Yolo Yunlong Tang

E-mail: yunlong.tang@rochester.edu | Phone: (+1)585-616-0074

Homepage: yunlong10.github.io | Google Scholar | GitHub | LinkedIn

Research Area: *Multimodal Learning, LMMs/Agents for Video Understanding*

Education

University of Rochester — *Aug. 2023 - Present*
Ph.D. Student in Computer Science, advised by Prof. Chenliang Xu — *Rochester, NY*

Southern University of Science and Technology (SUSTech) — *Aug. 2019 - Jun. 2023*
B.Eng. in Intelligence Science and Technology, advised by Prof. Feng Zheng — *Shenzhen, China*

Work & Internships

Amazon Ring AI — *May 2025 - Aug. 2025*
Applied Scientist Intern. Host: Wei Wang, Liqiang He — *Bellevue, WA*

ByteDance Multimedia Lab — *May 2024 - Aug. 2024*
Research Intern, mentored by Gen Zhan and Yiting Liao — *San Jose, CA*

SUSTech Visual Intelligence & Perception Lab — *Aug. 2022 - Jul. 2023*
Undergrad. Research Assistant, worked with Teng Wang and Prof. Feng Zheng — *Shenzhen, China*

Tencent Data Platform — *Sept. 2021 - Aug. 2022*
Research Intern, mentored by Qin Lin and Wenhao Jiang — *Shenzhen, China*

Publications & Preprints

[1] **Yolo Y. Tang**, Daiki Shimada, Hang Hua, Chao Huang, Jing Bi, Rogerio Feris, and Chenliang Xu. *Video-R4: Reinforcing Text-Rich Video Reasoning with Visual Rumination*. In: *Proceedings of the Computer Vision and Pattern Recognition Conference* (***CVPR***) *Findings*. 2026.

[2] **Yolo Y. Tang**, Chao Huang, Susan Liang, Jing Bi, Yicheng Wang, Daiki Shimada, and Chenliang Xu. *Asynchronous Temporal Modeling with Two-Agent Framework for Streaming Dense Video Captioning*. In: *Proceedings of the Computer Vision and Pattern Recognition Conference* (***CVPR***). 2026.

[3] **Yunlong Tang**, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. *Video Understanding with Large Language Models: A Survey*. In: *IEEE Transactions on Circuits and Systems for Video Technology* (***TCSVT***) (2025).

[4] **Yunlong Tang***, Pinxin Liu*, Zhangyun Tan*, Mingqian Feng, Rui Mao, Chao Huang, Jing Bi, Yunzhong Xiao, Susan Liang, Hang Hua, Ali Vosoughi, Luchuan Song, Zeliang Zhang, and Chenliang Xu. *MMPerspective: Do MLLMs Understand Perspective? A Comprehensive Benchmark for Perspective Perception, Reasoning, and Robustness*. In: *The 39th Annual Conference on Neural Information Processing Systems* (***NeurIPS***) *Datasets and Benchmarks Track*. 2025.

[5] **Yunlong Tang***, Junjia Guo*, Hang Hua, Susan Liang, Mingqian Feng, Xinyang Li, Rui Mao, Chao Huang, Jing Bi, Zeliang Zhang, Pooyan Fazli, and Chenliang Xu. *VidComposition: Can MLLMs Analyze Compositions in Compiled Videos?* In: *Proceedings of the Computer Vision and Pattern Recognition Conference* (***CVPR***). 2025.

[6] **Yunlong Tang**, Jing Bi, Chao Huang, Susan Liang, Daiki Shimada, Hang Hua, Yunzhong Xiao, Yizhi Song, Pinxin Liu, Mingqian Feng, Junjia Guo, Zhuo Liu, Luchuan Song, Ali Vosoughi, Jinxi He, Liu He, Zeliang Zhang, Jiebo Luo, and Chenliang Xu. *Caption Anything in Video: Fine-grained Object-centric Captioning via Spatiotemporal Multimodal Prompting*. In: ***AAAI** Demonstration Program* (2026). *Best Demo Award Runner-up*.

[7] **Yunlong Tang**, Daiki Shimada, Jing Bi, Mingqian Feng, Hang Hua, and Chenliang Xu. *Empowering LLMs with Pseudo-Untrimmed Videos for Audio-Visual Temporal Understanding*. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (***AAAI***). 2025.

[8] Hang Hua*, **Yunlong Tang***, Chenliang Xu, and Jiebo Luo. *V2Xum-LLM: Cross-modal Video Summarization with Temporal Prompt Instruction Tuning*. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (***AAAI***). 2025.

[9] **Yunlong Tang**, Gen Zhan, Li Yang, Yiting Liao, and Chenliang Xu. *CaRDiff: Video Salient Object Ranking Chain of Thought Reasoning for Saliency Prediction with Diffusion*. In: *Proceedings of the AAAI Conference on Artificial Intelligence* (***AAAI***). 2025.

[10] **Yolo Y. Tang**, Jing Bi, Pinxin Liu, Zhenyu Pan, Zhangyun Tan, Qianxiang Shen, Jiani Liu, Hang Hua, Junjia Guo, Yunzhong Xiao, Chao Huang, Zhiyuan Wang, Susan Liang, Xinyi Liu, Yizhi Song, Junhua Huang, Jia-Xing Zhong, Bozheng Li, Daiqing Qi, Ziyun Zeng, Ali Vosoughi, Luchuan Song, Zeliang Zhang, Daiki Shimada, Han Liu, Jiebo Luo, and Chenliang Xu. *Video-LMM Post-Training: A Deep Dive into Video Reasoning with Large Multimodal Models*. In: *arXiv* (2025).

[11] Jing Bi, Junjia Guo, **Yolo Y. Tang**, Lianggong Bruce Wen, Zhang Liu, and Chenliang Xu. *Unveiling Visual Perception in Language Models: An Attention Head Analysis Approach*. In: *Proceedings of the Computer Vision and Pattern Recognition Conference* (***CVPR***). 2025.

[12] Jing Bi, **Yunlong Tang**, Luchuan Song, Ali Vosoughi, Nguyen Nguyen, and Chenliang Xu. *EAGLE: Egocentric AGgregated Language-video Engine*. In: *Proceedings of the 32nd ACM International Conference on Multimedia* (***ACM MM***). 2024.

[13] **Yunlong Tang**, Jinrui Zhang, Xiangchen Wang, Teng Wang, and Feng Zheng. *LLMVA-GEBC: Large Language Model with Video Adapter for Generic Event Boundary Captioning*. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (***CVPR***) *Workshops*. 2023.

[14] Teng Wang*, Jinrui Zhang*, Junjie Fei*, Hao Zheng, **Yunlong Tang**, Zhe Li, Mingqi Gao, and Shanshan Zhao. *Caption Anything: Interactive Image Description with Diverse Multimodal Controls*. In: *arXiv* (2023).

[15] **Yunlong Tang**, Siting Xu, Teng Wang, Qin Lin, Qinglin Lu, and Feng Zheng. *Multi-modal Segment Assemblage Network for Ad Video Editing with Importance-Coherence Reward*. In: *Proceedings of the Asian Conference on Computer Vision* (***ACCV***). 2022.

High-Impact Projects

1. **Awesome LLMs for Video Understanding** (`GitHub Stars 3.1k+`)
   Latest papers, codes, and datasets on Video-LLMs. Repository for the survey paper [3].
   `https://github.com/yunlong10/Awesome-LLMs-for-Video-Understanding`
2. **Caption-Anything** (`GitHub Stars 1.8k+`)
   Open-source implementation of Caption-Anything [14], a versatile image processing tool that combines Segment Anything, Visual Captioning, and ChatGPT.
   `https://github.com/ttengwang/Caption-Anything`

Honors and Awards

Best Demo Award Runner-Up at AAAI 2026 Demonstration Program *2026*
The First Place in the AIM Challenge on Video Saliency Prediction at ECCV 2024 Workshop *2024*
The First Place in the GEBC Track of LOVEU Challenge at CVPR 2023 Workshop *2023*
Excellent Graduate for Exceptional Performance, SUSTech *2023*

Excellent Undergraduate Thesis, Department of Computer Science and Engineering, SUSTech *2023*
The First Class of Merit Student Scholarship for Exceptional Performance, SUSTech *2022*
Research Innovation Award, Shude College, SUSTech *2021*

Teaching

Teaching Assistant at University of Rochester
CSC 242 Artificial Intelligence *Instructor: Prof. Thaddeus E. Pawlicki* *Fall 2025*
CSC 249/449 Machine Vision *Instructor: Prof. Chenliang Xu* *Spring 2025*
CSC 245/445 Deep Learning *Instructor: Prof. Chenliang Xu* *Fall 2024*

Teaching Assistant at SUSTech
CS308 Computer Vision *Instructor: Prof. Feng Zheng* *Spring 2023*
CS308 Computer Vision *Instructor: Prof. Feng Zheng* *Fall 2022*

Service

Conference Reviewer
CVPR 2026, ICML 2026, AAAI 2026, NeurIPS 2025, ICML 2025, ICASSP 2025, ICLR 2025, NeurIPS 2024, ACL 2024, ACM MM 2024, CVPR 2024

Journal Reviewer
TPAMI, TMM, TIP, TCSVT

Skills

Programming Languages:
Proficient: Python, C/C++, Linux Shell
Capable: JavaScript, Java, SQL, MATLAB

Natural Languages:
Mandarin Chinese (*native*), *English* (*fluent*), *Japanese* (*beginner*)

Tools & Frameworks:
PyTorch, Git, LaTeX, *OpenCV, FFmpeg, HuggingFace, Claude Code*