Wenhan Yang

hangeryang18@g.ucla.edu
Google Scholar
https://hangeryang.github.io/

Research Summary

My research focuses on multimodal foundation models. It spans data curation, post-training (SFT, RL, and distillation), defenses against poisoning and backdoor attacks, and efficient training and test-time inference. More recently, I train lightweight orchestrators that decompose long-horizon tasks and delegate to sub-agents, extending the effective execution horizon at a fraction of frontier-model cost.

Education

2022 – Present **Ph.D. in Computer Science, University of California, Los Angeles**
Advisor: Baharan Mirzasoleiman; Expected: June 2027

2019 – 2022 **B.S. Mathematics of Computation, University of California, Los Angeles**
*Minor in **Statistics**; Magna Cum Laude; Dean's Honors List; GPA: 3.941/4.0*

Research Publications

* Equal contribution

1. H. H. Jin*, **Yang, Wenhan***, M. Ghaffari, C. Morato, and B. Mirzasoleiman, "From weeks to hours: Fast and principled SFT curation for LLM," in *The Fortieth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2026.

2. H. H. Jin*, **Yang, Wenhan***, M. Ghaffari, C. Morato, and B. Mirzasoleiman, "Reasoning quality emerges early: Data curation for reasoning models," in *The Forty-Third International Conference on Machine Learning (ICML)*, 2026.

3. **Yang, Wenhan**, N. Naharas, A. Payani, and B. Mirzasoleiman, "Distilling visual reasoning into text space," *Under review; arXiv:2609.34408*, 2026.

4. **Yang, Wenhan**, A. Rao, and A. Chandra, "HAWK: Rethinking multimodal drafting for speculative decoding," *Under review; arXiv:2610.00623*, 2026.

5. D. Nguyen*, **Yang, Wenhan***, R. Anand, Y. Yang, and B. Mirzasoleiman, "Memory-efficient training of LLMs with larger mini-batches," in *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

6. **Yang, Wenhan**, S. Stice, A. Payani, and B. Mirzasoleiman, "Bootstrapping LLM robustness for VLM safety via reducing the pretraining modality gap," *Under review; arXiv:2505.24208*, 2025.

7. **Yang, Wenhan**, J. Gao, and B. Mirzasoleiman, "Better safe than sorry: Pre-training CLIP against targeted data poisoning and backdoor attacks," in *The Forty-First International Conference on Machine Learning (ICML)*, 2024.

8. **Yang, Wenhan** and B. Mirzasoleiman, "Graph contrastive learning under heterophily via graph filters," in *The Fortieth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024.

9. S. Joshi, Y. Yang, Y. Xue, **Yang, Wenhan**, and B. Mirzasoleiman, "Towards mitigating spurious correlations in the wild: A benchmark & a more realistic dataset," *Under review; arXiv:2306.11957*, 2023.

10. **Yang, Wenhan**, J. Gao, and B. Mirzasoleiman, "Robust contrastive language-image pretraining against data poisoning and backdoor attacks," in *The Thirty-Seventh Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2023. Effectively protects CLIP pre-training against up to 1% poison rate.

Employment History

Research Scientist Intern, **Samsung Research America (Visual Display Intelligence Lab)** Jun – Sep 2026

- Achieved 2.60× lossless speedup (vs. 2.19× for EAGLE-3) and 23% longer average acceptance length across 10 multimodal benchmarks.
- Proposed HAWK, a target-aware drafter for speculative decoding in VLMs that combines compressed visual states with features from selected target-model layers.
- Introduced a dynamic logit-shifting loss that adjusts supervision when draft and target disagree, keeping them aligned across multi-step drafting.

Student Researcher, **Google Research (Algorithms & Optimization)** Jan – Jun 2026

- Post-trained small LMs as planners that decompose long-horizon tasks and coordinate workers in multi-agent systems, replacing much larger LMs and costly explicit search.
- Improved agent performance by 6.5% on multi-turn medical diagnosis and 3.6% on coding tasks.
- Extended planners to multi-turn diagnostic agents whose plans adapt to new observations.

Research Scientist Intern, **Cisco Research, Outshift (AI & ML)** Jun – Sep 2025

- Proposed V2T, a teacher–student framework that distills interleaved visual–textual reasoning into text-only chain-of-thought, with no intermediate images at inference.
- Built a two-stage pipeline: logit distillation with attention alignment, then GRPO rewarding task correctness and visual grounding.
- Improved average accuracy by 14.3% on held-out tasks and 2.7% on a broader suite, outperforming the teacher and competing baselines.

Research Scientist Intern, **Samsung Research America (NLU & On-Device AI Lab)** Jun – Sep 2024

- Extended the Matryoshka Query Transformer to SALMONN, representing audio with nested query-token subsets so one model supports multiple compression levels.
- Introduced a learned, input-adaptive token budget that allocates audio tokens per input, enabling variable inference cost without training separate models.

Technical Skills

Programming	Python, Bash
ML Frameworks	PyTorch, Hugging Face (Transformers, TRL, Accelerate), vLLM
Infrastructure	Multi-node training (FSDP, DeepSpeed), SLURM, Weights & Biases

Awards & Leadership

2025	**Summer Mentored Research Award,** University of California, Los Angeles
2023 – 2025	**President,** DanceSport Club at UCLA

Academic Service

Conference Reviewer	NeurIPS, ICML (2023–26); ICLR (2023, 2026–27); AISTATS, UAI (2025–26)
Journal Reviewer	Transactions on Machine Learning Research (TMLR)
Workshop Reviewer	AdvML-Frontiers @ ICML 2023, AdvML-Frontiers @ NeurIPS 2024

Teaching Experience

Teaching Assistant	**CS 260D: Large-Scale Machine Learning**, UCLA	Fall 2023, 2025, 2026
	CS M148: Introduction to Data Science, UCLA	Winter 2023, 2025
	CS 188: Introduction to Human-Computer Interaction, UCLA	Spring 2025