Learning Human Action from Different Modalities

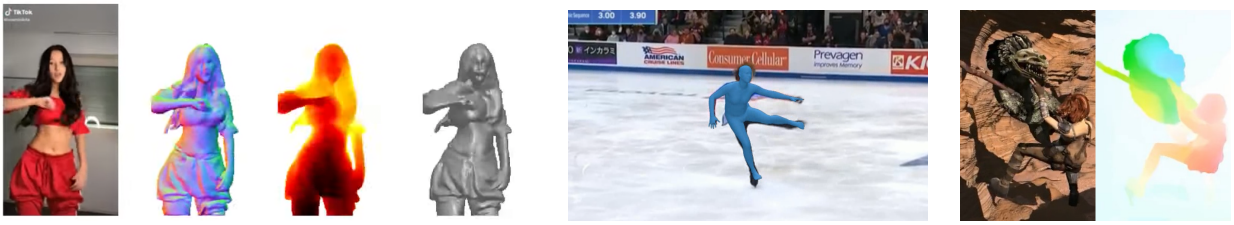
Research Proposal

Yuanhao Zhai

yzhai6@buffalo.edu · https://www.yhzhai.com

1 Introduction

Human actions are the cornerstone of our daily lives, serving as the primary means through which we interact with the world and each other. From simple gestures to complex interactions, our actions convey a wealth of information about our intentions, emotions, and behaviors.



(a) Video, normal, depth and 3D surface [4]. (b) Human body mesh [7]. (c) Optical flow.

Figure 1: Different modalities of human action representation.

Unlike humans, who perceive actions through a combination of visual, auditory, and tactile senses, in the field of computer vision, human actions have traditionally been represented using a single modality, typically visual [2, 3]. This approach, while straightforward, can introduce biases and noise, potentially limiting the accuracy and generalizability of action analysis and synthesis [5, 11, 9, 6]. To address these challenges, integrating multiple modalities—such as 3D, kinematic and motion information—offers significant advantages (fig. 1). This multi-modal strategy not only mimics the comprehensive way humans process complex actions but also significantly enhances the robustness and accuracy of computer vision systems [14, 17, 15, 16, 10]. Leveraging diverse data sources enables more effective human action learning, facilitating advanced applications in diverse fields such as interactive media, healthcare, and public safety.

Building on these observations, my research focuses on *learning human action from different modal representations*. Specifically, my work is directed along two principal axes: enhancing generalizability within single-modal paradigm; leveraging the synergistic potential of multi-modal representations. Below I explain the two directions in detail.

2 Enhancing generalizability within single-modal paradigm

The use of single-modal data, particularly visual, for learning human actions is well-established in computer vision, primarily due to its simplicity and data availability. Traditionally, systems under this paradigm have heavily depended on this type of data, which, though rich in information, is often susceptible to noise and bias. Issues such as varying lighting conditions, occlusions, and background can significantly degrade the quality of visual data. Such reliance on a single sensory input potentially fails to capture the multifaceted nature of human actions, leading to misinterpretations and limitations in generalization across environments To address these challenges and enhance the robustness of single-modal action learning models, I have introduced two methods: one focused on refining action synthesis and the other on improving action recognition, each addressing specific limitations of single-modal representations.

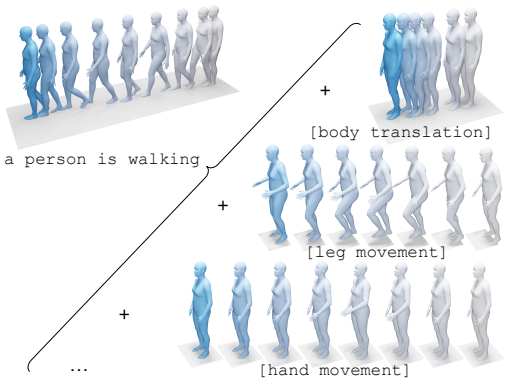
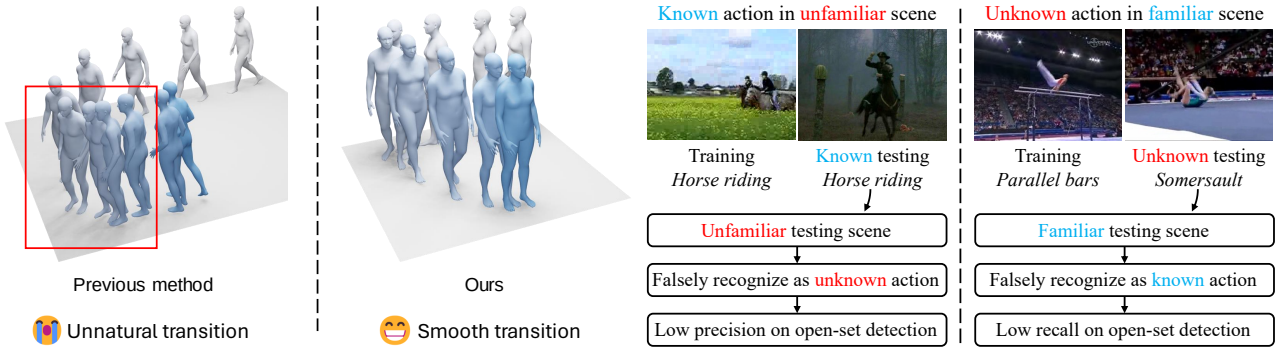


Figure 2: Atomic action assembly.

Learning atomic action representation. Human motion synthesis, essential for enhancing immersive experiences in virtual reality, video



(a) Unnatural action transitions in action synthesis.

(b) Scene bias in action recognition.

Figure 3: Limitations in the single-modal paradigm.

games, and animation, often grapples with the challenge of limited data diversity, leading to unnatural action transitions (fig. 3a left). To tackle this, I proposed ATOM (ATomic mOtion Modeling) [9], which breaks new ground by decomposing actions into distinct atomic components (fig. 2). This strategy not only reflects a naturalistic approach that mirrors human perception of actions as discrete, identifiable units but also ensures that each component maintains its uniqueness through carefully tailored learning objectives aimed at preserving diversity and atomicity. By doing so, the proposed ATOM significantly enhances the coherence and diversity of motion representations, enabling the synthesis of smooth and realistic human action sequences (fig. 3a right). This method brings us closer to replicating the intricate dynamics of human movement, thus improving the authenticity and engagement of digital human interactions in various media

Scene-bias mitigation. The reliance on spurious information like scene context still limits performance of action recognition models, where the background can be unpredictable and varied (fig. 3b). The proposed method, Scene-debiasing Open-set Action Recognition (SOAR), targets this issue by mitigating the scene bias through adversarial learning [11]. SOAR integrates an adversarial scene reconstruction module that reduces scene-related information within features, coupled with an adaptive adversarial scene classification module designed to promote the extraction of scene-invariant action features. These targeted strategies not only reduce dependence on background cues but also improve the model's capacity to reliably recognize actions across a variety of scenes. By doing so, SOAR significantly elevates the reliability of open-set action recognition, making it more robust against environmental variabilities and applicable to a broader range of real-world scenarios.

3 Leveraging multi-modal representations

The use of multi-modal data in computer vision, especially through combining visual data with 3D representations and motion cues, offers a robust approach for learning human actions. This approach enables a comprehensive representation of human actions by capturing both spatial depth and dynamic changes, which surpasses the capabilities of single-modal systems. This section details my proposed methods on how to effectively leverage these modalities to improve action analysis and synthesis.

Cross-modality learning. Human action localization tasks, essential for understanding dynamic content in videos, typically require detailed annotations that are costly and time-consuming to obtain. Addressing this challenge, my approach simplifies the annotation process by utilizing video-level labels while still achieving precise action localization. I introduced an adaptive two-stream consensus network (A-TSCN) that significantly enhances the performance of action localization ([15, 13, 16]). This method synergistically combines RGB and optical flow features, employing an adaptive attention normalization loss and a iterative refinement strategy for cross-modal learning. This approach builds upon the insight that late fusion typically yields more precise results, thereby effectively addressing the inherent ambiguities of action localization with only video-level supervision (fig. 4). Moreover, the integration of cross-modal learning exploits the complementary strengths of various data modalities, substantially improving both the accuracy and adaptability of action localization across diverse video contexts.

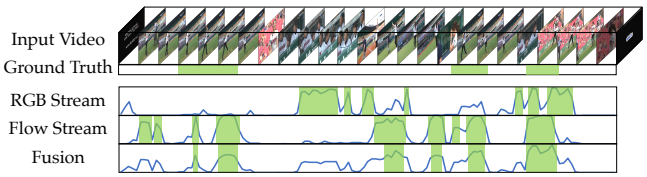


Figure 4: Cross-modal learning.

Learning to generate a holistic human action. The capacity to control and manipulate human-centric video contents has seen significant advancements with the evolution of generative models. Yet, most studies have focused on 2D

Figure 5: Joint generation of human-centric video and depth.

content, limiting applications that require depth perception such as in virtual and augmented reality [8]. I addressed this by proposing IDOL (unIfied Dual-mOdal Latent diffusion), which aims to jointly generate video and corresponding depth maps for human actions, enhancing both visual fidelity and spatial understanding [10] (fig. 5). This method leverages a dual-modal diffusion model that integrates video and depth generation, overcoming challenges posed by their distinct nature. The integration is facilitated by converting depth maps to RGB images, enabling stylized video generation and a more unified treatment of video and depth. By employing cross-modal learning strategies and a unified denoising process within a dual-modal U-Net, my approach not only improves alignment and consistency between the generated video and depth but also ensures efficient parameter usage. Additionally, to achieve precise spatial alignment, I introduced motion consistency and cross-attention map consistency losses, significantly enhancing the realism and applicability of the generated content across diverse settings.

4 Future research

My goal for research is to *build a unified system that comprehensively grasps the nuances of human actions*—capable not only of accurately recognizing and categorizing different movements but also interpreting their underlying intentions and generating corresponding actions in a responsive and realistic manner. This system aims to seamlessly integrate the three core capabilities: understanding the complex dynamics of human actions, interpreting their context and significance, and generating new action sequences that are indistinguishable from natural human behavior. Moving forward, I plan to develop this concept through two specific projects detailed below.

Human action forensics. Recent video generation methods like Sora [1] allow for the creation of hyper-realistic human-centric videos that can be nearly indistinguishable from genuine footage. While these technologies hold immense potential for positive applications, they also pose significant risks. Realistically generated videos can be used to create misleading or harmful content, such as deepfakes, which can spread misinformation or manipulate public opinion. As the visual fidelity of generated videos approaches that of pristine recordings, traditional pixel-based analysis methods may become insufficient for distinguishing between real and synthetic content [12]. To address this challenge, I propose a dual-aspect approach focusing on structure from motion (SfM) and motion pattern analysis. First, SfM will be employed to reconstruct the 3D geometry of scenes from video sequences. By examining the consistency and accuracy of these reconstructions, particularly around human subjects, we can detect anomalies typically absent in natural videos. This method helps identify discrepancies in depth and spatial relationships that are indicative of synthetic generation. Secondly, motion pattern analysis will concentrate on the dynamics of human movements, analyzing how human figures move and interact with their environments. This analysis will scrutinize temporal consistency and smoothness of motion to uncover unnatural patterns often found in generated content. Together, these methodologies form a comprehensive forensic strategy that enhances the detection of synthetic manipulations in human-centric videos.

Multi-view human-centric video generation. Despite my IDOL [10] is able to generate 2.5D effects human-centric videos, it currently lacks the multi-view ability that is critical for augmented reality (AR) and virtual reality (VR) applications. To bridge this gap, I aim to develop a multi-view human-centric video generation method that can produce dynamic video content viewable from multiple perspectives. This capability is essential for creating immersive environments where users can view and interact with digital content from various angles, mimicking real-world experiences. For this project, I plan to leverage SMPL (skinned multi-person linear) model, Gaussian splatting and video diffusion models. The SMPL model will serve as an articulated template for human bodies, ensuring coherent and realistic 3D representations. Gaussian splatting will be used for seamless rendering from multiple viewpoints, ensuring smooth visual transitions that are essential for a fully immersive experience. Additionally, priors from video diffusion models will be incorporated to maintain temporal consistency and realism across views. This combination aims to produce high-quality, consistent video sequences that maintain visual integrity from different perspectives.

References

[1] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 6299–6308, 2017.

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 6202–6211, 2019.

[4] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 12753–12762, 2021.

[5] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proc. Eur. Conf. Comput. Vis.*, pages 513–528, 2018.

[6] Yangcen Liu, Ziyi Liu, Yuanhao Zhai, Wen Li, David Doerman, and Junsong Yuan. Stat: Towards generalizable temporal action localization. *arXiv preprint arXiv:2404.13311*, 2024.

[7] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):1–16, 2015.

[8] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024.

[9] Yuanhao Zhai, Mingzhen Huang, Tianyu Luan, Lu Dong, Ifeoma Nwogu, Siwei Lyu, David Doermann, and Junsong Yuan. Language-guided human motion synthesis with atomic actions. In *Proc. ACM Int. Conf. Multimedia*, 2023.

[10] Yuanhao Zhai, Linjie Li, Kevin Lin, Chung-Ching Lin, Jianfeng Wang, Zhengyuan Yang, David Doermann, Junsong Yuan, Zicheng Liu, and Lijuan Wang. Idol: Unified dual-modal latent diffusion for human-centric joint video-depth generation. 2024.

[11] Yuanhao Zhai, Ziyi Liu, Zhenyu Wu, Yi Wu, Chunluan Zhou, David Doermann, Junsong Yuan, and Gang Hua. Soar: Scene-debiasing open-set action recognition. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2023.

[12] Yuanhao Zhai, Tianyu Luan, David Doermann, and Junsong Yuan. Towards generic image manipulation detection with weakly-supervised self-consistency learning. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2023.

[13] Yuanhao Zhai, Le Wang, David Doermann, and Junsong Yuan. Two-stream consensus network: Submission to HACS challenge 2021 weakly-supervised learning track. *arXiv preprint arXiv:2106.10829*, 2021.

[14] Yuanhao Zhai, Le Wang, Ziyi Liu, Qilin Zhang, Gang Hua, and Nanning Zheng. Action coherence network for weakly supervised temporal action localization. In *Proc. IEEE Int. Conf. Image Process.*, pages 3696–3700, 2019.

[15] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Junsong Yuan, and Gang Hua. Two-stream consensus network for weakly-supervised temporal action localization. In *Proc. Eur. Conf. Comput. Vis.*, pages 37–54, 2020.

[16] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Nanning Zheng, David Doermann, Junsong Yuan, and Gang Hua. Adaptive two-stream consensus network for weakly-supervised temporal action localization. In *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[17] Yuanhao Zhai, Le Wang, Wei Tang, Qilin Zhang, Nanning Zheng, and Gang Hua. Action coherence network for weakly-supervised temporal action localization. *IEEE Trans. Multimedia*, 24:1857–1870, 2022.