

Haoyi Zhu

✉ hyizhu1108@gmail.com | 🌐 <https://haoyizhu.site/> | 📄 Haoyi Zhu | 🔍 Google Scholar

EDUCATION

University of Science and Technology of China (USTC), China Sep. 2023 – Present
Ph.D. in Computer Science (Advisor: Prof. [Xiaogang Wang](#))
Shanghai Jiao Tong University (SJTU), China Sep. 2019 – Jun. 2023
B.E. in Artificial Intelligence Honor Class (Advisor: Prof. [Cewu Lu](#))

SELECTED AWARDS

XingQi Intern, Shanghai AI Lab (**6 winners in Embodied AI center**) Apr. 2025
Outstanding Paper Award, NeurIPS 2022 Dec. 2022

INTERNSHIPS

Shanghai AI Lab Sep. 2023 - Oct. 2025
Research Intern (Advisor: Prof. [Tong He](#) & Prof. [Wanli Ouyang](#))

- Foundation World Model
- General 3D Embodied Representation Learning

MVIG, Shanghai Jiao Tong University (SJTU) Sep. 2019 - Jun. 2023
Research Intern (Advisor: Dr. [Hao-Shu Fang](#) & Prof. [Cewu Lu](#))

- Human Pose Estimation
- General Robot Manipulation

NVIDIA AI Lab Feb. 2022 - Feb. 2023
Research Intern (Advisor: Dr. [Jim Fan](#) & Prof. [Anima Anandkumar](#))

- General Open-Ended Embodied Agent

Xu Lab, Carnegie Mellon University (CMU) Apr. 2021 – Feb. 2022
Research Intern (Advisor: Prof. [Min Xu](#))

- AI for Biology

PUBLICATIONS

Total Citations: 1637

Journal Papers

PonderV2: Pave the Way for 3D Foundation Model with A Universal Pre-training Paradigm [[paper](#)] [[code](#)]
Haoyi Zhu*, Honghui Yang*, Xiaoyang Wu*, Di Huang*, Sha Zhang, Xianglong He, Tong He, Hengshuang Zhao, Chunhua Shen, Yu Qiao, Wanli Ouyang
IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2025

AlphaTracker: A Multi-Animal Tracking and Behavioral Analysis Tool [[paper](#)] [[code](#)]
Zexin Chen, Ruihan Zhang, Hao-Shu Fang, Yu E Zhang, Aneesh Bal, Haowen Zhou, Rachel R Rock, Nancy Padilla-Coreano, Laurel R Keyes, **Haoyi Zhu**, Yong-Lu Li, Takaki Komiyama, Kay M Tye, Cewu Lu
Frontiers in Behavioral Neuroscience, 2023

AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time [[paper](#)] [[code](#)]
Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu*, **Haoyi Zhu***, Yuliang Xiu, Yong-Lu Li, Cewu Lu
[>8.3K GitHub Stars] *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022

Conference Papers (Selected)

π^3 : Scalable Permutation-Equivariant Visual Geometry Learning [[paper](#)] [[website](#)] [[code](#)]
Yifan Wang, **Jianjun Zhou**, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, Tong He
arXiv preprint, 2025

DeepVerse: 4D Autoregressive Video Generation as a World Model [[paper](#)] [[website](#)] [[code](#)]

Junyi Chen, **Haoyi Zhu**, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, Tong He
arXiv preprint, 2025

CoMo: Learning Continuous Latent Motion from Internet Videos for Scalable Robot Learning [paper]
Jiange Yang, Yansong Shi, **Haoyi Zhu**, Mingyu Liu, Kaijing Ma, Yating Wang, Gangshan Wu, Tong He, Liming Wang
arXiv preprint, 2025

VQ-VLA: Improving Vision-Language-Action Models via Scaling Vector-Quantized Action Tokenizers [paper]
[website][code]

Yating Wang, **Haoyi Zhu**, Mingyu Liu, Jiange Yang, Hao-Shu Fang, Tong He
International Conference on Computer Vision (ICCV), 2025

Aether: Geometric-Aware Unified World Modeling [paper] [website] [code]
Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, Tong He
International Conference on Computer Vision (ICCV), 2025

SPA: 3D Spatial-Awareness Enables Effective Embodied Representation [paper] [website] [code]
Haoyi Zhu, Honghui Yang, Yating Wang, Jiange Yang, Limin Wang, Tong He
International Conference on Learning Representations (ICLR), 2025

Tra-MoE: Learning Trajectory Prediction Model from Multiple Domains for Adaptive Policy Conditioning [paper] [code]
Jiange Yang, **Haoyi Zhu**, Yating Wang, Gangshan Wu, Tong He, Limin Wang
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025

Point Cloud Matters: Rethinking the Impact of Different Observation Spaces on Robot Learning [paper] [website] [code]
Haoyi Zhu, Yating Wang, Di Huang, Weicai Ye, Wanli Ouyang, Tong He
Advances in Neural Information Processing Systems (NeurIPS), 2024

RH20T: A Comprehensive Robotic Dataset for Learning Diverse Skills in One-Shot [paper] [website] [code]
Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, **Haoyi Zhu**, Cewu Lu
IEEE International Conference on Robotics and Automation (ICRA), 2024

UniPAD: A Universal Pre-Training Paradigm for Autonomous Driving [paper] [code]
Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, **Haoyi Zhu**, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, Xiaofei He, Wanli Ouyang
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024

X-NeRF: Explicit Neural Radiance Field for Multi-Scene 360° Insufficient RGB-D Views [paper] [code]
Haoyi Zhu, Hao-Shu Fang, Cewu Lu
IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2023

MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge [paper] [website] [code]
Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, **Haoyi Zhu**, Andrew Tang, De-An Huang, Yuke Zhu, Anima Anandkumar
[Outstanding Paper Award] *Advances in Neural Information Processing Systems (NeurIPS), 2022*

For the full publication list, please visit my [Google Scholar](#) page.

SELECTED PROJECTS

Aether

Scalable and Generalizable Foundation World Model

Nov. 2024 – Present

- Proposed **Aether**, a geometry-aware unified world model. **Aether** features 3 core capabilities: (1) 4D dynamic reconstruction, (2) action-conditioned video prediction, and (3) goal-conditioned visual planning. Building upon video generation models, our framework demonstrates unprecedented synthetic-to-real generalization despite *never* observing real-world data during training.
- Developed **DeepVerse**, a 4D auto-regressive world model. **DeepVerse** can fantasize the entire world behind images and enable free exploration through interaction. It demonstrates exceptional capabilities in modeling dynamics, geometry, and physical laws.
- Introduced π^3 , a novel feed-forward neural network that revolutionizes visual geometry reconstruction by eliminating the need for a fixed reference view. With a fully permutation-equivariant architecture, π^3 achieves SOTA performance on a wide range of tasks, including camera pose estimation, monocular/video depth estimation, and dense point map estimation.
- Mentored several undergraduate and junior Ph.D. students on robot learning topics, including action tokenizers, cross-embodiment policies, point cloud policies, *etc.*

SPA

Spatial intelligence and General 3D Embodied Representation Learning

Mar. 2023 – Oct. 2024

- Conducted extensive research on general 3D representation for embodied AI. Published three first-author papers.
- Proposed a universal 3D pre-training framework, **PonderV2**, which leverages differentiable neural rendering to learn point cloud representations. Achieved state-of-the-art (SOTA) performance across 11 indoor and outdoor benchmarks.
- Performed an in-depth investigation of observation spaces in robot learning, culminating in the discovery of the critical importance of **Point Cloud Matters**.
- Developed **SPA**, an innovative representation learning framework that highlights the significance of 3D spatial awareness in embodied AI. Delivered the most comprehensive evaluation of embodied representation learning to date, surpassing over 10 SOTA methods.

MineDojo

Building Open-Ended Embodied Agents with Internet-Scale Knowledge

Feb. 2022 – Jul. 2022

- Contributed to the **MineDojo** project, a new framework designed to build generally capable agents with internet-scale knowledge in Minecraft. Won the **Outstanding Paper Award** at NeurIPS 2022.
- Co-developed the YouTube database, which comprises over 730,000 narrated Minecraft videos, totaling approximately 300,000 hours of content and 2.2 billion English transcripts.
- Constructed MineCLIP, a video CLIP foundation model that acts as a learned reward function, enabling agents to solve diverse open-ended tasks specified in natural language.
- Authored APIs for three databases (YouTube, Minecraft Wiki, and Reddit), along with the corresponding documentation and part of the official project website.

ACADEMIC SERVICES

Reviewer for journals, including IJCV, TMLR, *etc.*

Reviewer for conferences, including ICCV, NeurIPS, ICLR, WACV, ICIP, *etc.*

INVITED TALKS

- 2025/06/12, 3D CV Workshop (3D视觉工坊), “Towards Generalizable and Scalable Spatial Embodied Intelligence” [\[link\]](#)
- 2025/05/22, Embodied Intelligent Mind (具身智能之心), “Towards Generalizable and Scalable Spatial Embodied Intelligence”
- 2025/04/22, Prof. Yanyong Zhang’s Lab, USTC, “Towards Scalable World Model for General Embodied AI”
- 2025/03/27, Graduate Academic Forum, USTC, “Towards Spatial Intelligence: From 3D Vision to Embodied AI”
- 2024/12/27, ZhiXingXing Embodied AI Frontier Lecture (智猩猩具身智能前沿讲座), “Towards Spatial Intelligence: From 3D Vision to Embodied AI” [\[link\]](#)