MDMP: Multi-modal Diffusion for supervised Motion Predictions with uncertainty

Supplementary Material

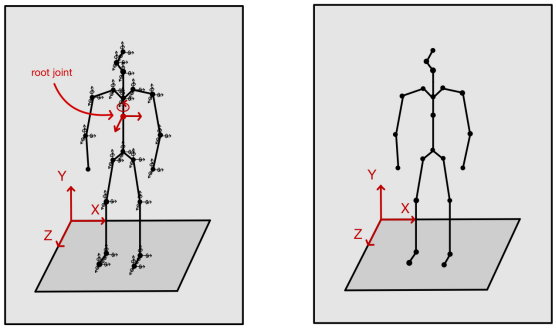

Figure 1. **Human Pose Representation. (Left)** HumanML3D features. **(Right)** 3D joint positions.

A. Data Processing for Evaluation, Visualization in 3D Plots and re-Training

This section details the data processing steps repeatedly performed in our study. Specifically, we discuss the feature transformation process necessary to convert the pose representation of the HumanML3D [3] dataset back into 3D coordinates for result evaluation and visualization in 3D plots. Furthermore, we describe the adaptations made to our model to work with real-time 3D joint position data.

A.1. Feature Transform

In our work, the feature transformation is a crucial step to obtain the 3D joint positions from the pose representation provided by the HumanML3D [3] dataset. This transformation is necessary because the dataset's pose representation includes various redundant features that must be processed to isolate the 3D joint coordinates required for quantitative evaluation of the model using the Mean Per Joint Position Error (MPJPE) and for qualitative evaluation through visualization of the predicted sequences (see Section **??** and supplementary video).

The pose representation in HumanML3D [3] consists of a tuple of features including root angular velocity, root linear velocities, root height, local joints positions, velocities, rotations, and binary foot contact features. Specifically, it provides 263 features per body frame. To extract the 3D joint positions, these features must be transformed because they include information in root space that needs to be converted into global coordinates.

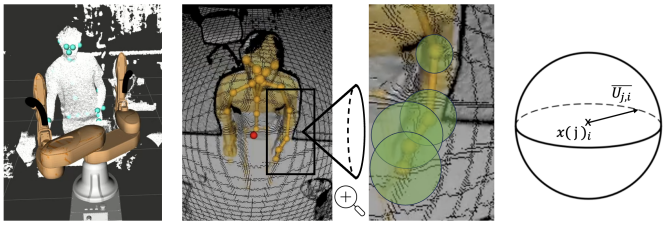

Figure 2. **Human-Robot Collaboration Experiment Setup & Conceptual Zones of Presence Representation** The first image is a representation of our Experiment Setup described in section B, taken in ROS. The second image is the real-time human pose estimation used as input to our model. The third and fourth images are conceptual representations of our predicted zones of presences, using the mean of the uncertainty features $\overline{U_{j,i}}$ as the radius around joint $x(j)_i$. The uncertainty factor "Mode Divergence" described in section **??** has proven to provide the best results in simulation.

The transformation process involves the following steps:

1. **Recover Root Rotation and Position:** The root rotational velocities are extracted and integrated over time to obtain the root rotation angles, which are then converted into quaternions. Simultaneously, the root positions are recovered by integrating the root linear velocities.

2. **Concatenate Rotations and Positions:** The local joint rotations and positions provided in the dataset are combined with the root rotations. The combined rotations are converted from quaternion representation to a continuous 6D rotation format.

3. **Forward Kinematics:** Using forward kinematics, the combined rotations and positions are processed to obtain the global 3D joint positions. This involves computing the global position of each joint based on its local rotation and position relative to the root and applying the root's global transformation.

A.2. Adaptation to 3D Joint Position Data

Due to the nature of our recorded motion capture data using real-time pose estimation (see section B), the pose data we have access to consists only of 3D joint positions. This results in a simplified feature representation of 96 features per skeleton (32 joints $\times$ 3 coordinates), compared to the 263 features per body frame provided by the HumanML3D [3] dataset. To address this discrepancy, we considered two po-

tential solutions:

1. **Transformation to Original Feature Space:** One approach was to transform the 3D joint positions back to the original feature space of 263 features per frame. However, this transformation involves estimating several parameters that are not directly observable from the joint positions alone, which would likely introduce inaccuracies into the data and could negatively impact the model's performance. For instance, approximating the root angular velocity can be complex, and computing local rotations typically requires sophisticated methods like inverse kinematics (IK).

2. **Retraining the Model:** Instead of transforming the data, we opted to retrain the model using the simplified feature representation of 3D joint positions. The only modifications involved adjusting the dimensions of the encoder and decoder. By training directly on the motion sequences represented as 3D joint positions, we avoid the inaccuracies associated with the transformation process and ensure that the model is trained on the most accurate representation of our data. However, the redundant pose representation can be useful for learning spatio-temporal motion patterns, and this lower-dimensional representation might result in a slight loss of information, potentially decreasing the model's performance. Another consideration is that the pose estimation data we access through pose estimation includes 32 joints, whereas the HumanML3D [3] dataset uses only 22 joints to represent the human body. Therefore, we need to filter out the 10 additional joints and predict motion sequences using only the 22 joints per frame.

We chose the second approach and retrained our model on the 3D joint position feature space. This retrained model was then used for the experiments in our lab, allowing us to work directly with the data collected through our real-time pose estimation system.

B. Experiment Setup

We provide details about our experimental setup used in our lab to predict the future motions of a human worker in a Human-Robot Collaborative (HRC) Workspace.

Physical Setup: Our collaborative workspace, as shown in Fig. 2, consists of a duAro1 Kawasaki robot and multiple desks where both the human and the robot can place and pass objects. The workspace is monitored by multiple Azure Kinect RGB-D cameras. All sensor data, along with the command signals for controlling the robot, are centralized in a ROS (Robot Operating System) setup. April Tags are used for calibration to ensure all spatial data (including the real-time position of the robot and data from the sensors) is aligned within the same reference frame.

Motion Tracking: Human Pose Estimation is performed in real-time to gather skeleton data and track the human worker within the HRC workspace using the Azure Kinect Body Tracking SDK. The skeleton data is transmitted to the ROS system via the Azure Kinect ROS driver, transformed into the robot's base frame, and then recorded in MarkerArray topics. These motion sequences are subsequently fed into our trained model.

Textual Actions: To leverage the benefits of our multimodal approach, a set of predefined Human-Robot collaborative actions is mapped to specific keys on a keyboard. This keyboard can be operated by either the human worker or an external observer who can also provide detailed textual descriptions of the actions being performed. If the model is not given textual information between actions, it relies solely on motion sequence data. As demonstrated in our Motion & Text ablation study in section C, our model can perform short-term predictions without contextual information. However, as described in the limitation section **??**, we are aware that this reliance on textual descriptions is not ideal for real-time Human-Robot Collaboration, as it can be burdensome and not every action is scripted in advance.

C. Additional Experimental Results

We present qualitative results by comparing our model to state-of-the-art Text2Motion methods such as MotionGPT [5] and MDM [7]. The comparisons are showcased in the video appendix, where our model's predictions are visualized against ground truth motion sequences. In every sequence, our model outperforms the baselines in terms of proximity to the ground truth.

Additionally, we visualize our predictions using meshes created with SMPL [6] and rendered in Blender. These visualizations transform the skeleton motions into human-like meshes performing the same actions, providing a clearer and more intuitive understanding of the predicted motions.

C.1. Video Representation

In the video appendix, we compare our model's predictions to those of MotionGPT [5] and MDM [7] on certain specific actions. The videos allow for a visual assessment of the proximity of the predicted sequences to the ground truth motion. Our model shows better performance in maintaining proximity to the ground truth. Even when the predictions differ from the ground truth, our model's predicted actions align with the intended textual actions, while the baseline models tend to diverge. Even when our predicted sequences differ from the ground truth, the actions correspond accurately to the textual descriptions, whereas the other baselines quickly diverge from the ground truth.

C.2. Path Trajectory Following

In Fig. 3, we also evaluate the trajectory following capability of our model through stop-motion images. These images track the motion sequences with faded colors for early frames and progressively darker colors for later frames. Additionally, a trajectory path projecting the root joint position

Model	NPSS			MPJPE (mm)				
	0-1s	1-2s	2-4s	1s	2s	3s	4s	5s
MDM (re-trained)	0.059	0.064	0.172	205.5	385.5	551.3	692.0	791.9
MDMP (Ours)	**0.034**	**0.043**	**0.132**	**186.7**	**341.8**	**474.8**	**592.5**	**669.8**

Table 1. Comparison of NPSS & MPJPE (mm) on HumanML3D.

Method	**R-Precision ↑**			**FID ↓**	**Diversity →**
	Top-1	**Top-2**	**Top-3**		
Real Motion	0.511 ±0.003	0.703 ±0.003	0.797 ±0.002	0.002 ±0.000	9.503 ±0.065
T2M [2]	0.457 ±0.002	0.639 ±0.003	0.740 ±0.003	1.067 ±0.002	9.188 ±0.002
MDM [7]	0.320 ±0.005	0.498 ±0.004	0.611 ±0.007	0.544 ±0.044	9.559 ±0.086
MotionGPT [5]	0.492 ±0.003	0.681 ±0.003	0.778 ±0.002	0.232 ±0.008	**9.528 ±0.071**
MoMask [4]	**0.521 ±0.002**	**0.713 ±0.002**	**0.807 ±0.002**	**0.045 ±0.002**	—
Our Method	0.445 ±0.002	0.692 ±0.006	0.775 ±0.005	0.437 ±0.698	8.335 ±0.025

Table 2. Comparison of our method with state-of-the-art text-to-motion models on the HumanML3D dataset. Metrics reported are R-Precision (higher is better), FID (lower is better), and Diversity (closer to Real Motion is better).

on the XZ-plane is included to precisely follow the predicted trajectory. This study further confirms our model's ability to accurately predict motion sequences that follow precise trajectories over long-term durations.

C.3. Additional Uncertainty Qualitative Comparison

In Figs. 4, 5, 6, and 7, we present additional results of our Uncertainty Parameters for visual comparison and evaluation. As shown in these figures, the "Mode Divergence" index is the only one that exhibits a notable increase over time, correlating closely with the error, particularly when the divergence between the prediction and ground truth becomes pronounced (see Figs. 4 and 7). In contrast, the "Predicted Variance" shows less temporal variation, while the "Mean Fluctuations" appear somewhat more unstable. These findings align with our previous analysis using the Sparsification Plot in Fig. **??**.

C.4. Additional Accuracy Quantitative Comparison

Table 1 presents additional comparative results between our method and the retrained MDM [7], evaluated using NPSS [1] and MPJPE metrics. These results further validate our contributions, highlighting improved accuracy, particularly in longer-term predictions.

To fairly benchmark our method against Text2Motion baselines, despite our distinct motion-conditioned framework, we evaluated our model using generic metrics (*R-Precision*, *Diversity*, and *FID*) and present the results in Table 2.

We argue that our lower *R-Precision* results compared to the latest benchmarks (MoMask [4] and MotionGPT [5]) reflects our model's prioritization of initial motion conditioning over textual alignment, confirming insights from our first ablation study (section **??**). Similarly, reduced *Diversity* arises naturally from constraining motions to coherent continuations of initial segments, which is advantageous in Human-Robot Collaboration settings where accuracy and confidence are prioritized over variability.

Finally, the *Frechet Inception Distance* (*FID*) metric is intended to evaluate the overall quality of generated motions by measuring the distributional difference between high-level features of the generated motions and those of real motions. We argue that as described by Guo et al. [2], the pretrained motion feature extractor used for computing *FID* was trained using a contrastive loss to produce geometrically close feature vectors for matched text-motion pairs. Hence, the motion encoder is specifically optimized for motions conditioned solely on text descriptions and may not accurately capture the features of motions generated by models conditioned on both text and an initial motion segment (motion prequel).

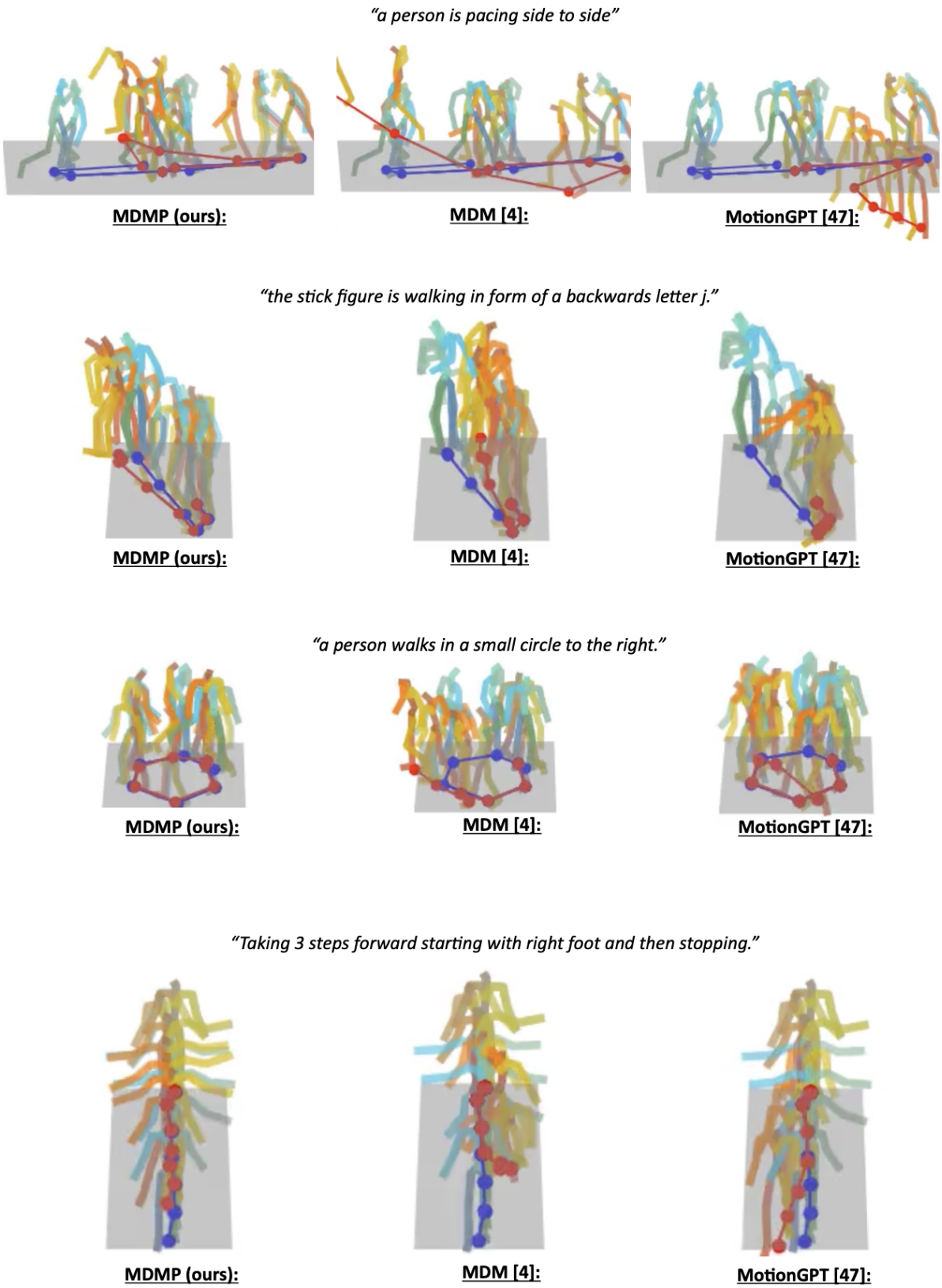

Figure 3. **Qualitative Comparisons on Path Following.** Ground-truth in **red**; Predictions in **blue**

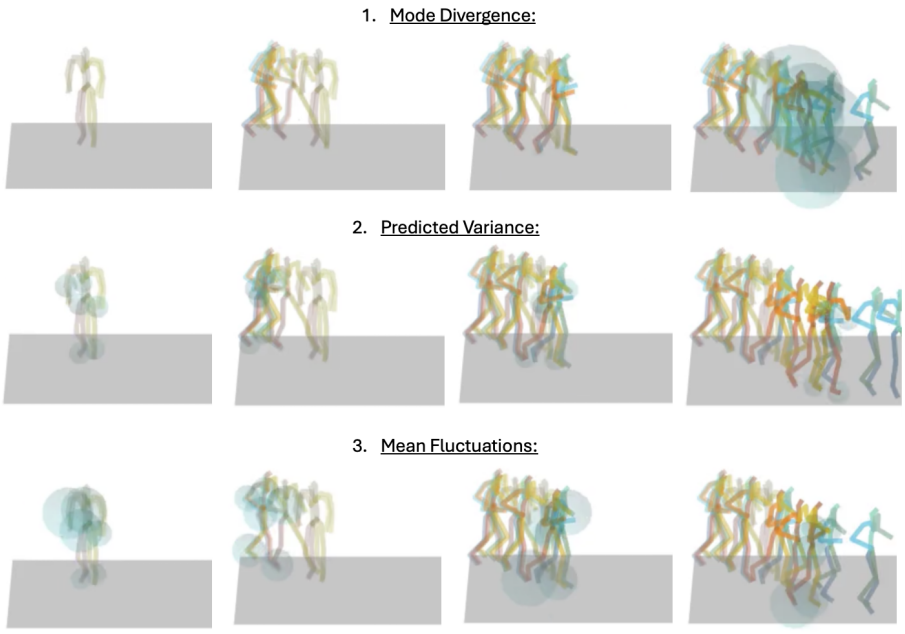

Figure 4. **Qualitative Comparisons on Uncertainty Parameters.** Textual Prompt: ”a person is jogging back and forth from where he has standing”; Ground-truth in **red**; Predictions in **blue**

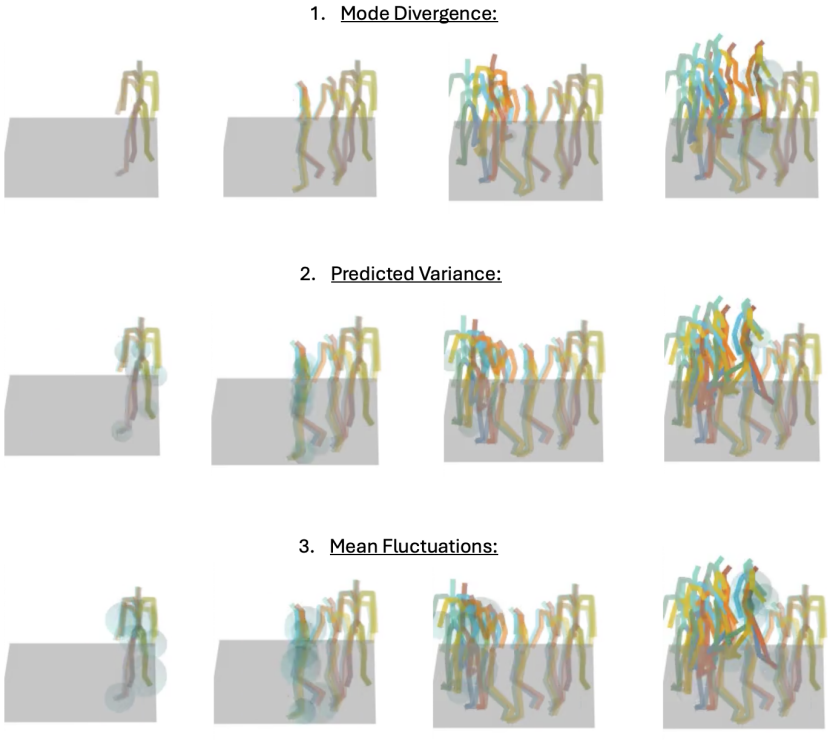

Figure 5. **Qualitative Comparisons on Uncertainty Parameters.** Textual Prompt: ”a person walks in a circle, clockwise.”; Ground-truth in **red**; Predictions in **blue**

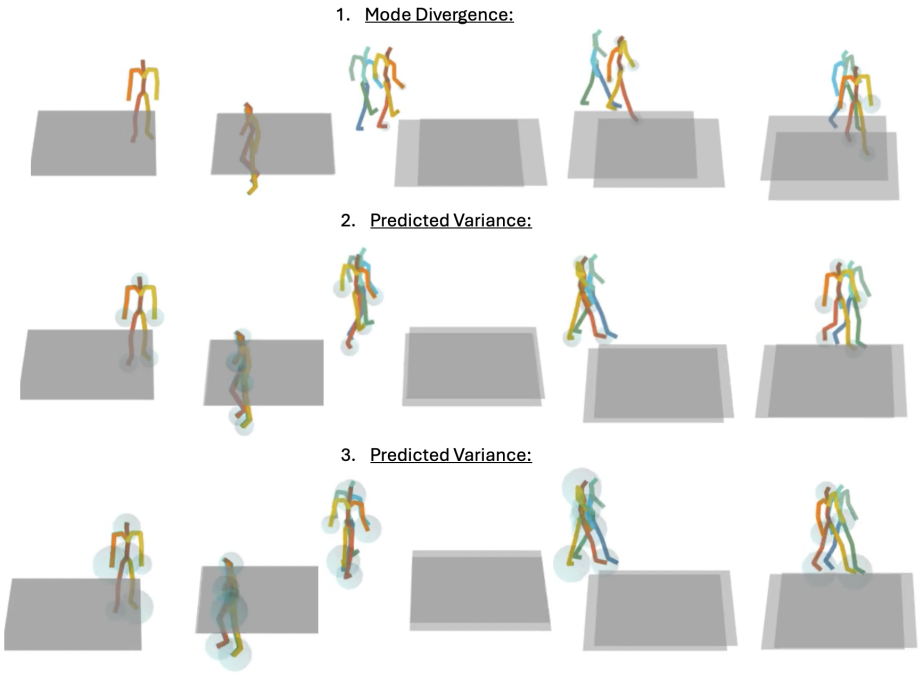

Figure 6. **Qualitative Comparisons on Uncertainty Parameters.** Textual Prompt: ”a person walks around in a circle.”; Ground-truth in **red**; Predictions in **blue**

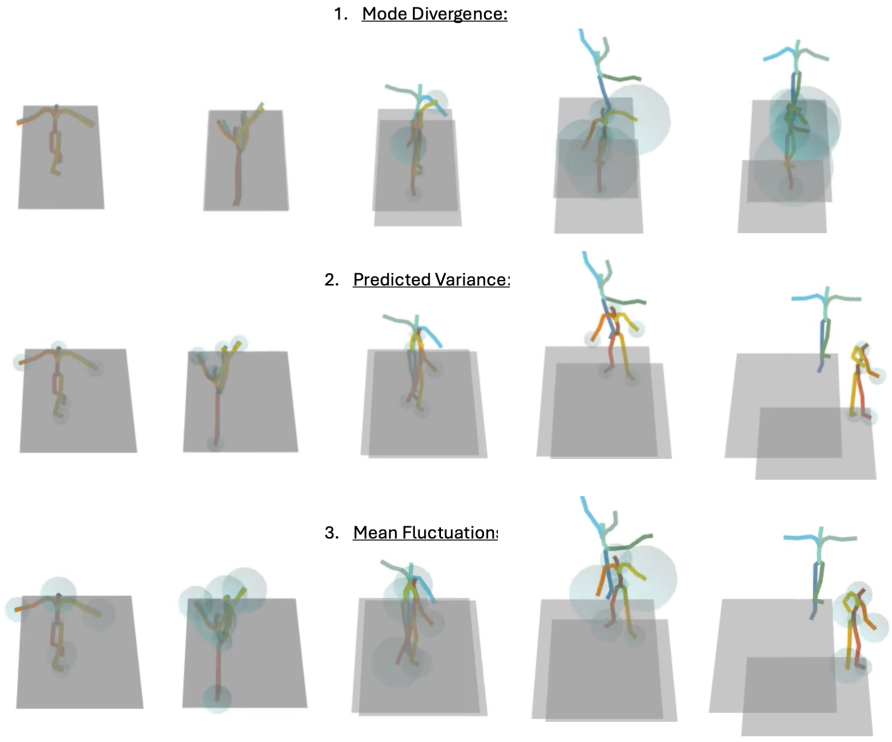

Figure 7. **Qualitative Comparisons on Uncertainty Parameters.** Textual Prompt: ”a person is doing a acrobatic dance.”; Ground-truth in **red**; Predictions in **blue**

References

[1] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, C. Lee Giles, and Alexander G. Ororbia. A neural temporal model for human motion prediction, 2019. 3

[2] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *ACM International Conference on Multimedia*, pages 2021–2029, 2020. 3

[3] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, 2022. 1, 2

[4] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[5] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems (NeurIPS)*, 37, 2023. 2, 3

[6] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. *SMPL: A Skinned Multi-Person Linear Model*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. 2

[7] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 2, 3