

Qinzhe Hu

SHANGHAI JIAO TONG UNIVERSITY

📍 Shanghai, China ✉ hugo0713@sjtu.edu.cn 🌐 hugo0713.com 📄 Hugo0713 🎓 Google Scholar

RESEARCH INTERESTS

I work on **test-time training** and **model architecture**, especially how a model stores and reads **memory** over a long context: unifying time and depth into a single recurrence, and reading fast-weight memory by generation. I have also worked on efficient reasoning with RL post-training and on sparse Mixture-of-Experts.

EDUCATION

Shanghai Jiao Tong University

Sept 2023 – Present

B.S. in **Computer Science**, ACM Honors Class, Zhiyuan College.

EXPERIENCE

HuLab, University of Wisconsin–Madison

Feb 2026 – Present

Research Intern, advised by Prof. [Junjie Hu](#)

Co-first author of **uTTT**: we identify fast-weight *ownership* as a design axis of test-time training and share one memory across all layers, unifying time and depth into a single recurrence. Co-developing **TTG** (second author), a generative read-out for fast-weight memory, and **LastKV**, where all layers share one last-layer KV history, cutting KV storage to roughly one layer’s worth while improving state tracking.

Advanced Network Laboratory, SJTU

Aug 2025 – Present

Research Intern, advised by Prof. [Shengzhong Liu](#) and Prof. [Fan Wu](#)

Co-developed **SmartThinker** (ICML 2026, second author), which lets a model think only as long as each question needs, cutting reasoning length by up to half. Now: on-policy distillation and adaptive test-time compute.

Audiocc, SJTU

Oct 2024 – Jun 2026

Research Intern, advised by Prof. [Chenda Li](#) and Prof. [Yanmin Qian](#)

First author of **TF-MoE** (INTERSPEECH 2026 Oral), a sparse Mixture-of-Experts routed over time frames and frequency bands that raises the capacity of speech-separation models at almost no extra inference cost.

PUBLICATIONS

Universal Test-Time Training

Qinzhe Hu*, Zefan Cai*, Ziqiao Ma, Hao Tan, Junjie Hu (*Equal contribution)

COLM 2026 Workshop on Efficient Reasoning (**Spotlight**); under review at ICLR 2027 [arXiv](#) [Code](#)

Test-Time Generation

Zefan Cai, **Qinzhe Hu**, Yuchen Zhu, Hao Tan, Tianyuan Zhang, Sai Bi, ..., Wen Xiao, Junjie Hu, Jiuxiang Gu

Under review at ICLR 2027 [PDF](#)

Last Layer KV Cache is All You Need

Zefan Cai, Sai Bi, Lin Zhang, ..., **Qinzhe Hu**, ..., Jiuxiang Gu, Junjie Hu, Hao Tan

Under review at ICLR 2027 [PDF](#)

TF-MoE: Time-Frequency Mixture-of-Experts for Efficient Speech Separation

Qinzhe Hu, Chenda Li, Wangyou Zhang, Shujie Liu, Yan Lu, Yanmin Qian

INTERSPEECH 2026 (**Oral**) [arXiv](#)

SmartThinker: Progressive Chain-of-Thought Length Calibration for Efficient Large Language Model Reasoning

Chenzhi Hu, **Qinzhe Hu**, Yuhang Xu, Junyi Chen, Ruijie Wang, Shengzhong Liu, Jianxin Li, Fan Wu, Guihai Chen

ICML 2026 [arXiv](#) [Code](#)

HONORS AND AWARDS

Zhiyuan Honor Scholarship

Zhiyuan College, SJTU

Awarded to **TOP 5%** students in SJTU annually.

SERVICE AND SKILLS

Reviewer: COLM 2026, ICLR 2027

Programming & Tools: Python, PyTorch, C/C++, CUDA, Go, Git, L^AT_EX, Docker, Kubernetes