Yunhai Hu

New York, NY | yunhaihu66@gmail.com | linkedin.com/in/yunhai-hu-3b3561323 | (551) 358-5392

EDUCATION

New York University Sep. 2024 - May 2026
M.S. in Computer Science Courant, GPA:3.7 Manhattan, NY
Shandong University Sep. 2017 - Jun. 2022
B.S. in Applied Mathematics, GPA:4.0 Shangdong, CN

TECHNICAL SKILLS

Languages: Python, Java, C++, Rust **Frameworks:** Transformers, Verl, DeepSpeed, vLLM, Flink, Spark, Kubernetes
AI Expertise: LLM & VLM; Agent & Multi-Agent; RAG & Reasoning; PEFT & RL; Speculative Decoding

PROFESSIONAL EXPERIENCE

Machine Learning Engineer Internship @ NEXA AI
On-Device AI Optimization with Speculative Decoding July 2025 – Present

- **Refactored and modularized** Qwen3-VL runtime for efficient **on-device deployment** across mobile platforms.
- Designed and implemented a **state-of-the-art speculative decoding architecture and training scheme** tailored for **vision–language models**, accelerating inference while preserving multimodal reasoning quality.
- Integrated speculative decoding into the **Qwen3-VL pipeline** with **visual–text alignment optimization** and **static KV-cache**, enabling faster token verification and reduced memory overhead during multimodal inference.
- Achieved up to **7× faster** inference and **30–40% lower energy consumption** on **mobile devices**.

Research Intern @ Cerebras System
DREAM: Entropy-Adaptive Cross-Attention for Multimodal Speculative Decoding Feb. 2025 – July 2025

- Achieved **2–4×faster inference** on Pixtral with tree-based speculative decoding and cross-attention draft models, showing robustness across ScienceQA, MMT-Bench, and related benchmarks while **preserving output quality**.
- Optimized draft training on **LLaVA-Mix-665K** instructions using layer-wise distillation with **dynamic mid-layer selection**, where both **final logits** and **intermediate features** from the target model provide supervisory signals.

Full-time Software Engineer @ Bilibili Technology Co., Ltd. May 2022 - Sep. 2024
AI-driven real-time data platform and stream-batch unification

- Built **stream–batch unified SQL pipelines** for Ads/AI models, handling **click–show joins and algorithm execution**, supporting both real-time serving and offline re-computation for training–serving consistency
- Developed a **cloud-native Flink+K8S platform**, improving scalability and reliability of algorithm data services
- Optimized **Flink RocksDB state backend**, cutting peak-time resource load by **15%** and boosting system stability

RESEARCH PROJECTS

MAICRL: Multi-Agent In-Context RL for Clinical Diagnosis
Collaborative Research, MIT Media Lab May. 2025 – Present

- Developed a **multi-agent** diagnostic workflow (initial diagnosis, specialist **multi-turn** refinement, final decision) and applied **In-Context Reinforcement Learning** to help agents adapt strategies using contextual examples with rewards.
- Designed a two-level **reward mechanism** using Hit@3 with turn-level and decayed global scoring, and built RareBench rollout memory with positive/negative exemplars to enhance in-context adaptation.
- Used **ICRL** to address key challenges in multi-agent diagnosis, aligning diagnostic styles across models, enriching diagnostic outcomes through specialist collaboration, and enhancing multi-turn communication quality.

Enhance Retrieval-Augmented Generation with Monte Carlo Tree Search
Collaborative Research, YaleNLP Dec. 2024 – Mar. 2025

- Developed **MCTS-RAG**, combining Monte Carlo Tree Search with retrieval-augmented generation, yielding **23% accuracy gain** on GPQA over leading baselines by enhancing search efficiency and factual grounding.
- Designed **concurrent expansion** (parallel rollouts) and **dynamic pruning** (cutting branches by low value estimates), preventing wasted search and reducing hallucinations, leading to **3.2× speedup and 45% fewer tokens**.
- Introduced **hallucination control** by pruning low-consistency nodes and enforcing grounding via retrieval verification
- Outperformed SOTA baselines (Search-o1, RAG-Star, DeepRAG) by **8%**, matching GPT-4o with a 7B model.

PipeSpec: Breaking Stage Dependencies in Hierarchical LLM Decoding
SAILAB Research, NYU Oct. 2024 – Dec. 2024

- Proposed a **hierarchical** pipeline-based speculative decoding framework enabling asynchronous parallel execution.
- Designed a prediction verification mechanism to break serial dependencies while **ensuring prediction correctness**.
- Achieved up to **2.54× speedup** on various tasks, offering a scalable acceleration strategy for multi-device deployments.

PUBLICATIONS

- **Hu, Y.**, *et al. DREAM: Drafting with Refined Target Features and Entropy-Adaptive Cross-Attention Fusion for Multimodal Speculative Decoding*. NeurIPS, 2025. *DREAM: Drafting with Refined Target Features and Entropy-Adaptive Cross-Attention Fusion for Multimodal Speculative Decoding*. NeurIPS, 2025.
- **Hu, Y.**, Zhao, Y., *et al. MCTS-RAG: Enhancing RAG with Monte Carlo Tree Search*. EMNLP Findings, 2025.
- **Hu, Y.**, *et al. Speculative Decoding and Beyond: An In-Depth Survey of Techniques*. EMNLP Findings, 2025.
- McDanel, B., Zhang, S. Q., **Hu, Y.**, *et al. PipeSpec:Break Stage Dependencies in LLM Decode*. ACL Findings, 2025.