Jun Zhan (詹俊)

jzhan24@m.fudan.edu.cn · +86 198 2167 0042 · Homepage

Education

Fudan University, Shanghai, China — Sep 2024 – Jun 2027 (Expected)
Ph.D. Student in Artificial Intelligence, Advisor: Prof. Xipeng Qiu

Fudan University, Shanghai, China — Sep 2022 – Jun 2024
M.Eng. in Electronic Information, Advisor: Prof. Xipeng Qiu

Huazhong University of Science and Technology, Wuhan, China — Sep 2018 – Jun 2022
B.Eng. in Software Engineering

Research Interests

Multimodal Interaction, Omni-modal Foundation Models

Publications

- **OmniVAE: An Audio-Video VAE with Cross-Modal Alignment for Joint Generation**
  Jun Zhan*, Chen Yang*, Yitian Gong*, et al.
  arXiv preprint, Project Lead, Co-first Author [Paper] [Code]
- **AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling**
  Jun Zhan*, Junqi Dai*, Jiasheng Ye*, Yunhua Zhou*, et al.
  ACL 2024, Project Lead, Co-first Author [Paper] [Code]
- **VStyle: A Benchmark for Voice Style Adaptation with Spoken Instructions**
  Jun Zhan*, Mingyang Han*, Yuxuan Xie*, et al.
  ICASSP 2026, Project Lead, Co-first Author [Paper] [Code]
- **SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities**
  Dong Zhang, Shimin Li, Xin Zhang, **Jun Zhan**, Pengyu Wang, Yaqian Zhou, Xipeng Qiu
  EMNLP 2023 Findings [Paper] [Code]

* *Equal contribution*

Research/Project Experience

OmniVAE — Dec 2025 – May 2026
Project Lead

- Proposed a jointly trained audio-video VAE, improving the learnability of both unimodal and cross-modal representations through unimodal semantic distillation and segment-level audio-video contrastive learning
- Led the development of the audio-video training data pipeline, defining an audio-video concept taxonomy, automatically generating retrieval queries, collecting videos, and filtering data for quality and semantic consistency
- Improved downstream text-to-audio-video generation quality and cross-modal alignment while largely preserving audio and video reconstruction quality

AnyGPT — Sep 2023 – Feb 2024
Project Lead

- Built an end-to-end omni-modal model with discrete representations, unifying the understanding and generation of images, speech, music, and text
- Constructed and released the first large-scale Any-to-Any multimodal instruction dataset, enabling omni-modal dialogue with arbitrary input and output modalities

- Open-sourced model weights, code, and data, receiving 800+ GitHub stars, 300+ citations, and over 80,000 model and dataset downloads

SpeechGPT Series Apr 2023 – Jan 2025

Group Lead

- Contributed to three SpeechGPT models: SpeechGPT pioneered intrinsic speech dialogue in LLMs; SpeechGPT-Gen enabled controllable timbre generation; SpeechGPT 2.0-preview introduced end-to-end audio modeling for multi-emotion, multi-style, and multi-timbre dialogue
- Responsible for TTS model training, supporting laughter, emphasis, pauses, instruction-based style control, and multi-speaker dialogue
- Responsible for speech dialogue data construction
- Open-sourced models and code, receiving 1.7k+ GitHub stars and 800+ citations across the SpeechGPT series

2nd Guangdong-Hong Kong-Macao Greater Bay Area International Algorithm Competition (RMB 1 Million Track) Oct 2023 – Apr 2024

Problem Setting Lead

- Designed the "Comprehensive Enhancement of Multimodal LLM Subject Capabilities" task, focusing on multimodal subject knowledge understanding and reasoning
- Built and released a LLaVA-like baseline based on MOSS2, providing training and evaluation references for participants
- Curated nearly a decade of college entrance examination practice questions to construct a multidisciplinary multimodal QA dataset and evaluation benchmark

MOSS Jan 2023 – Mar 2023

Speech Module Member

- Contributed to MOSS, the first open-source ChatGPT-like large language model in China
- Responsible for the text-to-speech module, supporting zero-shot voice cloning and extending the model's speech generation capabilities

Internship & Work Experience

Meta AI (FAIR) Jun 2026 – Present

Research Intern Paris, France

- Developing an end-to-end multimodal interaction model that jointly models speech, motion, and facial expressions to improve the naturalness and cross-modal consistency of virtual character interactions
- Developing a multimodal VAE for streaming body, hand, and facial motion encoding and decoding, and optimizing the training of the interaction model

OPPO Sep 2025 – Nov 2025

Industry-Academia Project Intern Beijing, China

- Curated and filtered pretraining data, constructed instruction data, and built the training framework using OPPO's proprietary speech data
- Trained an end-to-end speech dialogue model supporting multi-style, multi-emotion, and multi-timbre speech generation

Alibaba Group, Future Living Lab May 2025 – Aug 2025

Research Intern Beijing, China

- Trained an end-to-end speech dialogue model, improving its responsiveness to natural-language style instructions and implicit emotional expressions
- Built VStyle, a voice style adaptation benchmark covering acoustic attributes, natural language instructions, role-playing, and implicit empathy, with automated evaluation based on large audio language models