

# Yaoting Wang

+86 156-1553-8152 | yaoting.wang@outlook.com

## EDUCATION

<b>University of Edinburgh</b> MSc Speech and Language Processing <ul style="list-style-type: none"><li>Distinction MSc dissertation</li><li>Merit-class degree</li></ul>	2021.09 - 2022.09
<b>National University of Limerick, Ireland</b> BSc Computer Systems <ul style="list-style-type: none"><li>First-Class Honours with GPA 3.82 / 4.00</li><li>Full Awarded Scholarship (2020)</li><li>Half Awarded scholarship (2019)</li></ul>	2019.07 - 2021.07
<b>Shandong University of Technology</b> BSc Computer Science Bachelor <ul style="list-style-type: none"><li>First-class scholarship (2017, 2018)</li></ul>	2017.09 - 2021.06

## PROFESSIONAL EXPERIENCE

<b>King Abdullah University of Science and Technology</b> Visiting Student Research Program Vision-CAIR <ul style="list-style-type: none"><li>Advised by Prof. Mohamed h. Elhoseiny for vision-language research.</li></ul>	2024.03 - Present
<b>Renmin University of China (RUC)</b> Research Assistant Intern, GeWu-Lab <ul style="list-style-type: none"><li>Assist in research related to multimodal scene understanding: language, vision and audio.</li></ul>	2023.01 - Present Beijing
<b>Deepwise</b> NLP Intern, Deepwise AI Lab <ul style="list-style-type: none"><li>Responsible for the research and development of multimodal tasks, including the classification of pneumonia using medical text (electronic medical records) and medical images (Lung CT scan) on real patient data from the National Institutes for Food and Drug Control.</li></ul>	2023.01 - 2023.03 Beijing

## RESEARCH EXPERIENCE

<b>Emotional Multimodal LLM with Unified Perception and Generation</b> <ul style="list-style-type: none"><li>We aim to build a multimodal, multitask and multilingual LLM with emotion activated.</li><li>Unified perception and generation abilities allowing emotion recognition and emotional speech and talking face generation.</li><li>Advised by Prof. Mohamed Elhoseiny in Vision-CAIR, KAUST.</li></ul>	2024.04 - Present
<b>Ref-AVS: Refer and Segment Objects in Audio-Visual Scenes with Natural Language</b> <ul style="list-style-type: none"><li>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.</li><li>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.</li><li>Our work can inspire more methods to build better convenience and accessibility for the visually and hearing impaired population.</li><li><i>Accepted by ECCV'24 Main Track.</i></li></ul>	2023.12 - 2024.03
<b>Can Textual Semantics Mitigate Sounding Object Segmentation Preference?</b> <ul style="list-style-type: none"><li>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.</li><li>We designed task-specific few-shot prompt template with CoT to assist LLM reasoning step-by-step and obtain more accurate text cues.</li><li>We propose a Prompting Mask Queries with Semantics module to seamlessly introduce audio and semantic instructions into visual foundation model like Mask2Former.</li><li><i>Accepted by ECCV'24 Main Track.</i></li></ul>	2023.09 - 2023.11

## 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.*

## Incongruity-Aware Hierarchical Crossmodal Transformer with Dynamic Modality Gating: A Study on Affect Recognition

2022.08 - 2022.12

- We explore how affective information in different modalities can be influenced by each other, specifically highlighting the presence of latent inter-modal incongruity in crossmodal attention.
- To address this issue, we present the Hierarchical Crossmodal Transformer with Dynamic Modality Gating (HCT-DMG), a lightweight model that effectively reduces fusion times by dynamically choosing the primary modality.
- On five benchmarks: MUSTARD, UR-Funny, CMU-MOSI, CMU-MOSEI, and IEMOCAP, we not only performing better than baseline methods, but also significantly reducing trainable parameters (<1M).
- In TACL major revision.

## Edge Computing

2020.12 - 2021.02

- Assist tutors to expand mobile business selection in edge computing environment with Cuckoo Search algorithm.
- Responsible for obtaining RESTful API information through crawlers.
- Responsible for building a RESTful service intelligent invocation framework using Python.
- *Accepted by the IEEE SCC'21.*

## LEADERSHIP EXPERIENCE

### University Science and Technology Innovation Service Center

2018.05 - 2019.05

Vice Minister

### Science and Technology Service Center; Student Union of Computer Science School

2017.10 - 2018.03

Secretary

## Works

1. Can Textual Semantics Mitigate Sounding Object Segmentation Preference?. **Yaoting Wang\***, Peiwen Sun\*, Yuanchao Li, Honggang Zhang, Di Hu<sup>^</sup>. (2024). Accepted by The 18th European Conference on Computer Vision (ECCV 2024). [arxiv](#).
2. 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<sup>^</sup>. (2024). Accepted by The 18th European Conference on Computer Vision (ECCV 2024). [arxiv](#).
3. Prompting Segmentation with Sound is Generalizable Audio-visual Source Localizer. **Yaoting Wang\***, Weisong Liu\*, Guangyao Li, Jian Ding, Di Hu<sup>^</sup>, Xi Li. (2023). *Accepted by 38th AAAI conference on artificial intelligence (Main track) & ICCV'23 (Workshop)*. [arXiv](#).
4. 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). In 2021 IEEE International Conference on Services Computing (SCC) (pp. 394-400). IEEE.
5. 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<sup>^</sup>. (2023). *in TACL major revision*. [arxiv](#).