

Haotian Sun

haotian.sun@gatech.edu · Homepage · Google Scholar · GitHub

| Research Interests

Goal: Build foundation models through the *co-design of adaptive architectures and scalable learning systems*, advancing efficient pretraining, test-time refinement, and agentic learning.

Areas: Foundation Model Training · Agentic Learning & Infra · Adaptive & Self-Improving Systems

| Experience

ByteDance Seed

Student Researcher

San Jose, CA

Summer 2026

Long-Horizon Multimodal Data Engine: Built a scalable pipeline and trained embedding-based filters to curate nearly two billion long-horizon trajectories with coherent text, image, and action steps, supporting visual planning and coding-assisted visual generation.

Apple Foundation Model Team

AI/ML Research Intern · Worked with: Nan Du, Tao Lei, Ruoming Pang, Zhifeng Chen

Cupertino, CA

Architecture & Scaling Studies: AFM 3

Summer 2025–Spring 2026

Designed compute-controlled studies of attention capacity, feed-forward capacity, and depth–width balance to guide **architecture and compute-allocation decisions** for Apple’s third-generation foundation models. Used scaling fits and runtime analysis to map quality–efficiency trade-offs in **parallel-track MoE (PT-MoE)**, whose track-local computation reduces synchronization overhead.

Large-Scale Pretraining: EC-DIT (ICLR’25)

Summer 2024

Pretrained a **97B sparse diffusion transformer** with sequence-global expert-choice routing for adaptive compute and balanced expert utilization. Established **state-of-the-art GenEval performance at release**, surpassing Stable Diffusion 3 Large and its alignment-tuned variant while maintaining competitive inference speed.

| Selected Research & Systems

2026 STACX: Pluggable Infrastructure for End-to-End Agentic RL (Project, GitHub)

- **Training Infra Lead:** Architected a **pluggable Trainer Engine** that separates distributed backends from model/algorithm specializations for role-level batching and objectives.
- **Composable learning protocols:** Designed declarative recipes for roles, objectives, and dependencies, enabling new learning interactions without backend or synchronization rewrites.
- **Auto-research prototyping:** Connected editable recipes, sandbox feedback, and policy updates through scalable training and rollout services to shorten idea-to-evidence cycles.

2026 Forward-Free Diffusion Language Models with BPTT-Free Looped Refinement (Paper)

- **Looped refinement:** Introduced shared-weight looped refinement over confidence-aware token representations, learning from model-generated drafts without BPTT across iterations.
- **Quality–efficiency gains:** Continued pretraining from Qwen3-4B-Base to obtain FReDA-4B; surpassed 7–8B diffusion LMs by 5–15% performance gain on four reasoning and coding tasks while delivering a 1.5–1.8× average speedup through confidence-based early stopping.

2026 Matryoshka Agent: Learnable Harness for Long-Horizon MLE (Paper)

- **Learned orchestration:** Designed an orchestrator–sub-agent harness with isolated execution contexts; optimized refinement decisions using solution-refinement trees and online ranking-based RL.
- **MLE benchmark gains:** Enabled Qwen3-4B to nearly match o4-mini in orchestration; improved Qwen3-30B-Coder by up to 36.7% relative over Dojo Agent on MLE-Dojo.

- 2025 **Matryoshka Pilot: Self-Improving Guidance for Black-Box LLMs** ([Paper](#); [NeurIPS 2025](#))
- **Preference-optimized control:** Developed a lightweight controller for multi-turn guidance via on-policy iterative DPO, adapting black-box LLM behavior without updating generator weights.
 - **Planning & transfer:** Improved interactive planning success by 9.3% relative over the strongest evaluated baseline; transferred guidance across black-box models without retraining.
- 2025 **AMORLIP: Efficient Language–Image Pretraining via Amortization** ([Paper](#); [NeurIPS 2025](#))
- **Efficient training:** Proposed lightweight amortizer networks that retain cross-batch information and estimate normalization locally, reducing inter-GPU communication demands.
 - **Zero-shot transfer:** Outperformed CLIP by up to 12.24% relative across 38 classification and retrieval tasks.
- 2024 **HYDRA: Optimized Context Selection for Personalized Generation** ([Paper](#); [NeurIPS 2024](#))
- **Context engineering:** Proposed a preference-aware RAG framework for task-driven history selection and best-of-N response ranking, using shared encoders with user-specific heads.
 - **Preference alignment:** Averaged 9.01% relative gains over prompt-based baselines on five LaMP tasks.
- 2024 **BBox-Adapter: Lightweight Adaptation for Black-Box LLMs** ([Paper](#); [ICML 2024 Spotlight](#))
- **Test-time scaling:** Designed a lightweight ranking-based adapter that guides sentence-level beam search over black-box LLM proposals, directing test-time compute toward promising reasoning paths.
 - **Cost–quality gains:** Delivered up to 6.77% absolute gains in reasoning accuracy over a proprietary base model; averaged $31.3\times$ lower training and $1.84\times$ lower inference costs than its hosted SFT service.
- 2023 **AdaPlanner: Adaptive Planning from Feedback with Language Models** ([Paper](#); [NeurIPS 2023](#))
- **Closed-loop planning:** Proposed a closed-loop planning framework that allows LLM agents to adaptively refine self-generated plans from environmental feedback and execution failures.
 - **Sample efficiency:** Designed skill discovery that reuses successful plans as exemplars; achieved comparable MiniWoB++ performance to CC-Net with up to $600\times$ fewer demonstrations.

| Education

- 2022–Present **Ph.D. Candidate in Machine Learning**, Georgia Institute of Technology GPA: 4.0/4.0
 Advisors: [Bo Dai](#) and [Chao Zhang](#)
 Dissertation: *Towards Efficient and Adaptive Scaling of Foundation Models*
- 2015–2017 **Diplôme d’Ingénieur**, CentraleSupélec
 Dual-degree program with Xi’an Jiaotong University
- 2013–2020 **Combined B.S./M.S. in Electrical Engineering and Automation**, Xi’an Jiaotong University
 Honors Youth Program

| Publications

Foundation Model Scaling & Efficient Learning

EC-DIT: Scaling Diffusion Transformers with Adaptive Expert-Choice Routing

Haotian Sun, Tao Lei, Bowen Zhang, Yanghao Li, Haoshuo Huang, Ruoming Pang, Bo Dai, Nan Du
[ICLR 2025](#)

AMORLIP: Efficient Language–Image Pretraining via Amortization

Haotian Sun, Yitong Li, Yuchen Zhuang, Niao He, Hanjun Dai, Bo Dai
[NeurIPS 2025](#)

Forward-Free Diffusion Language Models with BPTT-Free Looped Refinement

Haotian Sun, Rushi Qiang, Yuqian Zheng, Bo Dai
[arXiv 2026](#)

BBox-Adapter: Lightweight Adapting for Black-Box Large Language Models

Haotian Sun*, Yuchen Zhuang*, Wei Wei, Chao Zhang, Bo Dai
ICML 2024, Spotlight

AdaPlanner: Adaptive Planning from Feedback with Language Models

Haotian Sun*, Yuchen Zhuang*, Ling kai Kong, Bo Dai, Chao Zhang
NeurIPS 2023

HYDRA: Model Factorization Framework for Black-Box LLM Personalization

Yuchen Zhuang, **Haotian Sun**, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, Bo Dai
NeurIPS 2024

Matryoshka Agent: Unfolding Sub-Agents for Long-Horizon Machine Learning Engineering

Rushi Qiang, Changhao Li, **Haotian Sun**, Yuchen Zhuang, Chao Zhang, Bo Dai
arXiv 2026

Matryoshka Pilot: Learning to Drive Black-Box LLMs with LLMs

Changhao Li, Yuchen Zhuang, Rushi Qiang, **Haotian Sun**, Hanjun Dai, Chao Zhang, Bo Dai
NeurIPS 2025

Towards Better Instruction Following Retrieval Models

Yuchen Zhuang, Aaron Trinh, Rushi Qiang, **Haotian Sun**, Chao Zhang, Hanjun Dai, Bo Dai
arXiv 2025

Exploration-Driven Optimization for Test-Time Large Language Model Reasoning

Changhao Li, Yuchen Zhuang, Chenxiao Gao, **Haotian Sun**, Rushi Qiang, Chao Zhang, Bo Dai
TMLR 2026

MedAdapter: Efficient Test-Time Adaptation of Large Language Models Towards Medical Reasoning

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, **Haotian Sun**, Hang Wu, Carl Yang, May Dongmei Wang
EMNLP 2024

ToolQA: A Dataset for LLM Question Answering with External Tools

Yuchen Zhuang, Yue Yu, Kuan Wang, **Haotian Sun**, Chao Zhang
NeurIPS 2023 Datasets & Benchmarks

Representation, Causality & Machine Learning

Spectral Representation for Causal Estimation with Hidden Confounders

Haotian Sun, Antoine Moulin, Tongzheng Ren, Arthur Gretton, Bo Dai
AISTATS 2025

Spectral Representation-based Reinforcement Learning

Chenxiao Gao, **Haotian Sun**, Na Li, Dale Schuurmans, Bo Dai
arXiv 2025

Two Birds with One Stone: Enhancing Uncertainty Quantification and Interpretability with Graph Functional Neural Process

Lingkai Kong*, **Haotian Sun***, Yuchen Zhuang, Haorui Wang, Wenhao Mu, Chao Zhang
AISTATS 2024

Autoregressive Diffusion Model for Graph Generation

Lingkai Kong, Jiaming Cui, **Haotian Sun**, Yuchen Zhuang, B. Aditya Prakash, Chao Zhang
ICML 2023

**Equal contribution. Full publication list available on [Google Scholar](#).*