Junlin Yang web: https://yangjl2003.github.io/ yangjl22@mails.tsinghua.edu.cn

Statement of Purpose
of Junlin Yang

My research focuses on **natural language processing**, **multimodal learning**, **reinforcement learning**, and **human-AI interaction**, aiming to use machine learning to solve open-world tasks for humans and with humans. Currently, I'm focused on building **embodied agents**, particularly **digital** ones, that excel in solving human tasks and collaborating effectively. Within this overarching question, I have explored or am eager to explore the following topics:

1. Building stronger embodied agents in real-world environments: Building strong embodied agents involves data curation, training methods, and inference strategies.

a. **Data curation**: Well-formed supervised GUI trajectories are limited, but there is abundant multimodal data on computer usage online. [1] effectively converts this indirect knowledge into scalable supervision. Additionally, current datasets lack real-world long-horizon tasks, which we addressed with **AgentNet**, a diverse dataset of real-world computer usage scenarios.

b. **Training method**: Most training methods rely on behavior cloning (BC), but BC has limitations: (1) BC loss doesn't align with real-world loss, focusing on trajectory displacement rather than task completion. (2) Predicting the next step is challenging for GUI agents, let alone the entire trajectory. Reinforcement learning offers more potential. Recent works like [2] and [3] explore using RL to train Vision-Language Models (VLMs), but there is still much to explore.

c. **Inference method**: Reactive-style GUI agents describe observations, reason over instructions, and take actions, but this limits their exploration of possibilities and the consequences of their steps. [4] first explored tree search's impact on agent reasoning, highlighting the potential of incorporating search. [5] and [6] investigated integrating world models to improve understanding of actions' consequences. With the potential for inference-time scaling, there is much to explore in improving embodied agent inference methods.

2. Facilitating better Human-Agent interaction in open-world tasks: How can agents better understand and respond to human goals to improve instruction-following and collaboration? Current research on computer use from an instructional perspective is limited. However, embodied agents, whether robots or digital, rely on multi-turn interactions to complete complex tasks with humans. Thus, studying multi-turn interactions and the chain of instructions in it are valuable research directions, with key areas for exploration including:

a. **Ambiguous Human Intent Understanding**: Human instructions in real-world scenarios often exhibit ambiguity[7]. To tackle this problem, agents can leverage prior instructions to infer unclear intentions and improve task execution[8][9].

b. **Task Decomposition**: For larger, more macro-level human goals, if the agent can break them down into smaller, executable tasks and progressively deepen its understanding of the human goal through interaction, this could enhance the agent's ability to execute complex tasks.

c. **Predicting Human Actions**: Predicting the next human action can reduce interaction costs.

This research could drive advancements in agent-human collaboration for long-horizon tasks.

In addition to these areas, I am **open to exploring a broader range** of research directions in NLP, multimodal learning, reinforcement learning, and human-AI interaction.

References:

1. Ou, Tianyue, Frank F. Xu, Aman Madaan, Jiarui Liu, Robert Lo, Abishek Sridhar, Sudipta Sengupta, Dan Roth, Graham Neubig, and Shuyan Zhou. "Synatra: Turning Indirect Knowledge into Direct Demonstrations for Digital Agents at Scale." ArXiv abs/2409.15637 (2024).

2. Bai, Hao, Yifei Zhou, Mert Cemri, Jiayi Pan, Alane Suhr, Sergey Levine, and Aviral Kumar. "DigiRL: Training In-The-Wild Device-Control Agents with Autonomous Reinforcement Learning." ArXiv abs/2406.11896 (2024).

3. Qi, Zehan, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Xinyue Yang, Jiadai Sun, Yu Yang, Shuntian Yao, Tianjie Zhang, Wei Xu, Jie Tang, and Yuxiao Dong. "WebRL: Training LLM Web Agents via Self-Evolving Online Curriculum Reinforcement Learning." (2024).

4. Koh, Jing Yu, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. "Tree Search for Language Model Agents." ArXiv abs/2407.01476 (2024).

5. Gu, Yu, Boyuan Zheng, Boyu Gou, Kai Zhang, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, Huan Sun, and Yu Su. "Is Your LLM Secretly a World Model of the Internet? Model-Based Planning for Web Agents." (2024).

6. Chae, Hyungjoo, Namyoung Kim, Kai Tzu-iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. "Web Agents with World Models: Learning and Leveraging Environment Dynamics in Web Navigation." ArXiv abs/2410.13232 (2024).

7. Qian, Cheng, Bingxiang He, Zhuang Zhong, Jia Deng, Yujia Qin, Xin Cong, Zhong Zhang, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. "Tell Me More! Towards Implicit User Intention Understanding of Language Model Driven Agents." Annual Meeting of the Association for Computational Linguistics (2024).

8. Wan, Yanming, Jiayuan Mao, and Joshua B. Tenenbaum. "HandMeThat: Human-Robot Communication in Physical and Social Environments." ArXiv abs/2310.03779 (2023).

9. Wan, Yanming, Yue Wu, Yiping Wang, Jiayuan Mao, and Natasha Jaques. "Infer Human's Intentions Before Following Natural Language Instructions." ArXiv abs/2409.18073 (2024).