EN YU

PhD. Student at Huazhong University of Science and Technology (HUST), Hongshan District, Wuhan, P.R.China
Email: yuen_daniel@outlook.com | GitHub: https://github.com/Ahnsun
Homepage: https://Ahnsun.github.io

EDUCATION

Huazhong University of Science and Technology *Sep 2022 - Jun 2026 (expected)*
PhD., Department of Intelligence Science and Technology
- **Academic Advisor:** Prof. Wenbing Tao

Huazhong University of Science and Technology *Aug 2020 - July 2022*
M.S., Department of Intelligence Science and Technology
- **Academic Advisor:** Prof. Shoudong Han

Huazhong University of Science and Technology *Aug 2016 - July 2020*
B.Eng, School of Automation
- **Award:** First Prize in the 13th National College Students' Intelligent Car Competition.
- National Champion in the 14th National College Students' Intelligent Car Competition.

RESEARCH INTERN EXPERIENCE

MEGVII Technology Beijing, CN
Research Intern at Foundation Model Group - Mentor: Xiangyu Zhang *Sep 2022 - Mar 2024*
StepFun AI Beijing, CN
Research Intern at Multimodal LLM Group - Mentor: Zheng Ge *Mar 2024 - July 2024*
University of California, Santa Barbara Santa Barbara, California, USA
Visiting PhD at UCSB NLP Group - Mentor: William Wang *July 2024 - May 2025*

MAIN RESEARCH CONTRIBUTIONS

Visual Perception, Understanding and Reasoning with Multimodal LLMs

- Focus on Multimodal Foundation Models adept at extracting knowledge and common sense across real-world (images or video sequences), aiming for visual perception, understanding and advanced reasoning & planning. Pioneering advancements in Multimodal LLMs pre-training and post-training.
- *Representative contributions:* Multimodal LLM in image understanding [IJCAI'24 Long Oral], and video understanding and reasoning [ECCV'24 Poster] and scalable video-language modeling [ICLR'25 Poster]. Advanced Multimodal LLM alignment [NeurIPS'25 Submission],[ICML'25 Poster] and RL post-training [NeurIPS'25 Submission].

Spatial Intelligence of Visual and Multimodal Foundation Models

- Explore to enable models to develop cognition of the physical world through localization, tracking, and intentional reasoning of objects in real-world 2D/3D imagery and video data. By establishing spatial awareness and causal understanding of environmental dynamics, such models will be implemented in embodied agents or robotic systems to execute a series of human-specified tasks and goals.
- *Representative contributions:* Multiple Object Tracking [ICCV'25 Submission], [AAAI'23 Poster], [CVPR'22 Poster], [TMM], Open-Vocabulary Tracking[ICLR'25 Poster], Referring Tracking [AAAI'24 Poster], Reasoning-based MOT [ICCV'25 Submission] and 3D object deteciton and tracking [IROS'24 Poster], [RA-L].

PUBLICATION HIGHLIGHTS

Google Scholar: https://scholar.google.com/citations?user=rWCQMNgAAAAJ&hl=en
Citations: **727+** | h-index: **12+** | i10-index: **13+** (until 20th, May, 2025)

- **E. Yu**, K Lin, L. Zhao, Y Wei, H. Wei, J. Sun, Z. Ge, X. Zhang, J Wang, W. Tao, "Unhackable Temporal Reward for Scalable Video MLLMs", in *International Conference on Learning Representations (ICLR)*, 2025.

  ➢ *Serving as the first to reveal the "anti-scaling law" phenomenon in current video MLLMs from a reinforcement learning perspective and establish a comprehensive theory of* ***temporal hacking****.*

- **E. Yu***, L. Zhao*, Y Wei, J. Yang, D. Wu, L. Kong, H. Wei, T. Wang, Z. Ge, X. Zhang, W. Tao, "Merlin: Empowering Multimodal LLMs with Foresight Minds", in *European Conference on Computer Vision (ECCV)*, 2024.

  ➢ *The groundbreaking model capable of generating natural language responses that are intricately linked with object trajectories. Merlin excels in predicting and reasoning about future events based on initial observations, showcasing an unprecedented capability in future prediction and reasoning.*

- **E. Yu**, K Lin, L. Zhao, J. Yin, Y Wei, H. Wei, J. Sun, C. Han, Z. Ge, X. Zhang, D. Jiang, J Wang, W. Tao, "Perception-R1: Pioneering Perception Policy with Reinforcement Learning", submitted to *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

  ➢ *Perception-R1 (PR1) proposes **perception policy learning** with rule-based RL and achieves comparable even better performance than expert models on mainstream visual and visual-language tasks. Notably, PR1 serves as the first pure MLLM to surpass 30 AP on challenging COCO2017 `val` set.*

- **E. Yu**, T. Wang, Z. Li, Y. Zhang, X. Zhang, W. Tao, "MOTRv3: Release-Fetch Supervision for End-to-End Multi-Object Tracking", submitted to *IEEE InternationalConference on Computer Vision (ICCV)*, 2025.

  ➢ *The first work successfully build fully end-to-end multiple object tracking model that outperforms the not end-to-end tracking model, without access to any extra detector or post-processing.*

- **E. Yu***, S. Liu*, Z. Li, J. Yang, Z. Li, S. Han, W. Tao "Generalizing Multiple Object Tracking to Unseen Domains by Introducing Natural Language Representation", in *Association for the Advance of Artificial Intelligence (AAAI)*, 2023.

  ➢ *The first multiple-object tracking model supporting vision-language modality inputs. Thanks to the domain invariant of natural language representation, LTrack achieves SOTA performance on our established cross-domain MOT benchmark.*

- **E. Yu***, Z. Li*, S. Han "Towards discriminative representation: Multi-view trajectory contrastive learning for online multi-object tracking", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

  ➢ *The first MOT scheme that adopts multi-view trajectory contrastive learning, in which each trajectory is represented as a center vector. By maintaining all the vectors in a dynamically updated memory bank, a trajectory-level contrastive loss is devised to explore the inter-frame information in the whole trajectories. MTrack surpassed preceding trackers and established new SOTA performance.*

- **E. Yu***, Z. Li*, S. Han, H, Wang, "RelationTrack: Relation-aware Multiple Object Tracking with Decoupled Representation", in *IEEE Transactions on Multimedia (TMM)*, 2022.

  ➢ *Serving as the first to point out the optimization conflict between detection and identity recognition in the joint detection and tracking framework, and proposed a set of feature decoupling schemes that effectively resolved this optimization conflict. Additionally, a relation-aware identity recognition block was designed, significantly enhancing the model's multi-object representation ability, achieving SOTA performance on MOT17 and MOT20.*

- L. Zhao*, **E. Yu***, Z. Ge†, J. Yang, H. Wei, H. Zhou, J. Sun, Y. Peng, R. Dong, C. Han, X. Zhang, "ChatSpot: Bootstrapping Multimodal LLMs via Precise Referring Instruction Tuning", Long Oral in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

  ➢ *ChatSpot is the pioneering MLLM framework in the field capable of precise region-aware understanding and reasoning. Based on GPT-4, we meticulously constructed MGVILD, a regional understanding dialogue dataset, contributing high-value data for precise region referring learning.*

- J. Yang*, **E. Yu***, Z Li, X Li, W Tao, "Quality matters: Embracing quality clues for robust 3d multi-object tracking", in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

  ➢ *The first pure camera-based 3D multi-object tracking solution to surpass 50% AMOTA, which achieves 51.1%, 54.8% and 56.6% AMOTA tracking performance on the nuScenes test sets with BEVDepth, VideoBEV, and StreamPETR models, respectively, which significantly reduces the performance gap between pure camera and LiDAR-based trackers.*

- J. Li*, **E. Yu***, S Chen, W Tao, "OVTR: End-to-End Open-Vocabulary Multiple Object Tracking with Transformer", in *International Conference on Learning Representations (ICLR)*, 2025.

  ➢ *The first fully end-to-end open-vocabulary multi-object tracking framework based on transformers.*

- S. Chen*, **E. Yu***, W. Tao, "Cross-View Referring Multi-Object Tracking", in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2024.

  ➢ *The first multi-view referring multi-object tracking benchmark and framework.*

Professional Services

Programme committee members or conference reviewers
NeurIPS (2023), CVPR (2022, 2023, 2024, 2025), ICCV (2023, 2025), ECCV (2024), ICLR (2025), ICML (2025), ACM Multimedia (2025), etc.
Journal reviewers
IEEE TRANSACTIONS ON MULTIMEDIA (TMM), IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), Neurocomputing.

Full List of Publications

*: Co-First Author, †: Project lead

Conference Papers (peer-reviewed, including workshops)

[C1] **E. Yu***, K Lin*, L. Zhao, Y Wei, H. Wei, J. Sun, Z. Ge, X. Zhang, J Wang, W. Tao, "Unhackable Temporal Reward for Scalable Video MLLMs", in *International Conference on Learning Representations (ICLR)*, 2025.

[C2] **E. Yu**, L. Zhao, Y. Wei, J. Yang, D. Wu, L. Kong, H. Wei, T. Wang, Z. Ge, X. Zhang†, W. Tao, "Merlin: Empowering Multimodal LLMs with Foresight Minds", in *European Conference on Computer Vision (ECCV)*, 2024.

[C3] L. Zhao*, **E. Yu***, Z. Ge†, J. Yang, H. Wei, H. Zhou, J. Sun, Y. Peng, R. Dong, C. Han, X. Zhang, "ChatSpot: Bootstrapping Multimodal LLMs via Precise Referring Instruction Tuning", Long Oral in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[C4] J. Li*, **E. Yu***, S Chen, W Tao, "OVTR: End-to-End Open-Vocabulary Multiple Object Tracking with Transformer", in *International Conference on Learning Representations (ICLR)*, 2025.

[C5] S. Chen*, **E. Yu***, J. Li, W. Tao, "Delving into the Trajectory Long-tail Distribution for Muti-object Tracking", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[C6] J. Yang*, **E. Yu***, Z Li, X Li, W Tao, "Quality matters: Embracing quality clues for robust 3d multi-object tracking", in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[C7] **E. Yu***, Z. Li*, S. Han "Towards discriminative representation: Multi-view trajectory contrastive learning for online multi-object tracking", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[C8] **E. Yu***, S. Liu*, Z. Li, J. Yang, Z. Li, S. Han, W. Tao "Generalizing Multiple Object Tracking to Unseen Domains by Introducing Natural Language Representation", in *Association for the Advance of Artificial Intelligence (AAAI)*, 2023.

[C9] **E. Yu***, Z. Li*, S. Han, H, Wang, "RelationTrack: Relation-aware Multiple Object Tracking with Decoupled Representation", in *IEEE Transactions on Multimedia (TMM)*, 2022.

[C10] Y. Wei*, L. Zhao*, K. Lin, **E. Yu**, Y. Peng, R. Dong, J. Sun, H. Wei, Z. Ge, X. Zhang, V. Patel, "Perception in Reflection", in *International Conference on Machine Learning (ICML)*, 2025.

[C11] S. Han, P. Huang, H. Wang, **E. Yu**, D. Liu, X. Pan, "MAT: Motion-Aware Multi-Object Tracking", in *Neurocomputing*, 2022.

[C12] Z. Li, C. Han, Z. Ge, J. Yang, **E. Yu**, H. Wang, H. Zhao, X. Zhang, "GroupLane: End-to-End 3D Lane Detection with Channel-wise Grouping", in *IEEE Robotics and Automation Letters (RA-L)*.

[C13] R. Zhou, W. Hua, L. Pan, S. Cheng, X. Wu, **E. Yu**, WY. Wang, "Rulearena: A Benchmark for Rule-guided Reasoning with LLMs in Real-world Scenarios", in *International Conference on Learning Representations (ICLR), 2025, Workshop*.

Preprints (including papers under peer review)

[R1] **E. Yu**, T. Wang, Z. Li, Y. Zhang, X. Zhang, W. Tao, "MOTRv3: Release-Fetch Supervision for End-to-End Multi-Object Tracking", submitted to to *IEEE International Conference on Computer Vision (ICCV)*, 2025.

[R2] Z. Zhu*, L. Zhao*, K. Lin, J. Yang, **E. Yu**, C. Liu, Z. Ge, X. Zhang, "PerPO: Perceptual Preference Optimization via Discriminative Rewarding", submitted to *International Conference on Machine Learning (ICML)*, 2025.

[R3] S. Chen*, Y. Yu*, **E. Yu***, W. Tao, "Reasoning-based Multiple Object Tracking", submitted to *International Conference on Computer Vision (ICCV)*, 2025.

[R4] E. Liu*, **E. Yu***, W. Tao, "Disentangling Instance and Scene Contexts for 3D Semantic Scene Completion", submitted to *International Conference on Computer Vision (ICCV)*, 2025.

Current Research Interests

Multimodal Large Language Models (MLLM):

- Pre-training: High-quality Synthetic Multimodal Data
- Post-training: Using SFT / RL technologies for better alignment
- Novel Modeling Paradigm: Diffusion maybe...

Multimodal Agent:

- Real-world Navigation Agent
- MLLM-based Game Agent

Spatial Intelligence:

- Spatial Perception, *e.g.*, spatial distance, of MLLM and visual foundation model.

Personal Hobbies

Movie, Singing, Reading, Ball Games, Swimming and Skiing.