

Bing Zhou

 homepage  zhoubinwy@gmail.com  +1 631-7216031  google scholar

SUMMARY

Staff Research Engineer at Snap Research focused on **Generative Audio, Video, Multimodal LLM (MLLM)** modeling. Expertise in large-scale audio pretraining, audio-video joint generation, human motion generation, and multimodal conditioning (speech, text, vision) for generative systems. Successfully delivered **Snap's first Audio-Video joint generation model** to product team and launched **Snap's first Text-to-Motion model**; leading high-fidelity generative audio models for conversational AI and video synthesis. Built streaming talking-video generation with an **Multimodal LLM** backbone, fusing speech audio, transcripts, and human imagery for low-latency audiovisual output.

EMPLOYMENT

- **Snap Research** New York, NY
Staff Research Engineer Oct 2024 - Present
Senior Research Engineer Oct 2021 - Oct 2024
- **IBM Research** Yorktown Heights, NY
Research Staff Member May 2019 - Oct 2021
Research Intern May 2018 - Aug 2018

EDUCATION

- **Stony Brook University** Stony Brook, NY
Ph.D. in Electrical and Computer Engineering Aug 2014 – May 2019
- **University of Chinese Academy of Sciences** Beijing, China
M.S. in Electronic and Communication Engineering Sept 2011 – May 2014
- **University of Science and Technology of China** Hefei, China
B.S. in Applied Physics (School of the Gifted Young) Sept 2007 – May 2011

RESEARCH EXPERIENCE

- **Audio Pretraining & Audio-Video Joint Generation (Research & Engineering)**
 - *Large-Scale Audio Pretraining* Leading development of large-scale generative audio models and speech language models using efficient diffusion DiT architectures. Architected data processing pipeline to ingest and filter hundreds of millions of raw video/audio clips, optimizing for diversity and high-fidelity reconstruction.
 - *Joint Video-Audio Generation.* Designed novel architectures for joint video-audio training and generation, achieving synchronized audio-visual outputs at scale. Scaled training datasets to hundreds of millions of clips through distributed data pipelines. Delivered the first audio-video joint generation model to product team.
 - *Multimodal LLM (MLLM) & Streaming Talking Video.* Developed a streaming audio-video generation model built on an LLM backbone, jointly conditioning on **speech samples**, **speech transcripts**, and a **reference human image** to synthesize synchronized talking videos in a streaming manner with temporally coherent multimodal fusion.
 - *Audio-Driven Video Generation.* Project lead for TalkVerse, a large-scale open-source corpus (2.3M clips, 6.3k hours) and 5B-parameter DiT baseline for minute-long audio-driven talking video generation. Focused on robust lip-sync, long-horizon consistency, and zero-shot generalization.
- **Multimodal 3D Animation Synthesis**
 - *Text-to-Animation.* Led and delivered Snap's first text-to-animation model for Bitmoji Avatars from scratch, enabling natural language-driven 3D character animation for production use.
 - *Audio-Driven 3D Talking Head.* Led the project CapTalk, a text-guided speech-driven 3D head animation framework with separate control over speaking style and emotion. Enables real-time generation of highly synchronized lip movements and dynamic facial expressions from audio and textual descriptions.
 - *Music-to-Dance Generation.* Developed Dancimation, a real-time music-to-dance prototype and DuetGen (SIGGRAPH'25), leveraging diffusion models and hierarchical masked modeling for realistic text-to-motion and music-to-dance synthesis.
- **Human Computer Interactions**

- **Human Egocentric Sensing.** Developed MI-Poser (UbiComp'23), a real-time full-body pose tracking system using magnetic and inertial sensor fusion with metal interference mitigation. Enables egocentric self-representation and human-human interaction modeling in AR/VR through on-body wearables.
- **Acoustic-Visual Multimodal Perception.** Developed EchoPrint (MobiCom'18) and AO-Finger (CHI'23), leveraging acoustic signals for secure authentication and fine-grained gesture recognition, bridging signal processing with deep learning.

SELECTED PUBLICATIONS AND PREPRINTS

* indicates that I served as the **Project Lead, Corresponding Author, Primary Mentor, or First Author.**

- RigMo: Unifying Rig and Motion Learning for Generative Animation CVPR'26
Hao Zhang, Jiahao Luo, Bohui Wan, Yizhou Zhao, Zongrui Li, Michael Vasilkovsky, Chaoyang Wang, Jian Wang, Narendra Ahuja, **Bing Zhou***
- HandX+: Scaling Up Text-Conditioned Bimanual Motion Generation CVPR'26
Zimu Zhang, Yucheng Zhang, Xiyan Xu, Ziyin Wang, Sirui Xu, Kai Zhou, **Bing Zhou**, Chuan Guo, Jian Wang, Yu-Xiong Wang, Liangyan Gui
- Unleashing Guidance Without Classifiers for Human-Object Interaction Animation ICLR'26
Ziyin Wang, Sirui Xu, Chuan Guo, **Bing Zhou**, Jiangshan Gong, Jian Wang, Yu-Xiong Wang, Liangyan Gui
- Text2Interact: High-Fidelity and Diverse Text-to-Two-Person Interaction Generation ICLR'26
Qingxuan Wu, Zhiyang Dou, Chuan Guo, Yiming Huang, Qiao Feng, **Bing Zhou**, Jian Wang, Lingjie Liu
- TalkVerse: Democratizing Minute-Long Audio-Driven Video Generation CVPR'26
Zhenzhi Wang, Jian Wang, Ke Ma, Dahua Lin, **Bing Zhou***
- Animated 3DGS Avatars in Diverse Scenes with Consistent Lighting and Shadows under review
Aymen Mir, Riza Alp Guler, Jian Wang, Gerard Pons-Moll, **Bing Zhou***
- AHA! Animating Human Avatars in Diverse Scenes with Gaussian Splatting under review
Aymen Mir, Jian Wang, Riza Alp Guler, Chuan Guo, Gerard Pons-Moll, **Bing Zhou***
- CapTalk: Text-Guided Stylization and Speech-Driven 3D Head Animation under review
Xuanguang Chu, Yuan Gan, Ziteng Cui, Shuhong Liu, Jian Wang, **Bing Zhou***, Tatsuya Harada
- SceneMI: Motion In-betweening for Modeling Human-Scene Interaction ICCV'25 (Highlight)
Inwoo Hwang, **Bing Zhou***, Young Min Kim, Jian Wang, Chuan Guo
- Ponimator: Unfolding Interactive Pose for Versatile Human-Human Interaction Animation ICCV'25
Shaowei Liu, Chuan Guo, **Bing Zhou***, Jian Wang
- DuetGen: Music Driven Two-Person Dance Generation via Hierarchical Masked Modeling SIGGRAPH'25
Anindita Ghosh, **Bing Zhou***, Rishabh Dabral, Jian Wang, Vladislav Golyanik, Christian Theobalt, Philipp Slusallek, Chuan Guo
- SnapMoGen: Human Motion Generation from Expressive Texts NeurIPS'25
Chuan Guo, Inwoo Hwang, Jian Wang, **Bing Zhou***
- Perspective-Aligned AR Mirror with Under-Display Camera SIGGRAPH Asia'24 (Journal, Best Paper Award)
Jian Wang, Sizhuo Ma, Karl Bayer, Yi Zhang, Peihao Wang, **Bing Zhou**, Shree Nayar, Gurunandan Krishnan
- MI-Poser: Human Body Pose Tracking using Magnetic and Inertial Sensor Fusion IMWUT/UbiComp'23
Riku Arakawa, **Bing Zhou***, Gurunandan Krishnan, Mayank Goel, and Shree Nayar
- AO-Finger: Hands-free Fine-grained Finger Gesture Recognition via Acoustic-Optic Sensor Fusing CHI'23
Chenhan Xu, **Bing Zhou***, Gurunandan Krishnan and Shree Nayar
- Fine-grained visual recognition in mobile augmented reality for technical support ISMAR'20
Bing Zhou*, Sinem Guven Kaya

TECHNICAL SKILLS

- **Generative Audio** Audio VAEs, Speech language models, Diffusion models (DiT), Large-scale audio pretraining
- **Distributed Training & Infrastructure** Hundreds of GPUs distributed training, PyTorch, Kubernetes, Ray, BigQuery, Large-scale data pipeline engineering
- **Multimodal LLMs (MLLM)** LLM-backbone generative stacks, Cross-modal fusion (speech, text, image), Streaming audiovisual decoding, Instruction and transcript conditioning
- **Audio-Video Modeling** Joint video-audio generation, Mixed audio-text language models, Speech and ambient sound synthesis, Streaming talking-video synthesis, Audio-driven animation, 3D avatar synthesis
- **Signal Processing & Sensing** Real-time signal processing, Acoustic sensing, Sensor fusion (IMU/magnetic), Multimodal perception