MoCap4D: A Synchronized Dataset Bridging Fine-grained Motion Tracking and High-Fidelity Multi-View Video

Abstract

With the development of virtual human technology, 3D motion estimation and synthesis have received unprecedented attention. Existing datasets generally have the problem that visual information and motion accuracy are difficult to achieve simultaneously, and this field urgently needs high-quality unified standards. This paper introduces MoCap4D, a comprehensive multimodal dataset that integrates high-fidelity motion capture data with multi-view human activity video recordings. The dataset encompasses full-body skeletal keypoint trajectories from 27 inertial sensors, complemented by high-definition imagery captured synchronously through a 24-camera array. The dataset comprises nearly 7.8 hours of synchronized recordings from 20 participants across 12 distinct activity scenarios encompassing sports, entertainment, and daily living activities, generating over 20.7 million frames of synchronized motion capture data and multi-view video footage. MoCap4D innovatively integrates two complementary data modalities, providing for the first time synchronized high-definition multi-view imagery for digital human reconstruction alongside sub-centimeter-level motion capture tracking, thereby advancing progress in human motion analysis, 3D pose estimation, and motion synthesis domains. We provide benchmark evaluations using state-of-the-art algorithms and demonstrate the dataset's utility across multiple computer vision and graphics applications. The dataset are available at https://anonymous.4open.science/r/MultiView-MoCap-11/.

CCS Concepts

• **Computing methodologies** → **Motion capture**; **Activity recognition and understanding**; **Reconstruction**; *Tracking*.

Keywords

Multimodal Dataset, Motion Capture, Multi-view Video, 3D Pose Estimate, Virtual Human

ACM Reference Format:

. 2018. MoCap4D: A Synchronized Dataset Bridging Fine-grained Motion Tracking and High-Fidelity Multi-View Video. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06

https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Understanding human motion remains a fundamental challenge across computer vision, graphics, biomechanics, and human-computer interaction. Despite significant progress in pose estimation, action recognition, and motion synthesis, research advancement is constrained by dataset limitations—existing datasets typically provide either high-quality motion capture without visual information or video recordings with limited 3D annotations.

Recent years have witnessed a surge in demand for accurate human motion analysis across diverse applications—from immersive virtual reality experiences and photorealistic character animation to comprehensive sports performance assessment—intensifying the need for comprehensive multi-modal datasets. While TotalCapture [46] demonstrated Inertial Measurement Unit (IMU) integration benefits for pose estimation, its visual quality and motion precision fall short of contemporary requirements. Conversely, emerging digital human datasets [1, 3, 32] lack multi-modal information, limiting realistic motion synthesis. Integrating these modalities would establish unified quality standards and provide essential guidance for fine-grained motion analysis tasks.

To address this critical gap, we introduce MoCap4D, a novel multi-modal dataset that uniquely combines high-fidelity motion capture data from a 27-marker full-body system with synchronized video recordings from a 24-camera array. This multi-modal approach provides researchers with unprecedented access to precisely aligned kinematic and visual data spanning diverse human activities, enabling holistic analysis of human motion dynamics. MoCap4D represents the first open dataset to provide multi-view imagery supporting high-definition digital human reconstruction while integrating real-time motion capture data. A detailed comparison with other related datasets is presented in Table 1.

MoCap4D encompasses 12 activity scenarios with over 20.7M video frames, featuring diverse actions including sports movements (basketball handling, football kicks and passes), recreational performances (guitar playing), athletic activities, and fundamental motion sequences (squats, stepping, turning). These actions were carefully selected to represent both common everyday movements and specialized skill-based activities, ensuring comprehensive coverage of the human motion space.

MoCap4D potentially contributes to diverse downstream applications spanning human pose estimation, motion synthesis, and digital human reconstruction. By uniquely integrating precise motion measurements with rich visual context, our dataset may provide a valuable foundation for researchers developing novel approaches to persistent challenges in human motion understanding. The multi-modal nature of our data could facilitate more robust algorithm development and enable new research directions in motion analysis and synthesis.

Against this backdrop, this paper makes several significant contributions to the field of human motion analysis. Specifically, our contributions are as follows:

Table 1: Dataset comparison on existing multi-view human-centric datasets. Our proposed dataset features more refined skeletal keypoint markers and IMU-derived data, facilitating high-precision development in human pose understanding. Additionally, we demonstrate competitive advantages in terms of frame count, resolution, and motion diversity.

Dataset	Venue	Markers	IMUs	Diversity	View	Frames	Resolution
Human3.6M [15]	TPAMI'2014	30	×	17 Activities	4	3.6M	1000P
CMU Panoptic [18]	ICCV'2015	×	×	65 Actions	31	15.3M	1080P
MPI-INF-3DHP [39]	3DV'2017	×	×	8 Activities	14	1.3M	2048P
TotalCapture [46]	BMVC'2017	—	6∼13	4 Activities	8	1.9M	1080P
3DPW [47]	ECCV'2018	—	9∼17	8 Activities	—	51K	1080P
ZJU-MoCap [43]	CVPR'2021	×	×	10 Activities	24	180K	1024P
Neural Actor [32]	TOG'2021	×	×	—	11∼100	250K	1280P
HuMMan [1]	ECCV'2022	×	×	500 Actions	10	60M	1080P
DNA-Rendering [3]	ICCV'2023	×	×	1187 Actions	60	67.5M	4096P
MoCap4D(Ours)	—	**27**	**27**	**12 Activities**	**24**	**20.7M**	**2048P**

- **A large-scale diverse multimodal dataset** combining real-time 27-point motion capture data with synchronized 24-view video recordings at 30 fps, providing multiple data formats (BVH, FBX for motion capture; calibrated raw and processed video) to support various research applications;
- Precise data annotation and post-processing protocols, particularly temporal alignment across data modalities and standardized skeletal markers, enabling the dataset to support various downstream human-centric tasks with effective applications;
- Comprehensive benchmark evaluations using state-of-the-art algorithms for 3D pose estimation, motion prediction, and action recognition.

2 Related Work

2.1 Motion Capture Datasets

Motion capture datasets have played a crucial role in advancing research on human movement analysis. The CMU Motion Capture Database [2] represents one of the earliest comprehensive collections, providing marker-based optical motion capture data across diverse activities. More recent contributions include the Human3.6M dataset [15], which combines motion capture with synchronized video but is limited to 4 camera views and controlled indoor activities.

The AMASS dataset [36] unifies multiple motion capture sources under a common parameterization using the SMPL body model [35], significantly expanding the available motion variety. However, AMASS does not include corresponding video data, limiting its applicability for vision-based tasks. Similarly, the HumanML3D dataset [12] offers a large collection of motion sequences with natural language descriptions but lacks visual context.

2.2 Multi-view Video Datasets

Multi-view human action datasets have grown in importance with advances in computer vision. The Panoptic Studio dataset [19] provides multi-view video recordings using 480 synchronized cameras, enabling detailed analysis of social interactions. However, it does not include full-body motion capture data with the same precision as dedicated motion capture systems.

The TotalCapture dataset [46] combines marker-based motion capture with 8 camera views, representing a step toward multimodal data collection. NTU RGB+D [31] offers Kinect-based skeleton data with RGB and depth information across multiple views, though the skeletal data lacks the precision of professional motion capture systems.

2.3 Human Motion Analysis

Research in human motion analysis spans multiple tasks, including 3D pose estimation, action recognition, and motion prediction. For 3D pose estimation, methods like VIBE [22] and SPIN [23] estimate human pose and shape from monocular video by leveraging temporal information and statistical body models. Multi-view approaches such as Learnable Triangulation [16] integrate information across camera views to improve accuracy.

Action recognition has evolved from hand-crafted features to deep learning approaches, with recent methods like ST-GCN [49] and MS-G3D [34] modeling the spatial-temporal relationships in skeletal data. Cross-modal approaches that combine visual and kinematic information have shown promising results but are constrained by the limited availability of synchronized datasets.

Motion prediction and synthesis represent another active research area, with approaches ranging from recurrent neural networks to transformer-based models like Motion Diffusion Models [45]. These methods benefit significantly from high-quality motion data but often struggle with complex activities and realistic human-object interactions.

Despite significant progress, research across these domains remains constrained by dataset limitations. Existing datasets typically excel in either motion precision or visual richness, but rarely both. MoCap4D addresses this gap by providing synchronized, high-quality data across both modalities.

3 Dataset Construction

3.1 Capture Setup

The MoCap4D dataset was collected using a comprehensive capture setup that integrates high-precision motion capture with multi-view video recording. Our data acquisition was primarily conducted in

Figure 1: The motion capture equipment used in our data collection process, featuring the optical tracking system and marker set that enables precise 3D tracking of body movements. The markers are strategically placed at anatomical landmarks to enable precise tracking of all major body segments during movement.

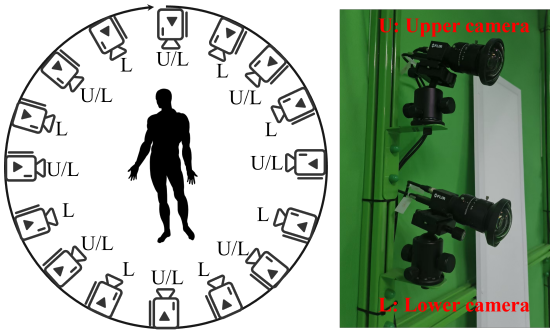

Figure 2: Our 24-camera array setup arranged in a 360-degree configuration around the capture volume. The cameras are set up with two different heights and pitch angles. Each camera is precisely calibrated and synchronized to ensure consistent capture across all viewpoints.

a cylindrical photography facility equipped with an array of high-resolution cameras and reflectors to ensure optimal imaging quality.

The entire capture volume measured $5m \times 5m \times 3m$, providing sufficient space for a wide range of movements including running, jumping, and extended sports activities. The integration of our motion capture system and multi-camera array within this volume creates a comprehensive sensing environment that captures both precise kinematic data and rich visual information simultaneously.

3.1.1 Motion Capture System. We employed a professional-grade optical motion capture system with 27 markers placed at key anatomical landmarks following the modified Helen Hayes marker set [20]. This configuration provides comprehensive tracking of all major body segments, including detailed hand and foot movement. The system captures data at 60 Hz with sub-millimeter precision, providing high-fidelity kinematic measurements of human movement.

To ensure consistency, all participants' marker placements were standardized according to human skeletal anatomical structure to guarantee accurate pose estimation. The anatomical landmark points were carefully positioned with particular attention to the following key anatomical markers:

- **Trunk and Head (3 markers)**:Head, Spine, Hip
- **Upper Extremities (8 markers)**: LeftShoulder, LeftUpArm, LeftForeArm, LeftHand, RightShoulder, RightUpArm, RightForeArm, RightHand
- **Lower Extremities (6 markers)**: LeftUpLeg, LeftLowLeg, LeftFoot, RightUpLeg, RightLowLeg, RightFoot
- **Finger joints (10 markers)**: Thumb, Index finger, Middle finger, Ring finger, Pinky(right and left)

Figure 1 shows a participant wearing the complete motion capture marker set. The right panel displays the skeletal keypoint markers from the original motion capture design and the standardized markers, respectively. This configuration allows for accurate reconstruction of full-body movements while minimizing interference with natural motion patterns.

3.1.2 Multi-camera Array. The video capture system consists of 24 synchronized cameras arranged in a 360-degree configuration around the capture volume, as shown in Fig. 2. Sixteen camera array positions are evenly distributed around the horizontal circular ring, with selected positions equipped with two cameras at different heights and pitch angles. This configuration aims to achieve comprehensive angular coverage and precise visual capture. All cameras were calibrated to a common coordinate system aligned with the motion capture space.

The camera specifications include:

- Resolution: 2448 × 2048 pixels
- Frame rate: 30 frames per second
- Field of view: 85 degrees horizontal
- Time synchronization: Hardware-triggered with <1ms inter-camera latency

The cameras were arranged to maximize coverage while minimizing occlusions, with an average angular separation of 15 degrees at each height level. This arrangement ensures that any point within the capture volume is visible from at least 8 cameras simultaneously, facilitating robust multi-view geometry. The synchronized multi-view captures provide comprehensive visual coverage, facilitating detailed kinematic analysis from multiple perspectives while supporting robust 3D reconstruction algorithms.

3.2 Data Collection

3.2.1 Participants. We recruited 20 young volunteers with diverse body types to capture multi-view imagery, including 13 males and 7 females. For each recording session, participants were fitted with the motion capture markers as shown in Figure 1, with careful attention to consistent placement across sessions.

3.2.2 Action Categories. MoCap4D includes five primary action categories, further divided into specific movements and sequences:

(1) **Ball Sports**
- **Basketball**: Dribbling (static and moving), shooting (free throws, jump shots, layups), passing (chest pass, bounce pass), defensive slides, and combined sequences.
- **Football/Soccer**: Kicking (various techniques), ball control, passing, dribbling, and goal-keeping movements.

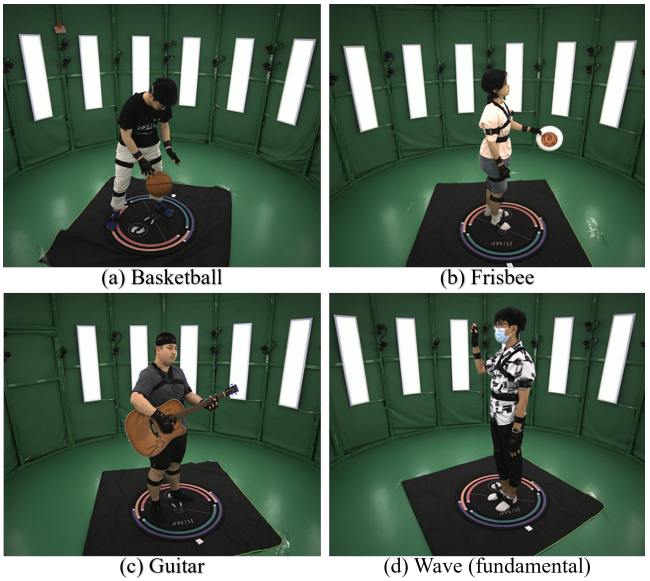

Figure 3: Representative synchronized frame sequences obtained from a multi-camera array system illustrate participants' behavioral demonstrations under different activity scenarios. The 24-camera setup captures the movement from all angles simultaneously, enabling comprehensive visual analysis.

- **Tennis**: Serving, forehand and backhand strokes, volleys, and court movement patterns.
- **Table Tennis**: Bouncing, juggling, and ball control.

(2) **Traditional and Mind-Body Exercises**
- **Tai Chi**: Slow, flowing movements emphasizing balance, coordination, and traditional forms.
- **Yoga**: Various poses (asanas), transitions, breathing sequences, and flexibility movements.

(3) **Fitness and Athletic Training**
- **Strength Training**: Push-ups, dumbbell exercises, and resistance-based movements.
- **Cardio and Agility**: Running, jumping (vertical and broad), agility drills, and sport-specific athletic movements.

(4) **Combat Sports**
- **Boxing**: Punching combinations, defensive movements, footwork, and sparring techniques.

(5) **Recreational Activities**
- **Frisbee**: Throwing techniques, catching movements, and recreational play patterns.
- **Guitar Performance**: Basic chords, strumming patterns, fingerpicking techniques (hammer-ons, pull-offs, bends), and song excerpts across different musical styles.

(6) **Daily Life Activities**
- **Fundamental Movement Sequences**: Walking (various speeds and styles), waving, pacing, squatting, stepping, turning, reaching, and basic locomotion patterns.

Each participant performed a subset of activities from these categories according to their expertise level while wearing full-body motion capture equipment, with all participants completing the basic movement sequences. For each activity, participants performed both isolated actions and continuous sequences, thereby providing data for segmented action recognition and continuous motion analysis. The Fig. 3 illustrates samples from our recording process, where 24-channel high-definition multi-view images were captured.

3.2.3 Recording Procedure. Each recording session followed a standardized protocol:

(1) Participant preparation: Anthropometric measurements, marker placement, and familiarization with the capture environment.
(2) Range of motion trials: Standardized movements to capture joint ranges and facilitate subsequent skeletal calibration.
(3) Activity performance: Guided execution of the action categories, with both predefined and freestyle components.
(4) Validation trials: Repetition of selected movements to assess consistency and provide redundancy.

For each participant, recording sessions lasted approximately 2-3 hours, resulting in 20-30 minutes of processed data per participant. Rest periods were incorporated to minimize fatigue effects. A research assistant provided verbal instructions and demonstrations when needed, ensuring consistent execution across participants.

During recording, participants performed activities within the calibrated volume captured by both the motion capture system and the multi-camera array, as illustrated in Figure 3. Participants were required to turn around at irregular intervals during movement with the repetition of actions allowed. This synchronized capture approach ensures that each movement is documented simultaneously through both precise kinematic measurements and comprehensive visual recordings from multiple viewpoints.

3.3 Data Annotation

To facilitate the advancement of applications in 2D/3D human pose understanding and reconstruction, our dataset provides comprehensive and diverse annotations alongside the raw data. Due to the excessive storage demands and handling challenges associated with raw image data, we utilized H.264-based video encoders for data compression. The subsequent annotations include camera calibrations, temporal synchronization, human segmentation, and standardized 3D skeleton markers. The annotation pipeline, as illustrated in Fig. 4, provides an overview of the entire process.

3.3.1 Camera and Spatial Calibration. We employed a commercial solution based on ChArUco calibration boards to achieve rapid and efficient camera calibration. Specifically, we positioned the calibration board with ChArUco patterns at the center of the capture area, ensuring that each camera could obtain a clear and complete view of the calibration target. Using specialized software, we acquired the intrinsic parameters, extrinsic parameters, and distortion coefficients for each camera. Additionally, we carefully adjusted other parameters, including illumination, exposure, and white balance, to ensure high-quality data acquisition. We used a combination of a calibration wand with known marker distances and a static calibration frame to establish the coordinate system relationship between the motion capture space and each camera's reference

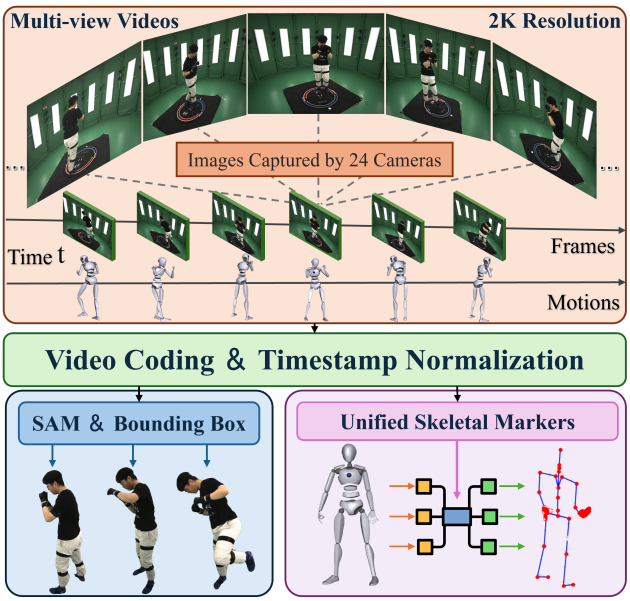

Figure 4: Our data annotation post-processing pipeline processes the recorded videos and captured motion data through a series of operations: video coding, temporal segmentation, human segmentation, and standardized 3D skeleton markers.

frame. The resulting calibration achieved a mean reprojection error of 0.42 pixels.

3.3.2 Temporal Synchronization. A hardware synchronization system provided timing signals to both the motion capture and camera systems. Additional visual time codes were recorded at the beginning of each session for verification. Moreover, due to the hardware configuration limitations of our equipment, the sampling frequencies of the two systems were not synchronized. Achieving precise temporal alignment between the motion capture and video systems was critical to the dataset's value. We implemented a rigorous temporal synchronization protocol to ensure minimal temporal discrepancy (less than 10ms) between the two modalities. The aligned data was standardized to a 30Hz sampling rate for subsequent development.

3.3.3 Human Segmentation. Our dataset comprises over ten million human images captured from diverse viewpoints. To facilitate development and usage, we performed human segmentation to eliminate extraneous background elements. We employed an automated image segmentation approach based on the Segment Anything Model (SAM) [21]. Bounding boxes were utilized to ensure that SAM segmentation focused primarily on human subjects rather than other scene elements.

3.3.4 Unified Skeletal Markers. The kinematic reconstruction phase utilizes Axis Studio's proprietary sensor fusion algorithms to translate raw IMU data into coherent skeletal representations. This process integrates quaternion-based orientation data with a biomechanical model that enforces anatomical constraints, effectively addressing sensor drift and occlusion issues common in inertial motion capture. The software applies an inverse kinematics solver that optimizes joint rotations to maintain skeletal integrity. These processing steps yield a standardized motion representation featuring aligned 3D joint positions, anatomically consistent joint angles, and derived kinematic features including velocities and accelerations—all essential for subsequent computational modeling and analysis. Following reconstruction, we extract complete human pose data from the processed FBX files including positional coordinates for all skeletal joints.

4 Experimental Benchmarks on MoCap4D

4.1 Pose Estimation Benchmark Results

Table 2: Pose Estimation Methods Evaluation on MoCap4D.

Method	Venue	MPJPE ↓	PA-MPJPE ↓
RepNet [48]	CVPR'19	76.5	68.3
VPose (1-frame) [40]	CVPR'19	71.4	49.6
EvoSkeleton [28]	CVPR'20	56.4	43.2
DH-AUG [14]	ECCV'22	53.5	43.5
MHFormer [30]	CVPR'22	46.3	38.2
MixSTE [52]	CVPR'22	43.5	36.7
P-STMO [44]	ECCV'22	45.1	35.1
Stridedformer [29]	TMM'22	46.2	37.4
PoseAug [51]	TPAMI'23	57.3	42.7
PoseGU [10]	CVIU'23	59.2	45.3
CEE-Net [26]	AAAI'23	51.3	40.2
Uplift [5]	WACV'23	47.7	37.2
STGFormer [33]	CVPR'23	44.2	36.3
PoseFormerV2 [53]	CVPR'23	47.4	37.8
UPS [6]	CVPR'23	45.5	37.9
GLA-GCN [50]	ICCV'23	46.0	36.4
MotionBERT [54]	ICCV'23	44.1	36.8
DAF-DG [42]	CVPR'24	49.7	37.3
MotionAGFormer [38]	WACV'2024	42.5	36.5

To demonstrate the utility and effectiveness of our MoCap4D dataset, we conducted comprehensive benchmarking experiments using state-of-the-art 3D human pose estimation methods. These evaluations not only validate the dataset's quality but also establish baseline performance metrics for future research.

We evaluated 19 leading human pose estimation approaches on our dataset, spanning from earlier methods such as RepNet [48] to recent innovations like MotionAGFormer [38]. For consistent comparison, we report two standard evaluation metrics: Mean Per Joint Position Error (MPJPE), which measures the average Euclidean distance between predicted and ground-truth joint positions in millimeters, and Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE), which applies rigid alignment before computing the error to focus on pose structure rather than absolute positioning.

As shown in Table 2, the performance of various methods reveals several notable trends. Earlier approaches such as RepNet and VPose exhibit relatively high error rates (MPJPE of 76.5mm and 71.4mm respectively), highlighting the challenging nature of our

dataset. Methods from 2022-2023 show substantial improvements, with MixSTE [52] achieving an MPJPE of 43.5mm and P-STMO [44] reaching the lowest PA-MPJPE of 35.1mm among methods from that period.

The most recent approach, MotionAGFormer [38], demonstrates the best overall performance with an MPJPE of 42.5mm, suggesting that the integration of transformer architectures with graph convolutional networks is particularly effective for capturing the complex spatial-temporal relationships in human motion. However, its PA-MPJPE of 36.5mm does not surpass P-STMO's 35.1mm, indicating that while global positioning accuracy has improved, the structural alignment precision still has room for advancement.

We observe that transformer-based methods (MHFormer [30], MixSTE [52], STGFormer [33], MotionBERT [54]) generally outperform earlier convolutional approaches, confirming the effectiveness of attention mechanisms for modeling long-range dependencies in human motion sequences. The varying performance across different architectural paradigms underscores the value of MoCap4D as a challenging benchmark that can differentiate between algorithmic approaches.

These benchmark results establish important baselines for the research community and highlight promising directions for future work. The persistent gap between current state-of-the-art performance and perfect reconstruction suggests that MoCap4D offers sufficient complexity to drive continued innovation in human pose estimation techniques. Additionally, the comprehensive nature of our dataset—combining high-precision motion capture with multi-view video—enables researchers to explore novel approaches that leverage both modalities for improved performance.

4.2 Motion Prediction Benchmark Results

Human motion prediction aims to forecast future poses based on observed movement sequences. This task is crucial for applications such as autonomous driving, human-robot interaction, and healthcare monitoring, where anticipating human behavior enhances system safety and performance. To evaluate the effectiveness of our proposed dataset for action sequence prediction, we compared several state-of-the-art human motion prediction methods on the MoCap4D dataset. Table 3 presents a comprehensive comparison using Mean Per Joint Position Error (MPJPE) across different prediction time spans.

The benchmark evaluates prediction accuracy across four time horizons: short-term (80ms), medium-term (160ms), and long-term (320ms and 400ms) predictions. These intervals are particularly relevant in motion prediction contexts, where even milliseconds of prediction accuracy can significantly impact user experience. Several insights emerge from these benchmark results:

- Simple baselines like ZeroV [37], which assumes static pose continuation, perform poorly even at short time horizons, confirming the dynamic nature of VR-induced movements.
- Recurrent models (ERD [7], Lstm3LR [7]) struggle with capturing the complex dependencies between visual stimuli and resulting motion.
- Adversarial approaches (AGED [11], BiHMP-GAN [24]) show competitive performance, particularly at longer horizons, suggesting the importance of learning natural motion distributions.

Table 3: Motion Prediction Results

Method	Venue	Prediction time span (ms)			
		80	160	320	400
ERD [7]	ICCV 2015	0.81	0.96	1.41	1.56
Lstm3LR [7]	ICCV 2015	0.84	1.23	1.41	1.58
SRNN [17]	CVPR 2016	0.86	0.98	1.05	1.35
ZeroV [37]	CVPR 2017	0.46	0.77	0.93	1.40
DropAE [8]	3DV 2017	0.96	1.17	1.43	1.85
Samp-loss [37]	CVPR 2017	0.74	0.97	1.17	1.26
Res-sup [37]	CVPR 2017	0.34	0.46	0.77	0.89
CSM [25]	CVPR 2018	0.41	0.68	0.74	0.89
TP-RNN [4]	WACV 2019	0.28	0.45	0.65	0.79
AGED [11]	ECCV 2018	0.27	0.44	0.53	0.62
BiHMP-GAN [24]	AAAI 2019	0.36	0.50	0.67	0.71
Skel-TNet [13]	AAAI 2019	0.34	0.53	0.66	0.71
VGRU-r1 [9]	CVPR 2019	0.38	0.63	0.67	0.79
Sybio-GNN [27]	TPAMI 2022	**0.26**	**0.35**	**0.45**	**0.57**

Notably, methods that incorporate structural information about human kinematics (QuaterNet [41], AGED [11], and Sybio-GNN [27]) consistently outperform generic sequence models, highlighting the importance of domain-specific architectural design.

These results validate the utility of the MoCap4D dataset for developing and evaluating motion prediction models, while demonstrating that significant performance gains are possible through specialized architectures that bridge visual and kinematic domains.

5 Conclusion

In conclusion, we have presented MoCap4D, a comprehensive multimodal dataset that addresses the critical gap in human motion analysis research by uniquely combining high-fidelity 27-marker motion capture data with synchronized 24-camera video recordings. This dataset encompasses diverse human activities ranging from sports movements to everyday actions, providing researchers with unprecedented access to aligned kinematic and visual data at 30 fps across multiple formats (BVH, FBX, and calibrated video). Through rigorous data annotation protocols and temporal alignment procedures, MoCap4D establishes a robust foundation for advancing research in 3D pose estimation, motion prediction, and action recognition. Our comprehensive benchmark evaluations demonstrate the dataset's effectiveness in supporting state-of-the-art algorithms across these tasks. By bridging the gap between precise motion measurements and rich visual context, MoCap4D not only facilitates the development of more accurate and holistic human motion analysis methods but also opens new avenues for applications in virtual reality, character animation, sports analysis, and clinical movement assessment, ultimately accelerating progress in the interdisciplinary field of human motion understanding.

References

[1] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. 2022. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*. Springer, 557–577.

[2] Carnegie Mellon University. 2003. CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu/. Accessed: 2025-05-30.

[3] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. 2023. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19982–19993.

[4] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. 2019. Action-agnostic human pose forecasting. In *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1423–1432.

[5] Moritz Einfalt, Katja Ludwig, and Rainer Lienhart. 2023. Uplift and upsample: Efficient 3d human pose estimation with uplifting transformers. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2903–2913.

[6] Lin Geng Foo, Tianjiao Li, Hossein Rahmani, Qiuhong Ke, and Jun Liu. 2023. Unified pose sequence modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13019–13030.

[7] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. 2015. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*. 4346–4354.

[8] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. 2017. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*. IEEE, 458–466.

[9] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. 2019. A neural temporal model for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12116–12125.

[10] Shannan Guan, Haiyan Lu, Linchao Zhu, and Gengfa Fang. 2023. Posegu: 3d human pose estimation with novel human pose generator and unbiased learning. *Computer Vision and Image Understanding* 233 (2023), 103715.

[11] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. 2018. Adversarial geometry-aware human motion prediction. In *Proceedings of the european conference on computer vision (ECCV)*. 786–803.

[12] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5152–5161.

[13] Xiao Guo and Jongmoo Choi. 2019. Human motion prediction via learning local structure representations and temporal dependencies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2580–2587.

[14] Linzhi Huang, Jiahao Liang, and Weihong Deng. 2022. Dh-aug: Dh forward kinematics model driven augmentation for 3d human pose estimation. In *European Conference on Computer Vision*. Springer, 436–453.

[15] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (jul 2014), 1325–1339.

[16] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. 2019. Learnable Triangulation of Human Pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[17] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. 2016. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 5308–5317.

[18] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2015. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE international conference on computer vision*. 3334–3342.

[19] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. 2015. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[20] M. P. Kadaba, H. Ramakrishnan, and M. E. Wootten. 1990. Measurement of lower extremity kinematics during level walking. *Journal of Orthopaedic Research* 8 (1990), 383–392. Issue 3. doi:10.1002/jor.1100080310

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4015–4026.

[22] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[23] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. 2019. Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[24] Jogendra Nath Kundu, Maharshi Gor, and R Venkatesh Babu. 2019. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 8553–8560.

[25] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. 2018. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5226–5234.

[26] Haolun Li and Chi-Man Pun. 2023. Cee-net: complementary end-to-end network for 3d human pose generation and estimation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 1305–1313.

[27] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2021. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE transactions on pattern analysis and machine intelligence* 44, 6 (2021), 3316–3333.

[28] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. 2020. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6173–6183.

[29] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang. 2022. Exploiting temporal contexts with strided transformer for 3d human pose estimation. *IEEE Transactions on Multimedia* 25 (2022), 1282–1293.

[30] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. 2022. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13147–13156.

[31] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. 2020. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 10 (2020), 2684–2701. doi:10.1109/TPAMI.2019.2916873

[32] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. 2021. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM transactions on graphics (TOG)* 40, 6 (2021), 1–16.

[33] Yang Liu and Zhiyong Zhang. 2024. STGFormer: Spatio-Temporal GraphFormer for 3D Human Pose Estimation in Video. *arXiv preprint arXiv:2407.10099* (2024).

[34] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. 2020. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.* 34, 6, Article 248 (Oct. 2015), 16 pages. doi:10.1145/2816795.2818013

[36] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. 2019. AMASS: Archive of Motion Capture As Surface Shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[37] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2891–2900.

[38] Soroush Mehraban, Vida Adeli, and Babak Taati. 2024. Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 6920–6930.

[39] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. 2017. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*. IEEE, 506–516.

[40] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7753–7762.

[41] Dario Pavllo, David Grangier, and Michael Auli. 2018. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485* (2018).

[42] Qucheng Peng, Ce Zheng, and Chen Chen. 2024. A dual-augmentor framework for domain generalization in 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2240–2249.

[43] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*.

[44] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. 2022. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *European Conference on Computer Vision*. Springer, 461–478.

[45] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. 2022. Human Motion Diffusion Model. arXiv:2209.14916 [cs.CV] https://arxiv.org/abs/2209.14916

[46] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John Collomosse. 2017. Total capture: 3d human pose estimation fusing video and

inertial sensors. In *Proceedings of 28th British Machine Vision Conference*. 1–13.
[47] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. 2018. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*. 601–617.
[48] Bastian Wandt and Bodo Rosenhahn. 2019. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7782–7791.
[49] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018). doi:10.1609/aaai.v32i1.12328
[50] Bruce XB Yu, Zhi Zhang, Yongxu Liu, Sheng-hua Zhong, Yan Liu, and Chang Wen Chen. 2023. Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8818–8829.
[51] Jianfeng Zhang, Kehong Gong, Xinchao Wang, and Jiashi Feng. 2023. Learning to augment poses for 3D human pose estimation in images and videos. *IEEE transactions on pattern analysis and machine intelligence* 45, 8 (2023), 10012–10026.
[52] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. 2022. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13232–13242.
[53] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. 2023. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8877–8886.
[54] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. 2023. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15085–15099.