BiPO: Bidirectional Partial Occlusion Network for Text-to-Motion Synthesis

Supplementary Material

1. Overview

The supplementary material is structured as follows:

- Section 2: Implementation Details.
- Section 3: Body part division.
- Section 4: Reconstrcution of VQ-VAE.
- Section 5: Effect of Dual-iteration Cascaded Part-based Motion Decoding.
- Section 6: Feature extractor for evaluation.
- Section 7: Evaluation for Motion editing.
- Section 8: Evaluation Metrics details.
- Section 9: Additional qualitative test.
- Section 10: Motion Representations.
- Section 11: Dataset Details.
- Section 12: Full Experiments of HumanML3D and KIT-ML.
- Section 13: User study details.
- Section 14: Further ablation study for PO.
- Section 15: Discussion of Inference Speed.
- Section 16: Evaluation on the MARDM Benchmark
- Section 17: Long-Term Motion Generation Evaluation
- Section 18: Evaluation on Motion-X

2. Implementation Details

2.1. Architecture

Our method utilizes 6 lightweight VQ-VAEs [17] to discretize part motions and 6 small transformers [18] equipped with Part Coordination modules for generating text-driven motions [26]. Each VQ-VAE contains a codebook with 512 entries. For most parts, the code dimension is set to 128, while the Root part has a reduced dimension of 64. The encoder applies a downsampling rate of $r = 4$ to reduce the motion sequence length. The transformers consist of 14 layers, with each layer having a token dimension of 256. A Selective Part Coordination Layer is added before all remaining layers except the first transformer layer. Selective Part Coordination Layer within the same layer of each transformer share their weights, and each Selective Part Coordination Layer includes 3 MLP layers. For text-to-motion generation, we use the CLIP model [16] with the ViT-B/32 variant to encode text features, enabling robust alignment between textual descriptions and motion representations. We used 1D Sinusoidal Positional Encoding as the positional embedding.

2.2. Hyperparameters

During training, we employ a masking probability of 40% within the Selective Part Coordination Layer to randomly mask out tokens from other parts, encouraging the model to generate coordinated motions even with incomplete information. For training the VQ-VAE, we use a learning rate of 2×10^{-4} and 3×10^{-4} for the first 200K steps, and 1×10^{-5} and 0.5×10^{-5} after 100k steps on HumanML3D and KIT-ML each. The AdamW [9] optimizer is applied with $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and the batch size is set to 256 and 196 on HumanML3D and KIT-ML each. The commitment loss weight β is fixed at 1.0. For transformer training, we use a learning rate of 1×10^{-4} for the first 150K steps, decreasing to 5×10^{-6} after 150K. The AdamW optimizer is also used for the transformer with $\beta_1 = 0.5$ and $\beta_2 = 0.99$, and the batch size is set to 64 and 48 on HumanML3D and KIT-ML each. All experiments are conducted using a single NVIDIA A6000 GPU and INTEL XEON(R) PLATINUM 8568Y+ in Ubuntu 22.04. The training of the VQ-VAE takes approximately 20 hours, while the training for text-to-motion generation takes around 64 hours.

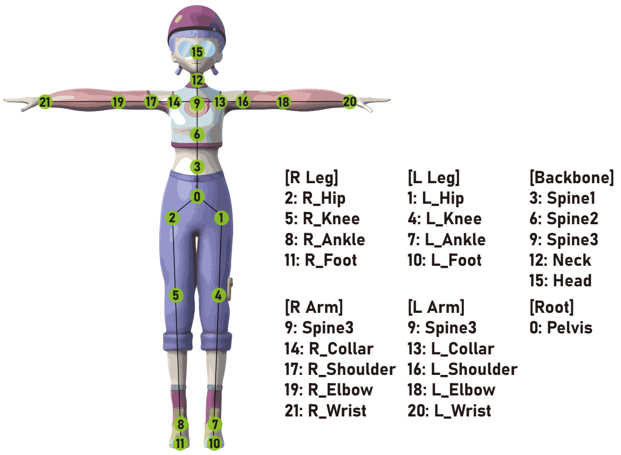


Figure 1. Visualization of body part division.

3. Body Part Division

Our method follows the same body partitioning strategy as ParCo [26], dividing the whole body into six parts: R.Leg, L.Leg, R.Arm, L.Arm, Backbone, and Root. The body partitioning is illustrated in Figure 1. For our experiments, we exclusively used the HumanML3D dataset [4]. Specifically, both R.Arm and L.Arm include the 9-th joint, as it serves as a critical key point connecting the arms to the backbone. This joint provides essential positional information for the arms relative to the connection point with the backbone.

During whole-body motion reconstruction from part motions, we generate three predictions for this joint: one from

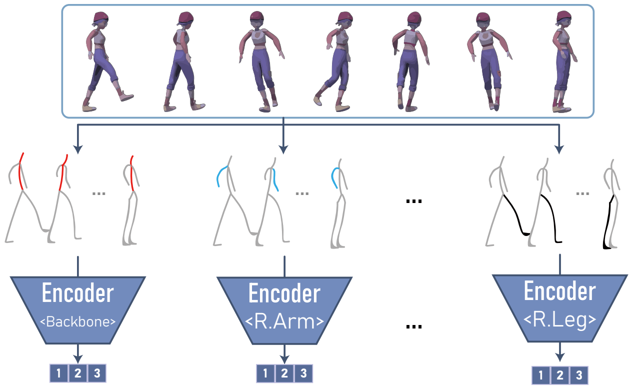


Figure 2. Part-based VQ-VAE architecture.

R.Arm, one from L.Arm, and one from Backbone. The final prediction for this joint is obtained by averaging these three values, ensuring consistent and accurate integration of part motions into the whole-body motion.

4. Reconstruction of VQ-VAE

We utilized part-based VQ-VAEs as proposed by ParCo [26], illustrated in Figure 2. The reconstruction performance is presented in Table 1. For evaluation, the reconstructed motions are integrated into whole-body motion sequences. While ParCo employs the VQ-VAE trained at the final iteration during training, we selected the VQ-VAE that achieved the best FID performance on the validation dataset. Experimentally, ParCo demonstrated the highest text-to-motion performance using the VQ-VAE from the final iteration. However, in our model, BiPO, selecting the VQ-VAE with the best FID performance on the validation dataset yield better overall performance, leading us to adopt this approach.

5. Effect of Dual-iteration Cascaded Part-based Motion Decoding

We conduct performance evaluations on the effectiveness of Dual-iteration Cascaded Part-based Motion Decoding. Dual-iteration Cascaded Motion Decoding, proposed by BAMM [13], is reported to outperform other masking strategies by masking even indexed motion tokens and predicting them again. Based on this, our model, BiPO, also adopt this masking strategy. Additionally, we conduct further experiments to evaluate the effectiveness of this approach. The performance results, present in Table 2, demonstrate that Dual-iteration Cascaded Part-based Motion Decoding is indeed effective.

6. Feature extractor for evaluation

For evaluation, we utilize a motion feature extractor and a text feature extractor trained using contrastive learning introduced by T2M [4]. These extractors are specifically designed to map text and motion features into a shared embedding space, where matching pairs are positioned closer together, and non-matching pairs are separated. This approach enables effective alignment of text-motion pairs for accurate evaluation. By employing the contrastive learning-based feature extractor, our evaluation framework aligns with established benchmarks, ensuring a rigorous assessment of text-to-motion alignment and generation quality.

7. Evaluation for Motion editing

We further evaluated our model's capabilities in motion editing tasks, with the results presented in Table 3. We conducted experiments on four motion editing scenarios to assess the model's adaptability and robustness: Temporal Inpainting, Temporal Outpainting, Prefix, and Suffix. Temporal Inpainting involves filling in the middle 50% of the motion sequence. Temporal Outpainting involves generating the outer 25% at both the beginning and the end of the sequence, given the middle 50%. Prefix involves generating the final 50% of the sequence based on the initial 50%, while Suffix involves generating the initial 50% of the sequence based on the final 50%. Since motion editing aligns more closely with prediction than generation, we did not apply the MModality metric.

In all scenarios, our model outperforms existing methods, particularly in FID, indicating superior quality and realism in the generated motions.

We visualize motion editing results. Visualizations for Motion editing are illustrated in Figure 3. In all four cases (Temporal Inpaintin, Temporal Outpainting, Prefix and Suffix), the remaining motions are appropriately generated to align with the given condition, demonstrating the effectiveness of our proposed model, BiPO, in performing the motion editing task.

8. Evaluation Metrics details

We use several evaluation metrics, as proposed in T2M [4], to measure the performance of our model. Below, we provide detailed formulations for these metrics.

8.1. Fréchet Inception Distance

Fréchet Inception Distance (FID) evaluates the quality of the generated motions by comparing the distribution of their features to the distribution of ground-truth motion features. It is calculated as follows:

$$\text{FID} = \|\mu_{\text{gt}} - \mu_{\text{pred}}\|^2 + \text{Tr}(\Sigma_{\text{gt}} + \Sigma_{\text{pred}} - 2(\Sigma_{\text{gt}}\Sigma_{\text{pred}})^{1/2}), \tag{1}$$

where μ_{gt} and μ_{pred} are the mean feature vectors of the ground-truth and predicted motions, respectively. Σ_{gt} and Σ_{pred} represent their covariance matrices, and Tr denotes the trace of a matrix.

Datasets	Methods	FID ↓	R-Precision (T2M) ↑			MM-Dist ↓	Diversity →	MModality ↑
			Top-1 ↑	Top-2 ↑	Top-3 ↑			
	Real motion	$0.002^{\pm.000}$	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
Human ML3D	BiPO (R)	$0.020^{\pm.000}$	$0.500^{\pm.003}$	$0.690^{\pm.002}$	$0.787^{\pm.002}$	$3.026^{\pm.006}$	$9.430^{\pm.094}$	-
	BiPO (G)	$0.030^{\pm.002}$	$0.523^{\pm.003}$	$0.714^{\pm.002}$	$0.809^{\pm.002}$	$2.880^{\pm.009}$	$9.556^{\pm.076}$	$1.374^{\pm.047}$
	Real motion	$0.031^{\pm.004}$	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
KIT-ML	BiPO (R)	$0.091^{\pm.005}$	$0.414^{\pm.007}$	$0.640^{\pm.004}$	$0.767^{\pm.005}$	$2.805^{\pm.014}$	$10.969^{\pm.093}$	-
	BiPO (G)	$0.164^{\pm.008}$	$0.444^{\pm.005}$	$0.674^{\pm.006}$	$0.803^{\pm.005}$	$2.658^{\pm.015}$	$10.833^{\pm.111}$	$1.098^{\pm.047}$

Table 1. Reconstruction and Generation results. BiPO (R) represents the reconstruction performance of VQ-VAE, while BiPO (G) represents the performance of text-to-motion generation.

Datasets	Methods	FID ↓	R-Precision (T2M) ↑			MM-Dist ↓	Diversity →	MModality ↑
			Top-1 ↑	Top-2 ↑	Top-3 ↑			
	Real motion	$0.002^{\pm.000}$	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
Human ML3D	BiPO (- DC)	$0.034^{\pm.002}$	$0.519^{\pm.003}$	$0.712^{\pm.003}$	$0.808^{\pm.002}$	$2.895^{\pm.009}$	$9.496^{\pm.076}$	**1.406**$^{\pm.052}$
	BiPO (S)	$0.517^{\pm.003}$	$0.708^{\pm.002}$	$0.805^{\pm.002}$	$0.043^{\pm.001}$	$2.932^{\pm.003}$	$9.374^{\pm.089}$	$1.393^{\pm.061}$
	BiPO (T)	$0.035^{\pm.001}$	$0.521^{\pm.002}$	**0.715**$^{\pm.002}$	**0.810**$^{\pm.002}$	$2.905^{\pm.003}$	$9.529^{\pm.073}$	$1.374^{\pm.048}$
	BiPO (L)	$0.037^{\pm.001}$	$0.520^{\pm.002}$	$0.712^{\pm.002}$	$0.809^{\pm.002}$	$2.913^{\pm.004}$	**9.497**$^{\pm.121}$	$1.385^{\pm.004}$
	BiPO	**0.030**$^{\pm.002}$	**0.523**$^{\pm.003}$	$0.714^{\pm.002}$	$0.809^{\pm.002}$	**2.880**$^{\pm.009}$	$9.556^{\pm.076}$	$1.374^{\pm.047}$
	Real motion	$0.031^{\pm.004}$	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
KIT-ML	BiPO (- DC)	$0.268^{\pm.026}$	$0.439^{\pm.006}$	$0.664^{\pm.004}$	$0.792^{\pm.005}$	$2.705^{\pm.016}$	**10.952**$^{\pm.129}$	**1.188**$^{\pm.099}$
	BiPO	**0.164**$^{\pm.008}$	**0.444**$^{\pm.005}$	**0.674**$^{\pm.006}$	**0.803**$^{\pm.005}$	**2.658**$^{\pm.015}$	$10.833^{\pm.111}$	$1.098^{\pm.047}$

Table 2. Ablation for Dual-iteration Cascaded Part-based Motion Decoding. BiPO (- DC) represents the results generated without employing Dual-iteration Cascaded Part-based Motion Decoding. Better result is highlighted in bold.

8.2. R-Precision

R-Precision evaluates the semantic alignment between text descriptions and generated motions by measuring the retrieval accuracy of the most relevant matches. Each text description's corresponding motion feature is expected to appear within the top k nearest neighbors of the motion features retrieved from the generated data.

To compute R-Precision, let f_{pred} and f_{text} represent the generated motion and the features of the text description, respectively. The distance matrix between all pairs of f_{pred} and f_{text} is defined as:

$$\mathbf{D}(i,j) = \|f_{\text{pred},i} - f_{\text{text},j}\|, \tag{2}$$

where $\mathbf{D}(i,j)$ denotes the Euclidean distance between the i-th generated motion feature and the j-th text feature.

For a given generated motion feature, we randomly sample 31 text descriptions from the test dataset. Along with the text description $f_{\text{text},i}$ matched to $f_{\text{pred},i}$, these 32 text descriptions form the search pool. The top k nearest text descriptions are retrieved by sorting the distances in ascending order. The R-Precision at top-k is calculated as:

$$\text{R-Precision@k} = \frac{1}{N}\sum_{i=1}^{N} \nVdash\{i \in \text{Top-}k(\mathbf{D}(i,:))\}, \tag{3}$$

where N is the total number of text-motion pairs, $\nVdash\{\cdot\}$ is the indicator function that equals 1 if the condition is true and 0 otherwise, and Top-$k(\mathbf{D}(i,:))$ represents the indices of the top-k closest text descriptions from the search pool to the i-th generated motion feature.

In our experiments, we report R-Precision for $k = 1, 2$, and 3 to provide a comprehensive evaluation of the alignment between text and motion.

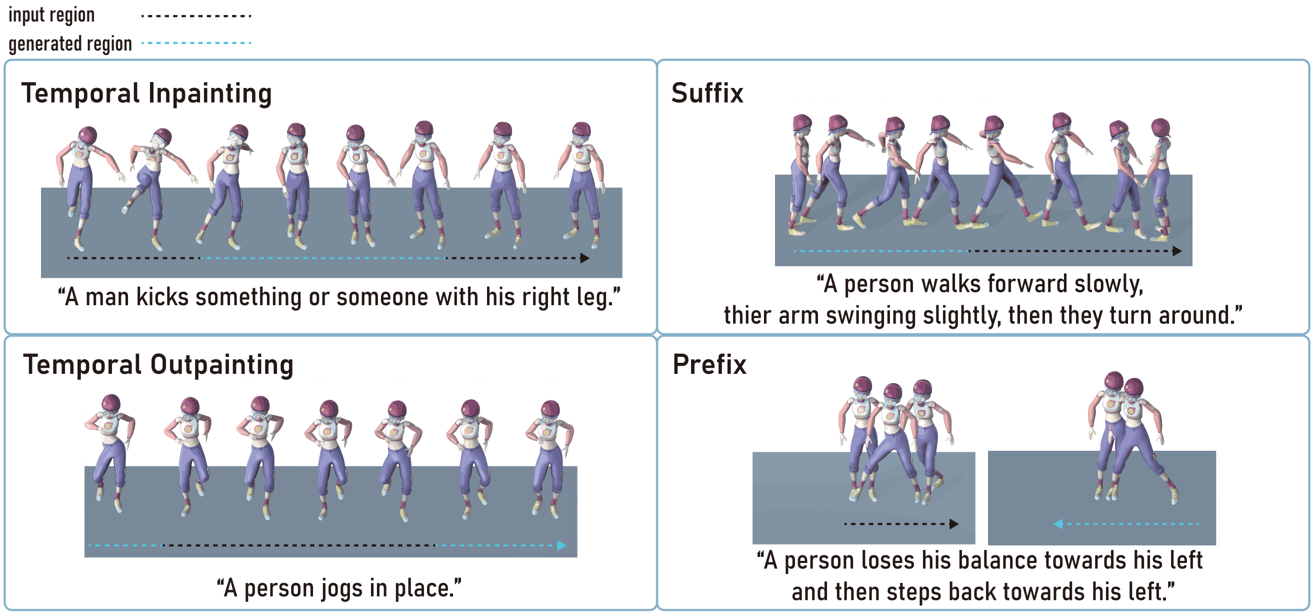
8.3. MultiModal Distance

MultiModal Distance (MM-Dist) measures the semantic alignment between the textual descriptions and the generated motions. It is defined as:

$$\text{MM-Dist} = \frac{1}{N}\sum_{i=1}^{N} \|f_{\text{pred},i} - f_{\text{text},i}\|, \tag{4}$$

Tasks	Methods	FID ↓	R-Precision (T2M) ↑			MM-Dist ↓	Diversity →
			Top-1 ↑	Top-2 ↑	Top-3 ↑		
Temporal Inpainting (In-betweening)	MDM	2.362	0.391	0.578	0.692	3.859	8.014
	MoMask	0.04	0.534	0.727	0.820	2.878	9.640
	BAMM	0.056	**0.535**	**0.729**	**0.821**	2.863	9.629
	BiPO	**0.029**	0.530	0.728	**0.821**	**2.837**	**9.617**
Temporal Outpainting	MDM	2.057	0.415	0.613	0.727	3.619	8.199
	MoMask	0.057	0.531	0.726	0.818	2.889	**9.619**
	BAMM	0.056	0.535	**0.730**	**0.822**	**2.856**	9.659
	BiPO	**0.052**	**0.536**	0.722	0.813	2.867	9.295
Prefix	MDM	1.460	0.420	0.613	0.725	3.563	8.972
	MoMask	0.060	**0.536**	**0.730**	**0.823**	2.875	**9.607**
	BAMM	0.058	0.532	0.727	0.821	**2.868**	9.612
	BiPO	**0.036**	0.522	0.717	0.808	2.876	9.357
Suffix	MDM	2.562	0.403	0.597	0.711	3.731	8.088
	MoMask	0.052	0.532	**0.726**	**0.819**	2.881	9.659
	BAMM	0.050	0.527	0.720	0.814	2.891	9.721
	BiPO	**0.046**	**0.533**	0.719	0.809	**2.861**	**9.513**

Table 3. Motion editing results. The target benchmark for Diversity is 9.503, corresponding to the Diversity of the real motion.



Temporal Inpainting

"A man kicks something or someone with his right leg."

Suffix

"A person walks forward slowly,
thier arm swinging slightly, then they turn around."

Temporal Outpainting

"A person jogs in place."

Prefix

"A person loses his balance towards his left
and then steps back towards his left."

Figure 3. Visualization for Motion editing. The input region represents the real motion used as a condition, while the generated region refers to the motion generated by Our model, BiPO.

where $f_{\text{pred},i}$ and $f_{\text{text},i}$ are the features of the i-th generated motion and its corresponding text description, respectively.

8.4. Diversity

Diversity measures the variance of the generated motion sequences. For S_{dis} randomly sampled motion pairs, the diversity is computed as:

$$\text{Diversity} = \frac{1}{S_{\text{dis}}} \sum_{i=1}^{S_{\text{dis}}} \|f_{\text{pred},i} - f'_{\text{pred},i}\|, \tag{5}$$

where $f_{\text{pred},i}$ and $f'_{\text{pred},i}$ are the features of the i-th pair of generated motions. In our experiments, S_{dis} is set to 300,

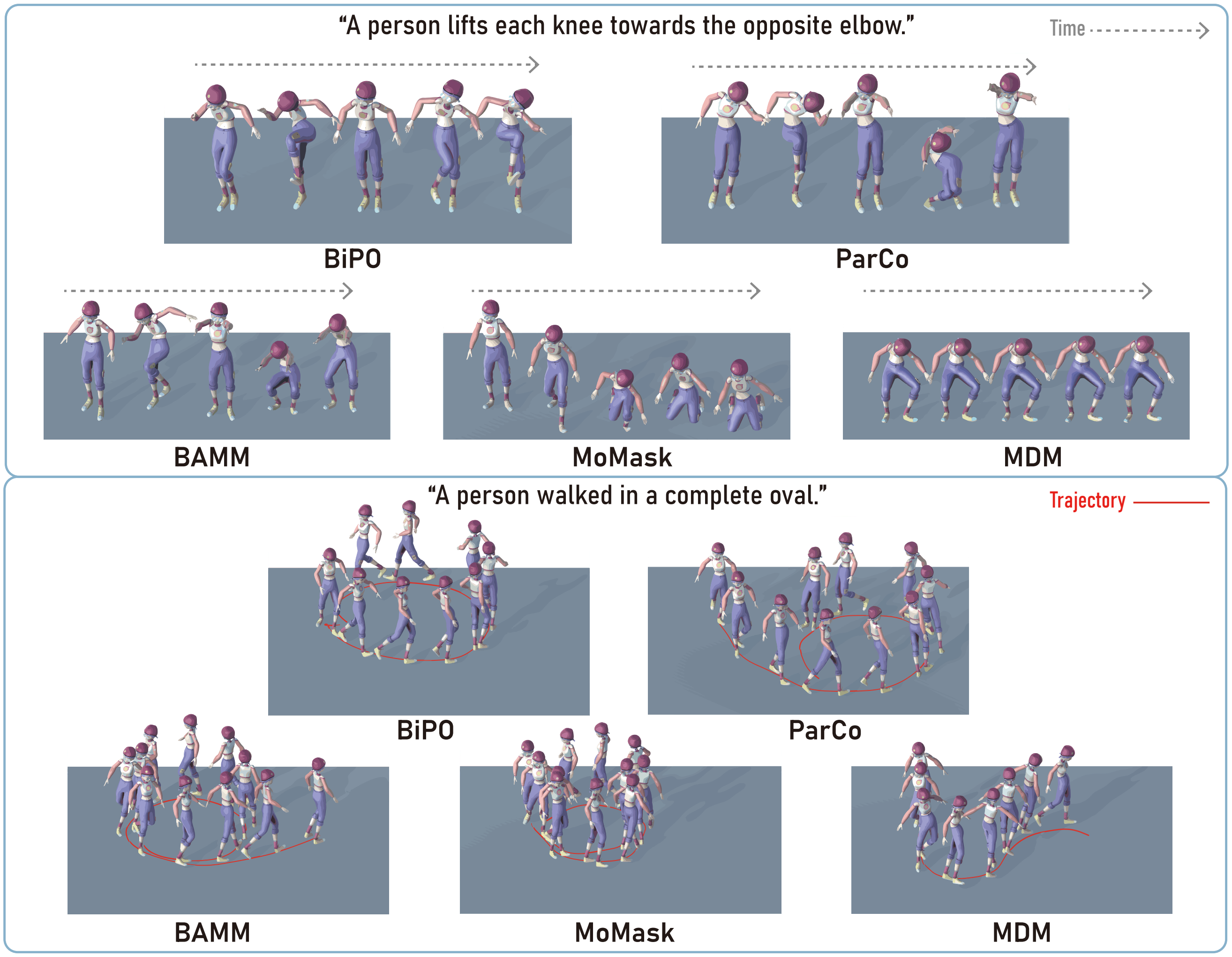

"A person lifts each knee towards the opposite elbow."
Time
BiPO
ParCo
BAMM
MoMask
MDM
"A person walked in a complete oval."
Trajectory
BiPO
ParCo
BAMM
MoMask
MDM

Figure 4. Additional qualitative test.

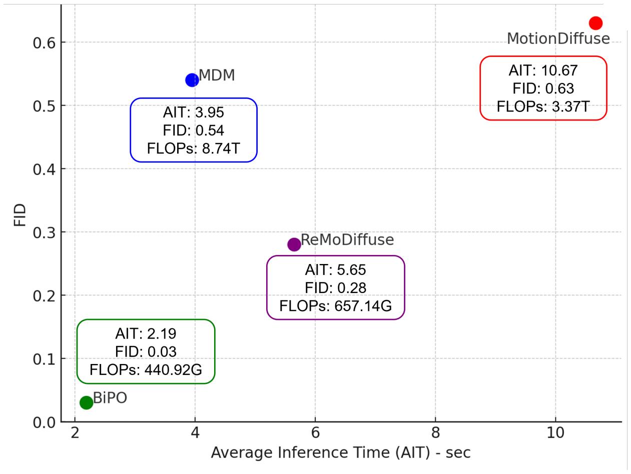


Figure 5. Efficiency analysis.

following T2M [4].

8.5. Multimodality

Multimodality (MModality) evaluates the diversity of motions generated from the same text description. For the r-th text prompt, we generate 30 motions and randomly sample two subsets, each containing 10 motions. The metric is calculated as:

$$\text{MModality} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{10}\sum_{j=1}^{10}\|f_{\text{pred},i,j} - f'_{\text{pred},i,j}\|, \quad (6)$$

where $f_{\text{pred},i,j}$ and $f'_{\text{pred},i,j}$ are the features of the i-th pair of generated motions for the r-th text description.

9. Additional qualitative test

We visualize additional qualitative tests. The examples are shown in Figure 4, featuring motions generated based on

Datasets	Methods	FID ↓	R-Precision (T2M) ↑			MM-Dist ↓	Diversity →	MModality ↑
			Top-1 ↑	Top-2 ↑	Top-3 ↑			
	Real motion	$0.002^{\pm.000}$	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
Human ML3D	T2M(2022) [4]	$1.087^{\pm.021}$	$0.455^{\pm.003}$	$0.636^{\pm.003}$	$0.736^{\pm.002}$	$3.347^{\pm.008}$	$9.175^{\pm.083}$	$2.219^{\pm.074}$
	TEMOS(2022) [12]	$3.734^{\pm.028}$	$0.424^{\pm.002}$	$0.612^{\pm.002}$	$0.722^{\pm.002}$	$3.703^{\pm.008}$	$8.973^{\pm.071}$	$0.368^{\pm.018}$
	TM2T(2022) [5]	$1.501^{\pm.017}$	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.002}$	$3.467^{\pm.011}$	$8.589^{\pm.076}$	$2.424^{\pm.093}$
	MotionDiffuse(2022)§ [22]	$0.630^{\pm.011}$	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$	$1.553^{\pm.042}$
	MDM(2022)§ [20]	$0.544^{\pm.044}$	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$5.566^{\pm.027}$	$\underline{9.559^{\pm.086}}$	$\underline{2.799^{\pm.072}}$
	MLD(2022)§ [1]	$0.473^{\pm.013}$	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
	T2M-GPT(2023) [21]	$0.116^{\pm.004}$	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
	M2DM(2023)§ [7]	$0.352^{\pm.005}$	$0.497^{\pm.003}$	$0.682^{\pm.002}$	$0.763^{\pm.003}$	$3.134^{\pm.010}$	$9.926^{\pm.073}$	$\mathbf{3.587}^{\pm.072}$
	ReMoDiffuse(2023)*	$0.281^{\pm.010}$	$0.450^{\pm.003}$	$0.638^{\pm.002}$	$0.743^{\pm.003}$	$3.271^{\pm.008}$	$9.236^{\pm.085}$	-
	Fg-T2M(2023)§ [19]	$0.243^{\pm.019}$	$0.492^{\pm.002}$	$0.683^{\pm.003}$	$0.783^{\pm.002}$	$3.109^{\pm.007}$	$9.278^{\pm.072}$	$1.614^{\pm.049}$
	AttT2M(2023) [25]	$0.112^{\pm.006}$	$0.499^{\pm.003}$	$0.690^{\pm.002}$	$0.786^{\pm.002}$	$3.038^{\pm.007}$	$9.700^{\pm.090}$	$2.452^{\pm.051}$
	FineMoGen(2023)§ [23]	$0.151^{\pm.008}$	$0.504^{\pm.002}$	$0.690^{\pm.002}$	$0.784^{\pm.002}$	$2.998^{\pm.008}$	$9.263^{\pm.094}$	$2.696^{\pm.079}$
	MoMask(2024)§ [6]	$\underline{0.045^{\pm.002}}$	$0.521^{\pm.002}$	$0.713^{\pm.003}$	$0.807^{\pm.002}$	$2.958^{\pm.008}$	-	$1.241^{\pm.040}$
	MMM(2024)§ [6]	$0.089^{\pm.005}$	$0.515^{\pm.002}$	$0.708^{\pm.002}$	$0.804^{\pm.002}$	$2.926^{\pm.007}$	$9.577^{\pm.050}$	$1.226^{\pm.035}$
	MotionMamba(2024)§ [24]	$0.281^{\pm.009}$	$0.502^{\pm.003}$	$0.693^{\pm.002}$	$0.792^{\pm.002}$	$3.060^{\pm.058}$	$9.871^{\pm.084}$	$2.294^{\pm.058}$
	ParCo(2024) [26]	$0.109^{\pm.005}$	$0.515^{\pm.003}$	$0.706^{\pm.003}$	$0.801^{\pm.002}$	$2.927^{\pm.008}$	$9.576^{\pm.088}$	$1.382^{\pm.060}$
	BAMM(2024) [13]	$0.055^{\pm.002}$	$\mathbf{0.525}^{\pm.002}$	$\mathbf{0.720}^{\pm.003}$	$\mathbf{0.814}^{\pm.003}$	$\underline{2.919^{\pm.008}}$	$9.717^{\pm.089}$	$1.687^{\pm.051}$
	BiPO (Ours)	$\mathbf{0.030}^{\pm.002}$	$\underline{0.523^{\pm.003}}$	$\underline{0.714^{\pm.002}}$	$\underline{0.809^{\pm.002}}$	$\mathbf{2.880}^{\pm.009}$	$\mathbf{9.556}^{\pm.076}$	$1.374^{\pm.047}$
	Real motion	$0.031^{\pm.004}$	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
KIT-ML	T2M(2022) [4]	$3.022^{\pm.107}$	$0.361^{\pm.006}$	$0.559^{\pm.007}$	$0.681^{\pm.007}$	$3.488^{\pm.028}$	$10.72^{\pm.145}$	$2.052^{\pm.107}$
	TEMOS(2022) [12]	$3.717^{\pm.051}$	$0.353^{\pm.006}$	$0.561^{\pm.007}$	$0.687^{\pm.005}$	$3.417^{\pm.019}$	$10.84^{\pm.100}$	$0.532^{\pm.034}$
	TM2T(2022) [5]	$3.599^{\pm.153}$	$0.280^{\pm.005}$	$0.463^{\pm.006}$	$0.587^{\pm.005}$	$4.591^{\pm.026}$	$9.473^{\pm.117}$	$\underline{3.292^{\pm.081}}$
	MotionDiffuse(2022)§ [22]	$1.954^{\pm.062}$	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$2.958^{\pm.005}$	$\mathbf{11.10}^{\pm.143}$	$0.730^{\pm.013}$
	MDM(2022)§ [20]	$0.497^{\pm.021}$	$0.164^{\pm.004}$	$0.291^{\pm.004}$	$0.396^{\pm.004}$	$9.191^{\pm.022}$	$10.85^{\pm.109}$	$1.907^{\pm.214}$
	MLD(2022)§ [1]	$0.404^{\pm.027}$	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$3.204^{\pm.027}$	$10.80^{\pm.117}$	$2.192^{\pm.071}$
	T2M-GPT(2023) [21]	$0.717^{\pm.041}$	$0.402^{\pm.006}$	$0.619^{\pm.005}$	$0.737^{\pm.006}$	$3.053^{\pm.026}$	$10.86^{\pm.094}$	$1.912^{\pm.036}$
	M2DM(2023)§ [7]	$0.515^{\pm.029}$	$0.416^{\pm.004}$	$0.628^{\pm.004}$	$0.743^{\pm.004}$	$3.015^{\pm.017}$	$11.417^{\pm.97}$	$\mathbf{3.325}^{\pm.37}$
	ReMoDiffuse(2023)*	$0.589^{\pm.022}$	$0.382^{\pm.005}$	$0.586^{\pm.007}$	$0.706^{\pm.006}$	$3.324^{\pm.030}$	$10.31^{\pm.065}$	-
	Fg-T2M(2023)§ [19]	$0.571^{\pm.047}$	$0.418^{\pm.005}$	$0.626^{\pm.004}$	$0.745^{\pm.004}$	$3.114^{\pm.015}$	$10.93^{\pm.083}$	$1.019^{\pm.029}$
	AttT2M(2023) [25]	$0.870^{\pm.039}$	$0.413^{\pm.006}$	$0.632^{\pm.006}$	$0.751^{\pm.006}$	$3.039^{\pm.021}$	$10.96^{\pm.123}$	$2.281^{\pm.047}$
	FineMoGen(2023)§ [23]	$0.316^{\pm.028}$	$0.404^{\pm.005}$	$0.621^{\pm.005}$	$0.744^{\pm.004}$	$2.977^{\pm.019}$	$10.910^{\pm.101}$	$1.232^{\pm.039}$
	MoMask(2024)§ [6]	$0.204^{\pm.011}$	$0.433^{\pm.007}$	$0.656^{\pm.005}$	$0.781^{\pm.005}$	$2.779^{\pm.022}$	-	$1.131^{\pm.043}$
	MMM(2024)§ [14]	$0.316^{\pm.028}$	$0.404^{\pm.005}$	$0.621^{\pm.005}$	$0.744^{\pm.004}$	$2.977^{\pm.019}$	$10.910^{\pm.101}$	$1.232^{\pm.039}$
	MotionMamba(2024)§ [24]	$0.307^{\pm.041}$	$0.419^{\pm.006}$	$0.645^{\pm.005}$	$0.765^{\pm.006}$	$3.021^{\pm.025}$	$11.02^{\pm.098}$	$1.678^{\pm.064}$
	ParCo(2024) [26]	$0.453^{\pm.027}$	$0.430^{\pm.004}$	$0.649^{\pm.007}$	$0.772^{\pm.006}$	$2.820^{\pm.028}$	$10.95^{\pm.094}$	$1.245^{\pm.022}$
	BAMM(2024) [13]	$0.183^{\pm.013}$	$\underline{0.438^{\pm.009}}$	$\underline{0.661^{\pm.009}}$	$\underline{0.788^{\pm.005}}$	$\underline{2.723^{\pm.026}}$	$\underline{11.008^{\pm.094}}$	$1.609^{\pm.065}$
	BiPO (Ours)	$\mathbf{0.164}^{\pm.008}$	$\mathbf{0.444}^{\pm.005}$	$\mathbf{0.674}^{\pm.006}$	$\mathbf{0.803}^{\pm.005}$	$\mathbf{2.658}^{\pm.015}$	$10.833^{\pm.111}$	$1.098^{\pm.047}$

Table 4. Comparative results on the HumanML3D and KIT-ML test set against current state-of-the-art methods. Metrics where "↑" indicates that a higher value is preferred, "↓" indicates that a lower value is favorable, and "→" indicates metrics optimized when closer to real motion score of 9.503 and 11.08 each. The top result is highlighted in bold, with the second-best result underlined. The symbol § indicates evaluations performed using the ground-truth motion length. The ReMoDiffuse* results are obtained using official checkpoints, with motion lengths randomly sampled in a uniform manner as input [26]. The order of listing is based on the date it was first published.

text prompts from the HumanML3D test set. These examples highlight the capability of our method to produce natural and well-coordinated motions that correspond closely to the input text descriptions.

10. Motion Representations

For motion representation, we follow T2M [4]. Each pose is described by:

$$(\dot{r}_a, \dot{r}_x, \dot{r}_z, r_y, j_p, j_v, j_r, c_f), \tag{7}$$

where $\dot{r}_a$ is the global root angular velocity; $\dot{r}_x, \dot{r}_z$ are the root velocities in the X-Z plane; j_p, j_v, j_r represent joint

Datasets	Methods	FID ↓	R-Precision ↑			MM-Dist ↓	MModality ↑	CLIP-score ↑
			Top-1 ↑	Top-2 ↑	Top-3 ↑			
HumanML3D	MARDM-DDPM [11]	$0.116^{\pm.004}$	$0.492^{\pm.006}$	$0.690^{\pm.005}$	$0.790^{\pm.005}$	$3.349^{\pm.010}$	$2.470^{\pm.053}$	$0.637^{\pm.005}$
	MARDM-SiT [11]	$0.114^{\pm.007}$	$0.500^{\pm.004}$	$0.695^{\pm.003}$	$0.795^{\pm.003}$	$3.270^{\pm.009}$	$2.231^{\pm.071}$	$0.642^{\pm.002}$
	BiPO	**0.112**$^{\pm.008}$	**0.501**$^{\pm.002}$	**0.698**$^{\pm.002}$	**0.796**$^{\pm.003}$	**3.267**$^{\pm.005}$	**3.185**$^{\pm.110}$	**0.644**$^{\pm.004}$
KIT-ML	MARDM-DDPM [11]	$0.340^{\pm.020}$	$0.375^{\pm.006}$	$0.597^{\pm.008}$	$0.739^{\pm.006}$	$3.489^{\pm.018}$	$1.479^{\pm.078}$	$0.681^{\pm.003}$
	MARDM-SiT [11]	$0.242^{\pm.014}$	$0.387^{\pm.006}$	$0.610^{\pm.006}$	$0.749^{\pm.006}$	$3.374^{\pm.019}$	$1.312^{\pm.053}$	$0.692^{\pm.002}$
	BiPO	**0.239**$^{\pm.018}$	**0.389**$^{\pm.004}$	**0.611**$^{\pm.006}$	**0.752**$^{\pm.006}$	**3.298**$^{\pm.015}$	**1.529**$^{\pm.098}$	**0.693**$^{\pm.003}$

Table 5. Evaluation on the MARDM Benchmark.

Methods	FID ↓	R-Precision ↑			MM-Dist ↓	Diversity →	MModality ↑
		Top-1 ↑	Top-2 ↑	Top-3 ↑			
Real	$0.004^{\pm.000}$	$0.437^{\pm.003}$	$0.622^{\pm.004}$	$0.721^{\pm.004}$	$3.343^{\pm.015}$	$8.423^{\pm.090}$	-
MDM [20]	$0.802^{\pm.044}$	$0.368^{\pm.005}$	$0.553^{\pm.006}$	$0.672^{\pm.005}$	$3.860^{\pm.025}$	$8.817^{\pm.068}$	-
MotionDiffuse [22]	$2.460^{\pm.062}$	$0.367^{\pm.004}$	$0.521^{\pm.004}$	$0.623^{\pm.004}$	$3.789^{\pm.005}$	$8.707^{\pm.143}$	$1.602^{\pm.013}$
MLD [1]	$0.952^{\pm.020}$	$0.403^{\pm.005}$	$0.584^{\pm.005}$	$0.690^{\pm.005}$	$3.580^{\pm.016}$	$9.050^{\pm.085}$	**2.711**$^{\pm.104}$
MotionMamba [24]	$0.668^{\pm.019}$	$0.417^{\pm.003}$	$0.606^{\pm.003}$	$0.713^{\pm.004}$	$3.435^{\pm.015}$	$9.021^{\pm.070}$	$2.373^{\pm.084}$
ParCo [26]	$0.488^{\pm.022}$	$0.454^{\pm.005}$	$0.637^{\pm.005}$	$0.738^{\pm.005}$	$3.230^{\pm.014}$	$8.921^{\pm.066}$	$1.489^{\pm.106}$
MoMask [6]	$0.317^{\pm.013}$	$0.455^{\pm.004}$	$0.639^{\pm.003}$	$0.740^{\pm.004}$	$3.287^{\pm.015}$	$8.914^{\pm.079}$	$1.390^{\pm.046}$
BAMM [13]	$0.308^{\pm.010}$	$0.458^{\pm.004}$	$0.645^{\pm.004}$	$0.746^{\pm.004}$	$3.232^{\pm.013}$	$8.981^{\pm.075}$	$2.021^{\pm.099}$
BiPO	**0.216**$^{\pm.005}$	**0.467**$^{\pm.003}$	**0.649**$^{\pm.002}$	**0.748**$^{\pm.003}$	**3.117**$^{\pm.004}$	**8.687**$^{\pm.062}$	$1.492^{\pm.111}$

Table 6. Long-Term Motion Generation Evaluation.

positions, velocities, and rotations, respectively; and c_f denotes foot contact features derived from the heel and toe joint velocities. Each pose is represented as a feature vector with a total dimension of 263.

11. Dataset Details

The HumanML3D dataset [4] is constructed by combining motion sequences from two large-scale publicly available datasets, HumanAct12 [3] and AMASS [10]. These datasets consist of various types of human actions, including everyday activities such as walking and jumping, sports like swimming and karate, acrobatic movements such as cartwheels, and artistic performances like dancing.

The dataset is processed to ensure consistency and usability. Motion sequences are normalized to 20 frames per second (FPS), and those exceeding 10 seconds in duration are randomly cropped to 10 seconds. Each motion sequence is retargeted to a standardized human skeletal template and oriented to initially face the positive Z-axis.

To provide textual descriptions for the motions, annotations are collected through Amazon Mechanical Turk (AMT). Annotators are required to describe each motion with at least five words, and three descriptions are provided for each motion clip by different individuals. These descriptions undergo a post-processing step to remove inconsistencies or errors, resulting in high-quality textual annotations.

The final HumanML3D dataset comprises 14,616 motion sequences with a total of 44,970 textual descriptions, featuring a vocabulary of 5,371 unique words. The dataset spans approximately 28.59 hours of motion data, with an average clip length of 7.1 seconds, ranging from 2 to 10 seconds. The average textual description length is 12 words, with a median of 10 words. This makes HumanML3D one of the most extensive datasets for research involving text-to-motion synthesis. The dataset was further augmented using mirroring techniques to increase diversity. For example, a motion described as “A man kicks something or someone with his left leg” was mirrored to create a new motion with the description “A man kicks something or someone with his right leg.” This approach ensures a balanced representation of left and right directional movements in the dataset.

The KIT-ML dataset [15] comprises 3,911 human motion sequences paired with 6,278 textual descriptions. It has a vocabulary of 1,623 unique words, excluding distinctions based on capitalization and punctuation. The motion data is sourced from the KIT [15] and CMU [2] datasets but is downsampled to 12.5 FPS. Each motion sequence is annotated with one to four descriptive sentences, with an average sentence length of approximately 8 words.

Methods	FID ↓	R-Precision ↑			MM-Dist ↓	Diversity →	MModality ↑
		Top-1 ↑	Top-2 ↑	Top-3 ↑			
Real	$0.004^{\pm.000}$	$0.424^{\pm.004}$	$0.639^{\pm.004}$	$0.760^{\pm.004}$	$3.190^{\pm.013}$	$9.356^{\pm.090}$	-
MDM	$1.406^{\pm.112}$	$0.426^{\pm.011}$	$0.621^{\pm.011}$	$0.726^{\pm.014}$	$3.756^{\pm.050}$	$12.027^{\pm.148}$	$2.552^{\pm.220}$
MLD	$0.702^{\pm.045}$	$0.425^{\pm.004}$	$0.623^{\pm.005}$	$0.732^{\pm.005}$	$3.947^{\pm.020}$	$11.088^{\pm.135}$	**2.600**$^{\pm.072}$
ParCo	$0.206^{\pm.012}$	$0.436^{\pm.003}$	$0.644^{\pm.003}$	$0.759^{\pm.003}$	$3.591^{\pm.016}$	$11.122^{\pm.093}$	$2.327^{\pm.103}$
MoMask	$0.219^{\pm.015}$	$0.467^{\pm.004}$	$0.674^{\pm.004}$	$0.780^{\pm.003}$	$3.384^{\pm.019}$	$11.497^{\pm.090}$	$1.279^{\pm.049}$
BAMM	$0.208^{\pm.011}$	$0.465^{\pm.004}$	$0.669^{\pm.003}$	$0.774^{\pm.003}$	$3.433^{\pm.012}$	$11.447^{\pm.068}$	$1.756^{\pm.057}$
BiPO	**0.123**$^{\pm.009}$	**0.478**$^{\pm.004}$	**0.676**$^{\pm.004}$	**0.781**$^{\pm.003}$	**3.300**$^{\pm.015}$	**11.052**$^{\pm.122}$	$2.291^{\pm.024}$

Table 7. Evaluation on Motion-X.

Datasets	Methods	FID ↓	R-Precision (T2M) ↑			MM-Dist ↓	Diversity →	MModality ↑
			Top-1	Top-2	Top-3			
Human ML3D	baseline	$0.108^{\pm.007}$	$0.506^{\pm.003}$	$0.701^{\pm.003}$	$0.796^{\pm.002}$	$2.952^{\pm.008}$	$9.544^{\pm.093}$	$1.401^{\pm.059}$
	with BA	$0.047^{\pm.002}$	$0.517^{\pm.002}$	$0.709^{\pm.003}$	$0.804^{\pm.002}$	$2.915^{\pm.010}$	$9.497^{\pm.081}$	$1.278^{\pm.059}$
	with PO (25%)	$0.065^{\pm.003}$	$0.511^{\pm.003}$	$0.704^{\pm.002}$	$0.800^{\pm.002}$	$2.934^{\pm.009}$	$9.563^{\pm.072}$	$1.518^{\pm.115}$
	with PO (30%)	$0.063^{\pm.004}$	$0.507^{\pm.003}$	$0.697^{\pm.002}$	$0.793^{\pm.002}$	$2.967^{\pm.008}$	$9.522^{\pm.072}$	$1.615^{\pm.075}$
	with PO (35%)	$0.060^{\pm.004}$	$0.507^{\pm.003}$	$0.699^{\pm.002}$	$0.795^{\pm.001}$	$2.962^{\pm.009}$	$9.552^{\pm.081}$	**1.674**$^{\pm.061}$
	with PO (40%)	$0.051^{\pm.004}$	$0.505^{\pm.003}$	$0.696^{\pm.003}$	$0.794^{\pm.002}$	$2.969^{\pm.010}$	$9.535^{\pm.077}$	$1.653^{\pm.070}$
	with BA + PO (25%)	$0.039^{\pm.003}$	$0.520^{\pm.002}$	$0.711^{\pm.002}$	$0.807^{\pm.002}$	$2.889^{\pm.007}$	**9.499**$^{\pm.065}$	$1.285^{\pm.054}$
	with BA + PO (30%)	$0.032^{\pm.002}$	$0.519^{\pm.003}$	$0.713^{\pm.002}$	$0.809^{\pm.002}$	$2.896^{\pm.009}$	$9.449^{\pm.089}$	$1.384^{\pm.010}$
	with BA + PO (35%)	$0.032^{\pm.001}$	$0.520^{\pm.003}$	**0.715**$^{\pm.002}$	**0.810**$^{\pm.002}$	**2.879**$^{\pm.008}$	$9.533^{\pm.078}$	$1.271^{\pm.090}$
	with BA + PO (40%)	**0.030**$^{\pm.002}$	**0.523**$^{\pm.003}$	$0.714^{\pm.002}$	$0.809^{\pm.002}$	$2.880^{\pm.009}$	$9.556^{\pm.076}$	$1.374^{\pm.047}$
KIT-ML	baseline	$0.376^{\pm.027}$	$0.434^{\pm.006}$	$0.653^{\pm.005}$	$0.773^{\pm.006}$	$2.825^{\pm.026}$	$10.929^{\pm.085}$	$1.397^{\pm.039}$
	with BA	$0.361^{\pm.036}$	$0.444^{\pm.006}$	$0.658^{\pm.007}$	$0.782^{\pm.006}$	$2.771^{\pm.028}$	$10.843^{\pm.109}$	**1.668**$^{\pm.023}$
	with PO (10 %)	$0.371^{\pm.036}$	$0.437^{\pm.006}$	$0.662^{\pm.009}$	$0.785^{\pm.006}$	$2.775^{\pm.019}$	$11.067^{\pm.106}$	$1.266^{\pm.069}$
	with BA + PO (25 %)	$0.289^{\pm.033}$	**0.446**$^{\pm.006}$	$0.670^{\pm.006}$	$0.794^{\pm.004}$	$2.707^{\pm.017}$	$10.945^{\pm.096}$	$1.196^{\pm.069}$
	with BA + PO (20 %)	$0.360^{\pm.035}$	$0.437^{\pm.007}$	$0.667^{\pm.006}$	$0.789^{\pm.006}$	$2.767^{\pm.022}$	$10.915^{\pm.092}$	$1.315^{\pm.113}$
	with BA + PO (15 %)	$0.273^{\pm.019}$	$0.442^{\pm.007}$	$0.671^{\pm.006}$	$0.794^{\pm.007}$	$2.729^{\pm.019}$	**11.081**$^{\pm.109}$	$1.265^{\pm.052}$
	with BA + PO (10 %)	**0.164**$^{\pm.008}$	$0.444^{\pm.005}$	**0.674**$^{\pm.006}$	**0.803**$^{\pm.005}$	**2.658**$^{\pm.015}$	$10.833^{\pm.111}$	$1.098^{\pm.047}$

Table 8. Ablation Study. The percentage values next to PO indicate the likelihood of occluding each part's information within the PO technique. The target benchmark for Diversity is 9.503 and 11.08 each, corresponding to the Diversity of the real motion.

12. Full Experiments of HumanML3D and KIT-ML

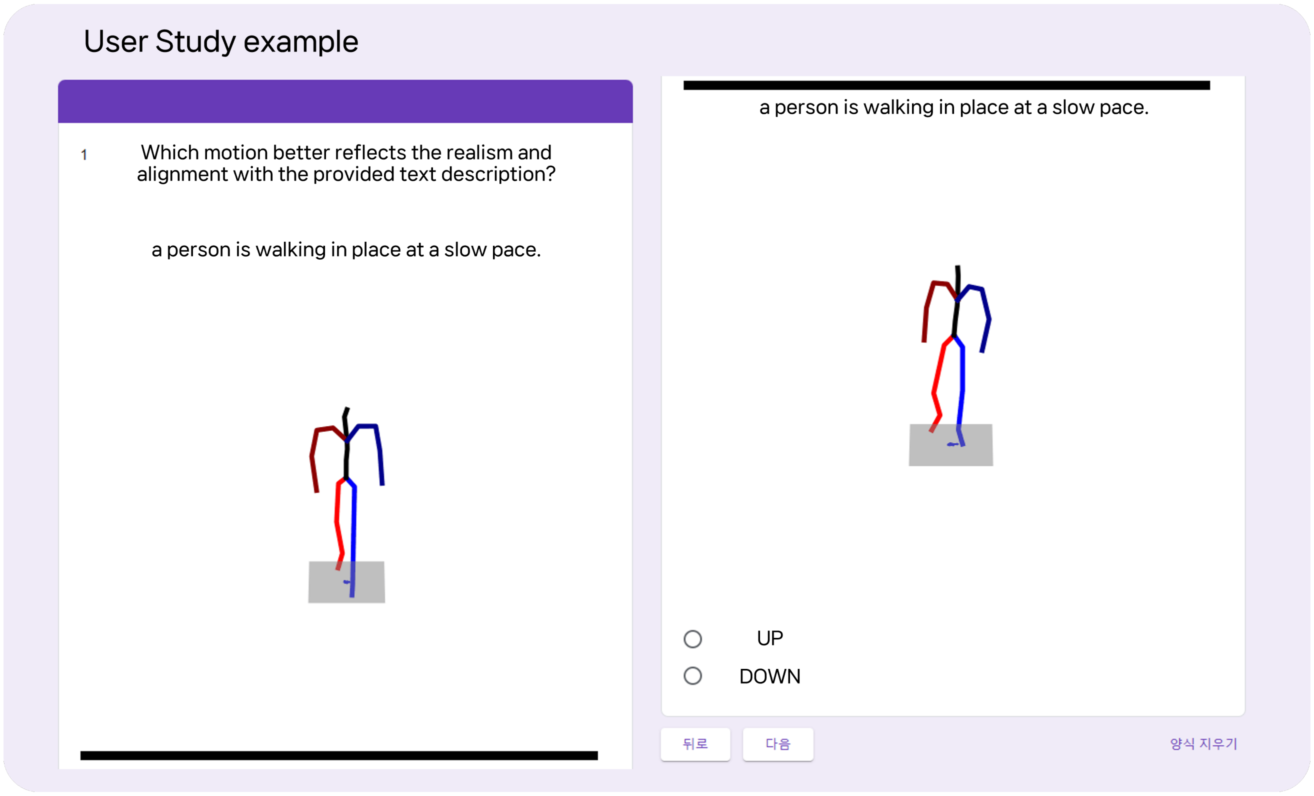
We conducted further experiments on KIT-ML, a well-known dataset in text-to-motion synthesis. The results are presented in the Table 4. As shown, our model, BiPO, achieves state-of-the-art performance, demonstrating its strong generalization capability.

13. User study details

For the user study, we utilize Google Forms. Examples of the survey are shown in Figure 6. We sample 30 motions generated from the same textual prompts in the test dataset. To ensure fairness, the models are anonymized, and their order is randomized for each question, with one model's motion displayed above and the other's below. This approach prevented users from knowing which motion corresponded to which model, allowing for an unbiased evaluation.

14. Further ablation study for PO

We conduct an additional ablation study to determine the optimal masking ratio for PO. To this end, we perform experiments with various masking ratios, and the results are presented in Table 8. Our findings consistently demonstrate that PO enhances model performance across all masking ratios, underscoring its robustness and effectiveness in improving feature learning. This consistent performance gain highlights the adaptability of PO, as it provides benefits regardless of the specific masking ratio used. Moreover, our empirical analysis shows that a 40% and 10% masking ratio



User Study example

1 Which motion better reflects the realism and alignment with the provided text description?

a person is walking in place at a slow pace.

a person is walking in place at a slow pace.

○ UP

○ DOWN

Figure 6. User study example.

yields the best performance on HumanML3D and KIT-ML each, further reinforcing the superiority of PO in optimizing model learning. These results suggest that PO not only contributes to performance improvements but also serves as a reliable and versatile enhancement for a wide range of settings.

15. Discussion of Inference Speed

A limitation of BiPO lies in its inference speed. As an autoregressive model, BiPO inherently relies on sequential generation, which is slower compared to non-autoregressive methods. BiPO, as an autoregressive model, inherently supports motion generation without requiring a predefined motion length. This is a crucial advantage, as demonstrated in BAMM [13], where MoMask [6] exhibits a severe quality drop when evaluated without a given motion length—its FID increasing drastically from 0.045 to 0.09. In contrast, BiPO maintains high-quality results under such conditions, proving its robustness and practical superiority.

Furthermore, BiPO significantly outperforms almost diffusion-based methods in inference speed. Following ParCo [26]'s protocol with a batch size of 100, it takes only 0.062 seconds on an A5000 GPU, and when Dual-iteration Cascaded Part-based Motion Decoding is omitted, the inference time further reduces to 0.043 seconds. This is overwhelmingly faster than diffusion-based approaches such as MDM (1.069 seconds) and MotionDiffuse (1.237 seconds). Given these speeds and its ability to generate motions without predefined lengths while preserving quality, BiPO stands as a highly practical and efficient solution.

In the subsequent analysis, we additionally measured the commonly used Average Inference Time (AIT) and FLOPs against diffusion models to examine the trade-off with FID. The results show that our model achieves an excellent balance in this trade-off, demonstrating both strong performance and suitability for general use. This is visualized in Figure 5.

16. Evaluation on the MARDM Benchmark

Recent studies have insisted that existing feature extractors are biased and therefore do not allow a fair comparison with continuous methods [11]. To ensure fairness, we adopted the feature extractor proposed in that work and directly compared state-of-the-art models on the benchmark. The results demonstrate that our model, BiPO, achieves superior performance over continuous methods. The results can be found in Table 5.

17. Long-Term Motion Generation Evaluation

To validate the long-range coherence of our model, we followed the MotionMamba [24] protocol and conducted additional experiments on HumanML3D using motions longer than 190 frames. Our model, BiPO, achieved the best performance among all methods, demonstrating its strong capability for long-range coherence. These results are presented in Table 6.

18. Evaluation on Motion-X

To demonstrate that our model performs well on larger and more diverse datasets, we conducted additional experiments on Motion-X [8], which is larger and more diverse than HumanML3D. The results confirm that our model maintains strong performance even in such challenging settings. Table 7 provides the corresponding results.

References

[1] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 6, 7

[2] Fernando De la Torre, Jessica Hodgins, Adam Bargteil, Xavier Martin, Justin Macey, Alex Collado, and Pep Beltran. Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. 2009. 7

[3] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 7

[4] 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, 2, 5, 6, 7

[5] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 6

[6] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 6, 7, 9

[7] Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, and Xinchao Wang. Priority-centric human motion generation in discrete latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14806–14816, 2023. 6

[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. 10

[9] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[10] 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. 7

[11] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation. *arXiv preprint arXiv:2411.16575*, 2024. 7, 9

[12] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 6

[13] Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. Bamm: Bidirectional autoregressive motion model. *arXiv preprint arXiv:2403.19435*, 2024. 2, 6, 7, 9

[14] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2024. 6

[15] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 7

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[17] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1

[18] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

[19] Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22035–22044, 2023. 6

[20] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16010–16021, 2023. 6, 7

[21] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 6

[22] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondif-

fuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 6, 7

[23] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems*, 36:13981–13992, 2023. 6

[24] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation. In *European Conference on Computer Vision*, pages 265–282. Springer, 2024. 6, 7, 10

[25] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 509–519, 2023. 6

[26] Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. Parco: Part-coordinating text-to-motion synthesis. *arXiv preprint arXiv:2403.18512*, 2024. 1, 2, 6, 7, 9