Proceedings of CONF-SEML 2026 Symposium: Importance of Machine Learning Methods and Analysis in Engineering
DOI: 10.54254/2755-2721/2026.AS32001

A Survey on Data-Driven Human Motion Prediction: From Deterministic Modeling to Generative Interaction

Wei Wei

School of Airspace Science and Engineering, Shandong University, Weihai, China
202300620418@mail.sdu.edu.cn

Abstract. Human motion prediction is a fundamental task in computer vision and computer graphics, aiming to accurately model temporal dependencies and multi-constraint coordination for applications including autonomous driving pedestrian prediction, human-machine collaboration, and virtual simulation. The field faces three core challenges: long-term prediction pose drift from temporal error accumulation, inherent motion multi-modal uncertainty, and difficult complex scene interaction modeling under extreme data scarcity. This paper synthesizes the field's technological evolution from deterministic to generative methods, focusing on physical prior injection and interaction mechanism design. It reveals its shift from pure data-driven kinematic regression to a data-prior integrated dynamic generative paradigm, laying a foundation for plausible predictions in highly constrained scenarios.

Keywords: Motion, Prediction, Generative, Computervision

1. Introduction

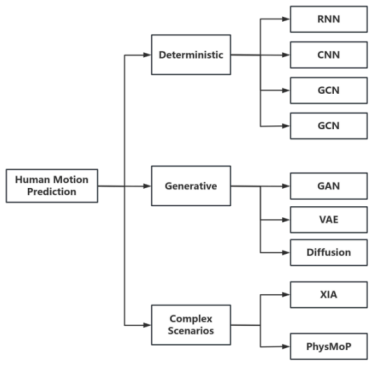

Figure 1. Human motion prediction methodology technical evolution framework

Human motion prediction figures out future motion sequences using historical postures. It's not the same as conventional coordinate regression. It needs both geometric and physical constraints. Geometric constraints [1] keep the skeleton whole and consistent. Physical constraints [2] make sure motion is smooth and makes mechanical sense. Their combined effect supports prediction reliability.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.AS32001

It also backs applications like autonomous driving [3], human-computer interaction [4], and pedestrian tracking [5]. It's a core cross-disciplinary task. It spans computer vision and intelligent systems. It's vital for relevant scenario safety and efficiency. Technological progress directly impacts real-world application reliability and naturalness.

We classify motion prediction tasks along two separate dimensions. This helps with later method analysis. One dimension is scene complexity. It includes single-person basic scenarios [6]. These focus on temporal consistency and basic constraints. There are no complex interactions here. The other is multi-person/scene interaction scenarios [7,8]. These need modeling of high-dimensional dependency and semantic constraints. The other dimension is prediction paradigm. It includes deterministic prediction [9]. This outputs a single sequence with high geometric accuracy. Generative prediction [10-13] is also part of it. It catches uncertainty in motion through

multiple reasonable multimodal sequences. Progress in this field has been obvious. Even so, three key challenges still exist. Long-term prediction has problems with errors piling up over time and pose drift. Human motion has many different modes. A single prediction can't cover all valid possibilities. Modeling complex interactions is hard. Extreme action data is scarce. These two issues limit model generalization.

Three key contributions come from our paper. Systematic review and deep analysis make this possible. The goal is to fix these challenges. Mainstream architectures' evolution gets a thorough look. The change from deterministic to generative models is tracked. The shift from purely data-driven to data-prior integrated methods is tracked too. Solutions for complex scenarios are analyzed. A common way to deal with them is pulled out. This way brings together interaction mechanism design and physical prior injection. Criteria for matching scenarios and models are set up. A layered

evaluation system is built too. It goes by the accuracy-quality-reasonableness framework. Metric and dataset rules are standardized. Principles for loss function design are added. Cross-study comparisons become easier because of this.

2. Problem formulation & representation

2.1. Problem definition

From the perspective of task essence, the core objective of human motion prediction is to infer a sequence of movements compliant with objective laws over a future time interval based on historically observed posture information. This process requires standardization through formal mathematical expressions and explicit constraints: from a mathematical perspective, the human motion prediction task can be formulated as follows, given a historical sequence of human poses $S_{obs} = P_1, P_2, \ldots, P_{T_{obs}}$ (where $P_t \in \mathbb{R}^{K\times 3}$, K being the number of key joints in the human body), the goal is to infer a future sequence of poses $S_{pred} = P_{T_{obs}+1}, \ldots, P_{T_{obs}+T_{pred}}$. Specifically, the input comprises a historical sequence of T frames of observed data, with the representation format for single-frame poses adaptable to diverse research requirements. The model output delivers a predefined sequence of K frames of predicted future poses.

Output sequences need to make sense. Prediction results have to meet three key constraints. Anatomical constraints set a rule. Skeletal length errors don't go beyond reasonable limits. Joint movement angles stay in physiological ranges too. That way, structurally wrong poses are avoided. Physical constraints ask for motion. It has to stick to basic rules. Things like momentum conservation and smooth angular velocity count. Joint angular velocity changes have to stay stable. Semantic constraints want actions. They have to fit scene interaction logic. Take human-object

DOI: 10.54254/2755-2721/2026.AS32001

contact scenarios. Key interactive joints and objects have to keep proper distances. Those distances match real-world interaction physics.

2.2. Motion representation

Motion representation is the key link for task definition and model training. Which one you pick changes how well a model can catch those constraints we talked about. Mainstream methods fall into two kinds. That's pose parameterization and structural prior encoding. Pose parameterization focuses on math descriptions for single-frame poses. 3D joint coordinates in absolute (x,y,z) form are easy to get. They also run fast when doing calculations. They can't pick up

rigid joint relationships, though. That brings on the "rubber man" problem. Bone lengths get weird. They go against anatomical constraints when making predictions. Representations based on joint rotational states fix this issue. We use quaternions [14] instead of Euler angles. They stop the gimbal lock that Euler angles naturally have [15]. Quaternions keep interpolation running smoothly. They also keep calculations quick. They even out rotational smoothness and computational continuity. They're the best pick for today's SOTA models. Look at QuaterNet [16]. It got way better pose prediction accuracy by using quaternion representation. Single-frame pose parameterization isn't everything. The human body has a natural tree-like hierarchical skeleton. It starts right at the spine. Rigid parent-child joint connections make up limbs and torso parts. This skeleton matters a lot for motion representation. We've made clear encoding strategies. They add this prior knowledge in. ARNet [17] sees the skeleton as a graph. It uses graph convolutional networks. These networks learn weights for parent-child joint connections. This cuts down on skeleton stretching in extreme poses. Recent work like MagicArticulate [18] shows something. Hierarchical processing based on the skeleton's natural structure makes learning joint relationships

faster. This shows how important structural prior encoding is. Picking a motion representation needs balance. It's between constraint adaptability and model compatibility. Quaternions work best for pose parameterization. They're good at showing rotations. Structural prior encoding methods like ARNet do one job. They make sure we follow anatomical constraints. Putting them together is now the main way to do representation. It's used for high-precision human motion prediction tasks.

3. Methodologies

Human motion prediction methodology evolves around three core objectives: enhancing accuracy, capturing diversity, and adapting to complex scenarios. It moves from early deterministic single-sequence models to generative multimodal approaches, to complex solutions integrating interaction and physical laws, forming a clear technological trajectory.

3.1. Deterministic temporal modeling

Deterministic temporal modeling aims to output a single precise future sequence, focusing on resolving temporal dependency and joint spatial structure modeling challenges to lay a foundation for human pose prediction advancements; its evolution progresses from early temporal correlation establishment to later efficiency, accuracy, and structural adaptation optimizations, with each generation addressing prior architecture limitations. Recurrent neural networks [19] laid the foundation, and their Seq2Seq derivative combined with LSTM [20] enabled the first end-to-end pose sequence prediction, but they suffer from inefficient serial computation and severe long-term pose drift, which residual connections [21] failed to fully resolve. Temporal convolutional networks

DOI: 10.54254/2755-2721/2026.AS32001

[22] resolved the serial computation bottleneck [23] via parallel temporal feature extraction but lack long-range temporal dependency modeling due to local receptive fields; MoPredNet [24] introduced deformable spatio-temporal convolutions yet still failed to break through long-term prediction accuracy and pose drift limitations. Graph Convolutional Networks [25] are compatible with skeletal structures but rely on manual adjacency matrices limiting generalization, which LTD-GCN [26] addressed with learnable adjacency matrices for accurate joint spatial dependency modeling. Transformer [27] resolved prior local dependency issues via self-attention but faces computational explosion with longer sequences, unsuitable for resource-constrained scenarios; MetaFormer [28] optimized this by simplifying architecture to reduce redundancy yet sacrificed flexibility and underperforms in complex motion scenarios. These innovations target core bottlenecks, support temporal coherence and structural consistency in constrained scenarios, and drive the field from local adaptation to global optimization.

3.2. Generative & probabilistic approaches

The limitations of mean-reverting deterministic models have made generative and probabilistic approaches the core for capturing motion multimodal uncertainty. Both model motion probability distributions via distinct technical paths, evolving from early adversarial training to contemporary diffusion-based generation with enhanced diversity, realism and plausibility. GANs [29] have inherent flaws: mode collapse, unstable training leading to gradient vanishing or oscillation, no explicit probability distribution, and high hyperparameter sensitivity; Social GAN introduced social pooling to mitigate group action collisions but failed to resolve training instability and limited expressiveness in complex motion scenarios. VAEs [30], another key generative branch for motion prediction, have drawbacks: fixed simple priors constrain inference and latent space expressiveness, easily inducing posterior collapse that results in low-fidelity, monotonous motions and aggravated mode collapse. DLow optimized reconstruction-distribution balance but cannot fundamentally overcome VAEs' poor motion detail control, with generated variations lacking flexibility in complex action scenarios.As current SOTA, diffusion models outperform GANs and VAEs but have critical flaws: iterative denoising leads to slow sampling and high computational overhead, making them incompatible with real-time and resource-constrained scenarios; high sensitivity to hyperparameters and noise schedules triggers semantic drift or structural deformation in generated motions. Their optimized variants also have limitations: MoFlow [31] enhances multimodal capture at the cost of motion plausibility to achieve faster sampling; MemoNet [32] boosts long-term consistency via memory but increases model complexity and training costs; EqMotion [33] ensures physical coherence with symmetry priors yet lacks adaptability to asymmetric complex motions; SingularTraj [34] accelerates sampling via dimensionality reduction but risks the loss of fine-grained motion details; MID integrates physical constraints but struggles to balance constraint strictness and generation diversity in dynamic interaction scenarios.

3.3. Interaction & dynamics in complex scenarios

High-dimensional interaction coupling and sudden physical disturbances are core obstacles to the practical deployment of human motion prediction; for three typical complex scenarios including multi-person strong interaction, scene perception, and physical dynamics, researchers have developed specialized solutions, with core limitations residing in balancing constraint strictness and adaptability plus limited generalization to unforeseen scenarios. In multi-person scenarios, methods such as XIA [35] exhibit distinct drawbacks: limited generalization to non-predefined random

DOI: 10.54254/2755-2721/2026.AS32001

unstructured interactions, degraded performance under motion speed-induced asynchrony due to overreliance on precise historical sequence alignment, and restricted real-time deployment caused by high computational complexity from bidirectional interaction features. In scene-aware scenarios, key deficiencies impede practicality: ScenePerturb [36] relies heavily on semantic segmentation accuracy, with inaccurate masks inducing invalid spatial constraints, model penetration, or floating artifacts; Contact-aware Forecasting [37] is incompetent in dynamic interactions as its static joint-scene graph cannot update for fast-changing scenes and inflexible interaction intensity results in rigid or loose constraints, and both lack adaptability to dynamic complex scenes. In extreme perturbation modeling, methods such as PhysMoP [38] have prominent deficiencies: dynamic adjustment of fusion weights between data-driven and physics-based models is challenging, resulting in either constrained diversity or mechanically implausible motions; heavy reliance on manual physical priors limits adaptability, dual-model fusion increases training costs and slows inference, and their fusion design still fails to balance mechanical plausibility and motion naturalness.

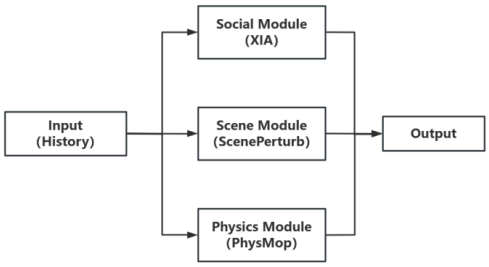

Figure 2. Multi-module architecture for human motion prediction in complex scenarios

4. Datasets and evaluation

4.1. Datasets

Table 1. Summarizes the mainstream datasets used in this field, categorized by scenario complexity

Dataset Type	Dataset Name	Core Features
Basic/Single-person	Human3.6M [39]	Classic indoor benchmark, standardized
	CMU Mocap [40]	Infrared capture, single-person focus
	HumanEva-I [41]	Motion/pose estimation, multi-view suitable
Generative/Large Model	AMASS [42]	Standardized DB, joint shape-pose info
	3DPW [43]	1st mobile in-the-wild, real-scene
Interactive/Complex	ExPI	Two-person interaction, robustness

4.2. Evaluation

To comprehensively assess model performance, we employ a layered metric system. Geometric accuracy is measured via MPJPE and PA-MPJPE for pose precision, along with FDE to quantify long-term drift [44]. Generation quality is evaluated using FID [45] for distribution distance and Diversity metrics to detect mode collapse. Furthermore, physical plausibility is validated through SLE for anatomical consistency, Penetration Rate for scene interaction validity, and Physics Score [46] for dynamic compliance.

DOI: 10.54254/2755-2721/2026.AS32001

5. Conclusion and future directions

5.1. Conclusion

This paper reviews human motion prediction's evolution, from data-driven kinematic regression to dynamic generation fusing statistical patterns with physical/interaction priors. Quaternion encoding and hierarchical skeleton modeling dominate for accuracy-anatomical constraint balance; the model trajectory evolves from deterministic temporal to generative modeling, then to complex scenario enhancement. A hierarchical accuracy-quality-reasonableness evaluation system unifies cross-method comparisons, and these breakthroughs extend the field from laboratory validation to real-world high-constraint scenarios, underpinning domain implementation.

Table 2. Core methods and strengths/weaknesses of different methodological categories in human motion prediction

Methodology Category	Core Method	Strengths	Weaknesses
Deterministic Temporal Modeling	RNN	Foundational; end-to-end	Low efficiency; drift
	TCN	Parallel; efficient	Poor spatial adaptability
	GCN	Skeleton-fit; strong generalization	Needs manual adjacency matrix
	Transformer	Global capture; balanced	High vanilla complexity
Generative & Probabilistic Approaches	GAN	Realistic samples	Poor diversity; unstable
	VAE	Probabilistic generation	Mode collapse; limited variety
	Diffusion Models	High quality; physics-compatible	Complex sampling; high cost
Interaction & Dynamics in Complex Scenarios	Multi-person Interaction	Sequence alignment	Multi-person complexity
	Scene-Aware	Avoids penetration; rational	Relies on quality segmentation
	Physical Dynamics	Physics-compliant	Complex fusion

5.2. Future directions

Future research needs to zero in on three frontiers. Mixing physics and data-driven models is key. It gets rid of pose drift in long-term predictions. Unified Interaction Frameworks are a must. They model human-object-scene coupling all at once. LLMs could be used for semantic reasoning. CLIP might help with visual-text alignment too. Dealing with scarce data in extreme scenarios needs a mix. Combine diffusion-based synthesis and few-shot learning strategies. This makes models more robust.

References

[1] Guo, X., & Choi, J. (2019). Human Motion Prediction via Learning Local Structure Representations and Temporal Dependencies. ArXiv, abs/1902.07367.
[2] Zhou, H., Guo, C., Zhang, H., & Wang, Y. (2021). Learning Multiscale Correlations for Human Motion Prediction. 2021 IEEE International Conference on Development and Learning (ICDL), 1–7.
[3] Chen, S., Liu, B., Feng, C., Vallespi-Gonzalez, C., & Wellington, C.K. (2020). 3D Point Cloud Processing and Learning for Autonomous Driving: Impacting Map Creation, Localization, and Perception. IEEE Signal Processing Magazine, 38, 68–86.
[4] Koppula, H., & Saxena, A. (2013). Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation. In International Conference on Machine Learning, 792–800.
[5] Bhattacharyya, A., Fritz, M., & Schiele, B. (2018). Long-term on-board prediction of people in traffic scenes under uncertainty. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4194–4202.

DOI: 10.54254/2755-2721/2026.AS32001

[6] Lehrmann, A.M., Gehler, P.V., & Nowozin, S. (2014). Efficient nonlinear markov models for human motion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1314–1321.
[7] Jeong, J., Park, D., & Yoon, K.J. (2024). Multi-agent long-term 3D human pose forecasting via interaction-aware trajectory conditioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1617–1628.
[8] Wang, J., Xu, H., Narasimhan, M., & Wang, X. (2021). Multi-person 3D motion prediction with multirange transformers. Advances in Neural Information Processing Systems, 34, 6036–6049.
[9] Liu, Z., Wu, S., Jin, S., Liu, Q., Lu, S., Zimmermann, R., & Cheng, L. (2019). Towards natural and accurate future motion prediction of humans and animals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10004–10012.
[10] Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2255–2264.
[11] Aliakbarian, S., Saleh, F.S., Salzmann, M., Petersson, L., & Gould, S. (2020). A Stochastic Conditioning Scheme for Diverse Human Motion Prediction. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5222–5231.
[12] Yuan, Y., & Kitani, K.M. (2020). DLow: Diversifying Latent Flows for Diverse Human Motion Prediction. In European Conference on Computer Vision.
[13] Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., & Bermano, A.H. (2022). Human Motion Diffusion Model. ArXiv, abs/2209.14916.
[14] Mao, Y., Yan, L., Guo, H., Hong, Y., Huang, X., & Yuan, Y. (2023). A Hybrid Human Activity Recognition Method Using an MLP Neural Network and Euler Angle Extraction Based on IMU Sensors. Applied Sciences, 13(18), 10529.
[15] Pavllo, D., Grangier, D., & Auli, M. (2018). QuaterNet: A Quaternion-based Recurrent Model for Human Motion.
[16] Zhu, Y., Chen, C., Yan, G., Guo, Y., & Dong, Y. (2020). AR-Net: Adaptive Attention and Residual Refinement Network for Copy-Move Forgery Detection. IEEE Transactions on Industrial Informatics, 16(10), 6714–6723.
[17] Song, C., Zhang, J., Li, X., Yang, F., Chen, Y., Xu, Z., Liew, J.H., Guo, X., Liu, F., Feng, J., & Lin, G. (2025). MagicArticulate: Make Your 3D Models Articulation-Ready. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15998–16007.
[18] Tang, Y., Ma, L., Liu, W., & Zheng, W. (2018). Long-term human motion prediction by modeling motion context and enhancing motion dynamic. arXiv: 1805.02513.
[19] Martinez, J., Black, M.J., & Romero, J. (2017). On human motion prediction using recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2891–2900.
[20] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.
[21] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778.
[22] Cui, Q., Sun, H., & Yang, F. (2020). Learning Dynamic Relationships for 3D Human Motion Prediction. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6518–6526.
[23] Lea, C., Flynn, M.D., Vidal, R., Reiter, A., & Hager, G.D. (2016). Temporal Convolutional Networks for Action Segmentation and Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 156–165.
[24] Zang, C., Li, M., & Pei, M. (2022). Few-shot human motion prediction using deformable spatio-temporal CNN with parameter generation. Neurocomputing, 513, 46–58.
[25] Mao, W., Liu, M., & Salzmann, M. (2019). Learning Trajectory Dependencies for Human Motion Prediction. In IEEE/CVF International Conference on Computer Vision.
[26] Cai, Y., Huang, L., Wang, Y., Cham, T.J., Cai, J., Yuan, J., Liu, J., Yang, X., Zhu, Y., & Shen, X. (2020). Learning progressive joint propagation for human motion prediction. In Computer Vision–ECCV 2020, 226–242.
[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 30.
[28] Yu, W., Luo, R., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., & Yan, S. (2022). MetaFormer is Actually What You Need for Vision. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10809–10819.
[29] Gui, L.Y., Wang, Y.X., Liang, X., & Zhu, S.C. (2018). Adversarial Geometry-Aware Human Motion Prediction. Springer, Cham.
[30] Kingma, D.P., & Welling, M. (2013). Auto-Encoding Variational Bayes. ArXiv, abs/1312.6114.
[31] Fu, Y., Yan, Q., Wang, L., Li, K., & Liao, R. (2025). MoFlow: One-Step Flow Matching for Human Trajectory Forecasting via Implicit Maximum Likelihood Estimation based Distillation. In 2025 IEEE/CVF Conference on

DOI: 10.54254/2755-2721/2026.AS32001

Computer Vision and Pattern Recognition, 17282–17293.
[32] Xu, C., Mao, W., Zhang, W., & Chen, S. (2022). Remember Intentions: Retrospective-Memory-based Trajectory Prediction. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6478–6487.
[33] Xu, C., Tan, R.T., Tan, Y., Chen, S., Wang, Y., Wang, X., & Wang, Y. (2023). EqMotion: Equivariant Multi-Agent Motion Prediction with Invariant Interaction Reasoning. In 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1410–1420.
[34] Bae, I., Park, Y., & Jeon, H. (2024). SingularTrajectory: Universal Trajectory Predictor Using Diffusion Model. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 17890–17901.
[35] Guo, W., Bie, X., Alameda-Pineda, X., & Moreno-Noguer, F. (2021). Multi-Person Extreme Motion Prediction. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13043–13054.
[36] Xing, C., Mao, W., & Liu, M. (2023). Scene-aware Human Motion Forecasting via Mutual Distance Prediction. ArXiv, abs/2310.00615.
[37] Mao, W., Liu, M., Hartley, R.I., & Salzmann, M. (2022). Contact-aware Human Motion Forecasting. ArXiv, abs/2210.03954.
[38] Zhang, Y., Kephart, J.O., & Ji, Q. (2024). Incorporating Physics Principles for Precise Human Motion Prediction. In 2024 IEEE/CVF Winter Conference on Applications of Computer Vision, 6152–6162.
[39] Ionescu, C., Papava, D., Olaru, V., & Sminchisescu, C. (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), 1325–1339.
[40] CMU Graphics Lab Motion Capture Database. http: //mocap.cs.cmu.edu/.
[41] Sigal, L., Balan, A.O., & Black, M.J. (2024). Dataset: HumanEva-I.
[42] Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., & Black, M.J. (2019). AMASS: archive of motion capture as surface shapes. In 2019 IEEE/CVF International Conference on Computer Vision, 5441–5450.
[43] von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., & Pons-Moll, G. (2018). Recovering accurate 3d human pose in the wild using imus and a moving camera. In Computer Vision - ECCV 2018, 614–631.
[44] Sun, J., & Chowdhary, G. (2023). Towards accurate human motion prediction via iterative refinement. arXiv: 2305.04443.
[45] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., & Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Nash Equilibrium. ArXiv, abs/1706.08500.
[46] Zhao, S., Wang, Z., Luan, T., Jia, J., Zhu, W., Luo, J., Yuan, J., & Xi, N. (2025). PP-Motion: Physical-Perceptual Fidelity Evaluation for Human Motion Generation. In Proceedings of the 33rd ACM International Conference on Multimedia.