

# Zhaokai Wang (王肇凯)

Ph.D. candidate at Shanghai Jiao Tong University

Email: "%s%s@sjtu.edu.cn" % (lastname,firstname)

Wechat: wz\_k\_1015 Homepage: <https://www.wzk.plus> Google Scholar

## EDUCATION

<b>Shanghai Jiao Tong University</b>	09/2022 - 2027 (expected)
Joint Ph.D. program with Shanghai AI Laboratory	Advisor: Prof. Jifeng Dai
<b>Peking University</b>	09/2019 - 06/2022
Double bachelor degree in Economics	
<b>Beihang University</b>	09/2018 - 06/2022
B.Eng in computer science	Overall GPA: 3.80

## EXPERIENCE

<b>OpenGVLab, Shanghai AI Laboratory</b>	Shanghai
Research Intern, collaborated with Prof. Jifeng Dai and Dr. Xizhou Zhu	11/2022 - Present
<b>TuTu. AI, Startup</b>	Shanghai
Co-founder & Research Scientist	06/2023 - 02/2024
<b>Fundamental Vision Group, SenseTime</b>	Beijing
Research Intern, collaborated with Prof. Jifeng Dai and Dr. Xizhou Zhu	02/2022 - 10/2022
<b>Sea AI Lab</b>	Remotely at Beijing
Research Intern, collaborated with Prof. Shuicheng Yan and Dr. Jibin Wu	08/2021 - 02/2022
<b>CoLab, Institute of Artificial Intelligence at Beihang University</b>	Beijing
Research Intern, collaborated with Prof. Si Liu	08/2019 - 06/2022

## PUBLICATIONS

1. Mono-InternVL-1.5: Towards Cheaper and Faster Monolithic Multimodal Large Language Models  
Gen Luo, Wenhan Dou, Wenhao Li, Zhaokai Wang, Xue Yang, Changyao Tian, Hao Li, Weiyun Wang, Wenhao Wang, Xizhou Zhu, Yu Qiao, Jifeng Dai  
Preprint
2. Multimodal Music Generation with Explicit Bridges and Retrieval Augmentation  
Baisen Wang, Le Zhuo, Zhaokai Wang, Chenxi Bao, Chengjing Wu, Xuecheng Nie, Jiao Dai, Jizhong Han, Yue Liao, Si Liu  
Preprint
3. TIDE: Temporal-Aware Sparse Autoencoders for Interpretable Diffusion Transformers in Image Generation  
Victor Shea-Jay Huang, Le Zhuo, Yi Xin, Zhaokai Wang, Peng Gao, Hongsheng Li  
Preprint
4. Parameter-Inverted Image Pyramid Networks for Visual Perception and Multimodal Understanding  
Zhaokai Wang, Xizhou Zhu, Xue Yang, Gen Luo, Hao Li, Changyao Tian, Wenhan Dou, Junqi Ge, Lewei Lu, Yu Qiao, Jifeng Dai  
**TPAMI 2025**
5. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?  
Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, Gao Huang  
**ICML 2025 AI4MATH Workshop Best Paper Award (2/172)**
6. Sparkle: Mastering Basic Spatial Capabilities in Vision Language Models Elicits Generalization to Composite Spatial Reasoning  
Yihong Tang\*, Ao Qu\*, Zhaokai Wang\*, Dingyi Zhuang\*, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, Jinhua Zhao  
**EMNLP 2025 Findings & IJCAI 2025 MKLM Workshop Best Paper Award**
7. Vision-to-Music Generation: A Survey  
Zhaokai Wang, Chenxi Bao, Le Zhuo, Jingrui Han, Yang Yue, Yihong Tang, Victor Shea-Jay Huang, Yue Liao  
**ISMIR 2025**
8. OS Agents: A Survey on MLLM-based Agents for Computer, Phone and Browser Use  
Xueyu Hu, Tao Xiong, Biao Yi, Zishu Wei, Ruixuan Xiao, Yurun Chen, Jiasheng Ye, Meiling Tao, Xiangxin Zhou, Ziyu Zhao, Yuhuai Li, Shengze Xu, Shawn Wang, Xinchun Xu, Shuofei Qiao, Zhaokai Wang, Kun Kuang, Tiejong Zeng, Liang Wang, Jiwei Li, Yuchen Eleanor Jiang, Wangchunshu Zhou, Guoyin Wang, Keting Yin, Zhou Zhao, Hongxia Yang, Fan Wu, Shengyu Zhang, Fei Wu  
**ACL 2025 Oral**

9. Mono-InternVL: Pushing the Boundaries of Monolithic Multimodal Large Language Models with Endogenous Visual Pre-training  
Gen Luo\*, Xue Yang\*, Wenhan Dou\*, Zhaokai Wang\*, Jiawen Liu, Jifeng Dai, Yu Qiao, Xizhou Zhu  
**CVPR 2025**

10. SynerGen-VL: Towards Synergistic Image Understanding and Generation with Vision Experts and Token Folding  
Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, Jifeng Dai  
**CVPR 2025**

11. Parameter-Inverted Image Pyramid Networks  
Xizhou Zhu\*, Xue Yang\*, Zhaokai Wang\*, Hao Li, Wenhan Dou, Junqi Ge, Lewei Lu, Yu Qiao, Jifeng Dai  
**NeurIPS 2024 Spotlight - Ranked Top 10 in NeurIPS 2024 (among 15671 submissions), Top 2 in Computer Vision Area)**

12. ITINERA: Integrating Spatial Optimization with Large Language Models for Open-domain Urban Itinerary Planning  
Yihong Tang\*, Zhaokai Wang\*, Ao Qu\*, Yihao Yan\*, Zhaofeng Wu, Dingyi Zhuang, Jushi Kai, Kebin Hou, Xiaotong Guo, Jinhua Zhao, Zhan Zhao, Wei Ma  
**EMNLP 2024 Industry Track & KDD 2024 UrbComp Workshop Best Paper Award**

13. Auto MC-Reward: Automated Dense Reward Design with Large Language Models for Minecraft  
Hao Li\*, Xue Yang\*, Zhaokai Wang\*, Xizhou Zhu, Jie Zhou, Yu Qiao, Xiaogang Wang, Hongsheng Li, Lewei Lu, Jifeng Dai  
**CVPR 2024**

14. Video Background Music Generation: Dataset, Method and Evaluation  
Le Zhuo\*, Zhaokai Wang\*, Baisen Wang\*, Yue Liao, Chenxi Bao, Stanley Peng, Songhao Han, Aixi Zhang, Fei Fang, Si Liu  
**ICCV 2023**

15. Video Background Music Generation with Controllable Music Transformer  
Shangzhe Di, Zeren Jiang, Si Liu, Zhaokai Wang, Leyan Zhu, Zexin He, Hongming Liu, Shuicheng Yan  
**ACM Multimedia 2021 Best Paper Award (1/1942)**

16. Confidence-aware Non-repetitive Multimodal Transformers for TextCaps  
Zhaokai Wang, Renda Bao, Qi Wu, Si Liu  
**AAAI 2021**

SELECTED AWARDS AND HONORS

<a href="#">Best Paper Award</a> , ICML AI4MATH Workshop 2025	2025
<a href="#">Best Paper Award</a> , KDD Urban Computing Workshop (UrbComp) 2024	2024
<a href="#">Best Zero to One Award</a> , Alibaba Creator@AI Entrepreneur Hackathon Finals	2024
Outstanding Graduate of Beihang University	2022
<a href="#">Best Paper Award</a> , ACM Multimedia 2021	2021
<a href="#">Best Video Award</a> , IJCAI 2021 Video Competition	2021
<a href="#">First Place</a> , CVPR TextCaps Challenge	2020

ADDITIONAL INFORMATION

Talks:

- 2025.4: Talk on Mono-InternVL at [Open Multimodal Gathering Workshop](#) hosted by NUS [ShowLab](#).

Activities

- **Conference Reviewer:** ICCV 2023 & 2025, ECCV 2024, CVPR 2024 & 2025, EMNLP 2024, NeurIPS 2024, ICLR 2025, ICML 2025, AAAI 2025.
- **Journal Reviewer:** ACM Computing Survey.
- **Teaching Assistant:** Fundamentals of Computers (2021), Software Engineering (2022).

Skills

- **Programming languages:** Python, C/C++, Java, JavaScript.
- **Scientific packages:** Numpy, Pytorch, Tensorflow.
- **English:** TOEFL 111 (S26) , GRE 327+4.0, CET-4 669, CET-6 612.