石枭昕

cialtion737410@sjtu.edu.cn 手机 17316716363 HaxxorCialtion

教育背景

上海交通大学 化学（计算化学）本科 2021 – 2025

上海创智学院 / 上海交通大学 LLM 方向博士在读 2025.09 – 至今

导师：黄增峰教授 **研究方向**：LLM 后训练 / 推理加速 / LLM 应用

核心亮点

- **游戏 / 数字人 NPC 全栈闭环**：独立完成“语音输入 → ASR → LLM 意图理解 → NPC 行为决策 → 角色动作与渲染”完整端到端 Demo（Echo Chronicles 纯 iPhone 端侧塔防游戏、SimpleLove VRM 数字人），实现 iOS / Windows / macOS Native 全栈跑通。
- **ICML 2026 (Accept, 一作)**：提出用于实时函数调用的多头并行解码架构，实现端到端 **3–6× 加速**（峰值 9.6×）；Qwen3-4B 在 4090 上达到 **61.2 ms / 16 Hz** 实时控制；iPhone 17 Pro Max 真机 **528 ms P50**。性能与速度均显著优于 Google FunctionGemma。
- **大规模训练实战**：具备**多节点分布式训练经验（最大规模 48×H200）**；掌握 Packing / 长度分桶 / 梯度累积 / FlashAttention-2 / FlexAttention / DDP 等高效训练技巧。
- **端侧推理 Infra**：为混合 SimpleTool 模型定制 llama.cpp 跨平台 KV Cache 共享与 Multi-seq Batching 改造；实现 nano-vllm 多头并行解码场景特化魔改；基于 ONNX 实现 DiT 框架高效推理；将 AI 协同工具（Claude / Cursor）作为开发加速器，单人闭环多模块整合。

项目经历

SimpleTool: Parallel Decoding for Real-Time LLM Function Calling（ICML 2026） 2025.10 – 至今

- **背景**：LLM 函数调用在端侧场景下延迟过高，难以满足游戏 NPC / 实时数字人 / 机械臂等 10 Hz+ 控制场景。
- **方法**：设计 17 个 Special Token 同时承担“结构 Token 压缩”与“模式选择器”双重角色，将 Function Call 输出空间压缩 **4–6**×；提出多头并行解码架构，利用 Decoding 阶段闲置算力，让 Function Name 与 Arguments 在不同头上同时解码。
- **个人工作**：独立完成 Idea → 数据合成 Pipeline → 训练框架 → 推理引擎魔改 → 多平台部署 → 论文撰写/投稿/Rebuttal 全流程闭环。
- **结果**：Qwen3-4B 在 4090 上 **61.2 ms P50**（16 Hz），并行 8 头平均效率 93%；Mobile Actions Unseen Benchmark 上 RT-Qwen-0.5B 达 **86.2%**（对比 FunctionGemma-270M 的 85.0%），通用能力完整保留（MMLU −0.29%，IFEval +2.78%）。论文已被 **ICML 2026 接收**，HuggingFace 与 ModelScope 已上线 7 个尺寸版本（0.5B–30B MoE）。

SimpleLove / NPC.exe ——端侧实时 AI-Native 数字人 & 游戏 NPC 引擎 2026.02 – 至今

- **背景**：现有数字人 / 游戏 NPC 系统普遍依赖云端大模型，延迟高、隐私差，且行为策略与角色身份割裂，体验类似传统云端语音助手。
- **方法**：基于 SimpleTool 构建端到端本地数字人引擎；采用 NPC Policy SFT / RL (Action as Cosplay) 使 LLM 直接生成角色 Actions；Simple-T2M 复用 SimpleTool LLM 中间层 Hidden States 作为 Text Condition，砍掉传统 8B 级独立 Text Encoder。
- **个人工作**：独立设计架构，训练 NPC Policy 及 50M DiT Flow Matching 动作生成模型，完成跨平台 Native 部署与 VRM 数字人渲染集成。
- **结果**：NPC Policy Eval Acc **16.2%** → **58.7%**（+42.5pp），输出分布 RL-friendly（Top-10 Cov 0.993，可直接接 PPO/GRPO）；Simple-T2M 在 4090 (Q8) 上 **~50 ms 出 Motion**，整套数字人系统仅占 **5 GB 显存**；Linux / Windows (DirectML) / 端到端 Native 全部打通，零 Python 依赖（macOS Metal+CoreML 正在制作），可直接嵌入游戏客户端。

Echo Chronicles / 回声战纪——语音驱动塔防游戏（SimpleTool 落地 Demo 之一） 2025.12 – 2026.02

- **完整游戏闭环**：玩家语音指令 → 端侧 ASR (Sherpa-onnx Paraformer, 5.9% WER, 115 ms) → SimpleTool 0.5B 意图理解与函数调用 → 游戏内塔防单位行为响应。
- **个人工作**：独立完成 iOS 端 llama.swiftui + Metal 推理集成、PixiJS WebGL 渲染层、五元素塔防系统与 Campaign 关卡设计。
- **结果**：iPhone 17 Pro Max 真机全程本地运行，无任何云端依赖；有效验证了 “LLM 作为游戏内 NPC / 角色控制器” 范式在端侧的可行性与延迟可控性。

技术栈

编程语言：熟练使用 Python；在 AI 协同下熟练开发 C++ / CUDA / Swift 应用

算法与训练：Transformer、SFT / RL、LoRA、Packing / 长度分桶、Curriculum Learning、多节点 DDP（48 卡训练经验）、FlashAttention-2

推理与架构：PyTorch、vLLM、llama.cpp（含魔改定制）、nano-vllm、ONNX Runtime（CUDA / DirectML）、GGUF 量化（Q4/Q8）、KV Cache 优化

端侧部署：Linux、Windows（DirectML）、macOS（Metal）、iOS LLM Native 开发（基于 ggml 与 onnx）、ASR 端侧部署

数据合成：多 Agent 协同 Pipeline、LLM-as-Judge；已生产 2M+ 工业级训练样本（覆盖游戏 NPC / 数字人 / 机械臂领域）

个人总结

科研 taste 喜欢能工业落地的研发，习惯从应用侧反推 Infra 和算法设计，坚信 “能落地的 idea 才是好 idea”。技术兴趣集中在 LLM 后训练、推理加速与端侧实时智能体，长期目标是落地从虚拟数字生命到物理机器人的具身智能。具备从论文 Idea 到跨平台 Native Demo 的全栈单人开发闭环能力，期望加入米哈游，参与 AI 数字人 / 游戏 NPC / 大模型训练与推理优化等方向的工业化落地实习。