Last updated in February 2026

Yuxuan Lou

National University of Singapore | yuxuanlou@u.nus.edu | +65 82600153

yuxuanlou.info | Google Scholar | Github

Research Interest

- Diffusion Large Language Models
- Efficient Large Language Model Scaling with Mixture of Experts
- Multimodal Foundation Model Adaptation from Large Language Models

Education

National University of Singapore, School of Computing, HPC-AI Lab — 2023 – Present
- Ph.D. in Computer Science, Advised by Prof. Yang You

National University of Singapore, School of Statistics and Probability — 2020 – 2022
- M.Sc. in Statistics

Harvard University, Computer Science Department, DAS Lab — 2019 – 2020
- Research Intern

Fudan University, School of Mathematical Science — 2016 – 2020
- B.Sc. in Applied Mathematics

Research Experiences

Diffusion-based Speech-Text Language Model, NUS - Tencent — Sep 2025 – Present
- Developed **DiffuSpeech**, the first diffusion-based speech-text language model supporting both understanding and generation, introducing a “Silent Thought, Spoken Answer” paradigm where internal text reasoning informs spoken responses
- Unified discrete text and tokenized speech under a single masked diffusion framework with modality-specific masking schedules, enabling joint generation of reasoning traces and speech tokens through iterative denoising
- Constructed **ThinkingTalk**, the first speech QA dataset with paired text reasoning traces (26K samples, 319 hours), achieving state-of-the-art speech-to-speech QA accuracy (+9 points over best baseline) and best TTS quality among generative models (6.2% WER)

Efficient Foundation Models with Mixture of Experts, NUS - Apple — Sep 2024 – May 2025
- Developed **MoST**, a novel speech-text foundation model featuring a Modality-Aware Mixture of Experts (MAMOE) architecture which directs tokens to specialized pathways for enhanced cross-modal understanding; achieved competitive performance across multiple speech-text benchmarks using exclusively open-source data
- Developed **MoRS** (Mixture of Reasoning Students), a four-stage distillation method that compresses large language models (70B parameters) into efficient mixture-of-experts architectures (12B parameters, 3B activated) while preserving specialized reasoning capabilities, achieving up to +14.5% on reasoning benchmarks
- Created the first framework to distill dense language models into MoE architectures without relying on pre-existing small models, using domain-specific expert specialization with a shared-expert design for optimal knowledge integration

Multimodal LLM Agent with Retrieval Augmented Planning, NUS - Panasonic — Oct 2023 - May 2024
- Developed **RAP**, a Multimodal planning agent which leverages past successful experiences to enhance decision-making process
- Developed **EnvBridge**, a Multimodal embodied agent which can transfer knowledge from diverse embodied environments and enhance planning ability
- SOTA results on text-only environments(ALFWorld, Webshop), Significant improvements on multimodal robotics benchmarks(Franka Kitchen, Meta-World, RLBench)

Vision Model Scaling with Mixture of Experts, HPC-AI Lab — Mar 2021 – Jan 2022
- Developed large-scale vision models: **Sparse-MLP**, **Widenet** based on Mixture of Experts

- Proposed a fully-MLP architecture with conditional computation in two directions and extended MoE to spatial dimension of image representation

Selected Publications

DiffuSpeech: Silent Thought, Spoken Answer via Unified Speech-Text Diffusion(2026)

Yuxuan Lou[*], Ziming Wu[*], Yaochen Wang, Yong Liu, Yingxuan Ren, Fuming Lai, Shaobing Lian, Jie Tang, Yang You

arxiv.org/abs/2601.22889

MoST: Modality-Aware Mixture of Experts for Efficient Speech-Text Foundation Model(2025)

Yuxuan Lou, Kai Yang, Yang You

arxiv.org/abs/2601.10272 · Github

EnvBridge: Bridging Diverse Environments with Cross-Environment Knowledge Transfer for Embodied AI(2024)

Tomoyuki Kagaya[*], **Yuxuan Lou**[*], Thong Jing Yuan[*], Subramanian Lakshmi[*], Jayashree Karlekar, Sugiri Pranata, Natsuki Murakami, Akira Kinose, Koki Oguri, Felix Wick, Yang You

arxiv.org/abs/2410.16919

RAP: Retrieval-Augmented Planning with Contextual Memory for Multimodal LLM Agents(2024)

Tomoyuki Kagaya[*], **Yuxuan Lou**[*], Thong Jing Yuan[*], Subramanian Lakshmi[*], Jayashree Karlekar, Sugiri Pranata, Natsuki Murakami, Akira Kinose, Koki Oguri, Felix Wick, Yang You

arxiv.org/abs/2402.03610

Cross-token modeling with conditional computation(2022)

Yuxuan Lou, Fuzhao Xue, Zangwei Zheng, Yang You

arxiv.org/abs/2109.02008

Open Source Projects

Colossal-AI: Making large AI models cheaper, faster, and more accessible 41k star

- A collection of parallel components for distributed training of large deep learning models
- Managed and contributed to Colossal-AI examples

awesome mixture-of-experts 1.2k star

- A collection of awesome Mixture of Experts papers and projects

MoST: Modality-Aware Mixture of Experts for Efficient Speech-Text Foundation Model

- Official implementation of MoST, a novel speech-text foundation model featuring a Modality-Aware Mixture of Experts (MAMOE) architecture which directs tokens to specialized pathways for enhanced cross-modal understanding

RAP: Retrieval-Augmented Planning with Contextual Memory for Multimodal LLM Agents

- Official implementation of RAP, a Multimodal planning agent which leverage past successful experiences to enhance decision-making process

Skills & Technologies

GPU Training: PyTorch, DeepSpeed, Megatron-LM, Colossal-AI, HuggingFace Transformers/Accelerate, vLLM, FlashAttention (NVIDIA GPU clusters)

TPU Training: TensorFlow, JAX/Flax, Keras (Google Cloud TPU pods)

Parallel Training & Optimization: Model parallel, tensor parallel, pipeline parallel, sequence parallel, data parallel, mixture-of-experts parallel training