

Xinyang Han

hanxinyang@berkeley.edu | Website | Scholar | GitHub | LinkedIn | 📍 Berkeley, CA

*I design benchmarks for long-horizon, economically valuable work by AI agents,
and study how agents can close the capability gaps those benchmarks reveal.*

Education

University of California, Berkeley

PhD Student

Berkeley Artificial Intelligence Research (BAIR)

University of Chinese Academy of Sciences

Computer Science Major

Concurrent Student in Computer Science, UC Berkeley

Berkeley, CA

Sep. 2024 to Present

Beijing, China

2020 to 2024

Jan. to Sep. 2023

Selected Research

Agents' Last Exam (ALE)

Co-Lead

Co-leading a living benchmark that tests whether AI agents can reliably complete long-horizon, economically valuable work across real professional domains.

Built with **300+ experts**, ALE has collected **1,500+ tasks** across **55 sub-industries**; most tasks interleave GUI and CLI work with verifiable outcomes, and ALE results appear as the opening figure in OpenAI's GPT-5.6 release.

JobBench: Aligning Agent Work With Human Will

Co-author

JobBench asks which occupational duties workers actually want to delegate, orienting agent evaluation toward human augmentation.

Grounded in WORKBank preferences from **1,500+ workers**, its **65-task main evaluation** spans **35 occupations** with heterogeneous reference files and fact-anchored rubric chains; JobBench has been adopted by Meta's Muse Spark 1.1 and Moonshot's Kimi K3.

MLS-Bench: Can AI Systems Build Better AI?

Core Contributor

Core contributor to a benchmark testing whether agents can invent ML methods that transfer across controlled settings and scales.

Across **140+ tasks** in **12 domains**, frontier agents generally remained below reproduced human methods; targeted analyses found that more search, compute, or context alone did not remove the method-discovery bottleneck.

SAP3D: Sparse-View 3D Reconstruction

First Author

Sparse-view 3D reconstruction must recover geometry and camera poses from a few unposed images while exploiting information beyond single-image priors.

Developed SAP3D to jointly refine camera poses and adapt a view-conditioned diffusion model at test time; with three GSO input views, adaptation improved novel view synthesis PSNR from **16.3 to 18.2** and LPIPS from **0.18 to 0.13**; **CVPR 2024 Highlight**.

U2Seg: Unsupervised Universal Image Segmentation

Co-first Author

Unified semantic, instance, and panoptic segmentation without dense human labels and established the first baseline for unsupervised panoptic segmentation.

Improved CutLER by **2.6 box AP** on COCO and STEGO by **7.0 PixelAcc** on COCOStuff; **CVPR 2024**.

Multimodal AI for Cancer Risk Prediction

PhD Research

Developing multimodal cancer risk models with physician collaborators from medical images and structured clinical data.

Defining clinically meaningful outcomes and prediction horizons, then evaluating discrimination, calibration, and robustness across institutions.

Selected Publications

Agents’ Last Exam
arXiv preprint, 2026. Yiyou Sun[‡], **Xinyang Han**^{†‡}, Weichen Zhang[‡], Yuanbo Pang[‡], Tianyu Wang[‡], et al.

JobBench: Aligning Agent Work With Human Will
arXiv preprint, 2026. Yuetai Li*, Yichen Feng*, Zhangchen Xu, Zixian Ma, ..., **Xinyang Han**, et al.

MLS-Bench: A Holistic and Rigorous Assessment of AI Systems on Building Better AI
ICML 2026 AI4Math Workshop. Bohan Lyu[‡], Yucheng Yang[‡], Siqiao Huang[‡], Jiaru Zhang[‡], ..., **Xinyang Han**[‡], et al.

The More You See in 2D, the More You Perceive in 3D
CVPR 2024, Highlight. **Xinyang Han***, Zelin Gao*, Angjoo Kanazawa, Shubham Goel, Yossi Gandelsman.

Unsupervised Universal Image Segmentation
CVPR 2024. Dantong Niu*, Xudong Wang*, **Xinyang Han***, Long Lian, Roei Herzig, Trevor Darrell.

MuNeRF: Robust Makeup Transfer in Neural Radiance Fields
IEEE TVCG, 31(3), 2025. Yu-Jie Yuan*, **Xinyang Han***, Yue He, Fang-Lue Zhang, Lin Gao.

* *Equal contribution.* † *Project lead.* ‡ *Core contributor.*

Leadership & Academic Service

AI Investment Summit 2025 at UC Berkeley Orchestrated the summit with a 100+ member cross-campus team for approximately 1,000 attendees , featuring UC Berkeley Chancellor Rich Lyons, Sequoia partner Konstantine Buhler, and GPT-5 contributor Tianfu Fu.	<i>Organizer</i>
Workshop on Agent Behavior at COLM 2026 Organizing an official COLM 2026 workshop on behavioral evaluation, agent interactions, interventions, and the social implications of AI agents; supported by Anthropic, Google, and OpenAI.	<i>Organizer</i>
Conference Organization Communications Chair, CHIL 2026; Local Chair, CHIL 2025.	<i>Chair Roles</i>
Peer Review ICLR 2025; NeurIPS 2025; KDD 2025; CHIL 2025.	<i>Reviewer</i>

Invited Talks

Agents’ Last Exam: Evaluating AI Agents on Real Work Presented at Snorkel AI; UC Berkeley; Moonshot AI (Kimi); WebAgentLab; Valkyrie	<i>June and July 2026</i>
3D Foundation Models for Generation and Reconstruction ICT Turing Class	<i>July 2024</i>

Technical Skills

Agent Evaluation	GUI and CLI environments; long-horizon tasks; rubric design; artifact evaluation; harness design; model evaluation
Programming	Python; C++
ML Frameworks	PyTorch; JAX; TensorFlow
Computer Vision	representation learning; sparse-view 3D reconstruction; unsupervised image segmentation