Ziyao (Adonis) Zeng | 曾子尧

Last Update: June 2026

Tel: (+1) 475-441-2636 | Email: ziyao.zeng@yale.edu | Homepage: https://adonis-galaxy.github.io/homepage

Research Interest

- **Multimodal Learning** inspired by human cognition for **Agents and Robotics,** especially **vision-language models** and **spatial understanding.**
- My line of work in **"Language for 3D Vision"** explores **how vision-language models can perceive and understand the world like humans do.**

Educations

Yale University **Aug. 2023 – [Expected] May. 2027**

-Ph.D. in Computer Science, advised by **Prof. Yuval Kluger** and **Prof. Daniel Rakita**

ShanghaiTech University **Sept. 2019 - June 2023**

-B.Eng. in Computer Science, Major | Innovation and Entrepreneurship, Minor

- Entrepreneurial Mindshift Program, Babson College, March 2021 | International Summer Program, Osaka University, June 2022

Publications (First Author / Co-first Author) *(*: equal contribution)*

- **SciAgentArena: Benchmarking AI Agents for Addressing Scientific Challenges Across Scales**
  Tianyu Liu*, Allen Xin Wang*, Antonia Panescu*, Lisa Xinyi Chen*, Wenxin Long*, Xinyu Wei*, Yueqian Jing*, **Ziyao Zeng***, Jihang Chen, Sihan Jiang, Ziqing Wang, Siyi Gu, Siyu Chen, Xinyang Hu, Haoran Shao, Leqi Xu, Wangjie Zheng, Zhiyuan Cao, Ada Fang, Botao Yu, Kunyang Sun, Rex Ying, Arman Cohan, Qingyu Chen, Lingzhou Xue, Kaize Ding, Yuanqi Du, Wengong Jin, Zhuoran Yang, Marinka Zitnik, James Zou, Hua Xu, Hongyu Zhao
  Under Review of Nature | Link: https://arxiv.org/abs/2606.12736
  - We introduce **SciAgentArena**, a benchmark with ~200 tasks for evaluating AI agents across diverse **real-world scientific research scenarios**.
  - Current agents handle structured data-analysis workflows well but struggle to **generate novel insights** and solve **open-ended research questions**.
- **Iris: Integrating Language into Diffusion-based Monocular Depth Estimation**
  Ziyao Zeng*, Jingcheng Ni*, Daniel Wang, Patrick Rim, Younjoon Chung, Fengyu Yang, Byung-Woo Hong, Alex Wong
  CVPR 2026 | NECV 2025 Oral Presentation (18.75%) | Link: https://adonis-galaxy.github.io/Iris-website/
  - **Language improves diffusion-based depth estimators** through the conditional distributions modelled in text-to-image pre-training.
  - Language improves depth in **ambiguous regions**, enables **iterative depth refinement**, and **speed up both training and inference**.
- **RSA: Resolving Scale Ambiguities in Monocular Depth Estimators through Language Descriptions**
  Ziyao Zeng, Yangchao Wu, Hyoungseob Park, Daniel Wang, Fengyu Yang, Stefano Soatto, Dong Lao, Byung-Woo Hong, Alex Wong
  NeurIPS 2024 | Link: https://arxiv.org/abs/2410.02924
  - We are the **first** to use **language description to infer scales of 3D scenes.**
  - Trained on multiple datasets, RSA can serve **for general relative to metric depth alignment in a zero-shot setting.**
- **WorDepth: Variational Language Prior for Monocular Depth Estimation**
  Ziyao Zeng, Daniel Wang, Fengyu Yang, Hyoungseob Park, Yangchao Wu, Stefano Soatto, Byung-Woo Hong, Dong Lao, Alex Wong
  CVPR 2024 | Link: https://arxiv.org/abs/2404.03635
  - We formulate **language as a variational prior** to resolve the scale ambiguity in monocular depth estimation.
  - We train a text-VAE generating plausible depth map given a text as prior, then use **image as condition** to select most likely depth as **posterior**.
- **Can Language Understand Depth?**
  Renrui Zhang*, **Ziyao Zeng***, Ziyu Guo, Yafeng Li
  ACM Multimedia 2022 (Brave New Idea, 12.5%) | Link: https://arxiv.org/abs/2207.01077
  - We are the **first** to conduct **zero-shot training-free adaptation** from semantic language to quantified vision tasks (monocular depth estimation).
  - **Our DepthCLIP surpasses existing unsupervised methods** and even **approaches the early fully-supervised networks**.
- **RuleSmith: Multi-Agent LLMs for Automated Game Balancing**
  Ziyao Zeng, Chen Liu, Tianyu Liu, Hao Wang, Xiatao Sun, Fengyu Yang, Xiaofeng Liu, Zhiwen Fan
  arXiv technical report, 2026 | NE Agents Day 2026 | Link: https://adonis-galaxy.github.io/RuleSmith-website/
  - Multi-agent LLMs self-play with Bayesian Optimization to automatically balance numerical parameters in asymmetric strategic games.
  - Demonstrate LLMs can play complex strategic games using only rule books, showing strong adaptation and generalization of RuleSmith.
- **Coffee: Controllable Diffusion Fine-tuning**
  Ziyao Zeng, Jingcheng Ni, Ruyi Liu, Alex Wong
  arXiv technical report, 2025 | Link: https://arxiv.org/abs/2511.14113
  - Coffee controls diffusion fine-tuning to **prevent learning and entangling undesired concepts**.
  - We use **language as a regularization**, enabling **training-free**, **concept-flexible**, and **architecture-agnostic** fine-tuning.
- **DSPoint: Dual-scale Point Cloud Recognition with High-frequency Fusion**
  Renrui Zhang*, **Ziyao Zeng***, Ziyu Guo*, Xinben Gao, Kexue Fu, Jianbo Shi
  SMC 2023 | Link: https://arxiv.org/abs/2111.10332
  - We proposed DSPoint to **conduct dual-scale processing** with local point convolution and global voxel attention for efficient point cloud processing.
  - For better feature blending, we conduct inter-scale cross-modality interaction by **communicating high-frequency coordinates information**.

Publications (Co-author) *(*: equal contribution)*

- **PointCLIP V2: Adapting CLIP for Powerful 3D Open-world Learning**
  Xiangyang Zhu*, Renrui Zhang*, Bowei He, **Ziyao Zeng**, Shanghang Zhang, Peng Gao
  ICCV 2023 | Link: https://arxiv.org/abs/2211.11682
  - We introduce a **realistic shape projection** module, and leverage large-scale language models to **automatically design 3D-semantic prompt**.
  - Our approach significantly surpasses PointCLIP by **+42.90%, +40.44%, and +28.75% accuracy** on three datasets for zero-shot 3D classification.

- **Binding Touch to Everything: Learning Unified Multimodal Tactile Representations**
  Fengyu Yang*, Chao Feng*, Ziyang Chen*, Hyoungseob Park, Daniel Wang, Yiming Dou,
  Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, Alex Wong
  CVPR 2024 | Link: https://cfeng16.github.io/UniTouch/
    - We align touch embedding with a pre-trained vision-image embedding, using sensor-specific tokens for **multi-sensor** training.
    - We learn **a unified multimodal tactile representation**, enabling touch-LLM, touch-to-image generation, image texture manipulation.
- **ProtoDepth: Unsupervised Continual Depth Completion with Prototypes**
  Patrick Rim, Hyoungseob Park, S. Gangopadhyay, **Ziyao Zeng**, Younjoon Chung, Alex Wong
  CVPR 2025 | Link: https://arxiv.org/abs/2503.12745
    - We presente a **prototype-based** approach for **continual learning** of unsupervised depth completion.
    - Propose to learn **domain descriptors** that enable the model to select the appropriate prototype set for inference.
- **iQuery: Instruments as Queries for Audio-Visual Sound Separation**
  Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, **Ziyao Zeng**, Jianbo Shi
  CVPR 2023 | Link: https://arxiv.org/abs/2212.03814
    - We re-formulate visual-sound separation task and **propose Instrument as Query (iQuery) with a flexible query expansion mechanism**.
    - We demonstrate state-of-the-art performance, with **up to 44.2 % improvement** of SDR on MUSIC benchmark.
- **ETA: Energy-based Test-time Adaptation for Depth Completion**
  Younjoon Chung, Hyoungseob Park, Patrick Rim, Xiaoran Zhang, Jihe He, **Ziyao Zeng**, Safa Cicek, Byung-Woo Hong, James S. Duncan, Alex Wong
  ICCV 2025 | Link: https://arxiv.org/abs/2508.05989
    - Energy-based test-time adaptation of pretrained depth completion models by **quantifying the likelihood**.
    - **Train an energy model** that scores local regions of depth predictions by **exploring the data space with adversarial perturbations**.
- **4DP-QA: Scalable QA for 4D Perception in Vision Language Models**
  Seokju Cho, Abhishek Badki, Hang Su, Jindong Jiang, **Ziyao Zeng**, Seungryong Kim, Sifei Liu, Orazio Gallo
  CVPR 2026 | Link: https://arxiv.org/abs/2606.11568
    - We propose **True-Motion Tracking for VLM** to entanglement of camera and object motion, targeting **motion-related scene understanding**.
    - We present a QA generation pipeline that generates **large-scale 400K training samples** and a 2.2K-sample benchmark.
- **UniTac: A Unified Multimodal Model for Cross-Sensor Tactile Understanding and Generation**
  Jiahang Tu, Fengyu Yang, Chenyang Ma, Xihang Yu, **Ziyao Zeng**, Shaokai Wu, Hanbin Zhao, Zhi Tao, Chao Zhang, Hui Qian, Alex Wong
  ECCV 2026 | Link: to appear on arxiv soon
    - We propose UniTac, the **first Unified multimodal models designed for tactile understanding and generation**.
    - UniTac enhances reasoning over physical and cross-sensor information, with a sensor-prior-based sampling that simulates real contact for generation.
- **A Physics-Grounded Benchmark for Multi-Agent Dynamics in World Models**
  Nuo Chen, Lulin Liu, Zihao Li, **Ziyao Zeng**, Zihao Zhu, Wenyan Cong, Junyuan Hong, Yunhao Yang, Zhengzhong Tu, Yan Wang, Boris Ivanovic, Marco Pavone, Zhangyang Wang, Yang Zhou, Zhiwen Fan
  ECCV 2026 | Link: to appear on arxiv soon
    - We introduce **CrashTwin**, a **physics-grounded** world model evaluation framework with **25K synthetic** and **12K in-the-wild** collision sequences.
    - We find that **high perceptual quality frequently masks severe physical violations** in world models during complex multi-agent interactions.
- **OpenLongTail: Generative Scaling of Long-Tail Driving Data**
  Lulin Liu, Nuo Chen, Yan Wang, Bangya Liu, Wenyan Cong, Hezhen Hu, Boris Ivanovic, Hao Wang, **Ziyao Zeng**, Xinyu Gong, Yang Zhou, Zixiang Xiong, Dilin Wang, Zhangyang Wang, Weisong Shi, Ruohan Zhang, Marco Pavone, Zhiwen Fan
  Submit to a top-tier conference | Link: Coming soon
    - We introduce OpenLongTail, **scaling monocular in-the-wild long-tail driving videos into multi-view assets** for long-tail training.
    - Scaling with monocular in-the-wild long-tail driving videos significant **gains in closed-loop driving robustness on long-tail events**.
- **Artificial Foveated Perception for Mitigating Shortcut Learning in Robotic Foundation Models**
  Xiatao Sun, Yuan Zhuang, Mateo Sanchez Lopez Negrete, Matei-Victor Coldea, Chen Liang, Haoyang Zhang, Che Liu, **Ziyao Zeng**, Shawn Li, Qian Wang, Fei Miao, Daniel Rakita
  Submit to a top-tier conference | Link: Coming soon
    - We propose AFP (Artificial Foveated Perception), **predicting task-conditioned masks over action-relevant regions** (objects, robot, critical areas).
    - AFP can be integrated into VLA and World Action Model to **suppress spurious scene-level correlations** to improve generalization.
- **HOMER: Homography-Based Efficient Multi-view 3D Object Removal**
  Jingcheng Ni*, Weiguang Zhao*, Daniel Wang, **Ziyao Zeng**, Chenyu You, Alex Wong, Kaizhu Huang
  arXiv technical report, 2025 | Link: https://arxiv.org/abs/2501.17636
    - **Remove objects** across multi-view images with region-based iteration in a single frame with **homography-based mapping**.
    - Present **a 3D multi-object removal dataset** with greater object diversity and viewpoint variation than existing datasets.
- **NeuroBind: Towards Unified Multimodal Representations for Neural Signals**
  Fengyu Yang*, Chao Feng*, Daniel Wang*, Tianye Wang, **Ziyao Zeng**, Zhiyang Xu,
  Hyoungseob Park, Pengliang Ji, Hanbin Zhao, Yuanning Li, Alex Wong
  arXiv technical report, 2024 | Link: https://arxiv.org/abs/2407.14020
    - We propose **a unified neural representation** that aligns multiple neural signal modalities, with pre-trained vision-language embeddings.
    - We enhances downstream performance on brain signal classification, retrieval, image reconstruction, and Neuro-LLM.

- **VT-CLIP: Enhancing Vision-Language Models with Visual-guided Texts**
  Longtian Qiu, Renrui Zhang, Ziyu Guo, **Ziyao Zeng,** Yafeng Li, Guangnan Zhang
  arXiv technical report, 2021 | Link: https://arxiv.org/abs/2112.02399
  - To adapt CLIP to few-shot classification, we enhance vision-language modeling **via visual-guided texts**.
  - Specifically, we **guide text feature to adaptively explore informative regions on the image** and aggregate the visual feature by cross-attention.

Research / Working Experience

- **Nvidia Research** May 2025 – March. 2026
  - Research Intern | Work with **Dr. Hang Su**, **Dr. Jindong Jiang**, **Dr. Abhishek Badki**, **Dr. Orazio Gallo**, and **Dr. Sifei Liu**
  - **Ziyao Zeng** et al. "Boosting Spatial Understanding in VLMs with Spatial State Tokens." U.S. Patent Application (Pending), NVIDIA Corp., 2026.
- **Shanghai AI Laboratory** Oct. 2022 – March 2023
  - Intern Researcher | Project: Visual Grounding, Prototype-based Segmentation
- **GRASP Lab, University of Pennsylvania** April 2021 - Dec. 2022
  - Work with **Prof. Jianbo Shi** and **Dr. Renrui Zhang** | Project: DSPoint, iQuery, Reconstruction-based Image Segmentation
- **UISEE AI Lab, UISEE Technology** July 2021 - April 2022
  - Intern Algorithm Engineer | Develop point cloud recognition & monocular 3D object detection algorithm for self-driving cars
- **PLUS Lab, ShanghaiTech University** July 2020 - June 2021
  - Working with **Prof. Xuming He** and **Zhitong Gao** | Project: Mix-data Anomaly Segmentation, Noisy Labels Classification, Uncertainty Estimation

Service / Tutoring Experience

Reviewer: CVPR (2022, 2025 **[Outstanding Reviewer]**, 2026), ICCV (2023, 2025), ECCV (2024, 2026), ICML (2025, 2026), ICLR (2025, 2026), NeurIPS (2024, 2025), ACM MM (2023, 2025), AISTATS (2024, 2025), ICASSP (2024, 2025), BMVC (2026), TCSVT, TMLR
Volunteer of WWF China & Greenpeace; Yale CS PhD '24 Mentor;
ShanghaiTech Freshman Mentor; Intern Lecturer of Shanghai Advanced Research Institute;

Awards / Competitions

Sigma Xi, The Scientific Research Honor Society, Membership Nomination 2025
Merit Student of ShanghaiTech University 2021, 2022
State Grid AI Innovation Competition in Power Dispatch and Control (Top 10%, 24/242) 2021

Skills

Claude Code, Python/PyTorch, C/C++, Unity/C#, Bilingual (Chinese/English: CET6-607, GRE-327+4, TOEFL-103)