Jiaqi Wang

Research Scientist & Team Leader

(+86) 18604792323 | wjqdev@gmail.com | myownskyw7.github.io | Jiaqi Wang

Profile

I am currently a Research Scientist and Team Leader at Shanghai AI Laboratory, as well as an Adjunct Ph.D. Supervisor at Shanghai Jiao Tong University. I received my Ph.D. in 2021 from the Multimedia Laboratory at the Chinese University of Hong Kong (MMLab@CUHK), where I was advised by Prof. Dahua Lin and worked closely with Prof. Chen Change Loy. My research focuses on visual perception, vision-language models, and the development of benchmarks and datasets in these areas.

I have published over 60 papers in top-tier journals and conferences, including TPAMI, CVPR, ICCV, ECCV, NeurIPS, ICML, ICLR, and ACL with more than 16,000 citations and a h-index of 44 on Google Scholar. Two of my papers (MMBench and ShareGPT4v) were recognized among the Most Influential ECCV Papers of 2024, and one paper (MMStar) was listed as one of the Most Influential NeurIPS Papers of 2024. Another work (OmniObject3D) was nominated as a CVPR 2023 Best Paper Candidate. I was also named in the Stanford/Elsevier Top 2% Scientists list in 2024.

As the lead researcher of the InternLM-XComposer multimodal large model series, our project has achieved over 4 million downloads and once ranked fifth among over 750,000 models on HuggingFace Trending.

Education

Sun Yat-Sen University (SYSU) — *Guangzhou, China*
B.Eng. in Software Engineering — *Aug. 2013 - Jun. 2017*

The Chinese University of Hong Kong (CUHK) — *Hong Kong*
Ph.D. in Information Engineering — *Aug. 2017 - Nov. 2021*

Work Experience

Shanghai AI Laboratory — *Shanghai, China*
Research Scientist & Team Leader — *Nov. 2021 - Now*

- I am leading a team focused on developing Large Vision-Language Models (LVLMs), overseeing four full-time researchers and more than twenty Ph.D. interns and joint Ph.D. students affiliated with Shanghai AI Laboratory.

Shanghai Jiao Tong University — *Shanghai, China*
Adjunct Ph.D. Supervisor — *Sep. 2024 - Now*

- I am supervising several joint Ph.D. students affiliated with Shanghai AI Laboratory and Shanghai Jiao Tong University.

Microsoft Research Asia (MSRA) — *Beijing, China*
Research Intern — *Jun. 2016 - Feb. 2017*

- I worked on 3D vision research.

Honors & Awards

2024 **Standford/Elsevier Top 2% Scientists List**, Top 2%
2024 **One Most Influential NeurIPS Papers 2024 (MMStar),** Top 0.37% （15/4035 Accept Papers）
2024 **Two Most Influential ECCV Papers 2024 (MMBench, ShareGPT4V),** Top 0.62% （15/2395 Accept Papers）
2023 **CVPR 2023 Award Candidate**, Top 0.13%（12/9155 Submit Papers）
2019 **1st place entry in COCO 2019 Object Detection Challenge (without external data)** , Team: MMDet, **Team Leader**
2018 **1st place entry in COCO 2018 Object Detection Challenge**, Team: MMDet, Core Member
2017 **Hong Kong PhD Fellowship Award (HKPFS)** , Top 4%
14&15&16 **National Scholarship**, Top 2%

Research Interests

Large Vision Language Models (LVLMs)

Recent works:

- InternLM-XComposer: A Vision-Language Large Model for Advanced Text-image Comprehension and Composition (2023)
- InternLM-XComposer2: Mastering Free-form Text-Image Composition and Comprehension in Vision-Language Large Model (2024)
- InternLM-XComposer2-4KHD: A Pioneering Large Vision-Language Model Handling Resolutions from 336 Pixels to 4K HD (2024)
- InternLM-XComposer-2.5: A Versatile Large Vision Language Model Supporting Long-Contextual Input and Output (2024)
- DualToken: Towards Unifying Visual Understanding and Generation with Dual Visual Vocabularies (2025)
- Autoregressive Semantic Visual Reconstruction Helps VLMs Understand Better (2025)

Video Large Language Models (Video-LLMs)

Recent works:

- Streaming Long Video Understanding with Large Language Models (2024)
- InternLM-XComposer2.5-OmniLive: A Comprehensive Multimodal System for Long-term Streaming Video and Audio Interactions (2024)
- VideoRoPE: What Makes for Good Video Rotary Position Embedding? (2025)

RL for LVLMs

Recent works:

- InternLM-XComposer2.5-Reward: A Simple Yet Effective Multi-Modal Reward Model (2025)
- Visual-RFT: Visual Reinforcement Fine-Tuning (2025)
- Unified Reward Model for Multimodal Understanding and Generation (2025)
- Unified Multimodal Chain-of-Thought Reward Model through Reinforcement Fine-Tuning (2025)

Dataset & Benchmarks

Recent works:

- ShareGPT4V: Improving Large Multi-Modal Models with Better Captions (2023)
- MMBench: Is Your Multi-modal Model an All-around Player? (2023)
- Are We on the Right Way for Evaluating Large Vision-Language Models? (2024)
- MMDU: A Multi-Turn Multi-Image Dialog Understanding Benchmark and Instruction-Tuning Dataset for LVLMs (2024)
- OVO-Bench: How Far is Your Video-LLMs from Real-World Online Video Understanding? (2025)

Projects

InternLM-XComposer

Project Leader *2023 - now*

- InternLM-XComposer (IXC) is one of the most influential open-source multimodal large model series, with over **4 million** cumulative downloads on Hugging Face and once **ranked fifth among over 750,000 models** on HuggingFace Trending.
- **Project Link**: https://github.com/InternLM/InternLM-XComposer

MMDetection

Second author and Core Developer *2018 - now*

- MMDetection is one of the most popular object detection codebase. It has more than **30k stars** on Github.
- Github Link: https://github.com/open-mmlab/mmdetection
- Technical Report: MMDetection: Open MMLab Detection Toolbox and Benchmark

Publications

Region Proposal by Guided Anchoring *CVPR 2019*

***J. Wang**, K. Chen, S. Yang, C. C. Loy, D. Lin* ***Poster**, First Author*

CARAFE: Content-Aware ReAssembly of FEatures *ICCV 2019*

***J. Wang**, K. Chen, R. Xu, Z. Liu, C. C. Loy, D. Lin* ***Oral (Top 4.3%)**, First Author*

Side-Aware Boundary Localization for More Precise Object Detection *ECCV 2020*

***J. Wang**, W. Zhang, Y. Cao, K. Chen, J. Pang, T. Gong, J. Shi, C. C. Loy, D. Lin* ***Spotlight (Top 5%)**, First Author*

Seesaw Loss for Long-Tailed Instance Segmentation *CVPR 2021*

***J. Wang**, W. Zhang, Y. Zang, Y. Cao, J. Pang, T. Gong, K. Chen, Z. Liu, C. C. Loy, D. Lin* ***Poster**, First Author*

CARAFE++: Unified Content-Aware ReAssembly of FEatures *TPAMI 2021*

***J. Wang**, K. Chen, R. Xu, Z. Liu, C. C. Loy, D. Lin* ***Regular Paper**, First Author*

V3Det: Vast Vocabulary Visual Detection Dataset — *ICCV 2023*
J. Wang, P. Zhang, T. Chu, Y. Cao, Y. Zhou, T. Wu, B. Wang, C. He, D. Lin — ***Oral (Top 4%)**, First Author*

LAVT: Language-Aware Vision Transformer for Referring Image Segmentation — *CVPR 2022*
Z. Yang, J. Wang*, Y. Tang, K. Chen, H. Zhao, P. H. S. Torr* — ***Poster**, Co-first Author*

Language-Aware Vision Transformer for Referring Segmentation — *TPAMI 2024*
Z. Yang, J. Wang*, X. Ye*, Y. Tang, K. Chen, H. Zhao, P. H. S. Torr* — ***Regular Paper**, Co-first Author*

Few-Shot Object Detection via Association and DIscrimination — *NeurIPS 2021*
Y. Cao, J. Wang†, Y. Jin, T. Wu, K. Chen, Z. Liu, D. Lin — ***Poster**, Corresponding Author*

Semi-Supervised Semantic Segmentation via Gentle Teaching Assistant — *NeurIPS 2022*
Y. Jin, J. Wang†, D. Lin — ***Poster**, Corresponding Author*

UPop: Unified and Progressive Pruning for Compressing Vision-Language Transformers — *ICML 2023*
D.Shi, C. Tao, Y. Jin, Z. Yang, C. Yuan, J. Wang† — ***Poster**, Last Author*

Voxurf: Voxel-based Efficient and Accurate Neural Surface Reconstruction — *ICLR 2023*
T. Wu, J. Wang†, X. Pan, X. Xu, C. Theobalt, Z. Liu, D. Lin — ***Spotlight**, Corresponding Author*

Multi-level Logit Distillation — *CVPR 2023*
Y. Jin, J. Wang†, D. Lin — ***Poster**, Corresponding Author*

BUOL: A Bottom-Up Framework with Occupancy-aware Lifting for Panoptic 3D Scene Reconstruction From A Single Image — *CVPR 2023*
T. Chu, P. Zhang, Q. Liu, J. Wang — ***Poster**, Last Author*

CrossGET: Cross-Guided Ensemble of Tokens for Accelerating Vision-Language Transformers — *ICML 2024*
D. Shi, C. Tao, A. Rao, Z. Yang, C. Yuan, J. Wang† — ***Poster**, Last Author*

GPT4Point: A Unified Framework for Point-Language Understanding and Generation — *CVPR 2024*
Z. Qi, Y. Fang, Z. Sun, X. Wu, T. Wu, J. Wang†, D. Lin, H. Zhao — ***Highlight**, Corresponding Author*

Alpha-CLIP: A clip model focusing on wherever you want — *CVPR 2024*
Z. Sun, Y. Fang, T. Wu, P. Zhang, Y. Zang, S. Kong, Y. Xiong, D. Lin, J. Wang† — ***Poster**, Last Author*

Long-CLIP: Unlocking the Long-Text Capability of CLIP — *ECCV 2024*
B. Zhang, P. Zhang, X. Dong, Y. Zang, J. Wang† — ***Poster**, Last Author*

InternLM-XComposer2-4KHD: A Pioneering Large Vision-Language Model Handling Resolutions from 336 Pixels to 4K HD — *NeurIPS 2024*
X. Dong, P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, S. Zhang, H. Duan, W. Zhang, Y. Li, H. Yan, Y. Gao, Z. Chen, X. Zhang, W. Li, J. Li, W. Wang, K. Chen, C. He, X. Zhang, J. Dai, Y. Qiao, D. Lin, J. Wang† — ***Poster**, Last Author*

Are We on the Right Way for Evaluating Large Vision-Language Models? — *NeurIPS 2024*
L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang†, Y. Qiao, D. Lin, F. Zhao — ***Poster**, Corresponding Author*

Streaming Long Video Understanding with Large Language Models — *NeurIPS 2024*
R. Qian, X. Dong, P. Zhang, Y. Zang, S. Ding, D. Lin, J. Wang† — ***Poster**, Last Author*

MMDU: A Multi-Turn Multi-Image Dialog Understanding Benchmark and Instruction-Tuning Dataset for LVLMs — *NeurIPS 2024 D&B Track*

Z. Liu, T. Chu, Y. Zang, X. Wei, X. Dong, P. Zhang, Z. Liang, Y. Xiong, Y. Qiao, D. Lin, ***J. Wang***† — ***Poster***, *Last Author*

ShareGPT4Video: Improving Video Understanding and Generation with Better Captions — *NeurIPS 2024 D&B Track*

L. Chen, X. Wei, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, B. Lin, Z. Tang, L. Yuan, Y. Qiao, D. Lin, F. Zhao, ***J. Wang***† — ***Poster***, *Last Author*

MIA-DPO: Multi-Image Augmented Direct Preference Optimization For Large Vision-Language Models — *ICLR 2025*

Z. Liu, Y. Zang, X. Dong, P. Zhang, Y. Cao, H. Duan, C. He, Y. Xiong, D. Lin, ***J. Wang***† — ***Poster***, *Last Author*

PyramidDrop: Accelerating Your Large Vision-Language Models via Pyramid Visual Redundancy Reduction — *CVPR 2025*

L. Xing, Q. Huang, X. Dong, J. Lu, P. Zhang, Y. Zang, Y. Cao, C. He, ***J. Wang***†*, F. Wu, D. Lin* — ***Poster***, *Corresponding Author*

Dispider: Enabling Video LLMs with Active Real-Time Interaction via Disentangled Perception, Decision, and Reaction — *CVPR 2025*

R. Qian, S. Ding, X. Dong, P. Zhang, Y. Zang, Y. Cao, D. Lin, ***J. Wang***† — ***Poster***, *Last Author*

OVO-Bench: How Far is Your Video-LLMs from Real-World Online Video Understanding? — *CVPR 2025*

Y. Li, J. Niu, Z. Miao, C. Ge, Y. Zhou, Q. He, X. Dong, H. Duan, S. Ding, R. Qian, P. Zhang, Y. Zang, Y. Cao, C. He, ***J. Wang***† — ***Poster***, *Last Author*

ByTheWay: Boost Your Text-to-Video Generation Model to Higher Quality in a Training-free Way — *CVPR 2025*

J. Bu, P. Ling, P. Zhang, T. Wu, X. Dong, Y. Zang, Y. Cao, D. Lin, ***J. Wang***† — ***Poster***, *Last Author*

VideoRoPE: What Makes for Good Video Rotary Position Embedding? — *ICML 2025*

X. Wei, X. Liu, Y. Zang, X. Dong, P. Zhang, Y. Cao, J. Tong, H. Duan, Q. Guo, ***J. Wang***†*, X. Qiu, D. Lin* — ***Oral***, *Corresponding Author*

SongGen: A Single Stage Auto-regressive Transformer for Text-to-Song Generation — *ICML 2025*

Z. Liu, S. Ding, Z. Zhang, X. Dong, P. Zhang, Y. Zang, Y. Cao, D. Lin, ***J. Wang***† — ***Poster***, *Last Author*

SongComposer: A Large Language Model for Lyric and Melody Generation in Song Composition — *ACL 2025*

S. Ding, Z. Liu, X. Dong, P. Zhang, R. Qian, J. Huang, C. He, D. Lin, ***J. Wang***† — ***Main***, *Last Author*

InternLM-XComposer2.5-Reward: A Simple Yet Effective Multi-Modal Reward Model — *ACL 2025*

Y. Zang, X. Dong, P. Zhang, Y. Cao, Z. Liu, S. Ding, S. Wu, Y. Ma, H. Duan, W. Zhang, K. Chen, D. Lin, ***J. Wang***† — ***Findings***, *Last Author*

Visual-RFT: Visual Reinforcement Fine-Tuning — *ICCV 2025*

Z. Liu, Z. Sun, Y. Zang, X. Dong, Y. Cao, H. Duan, D. Lin, ***J. Wang***† — ***Poster***, *Last Author*

SAM2Long: Enhancing SAM 2 for Long Video Segmentation with a Training-Free Memory Tree — *ICCV 2025*

S. Ding, R. Qian, X. Dong, P. Zhang, Y. Zang, Y. Cao, Y. Guo, D. Lin, ***J. Wang***† — ***Poster***, *Last Author*

X-Prompt: Generalizable Auto-Regressive Visual Learning with In-Context Prompting — *ICCV 2025*

Z. Sun, Z. Chu, P. Zhang, T. Wu, X. Dong, Y. Zang, Y. Xiong, D. Lin, ***J. Wang***† — ***Poster***, *Last Author*

Bootstrap3D: Improving Multi-view Diffusion Model with Synthetic Data — *ICCV 2025*

Z. Sun, T. Wu, P. Zhang, Y. Zang, X. Dong, Y. Xiong, D. Lin, ***J. Wang***† — ***Poster***, *Last Author*

MM-IFEngine: Towards Multimodal Instruction Following — *ICCV 2025*
*S. Ding, S. Wu, X. Zhao, Y. Zang, H. Duan, X. Dong, P. Zhang, Y. Cao, D. Lin, **J. Wang**†* — ***Poster***, *Last Author*

OCBEV: Object-Centric BEV Transformer for Multi-View 3D Object Detection — *3DV 2023*
*Z. Qi, **J. Wang**†, X. Wu, H. Zhao* — ***Poster***, *Corresponding Author*

Zero-shot Skeleton-based Action Recognition via Mutual Information Estimation and Maximization — *MM 2023*
*Y. Zhou, W. Qiang, A. Rao, N. Lin, B. Su, **J. Wang*** — ***Poster***, *Last Author*

Self-supervised Action Representation Learning from Partial Spatio-Temporal Skeleton Sequences — *AAAI 2022*
*Y. Zhou, H. Duan, A. Rao, B. Su, **J. Wang*** — ***Oral***, *Last Author*

OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation — *CVPR 2023*
*T. Wu, J. Zhang, X. Fu, Y. Wang, J. Ren, L. Pan, W. Wu, L. Yang, **J. Wang**, C. Qian, D. Lin, Z. Liu* — ***Award Candidate***

Optimizing Video Object Detection via a Scale-Time Lattice — *CVPR 2018*
*K. Chen, **J. Wang**, S. Yang, X. Zhang, Y. Xiong, C. C. Loy, D. Lin* — ***Poster***

Hybrid Task Cascade for Instance Segmentation — *CVPR 2019*
*K. Chen, J. Pang, **J. Wang**, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, D. Lin* — ***Poster***

Texture Memory-Augmented Deep Patch-Based Image Inpainting — *TIP 2021*
*R. Xu, M. Guo, **J. Wang**, X. Li, B. Zhou, C. C. Loy* — ***Regular Paper***

Semantics-Aware Dynamic Localization and Refinement for Referring Image Segmentation — *AAAI 2022*
*Z. Yang, **J. Wang**, Y. Tang, K. Chen, H. Zhao, P. H. S. Torr* — ***Poster***

Pyskl: Towards good practices for skeleton action recognition — *MM 2022*
*H. Duan, **J. Wang**, K Chen, D Lin* — ***Poster***

Dense Distinct Query for End-to-End Object Detection — *CVPR 2023*
*S. Zhang, X. Wang, **J. Wang**, J. Pang, C. Lyu, W. Zhang, P. Luo, K. Chen* — ***Poster***

HyperDreamer: Hyper-Realistic 3D Content Generation and Editing from a Single Image — *SIGGRAPH Asia 2023*
*T. Wu, Z. Li, S. Yang, P. Zhang, X. Pan, **J. Wang**, Z. Liu, D. Lin* — ***Poster***

VIGC: Visual Instruction Generation and Correction — *AAAI 2024*
*B. Wang, F. Wu, X. Han, J. Peng, H. Zhong, P. Zhang, X. Dong, W. Li, W. Li, **J. Wang**, C. He* — ***Poster***

OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation — *CVPR 2024*
*Q. Huang, X. Dong, P. Zhang, B. Wang, C. He, **J. Wang**, D. Lin, W. Zhang, N. Yu* — ***Highlight***

OneLLM: One Framework to Align All Modalities with Language — *CVPR 2024*
*J. Han, K. Gong, Y. Zhang, **J. Wang**, K. Zhang, D. Lin, Y. Qiao, P. Gao, X. Yue* — ***Poster***

Mmbench: Is your multi-modal model an all-around player? — *ECCV 2024*
*Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, **J. Wang**, C. He, Z. Liu, K. Chen, D. Lin* — ***Oral***

ShareGPT4V: Improving Large Multi-Modal Models with Better Captions — *ECCV 2024*
*L. Chen, J. Li, X. Dong, P. Zhang, C. He, **J. Wang**, F. Zhao, D. Lin* — *Poster*

Make-it-Real: Unleashing Large Multimodal Model for Painting 3D Objects with Realistic Materials — *NeurIPS 2024*
*Y. Fang, Z. Sun, T. Wu, **J. Wang**, Z. Liu, G. Wetzstein, D. Lin* — *Poster*

Prism: A Framework for Decoupling and Assessing the Capabilities of VLMs — *NeurIPS 2024*
*Y. Qiao, H. Duan, X. Fang, J. Yang, L. Chen, S. Zhang, **J. Wang**, D. Lin, K. Chen* — *Poster*

FiVA: Fine-grained Visual Attribute Dataset for Text-to-Image Diffusion Models — *NeurIPS 2024 D&B Track*
*T. Wu, Y. Xu, R. Po, M. Zhang, G. Yang, **J. Wang**, Z. Liu, D. Lin, G. Wetzstein* — *Poster*

MMLongBench-Doc: Benchmarking Long-context Document Understanding with Visualizations — *NeurIPS 2024 D&B Track*
*Y. Ma, Y. Zang, L. Chen, M. Chen, Y. Jiao, X. Li, X. Lu, Z. Liu, Y. Ma, X. Dong, P. Zhang, L. Pan, Y. Jiang, **J. Wang**, Y. Cao, A. Sun* — *Spotlight*

Utilize the Flow before Stepping into the Same River Twice: Certainty Represented Knowledge Flow for Refusal-Aware Instruction Tuning — *AAAI 2025*
*R. Zhu, Z. Ma, J. Wu, J. Gao, **J. Wang**, D. Lin, C. He* — *Poster*

MotionClone: Training-Free Motion Cloning for Controllable Video Generation — *ICLR 2025*
*P. Ling, J. Bu, P. Zhang, X. Dong, Y. Zang, T. Wu, H. Chen, **J. Wang**, Y. Jin* — *Poster*

IDArb: Intrinsic Decomposition for Arbitrary Number of Input Views and Illuminations — *ICLR 2025*
*Z. Li, T. Wu, J. Tan, M. Zhang, **J. Wang**, D. Lin* — *Poster*

OmniAlign-V: Towards Enhanced Alignment of MLLMs with Human Preference — *ACL 2025*
*X. Zhao, S. Ding, Z. Zhang, H. Huang, M. Cao, **J. Wang**, W. Wang, X. Fang, W. Wang, G. Zhai, H. Yang, H. Duan, K. Chen* — *Main*

Towards Storage-Efficient Visual Document Retrieval: An Empirical Study on Reducing Patch-Level Embeddings — *ACL 2025*
*Y. Ma, J. Li, Y. Zang, X. Wu, X. Dong, P. Zhang, Y. Cao, H. Duan, **J. Wang**, Y. Cao, A. Sun* — *Findings*

Light-A-Video: Training-free Video Relighting via Progressive Light Fusion — *ICCV 2025*
*Y Zhou, J. Bu, P. Ling, P. Zhang, T. Wu, Q. Huang, J. Li, X. Dong, Y. Zang, Y. Cao, A. Rao, **J. Wang**, L. Niu* — *Poster*

Deciphering Cross-Modal Alignment in Large Vision-Language Models via Modality Integration Rate — *ICCV 2025*
*Q. Huang, X. Dong, P. Zhang, Y. Zang, Y. Cao, **J. Wang**, W. Zhang, N. Yu* — *Poster*