DisCoRD: Discrete Tokens to Continuous Motion via Rectified Flow Decoding

Supplementary Material

This supplementary material is organized as follows: Section A details the implementation of DisCoRD. Section B provides additional information on the datasets and evaluation metrics. Section C offers a comprehensive analysis of sJPE. Section D presents quantitative results excluded from the main paper. Section E includes additional qualitative results. ***We highly recommend viewing the accompanying video***, as static images are insufficient to fully convey the intricacies of motion.

A. Implementation Details

Table A provides an overview of the implementation details for our method. These configurations were employed to train the DisCoRD decoder using the pretrained Momask [12] quantizer. Specifically, the 512-dimensional codebook embeddings from MoMask are projected into the conditional channel dimension. This projection is concatenated with Gaussian noise of the same dimensionality as the output channel. The concatenated representation is subsequently projected into the input channel dimension of the U-Net architecture. The U-Net processes this input and transforms it back into the output channel dimension, generating the final output shape. For training, we used an input window size of 64 and trained the model for 35 hours on a single NVIDIA RTX 4090 Ti GPU.

Training Details	
Optimizer	AdamW $(0.9, 0.999)$
LR	0.0005
LR Decay Ratio	0
LR Scheduler	Cosine
Warmup Epochs	20
Gradient Clipping	1.0
Weight EMA	0.999
Flow Loss	MSE Loss
Batch Size	768
Window Size	64
Steps	481896
Epochs	200
Model Details	
Input Channels	512
Output Channels	263
Condition Channels	256
Activation	SiLU
Dropout	0
Width	(512, 1024)
# Resnet / Block	2
# Params	66.9M

Table A. **Implementation details** for training the DisCoRD decoder on the HumanML3D dataset using the pretrained Momask quantizer.

B. Datasets and Evaluations

In this section, we provide additional explanations regarding the co-speech gesture generation and music-to-dance generation tasks that we were unable to describe in detail in the main paper.

B.1. Datasets.

For the co-speech gesture generation task, we utilized the SHOW dataset [65], a 3D holistic body dataset comprising 26.9 hours of in-the-wild talking videos. For the music-driven dance generation task, we used a mixed dataset combining AIST++ [26] and HumanML3D [11] where AIST++ is a large-scale 3D dance dataset created from multi-camera videos accompanied by music of varying styles and tempos, containing 992 high-quality pose sequences in the SMPL format.

B.2. Evaluations.

To evaluate the co-speech gesture generation task, we used Frechet Gesture Distance (FGD) [34], which measures the difference between the latent distributions of generated and real motions. Since our focus is on body movements, we reported body FGD, which quantifies differences specifically for the body part, for ProbTalk [31]. For TalkSHOW [65], which only utilizes holistic FGD—a metric that measures differences across the entire motion, including the face and hands—we reported the holistic FGD. To evaluate the music-to-dance generation task, we utilized $\mathrm{Dist_k}$, which quantifies the distributional spread of generated dances based on kinetic features, and $\mathrm{Dist_g}$, which does the same for geometric features, as proposed in [54]. A smaller difference between the distributions of the generated motion and the ground truth motion indicates that the $\mathrm{Dist_k}$ and $\mathrm{Dist_g}$ values of the generated motion align closely with those of the ground truth, reflecting a similar level of distributional spread.

C. Additional Analysis on sJPE

To evaluate the sample-wise naturalness of reconstructed motions, we introduce the symmetric Jerk Percentage Error (sJPE), as defined in Equation 5 of the main paper. We present detailed formulations of *Noise sJPE* and *Static sJPE*, supported by analysis using generated motion samples. Furthermore, qualitative comparisons highlight the effectiveness of DisCoRD against state-of-the-art discrete methods. Finally, we investigate the alignment of sJPE with human preference to validate its perceptual relevance.

C.1. Visualization of Fine-Grained Motion

To analyze fine-grained motion trajectories, we follow a three-step procedure. First, we select a joint for visualization, typically hand joints due to their high dynamism, and track their positional changes over time. Second, we apply a Gaussian filter to smooth the trajectory, reducing noise. Finally, we compute the difference between the smoothed and original trajectories to isolate fine-grained motion components. This method allows for detailed evaluation of frame-wise noise and under-reconstructed regions in motion trajectories. The visualizations in Figure 5 of the main paper and the qualitative samples in the supplementary material are generated using this process.

C.2. Details on sJPE.

Within the symmetric Jerk Percentage Error (sJPE), we define two components: Noise sJPE and Static sJPE. These isolate the instances where the predicted jerk overestimates or underestimates the ground truth jerk, respectively.

Noise sJPE and Static sJPE. *Noise sJPE* captures the average overestimation of jerk in the predicted motion signal, meaning frame-wise noise, corresponding to cases where $J_{\text{pred},t} > J_{\text{true},t}$. It is defined as:

$$\text{Noise sJPE} = \frac{1}{n}\sum_{t=1}^{n} \frac{\max\left(0, J_{\text{pred},t} - J_{\text{true},t}\right)}{|J_{\text{true},t}| + |J_{\text{pred},t}|}. \tag{6}$$

The operator $\max(0, x)$ ensures that only positive differences contribute to Noise sJPE, separating overestimations from underestimations.

Noise sJPE can be seen on the red box of Figure A. Time steps where motion trajectory is noisy compared to ground truth motion show bigger jerk. The area under the predicted jerk and above the ground truth jerk, shown in blue area, is proportional to the Noise sJPE, meaning frame-wise noise.

Static sJPE measures the average underestimation of jerk, meaning lack of dynamism in the predicted motion, corresponding to cases where $J_{\text{pred},t} \leq J_{\text{true},t}$. It is defined as:

$$\text{Static sJPE} = \frac{1}{n}\sum_{t=1}^{n} \frac{\max\left(0, J_{\text{true},t} - J_{\text{pred},t}\right)}{|J_{\text{true},t}| + |J_{\text{pred},t}|}. \tag{7}$$

Static sJPE can be seen on the green box of Figure A. Time steps where motion trajectory is under-reconstructed compared to ground truth motion show smaller jerk. The area above the predicted jerk and under the ground truth jerk, shown in red area, is proportional to the Static sJPE, meaning under reconstructed motions.

The overall sJPE can be expressed as the sum of *Noise sJPE* and *Static sJPE*:

$$\text{sJPE} = \text{Noise sJPE} + \text{Static sJPE}. \tag{8}$$

These formulations provide a measure of prediction accuracy by separately accounting for the tendencies of the predictive model to overestimate or underestimate the true motion jerks.

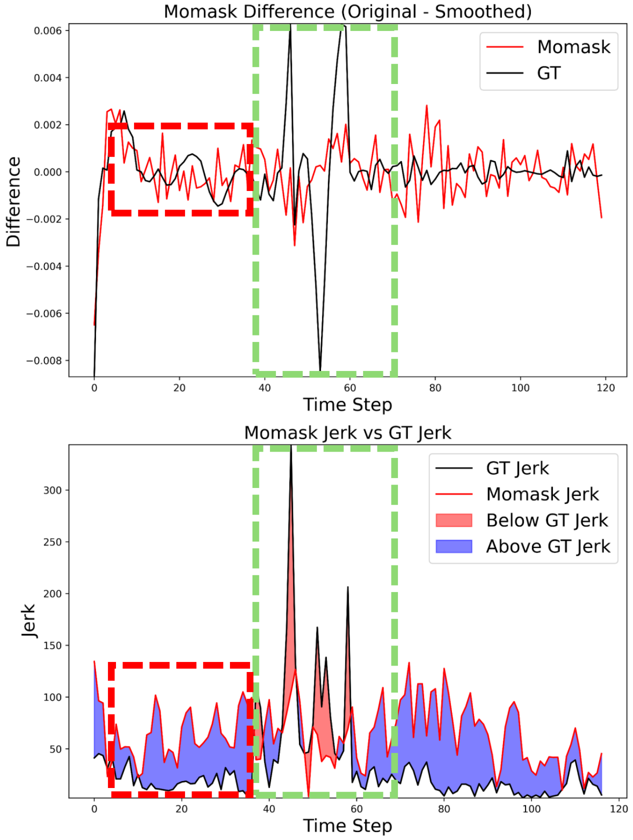


Figure A. **Relationship between fine-grained trajectory and jerk:** Frame-wise noise in predicted motions, highlighted in the red box, results in higher jerk values compared to the ground truth, represented by the blue areas. The sum of the blue areas corresponds to Noise sJPE. Conversely, under-reconstruction in predicted motions, highlighted in the green box, leads to lower jerk values compared to the ground truth, represented by the red areas. The sum of the red areas corresponds to Static sJPE.

C.3. Qualitative Results on Joint Trajectory and Jerk

We present a series of figures demonstrating the effectiveness of DisCoRD in reconstructing smooth and dynamic motion. For each sample, the first row visualizes the motion trajectory, while the second row plots the corresponding jerk at each time step, with the calculated sJPE displayed

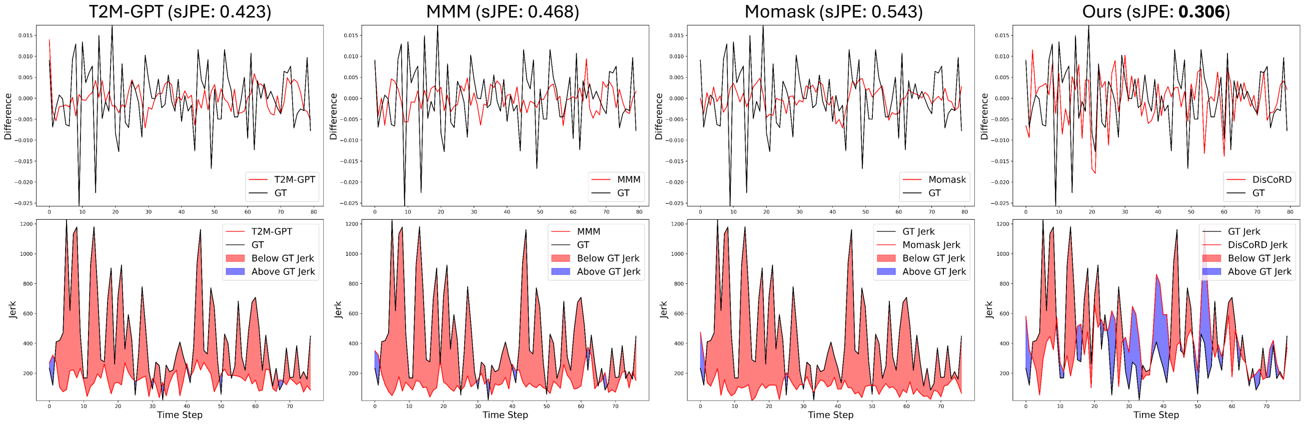


Figure B. **Joint Trajectory and Jerk: Under-Reconstruction in Discrete Methods** DisCoRD effectively reduces the red area, demonstrating its capability to reconstruct dynamic motion accurately. This improvement is also reflected in the lower sJPE value.

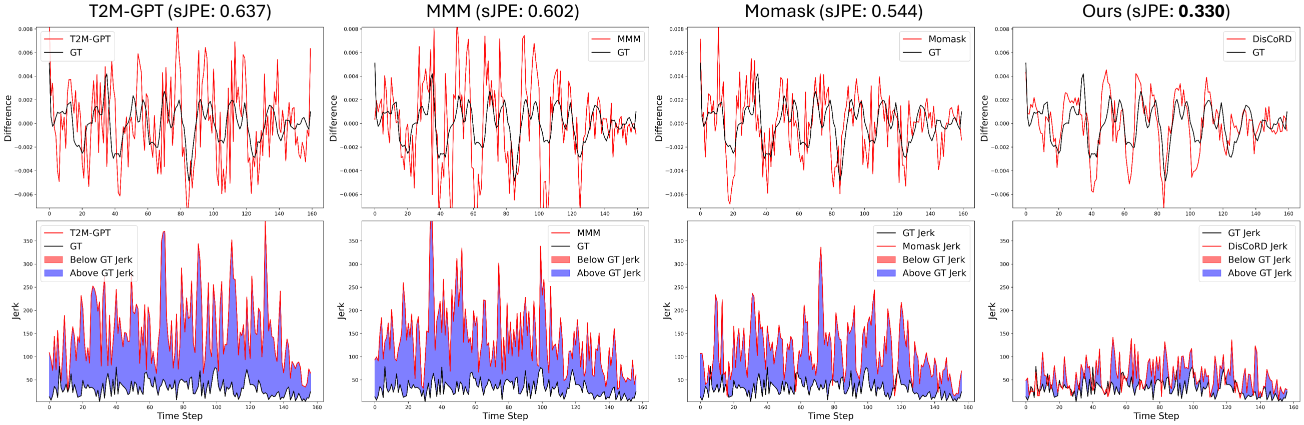


Figure C. **Joint Trajectory and Jerk: Frame-Wise Noise in Discrete Methods** DisCoRD significantly reduces the blue area, indicating its ability to generate smooth motions that closely resemble the ground truth. This improvement is further reflected in the lower sJPE value.

at the top. This visualization enables detailed analysis of fine-grained trajectories in predicted motions and highlights the contributions of Noise sJPE and Static sJPE to the overall sJPE.

We compare DisCoRD with recent discrete methods, including T2M-GPT [67], MMM [42], and Momask [12]. Motion samples that illustrate under-reconstruction in discrete methods are presented in Figure B, while those that exhibit frame-wise noise are shown in Figure C. Samples showing both issues in discrete models are displayed in Figure D. DisCoRD effectively reduces frame-wise noise while accurately reconstructing dynamic, fast-paced motions. This is shown in both the visualizations and the sJPE results.

C.4. Correlation between sJPE and human perception.

To further verify that sJPE aligns with human judgment of naturalness, we conducted an additional user study. We asked participants to rank three models—MLD, MoMask, and ours—in order of naturalness, guided as Figure J. The user interface for this user study is shown in Figure L. The rankings were scored such that the first place received 1 point, the second place 2 points, and the third place 3 points. Using these human scores, we calculated Pearson's correlation between the human scores and two metrics—MPJPE and sJPE— for each sample. During this process, we excluded the lowest 10% of samples in terms of human score standard deviation among models, as these were considered indistinguishable by human evaluators. Our analysis revealed that the average Pearson's correlation between MPJPE and human scores was 0.181, whereas the correlation between sJPE and human scores was significantly higher at 0.483. This result demonstrates the effectiveness of sJPE in evaluating sample-wise naturalness.

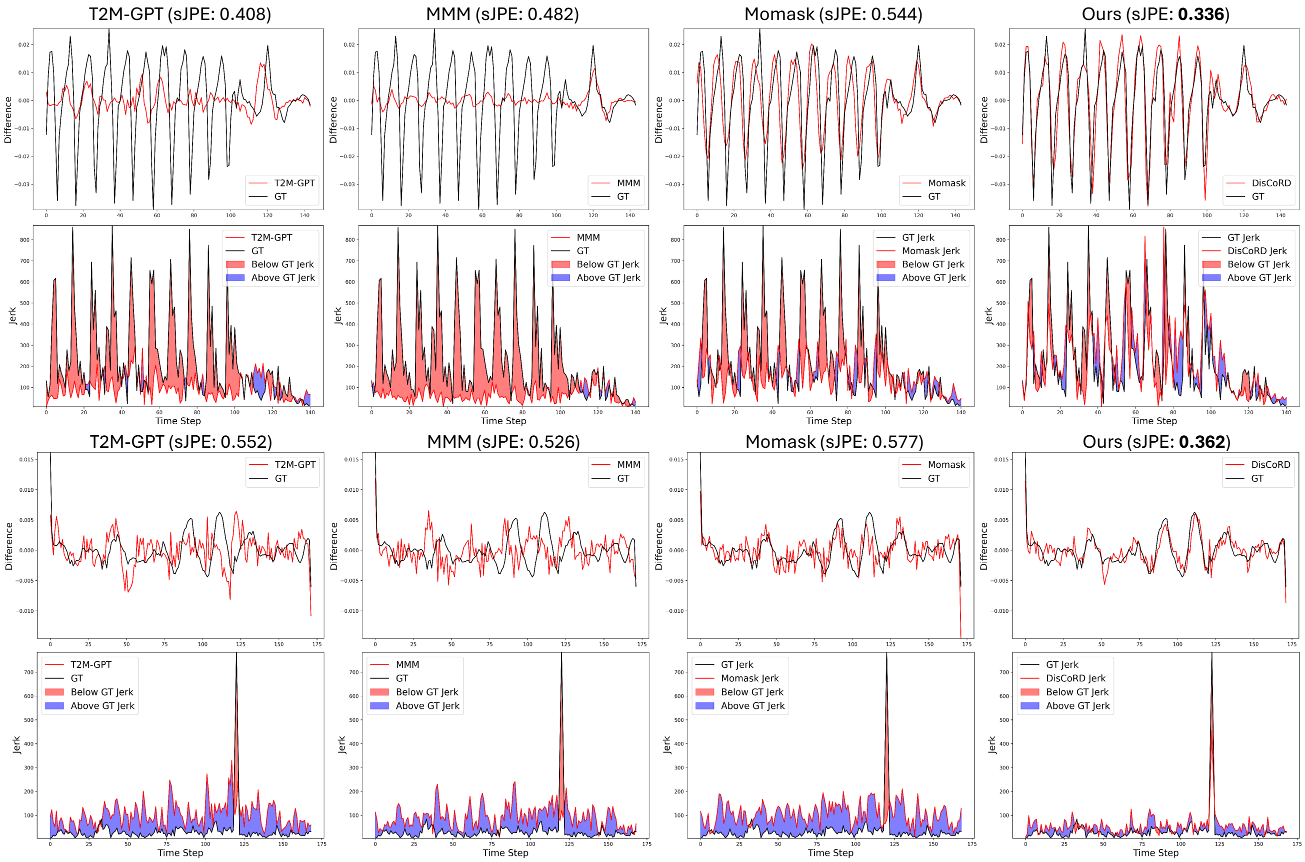


Figure D. **Joint Trajectory and Jerk: Both Frame-Wise Noise and Under-Reconstruction in Discrete Methods** DisCoRD addresses both frame-wise noise and under-reconstruction by simultaneously reducing the blue and red areas. This demonstrates its ability to generate smooth and dynamic motions, closely aligning with the ground truth. This further supported by the lower sJPE values.

D. Additional Quantitative Results

D.1. Performance on Text-to-Motion Generation.

In Table B, we present a comparison of our method against additional results from various text-to-motion models. Our method consistently achieves strong performance on the HumanML3D and KIT-ML [43] test sets, even when evaluated alongside these additional models. While ReMoDiffuse achieves particularly strong performance on KIT-ML, it is worth noting that its performance benefits from the use of a specialized database for high-quality motion generation, which makes direct comparisons less appropriate.

D.2. Performance on Various Tasks.

In Table C, we present additional evaluation results for co-speech gesture generation. Following [65], we additionally report Diversity, which measures the variance among multiple samples generated from the same condition, and Beat Consistency (BC), which evaluates the synchronization between the generated motion and the corresponding audio. In Table D, we provide additional evaluation results for music-to-dance generation. Following [47], we report FID_k to measure differences in kinetic motion features and FID_g for geometric motion features. Additionally, we include the Beat Align Score (BAS) to assess the synchronization between motion and music. While [54] has shown that these metrics are not fully reliable and often fail to align with actual output quality, we include them to follow established conventions.

E. Additional Qualitative Results

E.1. Motion visualization.

In Figure F and Figure G, we present qualitative comparisons between our model and other leading approaches. In Figure H, we additionally display more qualitative results of our method. We observed that our method effectively follows the text prompts while maintaining naturalness in the generated outputs. Again, we highly recommend viewing the accompanying video, as static images are insufficient to fully convey the intricacies of motion.

Datasets	Methods	R Precision ↑			FID ↓	MultiModal Dist ↓	MultiModality ↑
		Top 1	Top 2	Top 3			
Human ML3D	MDM [53]	-	-	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$\underline{2.799}^{\pm.072}$
	MLD [6]	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$2.413^{\pm.079}$
	MotionDiffuse [68]	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$1.553^{\pm.042}$
	ReMoDiffuse [69]	$0.510^{\pm.005}$	$0.698^{\pm.006}$	$0.795^{\pm.004}$	$0.103^{\pm.004}$	$2.974^{\pm.016}$	$1.795^{\pm.043}$
	Fg-T2M [59]	$0.492^{\pm.002}$	$0.683^{\pm.003}$	$0.783^{\pm.024}$	$0.243^{\pm.019}$	$3.109^{\pm.007}$	$1.614^{\pm.049}$
	M2DM [22]	$0.497^{\pm.003}$	$0.682^{\pm.002}$	$0.763^{\pm.003}$	$0.352^{\pm.005}$	$3.134^{\pm.010}$	$\mathbf{3.587}^{\pm.072}$
	M2D2M [7]	-	-	$0.799^{\pm.002}$	$0.087^{\pm.004}$	$3.018^{\pm.010}$	$2.115^{\pm.079}$
	MotionGPT [70]	$0.364^{\pm.005}$	$0.533^{\pm.003}$	$0.629^{\pm.004}$	$0.805^{\pm.002}$	$3.914^{\pm.013}$	$2.473^{\pm.041}$
	MotionLLM [62]	$0.482^{\pm.004}$	$0.672^{\pm.003}$	$0.770^{\pm.002}$	$0.491^{\pm.019}$	$3.138^{\pm.010}$	-
	MotionGPT-2 [60]	$0.496^{\pm.002}$	$0.691^{\pm.003}$	$0.782^{\pm.004}$	$0.191^{\pm.004}$	$3.080^{\pm.013}$	$2.137^{\pm.022}$
	AttT2M [72]	$0.499^{\pm.003}$	$0.690^{\pm.002}$	$0.786^{\pm.002}$	$0.112^{\pm.006}$	$3.038^{\pm.007}$	$2.452^{\pm.051}$
	MMM [42]	$0.504^{\pm.003}$	$0.696^{\pm.003}$	$0.794^{\pm.002}$	$0.080^{\pm.003}$	$2.998^{\pm.007}$	$1.164^{\pm.041}$
	T2M-GPT [67]	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$3.118^{\pm.011}$	$1.856^{\pm.011}$
	+ DisCoRD (Ours)	$0.476^{\pm.008}$	$0.663^{\pm.006}$	$0.760^{\pm.007}$	$0.095^{\pm.011}$	$3.121^{\pm.009}$	$1.831^{\pm.048}$
	BAMM [41]	$\mathbf{0.525}^{\pm.002}$	$\mathbf{0.720}^{\pm.003}$	$\mathbf{0.814}^{\pm.003}$	$0.055^{\pm.002}$	$\mathbf{2.919}^{\pm.008}$	$1.687^{\pm.051}$
	+ DisCoRD (Ours)	$0.522^{\pm.003}$	$\underline{0.715}^{\pm.005}$	$\underline{0.811}^{\pm.004}$	$\underline{0.041}^{\pm.002}$	$\underline{2.921}^{\pm.015}$	$1.772^{\pm.067}$
	MoMask [12]	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$0.045^{\pm.002}$	$2.958^{\pm.008}$	$1.241^{\pm.040}$
	+ DisCoRD (Ours)	$\underline{0.524}^{\pm.003}$	$\underline{0.715}^{\pm.003}$	$0.809^{\pm.002}$	$\mathbf{0.032}^{\pm.002}$	$2.938^{\pm.010}$	$1.288^{\pm.043}$
KIT-ML	MDM [53]	-	-	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.191^{\pm.022}$	$1.907^{\pm.214}$
	MLD [6]	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$0.404^{\pm.027}$	$3.204^{\pm.027}$	$2.192^{\pm.071}$
	MotionDiffuse [68]	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$1.954^{\pm.062}$	$2.958^{\pm.005}$	$0.730^{\pm.013}$
	ReMoDiffuse [69]	$0.427^{\pm.014}$	$0.641^{\pm.004}$	$0.765^{\pm.055}$	$\mathbf{0.155}^{\pm.006}$	$2.814^{\pm.012}$	$1.239^{\pm.028}$
	Fg-T2M [59]	$0.418^{\pm.005}$	$0.626^{\pm.004}$	$0.745^{\pm.004}$	$0.571^{\pm.047}$	$3.114^{\pm.015}$	$1.019^{\pm.029}$
	M2DM [22]	$0.416^{\pm.004}$	$0.628^{\pm.004}$	$0.743^{\pm.004}$	$0.515^{\pm.029}$	$3.015^{\pm.017}$	$\mathbf{3.325}^{\pm.037}$
	M2D2M [7]	-	-	$0.753^{\pm.006}$	$0.378^{\pm.023}$	$3.012^{\pm.021}$	$2.061^{\pm.067}$
	MotionGPT [70]	$0.340^{\pm.002}$	$0.570^{\pm.003}$	$0.660^{\pm.004}$	$0.868^{\pm.032}$	$3.721^{\pm.018}$	$2.296^{\pm.022}$
	MotionLLM [62]	$0.409^{\pm.006}$	$0.624^{\pm.007}$	$0.750^{\pm.005}$	$0.781^{\pm.026}$	$2.982^{\pm.022}$	-
	MotionGPT-2 [60]	$0.427^{\pm.003}$	$0.627^{\pm.002}$	$0.764^{\pm.003}$	$0.614^{\pm.005}$	$3.164^{\pm.013}$	$\underline{2.357}^{\pm.022}$
	AttT2M [72]	$0.413^{\pm.006}$	$0.632^{\pm.006}$	$0.751^{\pm.006}$	$0.870^{\pm.039}$	$3.039^{\pm.021}$	$2.281^{\pm.047}$
	MMM [42]	$0.404^{\pm.005}$	$0.621^{\pm.006}$	$0.744^{\pm.005}$	$0.316^{\pm.019}$	$2.977^{\pm.019}$	$1.232^{\pm.026}$
	T2M-GPT [67]	$0.398^{\pm.007}$	$0.606^{\pm.006}$	$0.729^{\pm.005}$	$0.718^{\pm.038}$	$3.076^{\pm.028}$	$1.887^{\pm.050}$
	+ DisCoRD (Ours)	$0.382^{\pm.007}$	$0.590^{\pm.007}$	$0.715^{\pm.004}$	$0.541^{\pm.038}$	$3.260^{\pm.028}$	$1.928^{\pm.059}$
	MoMask [12]	$\underline{0.433}^{\pm.007}$	$\underline{0.656}^{\pm.005}$	$\mathbf{0.781}^{\pm.005}$	$0.204^{\pm.011}$	$\mathbf{2.779}^{\pm.022}$	$1.131^{\pm.043}$
	+ DisCoRD (Ours)	$\mathbf{0.434}^{\pm.007}$	$\mathbf{0.657}^{\pm.005}$	$\underline{0.775}^{\pm.004}$	$\underline{0.169}^{\pm.010}$	$\underline{2.792}^{\pm.015}$	$1.266^{\pm.046}$

Table B. **Additional quantitative evaluation** on the HumanML3D and KIT-ML test sets. ± indicates a 95% confidence interval. +DisCoRD indicates that the baseline model's decoder is replaced with our DisCoRD decoder. **Bold** indicates the best result, while underscore refers the second best.

Methods	Diversity ↑	BC → (0.868)
TalkSHOW [65]	0.821	**0.872**
+DisCoRD(Ours)	**0.919**	0.876
ProbTalk [31]	0.259	0.795
+DisCoRD(Ours)	**0.331**	**0.866**

Table C. **Additional quantitative results** on each method's SHOW test set. The results demonstrate that our method performs on par with, or surpasses, the baseline models.

E.2. User preference study details.

We conduct two user studies to (1) validate our motivation and method effectiveness and (2) evaluate how well sJPE aligns with human perception. The first study, shown in Figure E, indicates that the discrete model Momask outperforms the continuous model MDM in faithfulness but lags in naturalness. In contrast, DisCoRD surpasses both, demonstrating its ability to generate motion that is both natural and faithful. In the second study, we find that sJPE exhibits 2.7 times higher correlation with human preference for naturalness compared to MPJPE, highlighting its effectiveness in evaluating sample-wise motion naturalness. Participants were guided to evaluate both faithfulness and naturalness, as shown in Figure I. Given two motion videos generated by two different models on the same prompt, participants were asked to choose a better one in terms of faithfulness and naturalness, as shown in Figure K. Total 41 participants participated in this user study.

Methods	FID$_k$ ↓	FID$_g$ ↓	BAS ↑
Ground Truth	17.10	10.60	0.2374
TM2D [10]	**19.01**	**20.09**	0.2049
+DisCoRD(Ours)	23.98	88.74	**0.2190**

Table D. **Additional quantitative results** on the AIST++ test set. The results demonstrate that, although our method shows performance degradation on FID$_k$ and FID$_g$, which are known to be unreliable, it achieves improvement in the Beat Align Score.

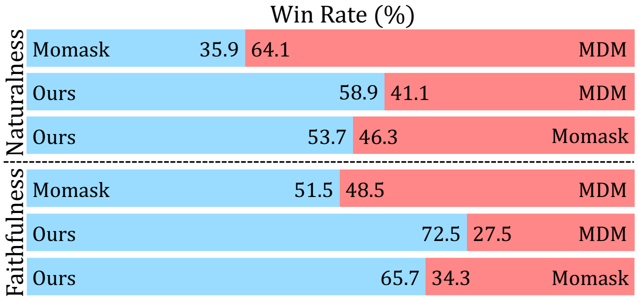


Figure E. **User study results on the HumanML3D dataset.** Each bar represents a comparison between two models, with win rates depicted in blue and loss rates in red, evaluated based on naturalness and faithfulness.

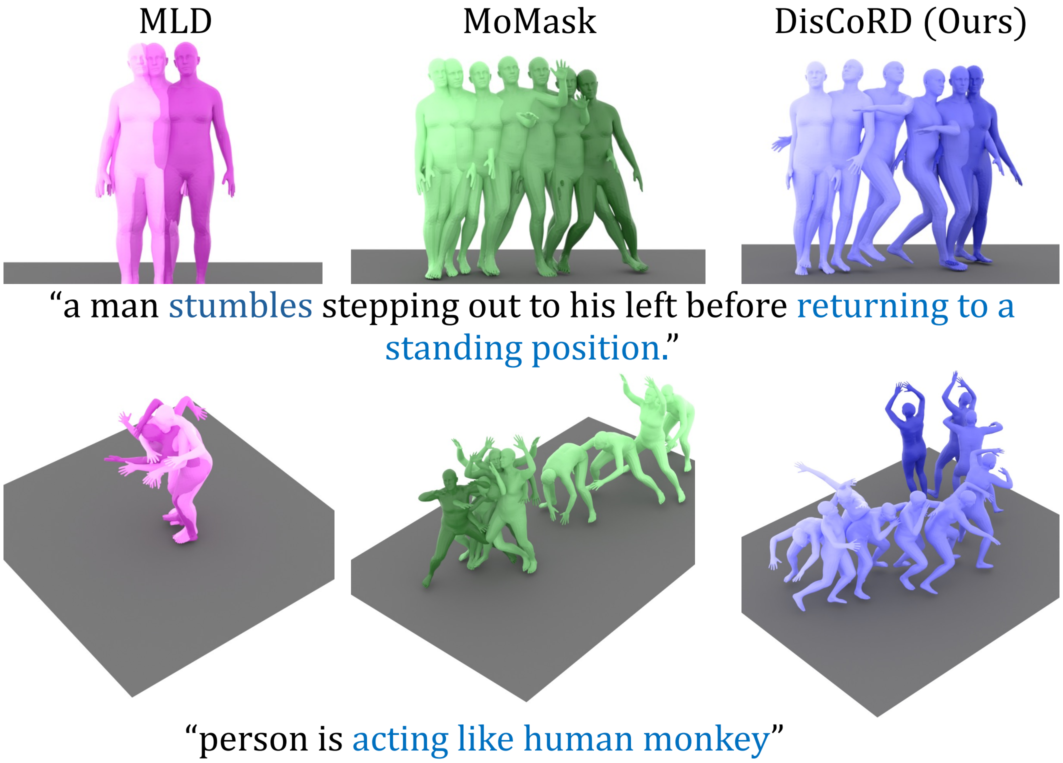
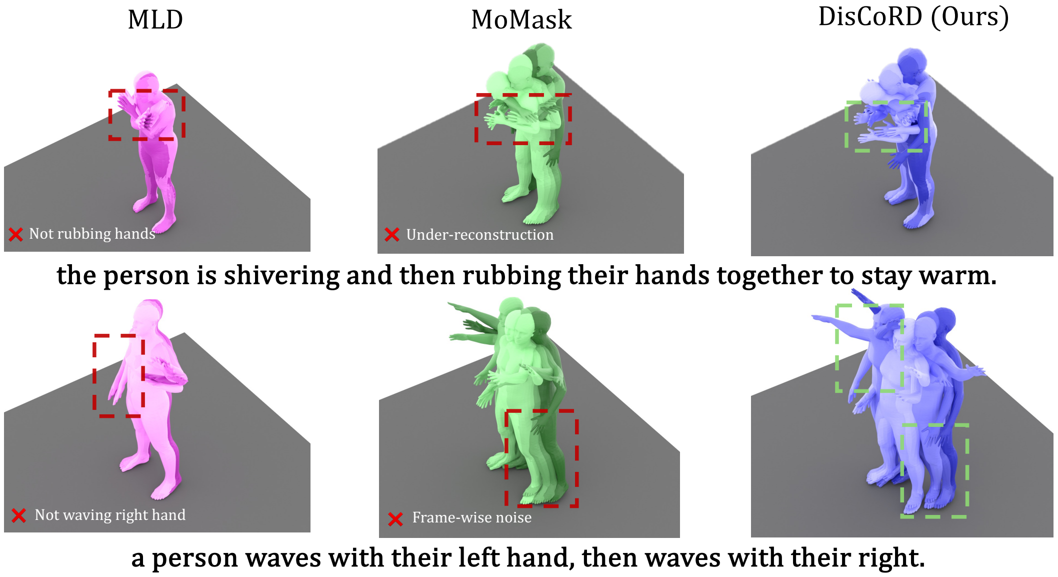

MLD
MoMask
DisCoRD (Ours)
“a man stumbles stepping out to his left before returning to a standing position.”
“person is acting like human monkey”

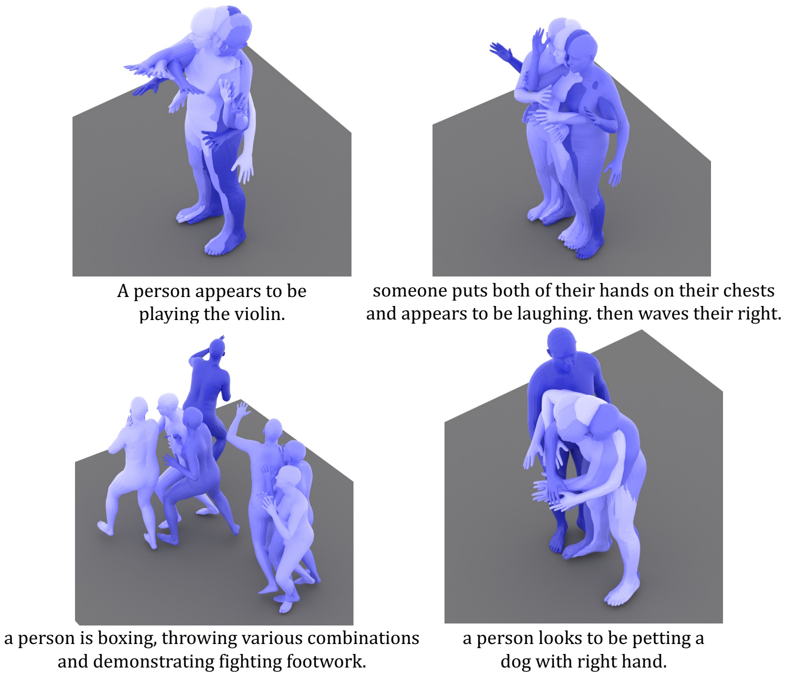
Figure F. **Qualitative comparisons** on the test set of HumanML3D.



the person is shivering and then rubbing their hands together to stay warm.

a person waves with their left hand, then waves with their right.

Figure G. **Additional qualitative comparisons** on the HumanML3D test set. The continuous method, MLD, often fails to perfectly align with the text consistently, while the discrete method, MoMask, exhibits issues such as under-reconstruction, resulting in minimal hand movement, or unnatural leg jitter caused by frame-wise noise.



A person appears to be playing the violin.

someone puts both of their hands on their chests and appears to be laughing. then waves their right.

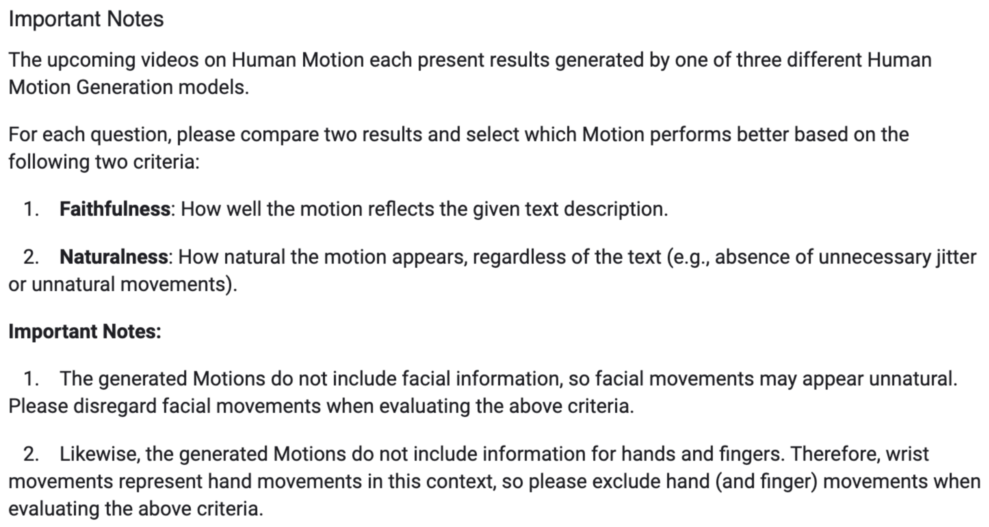
a person is boxing, throwing various combinations and demonstrating fighting footwork.

a person looks to be petting a dog with right hand.

Figure H. **Additional qualitative results** of our method on the HumanML3D test set.

Method	Sampling Steps			
	2	16	50	100
DDPM (DDIM sample)	0.055 / 0.007s	0.044 / 0.087s	0.038 / 0.331s	0.035 / 0.689s
Linear SDE	8.998 / 0.024s	0.047 / 0.157s	0.033 / 0.472s	0.032 / 0.955s
Ours	0.034 / 0.016s	**0.032 / 0.221s**	0.032 / 0.703s	0.032 / 1.426s

Table E. FID and decoding time (s) for different sampling steps (↓). The original **MoMask has a decoding time of 0.244 seconds**. We trained a diffusion decoder (DDPM) using the same architecture as our rectified flow decoder for a fair comparison. For the Linear SDE variant, we replaced only the sampler in our model with the Euler-Maruyama sampler, keeping all other components identical.



Important Notes

The upcoming videos on Human Motion each present results generated by one of three different Human Motion Generation models.

For each question, please compare two results and select which Motion performs better based on the following two criteria:

1. **Faithfulness**: How well the motion reflects the given text description.

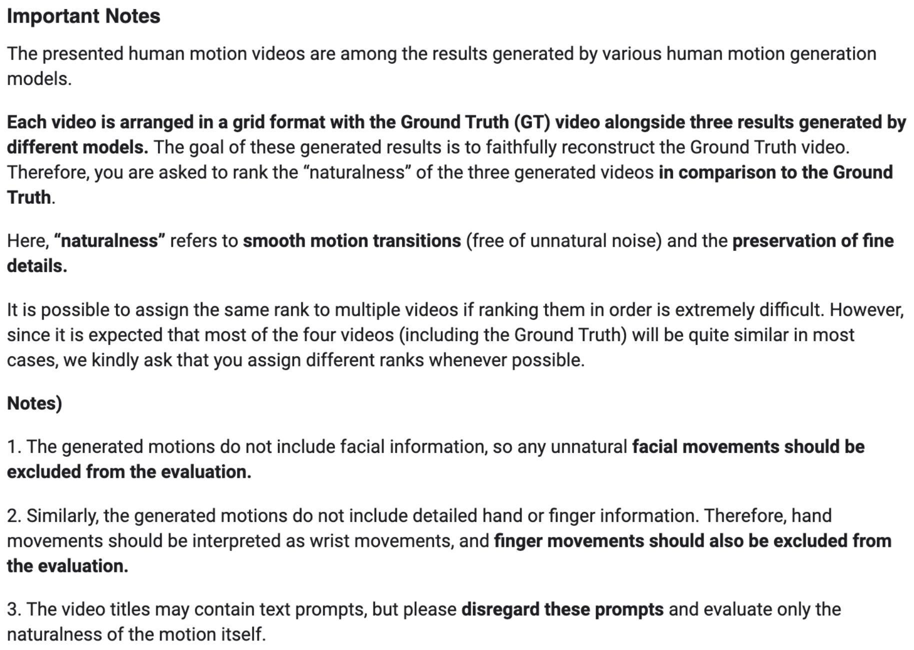
2. **Naturalness**: How natural the motion appears, regardless of the text (e.g., absence of unnecessary jitter or unnatural movements).

Important Notes:

1. The generated Motions do not include facial information, so facial movements may appear unnatural. Please disregard facial movements when evaluating the above criteria.

2. Likewise, the generated Motions do not include information for hands and fingers. Therefore, wrist movements represent hand movements in this context, so please exclude hand (and finger) movements when evaluating the above criteria.

Figure I. **Guidelines for user study in the Main paper:** participants were asked to evaluate Faithfulness and Naturalness, excluding hand and facial movements that are not included in HumanML3D.



Important Notes

The presented human motion videos are among the results generated by various human motion generation models.

Each video is arranged in a grid format with the Ground Truth (GT) video alongside three results generated by different models. The goal of these generated results is to faithfully reconstruct the Ground Truth video. Therefore, you are asked to rank the “naturalness” of the three generated videos **in comparison to the Ground Truth**.

Here, **“naturalness”** refers to **smooth motion transitions** (free of unnatural noise) and the **preservation of fine details.**

It is possible to assign the same rank to multiple videos if ranking them in order is extremely difficult. However, since it is expected that most of the four videos (including the Ground Truth) will be quite similar in most cases, we kindly ask that you assign different ranks whenever possible.

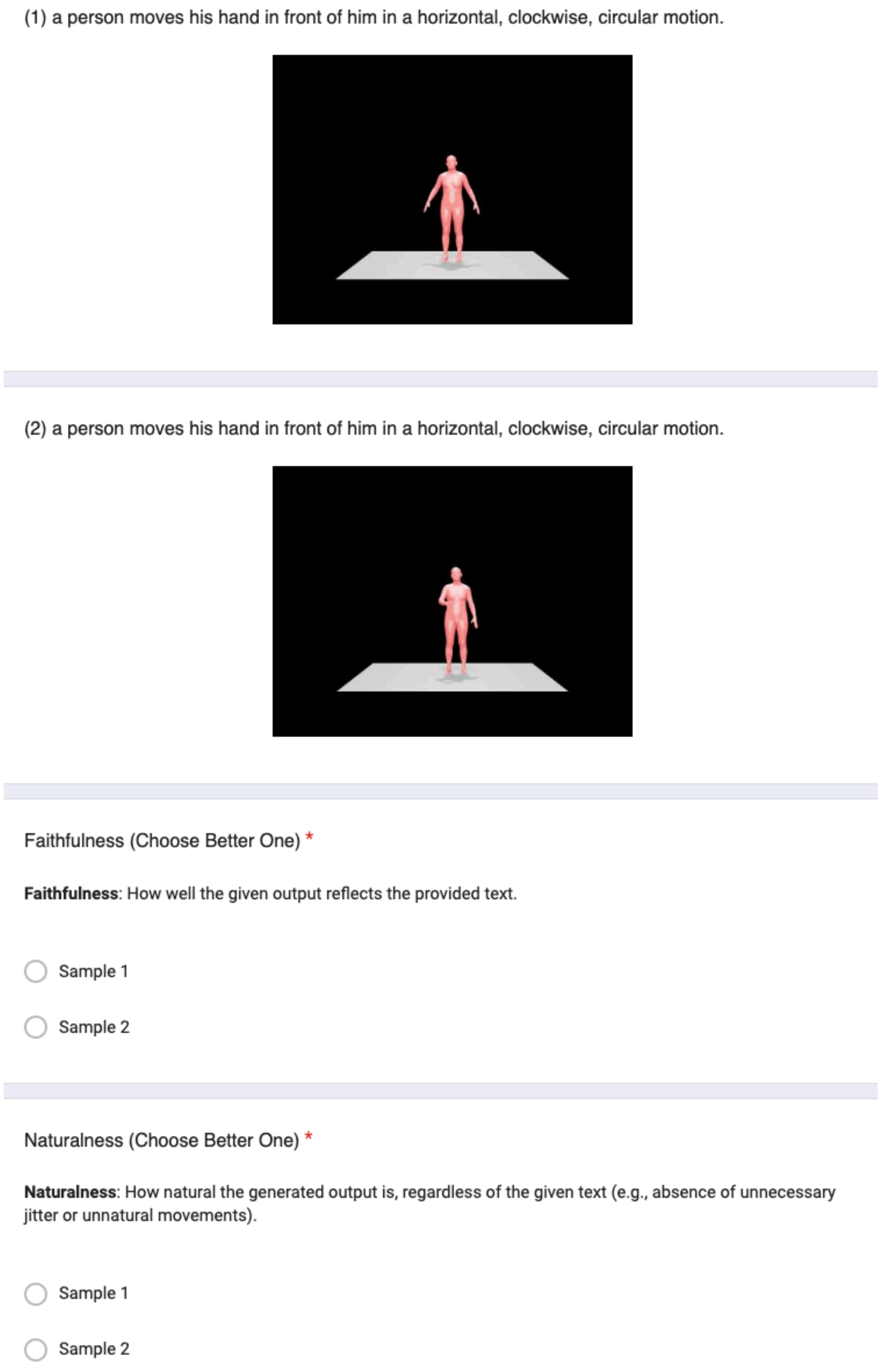
Notes)

1. The generated motions do not include facial information, so any unnatural **facial movements should be excluded from the evaluation.**

2. Similarly, the generated motions do not include detailed hand or finger information. Therefore, hand movements should be interpreted as wrist movements, and **finger movements should also be excluded from the evaluation.**

3. The video titles may contain text prompts, but please **disregard these prompts** and evaluate only the naturalness of the motion itself.

Figure J. **Guidelines for User Study in the Supplementary:** Participants were asked to evaluate Naturalness, excluding hand and facial movements that are not included in HumanML3D.



(1) a person moves his hand in front of him in a horizontal, clockwise, circular motion.

(2) a person moves his hand in front of him in a horizontal, clockwise, circular motion.

Faithfulness (Choose Better One) *

Faithfulness: How well the given output reflects the provided text.

○ Sample 1

○ Sample 2

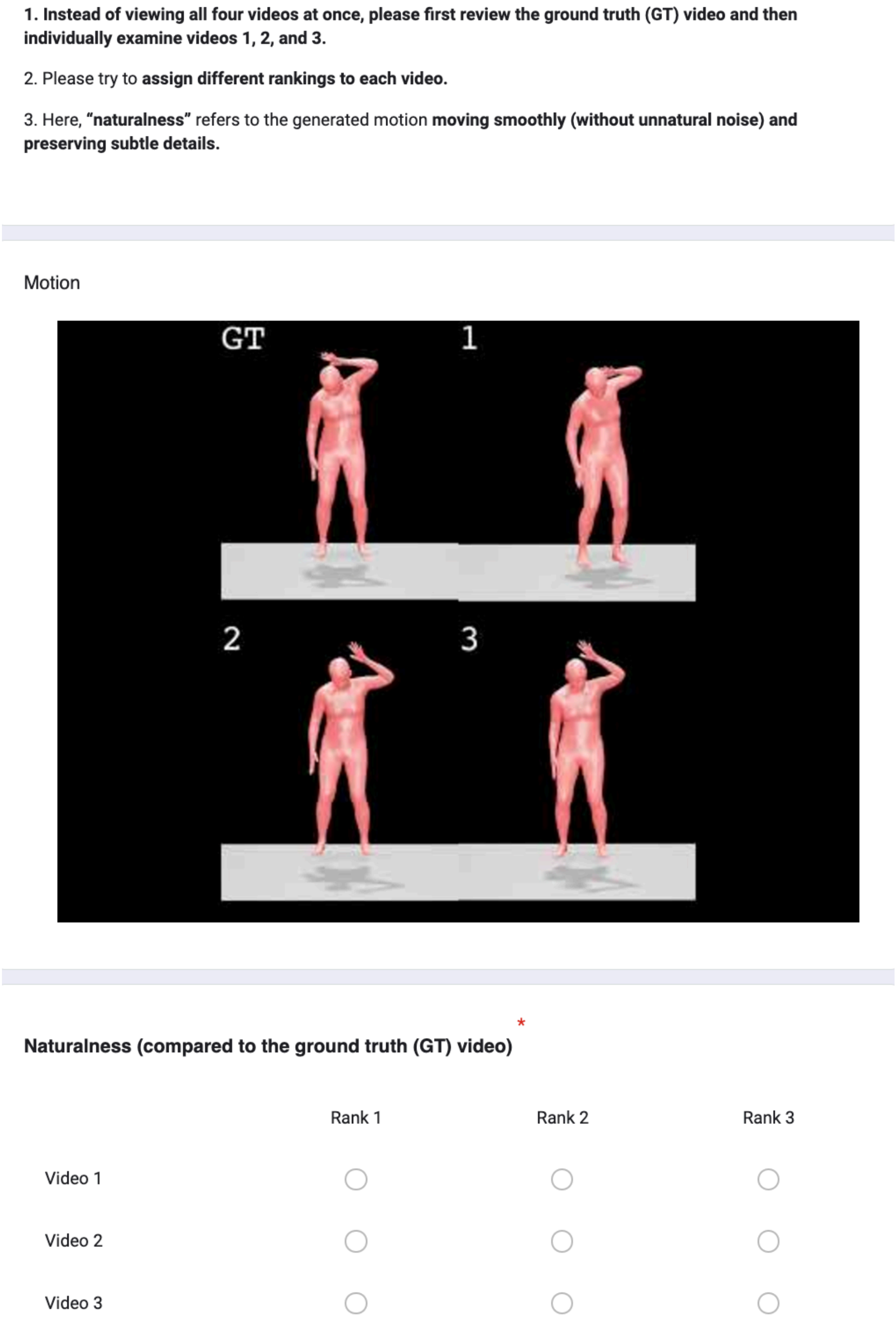
Naturalness (Choose Better One) *

Naturalness: How natural the generated output is, regardless of the given text (e.g., absence of unnecessary jitter or unnatural movements).

○ Sample 1

○ Sample 2

Figure K. **User evaluation interface** for the user study in the Main paper: participants were presented with two randomly selected videos and asked to choose the better sample in terms of faithfulness and naturalness.



1. Instead of viewing all four videos at once, please first review the ground truth (GT) video and then individually examine videos 1, 2, and 3.

2. Please try to **assign different rankings to each video.**

3. Here, **"naturalness"** refers to the generated motion **moving smoothly (without unnatural noise) and preserving subtle details.**

Motion

Naturalness (compared to the ground truth (GT) video) *

	Rank 1	Rank 2	Rank 3
Video 1	○	○	○
Video 2	○	○	○
Video 3	○	○	○

Figure L. **User evaluation interface** for the user study in the supplementary: participants were presented with a grid layout containing the GT video and three generated videos. Using the GT video as the upper bound, they were asked to rank the three generated videos in terms of naturalness.