VRMotion: A Large-Scale Dataset for Full-Body Motion Prediction in Ego-Vision Tasks

Dayou Zhang
zhangdayou@cnu.edu.cn
Capital Normal University
Beijing, China

Yi Song
Capital Normal University
Beijing, China
2251002088@cnu.edu.cn

Shufang Lin
123090335@link.cuhk.edu.cn
The Chinese University of Hong Kong, Shenzhen
Shenzhen, GuangDong, China

Zijian Cao
zijiancao1@link.cuhk.edu.cn
The Chinese University of Hong Kong, Shenzhen
Shenzhen, GuangDong, China

Rongrong Zhang
zhangrr@cnu.edu.cn
Capital Normal University
Beijing, China

Fangxin Wang*
wangfangxin@cuhk.edu.cn
The Chinese University of Hong Kong, Shenzhen
Shenzhen, GuangDong, China

Abstract

As artificial intelligence systems increasingly interact with humans in physical environments, understanding the causal relationship between visual perception and full-body motor responses becomes critical for safe and natural human-AI collaboration. However, existing motion datasets either lack visual context or are constrained by marker occlusion and limited capture volumes in large-scale scenarios. We present **VRMotion**, a large-scale multimodal dataset that captures temporally aligned egocentric visual stimuli and corresponding full-body kinematic responses. We leverage VR environments to safely simulate diverse task scenarios, while an omnidirectional treadmill combined with a 27-sensor IMU system enables occlusion-free capture of unconstrained locomotion with consistent precision. The dataset contains **21.6 million frames** across three task categories with varying cognitive and motor complexity: directive (Beat Saber), suggestive (Table Tennis), and explorative (Blade & Sorcery). Leveraging this rich data, we conduct a **systematic evaluation** of cross-modal motion prediction by benchmarking ten distinct combinations of visual backbones and temporal heads, establishing a rigorous foundation for the next generation of intelligent, predictive VR systems. Our dataset is available at https://naislab.cn/datasets/VRMotion/.

CCS Concepts

• **Computing methodologies** → **Motion capture**; **Machine learning algorithms**; • **Human-centered computing** → *Virtual reality*.

Keywords

Multimodal Dataset, Motion Capture, Motion Prediction, Virtual Reality, Benchmark

1 Introduction

The trajectory of machine learning has been toward systems that participate in human activity—from language models that assist in knowledge work to embodied agents that manipulate physical objects. Recent advances in vision-language models [30], large language models [8], and embodied AI [35] have brought this vision closer to reality. Yet a fundamental capability that humans perform effortlessly remains out of reach: planning and executing motor actions directly from visual input. While machines can now interpret complex scenes and generate human-like motion independently, they cannot predict how a person will move given what that person sees. This limitation—the inability to model the causal mapping from egocentric visual perception to full-body kinematic responses—prevents AI systems from truly anticipating human behavior, constraining them to reactive rather than proactive interaction across robotics, virtual reality, and human behavior modeling [7, 32].

However, mastering this capability remains fundamentally hindered because existing datasets cannot support such a cross-modal problem formulation. Current repositories fail to capture the essential causal relationship between perception and action due to three core limitations: (1) **Modality Isolation and Temporal Misalignment**: Motion capture repositories [19, 21] record rich kinematics but lack participants' visual experiences. Conversely, egocentric datasets [10] often omit body tracking, and critical temporal synchronization between modalities is either absent or achieved with precision insufficient for studying rapid sensorimotor responses. (2) **Spatial and Locomotion Constraints**: Datasets employing traditional marker-based capture [34] suffer from marker occlusion during self-interaction and limited capture volumes, while RGB-based pose estimation [44] experiences precision degradation proportional to camera distance. (3) **Absence of Diverse Elicitation Scenarios**: The field lacks diverse scenarios specifically designed to elicit natural, cognitively driven bodily responses from humans to specific visual stimuli.

To overcome the aforementioned challenges of modality isolation, marker occlusion, and spatial restrictions, we introduce **VRMotion**—a large-scale multimodal dataset engineered to capture the causal loop between egocentric visual stimuli and full-body motion responses (Fig. 1). In this study, we utilize Virtual Reality (VR) as a safe and convenient medium exclusively for acquiring egocentric visual inputs. Furthermore, to resolve the visual occlusion issues of traditional optical systems, we deploy a 27-sensor wireless inertial measurement unit (IMU) suite. To simulate expansive environments while preventing IMU signal degradation and precision

*Corresponding author.

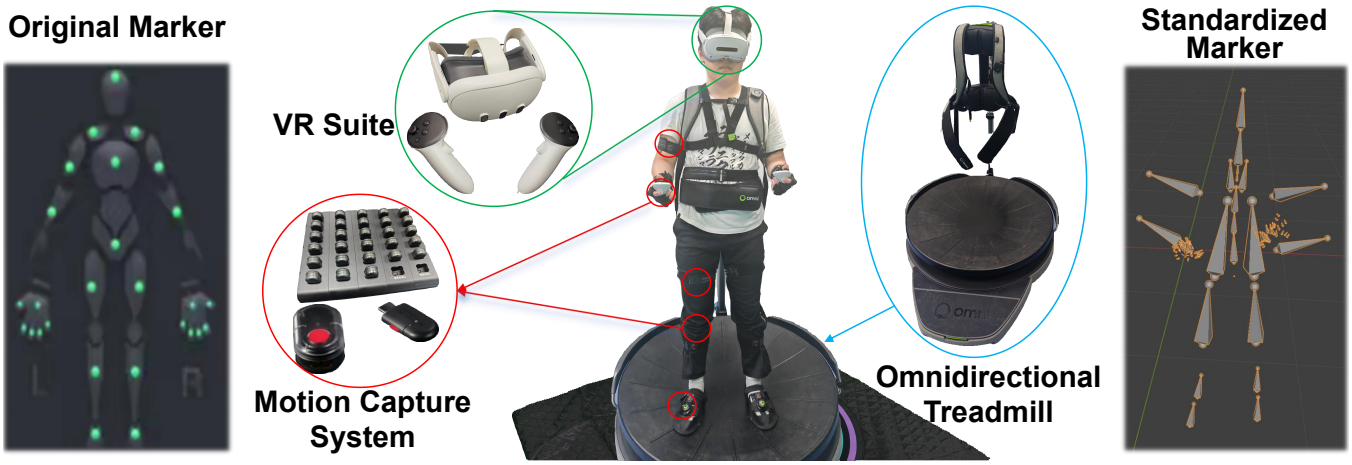

Figure 1: Visualization of the data acquisition equipment. The left part shows a schematic diagram of the wearable motion capture system, which consists of 17 inertial sensors with gyroscopes and two data gloves for capturing full-body and fine-grained finger movements. The middle part presents a real-world example of a participant wearing the complete setup. The right part displays the joint visualization generated by the industrial-grade software Axis Studio after calibration and reconstruction.

loss over long physical distances, we pair this suite with an omnidirectional treadmill [20]. In summary, this hardware combination successfully decouples natural human locomotion from physical space constraints, achieving continuous, high-fidelity kinematic data acquisition. Our core contributions are three-fold:

- **Large-Scale Multimodal Dataset**: We present VRMotion, comprising over **21.6 million precisely synchronized frames** of HD egocentric video and full-body IMU data. This extensive collection provides the essential data foundation for robust cross-modal learning across directive, suggestive, and explorative cognitive tasks within simulated environments [3, 42].
- **Unified Cross-Modal Predictive Framework**: Beyond the dataset, we propose an end-to-end, hierarchical research framework that formally defines the cross-modal synthesis task of mapping egocentric visual stimuli to biomechanically plausible future trajectories. This modular pipeline seamlessly integrates visual encoders with temporal motion dynamics, shifting the paradigm from reactive pose estimation to proactive physical anticipation.
- **Comprehensive Baseline Evaluations and Insights**: We establish a rigorous benchmark by systematically evaluating ten distinct visual-temporal models. Our extensive experiments demonstrate that integrating Large Vision-Language Models (LVLMs) achieves state-of-the-art accuracy, proving the critical role of advanced spatial-geometric understanding in forecasting complex reactive motions.

2 Related Work

2.1 Related Datasets

Motion Capture Datasets. Traditional datasets like CMU Mocap [21] and Human3.6M [19] provide high-quality 3D poses captured via marker-based systems, while HumanEva [34] offers synchronized video for evaluation. However, these datasets typically lack the immersive first-person context critical for VR and miss the extended locomotion patterns essential for navigation in virtual worlds.

Egocentric Vision Datasets. Existing first-person datasets like EPIC-KITCHENS [10] focus on hand-object interactions, while EgoBody [44] captures social interactions using HoloLens 2. Although some datasets explore head movements in VR [9, 39], they omit full-body motion or trajectory data.

Limitations of Existing Work. As shown in Table 1, current datasets either provide only egocentric vision without comprehensive body motion, or lack the causal relationship between visual stimuli and locomotion responses. Additionally, marker-based systems face occlusion issues in confined capture volumes, while RGB-based methods exhibit distance-dependent precision degradation. These constraints limit their applicability to large-scale environments. **VRMotion** addresses these limitations by combining IMU sensors with VR environments and an omnidirectional treadmill (occlusion-free, distance-independent), enabling unconstrained locomotion capture with consistent precision.

2.2 Human Motion Prediction

Human motion prediction has transitioned from statistical models to deep learning paradigms [12, 25, 27]. Early RNN-based methods [12] introduced residual connections to model velocities [27], but often suffered from long-term error accumulation [15]. Graph Convolutional Networks (GCNs) subsequently emerged to explicitly model skeletal structures [25], employing spatial-temporal decomposition and multi-scale representations [23]. Recently, attention mechanisms and transformers [1, 26] have further addressed long-range dependencies, while contemporary research focuses on equivariance constraints [40] and biomechanical consistency [16].

Parallel to these efforts, action anticipation from first-person vision has explored activity forecasting [6], temporal perception [46], and object-centric reasoning [14]. Advanced frameworks have incorporated gaze anticipation [43], reinforcement learning [33],

Table 1: Comparison with Existing Motion and VR Datasets

Dataset	HMD Screen Content	VR-Motion Alignment	Modality*	Movement Tracking		VR Environment	Size (Frames)
				Head Orientation	Locomotion Trajectory		
CMU Mocap [21]	✗	✗	Marker	✗	✗	✗	15.3M
HumanEva [34]	✗	✗	Marker	✗	✗	✗	0.08M
Human3.6M [19]	✗	✗	Marker	✗	✗	✗	3.6M
VR-Behavior [39]	✗	✗	–	✓	✗	✓	26M
TotalCapture [37]	✗	✗	IMU/RGB	✗	✗	✗	1.9M
EPIC-KITCHENS [10]	✓	✗	–	✗	✗	✗	11.5M
DIP-IMU [18]	✗	✗	IMU	✗	✗	✗	0.3M
EGO-CH [31]	✓	✗	–	✗	✗	✗	0.17M
Egobody [44]	✓	✗	RGB	✗	✗	✗	0.59M
VRMN-bD [42]	✓	✗	RGB	✓	✗	✓	0.97M
Questset [3]	✗	✗	RGB	✓	✗	✓	N/A
Movement & Traffic [4]	✗	✗	RGB	✓	✗	✓	N/A
VRMotion (Ours)	✓	✓	**IMU**	✓	✓	✓	**21.6M**

* Modality refers to the primary hardware or method for pose capture: Marker-based, Inertial Measurement Units (IMU), or CV-based (RGB). VRMotion is the first dataset that captures the causal relationship between VR visual stimuli and full-body responses through precise temporal alignment.

and future localization [29]. Standardized benchmarks like EPIC-KITCHENS [10, 13] have advanced discrete action prediction, yet continuous full-body trajectory forecasting remains a challenge.

While significant progress has been made in predicting motion from motion, predicting movements from visual stimuli in VR remains largely unexplored. Our work addresses this gap by providing a data foundation for cross-modal prediction, requiring an understanding of how specific virtual cues trigger distinctive locomotion patterns and gestural responses.

3 VRMotion Dataset

The VRMotion dataset is designed to provide high-fidelity, synchronized multimodal data of human motion in response to immersive VR stimuli. Our dataset encompasses over 21.6 million aligned frames, capturing the complex causal relationship between egocentric visual input and full-body kinematic responses. All data collection procedures were approved by the Institutional Review Board of The Chinese University of Hong Kong, Shenzhen (IRB No. CUHKSZ-D-20250059), and all participants provided written informed consent prior to participation.

3.1 Hardware Setup

Our acquisition system integrates three primary hardware components working in concert to produce synchronized multimodal data, as illustrated in Fig. 1:

- **Motion Capture System:** We implemented a professional-grade inertial motion capture system featuring 27 wireless IMU sensors strategically positioned on participants' bodies according to biomechanical landmarks. Each sensor integrates a triaxial gyroscope, accelerometer, and magnetometer, sampling at 60Hz with 0.1° rotational accuracy as specified by the hardware manufacturer.
- **Omnidirectional Treadmill:** We integrated the Virtuix Omni One treadmill, featuring a low-friction concave platform that allows participants to walk and run while remaining physically stationary. It captures walking direction, speed, and acceleration at 100Hz through joystick emulation.
- **VR System:** We selected the Meta Quest 3 and Pico 4 Ultra as our VR platform, featuring dual displays at 2160 × 2160 pixels per eye, a 90Hz refresh rate, and precise inside-out tracking.

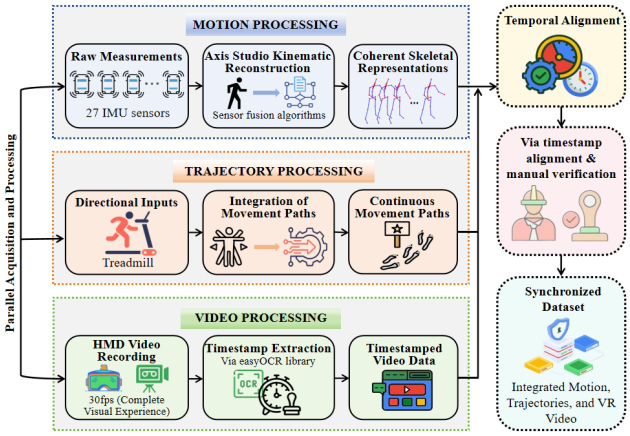

Figure 2: Data processing pipeline showing parallel acquisition of motion capture data and VR video data, locomotion trajectory recording, followed by temporal synchronization. The pipeline ensures high-fidelity capture of both human motion, movement trajectories, and immersive visual context.

3.2 Data Processing Pipeline

As shown in Fig. 2, our data collection pipeline involves the parallel acquisition and processing of motion capture data, locomotion trajectories, and VR video data, which are subsequently integrated through a rigorous synchronization and verification phase:

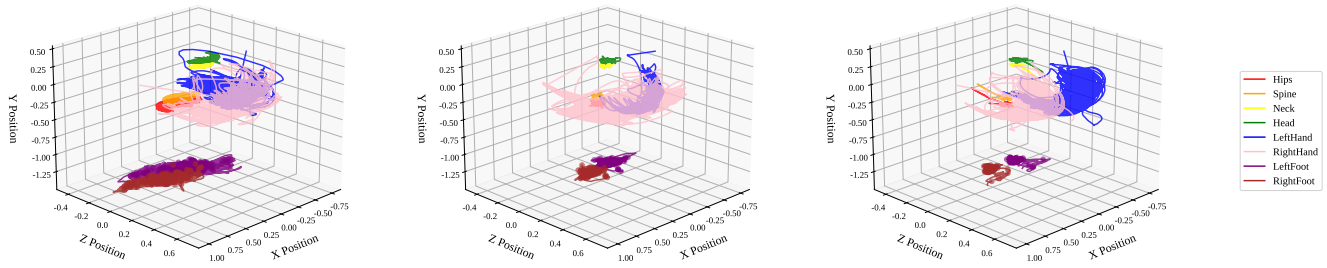

Figure 3: 3D joint trajectories (left to right: Blade & Sorcery, Table Tennis, Beat Saber). Spatial paths illustrate operative workspaces and coordination patterns across segments.

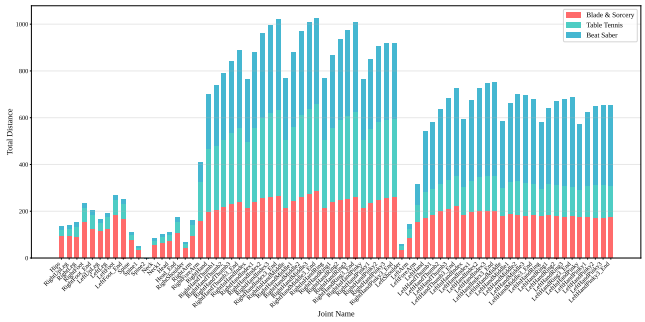

Figure 4: Cumulative per-joint total displacement across all tasks. The stacked bars quantify the combined joint-wise distances, with color segments representing the proportional contributions of Blade & Sorcery, Table Tennis, and Beat Saber, highlighting task-specific movement emphasis.

(1) **Motion Processing:** Raw measurements from 27 IMU sensors are captured at 60Hz, followed by kinematic reconstruction using Axis Studio's sensor fusion algorithms to generate coherent skeletal representations.
(2) **Trajectory Processing:** Directional inputs from the treadmill are integrated into continuous movement paths that accurately represent participants navigational intentions within the virtual environment.
(3) **Video Processing:** HMD Video Recording captures the participant's complete visual experience at 30fps, with timestamps extracted via the easyOCR library.
(4) **Temporal Alignment and Manual Verification:** Multimodal data streams are temporally aligned using the extracted timestamps, followed by rigorous manual verification to filter artifacts and guarantee reliable causal synchronization.

3.3 Task Categories

We design a complexity gradient across three task categories:

- **Directive (Beat Saber[5]):** Highly directive tasks with explicit visual cues that strongly correlate with expected user responses, primarily focusing on upper body movements in a relatively stationary position.
- **Suggestive (Table Tennis[11]):** Suggestive tasks providing contextual cues that require interpretation and personalized responses, introducing moderate locomotion requirements [34].
- **Explorative (Blade & Sorcery[38]):** Explorative tasks with open-ended interaction and minimal constraints, requiring substantial environmental awareness and spatial cognition enabled by the treadmill.

3.4 Dataset Analysis

We quantify motion characteristics across representative tasks using four complementary measurements (Figs. 3-5) to characterize the distinct motion patterns and reveal the statistical associations underlying full-body responses. More detailed dataset analysis can be seen in our Supplemental Materials.

Locomotion Fidelity and Workspace Scaling: 3D trajectories (Fig. 3) reveal task-specific spatial organization and the impact of hardware on movement realism. Blade & Sorcery exhibits enlarged trajectory envelopes across all joints, particularly the lower limbs, which is a direct result of natural locomotion on the omni-directional treadmill. Unlike controller-based datasets, VRMotion captures these high-fidelity navigational intentions, providing rich targets for models trained to forecast future trajectories from visual cues.

Kinematic Dominance and Predictive Weight: The stacked cumulative displacement (Fig. 4) identify segments that dominate overall movement, directly impacting prediction error weighting. As reflected in the proportional segments of the bars, in the directive Beat Saber task, hand joints overwhelmingly lead with displacements 18.3x that of body joints. Conversely, the explorative Blade & Sorcery task distributes effort more broadly (hands 2.2x body joints) due to integrated locomotion. These profiles indicate that predictive precision in high-displacement joints is critical for system-level performance.

Gaze-Trajectory Synergy: Analysis of the explorative scenario (Fig. 5) reveals a consistent speed-gaze trade-off: intensive exploratory scanning coincides with reduced forward speed, while speed increases when gaze aligns with the locomotion path. This relationship establishes egocentric field-of-view (FoV) orientation as a powerful contextual cue for predicting near-future velocity and path commitment, enabling the cross-modal predictive capabilities that form the core of the VRMotion framework.

4 Methodology

This section describes the **VRMotion** framework, an end-to-end modular system designed to anticipate long-term human motion in immersive VR scenarios. The core philosophy of VRMotion is

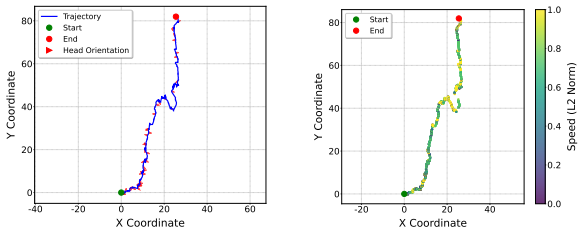

Figure 5: Randomly sampled exemplar trial in the exploratory scenario illustrating the relationship between viewing behavior and locomotion dynamics. (a) Path & FoV Divergence. (b) Locomotion Speed.

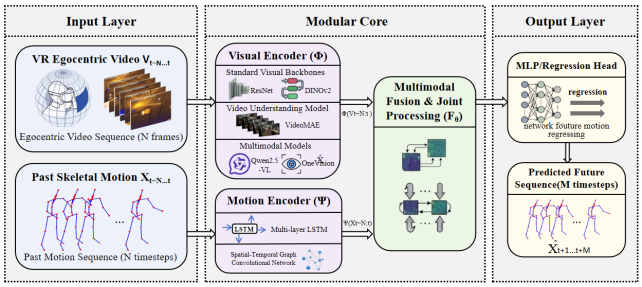

Figure 6: The proposed VRMotion framework. The architecture utilizes a modular core to fuse VR egocentric video and past motion inputs, ultimately regressing future 3D skeletal sequences.

to treat motion prediction as a cross-modal synthesis task, where egocentric visual stimuli and proprioceptive history are fused to generate bio-mechanically plausible future trajectories.

4.1 Framework Overview

The VRMotion framework is structured into a hierarchical pipeline consisting of three primary stages: the *Input Layer*, the *Modular Core*, and the *Output Layer*, as illustrated in Figure 6.

The **Input Layer** handles multimodal data streams: a sequence of VR egocentric video frames $V_{t-N\ldots t}$ and the corresponding past skeletal motion $X_{t-N\ldots t}$. The **Modular Core** serves as the central processing unit, where visual features are extracted and integrated with temporal motion dynamics. Finally, the **Output Layer** transforms the fused latent representations into a future motion sequence $\hat{X}_{t+1\ldots t+M}$. Formally, the system aims to optimize the parameters θ of a mapping function $\mathcal{F}$ such that:

$$\hat{X}_{t+1:t+M} = \mathcal{F}_\theta\left(\Phi(V_{t-N:t}), \Psi(X_{t-N:t})\right) \tag{1}$$

where Φ and Ψ denote the encoding operations for visual and motion modalities, respectively.

4.2 Visual Encoder & Temporal Modeling

The **Modular Core** is designed with high flexibility, allowing for the independent optimization of visual and temporal components. This modularity is essential for adapting to different VR tasks and computational constraints.

The **Visual Encoder** is responsible for extracting high-level semantic and spatial-geometric features from the raw VR video frames. We adopt Large Vision-Language Models (LVLMs), specifically the **Qwen2.5-VL** backbone, as our primary visual extractor. Unlike traditional CNNs, the LVLM-based encoder leverages a massive pre-trained parameter space to understand complex 3D environments, such as identifying the trajectories of virtual objects and their spatial relationship with the user. The extracted features are passed through a projection layer to align the vision-domain dimensions with the temporal-modeling requirements, resulting in a condensed visual stimuli representation $f_{vis} \in \mathbb{R}^D$.

The **Temporal Modeling** component acts as the temporal reasoning engine that integrates the visual stimuli with the user's historical movement momentum. Within this modular component, we consider two distinct architectural approaches:

- **Recurrent Modeling (LSTM):** This approach treats the prediction as a sequential regression task. By maintaining a hidden state that evolves over time, the LSTM module captures the temporal dependencies within the motion sequence while prioritizing new visual cues extracted from the VR scene. This ensures that the generated motion is continuous and maintains physical momentum.
- **Graph-based Spatio-Temporal Modeling (ST-GCN):** Alternatively, we model the human body as a spatio-temporal graph where the 24 joints represent nodes and the bones represent edges. The ST-GCN [41] module performs graph convolutions across both spatial and temporal dimensions, explicitly enforcing the skeletal topology and bio-mechanical constraints. This allows the system to focus on the interconnected nature of human joints during complex reactive movements.

The fused information from these two components is then passed to the Output Layer, which utilizes a series of fully connected layers to regress the final 3D coordinates for the entire future horizon M. This decoupled design allows the VRMotion framework to effectively bridge the gap between "seeing" a virtual event and "predicting" the subsequent physical response.

5 Experiments

In this section, we conduct a comprehensive evaluation of the **VRMotion** framework using the proposed benchmark. We analyze the performance of ten distinct visual-temporal combinations to demonstrate the effectiveness of Large Vision-Language Models (LVLMs) in anticipating complex human motions within VR environments.

5.1 Experimental Setup

Implementation Details. All baselines are initialized with pre-trained weights. Specifically, LVLM variants (Qwen2.5-VL-7B [2], OneVision [22]) utilize 8-bit quantization for LSTM models and FP16 for ST-GCN [41] models to ensure gradient stability. The models undergo supervised training on VRMotion for 100 epochs using the AdamW optimizer [24] on RTX 4090 GPUs, with LVLM backbones following a full-parameter instruction-tuning paradigm [45]. LSTM-based models use a constant learning rate of 10^{-4}, while ST-GCN-based models adopt 5×10^{-5} with a Cosine Annealing scheduler. Models are trained with a batch size of 8–32 and 2-step

Table 2: Quantitative comparison of ten visual-temporal baseline models on the VRMotion dataset. Models are grouped by their temporal modeling head and sorted by MPJPE. Bold indicates the best performance in each category.

Group	Visual Encoder	Temporal Model	MPJPE (mm) ↓	PA-MPJPE (mm) ↓	Vel Error (mm/f) ↓
LSTM-based	**Qwen2.5-VL** [2]	**LSTM** [27]	**44.25**	**40.47**	10.87
	OneVision [22]	LSTM [27]	48.17	44.36	**10.05**
	DINOv2 [28]	LSTM [27]	54.66	48.63	12.38
	ResNet [17]	LSTM [27]	63.48	52.92	12.43
	VideoMAE [36]	LSTM [27]	90.46	64.83	15.04
ST-GCN-based	**Qwen2.5-VL** [2]	**ST-GCN** [41]	**127.62**	**96.53**	**48.49**
	OneVision [22]	ST-GCN [41]	202.84	99.91	104.26
	DINOv2 [28]	ST-GCN [41]	215.46	113.36	108.26
	VideoMAE [36]	ST-GCN [41]	268.80	144.12	108.44
	ResNet [17]	ST-GCN [41]	409.68	252.03	206.56

gradient accumulation to predict a 16-frame future horizon (0.53 seconds).

Baseline Models. We evaluate a range of state-of-the-art vision backbones paired with different temporal modeling heads. The visual encoders include general-purpose models (ResNet, DINOv2), video-centric models (VideoMAE), and advanced LVLMs (OneVision, Qwen2.5-VL). Each encoder is tested with two types of temporal heads: a recurrent Long Short-Term Memory (LSTM) network and a topology-aware Spatio-Temporal Graph Convolutional Network (ST-GCN).

Evaluation Metrics. Following the standard protocols in human motion analysis, we evaluate the benchmark across three dimensions:

- **MPJPE (mm):** Mean Per-Joint Position Error, measuring the average Euclidean distance between predictions and ground truth.
- **PA-MPJPE (mm):** Procrustes-aligned MPJPE to assess pose similarity regardless of global orientation.
- **Vel Error (mm/f):** Velocity Error, assessing the temporal smoothness of generated sequences.

5.2 Quantitative Results

Table 2 presents a quantitative comparison of ten baseline models on the VRMotion dataset. The results reveal a clear performance hierarchy across different architectures.

LVLM Superiority: Our fine-tuned **Qwen2.5-VL + LSTM** model achieves the state-of-the-art accuracy with an MPJPE of **44.25 mm**. This represents a significant margin over traditional CNN backbones like ResNet (63.48 mm) and video-native models like VideoMAE (90.46 mm), proving that the massive pre-trained spatial knowledge in LVLMs is crucial for interpreting causal links between complex VR stimuli and human motor responses.

Temporal Head Effectiveness (LSTM vs. ST-GCN): Notably, LSTM heads consistently outperform ST-GCN across all visual backbones. For instance, Qwen2.5-VL degrades from an MPJPE of 44.25 mm (LSTM) to 127.62 mm (ST-GCN). We attribute this to a classic domain mismatch: modern visual encoders, especially LVLMs, extract highly abstract, global semantic features. While LSTM serves as a robust implicit sequence decoder that effectively maps these representations to motion trajectories, ST-GCN explicitly forces them into a rigid 24-joint physical graph. This rigid spatial routing disrupts holistic semantic flows, indicating that implicit recurrent modeling is more compatible with large-scale vision foundation models for highly dynamic, reactive tasks.

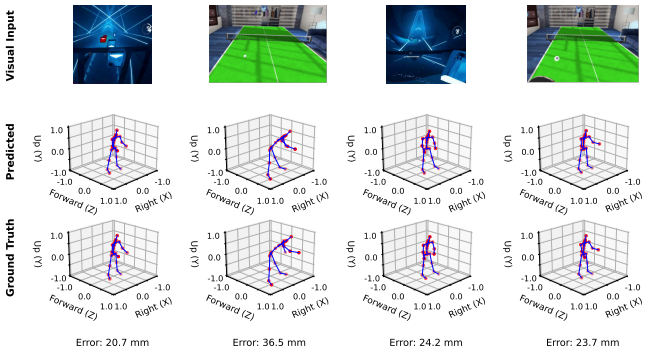

Figure 7: Qualitative results of our VRMotion framework across four test cases. The Predicted pose exhibits high fidelity to the Ground Truth even at a 0.53s future horizon, with errors ranging from 20.7 mm to 36.5 mm.

5.3 Qualitative Results

To intuitively demonstrate the predictive capability of our framework, we present qualitative results in Fig. 7. The visualizations across four representative cases illustrate how the model accurately aligns future poses with the visual input.

As shown in the second and third rows of Fig. 7, our model maintains biomechanical consistency, avoiding joint distortion even during complex movements. For instance, in the first case, the model achieves a minimum error of **20.7 mm**, capturing the precise orientation of the player's arms as they prepare to interact with upcoming virtual blocks. The structural alignment between the predicted skeleton and the ground truth indicates that the framework has successfully learned the underlying intent behind the visual cues provided by the VR egocentric view.

6 Conclusion

In this paper, we introduce **VRMotion**, the first large-scale multimodal dataset designed to bridge the gap between immersive visual stimuli and corresponding full-body human kinematic responses. The dataset encompasses over **21.6 million precisely aligned frames** across three task categories—directive, suggestive, and explorative—captured through the innovative integration of 27 wireless IMU sensors and an omnidirectional treadmill. We conduct a systematic benchmark evaluation of **ten distinct combinations** of visual backbones and temporal heads, establishing a rigorous foundation for cross-modal motion prediction. Our findings demonstrate that Large Vision-Language Models (LVLMs), particularly Qwen2.5-VL, exhibit superior spatial-geometric understanding and achieve state-of-the-art performance in anticipating complex reactive motions. By synchronizing VR visual input with natural locomotion and gaze dynamics, this work effectively transitions ego-motor learning from reactive to predictive paradigms.

References

[1] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. 2021. A Spatio-Temporal Transformer for 3D Human Motion Prediction. In *Proceedings of the International Conference on 3D Vision*. 565–574.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923* (2025).

[3] Sara Baldoni, Federica Battisti, Federico Chiariotti, Fabio Mistrorigo, Alfi Baqiatus Shofi, Paolo Testolina, Alessandro Traspadini, Andrea Zanella, and Michele Zorzi. 2024. Questset: A VR Dataset for Network and Quality of Experience Studies. In *Proceedings of the ACM Multimedia Systems Conference*. 408–415.

[4] Sara Baldoni, Salim Benhamadi, Federico Chiariotti, Michele Zorzi, and Federica Battisti. 2025. Movement-and Traffic-Based User Identification in Commercial Virtual Reality Applications: Threats and Opportunities. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces*. 72–81.

[5] Beat Games. 2019. Beat Saber. Video game. Available at: https://beatsaber.com/.

[6] Syed Zahir Bokhari and Kris M Kitani. 2016. Long-Term Activity Forecasting Using First-Person Vision. In *Proceedings of the Asian Conference on Computer Vision*. 346–360.

[7] Gianni Bremer, Niklas Stein, and Markus Lappe. 2021. Predicting Future Position from Natural Walking and Eye Movements with Machine Learning. In *Proceedings of the IEEE International Conference on Artificial Intelligence and Virtual Reality*. 19–28.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[9] Xavier Corbillon, Francesca De Simone, and Gwendal Simon. 2017. 360-Degree Video Head Movement Dataset. In *Proceedings of the ACM on Multimedia Systems Conference*. 199–204.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *Proceedings of the European Conference on Computer Vision*. 720–736.

[11] For Fun Labs. 2016. Eleven Table Tennis. Video game. Available at: https://elevenvr.com/.

[12] 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.

[13] Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. 2018. Leveraging Uncertainty to Rethink Loss Functions and Evaluation Measures for Egocentric Action Anticipation. In *Proceedings of the European Conference on Computer Vision Workshops*.

[14] Antonino Furnari, Sebastiano Battiato, Kristen Grauman, and Giovanni Maria Farinella. 2017. Next-Active-Object Prediction from Egocentric Videos. *Journal of Visual Communication and Image Representation* 49 (2017), 401–411.

[15] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. 2017. Learning Human Motion Models for Long-Term Predictions. In *Proceedings of the International Conference on 3D Vision*. 458–466.

[16] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. 2023. Back to MLP: A Simple Baseline for Human Motion Prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4809–4819.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[18] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. 2018. Deep Inertial Poser: Learning to Reconstruct Human Pose from Sparse Inertial Measurements in Real Time. *ACM Transactions on Graphics* 37, 6 (2018), 185:1–185:15.

[19] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. 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 (2013), 1325–1339.

[20] Julian Kreimeier, Sebastian Hammer, Timo Götzelmann, Andreas Braun, and Philipp Agethen. 2020. Evaluating the User Experience of Omnidirectional VR Walking Simulators. In *Proceedings of the International Conference on Multimodal Interaction*. 103–111.

[21] CMU Graphics Lab. 2007. CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu/ Accessed: 2025-05-24.

[22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-Onevision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326* (2024).

[23] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. 2020. Dynamic Multiscale Graph Neural Networks for 3D Skeleton Based Human Motion Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 214–223.

[24] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations*.

[25] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. 2019. Learning Trajectory Dependencies for Human Motion Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9489–9497.

[26] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. 2021. Multi-Level Motion Attention for Human Motion Prediction. *International Journal of Computer Vision* 129, 9 (2021), 2513–2535.

[27] 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.

[28] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Namya Goel, Alexey Dosovitskiy, Mahmoud Sayed, Natalia Fernandez, Tao Chu, Guillaume Pernot, et al. 2023. Dinov2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193* (2023).

[29] Hyun Soo Park, Jyh-Jing Hwang, Yedong Niu, and Jianbo Shi. 2016. Egocentric Future Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4697–4705.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*. PMLR, 8748–8763.

[31] Francesco Ragusa, Antonino Furnari, Sebastiano Battiato, Giovanni Signorello, and Giovanni Maria Farinella. 2020. EGO-CH: Dataset and Fundamental Tasks for Visitors Behavioral Understanding Using Egocentric Vision. *Pattern Recognition Letters* 131 (2020), 150–157.

[32] Lisa Rebenitsch and Charles Owen. 2016. Review on Cybersickness in Applications and Visual Displays. *Virtual Reality* 20, 2 (2016), 101–125.

[33] Nicholas Rhinehart and Kris M Kitani. 2017. First-Person Activity Forecasting with Online Inverse Reinforcement Learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 3696–3705.

[34] Leonid Sigal, Alexandru O Balan, and Michael J Black. 2010. Humaneva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *International Journal of Computer Vision* 87, 1 (2010), 4–27.

[35] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. 2021. Habitat 2.0: Training Home Assistants to Rearrange Their Habitat. In *Advances in Neural Information Processing Systems*, Vol. 34. 251–266.

[36] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *Advances in Neural Information Processing Systems*, Vol. 35. 10078–10091.

[37] Matt 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 the British Machine Vision Conference*.

[38] WarpFrog. 2018. Blade & Sorcery. Video game. Available at: https://www.warpfrog.com/.

[39] Chenglei Wu, Zhihao Tan, Zhi Wang, and Shiqiang Yang. 2017. A Dataset for Exploring User Behaviors in VR Spherical Video Streaming. In *Proceedings of the ACM on Multimedia Systems Conference*. 193–198.

[40] Chenxin Xu, Robby T Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, and Yanfeng Wang. 2023. Eqmotion: Equivariant Multi-Agent Motion Prediction with Invariant Interaction Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1410–1420.

[41] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[42] He Zhang, Xinyang Li, Yuanxi Sun, Xinyi Fu, Christine Qiu, and John M Carroll. 2024. VRMN-bD: A Multi-Modal Natural Behavior Dataset of Immersive Human Fear Responses in VR Stand-Up Interactive Games. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces*. 320–328.

[43] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. 2017. Deep Future Gaze: Gaze Anticipation on Egocentric Videos Using Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4372–4381.

[44] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. 2022. Egobody: Human Body Shape and Motion of Interacting People from Head-Mounted Devices. In *Proceedings of the European Conference on Computer Vision*. 180–200.

[45] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 400–410.

[46] Yipin Zhou and Tamara L Berg. 2015. Temporal Perception and Prediction in Ego-Centric Video. In *Proceedings of the IEEE International Conference on Computer Vision*. 4498–4506.