Jing Bi

jing.bi@rochester.edu jing.vision jing-bi Google Scholar

Research Interests

Multimodal foundation models for agentic/embodied intelligence across digital and physical worlds. I work on world models and spatial reasoning and scalable post-training + test-time optimization to learn language and physical actions from pixels. I'm interested in generative models that learn by acting, using interaction and verification to drive self-reinforcing improvement.

Education

University of Rochester **Ph.D.**, Computer Science 2020 – present
University of Rochester **Master**, Electrical & Computer Engineering 2017 – 2019
Shandong University **Bachelor**, Computer Engineering (Robotics) 2013 – 2017

Selected Publications

When to Think and When to Look: Uncertainty-Guided Lookback CVPR 2026
An early investigation of visual reasoning, tracing how CoT improves grounding, drifts, and can be recovered.
***Jing Bi**, Filippos Bellos, Junjia Guo, Yayuan Li, Chao Huang, Yunlong (Yolo) Tang, Luchuan Song, Susan Liang, Zhongfei Zhang, Jason J. Corso, Chenliang Xu*

Unveiling visual perception in language models: An attention head analysis approach CVPR 2025
A mechanistic study of attention heads in MLLMs, uncovering visual-head behaviors that consistently track model's visual competence.
***Jing Bi**, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bo Wang, Chenliang Xu*

EAGLE: Egocentric AGgregated Language-video Engine ACM MM 2024
The first unified multimodal LLM framework for egocentric video, bringing procedures and temporal grounding together, scaled by EAGLE-400K.
***Jing Bi**, Yunlong Tang, Luchuan Song, Ali Vosoughi, Nguyen Nguyen, Chenliang Xu*

MISAR: A multimodal instructional system with augmented reality ICCVW 2023
A unified AR AI assistant that fuses vision, speech, and context into task-aware guidance.
***Jing Bi**, Nguyen Manh Nguyen, Ali Vosoughi, Chenliang Xu*

Procedure planning in instructional videos via contextual modeling and model-based policy learning ORAL ICCV 2021
A goal-conditioned formulation of instructional videos, marrying Bayesian inference with model-based reinforcement learning for latent planning.
***Jing Bi**, Jiebo Luo, Chenliang Xu*

Learning from interventions using hierarchical policies for safe learning ORAL AAAI 2020
Hierarchical imitation learning framework that turns real-world human interventions into a reusable feedback signal for safer continual policy improvement.
***Jing Bi**, Vikas Dhiman, Tianyou Xiao, Chenliang Xu*

Research Deployment

Perceptual Copilot

- Low-latency perception-action loop coupling WebRTC with a tool-using multimodal agent, maintaining continuous visual grounding for open-ended queries over real-time video streams and running on any camera-equipped device.
- Agent memory integrated with asynchronous, distributed GPU execution for heavy and long-running vision tools, using task queues and request-aware routing to offload remote inference while keeping state consistent across tool results.

Re:Search

- Agentic RAG platform for long-context research: tool-calling workflow combining vector search and targeted lookups, and emitting citation-grounded answers with step-level traceability for reproducible evidence extraction.
- Prompt-as-function registry: versioned, schema-enforced functions that convert papers into normalized, queryable structures as methods, datasets, metrics and baselines, enabling reliable cross-paper comparison and grounded scientific discovery.
- Operated a 600K+ paper corpus at 99%+ uptime, exposed through a low-latency API backed by hybrid keyword and vector retrieval, with continuous ingestion of the latest papers.

Embodied Agentic Runtime

- Embodied multi-agent runtime translating natural-language goals into a structured loop of planning, communication, action, and verification, using re-perception and completion checks to keep execution robust under distribution shift.
- Built digital-twin world state with a self-reinforcing improvement loop for grounding and monitoring deployment, while recording requests, plans, tool calls, state snapshots, synchronized video, outcomes, and operator interventions for training.
- Built tool registry and skill library spanning heterogeneous hardware such as robot arms, AGVs, sensors, and services, decoupling agent logic from hardware with skills defined by specifications and recovery with realignment and escalation.

Internship

Corning Inc, Research Scientist Intern Corning, NY
- Research on Understanding Visual Perception in Language Models (CVPR 2025), ACTLLM. Summers 2021 & 2024
- Built visual manipulation system for real-world robot deployment, integrating multimodal perception, planning/control for multi-step execution in real environments.

University of Michigan, Student Researcher, DARPA PTG Research Team Ann Arbor, MI
- Conducted applied research for the DARPA PTG program, designing AR AI assistant. Summers 2022 & 2023

Publications (Full List)

Bridging Facial Understanding and Animation via Language Models CVPR 2026
*Luchuan Song, Pinxin Liu, Haiyang Liu, Zhenchao Jin, Yolo Yunlong Tang, Zichong Xu, Susan Liang, **Jing Bi**, Jason J. Corso, Chenliang Xu*

When to Think and When to Look: Uncertainty-Guided Lookback CVPR 2026
__Jing Bi__, Filippos Bellos, Junjia Guo, Yayuan Li, Chao Huang, Yunlong (Yolo) Tang, Luchuan Song, Susan Liang, Zhongfei Zhang, Jason J. Corso, Chenliang Xu

Video-R4: Reinforcing Text-Rich Video Reasoning with Visual Rumination (Findings) CVPR 2026
*Yolo Y. Tang, Daiki Shimada, Hang Hua, Chao Huang, **Jing Bi**, Rogerio Feris, Chenliang Xu*

I^2G: Generating instructional illustrations via text-conditioned diffusion CVPRW 2026
__Jing Bi__, Pinxin Liu, Ali Vosoughi, Jiarui Wu, Jinxi He, Chenliang Xu

Caption Anything in Video: Fine-grained Object-centric Captioning via Spatiotemporal Multimodal Prompting Best Demo AAAI 2026
*Yunlong Tang, **Jing Bi**, Chao Huang, Susan Liang, Daiki Shimada, Hang Hua, Yunzhong Xiao, Yizhi Song, Pinxin Liu, Mingqian Feng, Junjia Guo, Zhuo Liu, Luchuan Song, Ali Vosoughi, Jinxi He, Liu He, Zeliang Zhang, Jiebo Luo, Chenliang Xu*

Asynchronous Temporal Modeling with Two-Agent Framework for Streaming Dense Video Captioning CVPR 2026
*Yolo Yunlong Tang, Chao Huang, Susan Liang, **Jing Bi**, Yicheng Wang, Daiki Shimada, Chenliang Xu*

What to Do Next? Memorizing skills from Egocentric Instructional Video CVPRW 2026
__Jing Bi__, Yunlong Tang, Chao Huang, Chenliang Xu

Unveiling visual perception in language models: An attention head analysis approach CVPR 2025
__Jing Bi__, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bo Wang, Chenliang Xu

MMPerspective: Do MLLMs Understand Perspective? A Comprehensive Benchmark for Perspective Perception, Reasoning, and Robustness NeurIPS D&B 2025
*Yunlong Tang, Pinxin Liu, Zhangyun Tan, Mingqian Feng, Rui Mao, Chao Huang, **Jing Bi**, Yunzhong Xiao, Susan Liang, Hang Hua, Ali Vosoughi, Luchuan Song, Zeliang Zhang, Chenliang Xu*

Verify: A benchmark of visual explanation and reasoning for investigating multimodal reasoning fidelity arXiv 2025
__Jing Bi__, Junjia Guo, Susan Liang, Guangyu Sun, Luchuan Song, Yunlong Tang, Jinxi He, Jiarui Wu, Ali Vosoughi, Chen Chen, Chenliang Xu

Why reasoning matters? a survey of advancements in multimodal reasoning arXiv 2025
__Jing Bi__, Susan Liang, Xiaofei Zhou, Pinxin Liu, Junjia Guo, Yunlong Tang, Luchuan Song, Chao Huang, Guangyu Sun, Jinxi He, Jiarui Wu, Shu Yang, Daoan Zhang, Chen Chen, Lianggong Bruce Wen, Zhang Liu, Jiebo Luo, Chenliang Xu

ZeroSep: Separate Anything in Audio with Zero Training NeurIPS 2025
*Chao Huang, Yuesheng Ma, Junxuan Huang, Susan Liang, Yunlong Tang, **Jing Bi**, Wenqiang Liu, Nima Mesgarani, Chenliang Xu*

Vidcomposition: Can mllms analyze compositions in compiled videos? CVPR 2025
*Yunlong Tang, Junjia Guo, Hang Hua, Susan Liang, Mingqian Feng, Xinyang Li, Rui Mao, Chao Huang, **Jing Bi**, Zeliang Zhang, Chenliang Xu*

Video understanding with large language models: A survey TCSVT 2025
*Yunlong Tang, **Jing Bi**, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, Chenliang Xu*

ACTLLM: Action Consistency Tuned Large Language Model arXiv 2025
__Jing Bi__, Lianggong Bruce Wen, Zhang Liu, Chenliang Xu

rQdia: Regularizing Q-Value Distributions With Image Augmentation arXiv 2025
*Sam Lerman, **Jing Bi***

Oscar: Object state captioning and state change representation NAACL 2024
*Nguyen Nguyen, **Jing Bi**, Ali Vosoughi, Yapeng Tian, Chenliang Xu*

Avicuna: Audio-visual llm with interleaver and context alignment for temporal referential dialogue AAAI 2024
*Yunlong Tang, Daiki Shimada, **Jing Bi**, Chenliang Xu*

EAGLE: Egocentric AGgregated Language-video Engine ACM MM 2024
__Jing Bi__, Yunlong Tang, Luchuan Song, Ali Vosoughi, Nguyen Nguyen, Chenliang Xu

MISAR: A multimodal instructional system with augmented reality ICCVW 2023
__Jing Bi__, Nguyen Manh Nguyen, Ali Vosoughi, Chenliang Xu

Audio-Visual Action Prediction with Soft-Boundary in Egocentric Videos ICCVW 2023
*Luchuan Song, **Jing Bi**, Chao Huang, Chenliang Xu*

Procedure planning in instructional videos via contextual modeling and model-based policy learning ICCV 2021
__Jing Bi__, Jiebo Luo, Chenliang Xu

Learning from interventions using hierarchical policies for safe learning AAAI 2020
__Jing Bi__, Vikas Dhiman, Tianyou Xiao, Chenliang Xu

Cubic Spline Smoothing Compensation for Irregularly Sampled Sequences arXiv 2020
__Jing Bi__, Chenliang Xu, Yunlong Tang, Siyu Huang, Zhenfeng Zhu, Zhenhui Ye, Zhengchang Su, Xiaohan Xue, Jiebo Luo

Navigation by imitation in a pedestrian-rich environment arXiv 2018
__Jing Bi__, Chenliang Xu, Jiebo Luo

Awards & Projects

- CAT-V - AAAI 2026 Best Demo
- Claude Research Project Fund for LLM Agent Research Assistant (2025)
- OpenAI Research Grant for Video Understanding with LLM (2024)
- DARPA PTG project "AR AI Assistant for Task Guidance" (2022-2024)
- ICCV 2021 - Procedure Planning in Instructional Videos Oral Presentation
- AAAI 2020 - Learning from Interventions Oral Presentation