Hao Wang

Ph.D. Candidate in Computer Vision, Multimodal AI, and Agentic AI

HCP Lab, Sun Yat-sen University & Pengcheng Laboratory
Expected graduation: December 2026
Open to industry research roles and collaborations

wanghao9610@gmail.com
wanghao9610.github.io
+86 183 0102 8055

PROFILE

Ph.D. candidate focused on open-ended visual perception, multimodal large language models, and agentic AI. My recent work builds unified models for image and video segmentation, combining language instructions, visual prompts, open-vocabulary recognition, and temporally consistent pixel-level understanding. I am particularly interested in research systems that connect foundation models with reliable pixel-level perception for practical vision applications. My future research interests will focus on the multi-modal agentic models.

EDUCATION

Sun Yat-sen University & Pengcheng Laboratory **2022.09 – Present**

Ph.D. student, School of Intelligent Systems Engineering

- Co-supervised by Prof. Xiaodan Liang and Associate Prof. Xiangyuan Lan.
- Research areas: multimodal large language models, open-ended visual understanding, image/video segmentation.
- Expected to graduate in December 2026; actively seeking research positions in industry.

University of Chinese Academy of Sciences & Institute of Automation, CAS **2019.09 – 2022.06**

Master student, School of Artificial Intelligence

- Supervised by Prof. Jing Liu.

Beijing Jiaotong University **2015.09 – 2019.06**

Bachelor student, School of Electronic and Information Engineering

RESEARCH CONTRIBUTIONS

- Built a line of first-author research on open-ended perception, moving from video semantic segmentation to open-vocabulary detection and multimodal any-segmentation.
- Developed multimodal segmentation frameworks that accept both natural-language instructions and visual prompts, aiming to bridge high-level reasoning and dense mask prediction.
- Explored unified image/video segmentation with memory mechanisms for temporally consistent pixel-level understanding across frames.
- Released project pages and code for multiple representative works, reflecting a strong preference for reproducible and open research.

RESEARCH EXPERIENCE

Meituan M17-MM **2025.01 – Present**

Research Intern

- Worked with Limeng Qiao, Lin Ma, and Guanglu Wan on multimodal segmentation and pixel-level perception.
- Developed X2SAM, a unified any-segmentation MLLM for images and videos with conversational instructions and visual prompts.
- Investigated mask memory for preserving object-level consistency in video segmentation and interactive perception.

Meituan Vision Intelligence Department **2022.07 – 2025.01**

Research Intern

- Worked with Zequn Jie and Lin Ma on open-vocabulary detection and multimodal segmentation systems.
- Contributed to OV-DINO and X-SAM, advancing open-vocabulary recognition and promptable segmentation.
- Studied language-aware fusion and multimodal supervision for scalable visual recognition and segmentation.

Tencent AI Platform Department **2021.05 – 2021.08**

Application Research Intern

- Conducted applied research in AI platform scenarios.

Huawei Photo Processing Department **2019.09 – 2020.07**

Application Project Intern

- Worked on applied computer vision projects for photo processing.

FIRST-AUTHOR PUBLICATIONS

X2SAM: Any Segmentation in Images and Videos **arXiv, 2026**
Hao Wang, Limeng Qiao, Chi Zhang, Guanglu Wan, Lin Ma, Xiangyuan Lan, Xiaodan Liang
Unified segmentation MLLM for images and videos, supporting conversational instructions, visual prompts, and temporally consistent mask memory. Extends any-segmentation from static images to videos while preserving dense pixel-level responses. Paper | Code

X-SAM: From Segment Anything to Any Segmentation **AAAI, 2026**
Hao Wang, Limeng Qiao, Zequn Jie, Zhijian Huang, Chengjian Feng, Qingfang Zheng, Lin Ma, Xiangyuan Lan, Xiaodan Liang
Unified MLLM framework extending segment-anything capabilities to any segmentation and pixel-level perceptual understanding. Supports flexible segmentation through multimodal inputs and instruction-driven interaction. Paper | Code

OV-DINO: Unified Open-Vocabulary Detection with Language-Aware Selective Fusion **arXiv, 2024**
Hao Wang, Pengzhen Ren, Zequn Jie, Xiao Dong, Chengjian Feng, Yinlong Qian, Lin Ma, Dongmei Jiang, Yaowei Wang, Xiangyuan Lan, Xiaodan Liang
Unified open-vocabulary detector pre-trained on diverse large-scale datasets with language-aware selective fusion. Designed to improve category generalization by aligning visual detection with language semantics. Paper | Code

TMANet: Temporal Memory Attention for Video Semantic Segmentation **ICIP, 2021**
Hao Wang, Weining Wang, Jing Liu
Temporal memory attention for long-range video semantic segmentation without optical-flow prediction. Paper | Code

ADDITIONAL PUBLICATION

WL-MSR: Watch and Listen for Multimodal Subtitle Recognition **ICASSP, 2023**
Jiawei Liu, **Hao Wang**, Weining Wang, Xingjian He, Jing Liu
Transformer-based multimodal subtitle recognition using OCR and ASR information with mask/crop strategies and multi-level identity embeddings.

PROFESSIONAL SERVICE

Conference Reviewer	AAAI 2026, ICCV 2023, ECCV 2024
Journal Reviewer	Proceedings of the IEEE

AWARDS

2021.09 1st place in the 1st VSPW Challenge Workshop, ICCV 2021.