Motion Reconstruction via Human Anatomy Diffusion from Sparse Tracking

Supplementary Material

Zehai Niu[1], Ke Lu[1,2], Kun Dong[1], Jian Xue[⋆][1], Xiaoyu Qin[3], and
Jinbao Wang[4]

[1] University of Chinese Academy of Sciences, Beijing, China
[2] Peng Cheng Laboratory, Shenzhen, China
{niuzehai18,dongkun22}@mails.ucas.ac.cn
{luk,xuejian}@ucas.ac.cn
[3] Tsinghua University, Beijing, China
xyqin@tsinghua.edu.cn
[4] Shenzhen University, Shenzhen, China
wangjb@szu.edu.cn

This supplementary material delves deeper into the architectural and experimental nuances that complement our main study. The sections are organized as follows: We initiate with detailed implementation Details, progressing to a comprehensive analysis of the BasicBlock Architecture. Subsequent segments are dedicated to extensive ablation experiments. These trials dissect the HAN and HAD components, shedding light on hyperparameters' effects, the components' order and structure, and the performance across varying frame lengths. Such meticulous examination is instrumental in fine-tuning our model for optimal efficacy. We conclude with additional qualitative results that demonstrate the robustness of our approach. Through a series of visual comparisons and analyses, we showcase the prowess of our method in capturing the subtleties of human motion, benchmarked against state-of-the-art methods and ground truth data. We conclude by discussing the limitations and failure cases, acknowledging our method's challenges.

These supplementary details aim to provide a comprehensive understanding of the research, reinforcing the validity and reliability of our findings.

1 Implementation Details

Due to its simplicity and continuity, we represent the joint by the 6D rotations. Therefore, for the sequences of body poses $\boldsymbol{x} \in \mathbb{R}^{N \times K}, K = 22 \times 6$. The observed joint features $\boldsymbol{c} \in \mathbb{R}^{N \times C}$ consist of the orientation, translation, orientation velocity, and translation velocity of the head and hands in a global

[⋆] This work was supported by the National Natural Science Foundation of China (62320106007, 62032022, 62236006), the Guangdong Provincial Key Laboratory (2023B1212060076) and the Scientific Research Program of Beijing Municipal Education Commission (KZ201911417048). (Corresponding auther: Jian Xue)

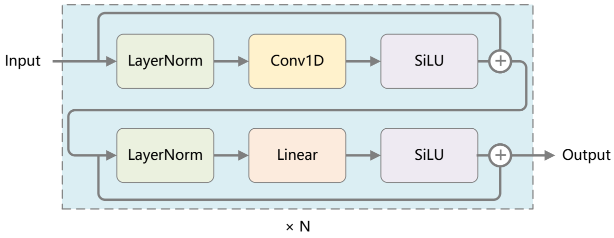
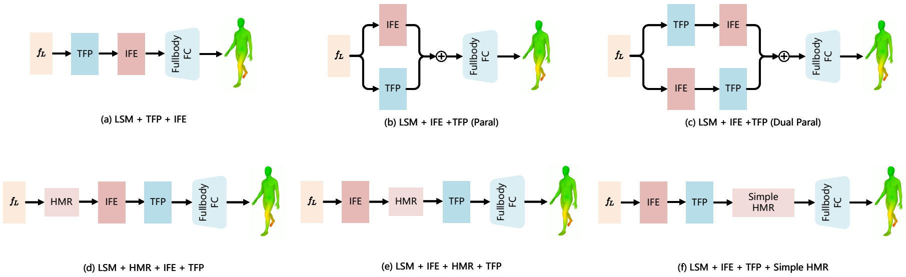


Fig. 1: Network architecture of BasicBlock layer.



(a) LSM + TFP + IFE

(b) LSM + IFE +TFP (Paral)

(c) LSM + IFE +TFP (Dual Paral)

(d) LSM + HMR + IFE + TFP

(e) LSM + IFE + HMR + TFP

(f) LSM + IFE + TFP + Simple HMR

Fig. 2: Variations in HAN component sequencing: exploring performance impacts. This figure illustrates different configurations of the Human Anatomy Network (HAN) components—Latent Space Mapping (LSM), Iterative Feature Enhancement (IFE), Temporal Feature Pyramid (TFP), and Hierarchical Motion Refinement (HMR)—and their influence on motion prediction performance. Configurations (a) through (f) depict the sequence from the standard approach to parallel processing and the integration of HMR.

coordinate system. We also adopt 6D representation for the orientation and velocity, thus $C = 18 \times 3$. We have selected specific values for the hyperparameter configuration within our network architecture to balance computational efficiency and predictive performance. The Iterative Feature Enhancement (IFE) module is configured with $M = 3$ iterations, which have been determined as optimal through our experiments. The Temporal Feature Pyramid (TFP) is set at $L = 5$, providing a multi-scale temporal analysis conducive to accurate motion prediction. We employ 12 layers within the BasicBlock to capture the complex spatiotemporal patterns effectively. Our Hierarchical Motion Refinement (HMR) has three branches, each with 4 BasicBlock layers. Unless otherwise stated, we set the frame number N to 196. We used PyTorch as the deep learning framework and trained our model on a computer with an Intel i9-9900K CPU and an NVIDIA GTX3090 GPU. In addition, our network used Adam optimizer with a base learning rate of 0.0003.

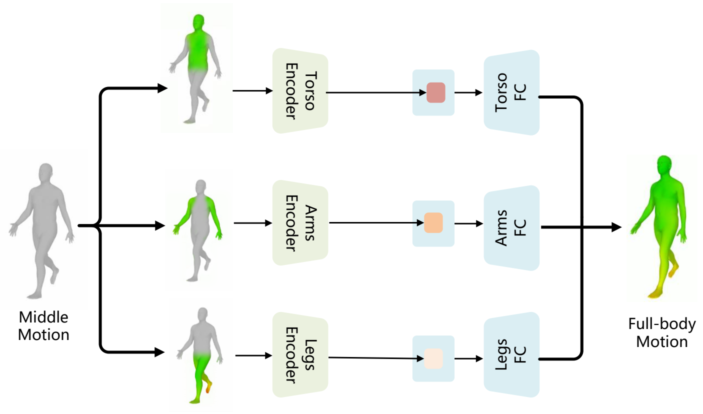


Fig. 3: Network Architecture Diagram of Simple Hierarchical Motion Refinement (Simple HMR).

Table 1: Ablation study on hyperparameter selection for Iterative Feature Enhancement (IFE) in HAN-SP and HAN-AP methods.

Method	M	MPJRE ↓	MPJPE ↓	MPJVE ↓	Hand PE ↓	Upper PE ↓	Lower PE ↓	Root PE ↓	Jitter ↓
HAN-SP	1	**2.39**	3.32	**16.56**	**1.72**	**1.50**	5.94	2.93	4.77
	2	2.40	**3.30**	16.60	1.74	1.51	**5.90**	2.91	4.82
	3	2.41	3.31	16.59	1.75	**1.50**	5.91	**2.87**	**4.69**
HAN-AP	1	2.43	3.30	17.19	1.38	1.45	5.96	2.95	7.78
	2	2.41	3.22	17.43	1.22	1.40	5.84	2.91	8.63
	3	**2.40**	**3.18**	**16.42**	**1.15**	**1.37**	**5.79**	**2.90**	**7.35**

Table 2: Ablation study on hyperparameter selection for Temporal Feature Pyramid (TFP) in HAN-SP and HAN-AP methods.

Method	L	MPJRE ↓	MPJPE ↓	MPJVE ↓	Hand PE ↓	Upper PE ↓	Lower PE ↓	Root PE ↓	Jitter ↓
HAN-SP	1	2.48	3.39	17.04	1.71	1.54	6.06	3.00	4.78
	2	2.48	3.41	17.12	1.77	1.55	6.11	3.01	4.86
	3	2.41	3.33	16.67	1.80	1.52	5.95	2.91	4.79
	4	2.40	3.29	16.80	**1.55**	**1.45**	5.95	2.90	4.86
	5	2.41	3.31	**16.59**	1.75	1.50	5.91	2.87	**4.69**
	6	2.44	3.38	17.02	1.98	1.59	5.98	2.94	4.84
	7	2.44	3.48	17.22	2.08	1.62	6.16	3.03	**4.69**
	8	**2.36**	**3.25**	16.82	1.79	1.50	**5.77**	**2.85**	4.77
HAN-AP	1	2.43	3.27	17.23	1.32	1.43	5.92	2.97	7.83
	2	2.44	3.26	17.13	1.25	1.42	5.92	3.00	7.68
	3	2.42	3.25	17.24	1.29	1.42	5.88	2.95	7.84
	4	2.41	3.23	17.30	1.29	1.41	5.86	2.94	7.93
	5	**2.40**	**3.18**	**16.42**	**1.15**	**1.37**	**5.79**	**2.90**	**7.35**
	6	2.41	3.25	17.95	1.28	1.43	5.87	2.96	9.74
	7	**2.40**	3.27	17.45	1.45	1.46	5.89	2.94	7.98
	8	2.42	3.25	17.80	1.30	1.42	5.89	2.92	9.26

Table 3: Comparative analysis of motion prediction performance using various configurations of the HAN method. The table presents the impact of component sequencing and variations in the HAN architecture.

Method	LSM	IFE	TFP	HMR	MPJRE ↓	MPJPE ↓	MPJVE ↓	Hand PE ↓	Upper PE ↓	Lower PE ↓	Root PE ↓	Jitter ↓
LSM+TFP+IFE	✓	✓	✓		2.49	3.40	20.24	1.50	1.47	6.20	2.96	11.73
LSM+IFE+TFP	✓	✓	✓		2.45	3.33	18.11	1.79	1.53	5.92	2.89	5.59
LSM+IFE+TFP (Paral)	✓	✓	✓		2.43	**3.30**	18.98	**1.44**	**1.44**	5.98	2.93	11.59
LSM+IFE+TFP (Dual Paral)	✓	✓	✓		2.43	3.41	18.85	1.72	1.52	6.14	2.92	7.60
LSM+HMR+IFE+TFP	✓	✓	✓	✓	2.61	3.89	18.14	2.49	1.86	6.83	3.41	5.26
LSM+IFE+HMR+TFP	✓	✓	✓	✓	2.45	3.48	16.99	1.87	1.60	6.20	3.09	4.91
LSM+IFE+TFP+HMR	✓	✓	✓	✓	2.41	3.31	**16.59**	1.75	1.50	**5.91**	**2.87**	4.69
LSM+IFE+TFP+Simple HMR	✓	✓	✓	✓	**2.40**	3.37	17.06	1.72	1.51	6.06	2.95	**4.53**

Table 4: Evaluating PMR and AMR Sequence Strategies in HAN Models. This table examines the impact of sequencing Parallel Motion Regression (PMR) and Alternate Motion Refinement (AMR), specifically focusing on the initial use of HAN-Smooth Prediction (SP priority) and HAN-Accurate Prediction (AP priority) within AMR on the performance metrics of HAN models.

Method	MPJRE	MPJPE	MPJVE	Hand PE	Upper PE	Lower PE	Root PE	Jitter
PMR-AMR Sequence (SP priority)	**2.29**	**3.03**	**15.45**	**1.15**	**1.30**	**5.52**	**2.71**	**4.61**
PMR-AMR Sequence (AP priority)	2.30	3.04	15.84	1.23	1.33	**5.52**	**2.71**	7.18
AMR-PMR Sequence (SP priority)	2.31	3.09	16.16	1.32	1.36	5.58	2.76	7.24
AMR-PMR Sequence (AP priority)	2.31	3.08	16.13	1.36	1.38	5.55	2.76	7.22

2 BasicBlock Architecture

The BasicBlock composition, including the BasicBlock Layer, is depicted in Figure 1. The BasicBlock is a core module within the Latent Space Mapping (LSM), IFE, TFP, and HMR components. The Basic Block Layer starts with Layer Normalization to maintain uniform data scaling across layers, essential for the model's stable learning progression. Following this, the Conv1D layer extracts temporal features, which are pivotal for capturing human motion's dynamic aspects. Incorporating the SiLU activation function introduces the necessary nonlinearity, enabling the model to grasp the complex patterns of human movement. This module utilizes a skip connection to effectively combine the processed output with the initial input, thereby preserving important data characteristics. The Linear layer further extends this mechanism, augmenting the network's capability to refine predictions.

3 Supplementary Ablation Studies

3.1 Hyperparameters

This section systematically explores the sensitivity and impact of various hyperparameters within the HAN. We can discern each influence on the model's performance by varying one hyperparameter at a time while keeping the others constant. The subsections are structured to address the ablation experiments for different components and parameters individually.

Table 5: Comprehensive ablation study evaluating the HAN method components' contribution to motion prediction. The table showcases the effects of LSM, IFE, and TFP.

Method	LSM	IFE	TFP	HMR	MPJRE ↓	MPJPE	MPJVE ↓	Hand PE ↓	Upper PE ↓	Lower PE ↓	Root PE ↓	Jitter ↓
LSM	✓				2.69	3.75	19.65	1.34	1.57	6.90	3.39	8.62
LSM + IFE	✓	✓			2.53	3.42	21.31	1.25	1.44	6.26	3.06	14.75
LSM + TFP	✓		✓		2.59	3.59	21.15	1.83	1.60	6.45	3.06	5.88
LSM + IFE + TFP (SP)	✓	✓	✓		2.45	3.33	18.11	1.79	1.53	5.92	2.89	5.59
LSM + IFE + TFP (AP)	✓	✓	✓		2.40	3.30	18.55	1.42	1.42	6.01	2.90	8.51
HAN-SP	✓	✓	✓	✓	2.41	3.31	16.59	1.75	1.50	5.91	2.87	4.69
HAN-AP	✓	✓	✓	✓	2.40	3.18	16.42	**1.15**	1.37	5.79	2.90	7.35
PMR	✓	✓	✓	✓	2.40	3.19	17.48	1.26	1.39	5.78	2.89	8.42
AMR	✓	✓	✓	✓	2.32	3.10	16.50	1.32	1.37	5.61	2.77	8.10
HAD	✓	✓	✓	✓	**2.29**	**3.03**	**15.45**	**1.15**	**1.30**	**5.52**	**2.71**	**4.61**

Table 6: Ablation study on input sequence length in HAN-SP, HAN-AP, and HAD methods.

Method	Input Size	MPJRE ↓	MPJPE ↓	MPJVE ↓	Hand PE ↓	Upper PE ↓	Lower PE ↓	Root PE ↓	Jitter ↓
HAN-SP	49	2.36	3.29	18.96	1.29	1.38	6.06	2.99	8.34
	98	2.25	3.13	16.50	1.52	1.38	5.65	2.78	5.81
	196	2.41	3.31	16.59	1.75	1.50	5.91	2.87	4.69
	392	2.86	4.07	18.62	2.30	1.86	7.26	3.53	4.33
HAN-AP	49	2.34	3.21	19.55	1.24	1.37	5.88	2.94	11.38
	98	2.30	3.11	16.85	1.22	1.36	5.64	2.87	8.36
	196	2.40	3.18	16.42	1.15	1.37	5.79	2.90	7.35
	392	2.86	3.90	18.53	1.64	1.72	7.06	3.55	7.73
HAD	49	2.25	3.04	18.10	**1.12**	1.28	5.59	2.77	8.26
	98	**2.16**	**2.90**	**15.42**	1.15	**1.25**	**5.29**	**2.63**	5.56
	196	2.29	3.03	15.45	1.15	1.30	5.52	2.71	4.61
	392	2.73	3.76	17.02	1.68	1.65	6.81	3.36	**4.30**

Ablation of IFE Parameters. As shown in Table 1, the ablation study for IFE parameters within the HAN provides critical insights into the performance optimization for both HAN-Smooth Prediction (HAN-SP) and HAN-Accurate Prediction (HAN-AP) methods. By adjusting the IFE hyperparameter M, which influences the iteration intensity of feature enhancement, the study reveals how different settings impact key performance metrics such as MPJRE, MPJPE, MPJVE, Hand PE, Upper PE, Lower PE, Root PE, and Jitter.

A focus on reducing Jitter is observed for HAN-SP, indicating the importance of motion smoothness in the evaluation. Meanwhile, for HAN-AP, the emphasis is on minimizing errors, reflecting a priority on motion prediction accuracy. The optimal setting for both HAN-SP and HAN-AP occurs when the IFE parameter M is set to 3, achieving a balanced outcome between model complexity and predictive performance.

Ablation of TFP Parameters. The TFP plays a crucial role in the HAN by integrating multi-scale temporal features to enhance motion prediction. We conduct an ablation study focusing on different temporal parameter settings within the TFP component for both HAN-SP and HAN-AP methods to optimize its configuration. The results, presented in Table 2, reveal the impact of temporal settings on various performance metrics, including MPJRE, MPJPE, MPJVE, as well as specific joint errors like Hand PE, Upper PE, Lower PE, Root PE, and overall motion smoothness measured by Jitter.

For HAN-SP, our findings indicate that adjusting the temporal parameter L to 8 frames results in the lowest MPJRE and MPJPE, signifying enhanced pose estimation accuracy. This setting also yields the minimum Lower PE and Root PE, indicating improved lower body motion prediction. However, the optimal balance between accuracy and smoothness, measured by Jitter, is achieved at a temporal setting of $L = 5$. Similarly, the HAN-AP method performs best in specific joint errors and overall smoothness at a temporal setting 5.

3.2 Component Sequencing

We conduct supplementary experiments to meticulously investigate the impact of sequencing and structural variations within the Human Anatomy Network (HAN). In particular, various permutations and combinations of the proposed components are analyzed to explore their individual and collective influence on performance. Relevant visual representations of these configurations are detailed in Figure 2, which illustrates the different sequences from standard approaches to parallel processing, including the integration of Simple HMR. The comparative analysis of motion prediction performance across various HAN configurations is also presented in Table 3.

The order of applying IFE and TFP. As shown in Figure 2 (a), the study first assesses the order of IFE and TFP. When IFE precedes TFP (LSM+IFE+TFP), there is a marked reduction in Jitter, decreasing from 11.73 to 5.59. This suggests refining features before applying temporal integration can lead to smoother motion predictions.

Parallel and Dual Parallel Configurations. The implementation of parallelism, as illustrated in Figure 2 (b) for LSM+IFE+TFP (Paral), slightly increases Jitter to 11.59, likely due to a lack of coordinated feature enhancement and temporal smoothing. In comparison, the dual parallel method, as shown in Figure 2 (c) LSM+IFE+TFP (Dual Paral), modestly improves Jitter to 7.60. This indicates that some parallel configurations can achieve a balance between feature refinement and smoothing. However, both parallel approaches do not surpass the sequential arrangement of LSM+IFE+TFP in terms of Jitter performance.

HMR Sequencing. The position of HMR in the sequence shows significant variability in performance. Through our ablation studies, the variations "LSM+HMR+IFE+TFP" (as depicted in Figure 2 (d), LSM+HMR+IFE+TFP) and "LSM+IFE+HMR+TFP" (as depicted in Figure 2 (e), LSM+IFE+HMR+TFP) were analyzed against the optimal arrangement "LSM+IFE+TFP+HMR". The results demonstrate a clear pattern: placing HMR later in the sequence consistently improves performance. Incorporating HMR at the end (LSM+IFE+TFP +HMR) achieves the lowest Jitter at 4.69, implying that performing a hierarchical refinement after feature mapping and enhancement and temporal smoothing is most effective.

HMR structure. A comparative analysis between the standard HMR and its simplified variant (as illustrated in Figure 3 and Figure 2 (f), LSM+IFE+TFP +Simple HMR) reveals nuanced performance dynamics. While the Simple HMR configuration marginally outperforms the standard HMR in minimizing Jitter—achieving a reduction to 4.53 from 4.69—the standard HMR demonstrates superior overall performance across several critical metrics.

Specifically, the standard HMR configuration, when placed as the concluding component (LSM+IFE+TFP+HMR), shows a competitive Jitter performance and significant improvements in MPJRE and MPJVE. Moreover, it reduces errors in Lower PE and Root PE, with the lowest scores of 5.91 and 2.87, respectively. This comprehensive performance advantage underscores the importance of a sophisticated hierarchical refinement approach considering the interrelations between different body parts' motions.

Each of these configurations and their sequence plays a pivotal role in the performance of the HAN method, with the best arrangement yielding substantial improvements in the critical metrics of Jitter, MPJRE, and MPJVE.

Sequencing Impact on PMR and AMR Performance. The final discussion centers on the structure design of HAD. As shown in Table 4, this ablation study explores the sequencing and prioritization between Parallel Motion Regression (PMR) and Alternate Motion Refinement (AMR) affecting motion prediction performance. The experiment reveals subtle differences across configurations. The PMR-AMR sequence with a preference for HAN-SP (PMR-AMR Sequence (SP priority)) demonstrates slightly better overall performance, especially in minimizing Jitter and improving MPJRE and MPJPE. The results suggest PMR is more suited for initiating a better denoising starting point early in the process while prioritizing HAN-SP in AMR can guide a more effective denoising path, resulting in smoother generated motion sequences.

For a complete overview of the ablation studies on all evaluation metrics, which were too extensive to detail in the main text, please refer to Table 5.

3.3 Input Sequence Length

In our focused ablation study regarding the effect of input sequence length on the performance of the Human Anatomy Network (HAN) models, we explored four different input sizes: 49, 98, 196, and 392 frames. This analysis aimed to ascertain the optimal balance between accuracy and motion smoothness for both the HAN Smooth Prediction (HAN-SP), HAN Accurate Prediction (HAN-AP), and Human Anatomy Diffusion (HAD) methods. Key findings from this investigation are summarized in Table 6.

We observed a reduction in Jitter as sequence length increased, indicating that longer sequences lead to smoother motion predictions. This effect was especially significant in the longest sequence of 392 frames, where the HAD method achieved the lowest Jitter of 4.30. However, the accuracy metrics like MPJPE and Hand PE showed a decline compared to shorter sequences. The 196-frame configuration was chosen as our study's standard, balancing motion accuracy

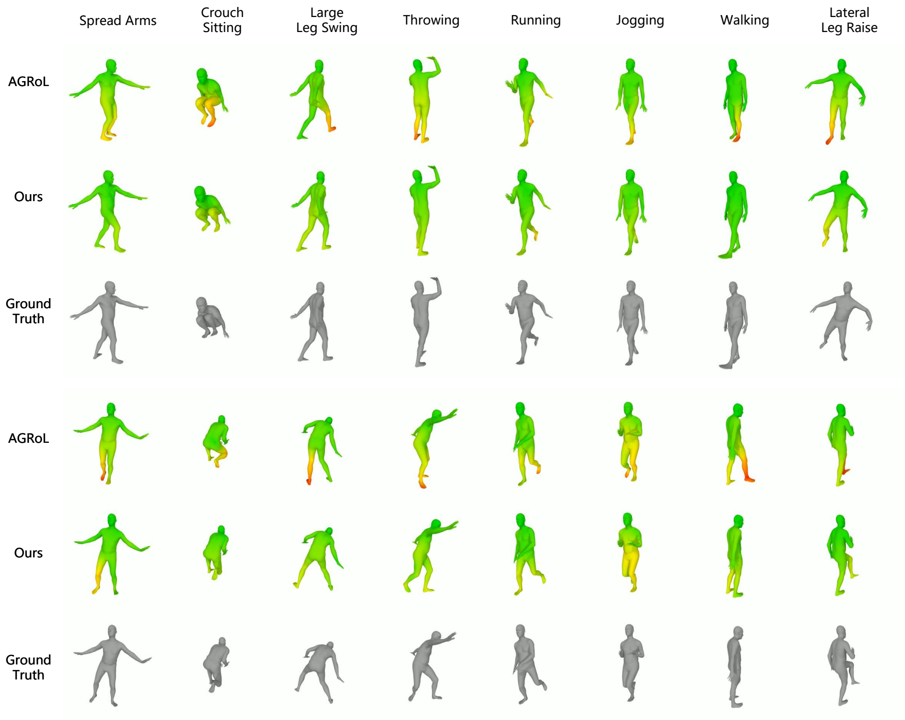


Fig. 4: Comparative visualization of static motion prediction between the AGRoL method (top), our proposed approach (middle), and the ground truth data (bottom).

and smoothness and ensuring a fair comparison with the AGRoL method. Interestingly, the 98-frame input size also proved to be a viable option, exhibiting the lowest MPJRE and MPJPE and significant reductions in Jitter, suggesting its potential for generating both accurate and smooth motion predictions.

4 Additional Qualitative Analysis

In this section, we provide a deeper look at the qualitative performance of our HAD method compared to the previous state-of-the-art AGRoL and ground truth data across static and dynamic movements.

4.1 Static Qualitative Analysis

In this section, the efficacy of our HAD method is thoroughly assessed through a bifurcated evaluation of complex and everyday motions. As illustrated in Figure 4, this segmentation allows for a detailed comparison against the AGRoL method and ground truth data.

The first half of the analysis scrutinizes complex motions: Spread Arms, Crouch Sitting, Large Leg Swing, and Throwing. These actions require precise joint coordination and balance, where HAD demonstrates superior pose estimation, capturing the nuanced positioning and movement that closely mirrors the ground truth.

The second half focuses on everyday motions: Running, Jogging, Walking, and Lateral Leg Raise. These activities are every day yet vital for assessing the practical applicability of motion prediction methods. HAD consistently delivers accurate predictions across these motions, showcasing its robustness and the ability to generalize well to daily human activities. The visual fidelity in these simpler movements further substantiates HAD's potential for real-world applications where natural and authentic motion capture is paramount.

5 Limitations and Failure Cases

Despite the cutting-edge advancements our method provides, it is not without its limitations and associated failure cases, which include:

(1) **Tracking Loss in Spatial Computing Devices**: Our model does not explicitly address scenarios where hand tracking is based on visual tracking in the latest spatial computing devices, potentially leading to tracking losses. It is important to note that this challenge is not isolated to our method but is a common issue across similar methods in the field.

(2) **Sequence Collection Delay**: Although our model operates at a fast speed, it requires the collection of sequences of a certain length. This necessity introduces a delay in sequence collection, a limitation that, once again, is shared by other competing methods.

(3) **Handling Challenging Samples**: Our method may encounter difficulties with particularly challenging samples, resulting in errors. This reflects a broader challenge within motion prediction and generation, where complex samples can sometimes trip up even the most advanced models. For visualization of related failure cases, please refer to the supplementary video.