Chenyi Xu

Phone: (+86) 153-2567-2086 — GitHub: https://github.com/ChenyiXu04
Email: chenyixu2004@gmail.com — Website: https://chenyixu04.github.io/

Education

Hangzhou Dianzi University Hangzhou, China
B.Sc. in Computer Science and Technology *GPA: 3.65 / 4.0* *2023.09 – 2027.06*
Core Courses: Advanced Mathematics, Probability Theory, Linear Algebra, Software Engineering, Computer Networks

Research & Projects

Embodied AI Research Hangzhou, China
Research Assistant, School of Computer Science, Hangzhou Dianzi University *2025.3 – Present*

- Investigated Vision-Language-Action (VLA) models for robotic manipulation, optimizing end-to-end policy learning with latent action representations and spatial grounding to improve generalization across unseen objects and environments.
- Explored Vision-and-Language Navigation (VLN) in photorealistic 3D environments, developing history-aware visual-text cross-modal matching and adaptive waypoint planning to reduce navigation drift in long-horizon tasks.

4D Vision Reconstruction Hangzhou, China
Visiting Student, AGI Lab, Westlake University *2025.07 – 2025.09*

- Proposed a monocular dynamic video 4D reconstruction framework based on 4DGS, using single-step diffusion models and information-gain-driven pose selection to generate high-quality pseudo-views for multi-view supervision.
- Modeled scene dynamics using a canonical-space 4DGS representation and a globally shared $SE(3)$ motion basis trajectory, integrating a Neural Texture Field to achieve stable long-sequence reconstruction.

JEPA-based VLA Learning Ningbo, China
Research Intern, HAI Lab, Eastern Institute of Technology *2026.07 – Present*

- Developed a JEPA-based Vision-Language-Action framework for embodied intelligence, leveraging predictive latent representations to learn task-relevant visual dynamics from large-scale video and robot interaction data.
- Modeled future state transitions in latent embedding space with self-supervised video prediction objectives, integrating learned representations into VLA policies to improve generalization and data efficiency in robotic manipulation tasks.

Publications

Yan, L.†, Wu, Y., **Xu, C.**, Yang, C. & Li, P. AR-Nav Benchmark: Augmented Reality Navigation with Vision and Language. *The 40th AAAI Conference on Artificial Intelligence* (†Mentor). (**AAAI 2026, Oral**).

- Proposed the AR-VLM model to construct high-fidelity 3D semantic fields via symbolic geometric alignment, resolving the long-term memory deficiency in AR navigation.
- Developed the ARN-Pilot module for lightweight topological reasoning and sub-second path planning, establishing the first standardized multi-scenario AR-Nav benchmark framework.

Yang, M., Yang, Y., **Xu, C.**, Song, C., Zuo, Y., Zhao, T., Li, R., & Zhang, C.† Fast3Dcache: Training-free 3D Geometry Synthesis Acceleration. *The 43rd IEEE/CVF Conference on Computer Vision and Pattern Recognition* (**CVPR 2026, Poster**).

- Proposed Fast3Dcache, a training-free 3D geometric caching framework leveraging voxel stabilization regularities during 3D diffusion generation to accelerate inference.
- Designed dynamic cache scheduling (PCSC) and stability discrimination (SSC) based on motion features, achieving a 27% speedup and 54% FLOPs reduction while preserving geometric quality.

Xu, C., Wu, Y., Yan, L.†, Yang, C., Zhang, J., Guan, F.& Li, P. LongDPM: Overlap-Aware 4D Reconstruction from Long Monocular Videos. In *The 9th Chinese Conference on Pattern Recognition and Computer Vision (**PRCV 2026, Poster**).*

- Proposed the LongDPM framework for long video 4D reconstruction via overlap-aware chunked inference, separating static anchors from dynamic features to ensure robust cross-chunk registration.
- Developed a geometrically-conditioned trajectory association and fusion mechanism to maintain identity continuity across chunks, achieving SOTA performance across multiple benchmarks.

Wu, Y., Cai, C., Yan, L.†, Li, M., **Xu, C.**, Yang, C., Guan, F., & Li, P. ShapeNav: 3D Geometric Perception and Reasoning for Multi-Floor Vision-and-Language Navigation. *ACM International Conference on Multimedia 2026 (**ACM MM 2026, Oral**).*

- Proposed ShapeNav, a 3D geometric navigation framework without 3D pre-training. It integrates point clouds and vision-language models, with GPE and MOT modules to align 3D geometric and 2D visual features for cross-modal fusion.
- SOTA navigation results are obtained on R2R and F2F benchmarks, with 3% higher success rate and 0.29 m lower navigation error than DUET. The framework retains strong performance even when only half of the training data is available.

C. Yang.*, **C. Xu.***, L. Yan.†, Z. Li., G. Min., F. Guan., J. Zhang., & P. Li. ToolUse: A Generative Robotic Benchmark for Language-Guided Open-World Tool Manipulation**(Under Preparation For ICASSP 2027).**

- Proposed Tool-VLA, a geometry-aware vision-language-action model for open-world robotic tool manipulation. It leverages semantic-geometric alignment (SGA) to extract tool functional frames, augmenting language prompts with grasp semantics, functional-part descriptions and execution direction for geometry-grounded cross-modal policy learning.

- Established the ToolUse generative benchmark covering multi-tier tool-use tasks. Experiments show that simulation pre-training on ToolUse brings over 40% real-world success improvement compared with real-data-only training, and achieves consistent gains under limited training data proportions.

Wu, Y.*, **Xu, C.***, Yan, L.†, Cai, C., Min, G., Guan, F., Lin, B., Zhang, J. & Li, P. BrainNav: Towards Dual-Brain Minimal Sufficient Representation for Vision-Language Navigation. (*Equal contribution). (**Under Preparation For ICRA 2027**).

- Proposed the BrainNav dual-brain framework for long-horizon navigation, utilizing instruction-driven attention filtering to eliminate irrelevant visual noise and enhance language-spatial alignment.
- Constructed a low-rank constrained alignment and a compact world model for state prediction within compressed latent spaces, achieving SOTA performance on R2R and RxR benchmarks.

Xu, C.*, Wu, Y.*, Yan, L.†, Bai, Z., Wang, Q., Guan, F., Zhang, J. & Li, P. Dense Soccer Video Captioning with Global Context and Player Position Awareness. (**Scientific Reports, Under Review**).

- Proposed the SCPR framework for dense soccer video captioning, leveraging Perceiver attention (GCPR/PCPR modules) to fuse global-local video context with multi-modal player positions.
- Integrated an SAA spatial enhancement module to inject pitch tactical statistics for LLM decoding, yielding optimal fine-grained event localization and SOTA performance on public benchmarks.

Awards, Honors & Competitions

CUMCM "Higher Education Press Cup" | Provincial First Prize, National Third Prize	2024.09
Service Outsourcing Innovation & Entrepreneurship Competition | National Third Prize	2024.12 – 2025.03
ICT Industry–Education Integration Innovation Competition | National Second Prize	2024.08 – 2024.10
ICT Industry–Education Integration Innovation Competition | National Second Prize	2025.08 – 2025.10
Zhejiang Provincial Xinmiao Talent Program | Project Leader	2025.12 – 2026.12
First-Class Academic Scholarship & Outstanding Student Leader Award | HDU	2024 & 2025

Skills & Leadership

- **Campus Leadership:** Class Monitor (Class 23052314) — Class Assistant (Class 25050417).
- **Languages:** English (CET-4 & CET-6 qualified).
- **Intellectual Property:** 7 Invention Patents filed/authorized (Public Nos.: CN122597679A, CN122524119A, CN122392071A, CN122322854A, CN122518363A, CN122378690A, CN120744017A).