XIANGYUE ZHANG

+86 1887-4623-697
zxyHAVPR@gmail.com
www.xiangyuezhang.com

EDUCATION

M.Sc. student in Computer Applications | ***Advised by: Prof.Zhigang Tu*** Sep. 2023 – Now
HAVPR Lab, Wuhan University

B.Eng. in Geomatics | ***Minor in Data Science and Big Data Technology*** Sep. 2019 – Jun. 2023
Central South University **Major GPA**:89.95/100 **CET(4/6)**: 586/577

RESEARCH: ALL ARE FIRST AUTHOR

SemTalk: Holistic Co-speech Motion Generation with Frame-level Semantic Emphasis

First Author, ICCV 2025 (accepted)

- We propose SemTalk, a novel framework for holistic co-speech motion generation that separately models general rhythm-related base motion and semantic-aware sparse motion, adaptively integrating them based on a learned semantic score.
- We introduce *rhythmic consistency learning* to incorporate latent face and latent hands features with rhythm-related priors, ensuring coherent and rhythm-related base motion. We then propose *semantic emphasis learning* to synthesize semantic gestures at certain frames, enhancing expressive semantic-aware sparse motion.
- Experimental results show that our model surpasses state-of-the-art methods qualitatively and quantitatively, achieving higher motion quality and richer semantics.

EchoMask: Speech-Queried Attention-based Mask Modeling for Holistic Co-Speech Motion Generation

First Author, ACM MM2025(accepted)

- We propose EchoMask for co-speech motion generation, a novel masked motion modeling framework that utilizes motion-aligned speech features to mask semantically important motion frames.
- We introduce MAM, a hierarchical cross-modal alignment module that embeds motion and audio into a shared latent space, producing motion-aligned speech features. We also propose SQA, a speech-queried attention mechanism that computes frame-level attention scores, enabling selective identification of semantically important motion frames.
- Extensive experiments demonstrate the superiority of EchoMask over state-of-the-art methods in terms of semantic alignment, motion quality, and generation diversity.

Robust 2D Skeleton Action Recognition via Decoupling and Distilling 3D Latent Features

First Author, TCSVT 2025 (accepted) IEEE Transactions on Circuits and Systems for Video Technology, 2025

- A new 2D skeleton action recognition paradigm, named $2D^3$, is proposed for decoupling and distilling latent pose and view features with the assistance of 3D skeletons, enhancing the robustness of the 2D skeleton models.
- A 2D-to-3D supervision strategy is designed for explicitly decoupling the pose and view features in 3D latent space using 2D skeleton inputs.
- Two cross-attention modules are utilized to distill discriminative motion features while considering the uncertainties of viewpoint and depth.

Rethinking Diffusion for Holistic Co-speech Motion Generation

First Author

WORK EXPERIENCE

Research Intern | ***Advised by: Prof.Liefeng Bo*** June 2024 – now
Alibaba Tongyi Lab Hangzhou, China

- During the internship, I participated in the code optimization and research of co-speech 3D motion generation. At the same time, investigated 2D work such as echomimic, hallo, liveportait, etc.
- Research Interest: Co-speech Motion Generation; 3D & 2D Motion Generation;