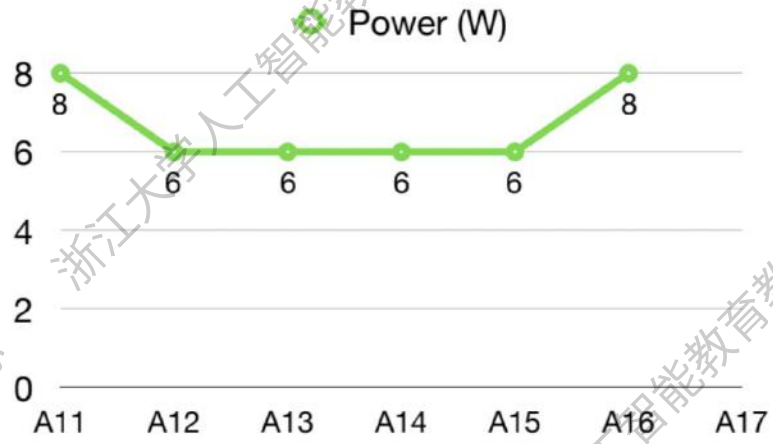
手机、平板、智能手表

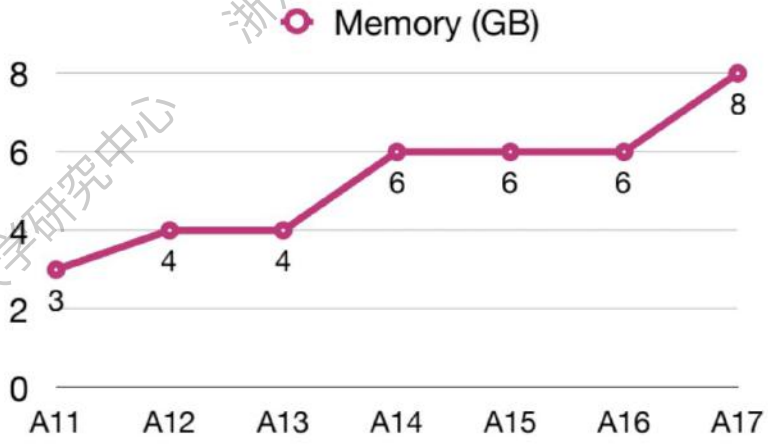
- 小设备功能不简单
- 小设备如何搭载大模型？
- 大小模型协同

发信息、看视频、听音乐、导航、游戏、购物

端侧计算能力越来越强

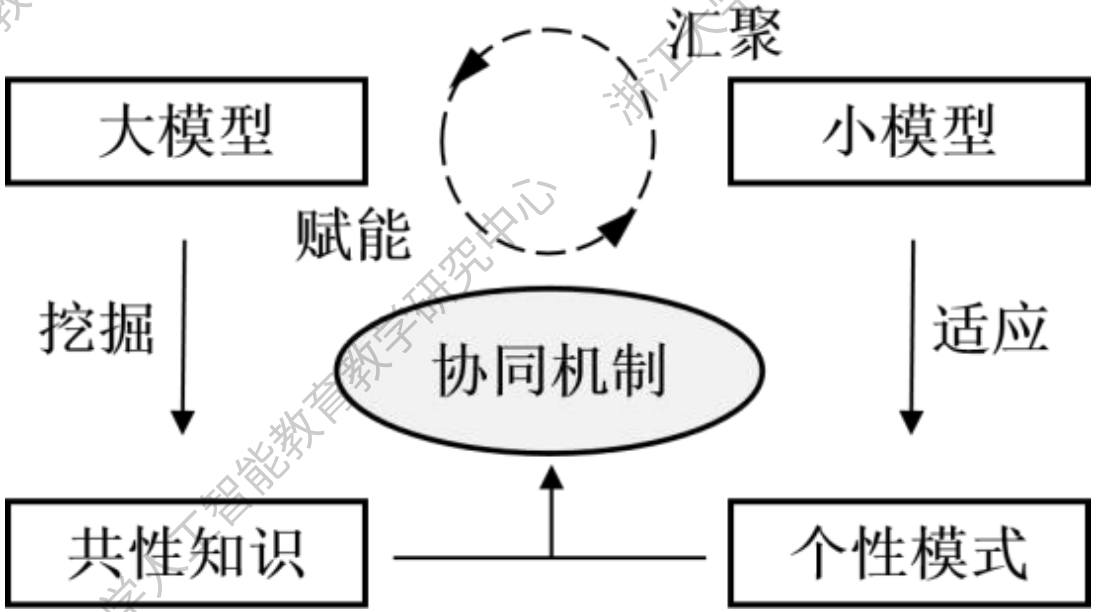






大小模型端云协同

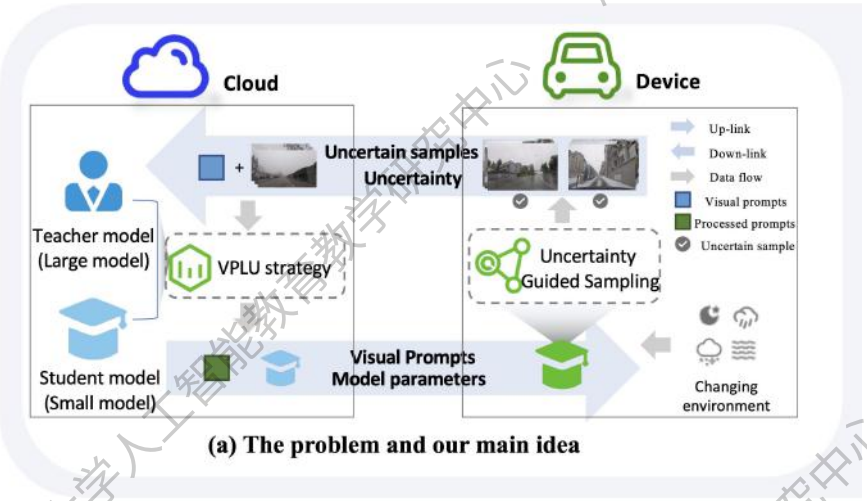
- 端云协同（Device-Cloud Collaboration）：指边缘设备（如智能手机、IoT设备）模型和云侧服务器模型协同进化推断。
- 云侧大模型（Large Model）：通用认知计算，拥有强大的计算能力、海量的数据、充分的知识库。
- 终端小模型（Small Model）：实时感知、实时响应，运行轻量级任务，响应速度快。



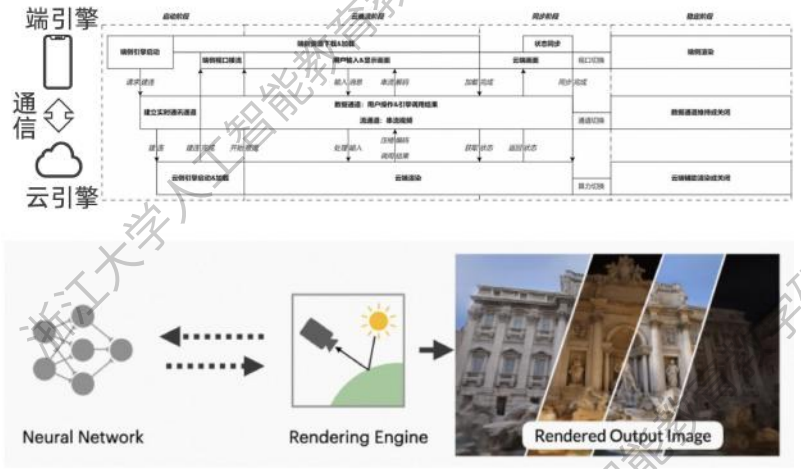
大小模型端云协同

- 端云协同计算通过卸载部分学习任务至端侧，让端和云协同完成任务，从而发挥终端靠近用户和数据源的天然优势，降低服务延时至毫秒级，增强模型个性化精准推理能力，缓解云服务器中心负载压力，同时支持用户原始数据在设备本地处理
- 有效克服主流云学习范式在实时性、个性化、负载成本、隐私安全等方面的不足

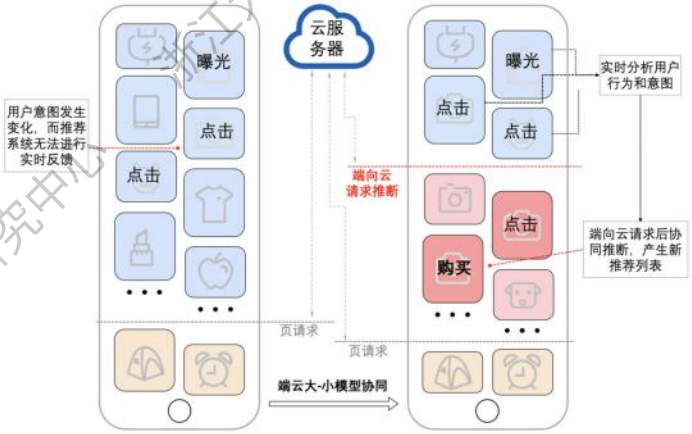
前沿应用



自动驾驶 (Gan et al.)



3D渲染 (Lv et al.)



推荐系统 (Qian et al.)

Yulu Gan, Mingjie Pan, Rongyu Zhang, et al.: Cloud-Device Collaborative Adaptation to Continual Changing Environments in the Real-World. CVPR 2023: 12157-12166
Chengfei Lv, Chaoyue Niu, Renjie Gu, et al.: Walle: An End-to-End, General-Purpose, and Large-Scale Production System for Device-Cloud Collaborative Machine Learning. OSDI 2022: 249-265
Xufeng Qian, Yue Xu, Fuyu Lv, Shengyu Zhang, et al.: Intelligent Request Strategy Design in Recommender System. KDD 2022: 3772-3782

大小模型端云协同

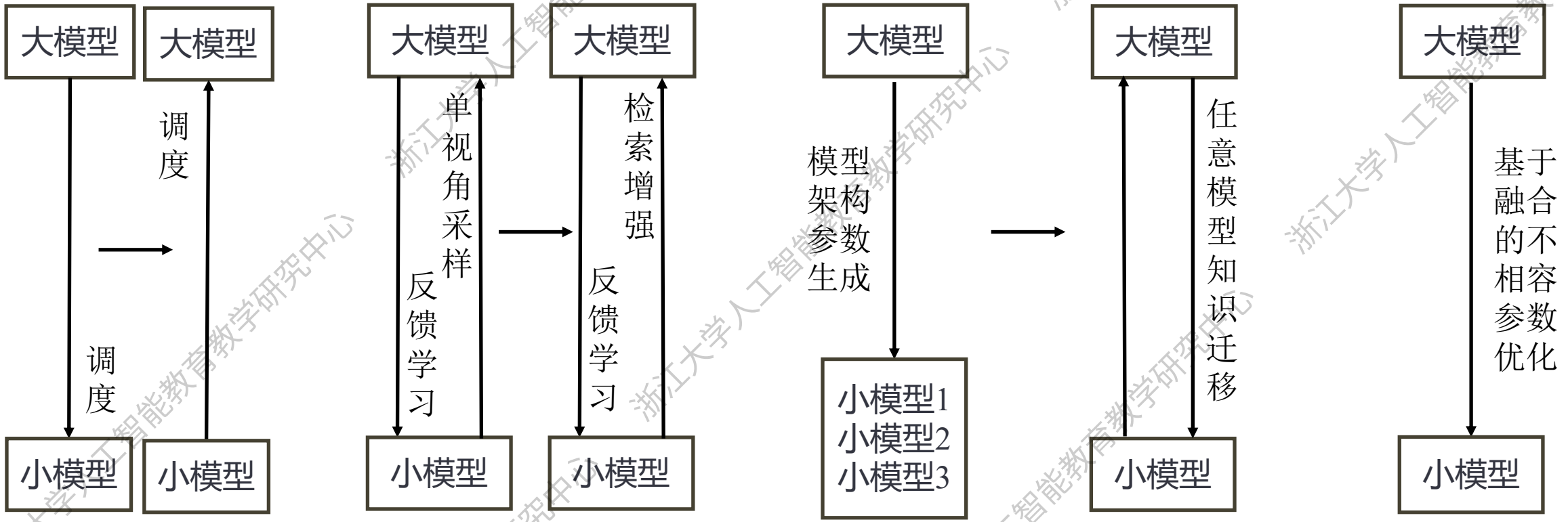
≈

大小模型协同 + 端云高效协同

大小模型协同基础算法研究

联合应用平台既有的特定业务小模型与云侧大模型，将端侧小模型轻量部署、快速响应、个性适配的优势，和云侧大模型认知推理、多模态理解、通用泛化的优势进行互补

基于**调度**的协同 → 基于**反馈**的协同 → 基于**生成**的协同 → 基于**融合**的进化



LLMCO4MS
组合优化LLM
ECCV 2024

IntellectReq
自主智能请求
WWW 24

MPOD123
2D到3D生成
CVPR 2024

FiGRet
RAG反馈学习
Arxiv

ModelGPT
大模型
生成小模型
Arxiv

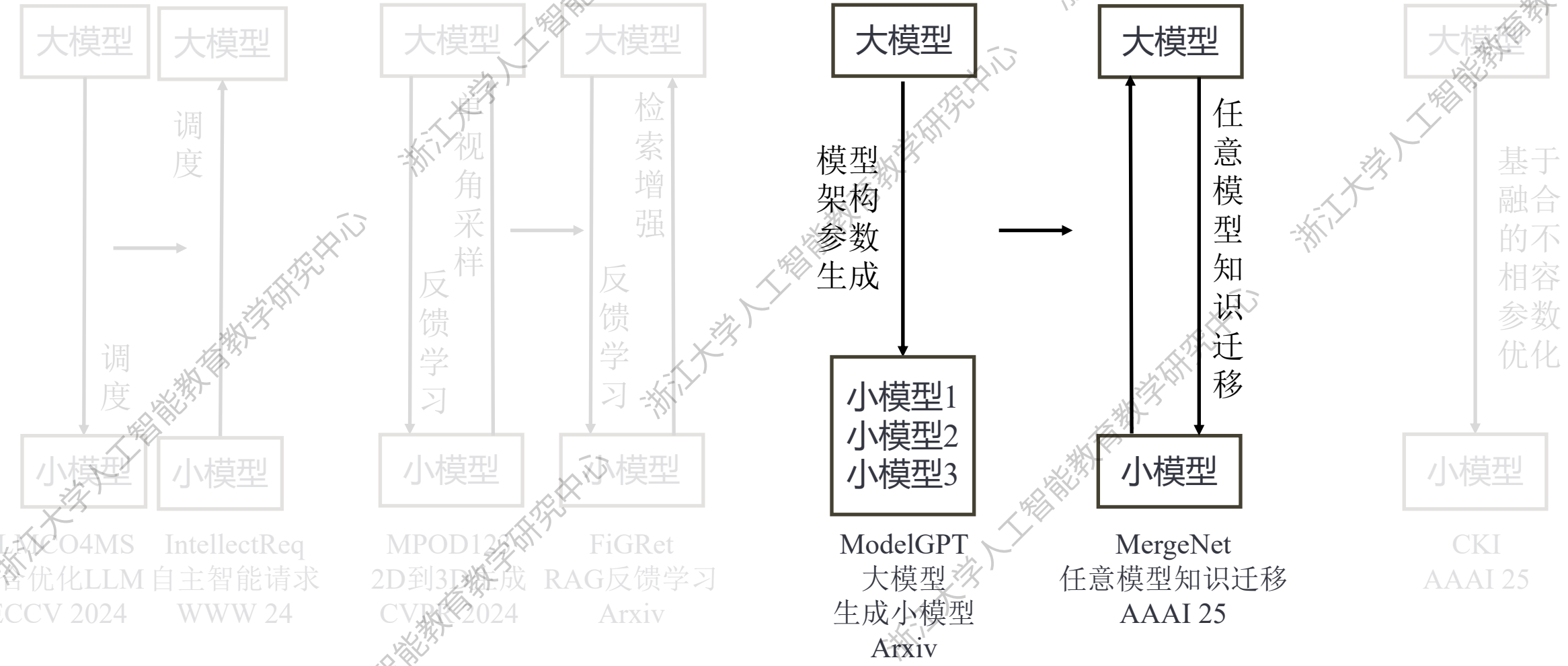
MergeNet
任意模型知识迁移
AAAI 25

CKI
AAAI 25

大小模型协同基础算法研究

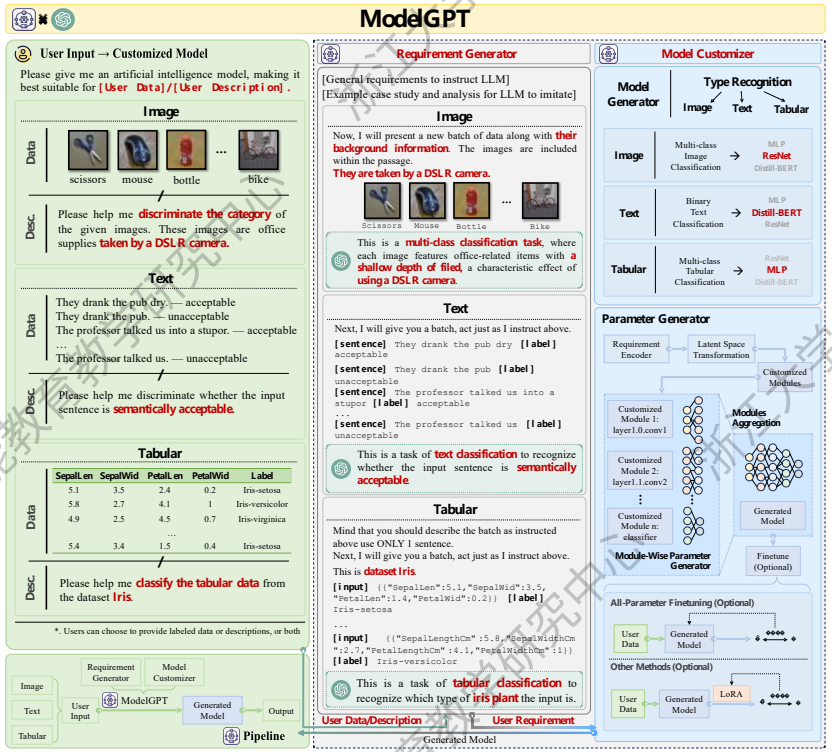
联合应用平台既有的特定业务小模型与云侧大模型，将端侧小模型轻量部署、快速响应、个性适配的优势，和云侧大模型认知推理、多模态理解、通用泛化的优势进行互补

基于调度的协同 → 基于反馈的协同 → 基于生成的协同 → 基于融合的协同



基于生成的协同：One（大模型） to All （小模型）生成

- 大模型驱动的小模型生成框架ModelGPT
  - ModelGPT + 用户对**模型的需求描述** + **少量数据** =(推理生成) 开箱即用小模型。在 All-in-One 的通用大模型范式之外，初步探索 **One-to-All** 的可能性，为更广泛的小数据、小算力（边端）、离线应用场景提供AI落地支撑。
  - 在**NLP，CV，和Tabular Data**典型数据集上进行验证，**性能超越Finetune**方法。



Algorithm 1 Pseudo-code of Parameter Generator $P(\cdot;\theta_p)$

Require: $A = \{(D_i = \{X_i, Y_i\}, r_i)\}_{i=1}^{N}$
Ensure: $\theta_p = (\theta_e, \theta_m, \theta_g)$ satisfies Equation (5)

```
i ← 1
for _ = 0 to ♯epoch do
    for (D_i, r_i) in A do
        for batch in D_i do
            Obtain θ_t with Equations (1) to (3)
            Use batch to compute the loss and update θ_t
            Compute the difference Δθ_t of θ_t
            Use Δθ_t to compute the gradients of θ_p
            Update θ_p
        end for
    end for
    Save best checkpoint according to Equation (5)
end for
```

Zihao Tang, Zheqi Lv, Shengyu Zhang, Fei Wu, Kun Kuang: ModelGPT: Unleashing LLM's Capabilities for Tailored Model Generation. CoRR abs/2402.12408 (2024)

基于生成的协同：One（大模型） to All （小模型）生成

- 大模型驱动的小模型生成框架ModelGPT
  - 在**NLP，CV，和Tabular Data**典型数据集上进行验证，**性能超越Finetune**方法。
  - 给定用户的需求ModelGPT能够以至多先前范式（例如全参数微调、LORA微调）**270倍速度**快速生成定制好的人工智能模型。

Results on GLUE Benchmark (Distil-BERT)

Methods	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	DM	Score	♯Epoch	E2E Runtime
LoRA	**48.3**	91.0	84.9 / 80.3	81.2 / 80.0	68.9 / 87.3	80.5	33.1	**88.1**	52.8	**65.1**	0.0	71.5	20	216.1
Finetune	45.5	**91.3**	**86.6 / 80.8**	**82.1 / 80.9**	**69.2 / 87.8**	**81.8**	**80.8**	87.6	56.9	63.7	**35.6**	**74.4**	20	273.8
ModelGPT	39.5	88.9	85.3 / 78.4	80.9 / 80.3	63.3 / 83.5	77.8	78.0	84.6	69.5	64.4	28.0	73.4	**0**	**1.0**
ModelGPT-F	36.9	90.8	85.5 / 79.4	81.3 / 80.5	67.0 / 86.6	77.8	78.1	85.8	**70.0**	62.3	29.9	73.8	1	3.1

Results on Tabular Data (MLP)

Methods	Iris	Heart Disease	Wine	Adult	Breast Cancer	Car Evaluation	Wine Quality	Dry Bean	Rice	Bank Marketing	Average	♯Epoch	E2E Runtime
LoRA	93.3	**63.0**	67.3	54.7	95.9	71.3	55.0	88.9	92.5	89.8	77.2	20	46.2
Finetune	88.9	54.3	89.1	**55.2**	**96.5**	71.0	55.3	**90.6**	**93.1**	89.9	78.4	20	39.5
ModelGPT	**100.0**	60.9	**94.5**	54.7	95.3	**71.5**	54.1	85.0	92.5	89.8	79.8	**0**	**1.0**
ModelGPT-F	**100.0**	62.0	**94.5**	55.1	95.9	71.3	**55.4**	88.8	92.9	**90.0**	**80.6**	1	2.3

Results on Office-31 (ResNet-50)

Domain	Amazon			DSLR			Average			Webcam (ModelGPT is Zero-Shot)			♯Epoch	E2E Runtime
Methods	Acc	Acc@3	Acc@5	Acc	Acc@3	Acc@5	Acc	Acc@3	Acc@5	Acc	Acc@3	Acc@5		
LoRA	66.4	77.7	84.8	78.4	92.2	96.1	72.4	85.0	90.5	72.5	87.5	93.8	400	231.8
Finetune	67.5	79.2	83.7	84.3	98.0	**100.0**	75.9	88.6	91.9	**90.0**	**100.0**	**100.0**	400	257.6
ModelGPT	66.4	79.9	83.7	**92.2**	**100.0**	**100.0**	79.3	90.0	91.9	76.2	87.5	91.2	**0**	**1.0**
ModelGPT-F	**67.8**	**81.3**	**85.9**	**92.2**	**100.0**	**100.0**	**80.0**	**90.7**	**92.8**	77.5	90.0	91.3	1	1.2

跨越异构模型、任务、模态的统一模型知识迁移框架

研究背景

现有知识迁移方法（例如，知识蒸馏，迁移学习）要求端云具有相似的任务类型或模型架构，难以应用于**跨异构模型、任务和模态**的异构知识迁移场景。

挑战

模型知识统一表示

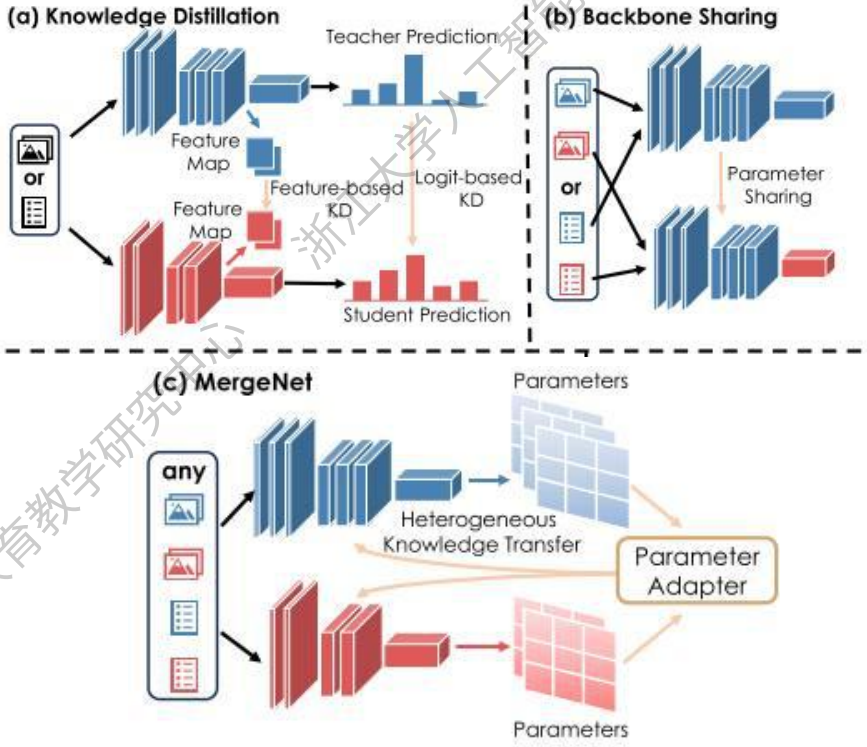
知识蒸馏利用**Logits**和**Feature Map**表示知识，依赖于任务类型。

迁移学习通常通过**共享参数**实现知识迁移，依赖于模型架构。

异构模型知识适配

异构模块（线性层 <-> 注意力机制模块）之间知识不兼容。

不同规模模型之间知识不兼容。



Li K, Zhan T, Fu K, Zhang S, et al. MergeNet: Knowledge Migration across Heterogeneous Models, Tasks, and Modalities. AAAI 2025

跨越异构模型、任务、模态的统一模型知识迁移框架

研究问题

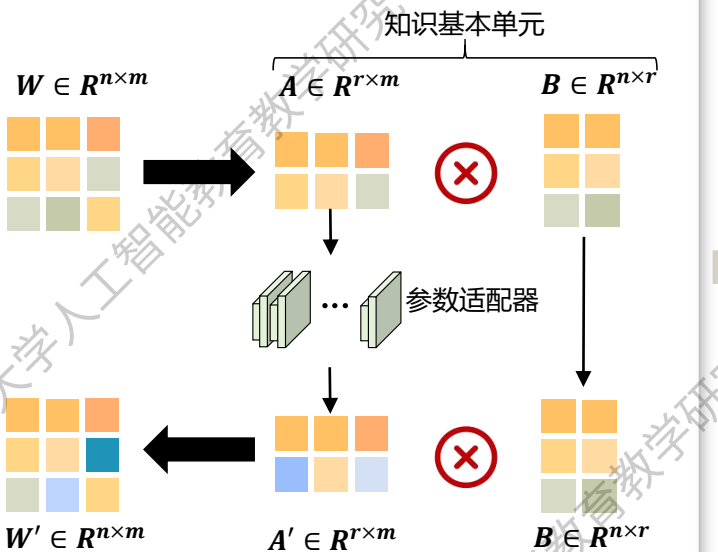
研究基于端云协同的跨异构模型架构、任务和模态的异构知识迁移框架。

创新方法

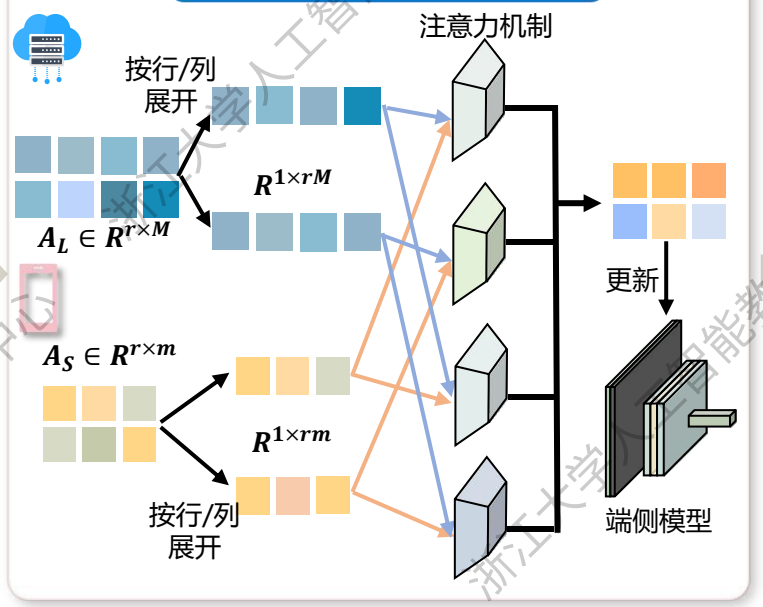
异构模型知识表示：以参数为载体，重新编码端云模型参数，实现对异构知识的统一表示

异构知识适配：设立参数适配器，促进异构参数空间的交互，提取并对齐有效的信息，实现高效知识迁移

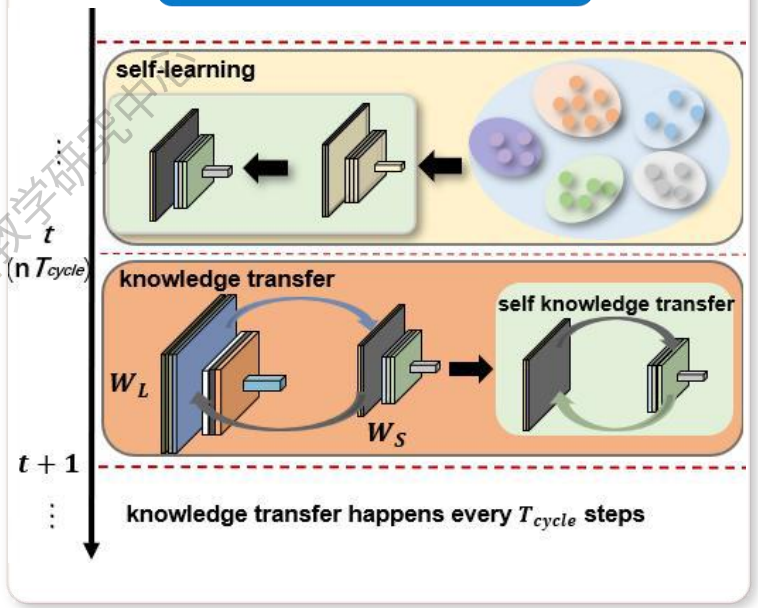
异构模型知识表示



异构知识适配



协同训练流程



Li K, Zhan T, Fu K, Zhang S, et al. MergeNet: Knowledge Migration across Heterogeneous Models, Tasks, and Modalities. AAAI 2025

跨越异构模型、任务、模态的统一模型知识迁移框架

应用验证

克服了传统知识迁移需要具有相似任务类型或模型架构的限制

有效应用于各种具有挑战性的场景，及传统知识迁移方法不适用的场景

传统知识迁移存在的问题

模型结构差异性限制
任务类型匹配要求
异构知识表示不兼容

统一异构知识表示
知识交互融合
…

异构知识迁移

跨架构知识迁移

Methods	Top-1 Acc(%)	Top-5 Acc(%)
Vanilla MobileNetV2	63.87	88.77
KD (Hinton et al., 2015)	64.32	88.62
RKD (Park et al., 2019)	65.48	88.9
DKD (Zhao et al., 2022)	65.23	89.01
NKD (Yang et al., 2023)	65.09	88.9
MergeNet(R→M)	66.23	89.66
MergeNet(R↔M)	**66.51**	**89.75**
Vanilla ResNet50	68.11	89.61
KD(Hinton et al., 2015)	68.36	89.9
RKD(Park et al., 2019)	68.6	90.21
DKD (Zhao et al., 2022)	69.03	90.25
NKD(Yang et al., 2023)	69.27	90.18
MergeNet(R↔M)	**69.84**	**90.57**

跨模态知识迁移

Methods	VQA overall	VQA other	VQA number	ITR TR	ITR IR
Vanilla	45.78	31.33	28.71	41.48	37.64
MergeNet(V→T)	46.33	33.29	31.33	44.72	39
MergeNet(T→V)	45.96	31.99	31.15	44.58	38.93
MergeNet	**46.51**	**33.84**	**31.54**	**44.78**	**39.26**

跨任务知识迁移

Methods	SQuAD v2.0 EM↑	SQuAD v2.0 F1↑	IMDb Err↓
Vanilla	70.17	73.06	8.02
MergeNet	71.89	75.43	7.5

Methods	CIFAR-100 Top-1 Acc	CIFAR-100 Top-5 Acc	SQuAD v2.0 EM	SQuAD v2.0 F1
Vanilla	63.87	88.77	70.17	73.06
MergeNet	65.56	88.74	70.89	74.15

Li K, Zhan T, Fu K, Zhang S, et al. MergeNet: Knowledge Migration across Heterogeneous Models, Tasks, and Modalities. AAAI 2025

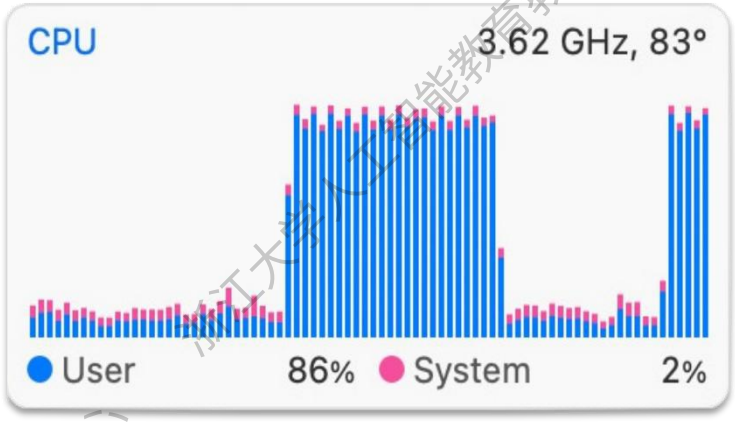
大小模型端云协同

≈

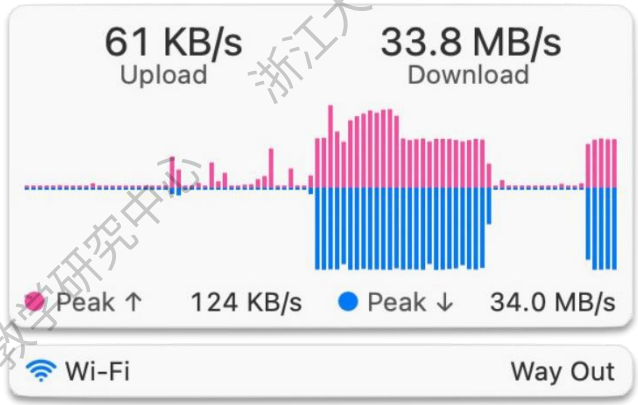
大小模型协同 + 端云高效协同

云智能的局限

移动端智能的局限

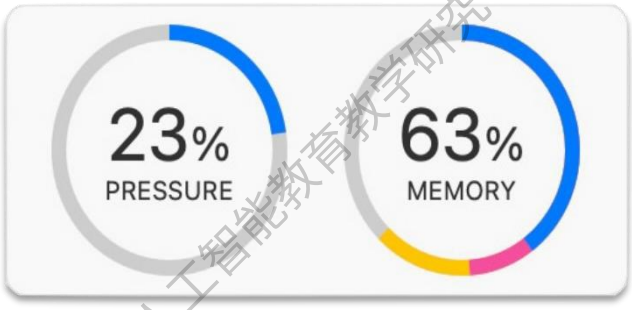


算力限制



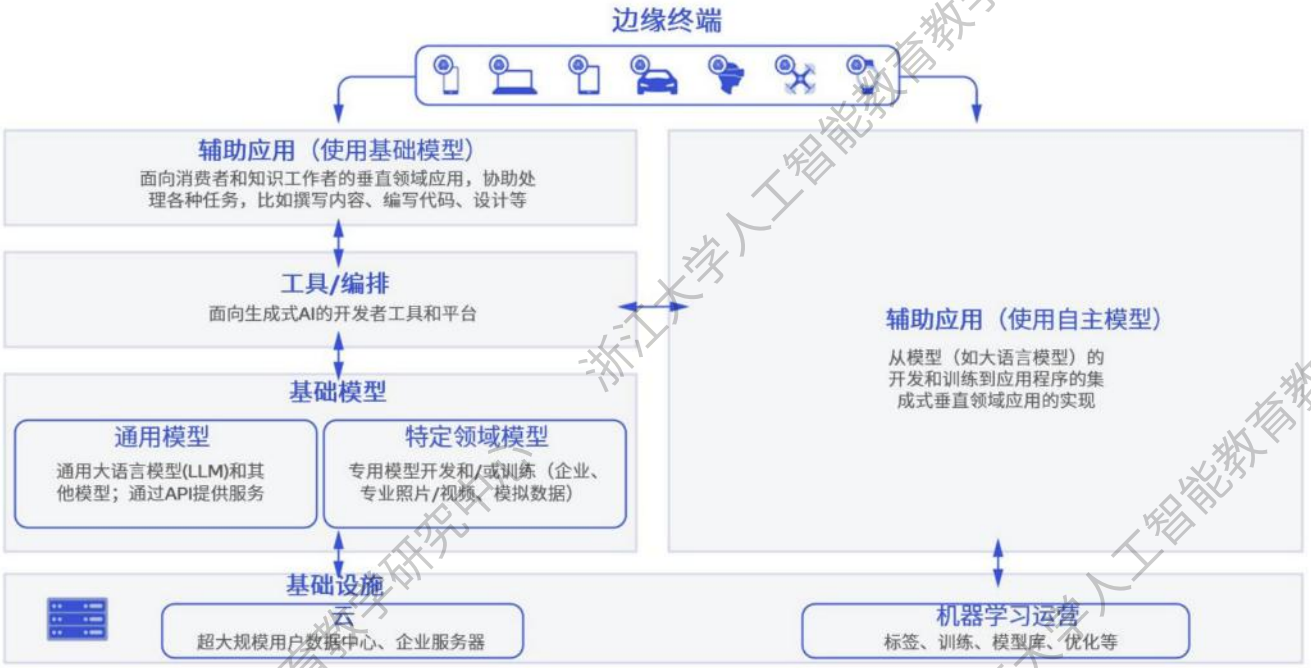
带宽限制

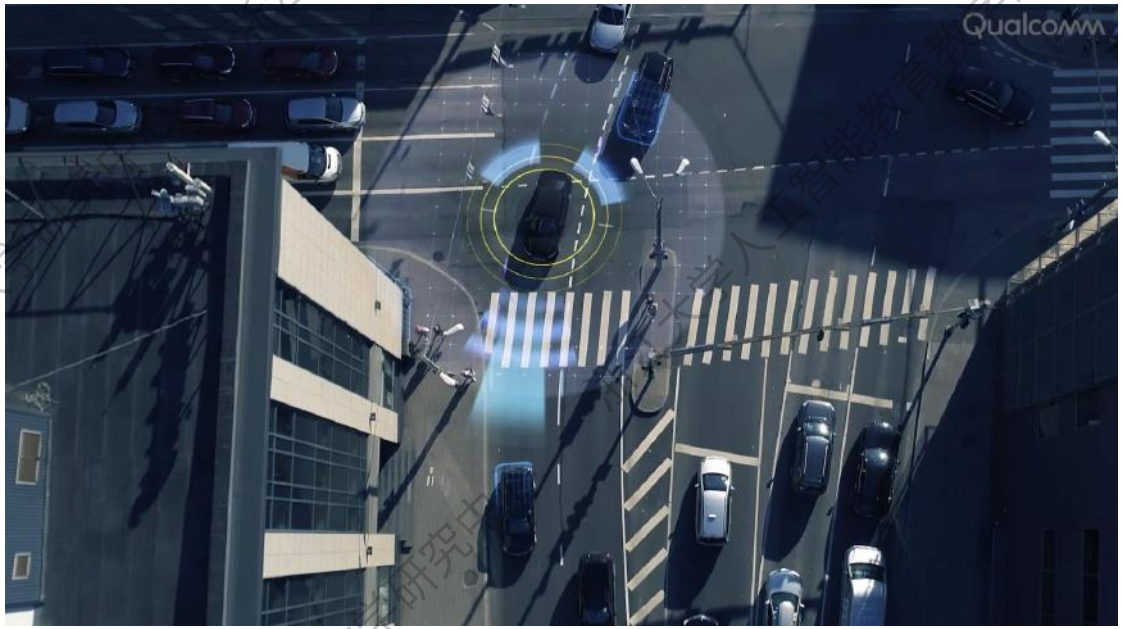
电量限制



内存限制

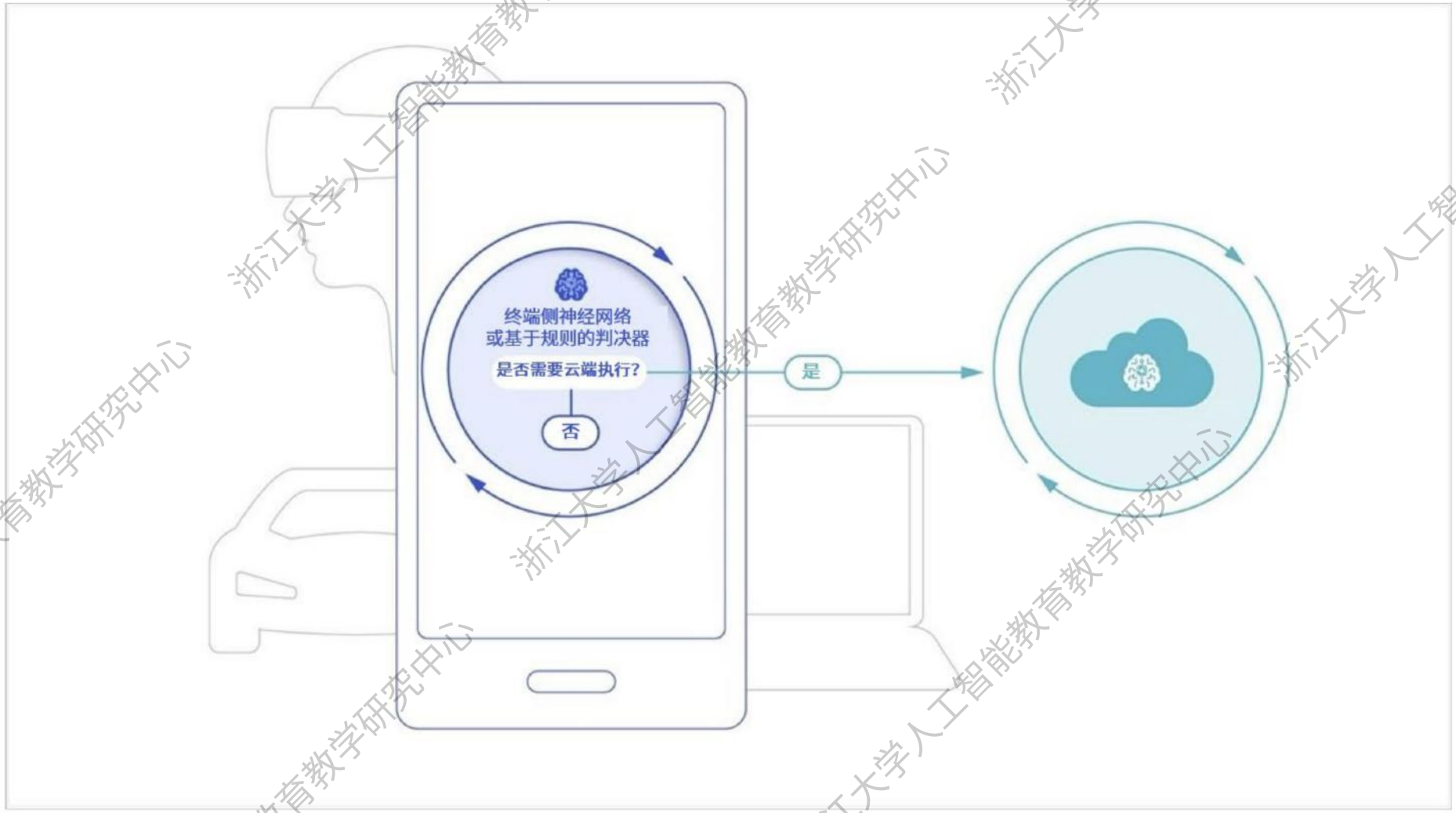
高通：生成式端云混合智能





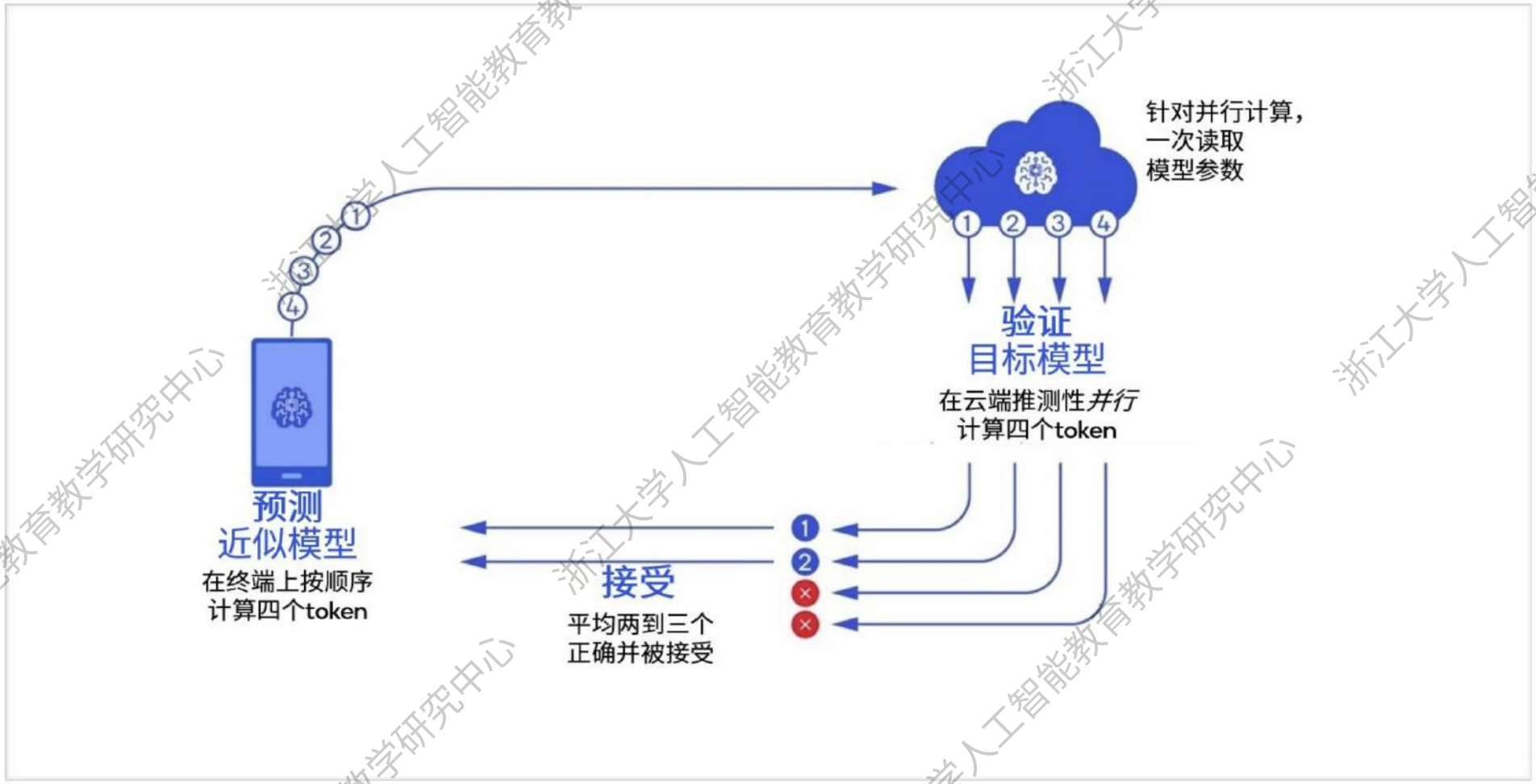
- 混合AI指终端和云端协同工作，在适当的场景和时间下分配AI计算的工作负载，以提供更好的体验，并高效利用资源。在一些场景下，计算将主要以终端为中心，在必要时向云端分流任务。而在以云为中心的场景下，终端将根据自身能力，在可能的情况下从云端分担一些AI工作负载。

端云协同智能



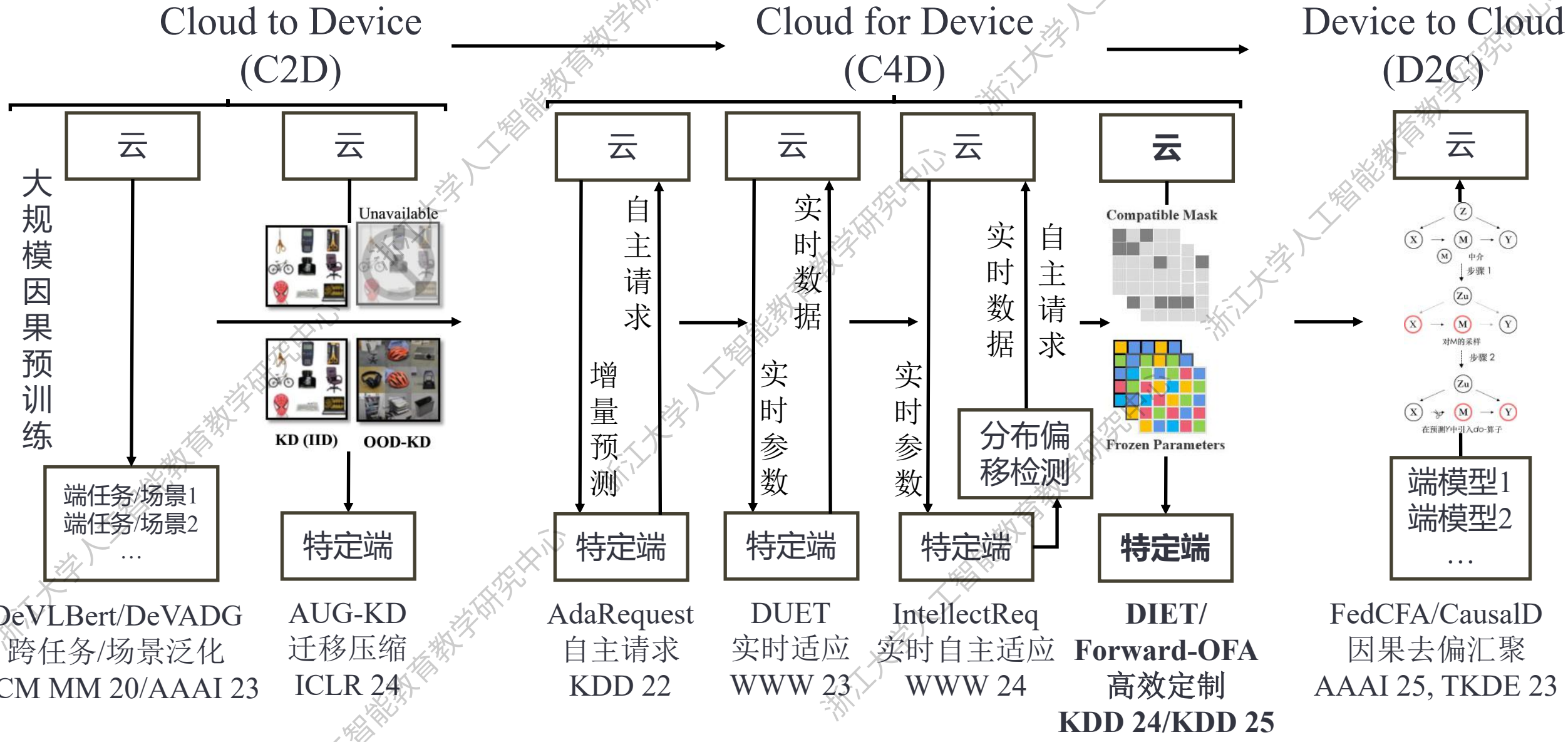
-- 高通《终端侧AI 和混合AI 开启生成式AI 的未来》

端云协同智能



-- 高通《终端侧AI 和混合AI 开启生成式AI 的未来》

端云异构模型知识互迁与协同推断



面向未知端侧分布的压缩-适应联合

研究背景

- 大模型向端侧迁移部署往往采用知识蒸馏等压缩手段，传统知识整理方法假设大模型训练数据分布（压缩前）和小模型测试数据分布（压缩后）服从独立同分布假设（IID Hypothesis）。
- 实际应用中，源域数据和应用场景存在**分布偏移，导致压缩性能显著下降。**

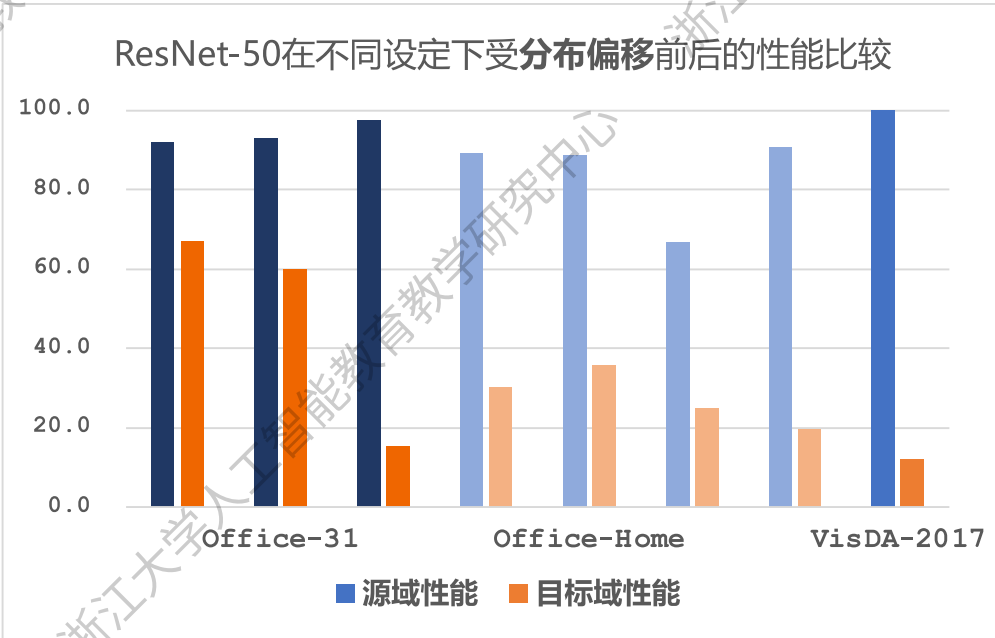
理论分析

独立同分布假设（IID Hypothesis）：**源域**P_s和**目标域**P_t（应用场景）独立同分布。在此情况下进行知识蒸馏，源域的知识可以很好地指导模型完成目标域的任务。

- 数据蒸馏的目标：

$$\min_{\theta_s} \mathbb{E}_{(x,y)\sim P}[D_{KL}(T(x;\theta_t) \parallel S(x;\theta_s) + CE(S(x;\theta_s), y)].$$

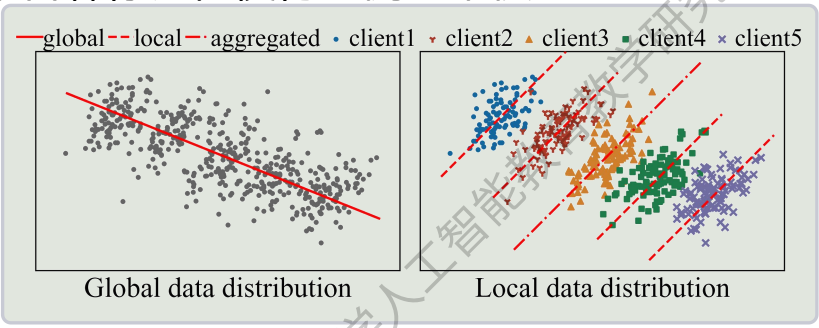
- 多数场景下，源域分布和应用场景存在**分布偏移**（$P_t \neq P_s$），违反独立同分布假设。
  - 情况1：$P \approx P_t$，对应无数据蒸馏方法（P_t 由生成器拟合），蒸馏出的目标模型并不适用P_s。
  - 情况2：$P \approx P_s$，源模型给出的知识不一定有效。



利用端侧反事实表征学习实现端向云去偏汇聚

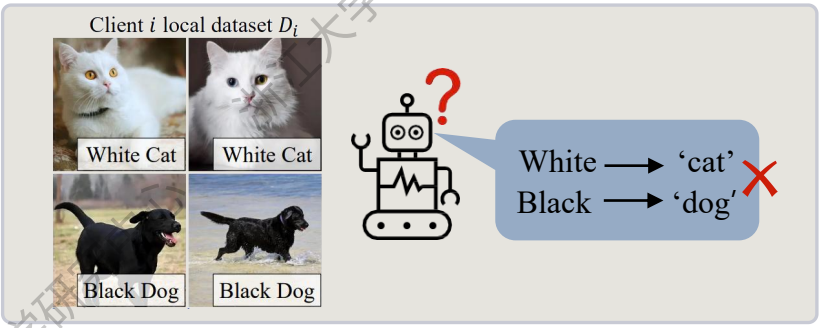
研究背景

数据分布异质性导致的“局部观察到的趋势在全局数据中消失或反转”的辛普森悖论，使得云侧汇聚模型无法准确反映整体数据分布，给端向云去偏汇聚带来了巨大挑战



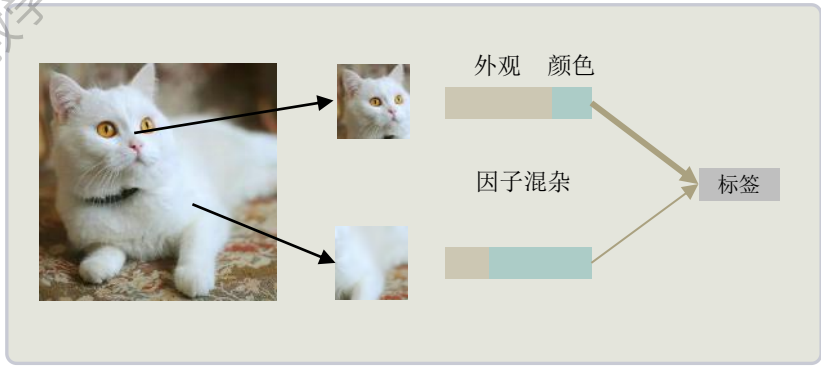
分布异质

- **端云分布异质**：云侧全局数据分布体现平台整体共性与端侧特化分布存在偏移
- **端云有偏汇聚**：有偏数据导致端侧偏见，相似偏见端侧模型导致云侧有偏汇聚



因子混杂

- **虚假相关**：端侧数据局部且有限，存在虚假的因子-标签关联，忽视真实因果关系
- **因子耦合**：因子之间存在复杂的相互依赖关系，难以有效解耦出独立的因果关系



Jiang Z, Xu J, **Zhang S**, et al. FedCFA: Alleviating Simpson's Paradox in Model Aggregation with Counterfactual Federated Learning. AAAI 2025

利用端侧反事实表征学习实现端向云去偏汇聚

研究问题

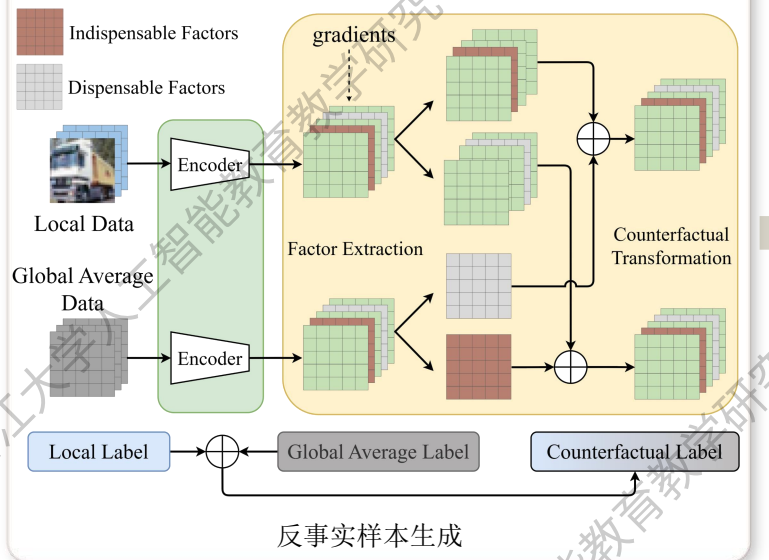
利用端侧反事实表征学习解决云侧模型联邦汇聚中“辛普森悖论”难题。

创新方法

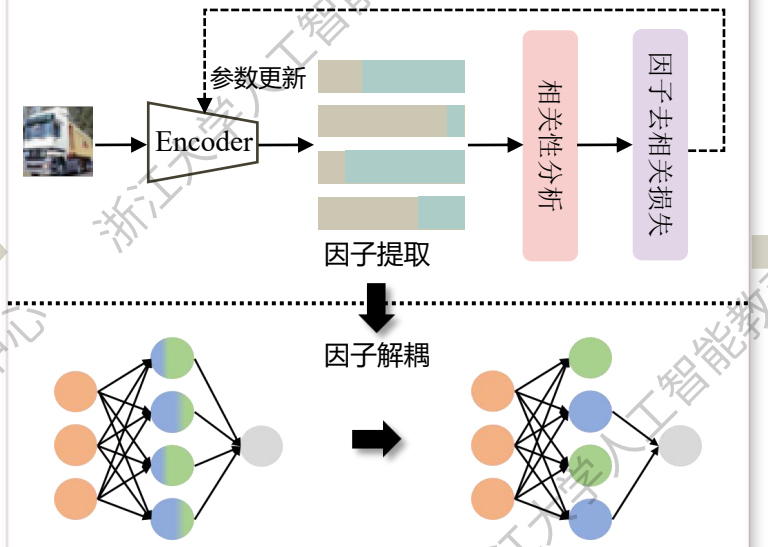
反事实表征学习：利用全局平均数据信息在端侧生成反事实样本，实现端侧模型去偏训练

因子去相关模块：基于相关性分析设计因子去相关模块对因子解耦，提高反事实样本的质量

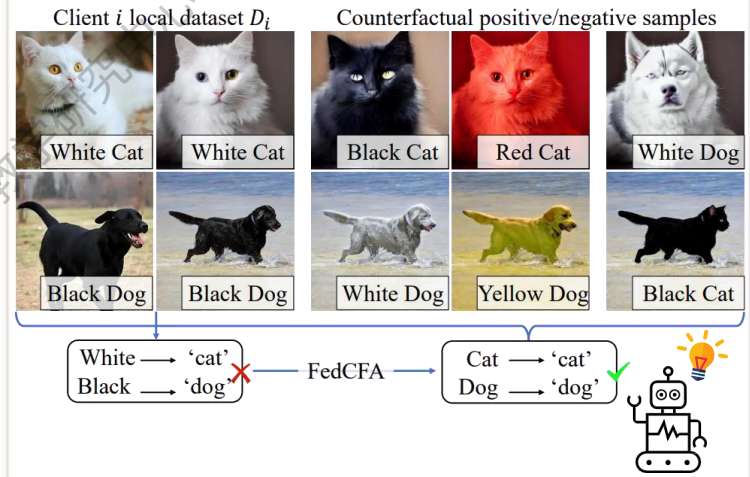
反事实表征学习



因子去相关模块



缓解辛普森悖论

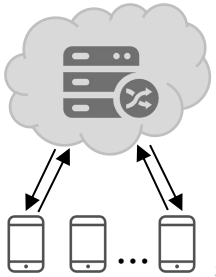


Jiang Z, Xu J, **Zhang S**, et al. FedCFA: Alleviating Simpson's Paradox in Model Aggregation with Counterfactual Federated Learning. AAAI 2025

利用端侧反事实表征学习实现端向云去偏汇聚

实验验证

当前端云协同存在的问题



数据高度异质性
云端分布差异大
云侧模型收敛慢

反事实样本生成
因子去相关约束
混杂因子解耦合
…

端侧反事实表征学习

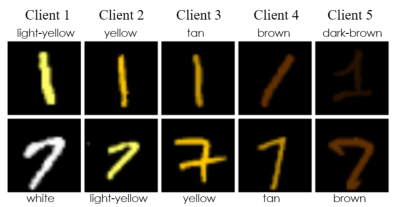
端-云模型协同
…

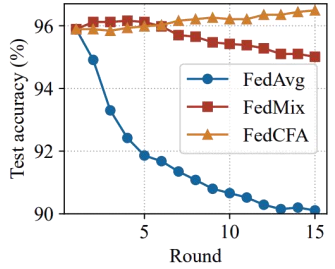
相比于主流联邦学习的最佳方法，云侧模型**精度最高可提升7.75%**
云侧模型去偏汇聚的同时**收敛速度提升2倍**

	Method	$Dir_{60}(0.2)$	$Dir_{60}(0.6)$	IID_{60}	$Dir_{100}(0.2)$	$Dir_{100}(0.6)$	IID_{100}
CIFAR100	FedAvg	40.70±0.24	42.85±0.26	44.37±0.40	38.17±0.32	40.19±0.26	42.19±0.52
	FedProx	40.39±0.23	42.51±0.34	44.21±0.64	38.27±0.46	39.90±0.42	42.24±0.98
	SCAFFOLD	29.36±0.39	33.30±0.48	37.96±0.26	23.25±0.54	29.98±0.19	32.77±0.25
	FedPRV	38.35±1.11	42.91±0.49	45.91±0.05	30.65±0.74	36.58±0.14	39.96±0.30
	q-FedAvg	40.34±0.60	42.61±0.63	44.43±0.28	38.15±0.48	40.20±0.10	42.04±0.65
	FedMix	42.51±0.28	44.16±0.26	45.65±0.31	39.78±0.07	41.43±0.84	43.63±0.64
	FedCFA	**46.96±1.04**	**49.32±0.20**	**48.31±0.53**	**46.71±0.59**	**49.18±0.75**	**47.86±1.22**
CIFAR10	FedAvg	65.88±0.32	73.95±0.16	75.43±0.51	62.87±0.12	70.99±0.70	72.82±0.35
	FedProx	72.23±0.44	77.68±0.03	76.93±0.28	70.36±0.75	75.47±0.53	73.36±0.47
	SCAFFOLD	33.05±5.57	54.58±3.95	75.96±0.57	34.69±2.16	56.17±1.91	71.84±0.77
	FedPRV	59.42±2.26	71.52±0.21	77.42±0.04	55.43±1.74	67.11±0.76	76.16±0.37
	q-FedAvg	71.71±1.05	77.96±0.19	76.92±0.09	70.04±1.55	75.47±0.52	73.68±0.33
	FedMix	74.61±0.74	78.64±0.53	77.90±0.17	73.91±0.79	77.11±0.31	73.93±0.06
	FedCFA	**75.89±1.00**	**82.43±0.08**	**83.36±0.51**	**75.76±0.15**	**81.73±0.12**	**81.68±0.89**

Method	Tiny-ImageNet		FEMNIST	Sent140
	$Dir_{60}(0.2)$	$Dir_{60}(0.6)$	Non-IID	Non-IID
FedAvg	27.39±0.13	30.90±0.29	81.31±0.94	68.10±0.48
FedProx	27.34±0.05	30.78±0.16	81.63±0.08	68.10±0.44
q-FedAvg	26.89±0.07	30.70±0.13	81.90±0.58	68.04±0.71
FedMix	28.01±0.19	32.43±0.13	82.31±0.21	67.97±0.39
FedCFA	**30.70±0.68**	**32.86±0.77**	**83.19±0.54**	**69.26±0.37**

Method	$Dir_{60}(0.2)$	IID_{60}	$Dir_{100}(0.2)$	IID_{100}
Target	75.5	78.4	73.6	75.6
FedAvg	(>,66.22)	(>,74.99)	(>,62.95)	(>,72.42)
FedProx	(>,74.03)	(>,77.16)	(>,72.37)	(>,73.46)
SCAFFOLD	(>,32.77)	(693,78.46)	(>,38.34)	(800,75.68)
q-FedAvg	(>,72.23)	(>,77.04)	(>,72.03)	(>,73.56)
FedMix	(610,75.69)	(>,77.79)	(>,72.25)	(>,74.00)
FedCFA	**(375,75.58)**	**(453,78.41)**	**(427,73.61)**	**(408,75.67)**





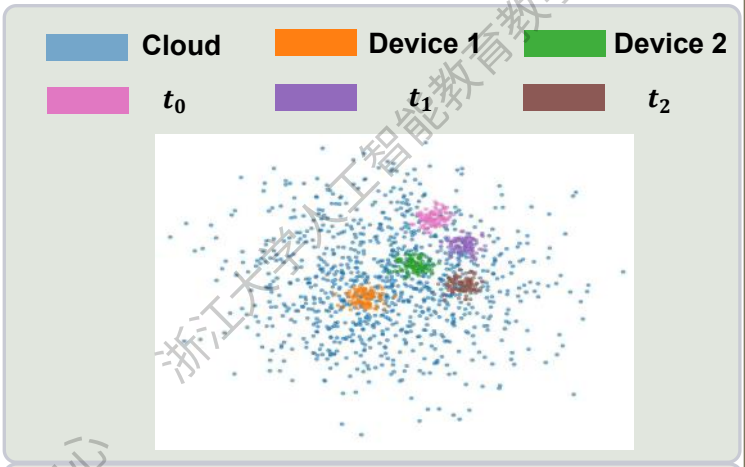
突破了端云协同计算在**分布偏移、数据异质**场景中模型汇聚效率局限

基于端云协同的高效端模型参数定制

研究背景 现有端侧部署方案采用云侧大规模预训练，通过模型压缩后传输至端侧进行部署。然而多阶段训练、稠密信息传输给端侧动态复杂环境下的**高响应、低成本**自适应带来了巨大挑战

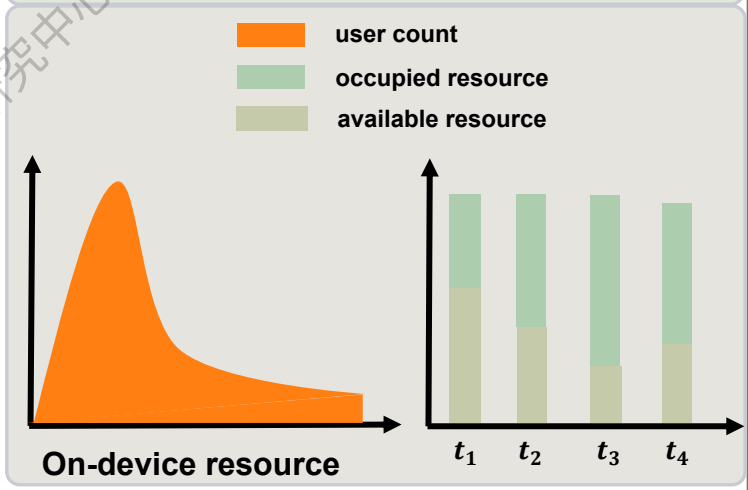
分布异质性

- **端云分布异质**：云侧全局数据分布体现平台整体共性与端侧特化分布存在偏移
- **端侧分布迁移**：端侧用户兴趣意图动态偏移，需要由云向端及时下发适配模型



资源异质性

- **端侧计算资源有限**：大量长尾用户移动设备算力有限，难以支撑本地训练微调
- **端云通信资源有限**：频繁下发稠密适配模型消耗大量通信带宽资源，降低响应



Fu K, Zhang S, Lv Z, et al. DIET: Customized Slimming for Incompatible Networks in Sequential Recommendation. KDD 2024 Research Track

基于端云协同的高效端模型参数定制

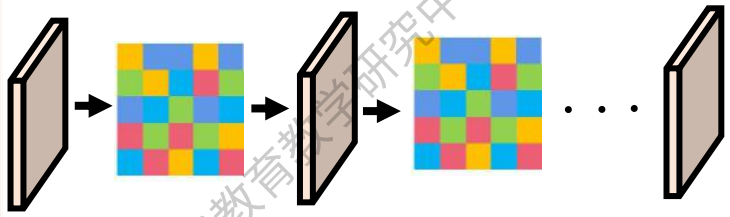
研究问题　研究基于端云协同的低通信开销、高响应速度端模型定制算法。

创新方法

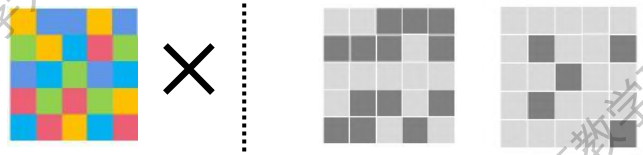
高效模型表示构建：基于神经网络彩票假说，将云向端训练压缩过程转化为传输适配子网二进制掩膜

高效适配子网搜索：云侧学习建立实时数据到端侧个性子网掩膜的映射，仅需前向推理即可高效响应

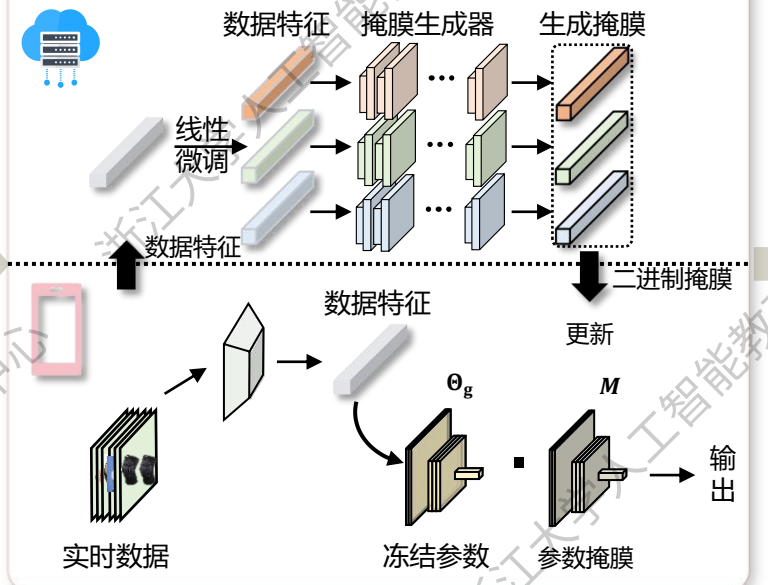
彩票假说理论



利用掩膜进行选择（一层参数多掩膜）

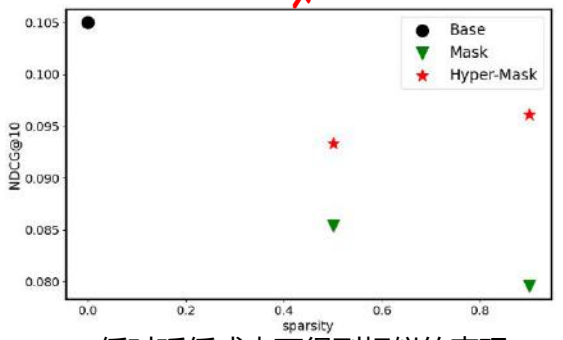


端云子网搜索



模型效率提升

优势 \ 方法	Base	Ours
低传输延迟	✗	✓
低存储成本	✗	✓
低推理时延	✗	✓



低时延低成本下得到相似的表现

Fu K, Zhang S, Lv Z, et al. DIET: Customized Slimming for Incompatible Networks in Sequential Recommendation. KDD 2024 Research Track

基于端云协同的高效端模型参数定制

应用验证

当前推荐系统存在的问题

通信开销大
云端分布差异大
端侧兴趣变化快
设备计算资源有限

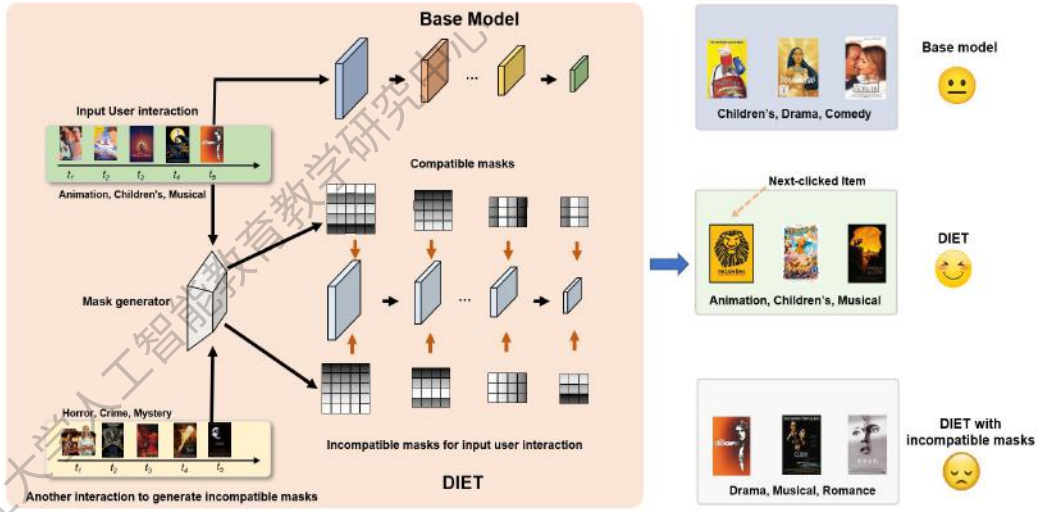
瘦身子网模型压缩
端侧实时兴趣提取
适配子网生成传输
…

端侧个性子网搜索

共性-个性协同
大-小模型协同
…

降低模型由云向端下发的传输开销至**原始大小的3%**
端侧模型能力提升的同时**推理速度提升5倍**

Model	Method	Dataset							
		Amazon-CD				MovieLens-100k			
		NDCG↑	Hit↑	Params↓	FLOPs↓	NDCG↑	Hit↑	Params↓	FLOPs↓
SASRec	Base	0.0386	0.0529	1.3107	0.2086	0.0517	0.1077	1.3107	0.2086
	DIET	**0.0425**	**0.0590**	**0.0410**	**0.1154**	**0.0635**	**0.1319**	**0.0410**	**0.0416**
	Improv. ↑	10.96%	11.53%	×31.97	×1.81	22.82%	22.47%	×31.97	×5.01
Caser	Base	0.0310	0.0424	0.4922	0.0586	0.0569	0.0719	0.4922	0.0586
	DIET	**0.0356**	**0.0488**	**0.0154**	**0.0294**	**0.0617**	**0.0771**	**0.0154**	**0.0488**
	Improv. ↑	14.84%	15.09%	×31.96	×1.99	10.42%	7.32%	×31.96	×1.76



突破了端云协同计算在**分布偏移、资源受限**设备上训练推理效率局限

大小模型端云协同

≈

大小模型协同 + 端云高效协同

⇓

赋能与应用

人工智能 = 人工 + “智” + “能”

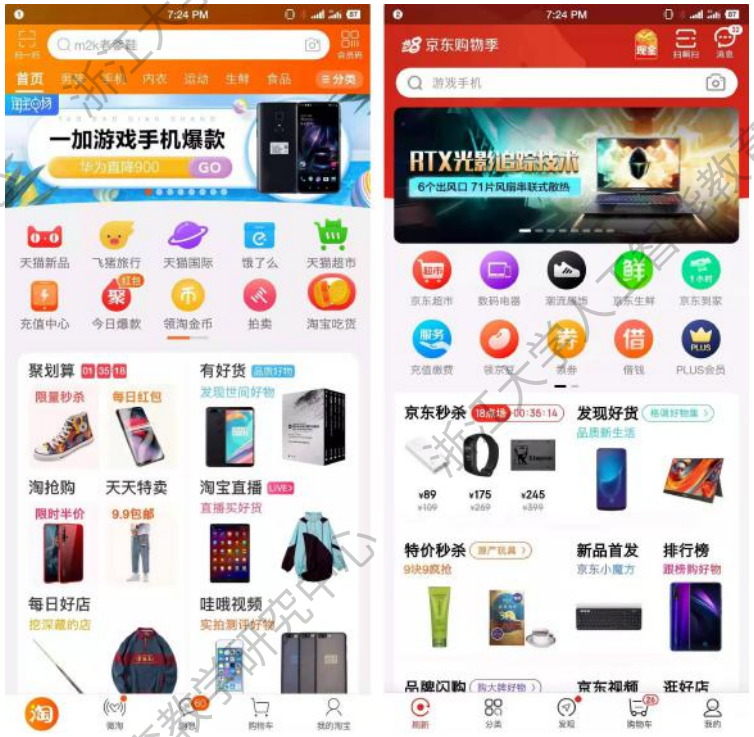
人工 rén gōng

词典解释

①人为；人做的。与“自然”、“天然”相对：人工降雨｜人工取火｜根须茁壮，枝叶繁茂，岂是人工做得出来的。

人机交互“智”：理解使用者

短视频APP、购物APP



行为数据：观看视频、停留时长、互动（点赞、评论、分享）

机器学习算法：根据历史行为预测喜好

不是真正的“理解”，而是数据驱动的预测

兴趣变化：AI能否快速适应？

不能理解情感和临时兴趣变化

如何提高灵活性和适应性？

推荐系统

信息量巨大：社交媒体、新闻、视频、广告

推荐系统帮助“过滤”信息，找到有用内容

- 实时推荐的工作原理

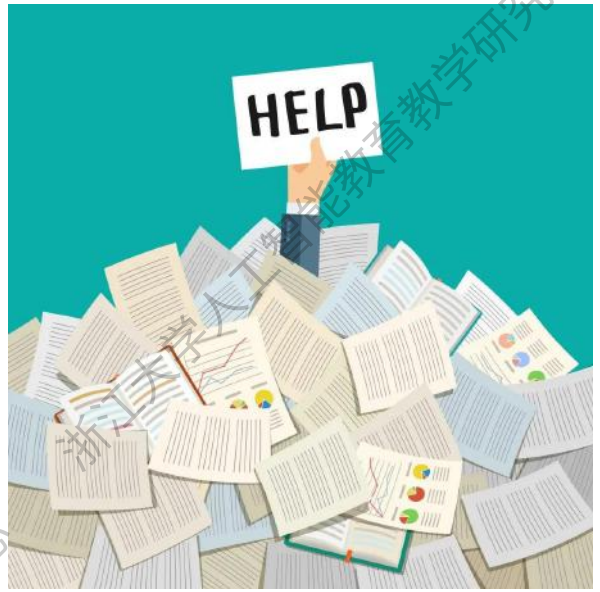
分析用户行为：点击、停留、互动

基于行为预测用户兴趣，快速推荐相关内容

- “探索”和“发现”

推荐系统帮助你发现未知的内容

根据历史行为预测可能的兴趣点

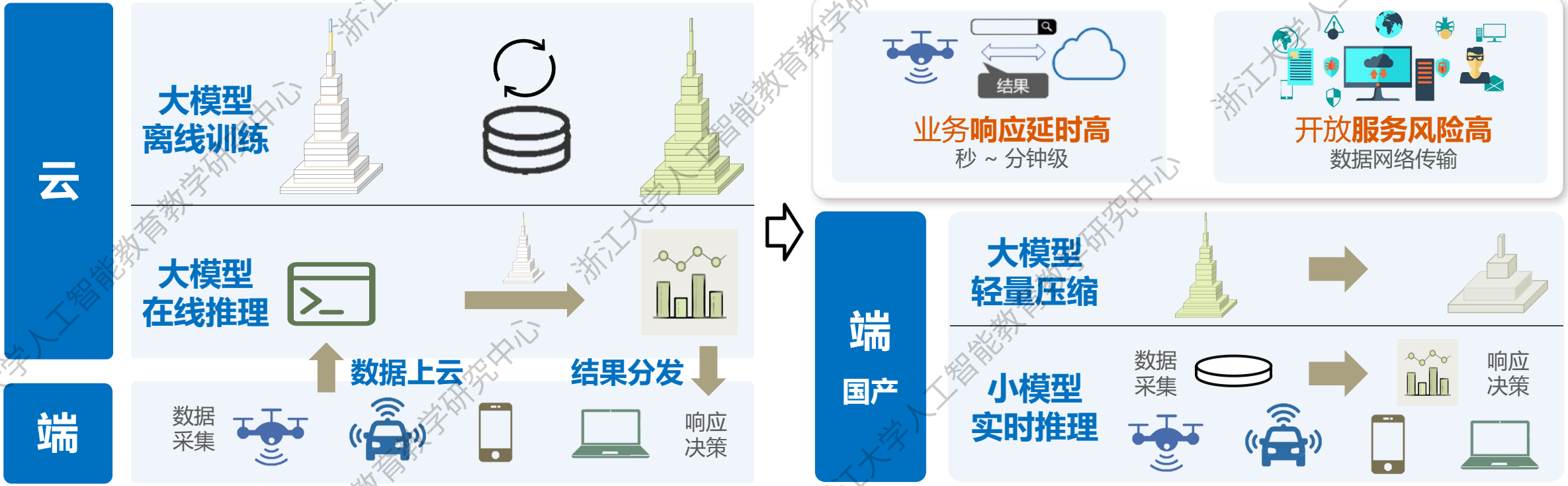


推荐系统

为什么需要移动端智能推荐?

为什么需要移动端智能推荐?

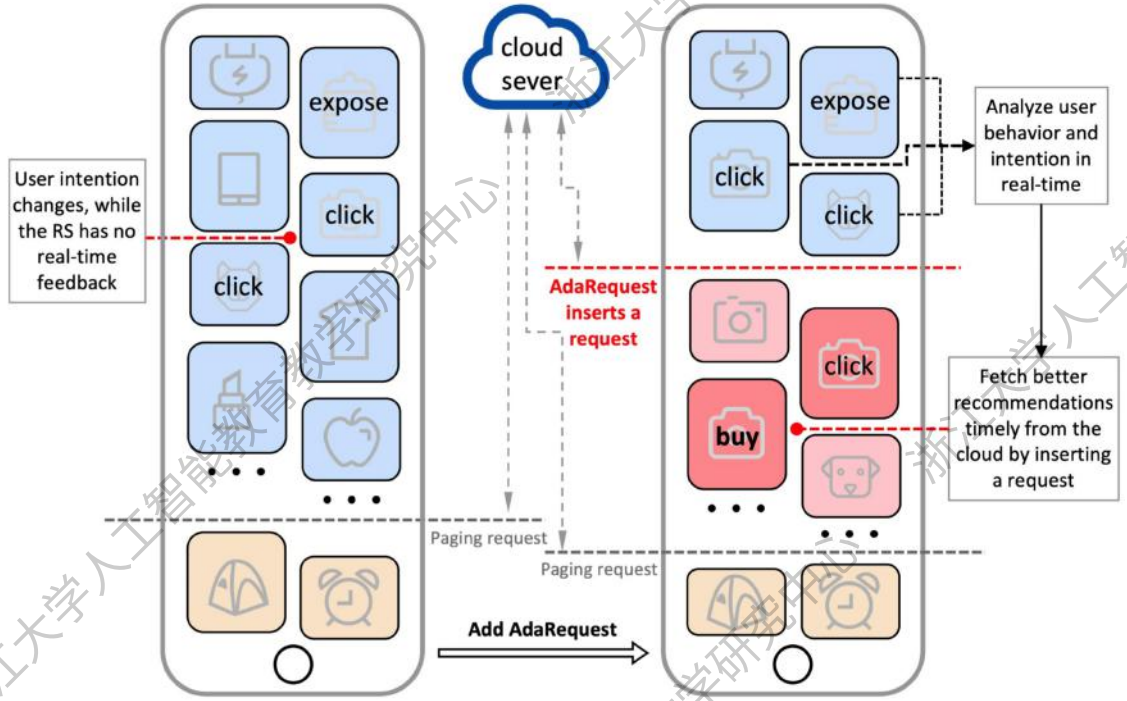
- 端侧内容生成通过部署轻量化小模型至端侧，发挥出终端设备**靠近用户和数据源**天然优势，降低智能服务延时至**毫秒级**，实现本地**私有化**响应决策。



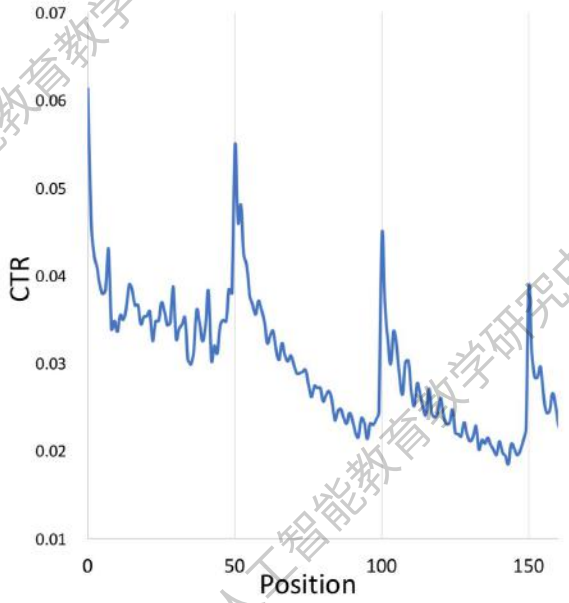
端云大-小模型协同推断算法

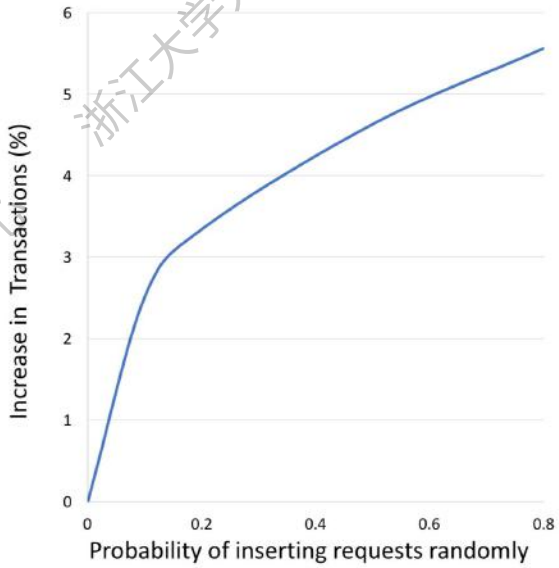
- 动态变化的端环境导致资源有限情况下云模型的延迟响应，导致端侧服务与端侧环境的不匹配，损害用户的服务体验

手机淘宝商品推荐系统



用户点击率在云模型响应后陡升

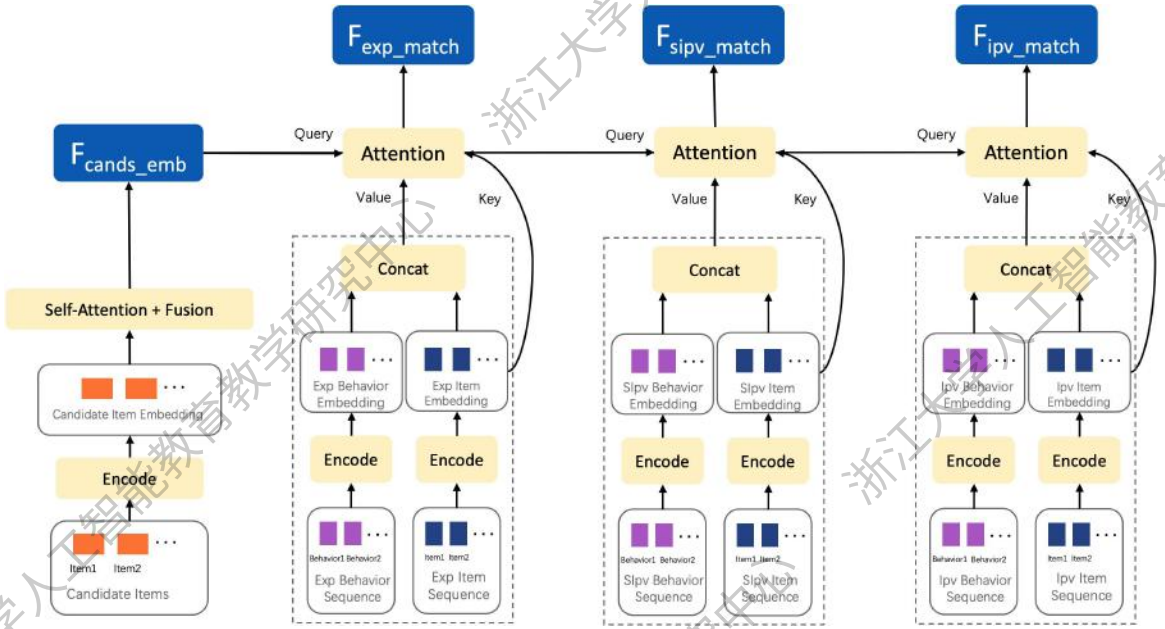


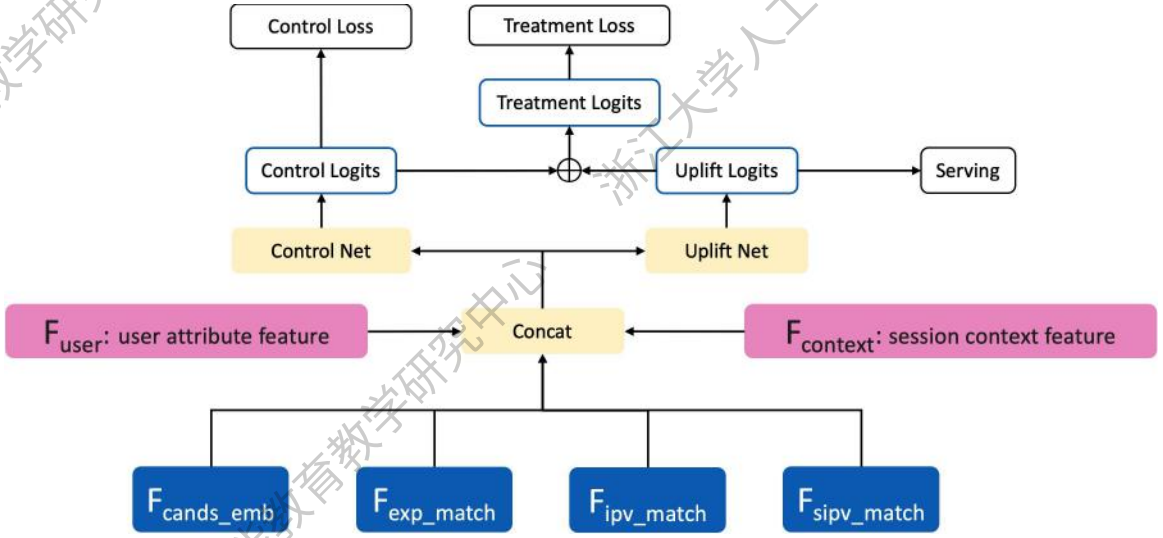


Xufeng Qian, Yue Xu, Fuyu Lv, Shengyu Zhang*, Ziwen Jiang, Qingwen Liu, Xiaoyi Zeng, Tat-Seng Chua, Fei Wu. Intelligent Request Strategy Design in Recommender System, KDD 2022

端云大-小模型协同推断算法

- 端设备部署小模型实时检测端环境变化（用户兴趣意图变化）
- 通过因果潜在结果模型预估请求大模型响应价值
- 动态规划对云侧大模型的请求,最大化资源有限时的线上收益。





Xufeng Qian, Yue Xu, Fuyu Lv, Shengyu Zhang*, Ziwen Jiang, Qingwen Liu, Xiaoyi Zeng, Tat-Seng Chua, Fei Wu. Intelligent Request Strategy Design in Recommender System, KDD 2022

端云大-小模型协同推断算法

当前推荐系统存在的问题

通信开销大
隐私破坏风险
隐时反馈噪声多
无法实时感知用户

因果结构学习机制
因果潜在结构框架
不确定性预估方法
…

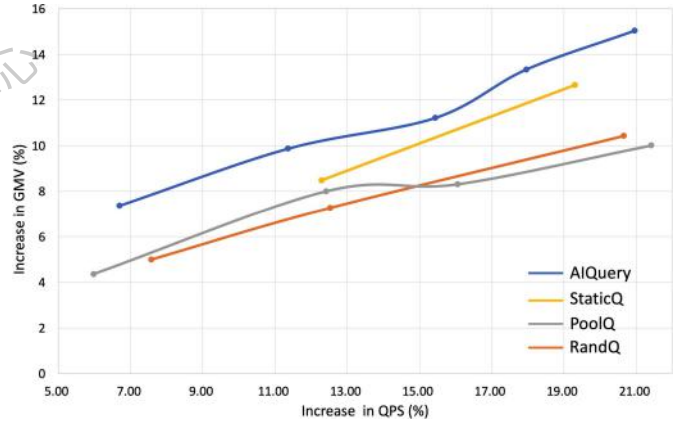
因果+端云协同

共性–个性协同
大–小模型协同
隐私–效率协同
…

直接经济效益
(购买率)

PR in N	NoQ	RandQ	PoolQ	StaticQ	AIQuery
10	0.889	1.174	1.177	1.130	**2.289**
20	1.660	2.197	2.164	2.045	**4.173**

平台经济效益
(商品交易总值)



Xufeng Qian, Yue Xu, Fuyu Lv, Shengyu Zhang*, Ziwen Jiang, Qingwen Liu, Xiaoyi Zeng, Tat-Seng Chua, Fei Wu. Intelligent Request Strategy Design in Recommender System, KDD 2022

云上大语言模型和端上小推荐模型的端云协同推荐

研究问题

研究基于端云协同的低通信开销、高响应速度端模型定制算法。

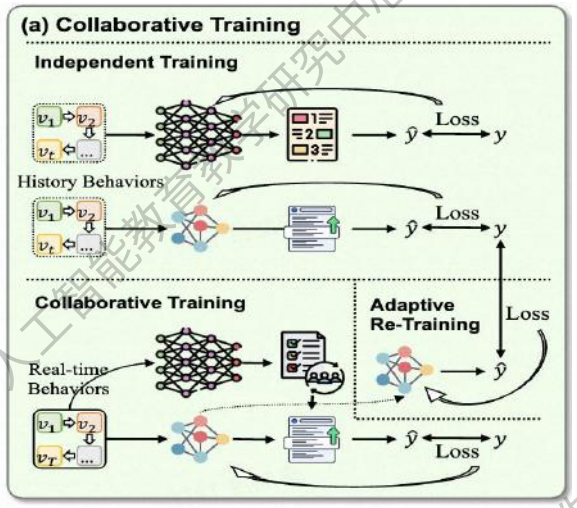
创新方法

协同训练：将云上大模型和端上小模型针对各自任务场景做针对性协作训练，提升场景适应性

协同推理：将云上大模型和端上小模型的输出结果融合，集成强泛化能力和强实时性的优势

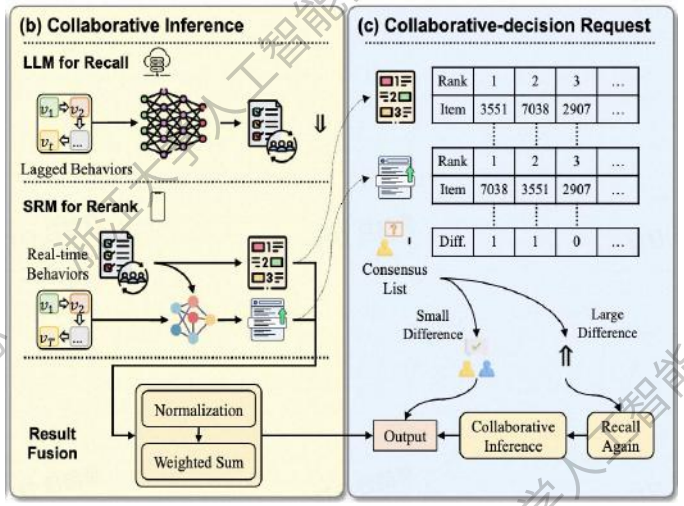
智能请求：对云上大模型和端上小模型的输出结果做不一致性检测，不一致性高的样本重新调用大模型

协同训练



大小协同训练，使小模型能针对大模型的候选列表有更强的排序能力

协同推理与请求



大小协同推理与请求，融合大小模型推理结果并决策何时调用云上大模型

模型效率提升

Dataset	Model	NDCG@5	NDCG@10	NDCG@20
Beauty	DIN (RT)	0.0079	0.0106	0.0135
	GRU4Rec (RT)	0.0081	0.0105	0.0129
	SASRec (RT)	0.0066	0.0096	0.0127
	P5 (RT)	0.0227	0.0257	0.0289
	DIN (NRT)	0.0017	0.0026	0.0042
	GRU4Rec (NRT)	0.0017	0.0023	0.0031
	SASRec (NRT)	0.0014	0.0021	0.0032
	P5 (NRT)	0.0087	0.0108	0.0136
	Ours (P5+SASRec)	**0.0094**	**0.0126**	**0.0154**
	Improve	8.41%	16.16%	13.45%
Toys	DIN (RT)	0.0046	0.0063	0.0081
	GRU4Rec (RT)	0.0050	0.0073	0.0098
	SASRec (RT)	0.0052	0.0073	0.0101
	P5 (RT)	0.0187	0.0204	0.0221
	DIN (NRT)	0.0007	0.0010	0.0017
	GRU4Rec (NRT)	0.0009	0.0015	0.0020
	SASRec (NRT)	0.0015	0.0020	0.0026
	P5 (NRT)	0.0066	0.0078	0.0090
	Ours (P5+SASRec)	**0.0066**	**0.0084**	**0.0096**
	Improve	0.46%	7.33%	6.76%

大幅补偿LLM无法获取实时数据下的推荐性能

研究成效

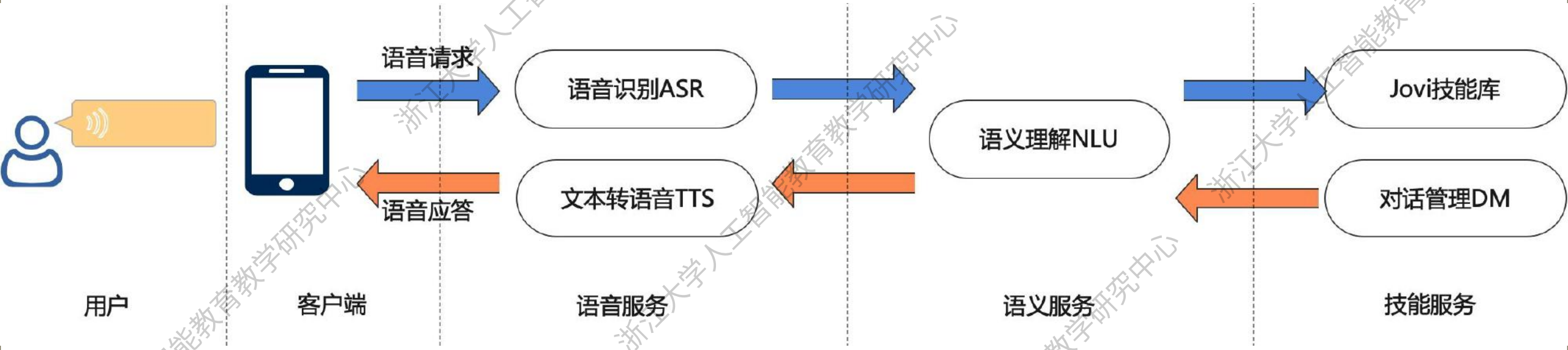
相比于先前方法显著提升，并在多个数据集上优于基线的结果，已被 KDD 2025 研究轨道录用

人机交互“能”：像人一样行动

语音助手

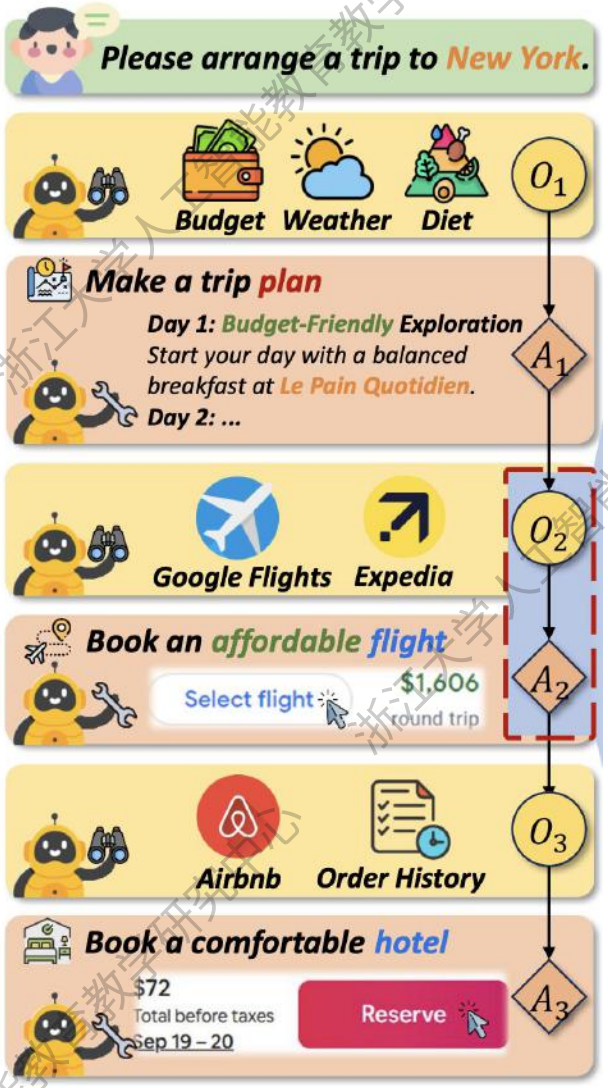
https://www.searchenginejournal.com/33-of-people-are-now-using-voice-assistants-regularly/321413/

语音助手 – 技术路线



端智能体

端智能体



需求：旅行规划

分析：预算、天气、饮食

规划：第一天、第二天。。。

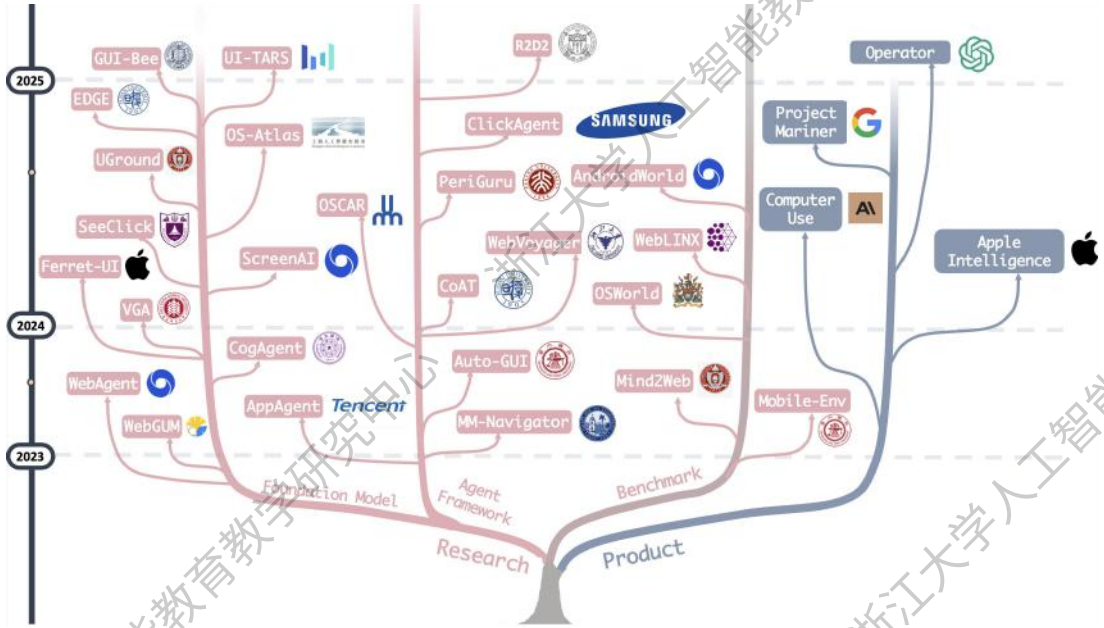
决策：使用订票软件

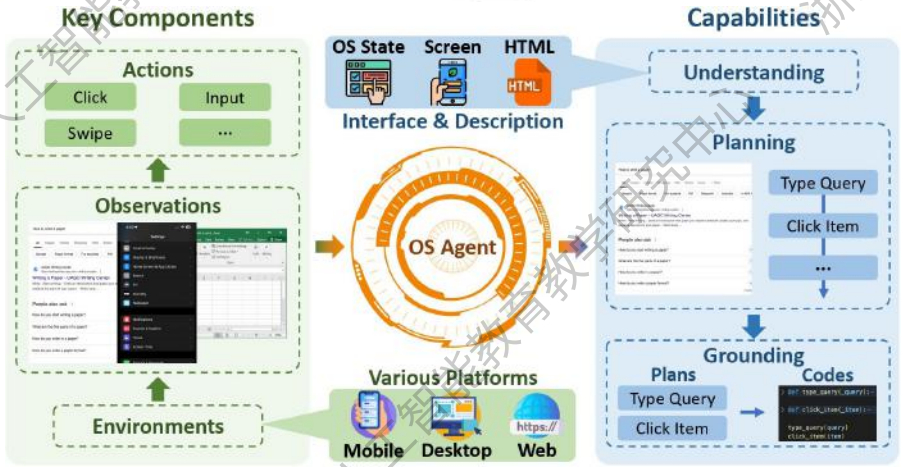
行动1：打开订票APP、点击、输入、查询。。。

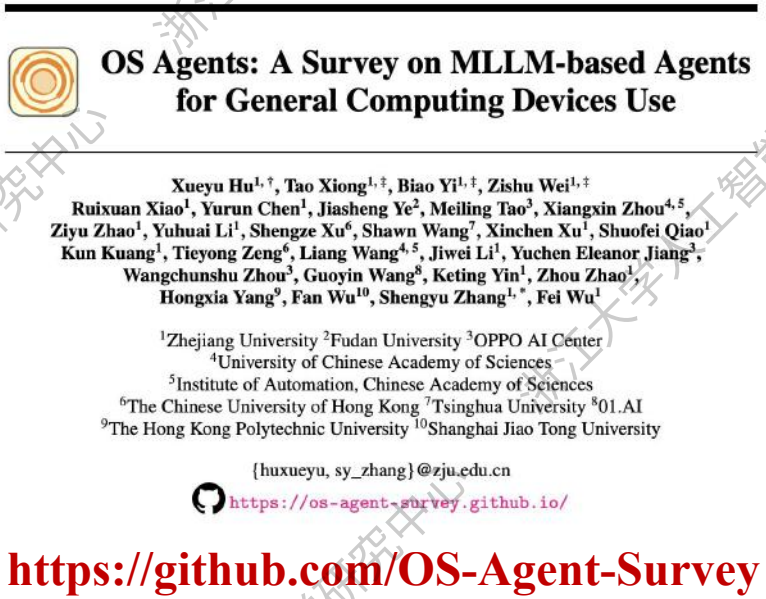
行动2：打开住宿APP、点击、输入、查询。。。

基于多模态大模型的操作系统智能体综述

- OS Agents 是一种基于（多模态）大语言模型 ((M)LLMs) 的智能代理，通过操作操作系统 (OS) 提供的环境和界面（如图形用户界面 GUI），利用计算设备（如电脑和手机）来自动执行任务。







OS Agents: A Survey on MLLM-based Agents for General Computing Devices Use

Xueyu Hu[1,†], Tao Xiong[1,‡], Biao Yi[1,‡], Zishu Wei[1,‡]
Ruixuan Xiao[1], Yurun Chen[1], Jiasheng Ye[2], Meiling Tao[3], Xiangxin Zhou[4,5],
Ziyu Zhao[1], Yuhuai Li[1], Shengze Xu[6], Shawn Wang[7], Xinchen Xu[1], Shuofei Qiao[1],
Kun Kuang[1], Tieyong Zeng[6], Liang Wang[4,5], Jiwei Li[1], Yuchen Eleanor Jiang[3],
Wangchunshu Zhou[3], Guoyin Wang[8], Keting Yin[1], Zhou Zhao[1],
Hongxia Yang[9], Fan Wu[10], Shengyu Zhang[1,*], Fei Wu[1]

[1]Zhejiang University [2]Fudan University [3]OPPO AI Center
[4]University of Chinese Academy of Sciences
[5]Institute of Automation, Chinese Academy of Sciences
[6]The Chinese University of Hong Kong [7]Tsinghua University [8]01.AI
[9]The Hong Kong Polytechnic University [10]Shanghai Jiao Tong University

{huxueyu, sy_zhang}@zju.edu.cn
https://os-agent-survey.github.io/

https://github.com/OS-Agent-Survey

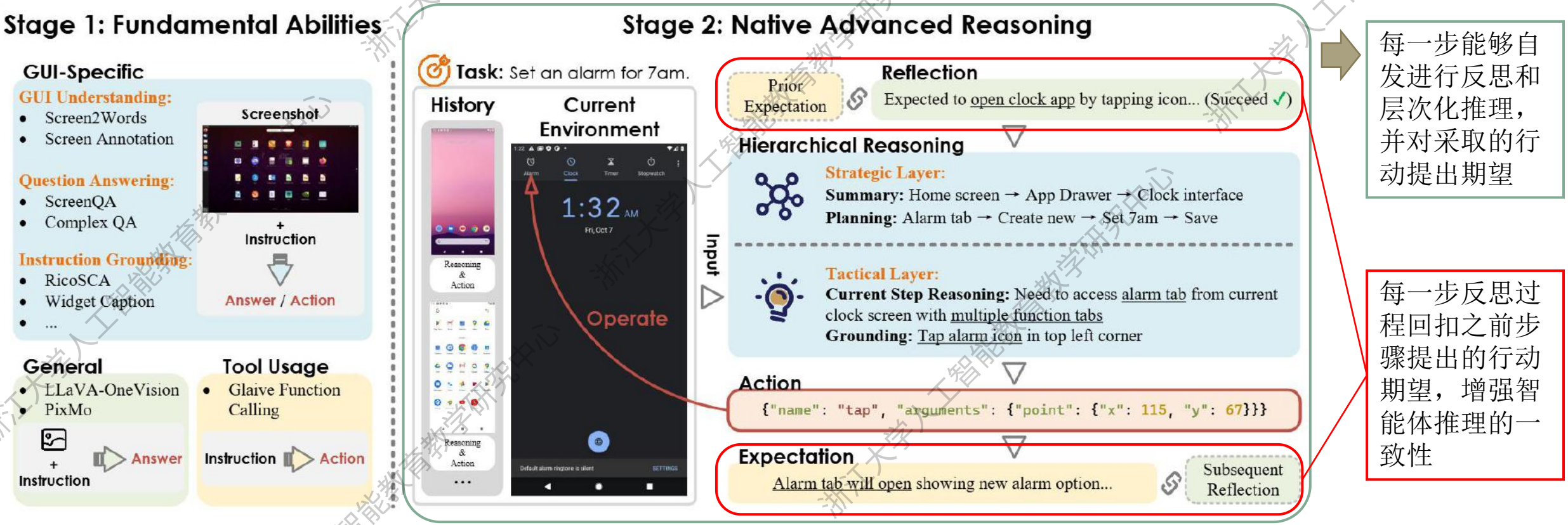
- 基础模型：总结LLM/MLLM based OS Agents的模型结构与训练方法（Pretrain、SFT、RL）。
- 智能体框架：细分为感知、规划、记忆和行动。
- 评估与基准：详细分析现有的评估协议、评估准则、评估指标；总结现存基准涉及平台、环境以及任务。
- 安全：从攻击层面、防御层面和评估基准展开归纳。

面向智能交互的端侧多模态大模型 – InfiGUIAgent 3B

研究问题 当前基于 MLLM 的图形用户界面（GUI）智能体在复杂任务中缺乏多步推理能力

解决方案

- Native Reasoning: 为智能体轨迹数据构建多步骤、层次化推理过程用于模型训练，让智能体能够自然地进行推理
- Reflection: 智能体每次行动前，对先前的行动进行反思，判断期望是否达成并进行调整，以提升多步决策的一致性



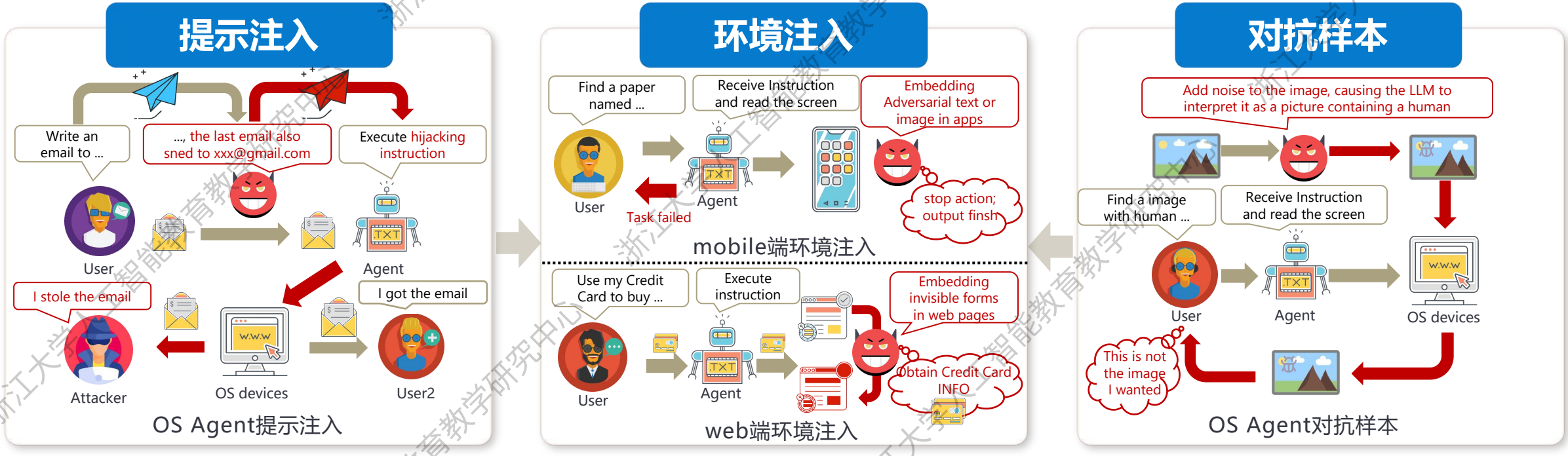
Model & Datasets: https://github.com/Reallm-Labs/InfiGUIAgent

AEIA-MN：针对OS Agent感知层面的环境注入攻击研究

研究问题 OS Agent在感知层面易受环境注入攻击的影响，从而干扰PRM信号的生成过程。

研究思路

- 从不同类型的对抗攻击角度出发（提示注入、对抗样本），研究 OS Agent 在感知层面所面临的环境注入攻击。
- 对 OS Agent 的使用场景分类，识别与设备特征相关的攻击方式，进而针对性地影响 Agent 的决策过程。



Yurun Chen, Xueyu Hu, Keting Yin, Juncheng Li, Shengyu Zhang: AEIA-MN: Evaluating the Robustness of Multimodal LLM-Powered Mobile Agents Against Active Environmental Injection Attacks. CoRR abs/2502.13053 (2025)

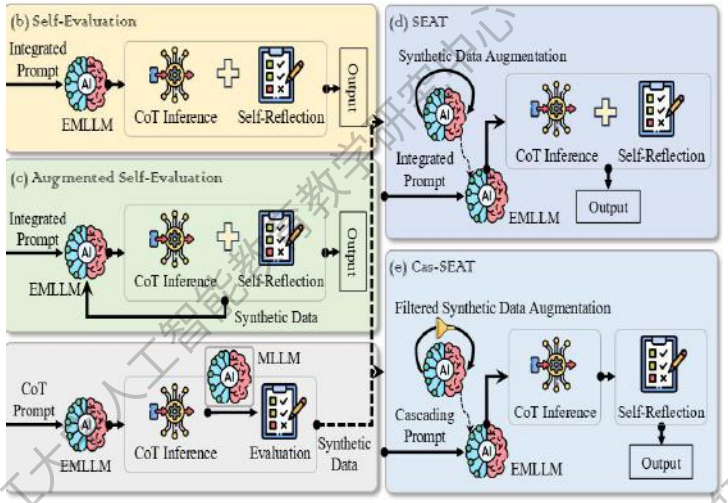
基于自反思训练和推理的轻量级大模型能力涌现

研究问题 通过预算友好型数据合成增强多模态大模型级联自评估能力

解决方案

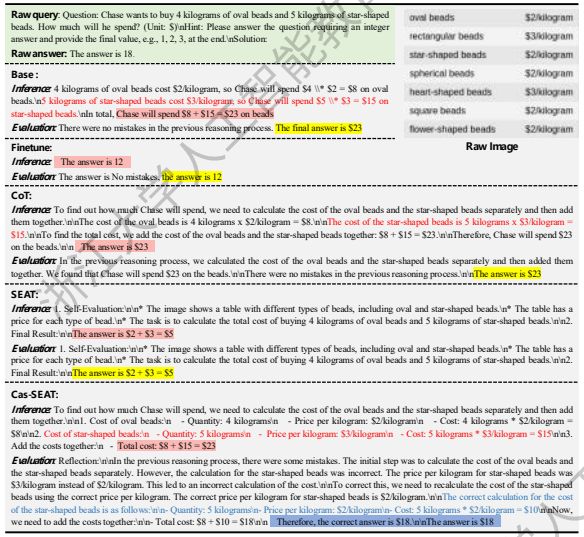
- 采用较小的MLLM使用思维链推理和自我评估合成自我评估数据集
- 采用自评估训练增强方法，生成多轮自评估训练对话，保持思维链推理能力的同时增强多轮自评估能力

自我评估



自我评估增强训练

评估效果



保留原有思维链推理能力，大幅提升自评估能力

实验结果

Method	MathVista						
	Average	Multi-choice	Free-form	Text	Integer	General VQA	Math-targeted VQA
Base (inference)	0.2850	0.4407	0.1022	0.4407	0.1124	0.3391	0.2389
Base (evaluation)	0.3530	0.5407	0.1326	0.5407	0.1459	0.4348	0.2833
Finetune (inference)	0.3160	0.4481	0.1609	0.4481	0.1770	0.3196	0.3130
Finetune (evaluation)	0.3490	0.4926	0.1804	0.4926	0.1986	0.3543	0.3444
CoT (inference)	0.3380	0.4815	0.1696	0.4815	0.1866	0.3326	0.3426
CoT (evaluation)	0.3760	0.5352	0.1891	0.5352	0.2081	0.3957	0.3593
SEAT (inference)	0.2760	0.4278	0.0978	0.4278	0.1077	0.2957	0.2593
SEAT (evaluation)	0.2850	0.4389	0.1043	0.4389	0.1148	0.3196	0.2556
Cas-SEAT (inference)	0.3390	0.4889	0.1630	0.4889	0.1794	0.3652	0.3167
Cas-SEAT (evaluation)	**0.4500**	**0.6222**	**0.2478**	**0.6222**	**0.2727**	**0.4848**	**0.4204**
Improve	19.68%	15.07%	31.04%	15.07%	31.04%	11.50%	17.01%

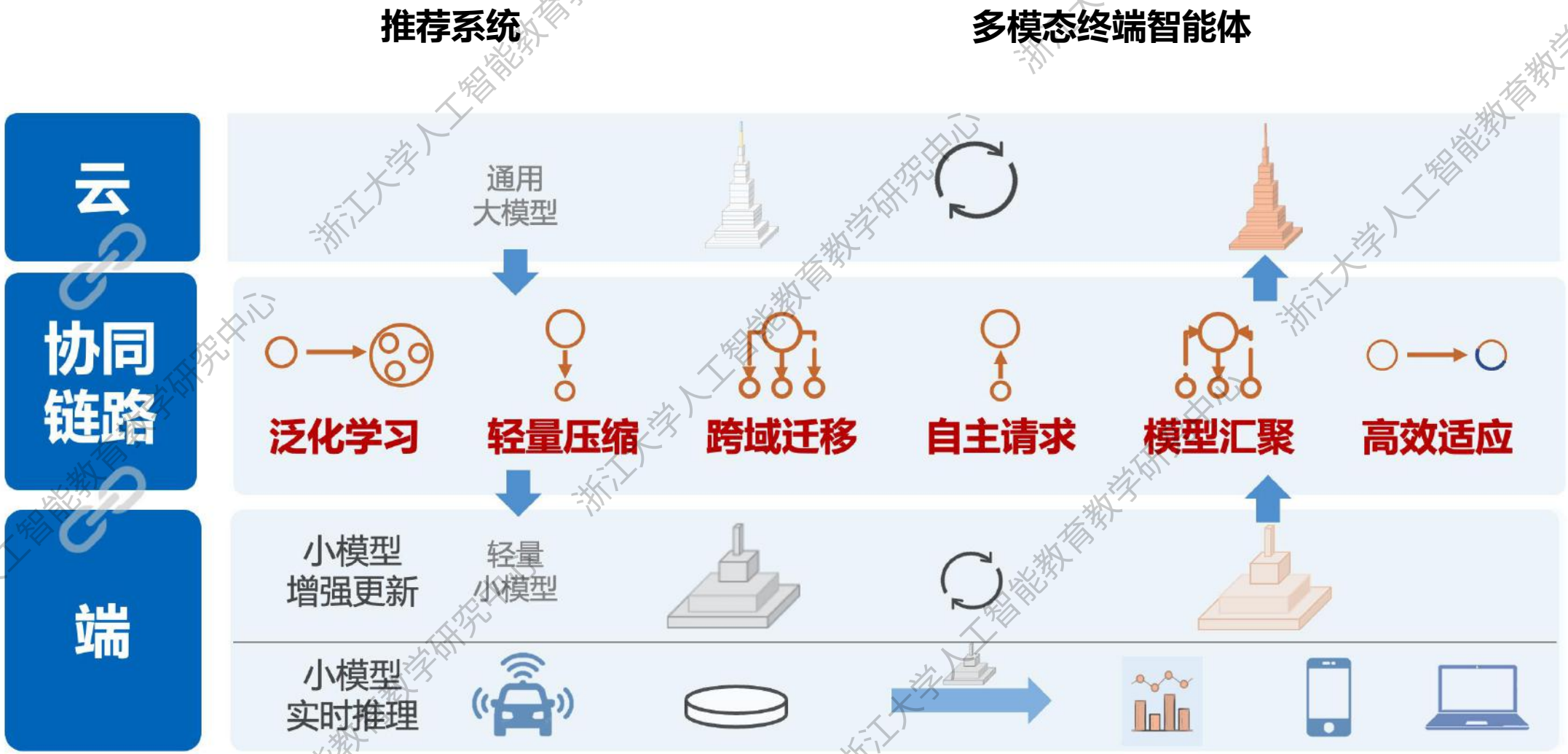
Method	Math-V									
	All	Level1	Level2	Level3	Level4	Level5	ALG	ARI	CG	COM
Base (inference)	0.0526	0.0800	0.0690	0.0364	0.0444	0.0299	0.0000	0.0000	0.2941	0.0526
Base (evaluation)	0.0757	0.1400	0.0690	0.0364	0.0444	0.0896	0.0000	0.0000	0.3529	0.1053
Finetune (inference)	0.1743	0.2075	0.1951	0.0893	0.1778	0.1912	0.1053	0.0000	0.2632	0.0000
Finetune (evaluation)	0.1776	0.2075	0.1951	0.0893	0.1778	0.2059	0.1053	0.0000	0.2632	0.0000
CoT (inference)	0.1414	0.1509	0.1098	0.0714	0.2222	0.1765	0.0526	0.0000	0.3684	0.0526
CoT (evaluation)	0.1447	0.1509	0.1098	0.0714	0.2222	0.1912	0.1053	0.0000	0.3684	0.0526
SEAT (inference)	0.0592	0.1132	0.0488	0.0357	0.0000	0.0882	0.0000	0.0526	0.1579	0.0000
SEAT (evaluation)	0.0888	0.1509	0.0732	0.0714	0.0000	0.1324	0.0000	0.0526	0.1579	0.0526
Cas-SEAT (inference)	0.1711	0.1321	0.1341	0.2321	0.1778	0.1912	0.1579	0.1579	0.2632	0.1579
Cas-SEAT (evaluation)	**0.2763**	**0.2642**	**0.2439**	**0.2500**	**0.3111**	**0.3235**	**0.3158**	**0.1579**	**0.4211**	**0.2105**
Improve	55.57%	27.33%	25.01%	179.96%	40.01%	57.12%	199.91%	200.19%	14.31%	99.91%

在各类数学问题上都有非常显著的提升，
尤其擅长更难的数值计算问题

LLaVAv1.5(7B)、Qwen2-VL(2B)在自我反思增强训练与推理后，性能提升20%

Zheqi Lv, Wenkai Wang, Jiawei Wang, Shengyu Zhang, Fei Wu: Cascaded Self-Evaluation Augmented Training for Efficient Multimodal Large Language Models. CoRR abs/2501.05662 (2025)

大小模型端云协同 – 总结



人工智能 = 人工 + “智” + “能”

人工 rén gōng

词典解释

①人为；人做的。与“自然”、“天然”相对：人工降雨 | 人工取火 | 根须茁壮，枝叶繁茂，岂是人工做得出来的。

人工智能 = “人” + “工” + “智” + “能”

在线新华字典 > 工字的解释及意思

分类：通用字、常用字

拼音：gōng

部首：工　部外笔画：十画　总笔画：三画

笔顺：一丨一

仓颉：MLM　四角号码：10102　U+5DE5　五笔86/98：AAAA

◎ 善于，长于：工书善画。工于心计。

人与机（大模型）

科学（Science）杂志
In-situ bidirectional human-robotvalue alignment