Chao Yang

Postdoctoral Researcher – Shanghai Artifical Intelligence Laboratory
Xuhui District, Shanghai, P. R. China

+86 135-4120-4722 • yangchao9264@126.com
Homepage: emigmo.github.io

Working Experience

Shanghai AI Lab — **Shanghai, China**
Postdoctoral Researcher, Leading large model safety & decision group — *April. 2022 – Current*
- Alignment and Safety of Large Language Model, Multi-modal LLM.
- Robotic Manipulation and Embodied Intelligence.
- Co-workers: Prof. Yu Qiao, Dr. Jing Shao, Dr. Yu Liu.

Education Experience

Tsinghua University — **Beijing, China**
Aug.2018 – July.2022
- Ph.D. Student, In Department of Computer Science and Technology
- Advisors: Prof. Fuchun Sun

University of Hamburg — **Hamburg, Germany**
Aug.2019 – Sep.2019
- Visiting Student, In UHH · Department of Informatics
- Advisors: Prof. Jianwei Zhang

Tsinghua University — **Beijing, China**
Sep.2015 – Jun.2018
- M.Eng, In Computer Science and Technology
- Advisors: Prof. Fuchun Sun

Sichuan University — **Chengdu, China**
Sep.2011 – Jun.2015
- B.Eng, In Electronics and Information Engineering
- Advisors: Prof. Qingchuan Tao

Research Interests

My research interest includes Large Language Model Safety, Multi-modal Large Model, and Robotic Embodied Intelligence. Some of my current research keywords can be found below:
- **Large Language Model**: LLM Alignment and Fine-tuning, LLM Attack and Defense.
- **Multi-modal LLM**: Vision and language Fusion, Video-QA, VQA.
- **Embodied Robotics**: Robotic Manipulation, Reinforcement and Imitation Learning.

Publications (Google Scholar¶)

*(equal contribution), ‡(corresponding author)

Emulated Disalignment: Safety Alignment for Large Language Models May Backfire!

- Zhanhui Zhou, Jie Liu, Zhichen Dong, Jiaheng Liu, Chao Yang‡.
- In The 62nd Annual Meeting of the Association for Computational Linguistics (ACL2024 Main)

SEER: Facilitating Structured Reasoning and Explanation via Reinforcement Learning

- Guoxin Chen, Kexin Tang, Chao Yang‡.
- In The 62nd Annual Meeting of the Association for Computational Linguistics (ACL2024 Main)

Beyond one-preference-for-all: Multi-objective direct preference optimization

- Zhanhui Zhou, Jie Liu, Chao Yang‡.
- In The 62nd Annual Meeting of the Association for Computational Linguistics (ACL2024 Findings)

Safety of Multimodal Large Language Models on Images and Text

- Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang‡, Yu Qiao
- In *International Joint Conference on Artificial Intelligence (IJCAI 2024 Survey Track)*.

Attacks, defenses and evaluations for llm conversation safety: A survey

- Zhichen Dong*, Zhanhui Zhou*, Chao Yang‡, Jing Shao, Yu Qiao
- In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2024)*.

LLaMA-Excitor: General Instruction Tuning via Indirect Feature Interaction

- Chao Yang*, Bo Zou*, Yu Qiao, Chengbin Quan, Youjian Zhao
- In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*.

VideoDistill: Language-aware Vision Distillation for Video Question Answering

- Chao Yang*, Bo Zou*, Yu Qiao, Chengbin Quan, Youjian Zhao
- In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)*.

Critic-guided decision transformer for offline reinforcement learning

- Chao Yang*, Yuanfu Wang*, Ying Wen, Yu Liu, Yu Qiao
- In *AAAI Conference on Artificial Intelligence (AAAI 2024)*.

SpaceCLIP: A Vision-Language Pretraining Framework With Spatial Reconstruction On Text

- Chao Yang*, Bo Zou*, Chengbin Quan, Youjian Zhao
- In *Proceedings of the 31st ACM International Conference on Multimedia (ACM-MM 2023)*.

Sim2Real Object-Centric Keypoint Detection and Description

- Chao Yang*, Chengliang Zhong*, Fuchun Sun, Xiaodong Mu, Wenbing Huang.
- In *AAAI Conference on Artificial Intelligence (AAAI 2022)*.

Evaluations of the Gap between Supervised and Reinforcement Lifelong Learning on Robotic Manipulation Tasks

- Fan Yang, Chao Yang, Huaping Liu, Fuchun Sun.
- In *5th Conference on Robot Learning (CoRL 2021)*.

Graph Topography-Aware Reinforcement Learning for Intelligent Traffic Signal Control

- Chao Yang, Bo Zou, Wenbing Huang, Fuchun Sun, Huaping Liu.
- in *IEEE Int. Conf. on CYBER Technology in Automation, Control, and Intelligent Systems.(IEEE-Cyber 2021).* ***(Finalist Best Poster Award)***

Fault-Aware Robust Control via Adversarial Reinforcement Learning

- Fan Yang, Chao Yang, Di Guo, Huaping Liu, Fuchun Sun.
- in *IEEE Int. Conf. on CYBER Technology in Automation, Control, and Intelligent Systems.(IEEE-Cyber 2021).* ***(Best Poster Award)***

Adversarial Skill Learning for Robust Manipulation

- Chao Yang*, Pingcheng Jian*, Di Guo, Huaping Liu, Fuchun Sun
- In *IEEE International Conference on Robotics and Automation.(ICRA 2021)*.

Visual-Tactile Fusion for Robotic Stable Grasping

- Bin Fang, Chao Yang*, Fuchun Sun, Huaping Liu
- In *Industrial Robotics-New Paradigms* 2020.

Multi-Modal Continual Learning using Online Dictionary Updating

- Fuchun Sun, Huaping Liu, Chao Yang, Bin Fang
- In *IEEE Transactions on Cognitive and Developmental Systems* 2020.

Reinforcement Learning from Imperfect Demonstrations under Soft Expert Guidance

- Mingxuan Jing, Xiaojian Ma, Wenbing Huang, Fuchun Sun, Chao Yang, Bin Fang, Huaping Liu
- In *AAAI Conference on Artificial Intelligence.(AAAI 2020)*.

Imitation learning from observations by minimizing inverse dynamics disagreement

- Chao Yang*, Xiaojian Ma*, Wenbing Huang, Fuchun Sun, Huaping Liu, Chuang Gan, Junzhou Huang
- In *Advances in Neural Information Processing Systems.(NeurIPS 2019)*. **(Spotlight Oral)**.

Multimodal grasp dataset: A novel visual–tactile data set for robotic manipulation

- Tao Wang, Chao Yang, Frank Kirchner, Peng Du, Fuchun Sun, Bin Fang
- In *International Journal of Advanced Robotic Systems* 2019.

Predict Robot Grasp Outcomes based on Multi-Modal Information

- Chao Yang, Peng Du, Fuchun Sun, Bin Fang, Jie Zhou
- In *IEEE International Conference on Robotics and Biomimetics.(ROBIO 2018)*.

A dual-modal vision-based tactile sensor for robotic hand grasping

- Bin Fang, Fuchun Sun, Chao Yang‡, Hongxiang Xue. (‡Corresponding author)
- In *IEEE International Conference on Robotics and Automation.(ICRA 2018)*.

Robotic grasping using visual and tactile sensing

- Di Guo, Fuchun Sun, Bin Fang, Chao Yang, Ning Xi
- In *Information Sciences* 2017.

A Vision-Based Tactile Sensor with Tactile Detection Using Neural Network

- Chao Yang, Fuchun Sun, Bin Fang, Luxuan Li.
- In *International Conference on Cognitive Systems and Signal Processing.(ICCSIP 2016)*.

Manuscripts

Beyond one-preference-for-all: Multi-objective direct preference optimization

- Zhanhui Zhou, Jie Liu, Chao Yang‡, et al. *under review*.

Emulated Disalignment: Safety Alignment for Large Language Models May Backfire!

- Zhanhui Zhou, Jie Liu, Zhichen Dong, Jiaheng Liu, Chao Yang‡, et al. *under review*.

SEER: Facilitating Structured Reasoning and Explanation via Reinforcement Learning

- Guoxin Chen, Kexin Tang, Chao Yang‡, et al. *under review*.

Workshop

Rethinking Mutual Information for Language Conditioned Skill Discovery. (Link)

- Chao Yang, Zhaoxun Ju. *RSS 2023 Workshop: Articulate Robots: Utilizing Language for Robot Learning.*

RGB-D Object Segmentation for Multi-Step Pick-and-Place in Open Cloud Robot Table

- Chao Yang, Chengliang Zhong, Mingxuan Jing, Yu Luo, Tianying Ji, Yikai Wang, Wenbing Huang, Xiaodong Mu, Fuchun Sun. *ICRA 2021 Workshop: Cloud-Based Competitions and Benchmarks for Robotic Manipulation and Grasping.*

Internship Experience

Xiaomi AI Lab — **Beijing, China**
Research Intern — *Dec. 2019 – Jun. 2020*

- Research on Generative Adversarial Networks and Domain Adaptation.
- Mentor: Dr. Deli Zhao

Tencent AI Lab — **ShenZhen, China**
Research Intern — *Aug. 2018 – Jun. 2019*

- Research on Reinforcement Learning and Imitation Learning from Observation.
- Mentor: Dr. Wenbing Huang

Honers and Awards

Shanghai Postdoctoral Excellent Program — **Shanghai, China**
- For our research on multi-modal perception and robotic skill learning. — *Dec. 2022*

'84' Future Innovation Scholarship (RMB 100,000) — **Tsinghua, China**
- For our research on multi-modal perception and robotic skill learning. — *Dec. 2020*

Open Cloud Robot Table Organization Challenge(OCRTOC) — **Online, China**
- **First place** in simulation stage and third place in real stage. — *Nov. 2020*

Robotic Grasping and Manipulation Competition@IROS 2019 — **Macau, China**
- **First place** in robotic logistics track. — *Nov. 2019*

Robotic Grasping and Manipulation Competition@IROS 2017 — **Vancouver, Canada**
- **Third place** in robotic service track. — *Nov. 2017*

Robotic Grasping and Manipulation Competition@IROS 2016 — **Daejeon, Korea**
- **First place** in robotic automatic pick-and-place track. — *Nov. 2016*

Excellent Undergraduate Student — **Chengdu, China**
- For their excellent performance during the four years of college life. — *Jun. 2015*

China National Scholarship (Best Undergraduate Scholarship) — **Chengdu, China**
- Reward the top 1% students during undergraduate study (every year). — *Oct.2012 – Oct.2014*

Academic Services/Teaching

- Conference Reviewer: IROS20-24, ICRA21-23, CoR21-23, ICLR22-24, NIPS20-24, CVPR23

- Journal Reviewer: IEEE Robotics and Automation Letters (RA-L).
- Head TA and Part-time Instructor: Intelligent Computing and Robot Learning 2018-2021; Artificial Intelligence Theory 2020; System Analysis and Control 2019-2021 in Tsinghua University.