STATEMENT OF PURPOSE

Wenhao Chai
wchai@uw.edu

In recent years, 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 for machines to achieve 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; 4; 5), 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 image understanding. Extending these models to video is a natural progression, but it introduces new challenges. Video processing requires models to handle more complex temporal information and much longer input sequences. My current research focuses on improving the *efficiency* of video-based LMMs, which is a critical bottleneck for scaling these models to real-world applications.

Existing LMMs typically employ ViT to encode visual information into tokens, which are then used as a prefix input to the LLM. The number of visual tokens is related to the patch size p_s, the resolution s, and the number of sampled frames f. Its approximate value can be expressed as $\frac{s^2}{p_s^2} \times f$. Simply put, even for a 15-second 360p video, if we sample just one frame per second as input, the number of visual tokens would exceed 20,000, which often surpasses the word count of a paper published at NeurIPS. Such a large input volume poses challenges for both model training and inference. More importantly, we cannot guarantee that the model will still be able to focus on the key parts within a long sequence of visual tokens. To this end, our work AuroraCap (2) can reduce the number of visual tokens to as low as 10% or even 1% of the original while incurring only fully acceptable performance loss across various tasks. To be specific, we gradually combine similar tokens in each ViT layer using a bipartite soft matching algorithm to reduce the number of visual tokens.

Although our proposed AuroraCap achieved the best performance-efficiency trade-off on video tasks, such a model is still insufficient for inference on long videos exceeding ten minutes or even an hour. To address this, we propose MovieChat (1), the first LMM capable of processing over 100,000 frames. We introduced token-level long-short term memory mechanisms into the inference process. Using a sliding window approach, video features are extracted and represented as tokens, which are sequentially fed into the short-term memory frame by frame. The short-term memory has a fixed length, and when it reaches its limit, the earliest tokens are popped and consolidated into the long-term memory. After passing through a projection layer, the video representation is then inputted into a large language model for interaction with the user. AuroraCap and MovieChat are training-free and end-to-end. Additionally, during further training, we observed performance improvements.

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 (2) video detailed captioning and MovieChat-1K (1) for long-form video understanding.

VDC is the first ever benchmark for video detailed captioning. We sampled videos from diverse sources and annotated them with captions averaging up to 500 words. In contrast, previous benchmarks typically had an average of only 10 words. We split the captions into five categories: short, camera, background, main object, and detailed, to provide a more comprehensive evaluation. In previous video tasks, we could typically find straightforward evaluation methods. For instance, in the video object segmentation task, we provide object segmentation masks and then calculate IoU or other deterministic metrics. For the previous video captioning task, we calculate metrics like CIDEr, which are based on term frequency. While these metrics are not perfect, they are sufficient for use. However, we observe that the LLM-based evaluation metric still struggles to differentiate the quality of detailed captions and tends to give lower scores. To address these challenges, we propose VDCscore, a novel captioning evaluation metric that leverages the reliability of LLMs by evaluating short visual question-answer pairs. We first decompose the ground-truth caption into a set of concise question-answer pairs using LLMs, then generate corresponding responses from the predicted caption. Finally, the LLM is used to assess the accuracy of each response to provide an overall score.

MovieChat-1K consists of 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 around 50% accuracy. Based on the MovieChat-1K benchmark, we organized a workshop[1] 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 designed for long-form video remains a challenging research topic.

Embodied Agent in Minecraft Building an embodied agent system with LLMs and LMMs as its core is a promising direction. Due to the significant costs and uncontrollable factors associated with deploying and training such agents in the real world, we start to build our agent in Minecraft, STEVE (5). It comprises three key components: vision perception, language instruction, and code action. Vision perception involves interpreting visual information in the environment, which is then integrated into the LLMs component with agent state and task instruction. Language instruction is responsible for iterative reasoning and decomposing complex tasks into manageable guidelines. Code action generates executable skill actions based on retrieval in skill database, enabling the agent to interact effectively within the Minecraft environment. Furthermore, we extend the single-agent framework to multiple agents in our follow-up work (3). We design the hierarchical architecture framework to automatically organize groups of agents for complex tasks. Finally, we explored the possibility of abandoning the complex agent framework in favor of using a single LMM (4). Through self-distillation and expert knowledge distillation, we demonstrate that, with sufficient training, a single model can achieve performance comparable to that of an agent-based system.

Generative Models and AI for Applied Science Before delving into LMMs and embodied agent, my research interests also include generative models for video (6) as well as human (7; 8) and animal (9) pose estimation. Additionally, I have experience applying deep learning techniques in other applied science area like civil engineering (10; 11) and medical imaging (12).

Future Research Plan I believe in sustainable, project-oriented research that goes beyond producing isolated publications. To promote reproducibility, all of my projects are fully open-sourced, earning over 2,400 stars on GitHub[2]. Moving forward, I aim to advance efficient LMMs for video 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 in real-world applications. I truly value the importance of producing high-quality projects, and I will make it my focus moving forward.

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 LMMs and AI for embodied tasks. Academia provides the intellectual freedom to explore the cutting edge of my field while offering opportunities to mentor the next generation of researchers, fostering long-term contributions to the broader AI community.

[1]CVPR 2024 workshop website: https://sites.google.com/view/loveucvpr24/track1
[2]GitHub profile: https://github.com/rese1f

References

[1] 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

[2] **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

[3] Zhonghan Zhao, Kewei Chen, Dongxu Guo, **Chai, Wenhao**†, Tian Ye, Yanting Zhang, and Gaoang Wang. Hierarchical auto-organizing system for open-ended multi-agent navigation. *The International Conference on Learning Representations LLM Agents Workshop*, 2024. 1, 2

[4] Zhonghan Zhao, Ke Ma, **Chai, Wenhao**†, Xuan Wang, Kewei Chen, Dongxu Guo, Yanting Zhang, Hongwei Wang, and Gaoang Wang. Do we really need a complex agent system? distill embodied agent into a single model. *arXiv preprint arXiv:2404.04619*, 2024. 1, 2

[5] 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

[6] **Chai, Wenhao**, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023. 2

[7] **Chai, Wenhao**, Zhongyu Jiang, Jenq-Neng Hwang, and Gaoang Wang. Global adaptation meets local generalization: Unsupervised domain adaptation for 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14655–14665, 2023. 2

[8] Yuan-Hao Ho, Jen-Hao Cheng, Sheng Yao Kuan, Zhongyu Jiang, **Chai, Wenhao**, Hsiang-Wei Huang, Chih-Lung Lin, and Jenq-Neng Hwang. Rt-pose: A 4d radar tensor-based 3d human pose estimation and localization benchmark. In *European Conference on Computer Vision*. Springer, 2025. 2

[9] Meiqi Sun, Zhonghan Zhao, **Chai, Wenhao**†, Hanjun Luo, Shidong Cao, Yanting Zhang, Jenq-Neng Hwang, and Gaoang Wang. Uniap: Towards universal animal perception in vision via few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5008–5016, 2024. 2

[10] Yasutaka Narazaki, Wendong Pang, Gaoang Wang, and **Chai, Wenhao**. Unsupervised domain adaptation approach for vision-based semantic understanding of bridge inspection scenes without manual annotations. *Journal of Bridge Engineering*, 29(2):04023118, 2024. 2

[11] Haojia Cheng, **Chai, Wenhao**†, Jiabao Hu, Wenhao Ruan, Mingyu Shi, Hyunjun Kim, Yifan Cao, and Yasutaka Narazaki. Random bridge generator as a platform for developing computer vision-based structural inspection algorithms. *Journal of Infrastructure Intelligence and Resilience*, 3(2):100098, 2024. 2

[12] Xuechen Guo, **Chai, Wenhao**, Shi-Yan Li, and Gaoang Wang. Llava-ultra: Large chinese language and vision assistant for ultrasound. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8845–8854, 2024. 2