Haotian Sun

E-mail: haotian.sun@gatech.edu | **Website:** haotiansun.tech

Research Interests: Large-scale multi-modal foundation models;
LLM adaptation and personalization; Agentic planning

Education

Georgia Institute of Technology | *GPA – 4.0/4.0* — Fall 2022 - Present
Ph.D. Student in Machine Learning — *Atlanta, GA*

École CentraleSupélec | *GPA – 3.9/4.0* — Fall 2015 – Summer 2017
Dual Degree Program in collaboration with Xi'an Jiaotong University — *Paris, France*

Xi'an Jiaotong University | *Average Grade – 91/100* — Fall 2013 – Summer 2020
Honors Youth Program — *Xi'an, China*
Highly selective nationwide program accepting under 120 students each year

Experience

Apple | *AI/ML Intern, Foundation Model Team* — May 2025 - August 2025
Manager: Nan Du — *Cupertino, CA*
Research Focus: Building unified diffusion models for multi-modal generation

Apple | *AI/ML Intern, Foundation Model Team* — May 2024 - August 2024
Manager: Nan Du — *Cupertino, CA*
Research Focus: Scaling up multi-modal generative models with efficient adaptive computation

Georgia Institute of Technology | *Graduate Research Assistant* — August 2022 - Present
Advisors: Bo Dai, Chao Zhang — *Atlanta, GA*
Research Focus: Multi-modal representation learning; Adaptation and personalization for black-box LLMs; Agentic planning with LLMs.

Featured Projects

AmorLIP: Efficient Language-Image Pretraining via Amortization — Spring 2025
- Proposed an amortized CLIP framework that reduces large-batch contrastive costs via lightweight networks;
- Delivered up to 12.24% relative improvement in zero-shot classification and retrieval across 38 tasks, outperforming standard CLIP baselines with higher efficiency.

EC-DIT: Scaling Diffusion Transformers with Adaptive Expert-Choice Routing — Fall 2024
- Scaled up 97B-parameter diffusion transformer with expert-choice routing, enabling adaptive heterogeneous compute for efficient text-to-image generation;
- Achieved state-of-the-art GenEval score of 71.68% with faster convergence, stronger text-image alignment, and superior scalability over conventional sparse models.

HYDRA: Model Factorization Framework for Black-Box LLM Personalization — Spring 2024
- Developed a model factorization framework that personalizes black-box LLMs by capturing user-specific patterns and shared knowledge without relying on access to the model's inherent parameters;
- Delivered 9.01% average improvement over SoTA prompt-based methods across diverse personalization tasks.

BBox-Adapter: Lightweight Adapting for Black-Box Large Language Models — Fall 2023
- Proposed an effective domain adaptation approach that is transparent, privacy-conscious, and cost-effective for customizing commercial black-box LLMs with only APIs;
- Achieved up to 6.77% improvement over the base model with 31.30x less training cost and 1.84x less inference cost than the official SFT service.

AdaPlanner: Adaptive Planning from Feedback with Language Models — Spring 2023
- Proposed AdaPlanner, a closed-loop planning approach allowing the LLM agent to refine its self-generated plan adaptively in response to environmental feedback;
- Designed a skill discovery mechanism that leverages successful plans as few-shot exemplars, boosting sample efficiency by up to 600x.

ToolQA: A Dataset for LLM Question Answering with External Tools — Spring 2023
- Proposed a new dataset to faithfully evaluate LLMs' ability to use external tools for question answering;
- Conducted an in-depth diagnosis of existing tool-use LLMs to highlight their strengths, weaknesses, and potential improvements.

Publications

[1] **Haotian Sun**, Yitong Li, Yuchen Zhuang, Niao He, Hanjun Dai, Bo Dai, "AMORLIP: Efficient Language-Image Pretraining via Amortization," *Thirty-ninth Conference on Neural Information Processing Systems (NeurIPS 2025)*.

[2] **Haotian Sun**, Tao Lei, Bowen Zhang, Yanghao Li, Haoshuo Huang, Ruoming Pang, Bo Dai, Nan Du. EC-DIT: Scaling Diffusion Transformers with Adaptive Expert-Choice Routing, *Thirteenth International Conference on Learning Representations (ICLR 2025)*.

[3] **Haotian Sun**, Yuchen Zhuang, Wei Wei, Chao Zhang, Bo Dai. BBox-Adapter: Lightweight Adapting for Black-Box Large Language Models, *Forty-first International Conference on Machine Learning (ICML 2024)*, **(Spotlight, top 3.5%)**.

[4] **Haotian Sun**, Yuchen Zhuang, Lingkai Kong, Bo Dai, Chao Zhang. AdaPlanner: Adaptive Planning from Feedback with Language Models, *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*.

[5] **Haotian Sun**, Antoine Moulin, Tongzheng Ren, Arthur Gretton, Bo Dai. Spectral Representation for Causal Estimation with Hidden Confounders, *The 28th International Conference on Artificial Intelligence and Statistics (AISTATS 2025)*.

[6] Yuchen Zhuang, **Haotian Sun**, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, Bo Dai. HYDRA: Model Factorization Framework for Black-Box LLM Personalization, *Thirty-eighth Conference on Neural Information Processing Systems (NeurIPS 2024)*.

[7] Lingkai Kong*, **Haotian Sun***, Yuchen Zhuang, Haorui Wang, Wenhao Mu, Chao Zhang. Two Birds with One Stone: Enhancing Calibration and Interpretability with Graph Functional Neural Process, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics (AISTATS 2024)*.

[8] Changhao Li, Yuchen Zhuang, Rushi Qiang, **Haotian Sun**, Hanjun Dai, Chao Zhang, and Bo Dai. "Matryoshka Pilot: Learning to Drive Black-Box LLMs with LLMs." *Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025)*.

[9] Yuchen Zhuang, Yue Yu, Kuan Wang, **Haotian Sun**, Chao Zhang. ToolQA: A Dataset for LLM Question Answering with External Tools, *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*.

[10] Lingkai Kong, Jiaming Cui, **Haotian Sun**, Yuchen Zhuang, B Aditya Prakash, Chao Zhang. Autoregressive Diffusion Model for Graph Generation, *Fortieth International Conference on Machine Learning (ICML 2023)*.

[11] Ran Xu, Yuchen Zhuang, Yue Yu, **Haotian Sun**, Hang Wu, Carl Yang, May D Wang. MedAdapter: Efficient Test-Time Adaptation of Large Language Models Towards Medical Reasoning. *Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*.