Zhishuo Zhao

Ph.D. Candidate in Computer Science, Sichuan University
Email: zhaozhishuo@stu.scu.edu.cn | GitHub: zhiishuo.github.io | Google Scholar: link

Research Interests

Multimodal Learning; Audio–Visual Speech Recognition; Emotion Understanding and Affective Computing; Multimodal Large Language Model (PEFT & CoT Reasoning)

Education

Ph.D. Candidate in Computer Science, Sichuan University — 2023 – Present
Department of Computer Science — Direct Ph.D. Program — *GPA: 3.79/4.00*
Advisor: Prof. Yi Lin

B.Eng. in Software Engineering, Sichuan University — 2017 – 2021
Wu Yuzhang Honors College, Department of Computer Science — *GPA: 3.18/4.00*

Publications

1. **Zhao, Z.**, Yi, L., Guo, D., and Fan, J.
   AV-RISE: Hierarchical Cross-Modal Denoising for Learning Robust Audio-Visual Speech Representation.
   Proceedings of the ACM International Conference on Multimedia (ACM MM), Dublin, Ireland, 2025. **(Oral)**
2. **Zhao, Z.**, Guo, D., Ou, W., *et al.*
   AMG-AVSR: Adaptive Modality Guidance for Audio-Visual Speech Recognition via Progressive Feature Enhancement.
   The 16th Asian Conference on Machine Learning (ACML, Conference Track), Hanoi, Vietnam, 2024. **(Oral)**
3. Ou, W., **Zhao, Z.**, Guo, D., *et al.*
   WinNet: Make Only One Convolutional Layer Effective for Time Series Forecasting.
   International Conference on Intelligent Computing (ICIC), Springer Nature, Singapore, 2024. **(Oral)**

Under Review / In Submission

4. **Zhao, Z.**, Yi, L., and Guo, D.
   Unity in Diversity: Dual-Branch Disentanglement of Consensus and Diversity for Multimodal Sentiment Analysis.
   IEEE Transactions on Multimedia (TMM), under review, 2026.
5. **Zhao, Z.**, Yi, L., and Guo, D.
   HEME: Hierarchical Emotion Modeling with Adaptive Multi-Level Mixture-of-Experts.
   IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), under review, 2026.
6. Ou, W., **Zhao, Z.**, Guo, D., *et al.*
   Logo-LLM: Local and Global Modeling with Large Language Models for Time Series Forecasting.
   IEEE Transactions on Neural Networks and Learning System (TNNLS), under review, 2026.
7. Hu, J., **Zhao, Z.**, Liao, C., Peng, R., Fan, J., and Lin, Y.
   Two-Stage Prototypical Prompting Framework for Voice Pathology Detection.
   IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), under review, 2026.

Patents

1. Yi Lin, **Zhishuo Zhao**, Dongji Yan, and Qinghai Zuo.
   Speech Recognition Method and Device with Progressive Audio–Visual Self-Tuning under Noisy Environments.
   China National Intellectual Property Administration (CNIPA), Patent No. CN119107945B, granted on Apr. 25, 2025.

2. Yi Lin, **Zhishuo Zhao**, Dongyue Guo, and Lei Bian.
   Audio–Visual Speech Recognition Method and Device Based on Multi-Layer Perceptual Fusion.
   China National Intellectual Property Administration (CNIPA), Patent No. 202510925715.6, filed on Jul. 6, 2025.

3. Yi Lin, **Zhishuo Zhao**, Tingting Zhang and Dongyue Guo.
   Emotion Recognition Method and Apparatus Based on Multimodal Consensus and Diversity Decoupling.
   China National Intellectual Property Administration (CNIPA), Patent No. 2025117096442, filed on Nov. 20, 2025.

Research Experience

Research Member, Intelligent Perception and Multimodal Learning Group, Sichuan University 2023 – Present
Advisor: Prof. Yi Lin
Engaged in research on *multimodal learning, audiovisual speech recognition, and air traffic control (ATC) situational awareness*, under several **National Natural Science Foundation of China (NSFC)** projects. Focused on multimodal representation learning, dataset construction, and model optimization for complex perceptual environments.

Representative Projects and Contributions:

- **NSFC Project (62371323)**: *Task-Driven Audio-Visual Speech Understanding in Complex ATC Environments.* Developed robust cross-modal alignment and denoising models for audiovisual speech recognition, resulting in first-author publications in *ACM MM 2025 (oral)* and *ACML 2024 (oral)*. Contributed to patent CN119107945B (granted, 2025).
- **NSFC Project (U20A20161)**: *Machine-Learning-Based ATC Route Prediction and Control under Complex Environments.*
  Designed time-series trajectory prediction and safety evaluation models to enhance air traffic situational awareness. Co-authored *WinNet (ICIC 2024)* for temporal forecasting, improving trajectory accuracy and response capability.
- **NSFC Project (U2333209)**: *Remote Platform Multimodal Data Acquisition and Management for ATC Scenarios.* Constructed large-scale multimodal ATC datasets covering perception understanding and audiovisual speech tasks, including over **1,127** minutes of remote tower video and **6,500** minutes of radar data. Assisted in system software registration (Audio-Visual Data Synchronization Platform V1.0, 2023SR1549420).

Selected Presentations

2025.10 **ACM Multimedia (ACM MM)**, Dublin, Ireland —Oral Presentation: *AV-RISE: Hierarchical Cross-Modal Denoising for Robust Audio-Visual Speech Representation*.

2024.12 **Asian Conference on Machine Learning (ACML)**, Hanoi, Vietnam —Oral Presentation: *AMG-AVSR: Adaptive Modality Guidance for Audio-Visual Speech Recognition via Progressive Feature Enhancement*.

Selected Honors & Awards

2024 – 2025 **National Scholarship**, Sichuan University
China's highest postgraduate honor recognizing outstanding academic excellence, innovation, and research.

2023 – 2024 **BYD Scholarship**, Sichuan University
Merit-based postgraduate award granted to 20 outstanding students university-wide.

2023 – 2025 **Excellent Graduate Student**, Sichuan University
Awarded twice consecutively for exceptional research and academic achievements.

Industry Experience

Backend Engineer, NetEase Inc., Guangzhou, China 2021 – 2023
Developed scalable backend microservices and data pipelines for large-scale user platforms, applying distributed system design and database optimization to improve throughput and fault tolerance.

Technical Skills

– **Machine Learning / AI:** PyTorch, Transformers, PEFT (LoRA/DoRA), Q-Former, Weights & Biases; experience in large-scale multimodal model training, distributed optimization, and prototype-based contrastive learning.

– **Audio – Visual Processing:** Librosa, Torchaudio, OpenFace, FFmpeg; proficient in audiovisual synchronization, noise augmentation, lip reading, and multimodal signal alignment for robust AVSR systems.

– **Vision & NLP:** OpenCV, tokenization, multimodal fusion, prompt engineering, feature disentanglement; skilled in integrating perception encoders with language models for visual reasoning and emotion understanding.

– **Systems & Tools:** Python, CUDA, Linux, Git, Docker, LaTeX; strong experience in HPC environments, GPU resource scheduling, and reproducible experiment management.