Xinyu Zhang

Tel: +1 848-228-6208

E-mail: xz653@rutgers.edu

Website: https://mlzxy.github.io

Introduction

I am a CS PhD candidate (end of 4th year, graduating this August) at **Rutgers University**, focusing on **computer vision and robotic policy learning**. I published 8 first-author papers at top robotics and vision venues. I worked as a research intern at **Meta Reality Lab** and an ML engineer at **Microsoft**. My publications span robot manipulation policy learning, world modeling, computer vision, 3D reconstruction, and dexterous hand video generation.

Publication

Xinyu Zhang, Zhengtong Xu, Yutian Tao, Yeping Wang, Yu She, Abdeslam Boularias.
Learning Visual Feature-based World Models via Residual Latent Action
NeurIPS 2026 Under Review [website] [code] [arxiv]

- Proposed RLA-WM, a simple and efficient state-of-the-art world model.
- The first demonstration of visual reinforcement learning inside a world model that learned from only offline videos, without online rollouts or handcrafted rewards.

Xinyu Zhang, Meta Reality Lab.
Glove2Hand: Synthesizing Natural Hand-Object Interaction from Multi-Modal Sensing Gloves.
CVPR 2026 Highlight [website] [arxiv]

- Proposed Glove2Hand, a 3d reconstruction + video generation framework that synthesizes temporally consistent, photorealistic real-hand object interaction from sensor gloves.
- It enables force sensing in human videos and aligns videos of multiple hand manipulators.

Xinyu Zhang, Haonan Chang, Yuhan Liu, Abdeslam Boularias.
Motion Blender Gaussian Splatting for Dynamic Scene Reconstruction.
CoRL 2025 [code & demo] [arxiv]

- Proposed Motion-Blender-GS, which reconstructs dynamic scenes into Gaussian splats that are controlled by motion graphs. The motion graph is an explicit and sparse motion representation, which enables digital twin and real2sim2real after reconstruction

Xinyu Zhang, Yuhan Liu, Haonan Chang, Liam Schramm, Abdeslam Boularias.
Autoregressive Action Sequence Learning for Robotic Manipulation.
RAL 2025 [code & demo] [arxiv]

- Proposed Autoregressive Policy, a unified LLM architecture for policy learning that uses variable-size chunking to support action prediction during token generation. This unified framework established SOTA in multiple environments at the time

Xinyu Zhang, Yuhan Liu, Yuting Wang, Abdeslam Boularias.
Detect Everything with Few Examples.
CoRL 2024 [code] [demo] [arxiv]

- Proposed DE-ViT, the first example-driven open-world detector that detects new categories instantly without any tuning
- Achieved SOTA on LVIS, COCO, and outperforms the second by +10 mAP at the time

Xinyu Zhang, Abdeslam Boularias.
One-Shot Imitation Learning with Invariance Matching for Robotic Manipulation.
RSS 2024 [code] [video] [arxiv]

- Proposed IMOP, a one-shot policy, which solves novel quasi-static manipulation tasks by matching 3D regions between demonstrations and testing scenes, and transferring

end-effector poses through matched correspondence

- The policy delivered non-trivial success rates (41%, +11% compared to retraining baseline) on 22 novel tasks on RLBench, with just a single demonstration per task

Xinyu Zhang, Yuhan Liu, Haonan Chang, Abdeslam Boularias.
Scaling Manipulation Learning with Visual Kinematic Chain Prediction.
CoRL 2024 [code] [demo] [arxiv]

- Proposed VKT, a generalist policy that effectively bridges embodiment and workspace gaps by forecasting the movements of the visual kinematic chain as a universal action space (e.g., from 10% to 55% success rate for a generalist on RLBench).

Yuhan Liu, **Xinyu Zhang**, Haonan Chang, and Abdeslam Boularias.
Failure Forecasting Boosts Robustness of Sim2Real Rhythmic Insertion Policies.
IROS 2025 [code, demo] [arxiv]

- Proposed an online estimation model that learns to predict policy success rates, which solves repeated insertion tasks with sim-to-real transfer

Haonan Chang, Kowndinya Boyalakuntla, Yuhan Liu, **Xinyu Zhang**, Liam Schramm, Abdeslam Boularias.
Diffusion-based Affordance Prediction for Multi-modality Storage.
IROS 2024 [code] [arxiv]

- Proposed a diffusion policy that predicts an affordance matching (similar to the region matching in IMOP) from multiple storage slots. The insertion pose is derived from the matched regions

Xinyu Zhang, Abdeslam Boularias.
Optical Flow boosts Unsupervised Localization and Segmentation.
IROS 2023 [code] [arxiv]

- Proposed Flow-DINO, which uses optical flow as a self-supervised learning objective, to encourage object-aware spatial features. The learned features improve downstream object localization and segmentation performance

Xinyu Zhang, Ian Colbert, Srinjoy Das.
Learning Low-Precision Structured Subnetworks Using Joint Layerwise Channel Pruning and Uniform Quantization.
MDPI 2022 [code] [paper]

- Proposed Recursive Pruning, which identifies inter-layer neuron dependencies, and prunes layer by layer in topological order, outperforms magnitude-based pruning consistently over classification, segmentation, and image generation

Xinyu Zhang, Srinjoy Das, Ken Kreutz-Delgado.
Tuning Confidence Bound for Stochastic Bandits with Bandit Distance.
Arxiv 2021 [paper]

- Proposed UCB-DT, which is a stochastic bandit method informed by estimating the distribution distance between the rewards of different arms, with better theoretical bound than UCB, and consistently outperforms other UCB variants such as UCB-dagger

Mengting Chen, Yuxin Fang, Xinggang Wang, Heng Luo, Yifeng Geng, **Xinyu Zhang**, Chang Huang, Wenyu Liu, Bo Wang.
Diversity transfer network for few-shot learning.
AAAI 2020 [paper]

- Classify any objects by generating features from example images

Education

Rutgers University 2022 - Present
PhD Candidate, Computer Science

University of California, San Diego 2015 - 2017
M.S., Computer Science

University of Science and Technology of China 2011 - 2015
B.E., Electrical Engineering

Industrial Experience

Meta 2025 Jun - Dec
Research Scientist Intern

- Proposed Glove2Hand, a 3D-2D hybrid generative framework that integrates reconstruction and video diffusion for consistent and flexible hand generation.
- My proposed method reconstructs photorealistic bare-hand video from tactile glove videos, which enables multi-modal sensing with egocentric bare hands for object manipulation

Microsoft 2020 - 2022
Software Engineer II

- Developed contextual bandit methods such as LinUCB and Thompson Sampling, for reranking ads, news, and widgets in the Edge browser, using C#, Kafka, and Flink

Megvii (Face++) 2018 - 2020
Machine Learning Engineer

- Developed large-scale face recognition networks, based on representation learning with contrastive loss, AMSoftmax, and self-training over billion-scale datasets

Skills

- Python, PyTorch, JavaScript, C/C++, CUDA, ROS 1, LaTeX, Slurm / Kubernetes, MuJoCo, SQL, MoCap System, OptiTrack, MANO, SMPL
- Imitation Learning, Model Compression, Reinforcement Learning, Self-Supervised Learning, Open-Vocabulary Detection/Segmentation
- Point Transformer, Spatially Sparse CNN, Point Cloud Registration, DINOv2, DETR, CLIP, MaskFormer, ViT, UNet, Mask-RCNN, Diffusion, VAE, GAN, ControlNet