Re^{2}MoGen: Open-Vocabulary Motion Generation via LLM Reasoning and Physics-Aware Refinement

Supplementary Material

7. A. Implementation Details

7.1. Hyperparameter

In the LLM planning, we employ **DeepSeek-R1** as the reasoning model, which exhibits superior spatial understanding and reasoning capabilities. The CLIP model mentioned in the paper uses the **CLIP-ViT-L/14** model from OpenCLIP [5]. The VLM used in our experiments is **Qwen-VL-Max**. The detailed hyperparameters involved in our method are presented in Table 4.

Table 4. Hyperparameters of Re^{2}MoGen.

Hyperparameter	Value
K_s	2
MCTS Iteration	30
Exploration Parameter α	0.05
Smoothing Parameter γ	0.1
Weight Parameter λ	0.01
MLD Fine-tune Batch Size	64
Learning Rate	$1e-4$
PPO Clip Threshold	$1e-3$
Buffer Size	3000
Samples Per Update Iteration	8
Policy Training Batch Size	128
KL Weight	0.01

7.2. Prompt Design

We use LLMs to plan keyframes for human motions. The quality of the planned keyframes depends not only on the model itself but also on the input prompt. Our prompt template is shown in Fig. 6, which includes four key points:

1. To simplify the reasoning task, we do not ask the LLM to plan motions for full-body joints. The reasons are: 1) excessive data processing would significantly degrade the reasoning quality of the model; 2) the strong interdependencies among full-body joints make the reasoning less tolerant to errors. Therefore, we only make the LLM infer motions for 5 key joints and plan only the displacement of each keyframe relative to the previous one.
2. We define a set of foundational information as follows:
   - Human skeleton information, derived from the SMPL [25] Neutral skeleton.
   - The displacement direction information is not represented in XYZ coordinates, but in directional terms such as "Forward-Backward", "Up-Down", and "Left-Right", which helps the LLM better grasp movement orientation.
   - The initial body pose is shown in Fig. 5, which helps standardize the planning's initialization conditions.
   - Task information, reminding the LLM of its objective.
3. Additionally, we require the LLM to output the reasoning behind its planned keyframes, encouraging more thorough deliberation during inference. We provide a formatted output template to facilitate information extraction and subsequent usage.
4. We include two examples to achieve few-shot fine-tuning, enabling the LLM to quickly adapt to our task during inference. Fig. 7 shows these two examples, and Fig. 8 explains the reasons.

With the above prompt template, the LLM's capability in planning such a task can be significantly enhanced.

7.3. Motion Representations

In our implementation, we utilize three distinct motion data representations:

1. **During LLM reasoning**, motion is represented by K planned keyframes. Each keyframe j_{key} contains X-Y-Z coordinates for 5 key joints (pelvis, l/r_ankle, l/r_wrist).
2. **During full-body poses optimization**, the optimized poses follow the SMPL-format [25], structured as a dictionary containing three components:
   - Global translation: Root joint X-Y-Z coordinates representing global body position.
   - Root rotation: Euler angles representing overall body rotation.
   - Body pose: Euler angles for 21 joints (excluding hands and facial joints).
3. **During MLD fine-tuning and post-training**, the pose of each frame for the model input uses the 263-dimensional format defined by HumanML3D, derived from SMPL-format data. This 263-dimensional format of data includes root angular/linear velocity, root height, joint rotation invariant position, joint rotation, joint linear velocity, and foot contact information. It must be noted that we maintain consistent pose notation with full-body pose optimization for clarity in the main text, and that all HumanML3D-format data can be converted to/from SMPL-format.

7.4. Physics-aware Reward Design

In the physics-aware refinement section, we use RL post-training with the following three reward functions:

- **Foot Sliding Reward.** This reward function imposes constraints by penalizing foot movements that deviate from expected locomotion patterns, formally defined as:

$$r_{\mathrm{S}}(m) = \frac{1}{L}\sum_{i=1}^{L}\exp(-\|(p_{ft}^{i} - p_{ft}^{i-1})\cdot p_{c}^{i}\cdot p_{c}^{i-1}\|_2), \quad (13)$$

  where $m = p^{1:L}$ is mentioned in the main text, p_{ft}^{i} and p_{c}^{i} represent the i-th pose's foot joint positions and the contact label, respectively.
- **Foot Floating Reward.** This reward function penalizes floating postures when foot-ground contact is required, which is formally defined as:

$$\begin{aligned} r_{\mathrm{F}}(m) = \frac{1}{L}\sum_{i=1}^{L}\exp(-||(\min_{v} p_i^v - h_{ground}) \\ \cdot \mathbb{I}[\min_{v} p_i^v > h_{\text{ground}}]||_2), \end{aligned} \quad (14)$$

  where $\min_v p_i^v$ represents the lowest body mesh vertex of the pose p_i, h_{ground} denotes the height of ground.
- **Ground Penetration Reward.** This reward function encourages all motions remain grounded, defined as:

$$\begin{aligned} r_{\mathrm{P}}(m) = \frac{1}{L}\sum_{i=1}^{L}\exp(-\|(h_{ground} - \min_{v} p_i^v) \\ \cdot \mathbb{I}[\min_{v} p_i^v < h_{\text{ground}}]\|_2). \end{aligned} \quad (15)$$

Figure 5. The initial pose for LLM planning.

8. B. Experiments Details

8.1. LLM-planned Keyframes

As shown in Figs.(10-16), we present the JSON data of keyframes planned by the LLM of different motion lengths alongside the rendered pose images after full-body optimization. These results demonstrate that LLM can reasonably plan actions based on text descriptions.

8.2. Evaluation Metrics

As mentioned in the main text, we evaluate the generated motions from two aspects: semantic alignment and physical plausibility. Thus, we use the following four metrics:

- CLIP score (CLIP_S): The average clip similarity between each frame of the rendered motion videos and the corresponding description, which is calculated as,

$$\text{CLIP\_S} = \frac{1}{L}\sum_{i=1}^{L}\cos\left(\phi_{\text{text}}(c), \phi_{\text{image}}(I_i)\right), \quad (16)$$

  where L is the motion length, c is the motion description, I_i represents the rendered image of each motion frame. ϕ_{text} and ϕ_{image} denote the CLIP text encoder and CLIP image encoder, respectively.
- VLM score (VLM_S): The weighted score combines the VLM's evaluations of semantic alignment and naturalness for the rendered motion video v, denoted as,

$$\text{VLM\_S} = \sigma_s \cdot \text{VLM}_{\text{S}}(v, c) + \sigma_n \cdot \text{VLM}_{\text{N}}(v), \quad (17)$$

  where VLM_{S} and VLM_{N} are distinct evaluation prompts designed to assess semantic alignment and naturalness, respectively, as illustrated in Fig. 9. The weights σ_s and σ_n are set to 0.6 and 0.4 respectively. For each motion, we compute the average of 10 VLM scores to ensure statistical robustness.
- Floating (Float): The degree of floating from the ground.

$$\text{Float} = \frac{1}{L}\sum_{i=1}^{L}|\min_{v} p_i^v - h_{\text{ground}}|\cdot\mathbb{I}[\min_{v} p_i^v > h_{\text{ground}}], \quad (18)$$

  where $\min_v p_i^v$ represents the lowest body mesh vertex of the pose p_i.
- Penetration (Pene): The degree of ground penetration.

$$\text{Pene} = \frac{1}{L}\sum_{i=1}^{L}|\min_{v} p_i^v - h_{\text{ground}}|\cdot\mathbb{I}[\min_{v} p_i^v < h_{\text{ground}}]. \quad (19)$$

9. C. Additional Experiments

Table 5. Results on SnapMotion dataset.

Methods	CLIP_S ↑	VLM_S ↑
MLD	$23.39_{\pm 0.40}$	$1.64_{\pm 0.24}$
MotionGPT	$22.36_{\pm 0.51}$	$1.42_{\pm 0.14}$
MoMask	$23.92_{\pm 0.36}$	$1.75_{\pm 0.17}$
Ours	$\mathbf{24.68}_{\pm \mathbf{0.64}}$	$\mathbf{2.17}_{\pm \mathbf{0.40}}$

You are an expert on Kinematics and Human Motion. You are tasked to plan the movement of the four end controllers (L_Ankle, R_Ankle, L_Wrist, R_Wrist) and the root joint (Pelvis) for some keyframes based on the motion description. The total number of motion frames is **{INPUT_LENGTH}**, please plan a keyframe every 10 frames. The FPS is 20 Hz.

Explanation of the human skeleton:
The length from Ankle to Hip: 0.7731 m.
The length from Ankle to Knee: 0.3979 m.
The length from Wrist to Shoulder: 0.5089 m.
The length from Wrist to Elbow: 0.2492 m.
The length from Pelvis to the ground: 0.9052 m.
The length from Head to the ground: 1.4811 m.
The length from Shoulder to the ground: 1.3537 m.

Explanation of the three movement directions:
Forward-Backward direction: If the joint moves forward, this is indicated by a positive number, and if the joint moves backward, this is indicated by a negative number.
Up-Down direction: If the joint moves up, this is indicated by a positive number, and if the joint moves down, this is indicated by a negative number.
Left-Right direction: If the joint moves left, this is indicated by a positive number, and if the joint moves right, this is indicated by a negative number.
If the joint does not move in a certain direction, it is indicated by 0. Movements in each direction are in meters (m).

Explanation of the initial pose: The initial posture is the body standing naturally. Arms naturally hang down at the sides of the body, legs are upright and shoulder-width apart, forming a standard standing posture. The wrist is 0.84 meters above the ground. The distance between wrist and head is 0.6363 m. The width between feet is 0.3 m.

Motion description: {DESCRIPTION}
Task: You need to utilize your understanding of human motion and help me plan the movement delta of the four end controllers (L_Ankle, R_Ankle, L_Wrist, R_Wrist) and the root joint (Pelvis) in three directions for some keyframes when executing the given motion description. The total number of motion frames is **{INPUT_LENGTH}**, please plan a keyframe every 10 frames. The FPS is 20 Hz. The first keyframe (F0) is the initial pose.
Please describe the text of each frame just like the following example, specifically what action has been performed compared to the previous frame.

The movement delta of the current keyframe is based on the previous keyframe (the first keyframe is represented by the initial pose). The movement process requires maintaining full body coordination. You should write all the keyframe together. **The response should follow the format:**
{"F0":{"Pelvis":{"Left-Right":"Movement Distance","Up-Down":"Movement Distance","Forward-Backward":"Movement Distance"},"L_Ankle":{"Left-Right":"Movement Distance","Up-Down":"Movement Distance","Forward-Backward":"Movement Distance"},"R_Ankle":{"Left-Right":"Movement Distance","Up-Down":"Movement Distance","Forward-Backward":"Movement Distance"},"L_Wrist":{"Left-Right":"Movement Distance","Up-Down":"Movement Distance","Forward-Backward":"Movement Distance"},"R_Wrist":{"Left-Right":"Movement Distance","Up-Down":"Movement Distance","Forward-Backward":"Movement Distance"}},
"F1":{"Pelvis":{"Left-Right":"Movement Distance","Up-Down":"Movement Distance","Forward-Backward":"Movement Distance"},"L_Ankle":{"Left-Right":"Movement Distance","Up-Down":"Movement Distance","Forward-Backward":"Movement Distance"},"R_Ankle":{"Left-Right":"Movement Distance","Up-Down":"Movement Distance","Forward-Backward":"Movement Distance"},"L_Wrist":{"Left-Right":"Movement Distance","Up-Down":"Movement Distance","Forward-Backward":"Movement Distance"},"R_Wrist":{"Left-Right":"Movement Distance","Up-Down":"Movement Distance","Forward-Backward":"Movement Distance"}},...}

Here are two examples: **EXAMPLE_1, EXAMPLE_1_REASON; EXAMPLE_2, EXAMPLE_2_REASON.**

Figure 6. Our prompt template for LLM reasoning.

9.1. Experiments on SnapMotion

Table 5 demonstrates the performance of our method on 100 text descriptions from the SnapMotion [15] dataset, comparing it with methods in motion generation, including MoMask [14]. The results highlight the superior capability of our approach in generating motions from unseen text descriptions.

9.2. Impact of Different LLMs

To evaluate the impact of different LLMs on the experimental results, we also conduct experiments on three other LLMs, i.e, Qwen-Plus, InternLM, and DeepSeek-V3. Qwen-Plus and DeepSeek-R1, used in our experiments, possess deep thinking capabilities, while InternLM and DeepSeek-V3 do not. The result is shown in Table 6. It can be observed that LLMs with deep thinking capabilities demonstrate superior performance, as they can more thoroughly analyze tasks, reflect on the rationality of their planning and reasoning steps, and correct their thought processes. Among these, the DeepSeek-R1 model we employed exhibits particularly outstanding abilities in this regard.

Table 6. Ablation study with different LLMs.

Methods	CLIP_S↑	VLM_S↑
DeepSeek-V3	$22.59_{\pm 0.75}$	$1.77_{\pm 0.16}$
InternLM	$22.1_{\pm 0.67}$	$1.54_{\pm 0.25}$
Qwen-Plus	$23.12_{\pm 0.51}$	$1.83_{\pm 0.24}$
DeepSeek-R1 (Ours)	$\mathbf{23.64_{\pm 0.49}}$	$\mathbf{2.72_{\pm 0.21}}$

9.3. Analysis of Smoothing Parameter γ

To analyze the effect of different smoothing parameter γ values, we conduct additional experiments with $\gamma = \{0.01,$

Example 1:

Input: Motion Description: kick something or someone with right leg.
Input length: 70.

Output:{
"F0": {"Pelvis": {"Left-Right": 0,"Up-Down": 0,"Forward-Backward": 0},
"L_Ankle": {"Left-Right": 0,"Up-Down": 0,"Forward-Backward": 0},
"R_Ankle": {"Left-Right": 0,"Up-Down": 0,"Forward-Backward": 0},
"L_Wrist": {"Left-Right": 0,"Up-Down": 0,"Forward-Backward": 0},
"R_Wrist": {"Left-Right": 0,"Up-Down": 0,"Forward-Backward": 0}},
"F1": {"Pelvis": {"Left-Right": 0.0173,"Up-Down": -0.0121,"Forward-Backward": 0.0054},
"L_Ankle": {"Left-Right": 0.0239,"Up-Down": 0.0331,"Forward-Backward": 0.1046},
"R_Ankle": {"Left-Right": 0.0026,"Up-Down": -0.0052,"Forward-Backward": 0.0131},
"L_Wrist": {"Left-Right": 0.1577,"Up-Down": 0.3209,"Forward-Backward": 0.0761},
"R_Wrist": {"Left-Right": 0.0827,"Up-Down": 0.2248,"Forward-Backward": 0.1716}},
"F2": {"Pelvis": {"Left-Right": 0.1443,"Up-Down": -0.0518,"Forward-Backward": 0.1249},
"L_Ankle": {"Left-Right": 0.1186,"Up-Down": -0.0330,"Forward-Backward": 0.2038},
"R_Ankle": {"Left-Right": -0.0323,"Up-Down": 0.0752,"Forward-Backward": 0.0735},
"L_Wrist": {"Left-Right": 0.0087,"Up-Down": -0.0729,"Forward-Backward": -0.3523},
"R_Wrist": {"Left-Right": 0.3162,"Up-Down": 0.4886,"Forward-Backward": -0.0922}},
"F3": {"Pelvis": {"Left-Right": 0.0673,"Up-Down": 0.0547,"Forward-Backward": 0.0448},
"L_Ankle": {"Left-Right": 0.0072,"Up-Down": 0.0092,"Forward-Backward": -0.0325},
"R_Ankle": {"Left-Right": 0.4944,"Up-Down": 0.9342,"Forward-Backward": 0.7092},
"L_Wrist": {"Left-Right": 0.1897,"Up-Down": -0.0138,"Forward-Backward": 0.2193},
"R_Wrist": {"Left-Right": 0.2870,"Up-Down": -0.2250,"Forward-Backward": 0.0127}},
"F4": {"Pelvis": {"Left-Right": 0.0261,"Up-Down": -0.0429,"Forward-Backward": 0.0093},
"L_Ankle": {"Left-Right": -0.0190,"Up-Down": -0.0252,"Forward-Backward": 0.0491},
"R_Ankle": {"Left-Right": -0.0112,"Up-Down": -0.3097,"Forward-Backward": 0.0877},
"L_Wrist": {"Left-Right": 0.0013,"Up-Down": -0.1470,"Forward-Backward": -0.0394},
"R_Wrist": {"Left-Right": 0.0674,"Up-Down": -0.2762,"Forward-Backward": 0.1991}},
"F5": {"Pelvis": {"Left-Right": -0.0551,"Up-Down": 0.0312,"Forward-Backward": -0.0651},
"L_Ankle": {"Left-Right": -0.0369,"Up-Down": 0.0080,"Forward-Backward": -0.0518},
"R_Ankle": {"Left-Right": -0.5850,"Up-Down": -0.4942,"Forward-Backward":-0.7889},
"L_Wrist": {"Left-Right": 0.0249,"Up-Down": 0.2698,"Forward-Backward": 0.1625},
"R_Wrist": {"Left-Right": -0.1766,"Up-Down": 0.0731,"Forward-Backward": 0.0094}},
"F6": {"Pelvis": {"Left-Right": -0.1989,"Up-Down": -0.0088,"Forward-Backward": -0.1649},
"L_Ankle": {"Left-Right": -0.0187,"Up-Down": 0.0241,"Forward-Backward": -0.0171},
"R_Ankle": {"Left-Right": 0.1365,"Up-Down": -0.1935,"Forward-Backward": -0.1971},
"L_Wrist": {"Left-Right": -0.2028,"Up-Down": 0.1424,"Forward-Backward": 0.1219},
"R_Wrist": {"Left-Right": -0.3250,"Up-Down": 0.0297,"Forward-Backward":-0.2011}},
"F7": {"Pelvis": {"Left-Right": -0.0056,"Up-Down": 0.0004,"Forward-Backward": -0.0410},
"L_Ankle": {"Left-Right": -0.0245,"Up-Down": -0.0110,"Forward-Backward":-0.0192},
"R_Ankle": {"Left-Right": 0.0460,"Up-Down": -0.0056,"Forward-Backward": -0.0026},
"L_Wrist": {"Left-Right": 0.0306,"Up-Down": -0.2492,"Forward-Backward": -0.0857},
"R_Wrist": {"Left-Right": -0.0997,"Up-Down":-0.0481,"Forward-Backward":-0.1939}}}

Example 2:

Input: Motion Description: walk forward.
Input length: 70

Output:{
"F0": {"Pelvis": {"Left-Right": 0.0,"Up-Down": 0.0,"Forward-Backward": 0.0},
"L_Ankle": {"Left-Right": 0.0,"Up-Down": 0.0,"Forward-Backward": 0.0},
"R_Ankle": {"Left-Right": 0.0,"Up-Down": 0.0,"Forward-Backward": 0.0},
"L_Wrist": {"Left-Right": 0.0,"Up-Down": 0.0,"Forward-Backward": 0.0},
"R_Wrist": {"Left-Right": 0.0,"Up-Down": 0.0,"Forward-Backward": 0.0}},
"F1": {"Pelvis": {"Left-Right": 0.0009,"Up-Down": 0.0007,"Forward-Backward": 0.0031},
"L_Ankle": {"Left-Right": -0.0018,"Up-Down": 0.0005,"Forward-Backward": -0.003},
"R_Ankle": {"Left-Right": 0.0026,"Up-Down": -0.0001,"Forward-Backward": -0.0014},
"L_Wrist": {"Left-Right": 0.0,"Up-Down": -0.0022,"Forward-Backward": 0.0041},
"R_Wrist": {"Left-Right": 0.0033,"Up-Down": -0.0031,"Forward-Backward": 0.0047}},
"F2": {"Pelvis": {"Left-Right": -0.0099,"Up-Down": -0.0007,"Forward-Backward": 0.0144},
"L_Ankle": {"Left-Right": 0.0023,"Up-Down": 0.0067,"Forward-Backward": -0.0112},
"R_Ankle": {"Left-Right": 0.0044,"Up-Down": 0.0019,"Forward-Backward": -0.0063},
"L_Wrist": {"Left-Right": -0.0073,"Up-Down": -0.0001,"Forward-Backward": 0.0175},
"R_Wrist": {"Left-Right": -0.0156,"Up-Down": -0.0017,"Forward-Backward": 0.0089}},
"F3": {"Pelvis": {"Left-Right": 0.007,"Up-Down": -0.0307,"Forward-Backward": 0.1741},
"L_Ankle": {"Left-Right": 0.0496,"Up-Down": 0.0186,"Forward-Backward": 0.5713},
"R_Ankle": {"Left-Right": 0.0069,"Up-Down": 0.0085,"Forward-Backward": -0.0158},
"L_Wrist": {"Left-Right": 0.0037,"Up-Down": -0.0287,"Forward-Backward": 0.0717},
"R_Wrist": {"Left-Right": -0.012,"Up-Down": -0.0334,"Forward-Backward": 0.2049}},
"F4": {"Pelvis": "Left-Right": 0.0995,"Up-Down": 0.0201,"Forward-Backward": 0.4949},
"L_Ankle": {"Left-Right": 0.0004,"Up-Down": -0.0131,"Forward-Backward": 0.0359},
"R_Ankle": {"Left-Right": 0.079,"Up-Down": 0.0086,"Forward-Backward": 0.9333},
"L_Wrist": {"Left-Right": 0.147,"Up-Down": 0.0564,"Forward-Backward": 0.6979},
"R_Wrist": {"Left-Right": 0.0843,"Up-Down": 0.0234,"Forward-Backward": 0.3881}},
"F5": {"Pelvis": {"Left-Right": 0.0257,"Up-Down": -0.0103,"Forward-Backward": 0.5889},
"L_Ankle": {"Left-Right": 0.1057,"Up-Down": 0.0026,"Forward-Backward": 0.9873},
"R_Ankle": {"Left-Right": 0.0389,"Up-Down": -0.0002,"Forward-Backward": 0.2136},
"L_Wrist": {"Left-Right": -0.0247,"Up-Down": -0.0441,"Forward-Backward": 0.3528},
"R_Wrist": {"Left-Right": 0.0127,"Up-Down": 0.0401,"Forward-Backward": 0.7884}},
"F6": {"Pelvis": {"Left-Right": 0.0446,"Up-Down": 0.0119,"Forward-Backward": 0.4636},
"L_Ankle": {"Left-Right": -0.0068,"Up-Down": 0.0024,"Forward-Backward": 0.0991},
"R_Ankle": {"Left-Right": 0.0604, "Up-Down": -0.0237, "Forward-Backward": 0.792},
"L_Wrist": {"Left-Right": 0.083,"Up-Down": 0.0853,"Forward-Backward": 0.7153},
"R_Wrist": {"Left-Right": 0.071,"Up-Down": -0.0538,"Forward-Backward": 0.2695}},
"F7": {"Pelvis": {"Left-Right": 0.003,"Up-Down": 0.0036,"Forward-Backward": 0.1456},
"L_Ankle": {"Left-Right": 0.026,"Up-Down": -0.0207,"Forward-Backward": 0.197},
"R_Ankle": {"Left-Right": -0.0087,"Up-Down": -0.0009,"Forward-Backward": -0.0255},
"L_Wrist": {"Left-Right": -0.0331,"Up-Down": -0.0757,"Forward-Backward": -0.0229},
"R_Wrist": {"Left-Right": 0.0105,"Up-Down": 0.0312,"Forward-Backward": 0.2614}}}

Figure 7. Examples for LLM reasoning.

0.05, 0.5, 1}. The results are shown in Table 7. It can be observed that our method remains robust across a wide range of γ values, and we ultimately set $\gamma = 0.1$ as the final choice.

Table 7. Results of different smoothing parameter values.

γ	CLIP_S↑	VLM_S↑
$\gamma = 0.01$	$23.30_{\pm 0.79}$	$2.38_{\pm 0.32}$
$\gamma = 0.05$	$23.52_{\pm 0.52}$	$2.65_{\pm 0.27}$
$\gamma = 0.5$	$\mathbf{23.68_{\pm 0.78}}$	$2.51_{\pm 0.18}$
$\gamma = 1$	$23.66_{\pm 0.82}$	$2.70_{\pm 0.26}$
$\gamma = 0.1$ (Ours)	$23.64_{\pm 0.49}$	$\mathbf{2.72_{\pm 0.21}}$

9.4. Additional Visualization Results

We present additional qualitative results comparing the motions generated by our method with those of other baselines in Fig. 17 and Fig. 18. It can be observed that our approach achieves superior performance. For instance, given the description "place something on the ground", MDM generates motions with meaningless walking. MotionGPT shows the intention of placing an object, but fails to put it on the ground. MLD, MotionCLIP and AnySkill produce ambiguous motions that don't convey the description. In contrast, our method accurately demonstrates the complete process of placing an object on the ground and standing back up as described.

Furthermore, we present additional results of the G1 robot on the MuJoCo platform in Fig. 19, imitating motions generated by our method, which further demonstrates the practical applicability of our approach.

Example_1 Reason: [
'F0': (Initial Pose) All joints are in the default standing position.
'F1': (Preparation) The pelvis shifts slightly to the right (0.0173 m) and upward (0.0121 m), indicating the beginning of weight transfer. The left ankle moves slightly upward (0.0331 m) and forward (0.1046 m), suggesting the left leg is preparing for action. The right ankle shows minor movement, slightly moving backward and upward. The wrists also start to shift, with the left wrist moving upward and forward (0.3209 m), and the right wrist moving upward and forward (0.1716 m), suggesting preparation for the motion.
'F2': (Kicking Phase) The pelvis continues to move rightward (0.1443 m) and forward (0.1249 m), shifting downward slightly (-0.0518 m). This indicates a greater forward momentum, likely as the body moves into the action phase. The left ankle moves more significantly (0.2038 m), moving downward and forward. The right ankle shifts slightly backward (0.0735 m), which may contribute to maintaining balance. The wrists experience a pronounced change, with the right wrist moving upward and forward (0.4886 m), while the left wrist moves backward (0.3523 m), likely indicating the start of the force buildup for the kick.
'F3': (Peak of Kick) The pelvis shifts upward (0.0673 m) and forward (0.0448 m), likely preparing to kick. The right ankle shows a significant upward and forward motion (0.9342 m and 0.7092 m), indicating the leg has fully raised for the kick. The left ankle moves slightly upward. The wrists are adjusting to the change in body posture, with the right wrist moving downward and backward, while the left wrist moves upward slightly.
'F4': (Follow-Through) The pelvis shifts downward (-0.0429 m) while maintaining some forward momentum. The right ankle shows a noticeable shift downward and backward (-0.3097 m and 0.0877 m), indicating the force of the kick being released. The left ankle also moves downward, supporting the body's stabilization. The wrists continue adjusting as the right wrist moves downward and forward (0.1991 m), likely preparing to regain balance after the action.
'F5': (Follow-Through) The pelvis shifts backward (-0.0651 m) and downward (0.0312 m), indicating the motion is completing and the body is starting to recover. The right ankle moves significantly backward and downward (-0.4942 m and -0.7889 m), likely as the foot returns from the kick. The left ankle shifts backward slightly, providing support. The wrists show some movement, with the left wrist moving forward and upward (0.2698 m).
'F6': (Post-Kick Stabilization) The pelvis shifts further backward and downward (-0.1649 m, -0.0088 m), as the body moves to stabilize after the kick. The right ankle moves downward (0.1365 m), while the left ankle shows little movement. The wrists continue adjusting, with the right wrist moving backward and downward (-0.2011 m) and the left wrist moving upward (0.1424 m), likely indicating recovery from the kicking motion.
'F7': (Balance and Reset) The pelvis shifts forward slightly (0.0866 m) and upward, while the right ankle moves backward slightly. The left ankle shows minimal movement. The wrists exhibit subtle adjustments, with the left wrist moving downward slightly and the right wrist adjusting forward and backward, maintaining balance during the reset phase.]

Example_2 Reason: [
"F0": (Initial Pose) All joints are in the default standing position. The pelvis is centered with no vertical displacement. Both ankles are planted on the ground, shoulder-width apart. Wrists remain static at the sides of the body, reflecting a neutral posture before motion begins.
"F1": (Initial Weight Shift) The pelvis shifts minimally rightward (0.0 m) and forward (0.0 m), initiating weight transfer to the left leg. The left ankle shifts slightly left (-0.0 m) as it prepares to bear weight, while the right ankle moves right (0.00 m) in preparation for stepping. Both wrists begin natural arm swing: the left wrist moves forward (0.0041 m) with slight downward adjustment (-0.0022 m), and the right wrist moves forward (0.0047 m) and rightward (0.0033 m), synchronizing with leg opposition.
"F2": (Left Leg Lift) The pelvis shifts left (-0.0099 m) while progressing forward (0.0144 m), lowering slightly (-0.0007 m) as weight transfers to the right leg. The left ankle lifts upward (0.0067 m) and pulls back (-0.0112 m), signaling toe-off. The right ankle stabilizes with minor backward motion (-0.0063 m). Wrists exhibit early swing phase: the left wrist swings forward (0.0175 m), and the right wrist moves leftward (-0.0156 m) to counterbalance the leg lift.
"F3": (Left Leg Swing) The pelvis drops (-0.0307 m) and shifts right (0.007 m) over the supporting right leg, advancing significantly forward (0.1741 m). The left ankle swings forward (0.5713 m) and upward (0.0186 m) for stride extension. The right ankle anchors backward (-0.0158 m). Arms amplify swing: the right wrist drives forward (0.2049 m) with downward motion (-0.0334 m) opposing the left leg, while the left wrist moves forward moderately (0.0717 m).
"F4": (Left Foot Strike & Right Leg Swing) Pelvis rises (0.0201 m) and shifts right (0.0995 m) as the left foot lands (forward: 0.0359 m). The right ankle initiates swing, thrusting forward (0.9333 m) and rightward (0.079 m). Arms reach peak swing: the left wrist surges forward (0.6979 m) and up (0.0564 m), while the right wrist follows (0.3881 m forward) in sync with the right leg's propulsion.
"F5": (Right Leg Propulsion) Pelvis maintains forward momentum (0.5889 m) with slight rightward tilt (0.0257 m). The left ankle pushes off, advancing forward (0.9873 m) as weight shifts. The right ankle prepares for landing (0.2136 m forward). Arms begin retraction: the right wrist swings forward strongly (0.7884 m) while rising (0.0401 m), and the left wrist pulls back slightly (forward: 0.3528 m) with downward correction (-0.0441 m).
"F6": (Right Foot Strike) Pelvis stabilizes vertically (0.0119 m up) and continues forward (0.4636 m). The right ankle lands (0.792 m forward) with downward adjustment (-0.0237 m). The left ankle settles (0.0991 m forward) for support. Arms coordinate with stance: the left wrist swings forward (0.7153 m) and up (0.0853 m), while the right wrist lowers (-0.0538 m) during weight acceptance.
"F7": (Recovery & Transition) Pelvis reduces forward motion (0.1456 m) as the step cycle concludes. The left ankle prepares for the next step (0.197 m forward), while the right ankle stabilizes (-0.0255 m backward). Arms reset: the right wrist moves forward (0.2614 m) for balance, and the left wrist retracts backward (-0.0229 m) and down (-0.0757 m) to initiate the next swing phase.]

Figure 8. Related reasons for the given examples.

Please evaluate the alignment between a given generative motion clip and the corresponding text description ('**{DESCRIPTION}**'). The motion clip is represented as a sequence of frames **{INPUT_LENGTH}**. The motion only represents the character's actions, and is not concerned with the presence or absence of objects. Rating from 0 to 5.

Scoring Criteria:
0-1: The motion does not match the text.
2-3: Motion partially matches the text description.
4-5: Motion matches the text well.

Output Format:
Rating: [Your score]
Rationale: [A brief explanation for the rating, no more than 40 words]

(a) semantic similarity evaluation prompt

Please evaluate the motion naturalness based on a given generative motion clip and the corresponding text description ('**{DESCRIPTION}**'). The motion clip is represented as a sequence of frames **{INPUT_LENGTH}**. The motion only represents the character's actions, and is not concerned with the presence or absence of objects. Rating from 0 to 5.

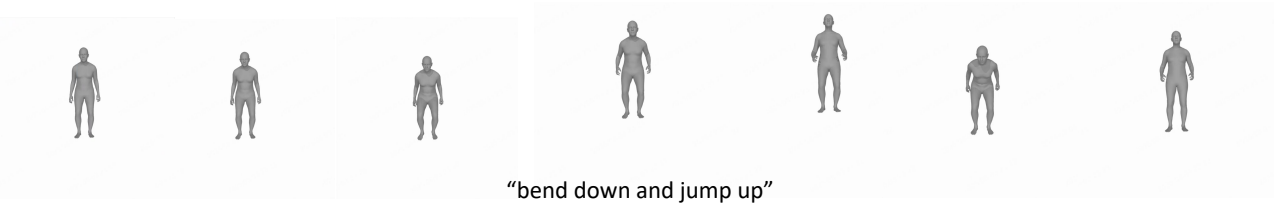
Scoring Criteria:
0-1: The motion exhibits unnatural behavior such as abnormal joint twisting, sudden teleportation, or complete stillness.
2-5: The motion is generally consistent with human movement patterns,you can evaluate based on fluency or smooth transitions.

Output Format:
Rating: [Your score]
Rationale: [A brief explanation for the rating, no more than 40 words]

(b) naturalness evaluation prompt

Figure 9. VLM evaluation prompt.

```
"F1": {
 "Pelvis": {
  "Left-Right": 0,
  "Up-Down": -0.03,
  "Forward-Backward": 0
 },
 "L_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0,
  "Forward-Backward": 0
 },
 "R_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0,
  "Forward-Backward": 0
 },
 "L_Wrist": {
  "Left-Right": 0,
  "Up-Down": -0.03,
  "Forward-Backward": -0.03
 },
 "R_Wrist": {
  "Left-Right": 0,
  "Up-Down": -0.03,
  "Forward-Backward": -0.03
 }
},
"F2": {
 "Pelvis": {
  "Left-Right": 0,
  "Up-Down": -0.05,
  "Forward-Backward": 0
 },
 "L_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0,
  "Forward-Backward": 0
 },
 "R_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0,
  "Forward-Backward": 0
 },
 "L_Wrist": {
  "Left-Right": 0,
  "Up-Down": -0.03,
  "Forward-Backward": -0.03
 },
 "R_Wrist": {
  "Left-Right": 0,
  "Up-Down": -0.03,
  "Forward-Backward": -0.03
 }
},
"F3": {
 "Pelvis": {
  "Left-Right": 0,
  "Up-Down": -0.07,
  "Forward-Backward": 0
 },
 "L_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0,
  "Forward-Backward": 0
 },
 "R_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0,
  "Forward-Backward": 0
 },
 "L_Wrist": {
  "Left-Right": 0,
  "Up-Down": -0.04,
  "Forward-Backward": -0.04
 },
 "R_Wrist": {
  "Left-Right": 0,
  "Up-Down": -0.04,
  "Forward-Backward": -0.04
 }
},
"F4": {
 "Pelvis": {
  "Left-Right": 0,
  "Up-Down": 0.28,
  "Forward-Backward": 0.05
 },
 "L_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0.26,
  "Forward-Backward": 0.05
 },
 "R_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0.26,
  "Forward-Backward": 0.05
 },
 "L_Wrist": {
  "Left-Right": 0,
  "Up-Down": 0.38,
  "Forward-Backward": 0.24
 },
 "R_Wrist": {
  "Left-Right": 0,
  "Up-Down": 0.38,
  "Forward-Backward": 0.24
 }
},
"F5": {
 "Pelvis": {
  "Left-Right": 0,
  "Up-Down": 0.12,
  "Forward-Backward": 0.02
 },
 "L_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0.1,
  "Forward-Backward": 0.02
 },
 "R_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0.1,
  "Forward-Backward": 0.02
 },
 "L_Wrist": {
  "Left-Right": 0,
  "Up-Down": 0.16,
  "Forward-Backward": 0.08
 },
 "R_Wrist": {
  "Left-Right": 0,
  "Up-Down": 0.16,
  "Forward-Backward": 0.08
 }
},
"F6": {
 "Pelvis": {
  "Left-Right": 0,
  "Up-Down": -0.4,
  "Forward-Backward": 0
 },
 "L_Ankle": {
  "Left-Right": 0,
  "Up-Down": -0.36,
  "Forward-Backward": 0
 },
 "R_Ankle": {
  "Left-Right": 0,
  "Up-Down": -0.36,
  "Forward-Backward": 0
 },
 "L_Wrist": {
  "Left-Right": 0,
  "Up-Down": -0.42,
  "Forward-Backward": -0.22
 },
 "R_Wrist": {
  "Left-Right": 0,
  "Up-Down": -0.42,
  "Forward-Backward": -0.22
 }
},
"F7": {
 "Pelvis": {
  "Left-Right": 0,
  "Up-Down": 0.15,
  "Forward-Backward": 0
 },
 "L_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0,
  "Forward-Backward": 0
 },
 "R_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0,
  "Forward-Backward": 0
 },
 "L_Wrist": {
  "Left-Right": 0,
  "Up-Down": 0.12,
  "Forward-Backward": 0.14
 },
 "R_Wrist": {
  "Left-Right": 0,
  "Up-Down": 0.12,
  "Forward-Backward": 0.14
 }
}
```

“bend down and jump up”

Figure 10. LLM-planned key joint positions and rendered images.


```
"F1": {
 "Pelvis": {
  "Left-Right": 0.01,
  "Up-Down": -0.15,
  "Forward-Backward": -0.05
 },
 "L_Ankle": {
  "Left-Right": -0.02,
  "Up-Down": 0.06,
  "Forward-Backward": 0.12
 },
 "R_Ankle": {
  "Left-Right": 0.02,
  "Up-Down": 0.06,
  "Forward-Backward": 0.12
 },
 "L_Wrist": {
  "Left-Right": -0.04,
  "Up-Down": 0.1,
  "Forward-Backward": 0.2
 },
 "R_Wrist": {
  "Left-Right": 0.04,
  "Up-Down": 0.1,
  "Forward-Backward": 0.2
 }
},
"F2": {
 "Pelvis": {
  "Left-Right": -0.02,
  "Up-Down": -0.18,
  "Forward-Backward": -0.08
 },
 "L_Ankle": {
  "Left-Right": -0.03,
  "Up-Down": 0.14,
  "Forward-Backward": 0.25
 },
 "R_Ankle": {
  "Left-Right": 0.03,
  "Up-Down": 0.14,
  "Forward-Backward": 0.25
 },
 "L_Wrist": {
  "Left-Right": -0.07,
  "Up-Down": 0.18,
  "Forward-Backward": 0.35
 },
 "R_Wrist": {
  "Left-Right": 0.07,
  "Up-Down": 0.18,
  "Forward-Backward": 0.35
 }
},
"F3": {
 "Pelvis": {
  "Left-Right": 0.01,
  "Up-Down": -0.07,
  "Forward-Backward": -0.03
 },
 "L_Ankle": {
  "Left-Right": -0.01,
  "Up-Down": -0.04,
  "Forward-Backward": 0.08
 },
 "R_Ankle": {
  "Left-Right": 0.01,
  "Up-Down": -0.04,
  "Forward-Backward": 0.08
 },
 "L_Wrist": {
  "Left-Right": -0.03,
  "Up-Down": -0.03,
  "Forward-Backward": 0.15
 },
 "R_Wrist": {
  "Left-Right": 0.03,
  "Up-Down": -0.03,
  "Forward-Backward": 0.15
 }
},
"F4": {
 "Pelvis": {
  "Left-Right": -0.01,
  "Up-Down": 0.03,
  "Forward-Backward": 0
 },
 "L_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0,
  "Forward-Backward": 0
 },
 "R_Ankle": {
  "Left-Right": 0,
  "Up-Down": -0.1,
  "Forward-Backward": 0.05
 },
 "L_Wrist": {
  "Left-Right": 0.08,
  "Up-Down": -0.05,
  "Forward-Backward": -0.04
 },
 "R_Wrist": {
  "Left-Right": -0.08,
  "Up-Down": 0.05,
  "Forward-Backward": -0.04
 }
},
"F5": {
 "Pelvis": {
  "Left-Right": 0,
  "Up-Down": -0.02,
  "Forward-Backward": 0
 },
 "L_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0,
  "Forward-Backward": 0
 },
 "R_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0.08,
  "Forward-Backward": -0.03
 },
 "L_Wrist": {
  "Left-Right": -0.06,
  "Up-Down": 0.04,
  "Forward-Backward": 0.03
 },
 "R_Wrist": {
  "Left-Right": 0.06,
  "Up-Down": -0.04,
  "Forward-Backward": 0.03
 }
},
"F6": {
 "Pelvis": {
  "Left-Right": 0.01,
  "Up-Down": 0.02,
  "Forward-Backward": 0
 },
 "L_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0,
  "Forward-Backward": 0
 },
 "R_Ankle": {
  "Left-Right": 0,
  "Up-Down": -0.09,
  "Forward-Backward": 0.04
 },
 "L_Wrist": {
  "Left-Right": -0.09,
  "Up-Down": 0.06,
  "Forward-Backward": -0.05
 },
 "R_Wrist": {
  "Left-Right": 0.09,
  "Up-Down": -0.06,
  "Forward-Backward": -0.05
 }
},
"F7": {
 "Pelvis": {
  "Left-Right": -0.01,
  "Up-Down": 0,
  "Forward-Backward": 0
 },
 "L_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0,
  "Forward-Backward": 0
 },
 "R_Ankle": {
  "Left-Right": 0,
  "Up-Down": 0.07,
  "Forward-Backward": -0.02
 },
 "L_Wrist": {
  "Left-Right": 0.07,
  "Up-Down": -0.05,
  "Forward-Backward": 0.04
 },
 "R_Wrist": {
  "Left-Right": -0.07,
  "Up-Down": 0.05,
  "Forward-Backward": 0.04
 }
}
```


“drive a car”

Figure 11. LLM-planned key joint positions and rendered images.

```
"F1": {
 "Pelvis": {
  "Left-Right": -0.015,
  "Up-Down": -0.008,
  "Forward-Backward": 0.005
 },
 "L_Ankle": {
  "Left-Right": -0.012,
  "Up-Down": 0.06,
  "Forward-Backward": 0.04
 },
 "R_Ankle": {
  "Left-Right": 0.002,
  "Up-Down": -0.002,
  "Forward-Backward": 0.0
 },
 "L_Wrist": {
  "Left-Right": -0.04,
  "Up-Down": -0.11,
  "Forward-Backward": 0.08
 },
 "R_Wrist": {
  "Left-Right": 0.04,
  "Up-Down": -0.11,
  "Forward-Backward": 0.08
 }
},
"F2": {
 "Pelvis": {
  "Left-Right": -0.025,
  "Up-Down": -0.012,
  "Forward-Backward": 0.007
 },
 "L_Ankle": {
  "Left-Right": -0.018,
  "Up-Down": 0.12,
  "Forward-Backward": 0.06
 },
 "R_Ankle": {
  "Left-Right": 0.003,
  "Up-Down": -0.001,
  "Forward-Backward": 0.001
 },
 "L_Wrist": {
  "Left-Right": -0.05,
  "Up-Down": -0.13,
  "Forward-Backward": 0.11
 },
 "R_Wrist": {
  "Left-Right": 0.05,
  "Up-Down": -0.13,
  "Forward-Backward": 0.11
 }
},
"F3": {
 "Pelvis": {
  "Left-Right": -0.018,
  "Up-Down": 0.015,
  "Forward-Backward": -0.003
 },
 "L_Ankle": {
  "Left-Right": -0.095,
  "Up-Down": 0.22,
  "Forward-Backward": 0.14
 },
 "R_Ankle": {
  "Left-Right": 0.004,
  "Up-Down": 0.001,
  "Forward-Backward": -0.002
 },
 "L_Wrist": {
  "Left-Right": -0.055,
  "Up-Down": -0.12,
  "Forward-Backward": 0.07
 },
 "R_Wrist": {
  "Left-Right": 0.055,
  "Up-Down": -0.12,
  "Forward-Backward": 0.07
 }
},
"F4": {
 "Pelvis": {
  "Left-Right": -0.012,
  "Up-Down": 0.018,
  "Forward-Backward": -0.008
 },
 "L_Ankle": {
  "Left-Right": -0.025,
  "Up-Down": 0.11,
  "Forward-Backward": -0.09
 },
 "R_Ankle": {
  "Left-Right": -0.002,
  "Up-Down": 0.003,
  "Forward-Backward": 0.004
 },
 "L_Wrist": {
  "Left-Right": -0.035,
  "Up-Down": 0.24,
  "Forward-Backward": -0.12
 },
 "R_Wrist": {
  "Left-Right": 0.035,
  "Up-Down": 0.24,
  "Forward-Backward": -0.12
 }
},
"F5": {
 "Pelvis": {
  "Left-Right": 0.008,
  "Up-Down": -0.004,
  "Forward-Backward": 0.002
 },
 "L_Ankle": {
  "Left-Right": 0.004,
  "Up-Down": -0.007,
  "Forward-Backward": 0.005
 },
 "R_Ankle": {
  "Left-Right": -0.003,
  "Up-Down": 0.002,
  "Forward-Backward": -0.003
 },
 "L_Wrist": {
  "Left-Right": 0.01,
  "Up-Down": 0.008,
  "Forward-Backward": -0.006
 },
 "R_Wrist": {
  "Left-Right": -0.01,
  "Up-Down": 0.008,
  "Forward-Backward": -0.006
 }
},
"F6": {
 "Pelvis": {
  "Left-Right": -0.003,
  "Up-Down": 0.002,
  "Forward-Backward": -0.001
 },
 "L_Ankle": {
  "Left-Right": -0.002,
  "Up-Down": 0.003,
  "Forward-Backward": -0.002
 },
 "R_Ankle": {
  "Left-Right": 0.001,
  "Up-Down": -0.001,
  "Forward-Backward": 0.001
 },
 "L_Wrist": {
  "Left-Right": -0.004,
  "Up-Down": -0.003,
  "Forward-Backward": 0.002
 },
 "R_Wrist": {
  "Left-Right": 0.004,
  "Up-Down": -0.003,
  "Forward-Backward": 0.002
 }
},
"F7": {
 "Pelvis": {
  "Left-Right": 0.006,
  "Up-Down": 0.003,
  "Forward-Backward": -0.002
 },
 "L_Ankle": {
  "Left-Right": 0.001,
  "Up-Down": -0.001,
  "Forward-Backward": 0.001
 },
 "R_Ankle": {
  "Left-Right": -0.002,
  "Up-Down": 0.002,
  "Forward-Backward": -0.001
 },
 "L_Wrist": {
  "Left-Right": -0.003,
  "Up-Down": 0.002,
  "Forward-Backward": -0.001
 },
 "R_Wrist": {
  "Left-Right": 0.003,
  "Up-Down": 0.002,
  "Forward-Backward": -0.001
 }
}
```


“hold left leg while standing with right one”

Figure 12. LLM-planned key joint positions and rendered images.


```
"F1": {
 "Pelvis": {
  "Left-Right": 0.0,
  "Up-Down": -0.005,
  "Forward-Backward": -0.01
 },
 "L_Ankle": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 },
 "R_Ankle": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 },
 "L_Wrist": {
  "Left-Right": 0.254,
  "Up-Down": 0.254,
  "Forward-Backward": 0.0
 },
 "R_Wrist": {
  "Left-Right": -0.254,
  "Up-Down": 0.254,
  "Forward-Backward": 0.0
 }
},
"F2": {
 "Pelvis": {
  "Left-Right": 0.0,
  "Up-Down": 0.005,
  "Forward-Backward": 0.01
 },
 "L_Ankle": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 },
 "R_Ankle": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 },
 "L_Wrist": {
  "Left-Right": 0.255,
  "Up-Down": 0.255,
  "Forward-Backward": 0.0
 },
 "R_Wrist": {
  "Left-Right": -0.255,
  "Up-Down": 0.255,
  "Forward-Backward": 0.0
 }
},
"F3": {
 "Pelvis": {
  "Left-Right": 0.0,
  "Up-Down": -0.05,
  "Forward-Backward": 0.1
 },
 "L_Ankle": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 },
 "R_Ankle": {
  "Left-Right": 0.0,
  "Up-Down": 0.15,
  "Forward-Backward": 0.3
 },
 "L_Wrist": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 },
 "R_Wrist": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 }
},
"F4": {
 "Pelvis": {
  "Left-Right": 0.0,
  "Up-Down": 0.05,
  "Forward-Backward": 0.1
 },
 "L_Ankle": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 },
 "R_Ankle": {
  "Left-Right": 0.0,
  "Up-Down": -0.15,
  "Forward-Backward": 0.3
 },
 "L_Wrist": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 },
 "R_Wrist": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 }
},
"F5": {
 "Pelvis": {
  "Left-Right": 0.0,
  "Up-Down": -0.17,
  "Forward-Backward": 0.05
 },
 "L_Ankle": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 },
 "R_Ankle": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 },
 "L_Wrist": {
  "Left-Right": -0.17,
  "Up-Down": -0.17,
  "Forward-Backward": 0.0
 },
 "R_Wrist": {
  "Left-Right": 0.17,
  "Up-Down": -0.17,
  "Forward-Backward": 0.0
 }
},
"F6": {
 "Pelvis": {
  "Left-Right": 0.0,
  "Up-Down": -0.17,
  "Forward-Backward": 0.05
 },
 "L_Ankle": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 },
 "R_Ankle": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 },
 "L_Wrist": {
  "Left-Right": -0.17,
  "Up-Down": -0.17,
  "Forward-Backward": 0.0
 },
 "R_Wrist": {
  "Left-Right": 0.17,
  "Up-Down": -0.17,
  "Forward-Backward": 0.0
 }
},
"F7": {
 "Pelvis": {
  "Left-Right": 0.0,
  "Up-Down": -0.19,
  "Forward-Backward": 0.0
 },
 "L_Ankle": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 },
 "R_Ankle": {
  "Left-Right": 0.0,
  "Up-Down": 0.0,
  "Forward-Backward": 0.0
 },
 "L_Wrist": {
  "Left-Right": -0.169,
  "Up-Down": -0.169,
  "Forward-Backward": 0.0
 },
 "R_Wrist": {
  "Left-Right": 0.169,
  "Up-Down": -0.169,
  "Forward-Backward": 0.0
 }
}
```


“raise T pose, step forward and lower both knees”

Figure 13. LLM-planned key joint positions and rendered images.

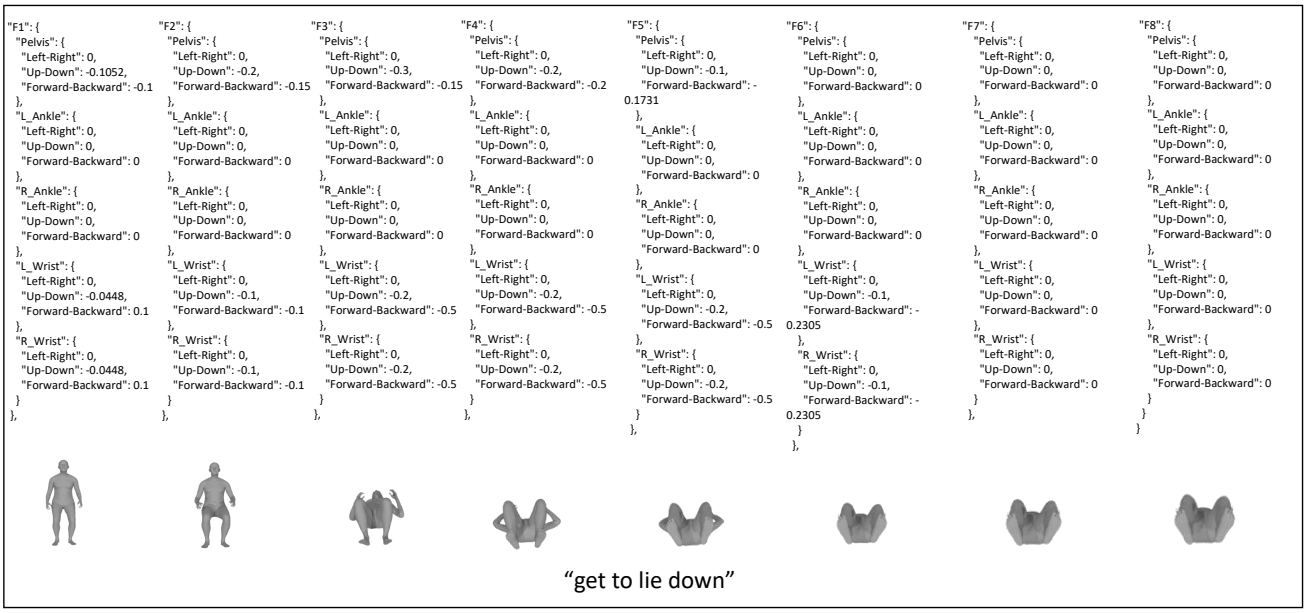


Figure 14. LLM-planned key joint positions and rendered images.

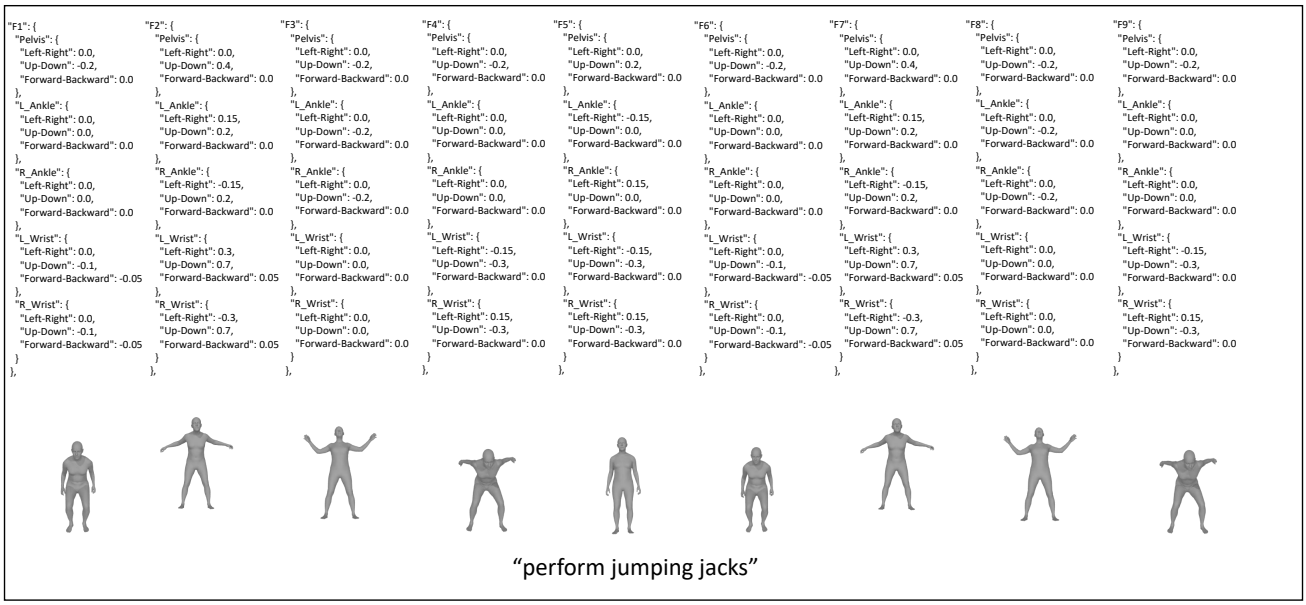


Figure 15. LLM-planned key joint positions and rendered images.

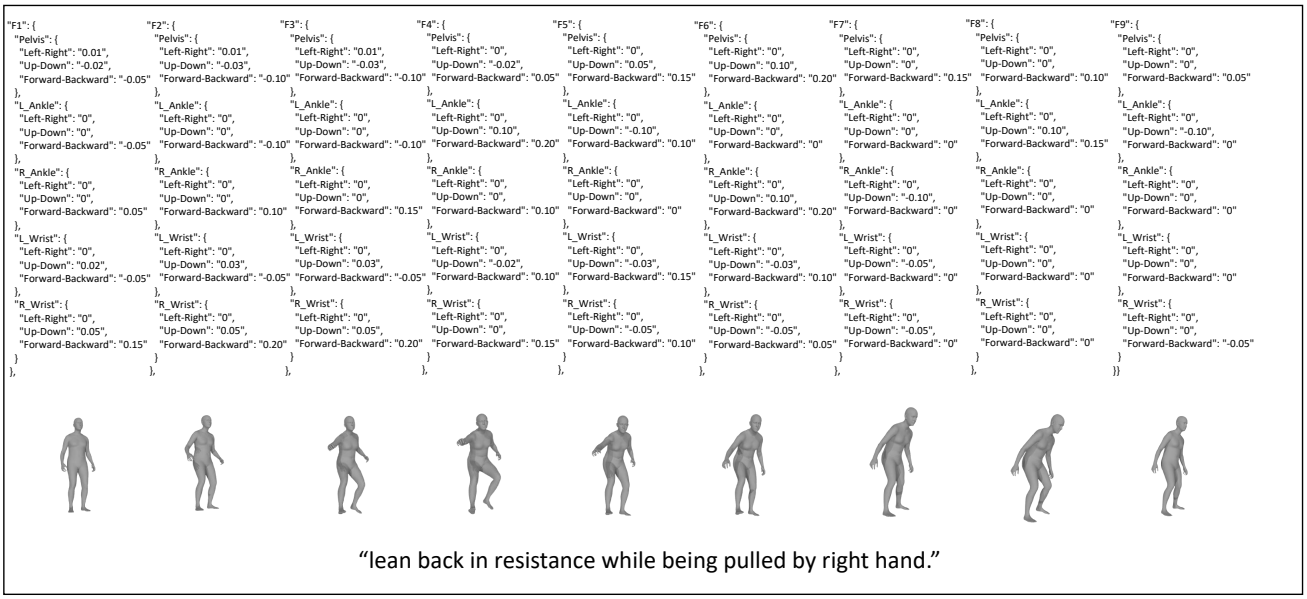


Figure 16. LLM-planned key joint positions and rendered images.

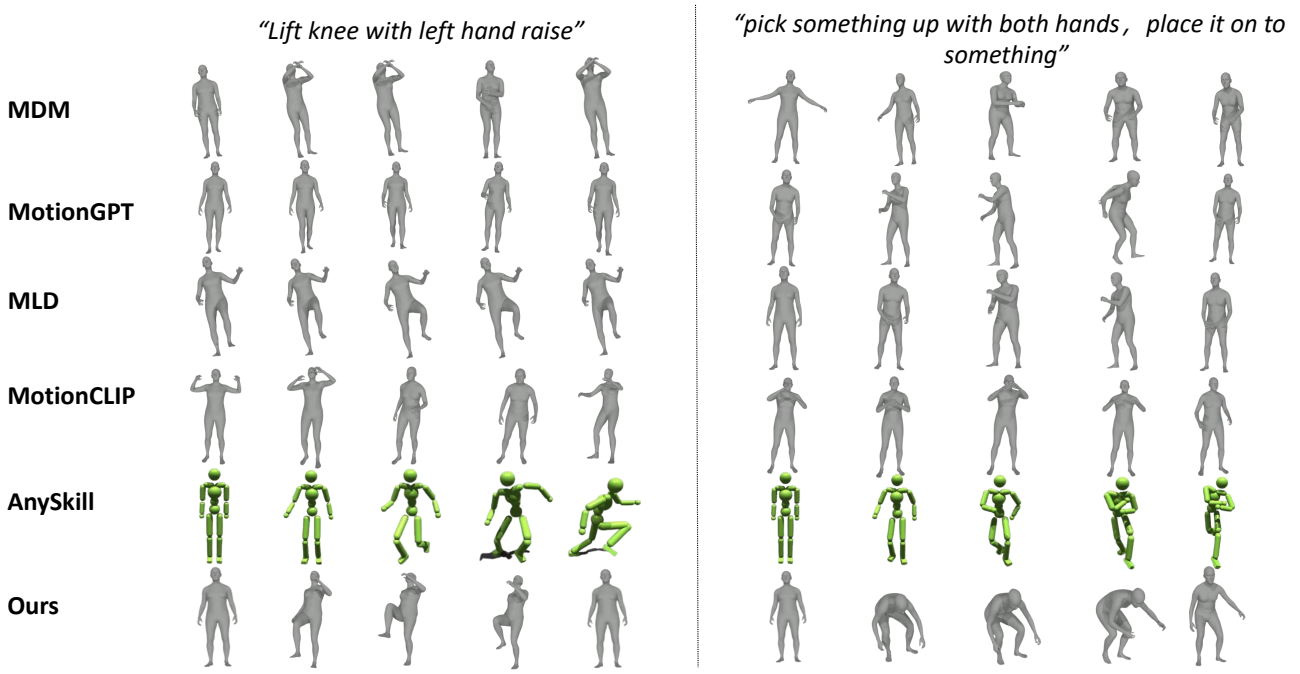


Figure 17. Additional qualitative comparison results of motions generated by different methods.

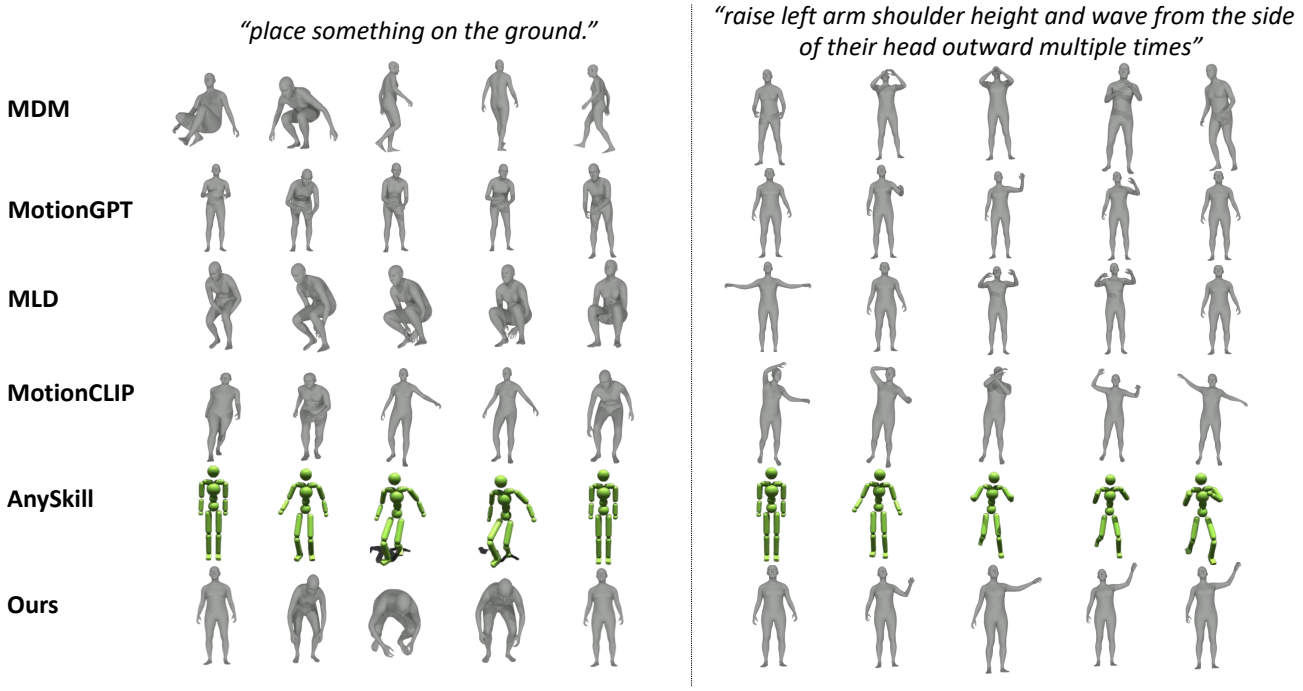
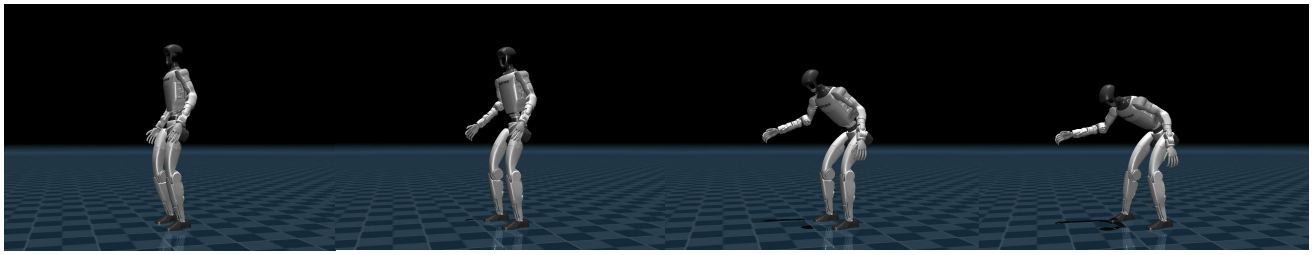
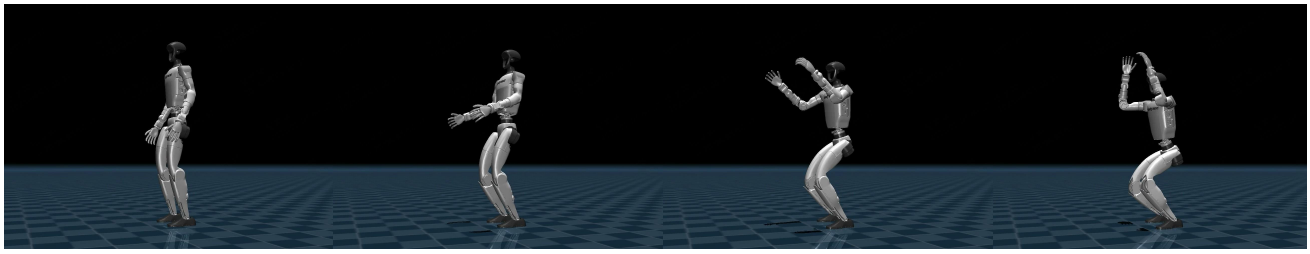


Figure 18. Additional qualitative comparison results of motions generated by different methods.



"lean back in resistance while being pulled by right hand"



"squat down and put hands above head"

Figure 19. Additional results on the MuJoCo platform.