Supplementary Material of
MotionCtrl: A Real-time Controllable Vision-Language-Motion Model

In this appendix, we provide additional details on our experimental setups in Section 1, including the datasets used for each benchmark and our adopted motion feature, HuMo263. Next, in Section 2, we elaborate on the design and implementation of our proposed part-level residual quantization (PRQ). Finally, in Section 3, we present further details and analysis of the HuMo100M dataset.

1. Experimental Setup

This section describes the datasets and motion features used in the experimental setup for evaluating the MotionCtrl model.

1.1. Datasets

To comprehensively evaluate MotionCtrl's performance on various motion-related tasks, we employ the following datasets:

- **Text-to-Motion:** We use the HumanML3D [5] and KIT-ML [13] datasets. HumanML3D includes 14,616 motion sequences with 44,970 corresponding text descriptions, and KIT-ML offers 3,911 motion sequences with 6,278 text descriptions. Both datasets are split into training (80%), validation (5%), and test sets (15%). In addition, we sample from HuMo100M to construct a new test set, HuMo-t2m, containing 200K samples.
- **Instruct-to-Motion: HuMo-I2M.** To construct this benchmark, we sample 1 million high-quality motion-instruction pairs from the HuMo100M dataset and divide them into training (80%), validation (5%), and test (15%) splits.
- **Instruct-to-Unseen: HuMo-unseen.** To evaluate the model's generalization ability on unseen motion data, we construct the HuMo-unseen dataset, which contains 200,000 new motion sequences sampled from HuMo100M that do not appear in any training data.
- **Instruct-to-PartMotion: HuMo-I2PM.** We use the same training data as HuMo-I2M for this benchmark. For evaluation, we carefully curate 200,000 instances with detailed part-level actions, allocating 150,000 for testing and 50,000 for validation.
- **Instruct-to-LongMotion: HuMo-I2LM.** As demonstrated in the main paper, we concatenate individual motion sequences to create longer sequences. For this task, we sample 500,000 such sequences to formulate the benchmark, dividing them into training, validation, and test splits using the same ratios as HuMo-I2M.
- **Motion Prediction and In-between:** We use the AMASS [11] and 3DPW [17] datasets. These datasets provide annotations based on the SMPL-X [12] or SMPL [10] formats. We extract the corresponding 3D keypoints using the provided beta and theta parameters. Following Zhang et al. [21], we focus on the motion prediction of six body parts (spine, left arm, right arm, left leg, right leg, and head), as the motion prediction task in the 3DPW dataset does not consider global translation. Facial expressions are not included in our analysis. Furthermore, we construct HuMo-MP on the HuMo100M dataset for evaluating the motion prediction task.
- **Action-to-Motion:** We evaluate the model's performance on this task on the UESTC [7] dataset, consistent with Zhang et al. [21].
- **Motion Reconstruction:** We evaluate motion reconstruction using HumanML3D [5], MotionX [8], and HuMo100M.
- **Motion-to-Text:** We evaluate on HumanML3D [5].

1.2. Motion Feature HuMo263

HM3D263-Format [5] is a widely adopted feature representation method in recent motion generation works. It includes relative joint positions, velocities, 6D rotations of key joints, and foot contact information. However, a key issue with HM3D-Format is that its rotation information is computed from joint position data via inverse kinematics (IK). This indirect computation not only loses original rotational details but also introduces significant computational complexity and latency due to the IK solving process, which is detrimental to real-time applications.

To address these issues, we propose and adopt the HuMo263 motion features. HuMo263 is based on the SMPL model [10], directly using SMPL parameters to represent human pose. Specifically, HuMo263 includes relative 6D rotations of key joints (126 dimensions), root node parameters (4 dimensions, including 1 dimension for angular velocity of rotation, 2 dimensions for xz-velocity, and

1 dimension for y-height), redundant joint position information derived from the SMPL model's forward kinematics (63 dimensions), and foot contact information (4 dimensions). Unlike HM3D-Format, HuMo263 directly preserves the rotation information output by the SMPL model, avoiding information loss and computational delay caused by IK calculations.

2. Part-level Residual Quantization

In PRQ, we divide the joints of the whole body into 5 parts, including:

- Left Hand: spine$_1$, spine$_2$, spine$_3$, left collar, left shoulder, left elbow, left wrist
- Right Hand: spine$_1$, spine$_2$, spine$_3$, right collar, right shoulder, right elbow, right wrist
- Left Leg: spine$_1$, spine$_2$, spine$_3$, left hip, left knee, left ankle, left foot
- Right Leg: spine$_1$, spine$_2$, spine$_3$, right hip, right knee, right ankle, right foot
- Torso: spine$_1$, spine$_2$, spine$_3$, neck, left collar, right collar, head

The pelvis, spine$_1$, spine$_2$, spine$_3$ are shared across all parts, as they remain relatively stable during human motion. Each joint is represented by relative 6D rotations and redundant 3D positions, resulting in a dimensionality of 63+8 per part, including 4D root node and 4D foot contact information. When aggregating part features into motion features, we average the shared joints.

3. The HuMo100M Dataset

In this section, we detail our large-scale multimodal human motion dataset, HuMo100M. HuMo100M not only integrates existing publicly available human motion datasets but also significantly expands the scale by extracting a large number of motion sequences from web videos using the WHAM[15].

3.1. Motion Data Collection

Motion Data Collection. We start by collecting over 20 million videos from publicly available datasets and online platforms, such as YouTube. To ensure motion quality, we first use a keyword-based filtering approach to discard videos whose associated textual metadata lacks human-related terms, such as *people*, *human*, *person*, *man*, and *woman*.

- **Video Boundary Detection:** Although the first-stage filtering ensures the presence of humans in the video, We observe that many segments in these videos lack human presence. To further improve the quality of the dataset, we employ YOLO[14] to track every individual throughout the video to obtain a precise human-related segments. Then, we conduct video clip segmentation. To ensure the integrity of the trajectory, we conduct clip segmentation utilizing the tracking results from YOLO.
- **Occlusion and Blur Filtering:** Occlusion and motion blur are common challenges in human-related videos. To mitigate these issues, we develop the following quality assessment framework to enhance the reliability of our dataset. First, we apply a pre-trained 2D keypoint detector to extract skeleton key points for each human and filter out detection results with low confidence. Next, we utilize the number of detected high-quality key points as the primary criterion for assessing potential occlusions. Specifically, a motion sequence is considered to contain significant occlusion if the number of visible key points is below a minimum threshold. Second, we implement a quality control method based on sequence duration. Short human motion sequences contain insufficient temporal information and often exhibit reduced estimation accuracy. Therefore, we have implemented a simple but effective sequence length filtering mechanism that removes motion sequences containing few frames to ensure data quality.

Motion Description Generation. We utilize Gemini-1.5-pro[16] with a carefully designed prompt and PoseScript[4] to generate hierarchical motion descriptions including: 1) body-level description, which provides a high-level and overall body movements description in a short sentence; 2) part-level description, which provides a description to describe the movements of upper body and lower body respectively. The upper body descriptions capture motions involving the arms and torso, while the lower body texts focus on the movement of the legs and feet; 3) rule-base description, which describes the relative position between different joints by utilizing several *posecodes* to extract semantic pose information, such as 'the left hand is below the right hand'.

3.2. Statistic Analysis of Data and Word Distribution

Data Distribution. Figure 1 shows the distribution of motion length across different subsets of HuMo100M. We can observe that HuMo100M integrates motion sequences from existing datasets (such as Motion-X[8], 3DPW[17], and MSCOCO[9]). Notably, Posetrack[1] contain the lowest average number of frames (16.12 frames), indicating a focus on short motion sequences. In contrast, PROX[6] and BEHAVE[3] contain high average frame numbers (994.19 frames and 975.78 frames, respectively), indicating that the motion sequences in the dataset are diverse in length. Figure 2 shows the distribution of the number of motions in different subsets of HuMo100M (logarithmic scale) and demonstrates the difference in the number of motion instances across subsets, ranging from 27 motions in PROX[6] to 2,376,376 motions in Webvid[2].

Word Distribution. To further investigate the annotated

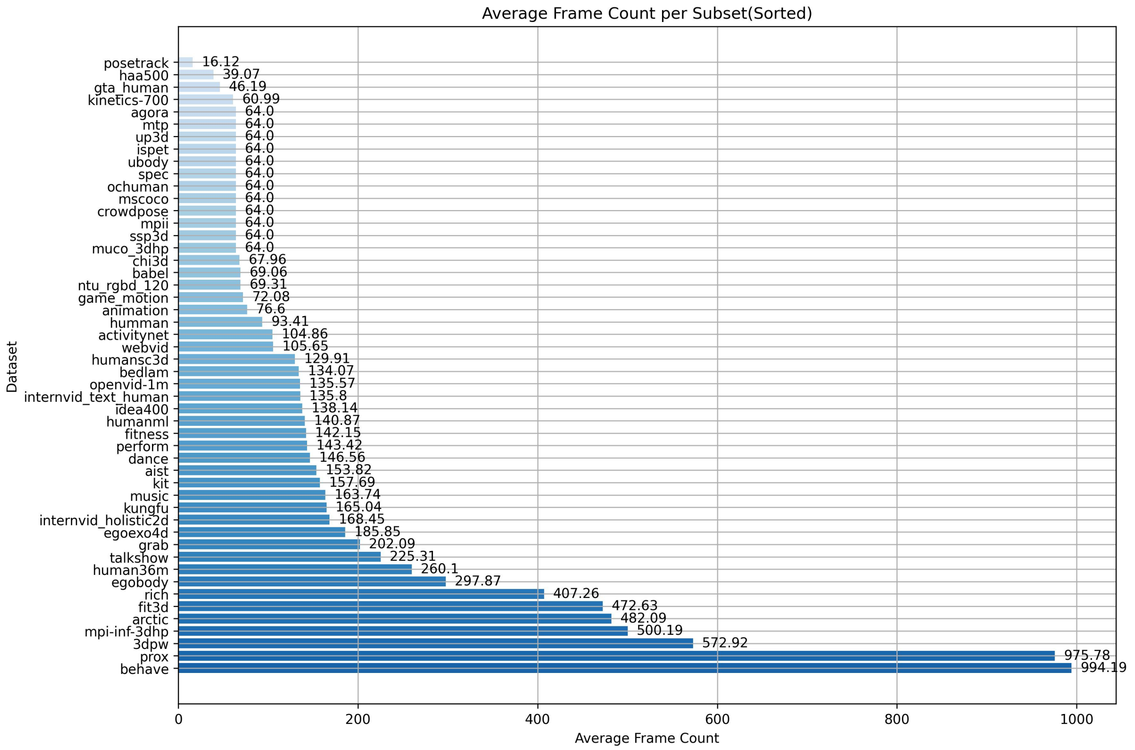


Figure 1. The length distribution across different subsets in HuMo100M

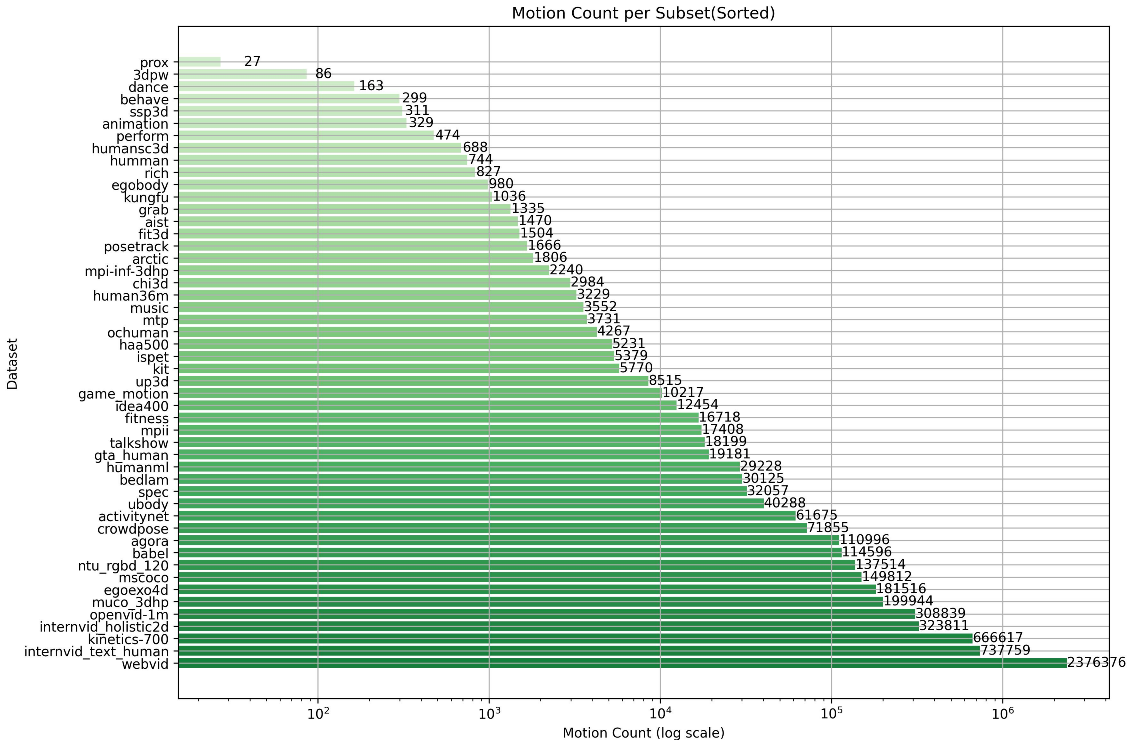


Figure 2. The motion count distribution across different subsets in HuMo100M

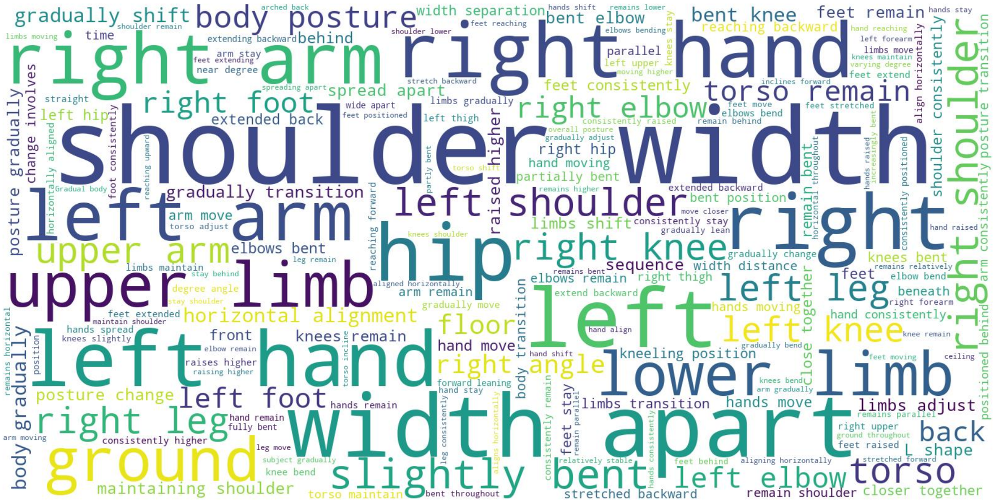


Figure 3. Word cloud of rule-base motion descriptions in HuMo100M

appear
sitting
standing
interacting
conversation
gesturing
walking
initially standing
holding
speaking
hand
performing
standing relatively
something
camera
likely
sitting
front
frame
talking
audience
seemingly engaged
playing
presentation
standing appearing
likely addressing
microphone
object
engaged
table
right hand
standing interacting
sitting seemingly
standing likely
small object
likely speaking
standing seemingly
explaining something
left
within
observing
appearing
listening
right
standing gesturing
walking toward
leaning forward
kneeling
crouching
dancing
lying
standing holding
seated
likely engaging
blue box
standing initially

Figure 4. Word cloud of body-level motion descriptions in HuMo100M

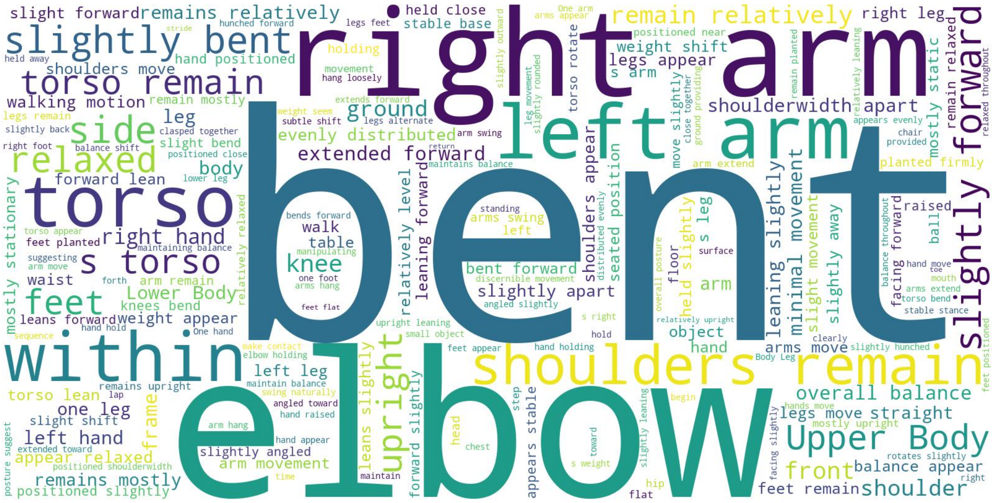


Figure 5. Word cloud of part-level motion descriptions in HuMo100M

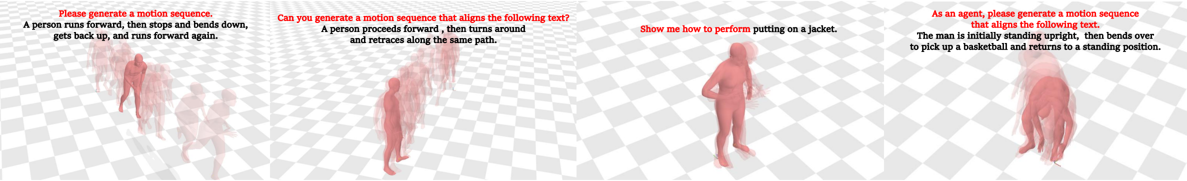

Please generate a motion sequence.
A person runs forward, then stops and bends down, gets back up, and runs forward again.
Can you generate a motion sequence that aligns the following text?
A person proceeds forward , then turns around and retraces along the same path.
Show me how to perform putting on a jacket.
As an agent, please generate a motion sequence that aligns the following text.
The man is initially standing upright, then bends over to pick up a basketball and returns to a standing position.

Figure 6. Visualization results generated by MotionCtrl given the random instruction.

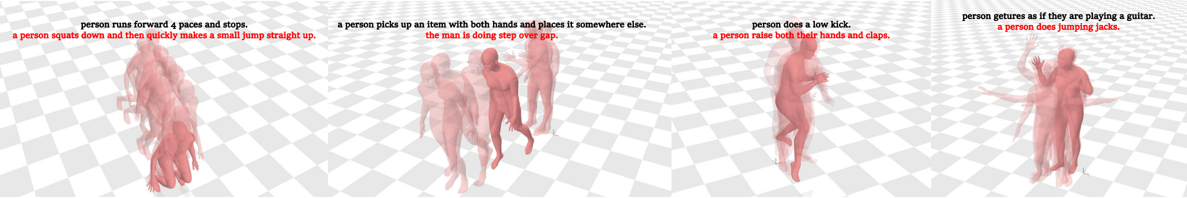

person runs forward 4 paces and stops.
a person squats down and then quickly makes a small jump straight up.
a person picks up an item with both hands and places it somewhere else.
the man is doing step over gap.
person does a low kick.
a person raise both their hands and claps.
person getures as if they are playing a guitar.
a person does jumping jacks.

Figure 7. Visualization results generated by MotionCtrl given the long-term instruction.

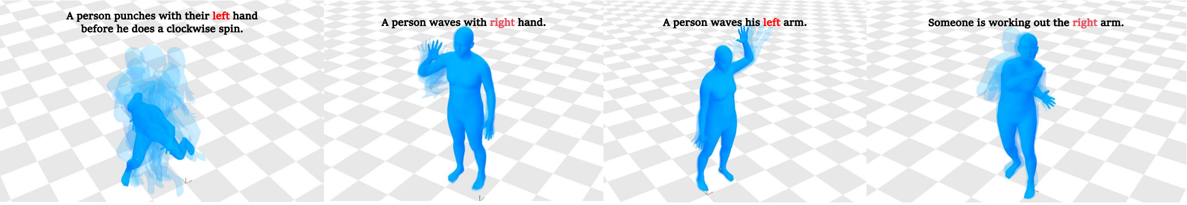

A person punches with their left hand before he does a clockwise spin.
A person waves with right hand.
A person waves his left arm.
Someone is working out the right arm.

Figure 8. Visualization results generated by MotionCtrl given the part-level instruction.

motion text within HuMo100M, we generate word clouds from the entire text corpus to visualize its linguistic patterns. Specifically, we compute three distinct word clouds for rule-based text, part-level text, and body-level text respectively. Figure 3 reveals that the rule-based text emphasizes semantic relationships between joints, such as positional constraints. In contrast, Figure 5 demonstrates that part-level descriptions focus on detailed movements of specific body parts, including the torso, shoulders, legs, and arms. Figure 4 highlights that body-level text predominantly captures high-level human activities, such as standing, sitting, and walking. This hierarchical structure of the text corpus, spanning rule-base, part-level, and body-level annotations, facilitates enhanced alignment in VLMMs.

3.3. Instruction Generation

To fully leverage the potential of the HuMo100M dataset and support diverse downstream tasks (such as part-level motion control and vision-based motion understanding), we generate rich and high-quality instructions for the dataset. These instructions not only include traditional whole-body motion descriptions but also incorporate fine-grained descriptions for individual body parts. We primarily generate these instructions in two ways:

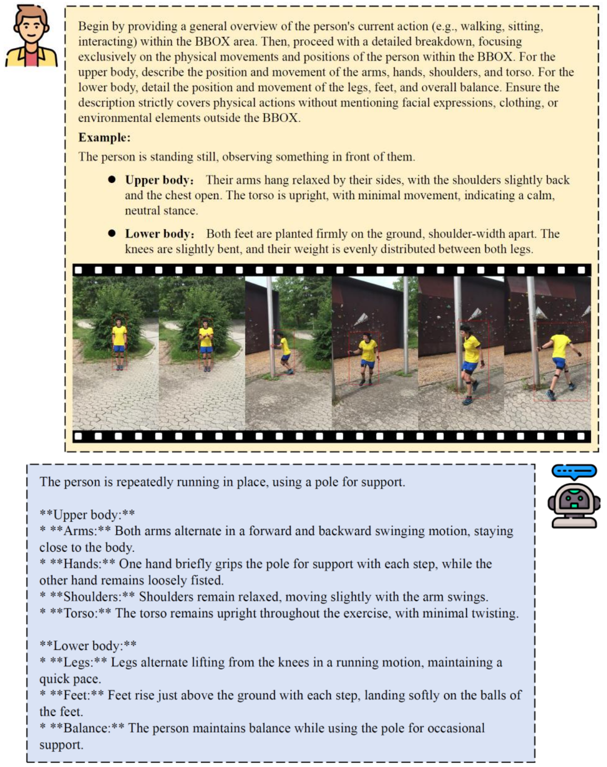
- **Generation based on Large Multimodal Models (LMMs):** We design detailed prompt templates (as shown in Figure 9) and utilize powerful LMMs, such as Gemini-1.5-pro [16], to extract part-level motion descriptions from videos. This approach enables the generation of semantically rich descriptions in natural language.
- **Rule-based Generation (PoseScript):** In addition to the descriptions generated by LMMs, we also utilize PoseScript [4] and *posecodes* to extract semantic pose information. This allows us to generate instructions that describe the relative positions between different joints, such as ”the left hand is below the right hand.”

3.4. Visualization Examples

Figures 6, 7, and 8 collectively present a series of motion visualization results generated by MotionCtrl from random instructions, initial pose, long-term text command, and part-level commands respectively. These visualizations demonstrate that MotionCtrl can effectively handle and process instructions in a variety of different formats, showcasing its versatility and robustness in motion generation tasks.

3.5. Long-term Motion Concatenation

Most existing human motion datasets are characterized by short-duration motion sequences, undermining the model’s ability to generate long-sequence motions from textual descriptions. To address this critical gap, we introduce two motion concatenation methods, integrating shorter motion



Begin by providing a general overview of the person's current action (e.g., walking, sitting, interacting) within the BBOX area. Then, proceed with a detailed breakdown, focusing exclusively on the physical movements and positions of the person within the BBOX. For the upper body, describe the position and movement of the arms, hands, shoulders, and torso. For the lower body, detail the position and movement of the legs, feet, and overall balance. Ensure the description strictly covers physical actions without mentioning facial expressions, clothing, or environmental elements outside the BBOX.

Example:

The person is standing still, observing something in front of them.

- **Upper body:** Their arms hang relaxed by their sides, with the shoulders slightly back and the chest open. The torso is upright, with minimal movement, indicating a calm, neutral stance.
- **Lower body:** Both feet are planted firmly on the ground, shoulder-width apart. The knees are slightly bent, and their weight is evenly distributed between both legs.

```
The person is repeatedly running in place, using a pole for support.

**Upper body:**
* **Arms:** Both arms alternate in a forward and backward swinging motion, staying
close to the body.
* **Hands:** One hand briefly grips the pole for support with each step, while the
other hand remains loosely fisted.
* **Shoulders:** Shoulders remain relaxed, moving slightly with the arm swings.
* **Torso:** The torso remains upright throughout the exercise, with minimal twisting.

**Lower body:**
* **Legs:** Legs alternate lifting from the knees in a running motion, maintaining a
quick pace.
* **Feet:** Feet rise just above the ground with each step, landing softly on the balls of
the feet.
* **Balance:** The person maintains balance while using the pole for occasional
support.
```

Figure 9. The Prompt template to generate part-level motion description in videos based on powerful large multimodal models (LMMs), such as Gemini-1.5-pro and GPT-4o-mini. For each sample in HuMo100M, we provide "body-level" (UP) and "part-level" (DOWN) labels to distinguish between whole-body and partial motion descriptions.

sequences into longer, consistent motion sequence.

Interpolation-based. To enable seamless motion concatenation, we implement a three-stage alignment process. First, we perform orientation alignment between the two motion sequences to ensure consistent directional coherence. Subsequently, we conduct global coordinate alignment through spatial translation of the motion sequences. Finally, we establish a smooth transition by selecting a standing pose as the reference pose and applying spherical linear interpolation (Slerp) between the motion sequences and the reference pose. This approach guarantees a natural transition where the character returns to a standard standing posture after completing the first motion sequence, maintaining continuity and physical plausibility throughout the concatenation process.

Learning-based. We introduce an in-between motion prediction model designed to generate smooth transitions between two motion sequences. Our approach builds upon a pre-trained text-to-motion model, which is fine-tuned specifically for this task. During training, we employ a masking strategy that masks approximately 50% of the motion sequence, enabling the model to obtain robust motion completion capability. During inference, the model is used to concatenate two motion sequences through a two-step process. First, an interpolation-based method generates initial key frames to bridge the sequences. These interpolated frames are then treated as a masked motion sequence and fed into the fine-tuned in-between model, which refines and completes the transition. This hybrid framework leverages the strengths of both data-driven and rule-based methods, ensuring that the generated motion transitions are not only smooth but also contextually consistent with the input sequences.

Method	specialized area	FID ↓	R@1 ↑	R@3 ↑	MMDist ↓
MotionDiffuse [20]	partial control	0.630	0.491	0.782	3.113
ParCo [22]	partial control	0.109	0.515	0.801	2.927
Fg-T2M++[18]	fine-grained text control	0.089	0.513	0.801	2.925
MotionCtrl-PRQ$_4$	-	**0.056**	**0.535**	**0.821**	**2.865**

Table 1. Effectiveness of multi-task motion training on HuMo-t2m testbed using different configuration.

4. Additional Experimental Analysis

Controllability Experiments. MotionCtrl tackles the challenge of part-specific motion control by representing the body with five part-level tokens instead of a single token. This design enables pretraining on part-level motion-text pairs and selective decoding of specific body parts during inference. While prior works have also explored partial body control or pose conditioning in isolation [18, 19, 22], our method demonstrates superior controllability. We evaluate MotionCtrl against recent part-specific, fine-grained control approaches on the HumanML3D dataset, where our model outperforms existing methods, as shown in Table 1. To further quantify control accuracy, we introduce human part awareness as a controllability metric. Following ParCo [22], we conduct left-right exchange tests on 50 sentences. Our method achieves a 76% success rate, significantly surpassing ParCo (64%) and T2M (46%), demonstrating stronger part-aware motion generation.

Inference Speed Comparison. We benchmark our method against Momask, a large-scale motion model. While both approaches generate initial motions in real time, our method incurs a slightly longer latency due to its $>3\times$ token output for part-level control and a larger model size. However, Momask's HM263 feature cannot produce directly applicable motions, as they require time-consuming inverse kinematics (IK) post-processing. In contrast, our proposed HuMo263 feature eliminates this bottleneck, enabling real-time, part-level controllable motion generation without additional refinement. Thus, compared to Momask, we achieve real-time performance, direct usability, and fine-grained part-level control simultaneously.

Motion Tokenizer Selection. We analyze the trade-offs in motion tokenizer selection. While PRQ adopted by our MotionCtrl achieves superior performance on R@# and MMDist compared to RQ used by Momask, it exhibits higher FID scores, which we attribute to its fewer quantized layers, which can be seen in Table 1 in the main paper (Row 4 vs. Row 6). In addition, the Table 4 of main paper demonstrates that FID is highly dependent on layer count: with equal layers, PRQ achieves competitive reconstruction FID. However, increasing layers introduces more motion tokens, complicating LLM-based decoding. To balance these factors, we select PRQ_4, preserving the LLM's advantages in tasks like I2M while maintaining efficient motion representation.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5167–5176, 2018. 2

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 2

[3] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 2

[4] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: Linking 3d human poses and natural language. *IEEE transactions on pattern analysis and machine intelligence*, 2024. 2, 5

[5] 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

[6] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 2

[7] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale rgb-d database for arbitrary-view human action recognition. In *Proceedings of the 26th ACM international Conference on Multimedia*, pages 1510–1518, 2018. 1

[8] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36: 25268–25280, 2023. 1, 2

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 2

[10] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 1

[11] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 1

[12] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 1

[13] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 1

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[15] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024. 2

[16] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2, 5

[17] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 1, 2

[18] Yin Wang, Mu Li, Jiapeng Liu, Zhiying Leng, Frederick WB Li, Ziyao Zhang, and Xiaohui Liang. Fg-t2m++: Llms-augmented fine-grained text driven human motion generation. *International Journal of Computer Vision*, pages 1–17, 2025. 6

[19] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 6

[20] 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. 6

[21] Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, et al. Large motion model for unified multi-modal motion generation. In *European Conference on Computer Vision*, pages 397–421. Springer, 2024. 1

[22] Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. Parco: Part-coordinating text-to-motion synthesis. In *European Conference on Computer Vision*, pages 126–143. Springer, 2024. 6