STATEMENT OF PURPOSE

Wenhao Chai
wchai@uw.edu

Since childhood, I have been deeply engaged with mathematics, a passion that blossomed as I consistently participated in math competitions and modeling contests from elementary school through university. This early exposure not only honed my analytical skills but also paved the way for my entry into deep learning. I vividly remember tackling a hackathon challenge focused on semantic segmentation of remote sensing images. Unable to solve it with traditional methodologies, I self-taught convolutional neural networks and successfully clinched the competition.

Initially, my research was centered on multimodal learning; I was captivated by the potential to integrate and interpret data from diverse sources, envisioning systems that could surpass human capabilities in sensory integration and decision making. Recently, large multimodal models (LMMs), which integrate pre-trained large language models (LLMs) with vision encoders, have exhibited human-like visual intelligence across various tasks. These models demonstrate strong generalization capabilities, handling diverse data inputs such as video. Furthermore, using these large foundational models to build agents is also a promising direction, making it possible achieving intelligence beyond human capabilities. This rapid progress has fueled my interest in two primary research directions:

I. What challenges arise in building more powerful LMMs for video? And how can we more effectively and accurately evaluate the capabilities of LMMs on video tasks?
II. How to build an embodied agent system with LMMs and LLMs?

My research to date has explored these questions, resulting in several papers (1; 2; 3), which partially address these topics. In this statement, I elaborate on my research focus and future plans.

LMMs for Video LMMs have made substantial advancements in many visual tasks. But it introduces new challenges when extending to video. It requires models to handle more complex temporal information and much longer input token sequences. My current research focuses on improving the efficiency, which is a critical bottleneck for scaling LMMs to real-world applications.

Existing LMMs often use ViT to encode visual information into tokens, which serve as input prefixes to LLMs. The number of visual tokens depends on the patch size p_s, resolution s, and sampled frames f, approximately $\frac{s^2}{p_s^2} \times f$. For instance, a 15-second 360p video sampled at 1 frame per second takes over 20,000 tokens. This large input size complicates training and inference and risks losing focus on key information in lengthy sequences. AuroraCap (1) addresses this by reducing visual tokens to 10% and even 1% of the original using a bipartite soft matching algorithm to merge similar tokens in each ViT layer, achieving minimal performance loss across tasks.

While AuroraCap achieves an excellent performance-efficiency trade-off for video tasks, it struggles with long videos exceeding 10 minutes. To address this, we propose MovieChat (2), the first LMM capable of processing over 10,000 frames. MovieChat employs a token-level long-short term memory mechanism and a sliding window approach. Video features are sequentially fed into a fixed-length short-term memory, with older tokens consolidated into long-term memory. The resulting video representation is projected and input into an LLM for user interaction. Both AuroraCap and MovieChat are training-free, end-to-end models, with performance improving during further training.

Evaluation Benchmark For any machine learning task, the benchmark is undoubtedly one of the most important components. We have established the first benchmark for these two novel tasks: VDC (1) video detailed captioning and MovieChat-1K (2) for long-form video understanding.

VDC features diverse videos with averaging 500 words captions - far exceeding the 10-word average in previous benchmarks. Captions are categorized into short, camera, background, main object, and detailed to provide a comprehensive evaluation. Standard n-gram metrics like CIDEr struggle with detailed captions. Therefore, we propose VDCscore, a novel metric, decomposing ground-truth captions into concise Q&A pairs, evaluates generated responses, and scores accuracy via LLMs.

MovieChat-1K consists of over 1,000 high-quality video clips from various movies and TV series, with 14,000 manually annotated question-answer pairs, capturing key moments throughout hour-long videos. We found that almost all LMMs struggle to understand such long-form video inputs. They are typically trained on videos only a few seconds long, resulting in an accuracy of only 50% accuracy. Based on the MovieChat-1K benchmark, we organize a workshop at CVPR 2024 to encourage more people to participate. We found that although people can use a divide-and-conquer approach, breaking the video into smaller segments for understanding and then summarizing, developing an end-to-end LMMs for long-form video remains a challenging research topic.

Embodied Agent in Minecraft Building embodied agent systems with LLMs and LMMs offers promising potential, but real-world deployment is costly and complex. We develop our agent, STEVE (3), in Minecraft, comprising vision perception, language instruction, and code action. Vision perception interprets environmental visuals, integrated with the agent state in the LLM. Language instruction enables iterative reasoning and task decomposition, while code action generates executable skills from a skill database for effective interaction. In follow-up work, we explore multi-agent systems with a hierarchical framework for coordinating agents on complex tasks. Additionally, we replace agent frameworks with a single LMM, demonstrating through self-distillation and expert knowledge distillation that a well-trained single model can match agent system performance.

Future Research Plan I believe it is possible to surpass human-level intelligence by combining different foundation models. And I also believe in sustainable, project-oriented research that goes beyond producing isolated publications. To promote reproducibility, my projects are fully open-sourced, earning over 4,000 stars on GitHub. Moving forward, I aim to advance efficient LMMs and explore their potential in embodied tasks, serving as the perceptual foundation for agents interacting with the physical environment. My research will continue to prioritize efficiency, scalability, and usability. And I truly value the importance of producing high-quality projects.

Future Career Plan Upon completing my Ph.D., I aspire to pursue a faculty position where I can lead independent research and build a collaborative team focused on advancing AI tasks. Academia provides the intellectual freedom to explore the cutting edge while offering opportunities to mentor the next generation of researchers, fostering long-term contributions to the broader AI community.

REFERENCES

[1] **Chai, Wenhao**†, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jeng-Neng Hwang, Saining Xie, and Christopher D. Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024. 1

[2] Enxin Song, **Chai, Wenhao**†, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 1

[3] Zhonghan Zhao, **Chai, Wenhao**†, Xuan Wang, Boyi Li, Shengyu Hao, Shidong Cao, Tian Ye, and Gaoang Wang. See and think: Embodied agent in virtual environment. In *European Conference on Computer Vision*, pages 187–204. Springer, 2025. 1, 2