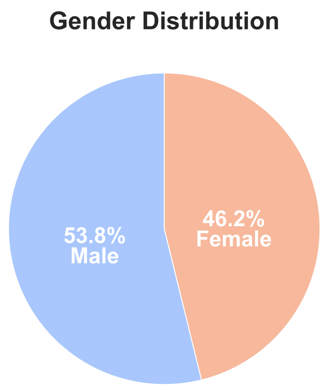
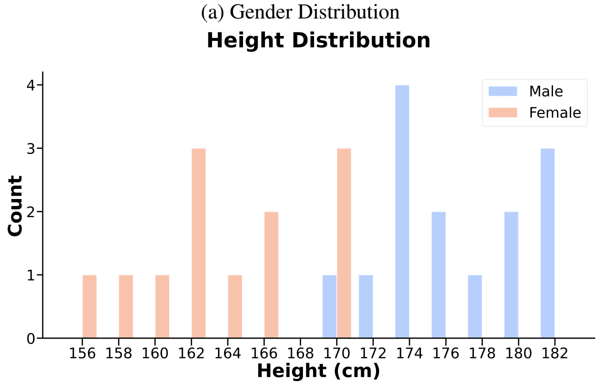
Pressure2Motion: Hierarchical Human Motion Reconstruction from Ground Pressure with Text Guidance

Supplementary Material

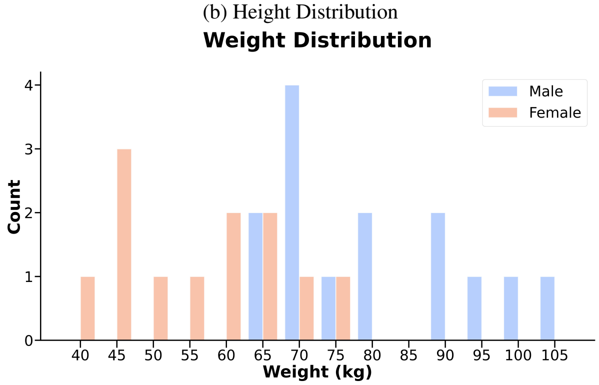
1. MPL Dataset Details



(a) Gender Distribution



(b) Height Distribution



(c) Weight Distribution

Figure 1. Distribution of gender, height, and weight among the 25 subjects in the *MPL* dataset.

The *MPL* dataset is developed to facilitate research in reconstructing full-body human motion from the highly sparse inputs of ground pressure and text prompts, building on top of the MotionPRO dataset [10], which contains a large-scale collection of human motion sequences captured using plantar pressure sensors. For our research, we extend this dataset by incorporating textual descriptions.

The raw motion sequences in MotionPro are generally around 10 minutes long and are not segmented based on action semantics. Additionally, actions within each long sequence are often repeated 2-3 times. To address this, we manually segmented the sequences based on clear semantic action boundaries, resulting in 20,944 motion sequences, amounting to approximately 2.3 million frames. Each sequence is temporally resampled to 20 FPS for consistency, lasting $2 \sim 8$ seconds, reflecting diverse temporal dynamics across action types.

1.1. Data Distribution

Our *MPL* dataset comprises motion sequences from 25 subjects with diverse physical characteristics, including a balanced distribution of gender, a wide range of heights and weights, and varying body types. We visualize the distribution of the following attributes:

- **Gender:** The dataset includes 12 females and 13 males, with a roughly balanced ratio.
- **Height:** The subjects range from 157 cm to 184 cm in height, with an average of 172.1 cm, covering both shorter and taller individuals.
- **Weight:** The weight distribution spans from 44.05 kg to 108.00 kg, with an average of 56.67 kg, ensuring the inclusion of both lightweight and heavier subjects.

Figure 1 illustrates the distribution for gender, height, and weight. This coverage enhances the robustness and applicability of our motion reconstruction model across real-world variations in body structure.

1.2. Data Processing

We reformatted the SMPL[6] parameters from the MotionPro dataset into a more comprehensive motion representation following the HumanML3D [2] convention. Each motion sequence of length N is transformed into a representation of shape $(N, 263)$ where each frame encodes the pelvis velocity, local joint positions, joint velocities, joint rotations (in pelvis space), and binary foot contact indicators.

During data processing, we intentionally exclude global operations such as "uniforming skeleton" , "put on floor," and "rotate to face $Z+$", which are part of the default HumanML3D preprocessing pipeline. Uniform skeleton retargeting enforces consistent bone lengths across subjects.

While these operations help standardize motion, they may distort the global positions of joints relative to the pressure data. To maintain spatial consistency, we retain the original global coordinates of both motion and pressure, ensuring accurate alignment between the two modalities during synthesis.

One notable characteristic of our processing pipeline is that the first frame of each motion sequence, i.e. the root joint of the initial pose is aligned to the origin of the XZ plane. However, this causes a spatial offset between the motion and the corresponding pressure map in the XZ plane. To account for this, we design a mechanism in pressure feature extraction module to predict and correct this offset using pressure information, ensuring precise spatial alignment for downstream tasks.

1.3. Caption Process

Text descriptions are an integral part of the dataset, providing semantic guidance for the motion synthesis process. To ensure diversity and semantic richness in textual prompts, descriptions are generated using Qwen2.5-VL [12], a vision-language model capable of processing long video sequences. Specifically, given a motion clip and a brief action keyword from the original MotionPRO dataset, we provide the RGB video frames and keyword as input to Qwen2.5-VL. The model interprets the human activity within the video context and generates five diverse captions at varying levels of detail. These descriptions range from simple high-level actions (e.g., "The person is walking") to more intricate and detailed descriptions (e.g., "The person is walking with a slight leftward tilt and right arm movement").

1.4. Augmentation

Specifically, given a pressure sequence, we apply spatial augmentations (translations and rotations) to simulate real-world variations in global orientation and position. We adjust the global offset of the motion accordingly to maintain spatial alignment with the augmented pressure data. This augmentation helps to simulate real-world variations in body posture and pressure signals.

2. Implementation Details

2.1. Training Setup

Our models are implemented in PyTorch and trained on 8 NVIDIA A800 GPUs for a total of 100,000 iterations. We adopt the AdamW optimizer [7] with a learning rate of 1×10^{-5}. The ControlNet is initialized with pretrained weights from MDM [11]. During training, the parameters of the Movement Trajectory extraction module $\mathcal{F}_{\text{traj}}$ and the pretrained MDM backbone $\mathcal{F}_\theta$ are frozen to retain their original representations.

2.2. Network and Feature Dimensions

We follow prior works and use CLIP [9] to encode text prompts into 512-dimensional embeddings. The output features of both the ControlNet and the Adapter modules are also of size 512 to ensure compatibility with the pretrained MDM architecture.

The Pressure-Inferred Movement Trajectory $\mathbf{T}_{\text{traj}}$ and Pressure-Induced Posture Shifts $\mathbf{S}_{\text{shift}}$ are extracted with output dimensions of $(B, L, 39)$ and $(B, L, 256)$ respectively, where B is the batch size and $L = 196$ is the sequence length. The 39-dimensional trajectory representation includes the global 3D positions (XYZ) of the root, left/right ankles, and left/right toes (total 5 joints × 3 = 15), as well as 6D rotation representations for the left/right ankles and toes (4 joints $\times$ 6 = 24).

2.3. Diffusion and Loss Hyperparameters

To improve robustness to text variations, we randomly drop 10% of the text conditions during training. This enables the use of Classifier-Free Guidance (CFG) [4] during inference, where we apply a CFG scale of 5. We adopt a standard DDPM [5] framework with $T = 1000$ denoising steps. The control strength τ for injecting pressure signals is defined as $\tau = \frac{20\hat{\Sigma}_t}{L}$, where $\hat{\Sigma}_t = \min(\Sigma_t, 0.01)$. We set $\lambda_{\text{diff}} = 1$ and $\lambda_{\text{cons}} = 5$ throughout all experiments.

2.4. Baseline Adaptations

For MDM[11] and MotionDiffuse[14], we concatenate the global and local pressure embeddings and append them to the noisy motion input at each denoising step. This allows these models to incorporate pressure signals at each frame, providing a consistent pressure-aware motion reconstruction. For OmniControl[13] and MaskControl[8], we replace the original spatial control inputs with the combined pressure embeddings, enabling these models to condition on pressure in a comparable manner to our approach.

2.5. Pressure Feature Extractor Details

The Pressure-Inferred Trajectory $\mathcal{F}_{\text{traj}}$ is essential for capturing the overall movement path and body alignment. To extract this information, we utilize a feature extraction module following the architecture from MotionPro [10], which includes a ResNet-based [3] pressure encoder, a temporal information processor, and a fully connected projection layer. Given the sparsity of pressure maps, where valid values are limited and primarily found under the feet during standing, the pressure encoder utilizes a compact ResNet architecture with small convolutional kernels to focus on the localized pressure regions, despite the large size of the pressure map. Temporal dynamics are captured using the temporal information processor, which combines a GRU [1] to model long-term dependencies with a self-attention mechanism to capture short-term correlations in the pressure sequence.

3. Evaluation Details

We adopt a text feature extractor and a motion feature extractor from HumanML3D and retrain it on our *MPL* dataset to adapt to the new data distribution. The resulting model is used to evaluate all the methods.

We evaluate motion quality, motion-pressure consistency, and semantic alignment of the reconstructed motions using the following metrics:

- **Center of Pressure (CoP) Error ↓:** This metric directly measures the physical consistency between the input pressure signal and the reconstructed motion. It is calculated as the mean L2 distance between two CoP time-series:
  Pressure CoP ($CoP_{Pressure}$): Calculated from the input pressure map $P_n \in \mathbb{R}^{H\times W}$ at frame n. The pixel-space CoP (geometric center) is computed as a weighted average:

$$CoP_{P,x}^{(n)} = \frac{\sum_{i,j} P_n(i,j)\cdot j}{\sum_{i,j} P_n(i,j)},$$

$$CoP_{P,z}^{(n)} = \frac{\sum_{i,j} P_n(i,j)\cdot i}{\sum_{i,j} P_n(i,j)}.$$

  This pixel-space CoP is then transformed into motion-space using a pre-calibrated scale S and offset O:

$$CoP_{Pressure}^{(n)} = [CoP_{P,x}^{(n)}, 0, CoP_{P,z}^{(n)}] \odot S + O.$$

  Motion CoP (CoP_{Motion}): Inferred from the reconstructed motion's lower-body joints. We use a softmax-weighted average of the K key foot joints' (e.g., ankles, toes) ground projections $j_k = [j_{k,x}, j_{k,y}, j_{k,z}]$, where the weight w_k is inversely related to the joint's height $j_{k,y}$:

$$w_k^{(n)} = \frac{\exp(-j_{k,y}^{(n)}/\tau)}{\sum_{l=1}^{K}\exp(-j_{l,y}^{(n)}/\tau)},$$

$$CoP_{Motion}^{(n)} = \sum_{k=1}^{K} w_k^{(n)} \cdot [j_{k,x}^{(n)}, 0, j_{k,z}^{(n)}].$$

  CoP Error ($\mathcal{L}_{CoP}$): The final error is the mean Euclidean distance over all N frames and B batch items:

$$\mathcal{L}_{CoP} = \frac{1}{B\cdot N}\sum_{b=1}^{B}\sum_{n=1}^{N}\left\|CoP_{Pressure}^{(b,n)} - CoP_{Motion}^{(b,n)}\right\|_2.$$

  A lower value indicates superior motion-pressure consistency.
- **Fréchet Inception Distance (FID) ↓:** FID measures the distributional distance between reconstructed motions and ground-truth motions in the feature space. In our setting, motions are encoded via a pre-trained motion encoder, and FID is computed on the extracted features. Lower FID indicates that the reconstructed motions are more realistic and distributionally similar to real data.
- **Foot Skating ↓:** This metric computes the ratio of frames in which a foot joint is supposed to be in contact with the ground but exhibits non-negligible motion, indicating physically implausible sliding. Specifically, for each frame, we check whether a foot is labeled as "in contact" and simultaneously has a velocity exceeding a small threshold. The ratio of such inconsistencies over all frames is reported. A lower value indicates better foot-ground contact realism and physical plausibility.
- **Mean Per Joint Position Error (MPJPE) ↓:** MPJPE measures the average Euclidean distance between corresponding joints in the predicted and ground-truth motions:

$$\text{MPJPE} = \frac{1}{N\cdot T}\sum_{t=1}^{T}\sum_{j=1}^{N}\|\hat{\mathbf{p}}_{t,j} - \mathbf{p}_{t,j}\|_2$$

  where T is the number of frames, N is the number of joints, and $\hat{\mathbf{p}}_{t,j}$ and $\mathbf{p}_{t,j}$ denote the predicted and ground-truth positions of joint j at time t. This metric evaluates the spatial alignment between reconstructed and real motions.
- **Lower-body MPJPE (L-MPJPE) ↓:** A variant of MPJPE that only considers lower-body joints (e.g., hips, knees, ankles, feet), which are most relevant to pressure-ground interactions. It reflects the model's ability to reconstruct physically grounded lower-body motion. Lower is better.
- **Trajectory Error (> 50cm) ↓:** This metric measures the ratio of motion sequences in which trajectory frames deviate from the ground truth by more than 50 cm. It reflects whether global body movement is consistently aligned with the physical signal.
- **R-Precision (Top-3) ↑:** R-Precision measures the semantic consistency between reconstructed motions and their associated text prompts. We use a joint motion-text encoder to compute the similarity between reconstructed motion features and ground-truth text embeddings. R-Precision@3 reflects whether the correct caption is ranked among the top-3 retrieved results for a reconstructed motion. Higher values indicate better semantic alignment.

4. Additional Ablation Study

In addition to the ablation study on pressure features, we further investigate the impact of model architecture by removing key components, namely the ControlNet and Adapter. Specifically, we modify the architecture by concatenating the two pressure features—Movement Trajectories and Posture Shifts—directly and feeding them into either the ControlNet or Adapter without using the hierarchical structure.

The results of this experiment are shown in Table 1, where we compare the full model with versions that exclude ControlNet and Adapter. For the version without ControlNet, we observe a significant increase in FID (1.3683) and MPJPE (0.1951), indicating that removing the ControlNet impairs the model's ability to properly align the motion with the pressure signals, resulting in less accurate motion reconstruction. Similarly, removing the Adapter leads to a noticeable degradation in performance, with FID increasing to 0.695 and MPJPE rising to 0.2092. These results demonstrate the critical role of both components in ensuring the high fidelity and physical plausibility of the reconstructed motions. Moreover, concatenating the two pressure features (Movement Trajectories and Posture Shifts) directly and feeding them into either the ControlNet or Adapter results in inferior performance compared to the full model. This suggests that the hierarchical structure, where ControlNet handles the overall movement trajectory and the Adapter fine-tunes the posture shifts, is essential for reconstructing realistic and semantically aligned motions.

Figure 2 provides visual comparisons of the motion sequences reconstructed by the different model variations. These visualizations further confirm that the full model consistently produces motions that are more physically plausible and aligned with the pressure data, especially in areas such as foot-ground contact.

Table 1. Additional Ablation study of Model Archtecture.

Method	FID↓	FS↓	Cop Err↓	L-MPJPE↓	MPJPE↓
w/o ctrlnet	1.3683	0.0621	0.6120	0.1694	0.1951
w/o Adapter	0.695	0.0634	0.6655	0.1702	0.2092
Full	**0.262**	**0.0553**	**0.4260**	**0.1273**	**0.1622**

5. More Visualization Results

We present more motion reconstruction results to further demonstrate the effectiveness of our method in reconstructing human motion from sparse pressure data. Figure 3 showcases a variety of reconstructed motions across different scenarios, highlighting the robustness of our approach in the pressure-to-motion task. The first few cases demonstrate typical human actions, such as walking, standing, and some daily activities. Our model reconstructs high-fidelity motions that align well with the pressure data and text prompts, maintaining both physical consistency and semantic plausibility.

Particularly interesting are the last two cases in Figure 3. The second-to-last case corresponds to a jumping motion, where **no pressure** is applied during the jump. Despite the absence of pressure in the air, our model successfully reconstructs a realistic jumping motion, demonstrating its ability to handle scenarios with **no foot-ground contact.**

The final case in Figure 3 shows a dynamic plank position, which involves complex pressure distributions from both the hands and feet. Our method effectively handles this rare and specialized pressure contact scenario, reconstructing a physically plausible motion that corresponds to the simultaneous pressure from all four limbs. This demonstrates the versatility of our model in **handling uncommon or intricate pressure patterns.**

6. Limitations

Despite the promising results, our approach is subject to several limitations.

First, the diversity of motion types in the dataset remains relatively limited. While the *MPL* dataset includes various basic motions such as walking, standing, and sitting, more complex activities—such as interactions on inclined surfaces or with dynamic real-world environments—are not yet covered. Extending the dataset to include more diverse and complex motion types, as well as scenarios involving pressure data from inclined or interacting surfaces (e.g., walking on stairs or engaging with objects), would significantly enhance the robustness and applicability of the model in real-world use cases.

Another major limitation of our model lies in its computational complexity. While our approach demonstrates high fidelity in reconstructing pressure-aware motions, the underlying architecture—specifically the pressure feature extraction module and the hierarchical pressure-modulated motion reconstruction framework—is relatively large and computationally demanding. Additionally, the motion reconstruction process, based on the diffusion model, involves multiple denoising steps, making the inference process slower. On a single NVIDIA A800 GPU, reconstructing one motion sequence takes approximately 180 seconds. This extended inference time can become a bottleneck when deploying the model in real-time applications. As a future direction, we plan to investigate strategies, such as more efficient pressure feature extraction techniques, and explore inference optimization to speed up the motion reconstruction process while maintaining high-quality results.

A third limitation concerns the nature of our textual guidance. The *MPL* dataset's text descriptions are entirely generated by a VLM. While this provides consistent and detailed annotations at scale, VLM-generated text tends to be homogeneous and overly descriptive, adhering to a specific stylistic pattern (e.Sg., "The person raises their left arm"). This clean, literal data distribution does not reflect the full diversity and ambiguity of real-world human language. Human prompts are often more abstract, colloquial, underspecified, or describe high-level goals rather than explicit kinematics (e.g., ”Look for something on the floor” vs. ”Bend over and turn head”). Consequently, our model may be overfitted to this VLM-specific text style and less robust to ”in-the-wild” human-authored prompts.

Future work will focus on addressing these limitations by: 1) Enhancing robustness and expanding the dataset to include more dynamic activities (e.g., inclined surfaces); 2) Investigating inference optimization strategies to enable real-time applications; 3) Improving robustness to diverse, human-authored text; and 4) Leveraging pressure data in simulated reinforcement learning or personalized motion modeling.

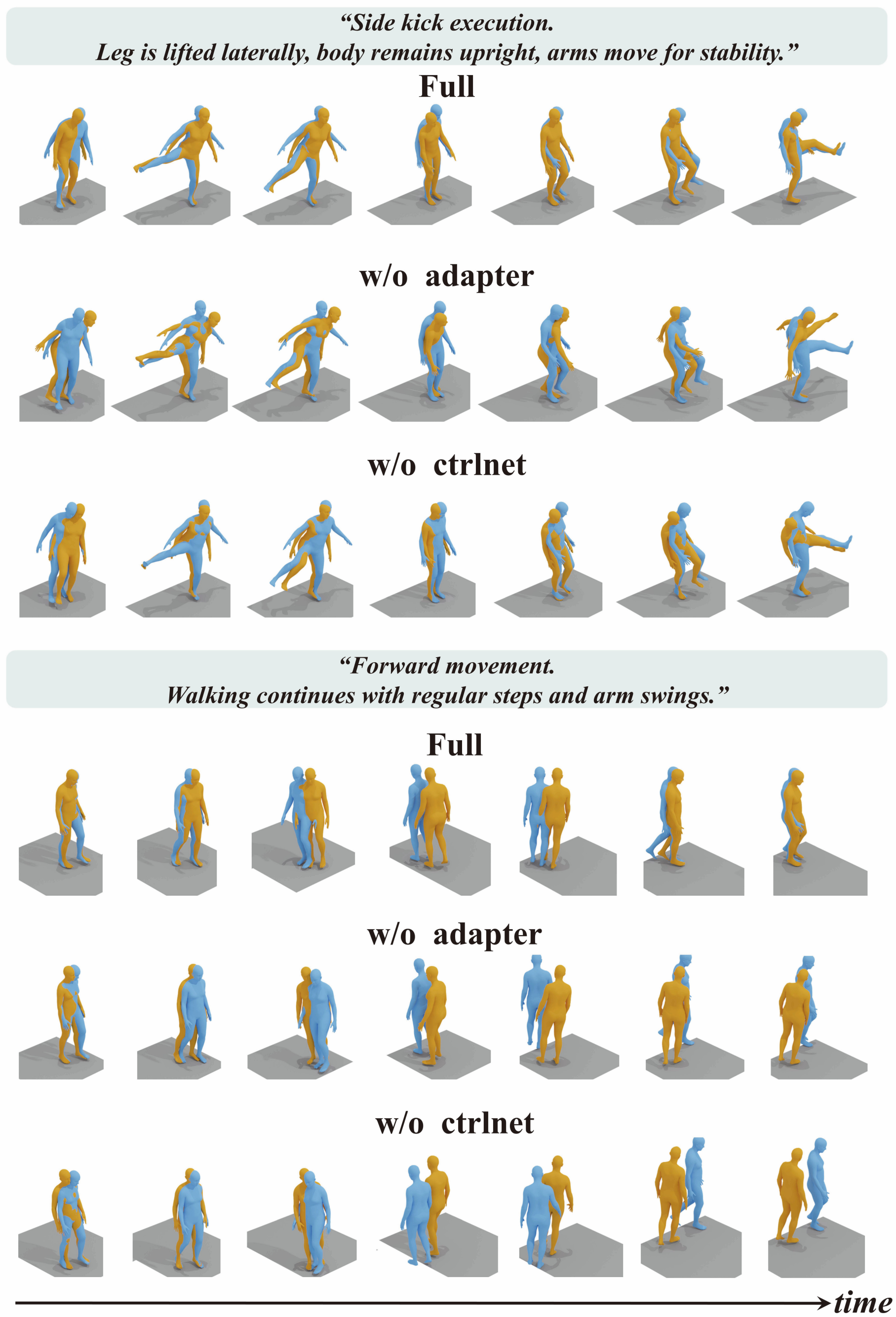


Figure 2. Additional ablation results on the MPL dataset. Yellow denotes the predicted results of different methods; blue represents the ground-truth motions.

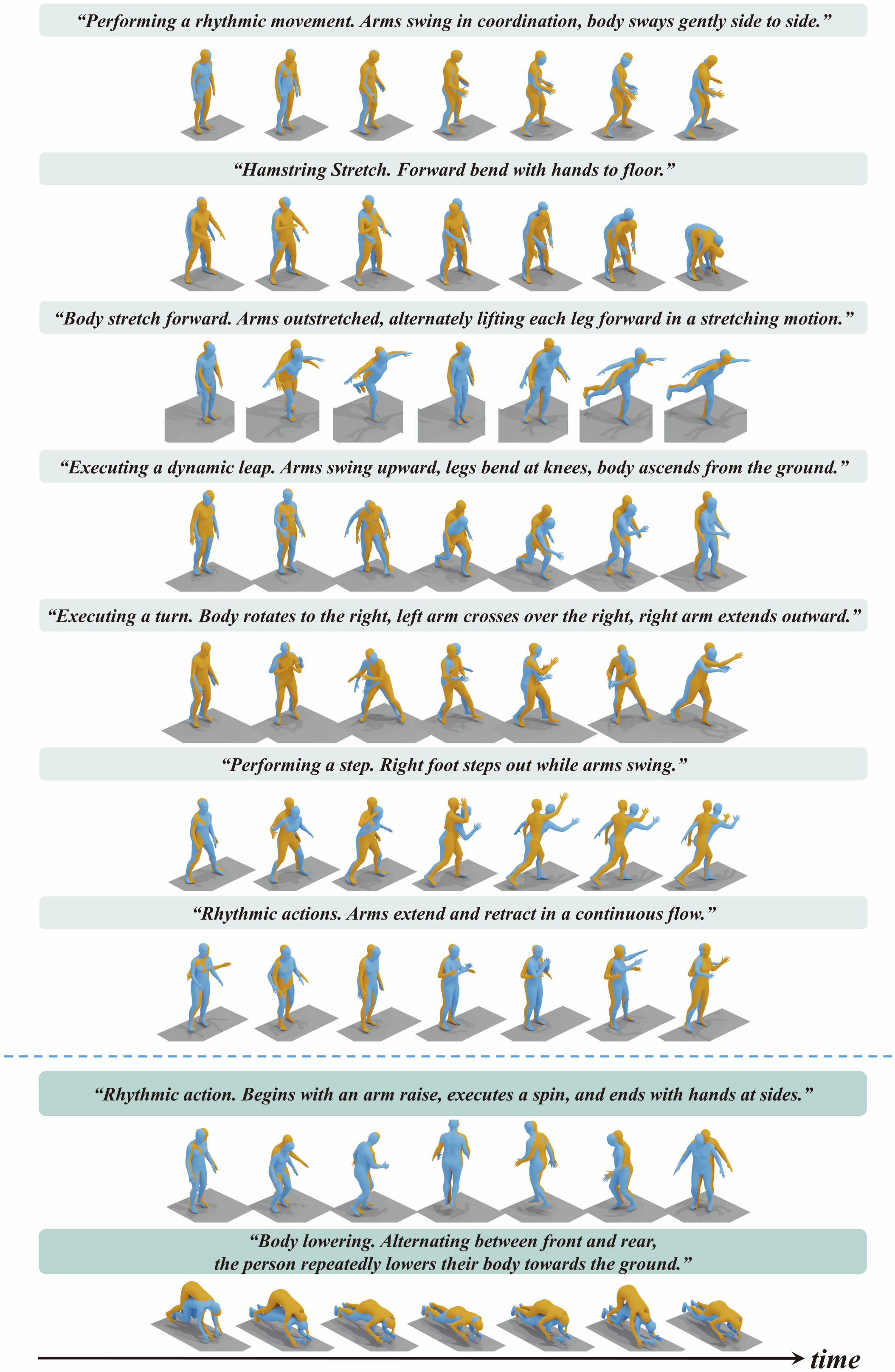


Figure 3. More visualization results on the MPL dataset. Yellow denotes the predicted results of different methods; blue represents the ground-truth motions.

References

[1] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 2

[2] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[6] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, 2015. 1

[7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[8] Ekkasit Pinyoanuntapong, Muhammad Saleem, Korrawe Karunratanakul, Pu Wang, Hongfei Xue, Chen Chen, Chuan Guo, Junli Cao, Jian Ren, and Sergey Tulyakov. Maskcontrol: Spatio-temporal control for masked motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9955–9965, 2025. 2

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[10] Shenghao Ren, Yi Lu, Jiayi Huang, Jiayi Zhao, He Zhang, Tao Yu, Qiu Shen, and Xun Cao. Motionpro: Exploring the role of pressure in human mocap and beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27760–27770, 2025. 1, 2

[11] 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*, 2023. 2

[12] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2

[13] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[14] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024. 2