Yaoting WANG

156-1553-8152 | yaoting.wang@outlook.com | yaotingwangofficial.github.io/

Fudan University, Shanghai, 200433, China

Education

- **Fudan University** *09.2025 - Present*
  Artificial Intelligence | PhD Shanghai, China
  - Advised by Prof. Henghui Ding.
- **University of Edinburgh** *09.2021 - 12.2022*
  Speech and Language Processing | MSc Edinburgh, UK
  - Master's Thesis Awarded Distinction.
- **University of Limerick** *09.2019 - 06.2021*
  Computer Systems | BSc Limerick, Ireland
  - GPA: 3.8/4.0
  - Full Tuition Scholarship
- **Shandong University of Technology** *09.2017 - 06.2021*
  Computer Science and Technology | BEng Zibo, China
  - GPA: 3.8/4.0

Experience

- **MUCG, ACM MM'25 [🌐]** *07.2025 - 10.2025*
  Co-Organizer, Program Chair Remote
  - Co-organizer and PC of the 1st International Workshop on MLLM for Unified Comprehension and Generation (MUCG) at ACM MM'25.
- **AIR, Tsinghua University [🌐]** *03.2025 - 09.2025*
  Research Intern Beijing, China
  - Advised by Prof. Yunxin Liu for multimodal LLMs research.
- **King Abdullah University of Science and Technology [🌐]** *03.2024 - 03.2025*
  Visiting Student Remote
  - Advised by Prof. Mohamed Elhoseiny for multimodal LLMs research.
- **GSAI, Renmin University of China [🌐]** *01.2023 - 03.2024*
  Research Assistant Beijing, China
  - Advised by Prof. Di Hu for multimodal scene understanding.

Projects

- **MCoT Survey** *12.2024 - 04.2025*
  Multimodal Chain-of-Thought Reasoning: A Comprehensive Survey [GH]
  - Conducted the first survey on multimodal CoT reasoning.
  - Proposed a comprehensive taxonomy to classify diverse methodologies and theoretical approaches.
  - Identified open challenges, future research directions and roadmap for advancing the field.
  - Maintained Awesome-MCoT, a widely used open-source resource that supports ongoing research.
- **Ref-AVS** *12.2023 - 03.2024*
  Refer and Segment Objects in Audio-Visual Scenes with Natural Language [GH]
  - Proposed Ref-AVS, a novel scene understanding task that segments objects of interest using multimodal natural language expressions.
  - Built the first Ref-AVS benchmark to enable systematic training and evaluation of models.
  - Designed an end-to-end framework that effectively integrates audio-visual cues with language.
  - Accepted at ECCV'2024.

- **GAVS** *04.2023 - 08.2023*

  Prompting Segmentation with Sound is Generalizable Audio-Visual Source Localizer [GH]

  - Proposed GAVS, an encoder–prompt–decoder framework for Audio-Visual Segmentation, against traditional encoder-fusion-decoder paradigm.
  - Introduced Correlation Adapter to improve crossmodal alignment.
  - Achieved state-of-the-art generalization across unseen classes and datasets.
  - Accepted at AAAI 2024.

PUBLICATIONS

A=ARXIV, C=CONFERENCE, J=JOURNAL

[A.1] **Multimodal Chain-of-Thought Reasoning: A Comprehensive Survey**
Yaoting Wang, Shengqiong Wu, Yuechen Zhang, Shuicheng Yan, Ziwei Liu, Hao Fei
arXiv preprint arXiv:2503.12605 arXiv, 2025

[C.1] **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.
In *International Conference on Machine Learning* (ICML Oral), 2025

[C.2] **AVTrustBench: Assessing Reliability and Robustness in Audio-Visual LLMs.**
Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, **Yaoting Wang**, Ruohan Gao, Mohamed Elhoseiny, Dinesh Manocha.
In *IEEE/CVF International Conference on Computer Vision* (ICCV), 2025

[C.3] **Can Textual Semantics Mitigate Sounding Object Segmentation Preference?**
Yaoting Wang*, Peiwen Sun*, Yuanchao Li, Honggang Zhang, Di Hu.
In *European Conference on Computer Vision* (ECCV), 2024

[C.4] **Ref-AVS: Refer and Segment Objects in Audio-Visual Scenes**
Yaoting Wang*, Peiwen Sun*, Dongzhan Zhou*, Guangyao Li, Honggang Zhang, Di Hu.
In *European Conference on Computer Vision* (ECCV), 2024

[C.5] **Prompting Segmentation with Sound Is Generalizable Audio-Visual Source Localizer**
Yaoting Wang*, Weisong Liu*, Guangyao Li, Jian Ding, Di Hu, Xi Li
In *Association for the Advancement of Artificial Intelligence* (AAAI), 2024