Jihai Zhang

E-mail: jihaizhang@link.cuhk.edu.hk **HomePage:** majordavidzhang.github.io

EDUCATION

Shanghai Jiao Tong University (Graduated with Honors) Sep. 2017 - Jul. 2021

Bachelor of Computer Science

National University of Singapore Aug. 2021 - Jun. 2024

Master of Science in Computer Science

The Chinese University of Hong Kong Aug. 2024 – Jun 2028 (expected)

PhD in Computer Science and Engineering

RESEARCH INTERESTS

Multimodal Large Language Model, Game World Model, World Action Model

RECENT PROJECT & RESEARCH

CLIP Mixture of Experts Sep. 2024 – May. 2025

Used the proposed Multi-stage Contrastive Learning (MCL) to obtain a series of CLIP's MLP parameter sets, each set corresponding to one stage. Used the different MLPs as different experts to initialize a Mixture of Experts (MoE), and successfully improved the model performance, especially when served as a vision encoder of Multimodal Large Language Models. The research was published in EMNLP 2025.

Sequence Compression for Autoregressive Video Generation Models Sep. 2024 – Jun. 2025

Addressed the issue of excessive redundancy in video representation sequences within autoregressive video generation models by proposing a sequence compression algorithm. Introduced an additional information bottleneck structure to compress the sequence length of the autoregressive video generation model by 4x, achieving a 10x improvement in inference speed while significantly reducing GPU memory consumption during both training and inference. By compressing redundant information, the algorithm alleviated model collapse during training, improved training quality, and achieved better generative performance compared to the uncompressed model.

The research paper is under review in AAAI 2025.

Generalization Across Understanding and Generation in Unified VLMs

Microsoft - Research Intern Oct. 2024 – Jun. 2025

Investigated the interplay between understanding and generation tasks in the unified training of Vision-Language Models (VLMs). Demonstrated that understanding and generation tasks can mutually enhance each other, with the

underlying mechanism attributed to cross-task knowledge transfer. Identified that the mutual benefit is closely tied to the alignment between the model's input visual space and output visual space. Validated the importance of integrating understanding and generation tasks for achieving optimal performance in VLMs.

Game Video Generation With Explicit Layout Representations

Tencent - **Research Intern** Jul. 2025 – Jan. 2026

Architected a novel, dual-model framework that decouples semantic layout prediction from pixel-level video rendering to achieve highly controllable game video generation. Designed a Transformer-based model to predict the dynamic state of object layouts, which are then used to guide a Diffusion Forcing model via cross-attention. Implemented a global "Memory Bank" to maintain a persistent world state, ensuring long-term spatial-temporal consistency and object permanence in generated videos.

Game World Model: Matrix Game 3.5

Skywork AI - **Research Intern** May. 2026 – Jul. 2026

Contributed to the Matrix Game 3.5 world model, with a focus on memory and consistency. Led the design and implementation of the memory module. Built a dynamic object filtering pipeline for Mosaic Memory, preventing moving objects from contaminating long-term scene reconstruction. Designed object-token conditioning to maintain protagonist consistency and support controllable protagonist generation during long-horizon rollouts.

World Action Model

Skywork AI - **Research Intern** Jul. 2026 – Present

Developing an embodied policy model based on LingBot-Video, an MoE video generation model. Designed and implemented a video-action Mixture-of-Transformers (MoT) architecture with autoregressive causal masking. Enabled action-only inference for interactive policy execution.

Game World Model: Matrix Game Next

Skywork AI - **Research Intern** Aug. 2026 – Present

Leading Matrix Game Next, a neural rendering project for real-time game visual enhancement. Developing a model that takes code-agent-generated-game frames as input and re-renders them into AAA-quality visuals while preserving gameplay-relevant structure and motion. Targeting approximately 300 ms end-to-end latency for real-time interactive rendering.

SELECTED PUBLICATIONS

- **JihaiZhang**, Tianle Li, Linjie Li, Zhengyuan Yang, Yu Cheng. *Cross-Task Generalization Between Understanding and Generation in Unified Vision-Language Models: A Controlled Study*. The 37th British Machine Vision Conference (BMVC2026)
- **Jihai Zhang**, Yucheng Zhou, Guanjie Chen, Jianbing Shen, Yu Cheng. *Less Is More: Vision Representation Compression for Efficient Video Generation with Large Language Models*. The 40th Annual AAAI Conference on Artificial Intelligence (AAAI2026)
- **Jihai Zhang**, Xiaoye Qu, Tong Zhu, Yu Cheng. *CLIP-MoE: Towards Building Mixture of Experts for CLIP with Diversified Multiplet Upcycling*. Conference on Empirical Methods in Natural Language Processing (EMNLP2025)
- Tianle Li, **Jihai Zhang**, Yongming Rao, Yu Cheng. *Unveiling the Compositional Ability Gap in Vision-Language Reasoning Model*. Conference on Neural Information Processing Systems (NeurIPS 2025)
- Jiashuo Sun, **Jihai Zhang**, Yucheng Zhou, et al. *SURf: Teaching Large Vision-Language Models to Selectively Utilize Retrieved Information*. Conference on Empirical Methods in Natural Language Processing (EMNLP2024)
- **Jihai Zhang**, Xiang Lan, Xiaoye Qu, et al. *Learning the Unlearned: Mitigating Feature Suppression in Contrastive Learning.* European Conference on Computer Vision (ECCV2024)

PROJECT

- Matrix Game 3.5 https://matrix-game-v3-5.github.io