Statement of Purpose

Haofei Yu *haofeiy@cs.cmu.edu*

Imagine a world where AI agents can effortlessly assist us in daily tasks, like buying shoes online, finding lost keys at home, and offering companionship in tough times. This advanced help promises to greatly improve our life's convenience and quality. Developing such practically useful AI agents is challenging due to the multimodal nature of the real world. In human tasks, we interact with our environment, but modeling and quantifying these complex multimodal interactions is still a complex task. Inspired by this vision of AI facilitating everyday tasks, my research interest is **building a multimodal agent for real-world applications**. My overarching research aim is to bridge the real world with pragmatic AI agent designs, striving to create an agent endowed with social, digital, and physical intelligence. To construct such an agent, I identify 4 principal challenges: the first two involve defining the agent's sensory inputs and behavioral outputs, the third focuses on its essential memory component, and the fourth addresses how to teach the agent to behave like humans.

Challenge1 Multimodal Fusion: Enable the agent to integrate and interpret varied types of data from the real world to understand complex communication forms, including sarcasm and deception.
Challenge2 Decision Making: Enhance the agent's capability to evaluate and rank potential actions based on predicted outcomes, ensuring contextually relevant and effective decision-making in real-world scenarios.
Challenge3 Memory Management: Create a dynamic memory system for the agent that adapts to new information while retaining essential past experiences, mirroring the evolving nature of the real world.
Challenge4 Interactive Learning: Equip the agent to learn from human's interactive experiences, emulating human social, physical, and digital behavior learning processes.

I am grateful to have had the opportunity to conduct preliminary research under the esteemed guidance of Professors Ruslan Salakhutdinov, Louis-Philippe Morency, Graham Neubig, Yue Zhang, and Zhenzhong Lan. In addition, my industrial experiences at **Apple** and **Tencent** have provided me experience to deal with real world user data. I am deeply passionate about continuing exploring these challenges throughout my Ph.D. studies.

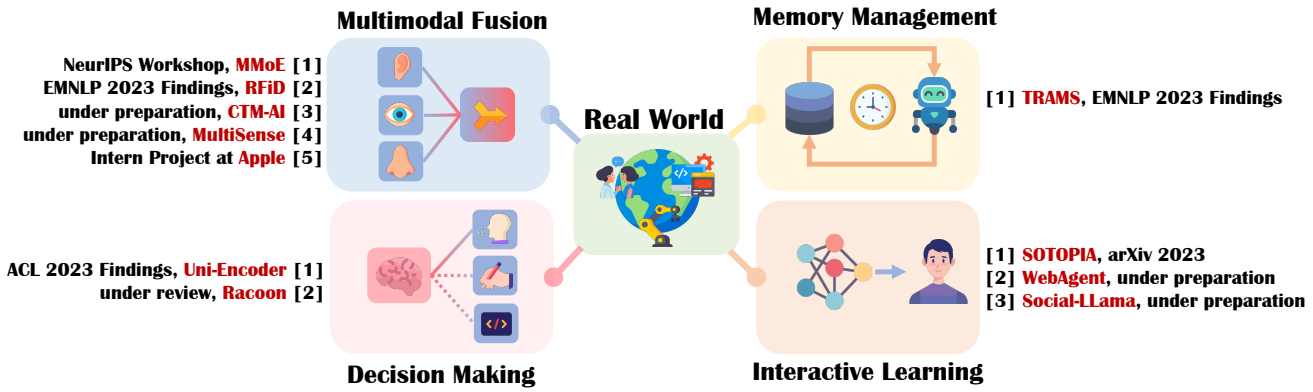

Figure 1: An overview of the 4 challenges in constructing a multimodal agent for real-world applications, accompanied by a snapshot of my current research progress on each of them, including ongoing work.

1 Progress on Challenge1: Multimodal Fusion

Multiple multimodal processors can be fused based on their interaction types. The key of multimodal fusion is to discuss how to handle interaction between modality information from multiple processors. Starting from *unimodal* fusion, I proposed RFiD [2] and proved that increasing interaction between positive data pairs and decreasing interaction between positive-negative data pairs helps text understanding. Subsequently, during my internship at Apple, I proposed a hierarchical prompting framework that groups and combines the summaries from multiple LLMs. Combining contradictory summaries and answers of different aspects of one questions provides a more satisfying answer to ambiguous and challenging questions. Extending it to *multimodal* data, the multimodal fusion process is more complicated due to diverse interaction types. Motivated by [3], interactions between different modalities can be quantified based on agreement, disagreement, synergy, redundancy, and uniqueness. Therefore, I propose MMOE [6], multimodal mixture of experts, to help separately handle different types of multimodal interaction and get 20%+ performance gain on the detected disagreed synergy part in the MUStARD [1] dataset. To push one step further, I am conducting Multisense: a multimodal foundation model that incorporates heterogeneous data including IoT, vision, language, and music into single language model outputs based on multimodal adapters that implicitly learn interactions in multimodal data.

2 Progress on Challenge2: Decision Making

Teaching agents to learn ranking benefits its decision making process. To enhance an agent's decision-making capabilities, particularly in generating what is aligned with intended outcomes, we teach it to learn to rank from bad ones to good ones. My collaboration with Prof. Zhenzhong Lan led to the development of the Uni-Encoder [4], which adeptly managed response ranking and demonstrated superior performance across 4 datasets. This achievement prompted a further question is whether it is possible to seamlessly integrate ranking functionalities without the necessity for a separate ranker. Pursuing this question, I proposed RACOON [5] that focused on code generation tasks. Our findings affirm that with the support of contrastive ranking loss, a generative model can indeed be trained to simultaneously rank and generate, employing a multi-task learning approach.

3 Progress on Challenge3: Memory Management

Training-free memory management is useful for long-range tasks. Building on the research into the fusion and generation processes, equipping an agent to perform real-world tasks necessitates the inclusion of memory as a core system component. It is crucial to design automatic memory updating strategies that function independently of the agent's intentions to cater to various tasks. Driven by this understanding, I uncovered that an automatic, query-independent approach to memory selection could significantly bolster performance on language modeling benchmarks, such as the enwik8 and wikitext-103. This method, called TRAMS [7], provides important perspectives on how to implement an automatic memory selection mechanism to manage working memory efficiently.

4 Progress on Challenge4: Interactive Learning

Interactive Learning helps agents learn task-specific skills. In-context learning is considered as basic methods for developing artificial intelligence agents, as demonstrated in the SOTOPIA [8] environment for simulating complex social interactions. In-context learning has shown that agents like GPT-4 can communicate in scenarios such as negotiation and collaboration, but struggle with social commonsense and strategic skills. To overcome these limitations, supervised training with human-generated web navigation trajectories has been utilized, allowing agents to learn and execute sub-tasks in web navigation. However, acquiring high-quality data for supervised training remains challenging. RL offers an alternative by creating an observation-feedback loop with rewards during environmental interactions, generating ample training data. This approach has been applied in SOTOPIA, using a GPT-4-based reward function, to enhance agents' general social skills, showing promise in adapting to unseen scenarios.

5 Future Plan

Based on my research interest, there are 4 main research directions that I am eagerly working on:

Grounded NLP My interest lies in integrating language agents into *grounded* or *embodied* environments. Such integration enables language agents to perform robotics or other grounded tasks. Moreover, I am keen on exploring more complicated tasks like *grounded social behaviour*. Typically, expanding language agents into a simulated room and the target social goal would be asking two embodied agents collaboratively finding an item in the room.

Social NLP LLMs have demonstrated remarkable proficiency in handling complex tasks across diverse scenarios. An intriguing area of research is exploring and enhancing the social capabilities of these LLMs. This includes devising training methodologies focused on acquiring strategic social skills such as negotiation, collaboration, and maintaining confidentiality. The goal is to understand and improve the ways in which these models can effectively mimic and engage in socially nuanced interactions.

Multimodal Learning Instead of projecting multimodal observation into text and utilize the power of LLM, another way is to develop pretrained *multimodal foundation model* capable of handling all real-world modalities including understudied modalities like thermal and touch. Exploring training mechanism specially designed for handling a large number of multimodal pretraining like *MoE-based* multimodal training is a super exciting direction.

Safety and Robustness When talking about using agents in our life, it is important to figure out testing method to guarantee its *safety*, particularly in areas where AI decisions directly influence physical aspects, like online shopping and code execution. Moreover, finding ways to provide safeguarding against *adversarial attacks* and guaranteeing consistent performance in various scenarios would be crucial as well.

References

[1] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards Multimodal Sarcasm Detection. *ACL*, 2019.

[2] Cunxiang Wang*, **Haofei Yu***, and Yue Zhang. RFiD: Towards Rational Fusion-in-Decoder for Open-Domain Question Answering. *Findings of ACL*, 2023.

[3] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling feature interactions: An information decomposition framework. *NeurIPS*, 2023.

[4] Chiyu Song*, Hongliang He*, **Haofei Yu**, Pengfei Fang, Leyang Cui, and Zhenzhong Lan. Uni-encoder: A fast and accurate response selection paradigm for generation-based dialogue systems. *Findings of ACL*, 2023.

[5] **Haofei Yu**, Uri Alon, and Graham Neubig. Racoon: Ranked Code Generation. ***Currently Under Preparation***, 2023.

[6] **Haofei Yu**, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal Mixture of Experts. *NeurIPS UniReps Workshop*, 2023.

[7] **Haofei Yu***, Cunxiang Wang*, Yue Zhang, Wei Bi, et al. TRAMS: Training-free Memory Selection for Long-range Language Modeling. *Findings of EMNLP*, 2023.

[8] Xuhui Zhou*, Hao Zhu*, Leena Mathur, Ruohong Zhang, **Haofei Yu**, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. *arXiv*, 2023.