Xiaoxin Shi

cialtion737410@sjtu.edu.cn +86 173-1671-6363 HaxxorCialtion

Education

Shanghai Jiao Tong University (SJTU) B.S. in Chemistry (Computational Chemistry) 2021 – 2025

Shanghai Institute of Intelligent Science / SJTU Ph.D. Student in Large Language Models Sept. 2025 – Present

Advisor: Prof. Zengfeng Huang **Research Interests:** LLM Post-training / Inference Acceleration / LLM Applications

Core Highlights

- **Full-Stack Integration for Game & Virtual Human NPCs:** Independently built a complete end-to-end pipeline ("Voice Input → ASR → LLM Intent Understanding → NPC Behavioral Decision → Character Animation & Rendering"). This includes *Echo Chronicles* (a pure iOS on-device tower defense game) and *SimpleLove* (a VRM virtual human demo), successfully achieving native full-stack deployment across iOS, Windows, and macOS.
- **ICML 2026 (Accepted, First Author):** Proposed a multi-head parallel decoding architecture for real-time function calling, achieving an end-to-end **3–6× speedup** (peak 9.6×). Realized real-time control at **61.2 ms / 16 Hz** for Qwen3-4B on an RTX 4090, and **528 ms P50** on a real iPhone 17 Pro Max device. Performance and speed significantly outperform Google's FunctionGemma.
- **Large-Scale Training Practice:** Possess **multi-node distributed training experience (up to 48×H200 GPUs)**; highly proficient in advanced training techniques including Packing, Length Bucketing, Gradient Accumulation, FlashAttention-2, FlexAttention, and DDP.
- **On-Device Inference Infrastructure:** Customized `llama.cpp` for cross-platform KV Cache sharing and Multi-seq Batching tailored for the hybrid SimpleTool model; modified `nano-vllm` specifically for multi-head parallel decoding; implemented highly efficient inference for DiT frameworks based on ONNX. Leveraged AI coding assistants (Claude / Cursor) to accelerate development, independently achieving full-stack multi-module integration.

Project Experience

SimpleTool: Parallel Decoding for Real-Time LLM Function Calling (ICML 2026) Oct. 2025 – Present

- **Background:** LLM function calling exhibits excessive latency in on-device scenarios, failing to meet the 10 Hz+ control requirements for game NPCs, real-time virtual humans, and robotic arms.
- **Methodology:** Designed 17 special tokens to concurrently serve as "structural token compressors" and "mode selectors," compressing the function call output space by **4–6×**. Proposed a multi-head parallel decoding architecture that exploits idle computing power during the decoding phase to generate function names and arguments simultaneously across different heads.
- **Key Contributions:** Independently led the entire lifecycle closed-loop: Idea → Data Synthesis Pipeline → Training Framework → Inference Engine Modification → Multi-Platform Deployment → Paper Writing, Submission, and Rebuttal.
- **Results:** Qwen3-4B reached **61.2 ms P50** (16 Hz) on an RTX 4090, with an average efficiency of 93% across 8 parallel heads. RT-Qwen-0.5B achieved **86.2%** accuracy on the Mobile Actions Unseen Benchmark (vs. 85.0% for FunctionGemma-270M) while seamlessly preserving general capabilities (MMLU −0.29%, IFEval +2.78%). The paper is **accepted to ICML 2026**. Seven model versions (0.5B–30B MoE) have been released on HuggingFace and ModelScope.

SimpleLove / NPC.exe — On-Device Real-Time AI-Native Virtual Human & Game NPC Engine Feb. 2026 – Present

- **Background:** Existing virtual human / game NPC systems predominantly rely on cloud-based LLMs, resulting in high latency, poor privacy, and disjointed behavioral strategies that feel like traditional voice assistants.
- **Methodology:** Built an end-to-end local virtual human engine powered by SimpleTool. Utilized NPC Policy SFT / RL (Action as Cosplay) to enable the LLM to directly generate character actions. Developed Simple-T2M, which reuses the intermediate hidden states of the SimpleTool LLM as text conditions, eliminating the need for a traditional 8B-parameter independent text encoder.
- **Key Contributions:** Independently designed the architecture, trained the NPC Policy and a 50M DiT Flow Matching motion generation model, and accomplished cross-platform native deployment alongside VRM virtual human rendering integration.
- **Results:** Boosted NPC Policy evaluation accuracy from **16.2% to 58.7%** (+42.5pp), yielding an RL-friendly output distribution (Top-10 Cov 0.993, seamlessly attachable to PPO/GRPO). Simple-T2M generates **motion in ∼50 ms** on an RTX 4090 (Q8). The complete system requires only **5 GB of VRAM**. Fully integrated across Linux, Windows (DirectML), and Native End-to-End with zero Python dependency (macOS Metal+CoreML in progress), ready for direct embedding into game clients.

Echo Chronicles — Voice-Driven Tower Defense Game (SimpleTool Landing Demo) Dec. 2025 – Feb. 2026

- **End-to-End Game Loop:** Player Voice Commands → On-Device ASR (Sherpa-onnx Paraformer, 5.9% WER, 115 ms) → SimpleTool 0.5B Intent Understanding & Function Calling → In-Game Tower Defense Unit Behavioral Response.
- **Key Contributions:** Independently developed iOS on-device `llama.swiftui` + Metal inference integration, PixiJS WebGL rendering layer, a five-element tower defense system, and campaign level design.
- **Results:** Runs entirely locally on an iPhone 17 Pro Max without any cloud dependency, effectively validating the feasibility and latency controllability of the "LLM as an In-Game NPC / Character Controller" paradigm on edge devices.

Technical Skills

Programming: Proficient in Python; adept at developing C++ / CUDA / Swift applications leveraging AI coding assistants.

Algorithm & Training: Transformer, SFT / RL, LoRA, Packing / Length Bucketing, Curriculum Learning, Multi-Node DDP (48-GPU training experience), FlashAttention-2, FlexAttention.

Inference & Architecture: PyTorch, vLLM, `llama.cpp` (including custom modifications), `nano-vllm`, ONNX Runtime (CUDA / DirectML), GGUF Quantization (Q4/Q8), KV Cache Optimization.

On-Device Deployment: Linux, Windows (DirectML), macOS (Metal), iOS LLM Native Development (via `ggml` and `onnx`), On-Device ASR Deployment.

Data Synthesis: Multi-Agent Collaboration Pipelines, LLM-as-a-Judge; generated over 2M industrial-grade training samples (covering Game NPCs, Virtual Humans, and Robotic Arms).

Personal Summary

My research taste favors industry-oriented R&D. I am accustomed to reverse-engineering infrastructure and algorithm designs from application requirements, firmly believing that "a good idea is one that can be deployed." My technical interests focus on LLM post-training, inference acceleration, and on-device real-time agents, with a long-term vision of actualizing embodied intelligence—from virtual digital lives to physical robots. Equipped with full-stack, solo-development closed-loop capabilities to translate theoretical ideas into cross-platform native demos, I aspire to join miHoYo for an internship focusing on the industrial deployment of AI virtual humans, game NPCs, and the optimization of large model training and inference.