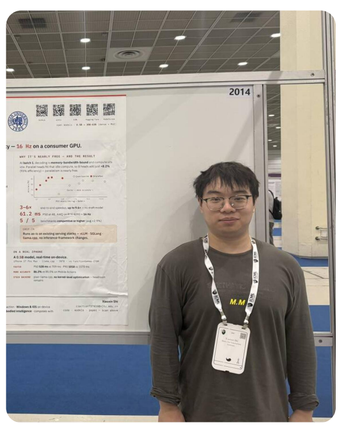
上海创智学院
Shanghai Innovation Institute

“研创学”生涯发展共创交流会



石枭昕 Xiaoxin Shi

上海交通大学 / 上海创智学院 · 25 级直博（LLM 后训练 & infra）· 导师 黄增峰教授

研究模型的架构、后训练与推理系统：设备与场景给定条件下，三者协同设计。

- **ICML 2026 一作** RealtimeTool
- **米哈游 × 宇树科技** G1 机器人交互 · 核心研发
- **创业** 上海简爱智生科技 · 创始人

教育背景
博士 · 上海交大 / 创智学院 · 2025–
学士 · 上海交大 计算化学 · 2021–25

研究方向
模型架构 · 后训练 · 推理系统
实时工具调用 · 具身智能 · 端侧多模态

技术栈
SFT / RL · LoRA · 蒸馏 · 数据合成
vLLM · SGLang · llama.cpp · MLX

开源
GitHub @HaxxorCialtion
HF / ModelScope @Cialtion

科研 论文与项目：具身实时智能体

RealtimeTool ICML 2026 第一作者

Parallel Decoding for Real-Time LLM Function Calling
arXiv 2603.00030 · 开源名 SimpleTool

压掉结构化输出的冗余 token（4–6× 压缩），函数名与各参数作为多个头**并行解码**，共享同一份 Prefix / KV Cache。

逐字串行 `{"name":"move_to","x":300,"y":150,"speed":"slow"}`

多头并行 `<fn> move_to` `<a1> 300` `<a2> 150` `<a3> slow`

时延由「所有 token 之和」变为「最长一路」；decode 受带宽限制，多路近乎免费

3–6×	16 Hz	0.5B–30B
端到端加速 · 最高 9.6×	4B 量化 · P50 61.2 ms	Qwen 系列开源模型族

单次决策最快 · RTX 4090 塔防 24.6 ms · 机械臂 39.9 ms · 数字人 29.1 ms

米哈游 × 宇树科技 真实机器人交互系统 · 核心研发 · 2026.06–10

- G1 Edu 上 0→1 复刻《星布谷地》AI 多人语音对话场景
- 语音 · 动作 · 线下交互全链路落地，长时间实时人机交互
- 参与 AI NPC 模型后训练：SFT · OPD · RL
- 米哈游内部、小红书 Redland、Tim Cook 到访时展示，获米哈游联创验收

SimpleTool-VLA

在研 LLM / VLM as Policy · 具身智能

具身操作 = 语言定义的 Tool Calling：类型约束 + 连续参数；LLM/VLM 高层策略 × 底层运控闭环；MuJoCo 仿真、数据蒸馏与 Sim-to-Real。最终实现 RealTime 的 LLM as policy for manipulation。

开源 SimpleTool-VLM（4B）

SimpleText2Motion 等

开源 实时文生动作 · 纯 C++ 跨平台

端侧 personal agent 场景中，冻结 SimpleTool-4B 隐状态，条件 DiT + Flow Matching 生成 SMPL-H 动作；约 1M 文本–动作对训练；llama.cpp + ONNX Runtime 推理。

3 s 动作 · M4 Pro 71.3 ms

创业 上海简爱智生科技 · 端侧模型架构与后训练

技术方向 Device-Native One-Model System

不是把云端模型压小搬进手机，而是从设备反推：**设备画像与场景约束**确定后，架构、后训练、推理配置与调度一起定制。

设备 × 场景
算力 F
带宽 B
内存 · 功耗
延迟预算
行为清单

› 模型架构 共享前缀 · 多行为头 零步选择器 · 单一主 KV
› 后训练 时间线 SFT · 分支掩码 设备时钟 RL · 弹性宽度
› 推理 Infra 分叉 KV · Tick 调度 宽度内合批 · 异步提交

› **一个模型** 听说想做 同时进行 永不阻塞

我们的技术判断 理解只算一次，行动并行展开；最优解由设备决定

1 理解一次，行动并行
Prefill 只理解一次，所有行为共享一份 KV；decode 受带宽限制，闲置算力用来并发多个行动头。

2 免费宽度由设备决定
并行宽度上限 $H^* \approx F \cdot \beta / 2B$，随算力带宽比变化；实测 4090 上 8 路并行单步只慢约 8%。

3 一个模型替代多智能体
单模型，端侧设备；Prefill 即理解，Decode 即行动；一次理解，多个并行、异步行动；系统层面让 one model 实现异步调度，最大化算力使用。

4 具身端侧智能体
依据单一端侧设备算力 & 带宽等特点，固定应用场景下，定制化设计模型架构和 infra 调度。

最新产品 SimpleFun 端侧 AI 角色 × 代码化小世界

上面这套方法的第一个落地尝试：**一台手机 · 一个角色 · 一个模型**。AI 角色完整跑在用户手机上，离线可用。

共享 Prefill › Speech 说 · Expression 演 · Inner 想 · Command 做

Qwen3.5-2B int4	冻结四头协议	280+ 小世界
iPhone 端侧 进程约 1.4 GB	24 表情 工具调用先校验	code4mini-app code4mv

欢迎交流 学术合作 · 产业合作 · 技术交流 · VC 合作

邮箱 cialtion737410@sjtu.edu.cn
个人主页 haxxorcialtion.github.io
公司官网 simplelove.fun
GitHub github.com/HaxxorCialtion
模型 huggingface.co/Cialtion

微信 扫码加我 →