Wanqi Zhong — Personal Statement for PhD in Embodied Artificial Intelligence

I am Zhong Wanqi, a student at Harbin Institute of Technology majoring in Computer Science and Technology. My research interests focus on vision-centered multimodal perception, with a passion for applying this to assist robot motion and manipulation. And my research goal is to develop active intelligent agents capable of understanding the world as comprehensively as humans, while moving and manipulating objects freely in the physical world. Driven by a pursuit of an embodied AI future, I am gradually approaching this goal through the following research experience and plans.

The starting point for active intelligence is a multidimensional understanding of the rich semantic information in the world. I aimed to link large language models with powerful visual models which excel at open-vocabulary segmentation and classification. By utilizing large language models as the ”brain” for unified information processing, this approach compensates for the lack of extensive common sense in traditional visual models, thereby helping to achieve a better understanding of the world.

The first research project I participated in, LMEye, represents my initial exploration of visual information extraction in this approach. After aligning the visual encoder hidden state to the semantic space of the language model, the language model can understand the information contained in the images. However, during the actual perception process, attention needs to be paid to specific details in different locations. At this time, the global static visual information is insufficient to provide information corresponding to the instructions, introducing the problem of hallucinations in the responses. Therefore, in this project, interacting with the image again based on input instructions was implemented to re-extract specific image information as needed, effectively improving the performance of VQA.

To better exploit the advantages of LLM as a “brain”, I further expanded the input modalities to promote comprehensive understanding of world information. In the Uni-MoE project, using image and speech encoders to process multimodal inputs, I trained multiple experts, each capable of efficient processing of single modality. Then integrating them through a MoE architecture that distributes different experts for various modal inputs. Through the design of auxiliary loss, I guided the router to select the top 2 most suitable experts to simultaneously participate in the cross-modal understanding of modal data. Compared to traditional Socratic models, encoding different modalities into a unified semantic space demonstrates more potential for cross-modal understanding. The Uni-MoE model can effective integrate video background sounds and frame information to respond to instructions by leveraging the strengths of different experts.

Through the above-mentioned scientific research experience, I have gradually explored visual information extraction and cross-modal universal understanding frameworks for world understanding, and in subsequent explorations, I am committed to finding solutions to the difficulties in the actual deployment of embodied intelligent robots. I found the defects in storage and retrieval of visual information within models result in severe hallucinations, hindering their use in real-world scenarios. So I utilized an additional visual memory unit to integrate the extraction and storage of original visual information, supervising its information retrieval capability in a visualized manner. Additionally, by segregating visual information extraction and text generation into separate units, I facilitated a more intuitive comparison of their respective utilization of image information and language knowledge in actual responses.

At the same time, the high latency characteristic of transitioning from understanding to

manipulation with large language models is not suitable for highly interactive scenarios like real-time human-machine interaction. During this period, the brain's dual-system mechanism was applied, with the large model acting as the slow system for human intention understanding and semantic map related information construction, while the fast system issued intuitive low-level control commands, thus hiding the response delay of the slow system. Additionally, a unified lower framework was designed, enabling both to adjust plans in the same cost map, which reduced the gap between the two planning spaces. strongly promoting the deployment of real-time active intelligence.

Throughout the aforementioned research endeavors, I continuously pondered the potential for optimizing various processes from perception to planning to manipulation. Therefore, my future research plans will focus on how to distill the specific knowledge needed in actual manipulation scenarios from LLMs, promoting the effective combination of LLMs with visual perception and task planner. Additionally, better integrating the intermediate spatial representations of task planner and manipulation is also a key focus of the research.

Visual Information Extraction One of my research focuses is constructing a visual information extractor that is rich in both detail and semantics. More and more works try to enhance the extraction of image information in visual language models, including improving visual understanding by replacing higher resolution visual encoders and integrating expert models to convert visual information into text form.

From my perspective, the approach of replacing visual encoders mainly increases the amount of information accessed by the visual encoders to help improve image understanding. However, CLIP-like models do not pay sufficient attention to details during training, and the hidden states contains more holistic understanding of images, posing challenges for subsequent grounding tasks. While using expert models can get the textual output of fine-grained segmentation which provide precise coordinates for large models. However, the gap between understanding and manipulation processes means that errors recognized by large models cannot effectively be fed back to the expert models for adjustments, hindering more attempts. Also, cumulative errors are inevitable if there are biases in the first extractor block.

So my research plan involves using a self-regressive generation method to call dual encoders to achieve both scene understanding and fine-grained understanding simultaneously. CLIP's global encoder will still be used for general understanding of the scene. Meanwhile, specific tasks and global perception information will be combined to output tokens in a self-regressive manner, requiring the extraction of specific area information. This way, a finer-grained encoder can be used to extract specific area information according to the output token, achieving in-depth understanding of both overall and fine-grained scene information. At the same time, it is also necessary to enhance fine-grained encoders for rich object descriptions, ensuring a deeper understanding and representation of different states of the same object across various scenarios, thereby better facilitating planning processes.

Intermediate Representation for Planning and Manipulation. After accurately perceiving and understanding the scene, task completion still requires a Task Planner and Motion Planner. Recent works use large language models as Task Planner to analyze and compute object information in the scene, ultimately encoding and calling parameterized basic skills. Meanwhile, Motion Planners often utilize RL methods to learn basic skills or generate corresponding paths using path planners. However, there is still a gap between the two, as the discontinuity in calling parameterized basic skills and the limitations in basic skill settings hinder the application of powerful skills learned by RL methods such as high-speed motion and

obstacle avoidance.

Consequently, the focus of my research is the connection between the two, requiring the definition of a better intermediate state representation that allows planners to accurately describe planning schemes and fully mobilize the flexibility of RL. One possible approach may be place online rewards based on the scene and requirements. The planner uses a tree structure to plan different schemes, assigning costs and rewards associated with tree depth in a three-dimensional scene, thus providing step-by-step execution clues for the Motion Planner. Online penalties for unreasonable behaviors are applied while encouraging the use of free execution methods by the lower-level control to obtain rewards. In simulation scenarios, rewards can be judged based on contact; in real scenarios, auxiliary tracking visual models can provide low-latency validation of execution effectiveness.

Reducing Latency and Specializing Knowledge. In further deployment, the latency problem caused by directly calling large models leads to unsuitability in actual scenarios. So my research plan in this part will attempt to extract the most relevant parts for robot manipulation from large models using distillation methods, achieving specialization in manipulation common sense. By defining knowledge areas such as physics, kinematics, and scene understanding, manipulation-related information will be extracted to obtain small models capable of fully understanding manipulation and motion scenarios, promoting the deployment of embodied intelligent robots. This allows to expand the collection of real-life data, continuously iterate and optimize small models by analyzing common skills, and promote the improvement of proficiency in frequent scenarios. Based on such an iterative process, it helps to further learn more skills on this basis.

Therefore, based on the aforementioned research plans, I will continue to explore these directions with unwavering dedication. I am genuinely eager to join IIIS, where I hope to contribute to propelling embodied intelligence into a rapidly advancing future.