詹俊

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

教育背景

复旦大学, 上海 2024.09 – 2027.06（预计）

博士研究生 人工智能，导师：邱锡鹏教授

复旦大学, 上海 2022.09 – 2024.06

硕士 电子信息，导师：邱锡鹏教授

华中科技大学, 武汉 2018.09 – 2022.06

学士 软件工程

研究兴趣

多模态交互，全模态基础模型，具身交互（语音-表情-动作联合建模）

能力总结

- **研究主线**：从语音到全模态、音视频，再到 3D 人体动作，专注端到端的多模态生成与交互
- **技术积累**：熟悉多模态表征（Codec / VAE）、离散 token 与扩散两类生成范式及全双工流式建模
- **数据与评测**：擅长大规模多模态数据构建与自动化评测，构建过多个数据集与评测基准
- **全流程能力**：可独立完成从选题、数据、模型训练、评测到论文与开源发布的完整研究流程

论文

- **OmniVAE: An Audio-Video VAE with Cross-Modal Alignment for Joint Generation**
  Jun Zhan*, Chen Yang*, Yitian Gong*, et al.
  arXiv preprint，项目负责人，共同一作 [Paper] [Code]
- **AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling**
  Jun Zhan*, Junqi Dai*, Jiasheng Ye*, Yunhua Zhou*, et al.
  ACL 2024，项目负责人，共同一作 [Paper] [Code]
- **VStyle: A Benchmark for Voice Style Adaptation with Spoken Instructions**
  Jun Zhan*, Mingyang Han*, Yuxuan Xie*, et al.
  ICASSP 2026，项目负责人，共同一作 [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]
- **其他参与工作**（GitHub Stars）：YuE **12.8k**、MOSS **12.3k**、MOSS-TTSD **1.4k**、MOSS-Audio **690+**

* 表示共同一作

实习 & 工作经历

Meta AI (FAIR) 2026 年 6 月 – 至今

研究型实习生 法国巴黎

- 基于 7B 全双工语音大模型与扩散模型的混合架构，构建**能听、能看、能说、能动**的单一流式模型，联合生成语音、表情、身体与双手动作，**RTF 0.8**
- 研发身体（SMPL）、双手（MANO）与面部表情的**流式动作 VAE**，双人相对距离误差仅 **1.7cm**，困难动作重建误差仅 **11.3mm**
- 在指令动作合成任务上引入 **CoT 机制**，思考时延仅 **0.32s**；在 HumanML3D 上 R-Precision **超过真实动作**，实现 **SOTA** 的流式动作合成性能
- 构建约 **3.1 万小时**“语音-表情-动作”全双工交互数据，涵盖真实双人交互、肢体与表情理解、指令控制的肢体动作与表情合成

OPPO 2025 年 9 月 – 2025 年 11 月

校企合作实习生 中国北京

- 基于 OPPO 自有语音数据，负责预训练数据清洗与过滤、指令数据构建及训练框架构建
- 训练端到端语音对话模型，支持多风格、多情感和多音色语音生成

阿里巴巴未来生活实验室 2025 年 5 月 – 2025 年 8 月

研究型实习生 中国北京

- 训练端到端语音对话模型，提升模型对自然语言风格指令和隐式情感表达的响应能力
- 构建语音风格自适应评测基准 VStyle，覆盖声学属性、自然语言指令、角色扮演和隐式共情，并基于大型音频语言模型实现自动化评测

研究/项目经历

OmniVAE 2025 年 12 月 – 2026 年 5 月

项目负责人

- 提出联合训练的音视频 VAE，通过单模态语义蒸馏和片段级音视频对比学习，提升单模态与跨模态表征的可学习性；开源后获得 GitHub Stars **100+**
- 主导音视频训练数据管线建设，定义音视频概念体系，自动生成检索词并采集视频，完成数据质量与语义一致性过滤
- 在基本保持音频与视频重建质量的同时，**明显提升下游文本到音视频生成质量与跨模态对齐水平**

AnyGPT 2023 年 9 月 – 2024 年 2 月

项目负责人

- 基于离散表示构建端到端全模态模型，统一建模图像、语音、音乐和文本的理解与生成
- 构建并发布**首个**大规模 Any-to-Any 多模态指令数据集，实现任意模态输入输出的全模态对话
- 开源模型权重、代码、数据，GitHub Stars **800+**，引用数 **350+**，模型和数据下载次数**八万余次**

SpeechGPT 系列 2023 年 4 月 – 2025 年 1 月

组长

- 参与 SpeechGPT 系列模型研发：SpeechGPT**首次**实现大语言模型的内生语音对话，SpeechGPT-Gen实现音色可控的语音生成，SpeechGPT 2.0-preview实现端到端音频建模及多情感、多风格、多音色对话
- 负责 TTS 模型训练，支持笑声、重读、停顿、指令风格控制和多人对话
- 负责语音对话数据的构建
- 开源模型与代码累计获得 GitHub Stars **1.7k+**，系列工作累计引用 **900+**

第二届粤港澳大湾区（黄埔）国际算法算例大赛（百万奖金赛题） 2023 年 10 月 – 2024 年 4 月

出题负责人

- 设计“多模态大模型学科能力综合强化”赛题，收集并处理近十年高考模拟题，构建覆盖多学科的多模态问答数据集与评测基准
- 基于 MOSS2 构建并发布 LLaVA-like 基线模型，为参赛队伍提供训练与评测基准

MOSS 2023 年 1 月 – 2023 年 3 月

语音模块成员

- 参与研发**国内首个**类 ChatGPT 开源大语言模型 MOSS，开源后获得 **12.3k+** GitHub Stars
- 负责语音合成模块，支持零样本音色克隆，扩展模型的语音生成能力