

Mu Cai

CONTACT INFORMATION

2300 Harrison St
Thinking Machines Lab
San Francisco, CA 94110 USA

Phone: (510) 280-4206
E-mail: im.mucai@gmail.com
WWW: <https://mu-cai.github.io/>

RESEARCH INTERESTS

I work on multimodal architecture and agents.

EMPLOYMENT

Thinking Machines Lab, San Francisco, CA, USA

Member of Technical Staff, April 2026 -

- Multimodal pretraining, vision encoder.
- Visual coding/Image2Code.

Google DeepMind, Mountain View, CA, USA

Research Scientist, March 2025 - April 2026

- Core contributor of Google Gemini.
- Worked on image-to-code generation, agentic vision, and multimodal post training since Gemini 2.5.

EDUCATION

University of Wisconsin–Madison, Madison, WI, USA

Ph.D. in Computer Sciences, 2020-2025

M.S. in Computer Sciences, 2024

- Advisor: Prof. Yong Jae Lee
- Thesis: “Toward Versatile and Efficient Multimodal Models”
- Minor: Mathematics
- GPA: 4.0/4.0

Xi'an Jiaotong University (XJTU), Xi'an, China

Bachelor in Electrical Engineering, July, 2020

- GPA: 4.04/4.3 (93.39/100, rank 1/350)
- Outstanding Student Pacesetter (highest honor, 10 out of ~4,000 undergraduates)
- National Scholarship

PUBLICATIONS & MANUSCRIPTS

Gemini Team, Google. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv*, 2025.

Gemma Team. Gemma 4 Technical Report. *arXiv*, 2026.

Center for AI Safety, Scale AI. Humanity’s Last Exam. *Nature*, 2026.

Jaden Park, **Mu Cai**, Feng Yao, Jingbo Shang, Soochahn Lee, Yong Jae Lee. Contamination Detection for VLMs using Multi-Modal Semantic Perturbation. In *International Conference on Learning Representations (ICLR)*, 2026.

Kele Shao*, Keda Tao*, Kejia Zhang, Sicheng Feng, **Mu Cai**, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, Huan Wang. When Tokens Talk Too Much: A Survey of Multimodal Long-Context Token Compression across Images, Videos, and Audios. In *Transactions on Machine Learning Research (TMLR)*, 2026. (*equal contribution)

Hyunsik Chae, Seungwoo Yoon, Jaden Park, Chloe Yewon Chun, Yongin Cho, **Mu Cai**, Yong Jae Lee, and Ernest K. Ryu. Decomposing Complex Visual Comprehension into Atomic Visual Skills for Vision Language Models. *Data-centric Machine Learning Research* (DMLR), 2026.

Bocheng Zou, **Mu Cai**, Mark Stanley, Dingfu Lu, Yong Jae Lee. MuRF: Unlocking the Multi-Scale Potential of Vision Foundation Models. *arXiv* 2026.

Fangrui Zhu, Yunfeng Xi, Jianmo Ni, **Mu Cai**, Boqing Gong, Long Zhao, Chen Qu, Ian Miao, Yi Li, Cheng Zhong, Huaizu Jiang, Shwetak Patel. EgoReasoner: Learning Egocentric 4D Reasoning via Task-Adaptive Structured Thinking. *arXiv* 2026.

Yuzhang Shang*, **Mu Cai***, Bingxin Xu, Yong Jae Lee[†], Yan Yan[†]. LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models. In *IEEE/CVF International Conference on Computer Vision* (ICCV), 2025. (*equal contribution, [†]equal advising)

Jianwei Yang*, Reuben Tan*, Qianhui Wu*, Ruijie Zheng[†], Baolin Peng[†], Yongyuan Liang[†], Yu Gu, **Mu Cai**, Seonghyeon Ye, Joel Jang, Yuquan Deng, Lar Liden, Jianfeng Gao. Magma: A Foundation Model for Multimodal AI Agents. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 2025. (*first author, [†]second author)

Mu Cai, Jianwei Yang, Jianfeng Gao, Yong Jae Lee. Matryoshka Multimodal Models. *Proceedings of the International Conference on Learning Representations* (ICLR), 2025.

Xiang Li, Cristina Mata, Jongwoo Park, Kumara Kahatapitiya, Yoo Sung Jang, Jinghuan Shang, Kanchana Ranasinghe, Ryan Burgert, **Mu Cai**, Yong Jae Lee, and Michael S. Ryoo. LLaRA: Supercharging Robot Learning Data for Vision-Language Policy. *Proceedings of the International Conference on Learning Representations* (ICLR), 2025.

Mu Cai*, Zeyi Huang*, Yuheng Li, Haohan Wang, Yong Jae Lee. An Investigation on LLMs’ Visual Understanding Ability Using SVG for Image-Text Bridging *IEEE/CVF Winter Conference on Applications of Computer Vision* (WACV) 2025. (*equal contribution)

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, Yong Jae Lee. Making Large Multimodal Models Understand Arbitrary Visual Prompts. *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 2024.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao[†], Yong Jae Lee[†], Jianwei Yang[†]. TemporalBench: Benchmarking Fine-grained Temporal Understanding for Multimodal Video Models. *NeurIPS 2024 Workshop on Video-Language Models*, 2024. ([†]equal advising)

Jianrui Zhang*, **Mu Cai***, Tengyang Xie, Yong Jae Lee. CounterCurate: Enhancing Physical and Semantic Visio-Linguistic Compositional Reasoning via Counterfactual Examples. *ACL Findings*, 2024. (*equal contribution)

Bocheng Zou*, **Mu Cai***, Jianrui Zhang, Yong Jae Lee. VGBench: Evaluating Large Language Models on Vector Graphics Understanding and Generation. *EMNLP main* 2024. (*equal contribution)

Mu Cai, Chenxu Luo, Yong Jae Lee, Xiaodong Yang. Cross-Modal Self-Supervised Learning with Effective Contrastive Units for LiDAR Point Clouds. *IEEE/RSJ Intelligent Robots and Systems Conference* (IROS), 2024.

Shubham Bharti, Shiyun Cheng, Jihyun Rho, Jianrui Zhang, **Mu Cai**, Yong Jae Lee, Martina

Rau, and Xiaojin Zhu. CHARTOM: A Visual Theory-of-Mind Benchmark for Multimodal Large Language Models. *arXiv*, 2024.

Jianrui Zhang*, **Mu Cai***, Yong Jae Lee. Vinoground: Scrutinizing LMMs over Dense Temporal Reasoning with Short Videos. *arXiv* 2024. (*equal contribution)

Thao Nguyen, Haotian Liu, Yuheng Li, **Mu Cai**, Utkarsh Ojha, Yong Jae Lee. Yo’LLaVA: Your Personalized Language and Vision Assistant. *NeurIPS* 2024.

Yuheng Li, Haotian Liu, **Mu Cai**, Yijun Li, Eli Shechtman, Zhe Lin, Yong Jae Lee, and Krishna Kumar Singh. Removing Distributional Discrepancies in Captions Improves Image-Text Alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Yuzhang Shang, Bingxin Xu, Weitai Kang, **Mu Cai**, Yuheng Li, Zehao Wen, Zhen Dong, Kurt Keutzer, Yong Jae Lee[†], Yan Yan[†]. Interpolating Video-LLMs: Toward Longer-sequence LMMs in a Training-free Manner. *arXiv* 2024. ([†]equal advising)

Yuexiang Zhai, Shengbang Tong, Xiao Li, **Mu Cai**, Qing Qu, Yong Jae Lee, Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *Conference on Parsimony and Learning* (Proceedings Track, oral), 2023.

Zeyi Huang, Andy Zhou, Zijian Ling, **Mu Cai**, Haohan Wang, Yong Jae Lee. A Sentence Speaks a Thousand Images: Domain Generalization through Distilling CLIP with Language Guidance. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Mu Cai, and Yixuan Li. Out-of-distribution Detection via Frequency-regularized Generative Models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, **Spotlight**

Haotian Liu, **Mu Cai**, and Yong Jae Lee. Masked Discrimination for Self-Supervised Learning on Point Clouds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Xuefeng Du, Zhaoning Wang, **Mu Cai**, and Yixuan Li. VOS: Learning What You Don’t Know by Virtual Outlier Synthesis. In *International Conference on Learning Representations (ICLR)*, 2022.

Mu Cai, Hong Zhang, Huijuan Huang, Qichuan Geng, Yixuan Li, Gao Huang. Frequency Domain Image Translation: More Photo-realistic, Better Identity-preserving. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Liting Sun*, **Mu Cai***, Wei Zhan, and Masayoshi Tomizuka. A Game-Theoretic Policy-Aware Interaction Strategy with Validation on Real Traffic Data. In *IEEE/RSJ Intelligent Robots and Systems Conference (IROS)*, 2020. (*equal contribution)

INTERNSHIP &
VISITING
EXPERIENCE

Microsoft Research, Redmond, WA, USA
Research Intern

May 2024 – Aug. 2024

Mentors: Dr. Jianwei Yang, Prof. Yong Jae Lee

- Developed [Matryoshka Multimodal Models \(M3\)](#) for adaptive multi-granularity visual representation in MLLMs (**ICLR 2025**).
- Created [TemporalBench](#), a benchmark for fine-grained temporal reasoning in multimodal video models.

Cruise AI Research, Sunnyvale, CA, USA

May 2023 – Aug. 2023

Research Intern *Mentors: Prof. Yong Jae Lee, Dr. Dennis Park, Dr. Siva Karthik Mustikovela*

- Proposed [ViP-LLaVA](#) to enable MLLMs to interpret arbitrary visual prompts via visual overlay

and multi-layer CLIP encodings (**CVPR 2024**; adopted in Meta’s [Llama-3-V](#)).

QCraft Inc., San Jose, CA, USA

May 2022 – Aug. 2022

Research Intern

Mentor: Dr. Xiaodong Yang

- Developed [cross-modal 2D-3D contrastive learning](#) to align LiDAR point clouds with images for self-supervised autonomous driving perception (**IROS 2024**).

University of California, Berkeley, Berkeley, CA, USA

July 2019 – Nov. 2019

Visiting Student Researcher

Advisor: Prof. Masayoshi Tomizuka

- Formulated [game-theoretic interaction policies](#) (Nash, Stackelberg) with MCTS for autonomous vehicle decision-making and trajectory planning (**IROS 2020**).

PROFESSIONAL ACTIVITIES

• Conference Reviewer:

- Computer Vision and Pattern Recognition (CVPR), 2023-26
- International Conference on Computer Vision (ICCV), 2023, 2025
- Neural Information Processing Systems (NeurIPS), 2023-24
- International Conference on Learning Representations (ICLR), 2024-26
- International Conference on Machine Learning (ICML), 2024
- European Conference on Computer Vision (ECCV), 2022, 2024, 2026
- Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH), 2024
- IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2024
- AAAI Conference on Artificial Intelligence (AAAI), 2023-24
- International Conference on 3D Vision (3DV), 2022
- The International Joint Conference on Artificial Intelligence (IJCAI), 2023
- IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023-25

• Journal Reviewer:

- IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
- IEEE Transactions on Multimedia (TMM)
- IEEE Transactions on Neural Networks and Learning Systems (TNNLS)

• Workshop Organizer:

- The 4th Workshop on Computer Vision in the Wild, CVPR 2025
- ICCV 2025 Workshop on Building Foundation Models You Can Trust

• Teaching Assistant:

- CS639 Deep Learning for Computer Vision, UW Madison, Spring 2023
- CS400 Programming III (Java), UW Madison, Fall 2020, Spring 2021, Fall 2021

HONORS AND AWARDS

- ICCV Doctoral Consortium, 2025
- CVPR Doctoral Consortium, 2025
- WACV Doctoral Consortium, 2025
- Qualcomm Innovation Fellowship Finalist, 2024
- UW-Madison, Conference Presentation Award, 2024
- UW-Madison, Computer Sciences Department Summer Research Award, 2021

- XJTU, First Prize, Huawei Large-scale Image Classification Competition, 2020
- XJTU, Outstanding Graduate, 2020
- XJTU, Thanksgiving Chinese Scholarship for Modern and Contemporary Scientists, 2019
- XJTU, Gold Prize, "Tengfei Cup" Student Technology Innovation Competition, 2018-2019
- XJTU, Ultra-High Voltage (UHV) Scholarship (Top 1%, highest honor in School of EE), 2018
- XJTU, First Prize, XJTU Mathematical Contest in Modeling, 2018
- XJTU, First Prize, The Chinese Mathematics Competitions, 2017

SKILLS

- Programming: Python, C/C++, Matlab, Java, SQL
- Misc: PyTorch, L^AT_EX, Markdown
- OS: Linux, macOS, Windows

INVITED TALKS

- UC Santa Cruz, April 2025
- [Twelve Labs](#), April 2024, January 2025
- [UW-Madison MLOPT Idea Seminar](#), March 2024

STUDENT ADVISING

- Jianrui Zhang (UW Madison BS → UW Madison PhD)
- Bocheng Zou (UW Madison BS → UW Madison PhD)
- Jaden Park (UW Madison PhD)

REFERENCES

- **Prof. Yong Jae Lee** (PhD Advisor)
Associate Professor
Department of Computer Sciences
University of Wisconsin-Madison
Office: Computer Sciences 5395
Email: yongjaelee@cs.wisc.edu
Phone: 608-262-1804
- **Dr. Jianwei Yang** (Mentor at Microsoft Research)
Member of Technical Staff
xAI
Email: jw2.yang@gmail.com