

# 石泉昕

Xiaoxin Shi

上海交通大学 / 上海创智学院 cialtion737410@sjtu.edu.cn GitHub 个人主页

研究方向：大语言模型架构与后训练、实时 Function Calling、推理系统优化、具身智能 Agent、端侧多模态模型

## 教育背景

|                               |                   |
|-------------------------------|-------------------|
| 上海交通大学 / 上海创智学院 博士研究生（LLM 方向） | 2025.09 – 至今      |
| 导师：黄增峰教授                      |                   |
| 上海交通大学 理学学士，化学专业（计算化学方向）      | 2021.09 – 2025.06 |

## 代表性学术成果

|                                                                                                                                                                                                                                                                                                                                |           |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| RealtimeTool: Parallel Decoding for Real-Time LLM Function Calling                                                                                                                                                                                                                                                             | ICML 2026 |
| 第一作者                                                                                                                                                                                                                                                                                                                           |           |
| <ul style="list-style-type: none"><li>提出面向实时 Function Calling 的结构化输出压缩与多头并行解码方法，通过共享 Prefix/KV Cache 支持函数名与参数并行生成，减少串行自回归解码开销。</li><li>在 Qwen 系列 0.5B–14B 模型上验证方法；在所报告的测试条件下实现典型 3–6× 端到端加速，最高 9.6×，同时保持有竞争力的准确率。</li><li>4B 量化模型在 RTX 4090 上测得 61.2 ms P50 Function Calling 延迟（约 16 Hz）；全面负责核心算法、训练与数据构建、系统实现和实验评测。</li></ul> |           |

## 科研与实习经历

|                                                                                                                                                                                                                                                                        |                   |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|
| 米哈游 × 宇树科技合作项目 核心研发                                                                                                                                                                                                                                                    | 2026.06 – 2026.10 |
| 真实机器人交互系统                                                                                                                                                                                                                                                              |                   |
| <ul style="list-style-type: none"><li>在 G1 Edu 完成从 0 到 1 的《星布谷地》AI 多人语音对话场景复刻，完成语音、动作、线下的设计与算法落地。</li><li>项目在公司内部、小红书 Redland, 以及苹果前 CEO 蒂姆库克访问米哈游等活动展示, 获得公司联创验收认可。</li><li>实现机器人实时、长时间人机交互；实现公司内多部门、跨公司技术协作项目推进与落地。</li><li>参与相关 AI NPC 模型的后训练：SFT、OPD、RL。</li></ul> |                   |

|                                                                                                                                                                                                                                                       |           |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| SimpleToolVLA LLM/VLM as Policy 具身智能研究                                                                                                                                                                                                                | 2026 – 至今 |
| 在研项目；机器人闭环实时决策与结构化动作接口                                                                                                                                                                                                                                |           |
| <ul style="list-style-type: none"><li>探索将具身操作建模为语言定义的 Tool Calling，以带类型约束和连续参数的工具调用统一表达机器人动作与技能。</li><li>研究 LLM/VLM 高层策略与底层运动控制器的分层集成，在感知–决策–执行闭环中考察动作参数、实时延迟与任务泛化。</li><li>基于 MuJoCo 等仿真环境搭建桌面操作实验及数据合成/蒸馏流程，关注多样性设计、闭环控制和 Sim-to-Real 问题。</li></ul> |           |

|                                                                                                                                                                                                                                                                                                                                                                 |              |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------|
| SimpleTool / SimpleLove 端侧模型训练与推理 Infra                                                                                                                                                                                                                                                                                                                         | 2025.10 – 至今 |
| 模型后训练、多头推理、跨平台部署                                                                                                                                                                                                                                                                                                                                                |              |
| <ul style="list-style-type: none"><li>基于 Qwen 系列模型搭建 SFT、蒸馏与结构化输出训练及评测流程，构建多轮对话与动作/工具调用数据。</li><li>探索 Speech、Expression、Action、Command 多并行输出头及共享计算机制，降低实时数字人/Agent 系统的串行输出延迟。</li><li>使用并定制 vLLM、SGLang、llama.cpp、MLX 等推理框架，开展 KV Cache、Batching、多路解码和量化优化。</li><li>在 NVIDIA GPU、Apple Silicon 与 iPhone 等平台部署和分析模型，分析并改进 TTFT、Decode TPS、峰值内存及用户感知延迟。</li></ul> |              |

## 技术能力

|       |                                                                                 |
|-------|---------------------------------------------------------------------------------|
| 模型与算法 | Transformer、LLM Post-training、SFT/RL、LoRA、模型蒸馏、结构化/并行解码、VLM Agent、LLM as Policy |
| 训练与推理 | PyTorch、Transformers、vLLM、SGLang、llama.cpp、MLX、量化部署、KV Cache 优化、推理性能分析          |
| 工程实践  | 大规模合成数据构建、自动化评测、推理链路 profiling、机器人真机调试、跨平台应用集成                                  |