Wei-Jin Huang

Ph.D. Student in Computer Science and Technology | Sun Yat-sen University Guangzhou, China

huangwj235@mail2.sysu.edu.cn vkgo.github.io Google Scholar github.com/vkgo

Research Profile

- **Research Focus:** Ph.D. student at Sun Yat-sen University, advised by Prof. Wei-Shi Zheng, working on **humanoid robot learning and video action understanding**.

Education

Sun Yat-sen University (Ph.D.) School of Computer Science and Engineering Sep. 2024 – Present
Computer Science and Technology | Research: Embodied Intelligence and Computer Vision (Humanoid Robot Learning, Video Action Understanding) | Advisor: Prof. Wei-Shi Zheng (Changjiang Scholar)

- **Honors:** Sun Yat-sen University President's Scholarship (top 5%)
- **Publications:** Multiple first/co-first-author papers at CVPR, ECCV, and ACM MM.
- **Internship Interests:** Combining VLA/WAM models, agentic planning, and continual learning with egocentric human data and robot experience for compositional generalization, rapid adaptation, and failure recovery.

South China University of Technology (B.E.) School of Computer Science and Engineering Sep. 2020 – Jul. 2024
Network Engineering | Major Rank: 3/66

- **Honors:** National Scholarship (1/66), First-Class Tencent Scholarship (1/66)
- **Competitions:** Kaggle H&M Personalized Fashion Recommendations, Silver (85/2952, 2022); Kaggle G2Net Gravitational Wave Detection, Silver (57/1219, 2021)

Selected Publications

Beyond Mimicry: Learning Whole-Body Human-Humanoid Interaction from Human-Human Demonstrations
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2026)
Wei-Jin Huang, Yue-Yi Zhang, Yi-Lin Wei, Zhi-Wei Xia, Juantao Tan, Yuan-Ming Li, Zhilin Zhao, Wei-Shi Zheng

- **Overview:** Despite abundant human–human data, high-quality human–humanoid demonstrations are scarce: retargeting breaks key contacts, while standard imitation favors trajectory reproduction over interaction reasoning. PAIR generates executable supervision via contact-preserving, physics-aware retargeting. D-STAR separates *when*/*where*: Phase Attention models stages, Multi-Scale Spatial locates targets, and a diffusion head fuses both into synchronized interactions.

ChoreoPlan: Hybrid Phrase Planning and Execution-Grounded Selection for Music-to-Humanoid Dance
ACM International Conference on Multimedia (ACM MM 2026)
Wei-Jin Huang, Jianhong Fan, Hao Huang, Zhi-Wei Xia, Jun-Yi Deng, Yuan-Ming Li, Kun-Yu Lin, Wei-Shi Zheng

- **Overview:** Reference-space plans may look optimal but execute unstably or off-beat under whole-body control (WBC)—a reference-to-execution gap. Since fixed-window planning cannot adapt to variable rhythmic/phrase structures and reference quality misrepresents execution, ChoreoPlan combines beat-aligned variable-length discrete–continuous phrase planning with an Embodied Selector trained on offline WBC rollouts to preserve executability, rhythm, and semantics.

Modeling Multiple Normal Action Representations for Error Detection in Procedural Tasks
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025)
Wei-Jin Huang, Yuan-Ming Li, Zhi-Wei Xia, Yu-Ming Tang, Kun-Yu Lin, Jian-Fang Hu, Wei-Shi Zheng

- **Overview:** Procedural error detection must handle multiple valid next steps after the same history. Fixed temporal rules or a single static prototype ignore these futures and may select the wrong comparison target as execution changes. AMNAR predicts all valid next actions, reconstructs an adaptive normal representation for each, matches the current action to the best candidate, and thresholds their distance—avoiding incorrect prototypes under ambiguous futures.

EgoExo-Fitness: Towards Egocentric and Exocentric Full-Body Action Understanding
European Conference on Computer Vision (ECCV 2024)
Yuan-Ming Li, **Wei-Jin Huang (co-first author)**, An-Lan Wang, Ling-An Zeng, Jing-Ke Meng, Wei-Shi Zheng

- **Overview:** Ego/Exo understanding lacks synchronized data and fine-grained labels for what/when/how well. EgoExo-Fitness records 1,276 sequences from 40 participants with six synchronized cameras, annotated with two-level temporal boundaries, technical keypoint verification, natural-language comments, and quality scores. It benchmarks classification/localization, cross-view sequence verification and skill determination, and guidance-based execution verification.

Less Static, More Private: Towards Transferable Privacy-Preserving Action Recognition by Generative Decoupled Learning
IEEE/CVF International Conference on Computer Vision (ICCV 2025)
Zhi-Wei Xia, Kun-Yu Lin, Yuan-Ming Li, **Wei-Jin Huang**, Xian-Tuo Tan, Wei-Shi Zheng

- **Overview:** Privacy-preserving action recognition transfers poorly across domains; privacy mainly lies in static appearance. GenPriv uses labeled source and unlabeled target videos with a generative ST-VAE to separate static/dynamic features and remove privacy from static ones. Mutual-information minimization decouples them; spatial consistency stabilizes static representations, and temporal alignment learns domain-consistent motion—preserving transferable action features.