Yunyi Shang: Research Statement

My research interests converge at the intersection of natural language processing, computer vision, human-computer interaction, cognitive computing, and reinforcement learning. My research goal is to enhance the comprehension of natural intelligence by leveraging machine intelligence, utilize insights from natural intelligence to advance machine intelligence technology, and leverage these advancements to develop safer and more foresighted applications. Below are some of the research directions I plan to pursue.

Multimodal Data Sources. In my past research, I focused on the fields of natural language processing, computer vision and computer graphics, realizing the unique advantages of language and images as primary perceptual input modalities for humans. Language, as a crucial tool for human learning and communication, carries rich semantic information, while images captivate interest due to their intuitive and complex nature. The complementary nature between language and images can generate a synergistic effect, making the overall information more comprehensive and profound. For instance, when textual descriptions mention "a red car," computer vision systems can target the location of a red car more precisely, enhancing object recognition accuracy. Exploring the collaborative effects between different perceptual modalities can lead to new cognitive understandings and information processing methods. By studying multimodal data, I aim to achieve an effect where 1+1>2. Additionally, signals such as sound, touch, and even physiological signals like brainwaves are rich sources of daily-life information. Through in-depth exploration of the cooperative relationships among these signal sources, I aim to construct more comprehensive multimodal models that approach the way humans process data, achieving a deeper understanding of information.

Interactive Scenarios. Taking child language acquisition as an example, human learning processes are often not isolated but full of interactivity. Rich implicit and explicit feedback signals are embedded in interactions, and utilizing these signals can lead to better training models. Children, when interacting with adults, peers, or their environment, receive feedback through language, actions, and facial expressions, containing contextual information and emotional factors. Such signals are crucial for children learning language, understanding object functionalities, and even forming moral values. By introducing elements of interaction, models can better adapt to changes in various scenarios and user behaviors. For example, in child language acquisition, through conversations and interactive games, children more easily grasp the usage of vocabulary and grammar rules. Introducing this interactivity into multimodal data processing allows models to adapt more flexibly to diverse real-world applications, enhancing model generalization and pioneering innovative and practical human-computer interaction methods.

Cognitive, Reasoning, Interpretability. When humans process diverse data sources, they employ various cognitive models. These models can provide valuable insights for designing multimodal data processing systems. For instance, the "Chain of Thought" model describes the human thought process, and we can draw on its information transmission methods to optimize the information flow in multimodal data processing. The "Counterfactual Model" helps predict possible scenarios, providing the system with more robust reasoning capabilities. "Theory of Mind" allows humans to infer others' intentions by observing their behavior, which is crucial for understanding user needs and intentions in interactive systems. By introducing principles from cognitive science and reasoning models, I aim to ensure that the model's behavior is transparent and interpretable. Transparency refers to the clarity of the model's decision-making and learning processes, while interpretability means understanding why the system makes specific decisions or learns certain patterns. This transparency and interpretability are critical to ensuring the safety and controllability of system behavior, as well as to promote trust between the model and users.

Reinforcement Deep Learning. I believe that the true value of a model lies not only in its performance in familiar scenarios but also in its ability to explore the uncertainty space, demonstrating genuine generalization capabilities. While deep learning has made significant strides in addressing various challenges in artificial intelligence, such as pattern recognition, by emulating human activities like vision, hearing, and cognition, it fundamentally imitates human performance through surface distribution. In contrast, reinforcement learning, through exploration in the environment and learning from experience, exhibits better generalization capabilities. In reinforcement deep learning, models adjust their strategies by interacting with the environment, utilizing reward and punishment signals to adapt to unknown and complex tasks. By introducing reinforcement learning into multimodal scenarios, specifically in few-shot scenarios, exploring uncertainty space through interaction and reasoning enables the model to better understand the correlation between different perceptual modalities. This learning approach allows the model to adapt more effectively to new domains and situations, showcasing higher flexibility.

Application. In the field of multimodal perception, its integrative nature offers tremendous potential in various application domains, especially in intelligent assistants, AGI robots, virtual/augmented reality (VR/AR), and healthcare. For example, Considering the realm of smart homes, instructions like "dim the living room lights" indicate the agent's simultaneous comprehension of both the language input and the spatial environment (such as identifying the living room and the lighting control system). Throughout the execution process, visual input plays a pivotal role, enabling the agent to recognize obstacles, strategize navigation paths, and provide visual feedback to confirm successful completion. Multimodal perception empowers intelligent agents to comprehend the surrounding environment more comprehensively and intricately, providing more intelligent and human-centric services. Drawing upon my experience in virtual environment development and robotics, I aim to contribute to the field of multimodal perception research by enabling immersive, natural, and intuitive user experiences through interactions with virtual or physical agents.

In summary, my research spans multiple domains to explore the synergies between natural and machine intelligence. By leveraging multimodal data sources, employing interaction as a pivotal training scenario, integrating cognitive models for effective data fusion, and utilizing reinforcement learning with reward and penalty signals to adjust strategies, I aims to elevate user experience, foster safety, and empower applications with robust generalization capabilities.