Zhang Jiahao

https://jiahaozhang-public.github.io/
Email: jiahao.zhang.public@gmail.com | jiahao.zhang@mbzuai.ac.ae

Research Focus

Explainable AI as a bidirectional, interactive system: I study *(i) understanding*—mechanistic, concept-level explanations of model decisions—and *(ii) intervention*—human-guided control to reliably steer model behavior. I validate these ideas through **AI-for-Science** applications (vision & biology), focusing on faithfulness, reliability, and downstream scientific utility.

Research Interests

- **Mechanistic Interpretability for Transformers:** concept/neuron-path discovery, cross-layer information flow tracing, faithful explanations with practical efficiency.
- **Human-in-the-Loop Intervention & Control:** test-time steering, concept-based interventions, robustness/reliability under weak or indirect supervision.
- **XAI for Scientific Discovery (AI4Science):** interpretable and controllable modeling for protein generation and biomedical/spatial-omics analysis.

Education

- **Ph.D. in Machine Learning**, MBZUAI — Aug. 2025 – Present
  Abu Dhabi, UAE — Supervisor: Lijie Hu
  Theme: interactive XAI (understanding + intervention) for vision and biology.
- **B.Eng in Electronic Information**, HUST (985) — Sept. 2021 – Jun. 2025
  Wuhan, China — GPA: 86/100
- **Visiting Student**, UC Berkeley — Jan. 2024 – May 2024
  Berkeley, CA — GPA: 3.57/4

Publications

** denotes equal contribution (co-first authors).* † *denotescorrespondingauthor*(*s*).

Accepted

- *Jiahao Zhang*, Zeqing Zhang*, Di Wang, Lijie Hu†. Controlling Repetition in Protein Language Models. ICLR 2026 (accepted), Poster.*
  XAI/Intervention angle: *diagnosed repetition failure modes in protein LMs and proposed an inference-time steering method (UCCS) that reduces degeneracy while preserving downstream foldability (AlphaFold confidence).*

Under Review

- *Jiahao Zhang, Wenshuo Dong, Jie Li, Lijie Hu†. From Neurons to Concepts: Coarse-to-Fine Neuron-Path Discovery for Interpretable Vision Transformers. ICML 2026 (under review).*
  Understanding angle: *concept-guided coarse-to-fine neuron-path discovery to trace cross-layer information flow with concept-level paths and neuron grounding, improving explanation faithfulness with practical speedups.*
- *Jiahao Zhang, Peng Cui, Hongbin Lin, Xinyue Xu, Lijie Hu†. Complementary Label Concept Bottleneck Models. ICML 2026 (under review).*
  Intervention angle: *concept learning under complementary-label supervision; introduced an intervention-consistency objective to stabilize concept-to-task mapping and improve intervention performance.*
- *Peng Cui*, Jiahao Zhang*, Lijie Hu†. Bayesian Gated Non-Negative Contrastive Learning. ICML 2026 (under review).*
  Representation/XAI angle: *variational Bayesian gating suppresses background/common factors to mitigate optimization conflicts and yield more interpretable representations.*

In Preparation

- *Jiahao Zhang†, Jerry Wang†. TIC: A Unified Framework for Temporal and Causal Inference in Tumor Microenvironments. Manuscript in preparation.*

Code: github.com/cellethology/tic. ***AI4Science angle:*** *a unified pipeline combining trajectory inference and causal analysis to study EMT progression and biomarker interactions in spatial transcriptomics.*

Research Experience

- ***Laboratory of Cell Ethology, CIS, Westlake University*** *Jan. 2025 – Jul. 2025*
  Research Assistant (Advisor: Dr. Jerry Wang), Hangzhou, China
  - *Designed and built* ***TIC****, an AI4Science pipeline for spatial transcriptomics in tumor microenvironments: LLM-assisted cell annotation, graph representation learning, trajectory inference (pseudotime), and causal analysis for EMT-related interactions.*
  - *Implemented graph-based trajectory modules and validated scientific findings via downstream analyses on EMT progression and biomarker interactions; released and maintained an open-source package (v2.0.0).*
- ***Representation Learning Lab, Westlake University*** *Jun. 2024 – Dec. 2024*
  Research Assistant, Hangzhou, China
  - *Developed a diffusion-inspired generative pipeline for* ***controllable protein affinity*** *design; implemented training/finetuning to generate high- vs. low-affinity candidates under controlled conditions.*
  - *Framed controllability as a reliability/intervention problem: enforced generation constraints and evaluated with structure-aware signals for downstream utility.*
- ***AI Lab (Chaowei Xiao), University of Wisconsin–Madison*** *Oct. 2023 – Dec. 2023*
  Research Assistant, Remote
  - *Built a benchmark for software vulnerability detection/repair; contributed to metric design, experiments on real-world datasets, and systematic analysis of model failure modes and robustness.*

Selected Project

- ***Interactive QA for Survey Insights (RAG + Evaluation)*** *Jan. 2024 – May 2024*
  Data Science Research Intern, Grapedata (UC Berkeley)
  - *Built a RAG-based QA system over B2B survey data; led a two-stage workflow (retrieval + answer verification/evaluation) to improve reliability and reduce hallucination in production settings.*
  - *Impact: prototype commercialized (£10,000+ per unit); received Cloud Computing Application Award (UCB Data Science Discovery Program).*

Honors

- *Cloud Computing Application Award — UCB Data Science Discovery Program (May 2024)*
- *Undergraduate Academic Excellence Scholarship (2023–2024)*
- *Second Prize — College Student Mathematics Competition, Hubei Division (Mar. 2023)*
- *Second Prize — Chinese Mathematics Competition (Jan. 2023)*
- *Third Prize — China Undergraduate Mathematical Contest in Modeling (Sept. 2022)*

Technical Skills

- ***Programming:*** *Python, C++, Java* ***Frameworks/Tools:*** *PyTorch, Git*
- ***Research Areas:*** *Mechanistic Interpretability (ViTs/Transformers), Concept-based Models & Intervention, Protein LMs / Generative Modeling, Graph Representation Learning, Causal/Temporal Inference (applied)*

Last Updated: Feb. 2026