# Xingrui WANG

## Education Background

**Whiting School of Engineering, Johns Hopkins University;** — **Baltimore, MD**
Ph.D. in Computer Science; **GPA**: 4.00 / 4.00; — *08/2023- Present*
**Advisor**: Prof. Alan L. Yuille.

**Viterbi School of Engineering, University of Southern California** — **Los Angeles, CA**
M.S. in Applied Data Science; **GPA**: 3.92 / 4.00 — *08/2021- 05/2023*

**School of Statistics, Renmin University of China** — **Beijing, China**
B.S. in Statistics; Minor in Data Science; **GPA**: 87.04 / 100 — *09/2017- 07/2021*

## Selected Publications

[1] **KeyVID: Keyframe-Aware Video Diffusion for Audio-Synchronized Visual Animation** [✗]
**Xingrui Wang**, Jiang Liu, Ze Wang, Xiaodong Yu, Jialian Wu, Ximeng Sun, Yusheng Su, Alan Yuille, Zicheng Liu, Emad Barsoum.
*Preprint 2025*
*TL; DR: Video generation model that learns synchronized visual motion from audio via keyframe awareness.*

[2] **Captain Safari: A World Engine** [✗]
Yu-Cheng Chou, **Xingrui Wang**, Yitong Li, Jiahao Wang, Hanting Liu, Cihang Xie, Alan Yuille, Junfei Xiao
*Preprint 2025*
*TL; DR: A world engine video generation model with camera control and 3D explicit memory conditions.*

[3] **XModBench: Tri-Modal Benchmark for Omni-Language Models** [✗]
**Xingrui Wang**, Jiang Liu, Chao Huang, Xiaodong Yu, Ze Wang, Ximeng Sun, Jialian Wu, Alan Yuille, Emad Barsoum, Zicheng Liu
*Preprint 2025*
*TL; DR: A large-scale tri-modal dataset (across text, vision, audio) for cross-modal consistency and reasoning stability of omni large language models.*

[4] **SpatialReasoner: Towards Explicit and Generalizable 3D Spatial Reasoning** [✗]
Wufei Ma*, Yu-Cheng Chou*, Qihao Liu*, **Xingrui Wang**, Jieneng Chen, Jianwen Xie, Alan Yuille
*Conference on Neural Information Processing Systems (**NeurIPS**) 2026*
*TL; DR: A novel framework for explicit 3D spatial reasoning on vision-language model that generalizes across diverse environments and tasks.*

[5] **Spatial457: A Diagnostic Benchmark for Comprehensive Spatial Reasoning of Large Multimodal Models** [✗]
**Xingrui Wang**, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, Alan Yuille.
*Conference on Computer Vision and Pattern Recognition (**CVPR**, Highlight) 2025*
*TL; DR: A benchmark for comprehensive 6D spatial reasoning of large vision language models.*

[6] **Compositional 4D Dynamic Scenes Understanding with Physics Priors for Video Question Answering** [✗]
**Xingrui Wang**, Wufei Ma, Angtian Wang, Shuo Chen, Adam Kortylewski, Alan Yuille.
*International Conference on Learning Representations (**ICLR**) 2025.*
*TL; DR: A video question answering benchmark and model for 4D physical properties of objects from 3D space.*

[7] **3D-Aware Visual Question Answering about Parts, Poses and Occlusions** [✗]
**Xingrui Wang**, Wufei Ma, Zhuowan Li, Adam Kortylewski, Alan Yuille.
*Advances in Neural Information Processing Systems (**NeurIPS**), 2023*
*TL; DR: A benchmark and model for 3D scene understanding in vision question answering, particularly parts, poses, and occlusions.*

[8] **Super-CLEVR: A Virtual Benchmark to Diagnose Domain Robustness in Visual Reasoning** [✗]
Zhuowan Li, **Xingrui Wang**, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, Alan Yuille.
*Conference on Computer Vision and Pattern Recognition (**CVPR**, Highlight), 2023*
*TL; DR: A diagnosis dataset analyzes the factors of domain shift in vision question answering models.*

[9] **Contributions of Shape, Texture and Color in Visual Recognition** [✗]
Yunhao Ge*, Yao Xiao*, Zhi Xu, **Xingrui Wang**, Laurent Itti.
*European Conference on Computer Vision (**ECCV**)*, 2022
*TL; DR: A human-inspired object recognition network which considers the disentangled shape, texture, and color from images.*
**(See google scholar for full paper list)**

## Working Experience

**Advanced Micro Devices, Inc. |** Research Intern *06/2024- 09/2025*

- Advisor: Dr. Jiang Liu. Remotely, US
- Research Topic: **Multimodal conditioning video generation and omni-large language model.**

Project Description: (1) Build a video generation diffusion model for dynamical motion conditioned on audio and image input. Evaluate the temporal alignment of given audio and generated video ; (2) Build a large-scale tri-modal dataset (across text, vision, audio) for cross-modal consistency and reasoning stability of omni large language models.

- .

**Samsung R&D Institute China-Beijing |** Research Intern *12/2020- 06/2021*

- Advisor: Dr. Yang Liu Beijing, China
- Research Topic: **Embodied AI; Reinforcement learning.**
- Project Description: (1) Human-Guided Reinforcement Learning**:** Proposed a method that combines language hints with an object template matching module, providing human coarse-grained pre-guided attention to improve the efficiency and performance of the reinforcement learning model. (2) ALFRED benchmark, Embodied AI @ CVPR 2021. Leveraged instance segmentation and depth estimation to ground object positions on the bird's-eye-view obstacle map, generate navigation paths to the grounded objects, and integrate these with language instructions.

## Teaching Experiences

***University of Southern California***

- Course Producer: DSCI 552 - Machine Learning for Data Science

***Johns Hopkins University***

- Course Producer: EN.601.673 - Cognitive Artificial Intelligence