Shihao Cheng (程世豪)

Phone: +86 19819662573 | Email: shihaocheng@whu.edu.cn | WeChat: wx19819662573
Address: Faculty of Information Science, Wuhan University, Wuhan 430072
My homepage: https://shihao-cheng.github.io

I am a Master's student at **Wuhan University**, advised by Prof. Zhigang Tu. China. Expected graduation: June 2027. Prior to this, I obtained my B.S. degree from **Harbin Institute of Technology** in 2024, ranking **first** in my class.
Currently, I am a Research Intern at **Tencent Hunyuan**, focusing on agentic streaming video generation and world modeling via closed-loop feedback. Before this, I interned at TeleAI, working on **streaming audio-video joint generation.**
My research interests lie in multimodal understanding and generation, with a particular focus on solving complex alignment and interaction challenges in generative models. I have published works to **CVPR (Highlight)**, **ECCV**, and **T-CSVT**.

EDUCATION

Wuhan University — Wuhan, Hubei
M.S. Student. Communication and Information Systems — Sep. 2024 – Now

Harbin Institute of Technology — Harbin, Heilongjiang
Bachelor of Information Engineering — Sep. 2020 – Jun. 2024

- Avg Score: 92.3/100, Rank: 1/29.
- **National Scholarship 2023. ¥8,000 RMB, Top 1%.** Provincial Outstanding Graduate, 2024, Top 1%.

SELECTED PUBLICATIONS

Unison: Harmonizing Motion, Speech, and Sound for Human-Centric Audio-Video Generation
European Conference on Computer Vision ***(ECCV 2026)*** | ***First Author***

- Resolved speech-SFX interference and motion-audio desynchronization in audio-video generation.

InteractiveAvatar: Real-Time Streaming Video Generation for Consistent and Intent-Aware Avatars
European Conference on Computer Vision ***(ECCV 2026)***

- Resolved the understanding interaction and long-term semantic drift in infinite-length video generation.

GeoMMBench and GeoMMAgent: Toward Expert-Level Multimodal Intelligence in Geoscience and Remote Sensing
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition ***(CVPR 2026,Highlight)*** | ***Co-First Author***

- Resolved the lack of expert-level evaluation benchmarks for interdisciplinary and multi-sensor scenarios, and built a multi-agent architecture to address this issue.

OwlSight: A Robust Illumination Adaptation Framework for Dark Video Human Action Recognition
IEEE Transactions on Circuits and Systems for Video Technology, 2025 ***(T-CSVT, IF: 11.1)*** | ***First Author***

- Resolved inefficient brightness utilization in low-light enhancement and action recognition end-to-end training.

EXPERIENCE

Research Intern, Tencent Hunyuan — Apr. 2026 – Now

- I participated in the research of agentic video world modeling to synergize reasoning and generation, focusing on streaming audio-visual generation with semantic-temporal alignment via Hierarchical World State Memory.

Research Intern, TeleAI — Aug. 2025 – Apr. 2026

- Advised by: Dr. Shansong Liu
- Participated in research on audio-video synchronous and streaming video generation models, contributing to the development of a real-time, human-centric video generation system with audio-visual synchronization, which has been successfully launched on **Telestudio**.

Research Project (SenseTime): Audio-Visual Semantic and Temporal Alignment in MLLMs — Mar. 2025 – Jun. 2025

- Advised by: Dr. Jiahao Wang
- Investigated Multimodal Large Language Models (MLLMs) for joint audio-visual understanding, focusing on resolving complex semantic and temporal alignment challenges. Designed and optimized multimodal fusion modules to enhance cross-modal representation learning and temporal synchronization.