Yaoting Wang

Phone： +86 156-1553-8152 | Email： yaoting.wang@outlook.com
Website： https://yaotingwangofficial.github.io

EDUCATION

Fudan University 2025.09 - Present
Artificial Intelligence | Doctor Shanghai

University of Edinburgh 2021.09 - 2022.12
Speech and Language Processing | Master
- Distinction MSc dissertation
- Merit-class degree

National University of Limerick, Ireland 2019.07 - 2021.07
Computer Systems | Bachelor
- First-Class Honours with GPA 3.82 / 4.00
- Full Awarded Scholarship (2020)
- Half Awarded scholarship (2019)

Shandong University of Technology 2017.09 - 2021.06
Computer Science | Bachelor
- First-class scholarship (2017, 2018)

PROFESSIONAL EXPERIENCE

MUCG, ACM MM'25 2025.07 - 2025.10
Program Chair
- Co-organizer and PC of the 1st Inter. Workshop on MLLM for Unified Comprehension and Generation (MUCG) at ACMMM'25.

Tsinghua University 2025.03 - 2025.09
Research Intern, Institute for AI Industry Research (AIR) Beijing
- Advised by Prof. Yunxin Liu for multimodal LLMs research.

King Abdullah University of Science and Technology 2024.03 - Present
Visiting Student Research Program Vision-CAIR
- Advised by Prof. Mohamed h. Elhoseiny for vision-language research.

Renmin University of China 2023.01 - Present
Research Assistant Intern, GeWu-Lab Beijing
- Assist in research related to multimodal scene understanding: language, vision and audio.

Deepwise 2023.01 - 2023.03
NLP Intern, Deepwise AI Lab Beijing
- Responsible for the research and development of multimodal pneumonia classification using medical text and images on real patient data from the National Institutes for Food and Drug Control.

RESEARCH EXPERIENCE

AVI-Bench: Toward Human-like Audio-Visual Intelligence of Omni-MLLMs 2025.03 - 2025.08

- We present AVI-Bench, a cognitively inspired benchmark spanning perception, understanding, and reasoning, along with AVI-Bench-PriSe for evaluating generalization to low-semantic inputs.
- Extensive evaluations on 28 both open- and closed-source Omni-MLLMs reveal key challenges in achieving robust and general audio-visual intelligence.
- We propose a four-level taxonomy to systematically characterize the audio-visual intelligence of Omni-MLLMs: Task adaptive, Modality adaptive, Stage adaptive, Domain adaptive.

Multimodal Chain-of-Thought Reasoning: A Comprehensive Survey 2024.12 - 2025.04

- First Survey: This paper represents the first survey dedicated to an inaugural thorough review of MCoT reasoning.
- Comprehensive Taxonomy: We propose a meticulous taxonomy that categorizes the diverse approaches in MCoT research.
- Frontiers and Future Directions: We discuss emerging challenges and outline promising avenues for future research.
- Awesome-MCoT (700+ stars)

MiniGPT-VSpeech-R1: Advancing Robust Speech Understanding in Audio-Visual Scenes 2024.04 - 2025.04

- We propose a novel audio-visual speech processing task: Audio-Visual Speech Understanding
- We build the first multimodal LLM for AVSU with the structured reasoning pattern (VSpeech-CoT) and reinforcement learning.
- We build the first dataset AVSU-Bench with 50k audio-visual speech question-answer pairs for training and accessing the robustness of audio-visual speech models.

Ref-AVS: Refer and Segment Objects in Audio-Visual Scenes with Natural Language 2023.12 - 2024.03

- We propose Ref-AVS as a challenging scene understanding task that segments objects of interest with multimodal-cue natural language expressions, and provide the corresponding Ref-AVS benchmark for performance training and validation.
- We design an end-to-end framework for Ref-AVS that efficiently processes the multimodal cues with a crossmodal transformer, serving as a feasible research framework for future development.
- Our work can inspire more methods to build better convenience and accessibility for the visually and hearing impaired population.
- *Accepted by ECCV'24 Main Track.*

Can Textual Semantics Mitigate Sounding Object Segmentation Preference? 2023.09 - 2023.11

- We use multimodal LLM for visual scene understanding and obtain potential sound objects as text cues to enhance audio-visual correlation with language as the bridge, providing further precise guidance for segmentation models.
- We designed task-specific few-shot prompt template with CoT to assist LLM reasoning step-by-step and obtain more accurate text cues.
- We propose a Prompting Mask Queries with Semantics module to seamlessly introduce audio and semantic instructions into visual foundation model like Mask2Former.
- *Accepted by ECCV'24 Main Track.*

Prompting Segmentation with Sound is Generalizable Audio-Visual Source Localizer 2023.04 - 2023.08

- Current methods for the Audio-Visual Segmentation task are based on the encoder-fusion-decoder paradigm and fail to address the challenges posed by limited data and varying data distributions as they do not leverage the prior knowledge of pre-trained models effectively.
- We propose our GAVS model follow the encoder-prompt-decoder paradigm. We introduce the Semantic-aware Audio Prompt to assist the visual foundation model in querying sounding objects from the visual space using audio cues.
- We propose the Correlation Adapter, which minimizes effort in adjusting the visual foundation model to establish cross-modal audio-visual correlation.
- Our method outperforms fusion-based methods significantly in both unseen classes and cross-dataset settings.
- *Accepted by AAAI'24 Main Track and ICCV'23 Workshop.*

WORKS

- Multimodal chain-of-thought reasoning: A comprehensive survey.
  > **Yaoting Wang**, Shengqiong Wu, Yuechen Zhang, Shuicheng Yan, Ziwei Liu, Hao Fei
- MiniGPT-VSpeech-R1: Advancing Robust Speech Understanding in Audio-Visual Scenes.
  > **Yaoting Wang**, Jian Ding, Jun Chen, Di Hu, Henghui Ding^, Mohamed Elhoseiny.
- On Path to Multimodal Generalist: General-level and General-bench.
  > Hao Fei*, Yuan Zhou*, Juncheng Li*, Xiangtai Li*, Qingshan Xu*, Bobo Li*, Shengqiong Wu*, **Yaoting Wang**, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Siliang Tang, Kaihang Pan, Yaobo Ye, Haobo Yuan, Tao Zhang, Weiming Wu, Tianjie Ju, Zixiang Meng, Shilin Xu, Liyu Jia, Wentao Hu, Meng Luo, Jiebo Luo, Tat-Seng Chua, Shuicheng Yan, Hanwang Zhang. (2025).
  Accepted by The 42nd International Conference on Machine Learning (ICML 2025 Oral). arxiv.
- AVTrustBench: Assessing Reliability and Robustness in Audio-Visual LLMs.
  > Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, **Yaoting Wang**, Ruohan Gao, Mohamed Elhoseiny, Dinesh Manocha. (2024). *Accepted by ICCV 2025*. arxiv.
- Can Textual Semantics Mitigate Sounding Object Segmentation Preference?.
  > **Yaoting Wang***, Peiwen Sun*, Yuanchao Li, Honggang Zhang, Di Hu^. (2024).
  Accepted by The 18th European Conference on Computer Vision (ECCV 2024). arxiv.
- Ref-AVS: Refer and Segment Objects in Audio-Visual Scenes with Natural Language.
  > **Yaoting Wang***, Peiwen Sun*, Dongzhan Zhou, Guangyao Li, Honggang Zhang, Di Hu^. (2024).
  Accepted by The 18th European Conference on Computer Vision (ECCV 2024). arxiv.
- Stepping Stones: A Progressive Training Strategy for Audio-Visual Semantic Segmentation.
  > Juncheng Ma, Peiwen Sun, **Yaoting Wang**, Di Hu^. (2024).
  Accepted by The 18th European Conference on Computer Vision (ECCV 2024). arxiv.
- Prompting Segmentation with Sound is Generalizable Audio-visual Source Localizer.
  > **Yaoting Wang***, Weisong Liu*, Guangyao Li, Jian Ding, Di Hu^, Xi Li. (2023).
  Accepted by 38th AAAI conference on artificial intelligence (Main track) & ICCV'23 (Workshop). arxiv.
- Incongruity-Aware Hierarchical Crossmodal Transformer with Dynamic Modality Gating: A Study on Affect Recognition.
  > **Yaoting Wang***, Yuanchao Li*, Paul Pu Liang, Louis-Philippe Morency, Peter Bell and Catherine Lai^. (2023).
  in TACL major revision. arxiv.
- Scaling up mobile service selection in edge computing environment with cuckoo optimization algorithm.
  > Ming Zhu, Feilong Yu, Xiukun Yan, Jing Li, **Yaoting Wang**. (2021, September).
  Accepted by 2021 IEEE International Conference on Services Computing (SCC) (pp. 394-400). IEEE.