Yusu Fang

Homepage | *Scholar* | *Email* : *ysfang*0306@*gmail.com*

Education

Peking University — Beijing, China
School of Electronics Engineering and Computer Science — Sep. 2022 — Jun. 2026
B.Sc. Electronic Information Engineering
GPA: 88/100, 3.71/4.0

Stanford University — California, United States
Computer Science Department, School of Engineering — June. 2024 — Present
Research Assistant, Supervised by Prof. Ehsan Adeli

Publications

HumanScore: Benchmarking Human Motions in Generated Videos
Yusu Fang*, Tiange Xiang*, Tian Tan, Narayan Schuetz, Scott Delp, Li Fei-Fei†, Ehsan Adeli†
Under Reiew of CVPR 2026

SocialGen: Modeling Multi-Human Social Interaction with Language Models
Heng Yu*, Juze Zhang*, Changan Chen, **Yusu Fang**, Tiange Xiang, Juan Carlos Niebles, Ehsan Adeli
3DV 2026

The Language of Motion: Unifying Verbal and Non-verbal Language of 3D Human Motion
Changan Chen*, Juze Zhang*, S. K. Lakshmikanth*, **Yusu Fang**, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, Ehsan Adeli
CVPR 2025

Research Experience

Egocentric Human–Object Interaction Reconstruction from Egocentric Video — Oct. 2025 — Present
Research Intern with ***Prof. Lingjie Liu*** *and* ***Prof. Kostas Daniilidis*** *at GRASP Lab,* ***University of Pennsylvania***
This project explored how to reconstruct high-precision full-body 3D human poses and object trajectories from egocentric videos of everyday interactions, with an eye toward powering imitation learning for humanoid robots.

- Developed a reconstruction pipeline that combines hand and object pose estimation in camera space, lifts them into world coordinates using head pose, and infers full-body SMPL-X motion consistent with these local interaction cues.
- Extended egocentric motion modeling to jointly reason about head, hand, and object poses, enabling detailed 3D reconstructions of human–object interactions from noisy first-person sensor streams.

Benchmarking Human Motions in Generated Videos — Aug. 2025 — Present
Research Intern with ***Prof. Ehsan Adeli*** *and* ***Prof. Li Fei-Fei*** *at Stanford Vision and Learning Lab,* ***Stanford University***
This project introduced the first systematic benchmark for human motion realism in AI-generated videos, bridging the gap between visual fidelity and biomechanical plausibility.

- Introduced a comprehensive evaluation framework centered on human figures and motions in generated videos, complementing existing appearance- and semantics-based benchmarks.
- Defined a suite of interpretable, biomechanics-grounded metrics for anatomy, kinematics, and kinetics, together with transparent baselines and reference implementations.
- Revealed consistent gaps between visual plausibility and motion fidelity in state-of-the-art video generators and outlined concrete directions for improving human motion quality.

Reusing Language Models for Motion Understanding and Generation — Apr. 2025 — Jul. 2025
Research Intern with ***Prof. Ehsan Adeli*** *at Stanford Vision and Learning Lab,* ***Stanford University***
This project proposed a Pose-as-Text framework that mapped 2D/3D human poses into pseudo text tokens so they could be processed directly by frozen language and vision–language encoders in multimodal foundation models.

- Analyzed the limitations of standard late-fusion multimodal architectures, identifying a behavior gap between motion tokenizers and web-scale text encoders driven by data imbalance and weak cross-modal alignment.
- Designed an early-fusion approach that learned lightweight mappers from 2D/3D/SMPL poses into text-like token sequences, enabling plug-and-play pose conditioning for existing CLIP/T5-style generative models.
- Demonstrated pose-to-image generation with strong alignment and novel-pose generalization, and showed how the same abstraction naturally extends to scalable "pose-to-anything" tasks such as pose-to-text and motion-to-video.

Modeling Multi-Human Social Interaction with Language Models — Dec. 2024 — Mar. 2025
Research Intern with ***Prof. Ehsan Adeli*** *at Stanford Vision and Learning Lab,* ***Stanford University***

This project addresses the challenge of modeling realistic multi-human social interactions by unifying motion and language in a single framework that scales beyond traditional two-person settings.

- Proposed a language-model-based framework for multi-human social interaction that supports both motion generation and diverse motion–language understanding tasks.
- Designed a scalable motion representation that handles varying numbers of interacting people, enabling effective modeling of complex social group dynamics.
- Built SocialX, a curated multi-human interaction benchmark with paired motion and language, establishing a standard testbed for evaluating social interaction models.

Unifying Verbal and Non-verbal Language of 3D Human Motion Aug. 2024 — Nov. 2024
*Research Intern with **Prof. Ehsan Adeli** at Stanford Vision and Learning Lab, **Stanford University***
This project unifies verbal and non-verbal communication by treating 3D human motion, speech, and text as a single "language," enabling flexible cross-modal understanding and generation across these modalities.

- Formulated expressive full-body human motion as a discrete sequence aligned with speech and text, allowing one model to jointly reason over all three modalities.
- Demonstrated that a single multimodal language model can support a wide range of motion–language tasks (e.g., co-speech gesture generation, motion captioning, emotion recognition) without task-specific architectures.
- Showed improved generalization and data efficiency over modality-specific baselines, especially in low paired-data regimes and under novel modality combinations at inference time.

Motion-Aware 4D Gaussian Human Avatars from a Single Image Jun. 2024 — Jul. 2024
*Research Intern with **Prof. Ehsan Adeli** at Stanford Vision and Learning Lab, **Stanford University***
This project explored how to generate a 4D, animatable human avatar and surrounding scene from a single reference image, focusing on temporal consistency and motion fidelity.

- Developed a two-stage pipeline that first generates a monocular human video from a single image and target motion, then reconstructs per-frame SMPL meshes from the video and uses them to build a temporally coherent 4D Gaussian avatar.
- Introduced motion-aware constraints by initializing the Gaussian point cloud from the reconstructed SMPL meshes and predicting only small residual updates, substantially improving stability of the geometry over time.

Awards

Research Excellence Award 2025
Awarded for outstanding research experience, top 10% of Peking University.

Outstanding Overseas Exchange Scholarship 2024
Awarded for academic performance during the exchange in Stanford.

Finalist of Interdisciplinary Contest in Modeling (ICM) 2024
Awarded for academic and innovative outstanding, top 0.3% worldwide.

Samsung Scholarship 2023
Awarded for academic and research performance, top 1% of EECS Department, Peking University.

Academic Excellence Award 2023
Awarded for academic and innovative outstanding, top 10% of Peking University.

Project Experience

Motion Matching — Python Controlled the character's actions such as running, turning, and walking through keyboard and mouse inputs, utilizing the motion matching algorithm to achieve natural transitions between different actions.
Link to project: https://github.com/Daydreamer-f/Motion-Matching.git.

Domain Adaptive Object Detection — Python Developed unsupervised domain-adaptive object detection methods using adversarial feature alignment and a teacher-student model to bridge the gap between source and target domains, significantly enhancing model performance in varied deployment environments without additional labeling.
Link to project: https://github.com/Daydreamer-f/Domain-Adaptive-Object-Detection.git.

Skills

- **Language:** TOEFL 104 (S25), GRE 320+3.5.
- **Programming:** Python, C++, MATLAB, LaTeX
- **Research Tools:** PyTorch, NumPy, OpenGL, OpenCV, Blender