

Meng Gao

Tel: (+86)15841323336 ◊ gaom24@mails.tsinghua.edu.cn ◊ [LinkedIn](#)

EDUCATION

Tsinghua University *Master of Engineering in Artificial Intelligence* Aug. 2024 – Jun. 2027

- **GPA:** 3.9/4.0 (Top 10%)
- **Advisor:** [Yujiu Yang](#) • **Research interests:** reasoning-oriented learning, reinforcement learning, efficient inference, and theoretical foundations of large models
- **Selected coursework:** Artificial Neural Network (4.0), Introduction to Statistical Learning Theory (4.0), Artificial Intelligence Programming (4.0), Selected Topics on Intelligent Information Processing (4.0)

Fudan University *Bachelor of Science in Mathematics (Talent Program)* Aug. 2020 – Jun. 2024

- **Advisor:** [Yingzhou Li](#) • **Research interests:** Fast algorithms and quantum computing
- **Selected coursework:** Functional Analysis (H) (A), Ordinary Differential Equations (A), Computational Methods / Numerical Analysis (A-), Python Programming (A), Deep Learning (A)

INTERNSHIP

Qwen-Omni Team, Alibaba *Research Intern* (Mentors: [Yunfei Chu](#), [Jin Xu](#)) Mar. 2026 – Now

- Researched long-form audio understanding and reasoning, including benchmark design, QA construction, and evaluation of large audio-language models.
- Investigated methods to improve temporal grounding and counting abilities in audio models.
- Designed a conversation-mode training paradigm with per-turn loss scaling for streaming proactive-response tasks, and built the associated benchmark and synthetic data pipeline.

Microsoft Research Asia (MSRA) *Research Intern* (Mentors: [Lei Ji](#), [Yeyun Gong](#)) Jun. 2025 – Feb. 2026

- Researched efficient multimodal retrieval with a focus on inference acceleration.
- Developed token and layer pruning methods to reduce redundant computation.
- Designed adaptive early-exit mechanisms based on input difficulty.

PUBLICATIONS

Skim and Skip: Hierarchical Adaptive Inference for Efficient Multimodal Retrieval.

Meng Gao*, Yizhen Zhang*, Yang Ding*, Ziqi Dai, Shuoshuo Zhang, Junjie Wang, Taiqiang Wu, Chufan Shi, Lei Ji, Jian Jiao, Linfeng Zhang, Yeyun Gong, Yujiu Yang **EMNLP 2026 (Main Conference)**

- Proposed Skim and Skip (SAS), a hierarchical adaptive inference framework for efficient multimodal retrieval.
- Designed token-level evidence selection and depth-adaptive early exit to reduce redundant computation.
- Achieved up to $1.64\times$ speedup and $2.97\times$ FLOPs reduction while retaining 99% retrieval performance.

AnyCap: Omni-Modal Captioning with Instruction Alignment.

Yiming Ren*, Zhiqiang Lin*, Yu Li*, Meng Gao*, Junjie Wang, Zicheng Lin, Jifeng Dai, Yujiu Yang, Wenhai Wang, Ruihang Chu **EMNLP 2026 (Main Conference)**

- Developed AnyCap, a plug-and-play framework for instruction-aligned captioning.
- Built the audio data pipeline and contributed to AnyCapEval for fine-grained content-style evaluation.
- Improved caption quality and instruction following across both open-source and API-based models, with gains validated on public benchmarks.

MANUSCRIPTS

AudioSpan: Spanning the Duration and Depth of Audio Comprehension.

Wen Huang*, Yunfei Chu*, Meng Gao, Haolin He, Jin Xu

- Built a benchmark for long-form audio comprehension over real-world audio from 10 minutes to 2+ hours.
- Developed an automated pipeline for QA construction and quality assurance.
- Evaluated 12 models, identifying long-context extraction and temporal grounding as key bottlenecks.

MLP: Modality-Aware Layer Pruning for Efficient Multimodal Retrieval

Minji Seo, Meng Gao, Zhiwei Yu, Fanhong Li, Jaesik Park, Yujiu Yang, Lei Ji

- Proposed MLP, a modality-aware layer pruning framework for efficient multimodal retrieval.
- Designed structure-guided layer selection and hardness-scaled self-distillation for pruned retrieval models.
- Reduced computation by 22% while retaining over 96% of retrieval performance.

ONGOING RESEARCH

Credit Assignment in RLVR for Reasoning

Ongoing

- Study credit assignment in RLVR for reasoning by repurposing speculative decoding from a lossless acceleration tool into a source of token-level training signals.
- Use draft–target comparisons in speculative decoding to derive auxiliary supervision that may improve credit assignment under sparse outcome rewards.

UNDERGRADUATE THESIS

Hash-Table Acceleration of 2RDM Computation in OptOrbFCI Algorithm

- Designed a sorted-hash-table algorithm with binary-search lookup to accelerate two-body reduced density matrix (2RDM) computation in full configuration interaction (OptOrbFCI), a quantum-chemistry solver.
- Analyzed the lookup scheme’s complexity and benchmarked it against the open-addressing hash baseline.
- Achieved a 45% average reduction in computation time with more stable runtime.

TEACHING ASSISTANT EXPERIENCE

Teaching Assistant, Mathematical Analysis II, Fudan University (Spring 2023)

AWARDS & HONORS

Stars of Tomorrow, Microsoft Research Asia Internship Program, 2026
Outstanding Undergraduate Graduate, Fudan University, 2024
National Endeavor Scholarship (three consecutive years), 2021–2023
Fudan University Merit Scholarship (Second & Third Class) and Academic Excellence Scholarship, 2021–2023
Merit Student & Outstanding Student Leader, Fudan University, 2020–2022
First Prize – Contemporary Undergraduate Mathematical Contest in Modeling (CUMCM), 2022
Second Prize (2021) & Third Prize (2022) – National College Student Mathematics Competition (Class A)

SKILLS & INTERESTS

Programming Languages: Python, LaTeX, MATLAB, C++, R
Languages: Chinese (native), English (fluent)
Interests: Singing, Choir, Chinese Classical Dance (Level 10), Swimming