

Zizhao Hu

CS PhD Student, University of Southern California

zizhaoh@usc.edu | 213-675-4878 | zizhaohu.com

scholar.google.com/citations?user=A8J42tQAAAAJ | linkedin.com/in/zizhaohu

Research Interests

Agentic memory (continual learning, in-context adaptation, unlearning, and world models); **Recursive self-improvement** (synthetic data, self-distillation, post-training, test-time training); and **AI alignment** (synthetic data risk, multi-agent safety, post-training guardrails).

Education

University of Southern California

Los Angeles, CA

Ph.D. in Computer Science (Artificial Intelligence)

2022 – Present (expected 2027)

Advised by Dr. Jesse Thomason and Dr. Mohammad Rostami

M.S. in Computer Science

2020 – 2022

Coursework: ML, CV, NLP, RL, continual learning, synthetic data, reasoning models, model fine-tuning.

Georgia Institute of Technology

Atlanta, GA

B.S. in Physics, Highest Honor

2016 – 2020

Coursework: remote sensing, optical signal processing, quantum mechanics, robotics.

Research Experience

Graduate Research Assistant — USC GLAMOR Lab

Advised by Dr. Jesse Thomason

2022 – Present

- **Agentic memory:** continual learning, in-context adaptation, unlearning, and world models for LLM agents. Developed SHRED, retain-set-free unlearning that demotes memorized logits at high-surprisal tokens via self-distillation.
- **Recursive self-improvement:** post-training pipelines that let agents improve from their own output — synthetic data, self-distillation, and test-time training. Developed PRISM, intent-based persona routing bootstrapped from the model's own self-modeling.
- **AI alignment:** safety alignment and guardrails for synthetic fine-tuning of VLMs and text-to-image models, including multimodal synthetic data risk and model collapse.
- Efficient neural architectures and cross-modal alignment for vision and vision-language tasks (text-to-image diffusion, VLMs).

Graduate Research Assistant — USC iLab

Advised by Dr. Laurent Itti

2021 – 2022

- Developed GalilAI, an RL algorithm for AI agents to understand the physics of environments.

Undergraduate Research Assistant — Georgia Tech Nanophotonics Lab

Advised by Dr. Ali Adibi

2019 – 2020

- Autoencoders for reverse-design of phononic metasurfaces.

Undergraduate Research Assistant — Georgia Tech Agile Systems Lab

Advised by Dr. Simon Sponberg

2017 – 2018

- Image-segmentation pipeline for hawk-moth flight analysis.

Professional Experience

Researcher — IARPA Bengal Program (U.S. Government)

Jan 2026 – Present

- Developed SHRED, a retain-set-free LLM unlearning method that removes memorized private, copyrighted,

or hazardous content by demoting memorized logits at high-surprisal tokens via self-distillation — no curated retain set required. PI: Dr. Robin Jia; Co-PI: Dr. Jesse Thomason.

Domain Lead — Handshake AI Fellowship

2025

- Domain lead (AI & Machine Learning, PhD level) for LLM and vision-language post-training data curation.
- Managed 78 PhD domain experts and built the automation tooling behind the curation pipeline.
- Ran logistics and technical communication with teams at Meta, Anthropic, and OpenAI, curating reasoning datasets targeting Chatbot Arena and Humanity’s Last Exam.

Domain Expert — Scale AI

2024

- STEM SFT data curation team; STEM domain expert labeling graduate-level STEM data for OpenAI.

Publications

71 citations, h-index 4 (Google Scholar, Aug 2026). Author names in submission order.

- **Zizhao Hu**, Ameya Godbole, Johnny Tian-Zheng Wei, Mohammad Rostami, Jesse Thomason, Robin Jia. *SHRED: Retain-Set-Free Unlearning via Self-Distillation with Logit Demotion*. Under review, NeurIPS 2026 (reviewer scores: 5, 4, 4). arXiv:2605.07482
- **Zizhao Hu**, Mohammad Rostami, Jesse Thomason. *Expert Personas Improve LLM Alignment but Damage Accuracy: Bootstrapping Intent-Based Persona Routing with PRISM*. Under review, EMNLP 2026 (reviewer scores: 4, 4, 4, 4). arXiv:2603.18507
- **Zizhao Hu**, Xiaolin Zhou, Mohammad Rostami. *Rethinking Attention in Vision Tasks: Is Dynamic Parameterization Always Necessary?* Under review, IEEE TPAMI.
- **Zizhao Hu**, Mohammad Rostami, Jesse Thomason. *Multimodal Synthetic Data Finetuning and Model Collapse: Insights from VLMs and Diffusion Models*. ACM ICMI 2025. doi:10.1145/3716553.3750806
- **Zizhao Hu**, Xiaolin Zhou, Mohammad Rostami. *Static Key Attention in Vision*. Preprint, 2024. arXiv:2412.07049
- **Zizhao Hu**, Mohammad Rostami. *Lateralization MLP: A Simple Brain-inspired Architecture for Diffusion*. Preprint, 2024. arXiv:2405.16098
- **Zizhao Hu**, Shaochong Jia, Mohammad Rostami. *An Intermediate Fusion ViT Enables Efficient Text-Image Alignment in Diffusion Models*. Preprint, 2024. arXiv:2403.16530
- **Zizhao Hu**, Shaochong Jia, Mohammad Rostami. *Efficient Multimodal Diffusion Models Using Joint Data Infilling with Partially Shared U-Net*. Preprint, 2023. arXiv:2311.16488
- **Zizhao Hu**, Mohammad Rostami. *Class-Incremental Learning Using Generative Experience Replay Based on Time-aware Regularization*. Preprint, 2023. arXiv:2310.03898
- **Zizhao Hu**, Mohammad Rostami. *Cognitively Inspired Cross-Modal Data Generation Using Diffusion Models*. Preprint, 2023. arXiv:2305.18433
- **Zizhao Hu**, Mohammad Rostami. *Encoding Binary Concepts in the Latent Space of Generative Models for Enhancing Data Representation*. Preprint, 2023. arXiv:2303.12255
- Sumedh A. Sontakke, Stephen Iota, **Zizhao Hu**, Arash Mehrjou, Laurent Itti, Bernhard Schölkopf. *GalilAI: Out-of-Task Distribution Detection using Causal Active Experimentation for Safe Transfer RL*. AISTATS 2022. arXiv:2110.15489
- Jeff Gau, Ryan Gemilere, LDS-VIP (FM subteam, incl. **Zizhao Hu**), James Lynch, Nick Gravish, Simon Sponberg. *Rapid frequency modulation in a resonant system: aerial perturbation recovery in hawkmoths*. Proceedings of the Royal Society B 288(1951), 2021. doi:10.1098/rspb.2021.0352
- **Zizhao Hu**, Ravikiran Chanumolu, Xingyu Lin, Nayela Ayaz, Vincent Chi. *Evaluating NLP Systems On a Novel Cloze Task: Judging the Plausibility of Possible Fillers in Instructional Texts*. Preprint, 2021. arXiv:2112.01867

Teaching

Teaching assistant across 10 semesters and 5 graduate courses at USC: lectured, ran sections, and supervised student deep-learning projects in computer vision and NLP.

- **DSCI 552** (Machine Learning for Data Science): Summer 2025, Summer 2023.
- **CSCI 576** (Multimedia Systems Design): Spring 2025, Fall 2024, Fall 2022.
- **CSCI 567** (Machine Learning): Summer 2024, Summer 2022.

- **CSCI 566** (Deep Learning and Its Applications): Spring 2024.
- **CSCI 544** (Applied Natural Language Processing): Fall 2023, Spring 2023.

Awards, Leadership & Service

- **Joyce M. and Glenn A. Burdick Prize**, Georgia Tech (2018).
- **2 Gold Medals**, 27th & 28th Chinese Mathematics Olympiads.
- **Reviewer**: NeurIPS (2024, 2025, 2026), ICML (2024, 2025), ICLR (2024, 2025).

Skills

- **Expertise**: Agentic memory, continual learning, in-context adaptation, unlearning, world models; post-training, self-distillation, test-time training, synthetic data; AI alignment, multi-agent safety, guardrails; LLM, VLM, VLA, diffusion, multimodal AI.
- **Languages**: English, Mandarin Chinese; Python, Java, JavaScript, SQL, C/C++.
- **Tools**: PyTorch, TensorFlow, CUDA, Git, SLURM, GNU/Linux, Cursor, Claude Code, Codex, Unity.
- **Agentic coding**: context management, agentic harness design, agentic skills, agentic coding workflows.