Scalable Vision-Language Understanding and Generation

Linchao Zhu

The College of Computer Science and Technology, Zhejiang University
zhulinchao@zju.edu.cn

Abstract

Recent advances in vision-language models have shown remarkable potential, yet creating scalable systems that can effectively understand and generate across modalities remains challenging. This talk presents our contributions to advancing scalable vision-language systems, focusing on three key research themes and their practical applications.

Efficient Vision-Language Understanding. Our work has made significant strides in video understanding through several key innovations. The development of temporal perceiving video-language pre-training has introduced a novel framework for understanding long-form videos through hierarchical temporal modeling (Ma et al. 2024). The ActBERT framework (Linchao and Yang 2020), widely adopted in both academia and industry, has demonstrated practical effectiveness by winning first place in multiple EPIC-Kitchen Action Recognition Challenges (2019, 2020). In the domain of efficient multimodal processing, we have developed innovative approaches to knowledge integration and zero-shot capabilities. Our knowledge-enhanced dual-stream zero-shot composed image retrieval system (Suo et al. 2024) has been successfully integrated into large-scale retrieval systems, while our whitening-based contrastive learning of sentence embeddings has gained significant traction in the research community.

Scalable Generation and Adaptation. Our research has advanced the field through novel approaches to generation and adaptation. The DECAP framework (Li et al. 2023) enables zero-shot captioning through text-only training, making it particularly valuable for low-resource languages. We developed test-time adaptation with CLIP reward for zero-shot generalization in vision-language models (Zhao et al. 2024), addressing the crucial challenge of model adaptation without fine-tuning. This technology has been enhanced through efficient multimodal fusion via interactive prompting and dynamic global-local prompt tuning. In the realm of multimodal generation, our SEEG co-speech gesture generation framework has been successfully deployed in virtual assistant systems, creating more natural human-AI interactions. Additionally, our fine-grained semantically aligned vision-language pre-training has been widely adopted by the research community.

Practical Applications and Social Impact. Our research has demonstrated significant real-world impact across multiple domains. We develop a gloss-free end-to-end sign language translation system, deployed in pilot programs to assist communication for the deaf and hard of hearing community (Lin et al. 2023). We implement efficient video classification in content moderation systems, contributing to safer online environments while reducing computational costs and environmental impact. Furthermore, we integrate vision-language technologies in educational and assistive technology applications.

Looking forward, my research agenda focuses on three interconnected directions: (1) Scalable Multimodal Understanding: Developing new frameworks for efficient processing and understanding of large-scale multimodal data, with particular emphasis on video understanding and cross-modal alignment. (2) Adaptive and Robust AI: Investigating new approaches for creating AI systems that can rapidly adapt to new domains and remain robust under various operating conditions. (3) AI for Real-world Applications: Expanding research into practical applications, particularly in areas such as sign language translation, co-speech gesture generation, and egocentric action recognition

This talk will provide both technical depth and broader context, making these advances accessible to a general AI audience while highlighting the fundamental innovations and practical implications of the work. Through discussion of these developments, we aim to foster collaboration and advance the field of scalable multi-modal systems.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

References

Li, W.; Linchao, Z.; Wen, L.; and Yang, Y. 2023. DeCap: Decoding CLIP Latents for Zero-Shot Captioning via Text-Only Training. In *ICLR*.

Lin, K.; Wang, X.; Linchao, Z.; Sun, K.; Zhang, B.; and Yang, Y. 2023. Gloss-Free End-to-End Sign Language Translation. In *ACL*.

Linchao, Z.; and Yang, Y. 2020. ActBERT: Learning Global-Local Video-Text Representations. In *CVPR*.

Ma, F.; Jin, X.; Wang, H.; Huang, J.; Linchao, Z.; Feng, J.; and Yang, Y. 2024. Temporal perceiving video-language pre-training. In *AAAI*.

Suo, Y.; Ma, F.; Linchao, Z.; and Yang, Y. 2024. Knowledge-Enhanced Dual-stream Zero-shot Composed Image Retrieval. In *CVPR*.

Zhao, S.; Wang, X.; Linchao, Z.; and Yang, Y. 2024. Test-Time Adaptation with CLIP Reward for Zero-Shot Generalization in Vision-Language Models. In *ICLR*.