Morph: A Motion-free Physics Optimization Framework for Human Motion Generation

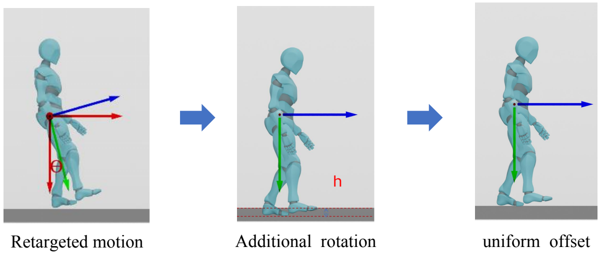
Zhuo Li*[1], Mingshuang Luo*[2,3,4], Ruibing Hou†[2], Xin Zhao[5],
Hao Liu[1], Hong Chang[2,4], Zimo Liu[3], Chen Li[1]
[1]WeChat, Tencent Inc, [2]State Key Laboratory of AI Safety, Institute of Computing Technology, CAS, China
[3]Peng Cheng Laboratory, China, [4]University of Chinese Academy of Sciences, China
[5]MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

Appendix

In this Appendix, we present more details for Morph, including data preprocess, additional experimental results, qualitative comparisons. First, we describe the data preprocessing procedure used for training the Motion Physics Refinement (MPR) module with generated motion data (Sec. A). Then, we present experimental results analyzing the impact of τ in the imitation selection operation (Sec. B), the effect of varying the quantity of noisy motion data for MPR training (Sec. C), and effect of the number of training rounds for Morph (Sec. E). Finally, we provide additional qualitative comparisons for text-to-motion and music-to-dance tasks (Sec. H).

A. Details for Data Preprocess

As discussed in the main text, the generated motion sequences may exhibit issues such as body leaning, floating and ground penetration. When imported into the simulator, these issues can cause instability in the robot, potentially causing it to fall, bounce off the ground, or drop from mid-air. To address this issue, we apply a preprocessing step to the motion sequences, detailed in Fig. 1 and Alg. 1. Specifically, we first compute the body's tilt angle, defined as the angle between the projection of the center of mass onto the ground and the line connecting both feet. If this angle exceeds 10°, we apply the necessary adjustment to the pelvis throughout the sequence. To correct floating and penetration, we determine the lowest mesh height and adjust the entire sequence by this offset. The preprocessed sequence is then used for training and inference.



Retargeted motion

Additional rotation

uniform offset

Figure 1. A flowchart illustrating the data preprocessing process. The parameters are calculated from the first frame and then applied to all generated motion sequences before they are fed into the MPR module.

Algorithm 1 Preprocessing Motion Sequences

```
Require: Motion sequence $S$ with frames $F_1, F_2, \ldots, F_n$
Ensure: Preprocessed motion sequence $S'$
    Step 1: Calculate the angle $\theta$
    (1) Compute the projection of the center of mass of $F_1$ onto the ground.
    (2) Determine the line connecting the pelvis point and the center of both feet in $F_1$.
    (3) Calculate the angle $\theta$ between the projection and the line.
    Step 2: Correct posture if $\theta > 10^\circ$
    (1) Apply an additional rotation to the pelvis for the entire sequence $S$.
    Step 3: Ensure $F_1$ is on the ground
    (1) Infer the lowest point height $h$ of the mesh in $F_1$.
    (2) Add a uniform offset to the entire sequence $S$.
    Step 4: Output the preprocessed sequence $S'$.
```

B. Effect of τ in Imitation Selection on Morph

In Tab. 1, we analyze the effect of the threshold τ in the imitation selection operation on Morph. Different values of τ are tested to assess the performance of Morph-MoMask†

*Equal contribution
†Corresponding author

Table 1. Hyper-parameter analysis of τ in Imitation Selection operation. Comparison with different values of τ based on Morph-MoMask† (combined with MoMask [1] motion generator, without fine-tuning motion generator) for text-to-motion task on HumanML3D dataset. The arrows (↑ / ↓) indicate that higher/smaller values are better.

Methods	Common Generation Metrics				Physical Plausibility Metrics				
	RTOP-1 ↑	RTOP-3 ↑	FID ↓	Diversity ↑	PFC ↓	Penetrate ↓	Float ↓	Skate ↓	IFR ↓
τ=0.0	**0.521**	**0.807**	**0.045**	**9.641**	1.058	23.152	10.660	5.262	-
τ=0.1	0.520	0.806	0.048	9.636	0.877	3.564	4.779	2.128	0.1281
τ=0.2	0.519	0.805	0.056	9.625	0.771	0.838	4.015	1.057	0.0540
τ=0.3	0.518	0.805	0.067	9.583	0.757	0.054	3.200	0.529	0.0258
τ=0.4	0.517	0.803	0.071	9.584	0.722	0.002	2.991	0.211	0.0158
τ=0.5	0.516	0.802	0.074	9.578	0.669	0.000	2.268	0.011	0.0153
τ=0.6	0.512	0.801	0.079	9.576	0.664	0.000	2.263	0.010	0.0144
τ=0.7	0.510	0.799	0.080	9.543	0.660	0.000	2.093	0.006	0.0128
τ=0.8	0.506	0.797	0.081	9.520	0.645	0.000	2.022	0.005	0.0124
τ=0.9	0.504	0.795	0.085	9.408	0.634	0.000	1.985	0.004	0.0117
τ=1.0	0.497	0.793	0.084	9.255	**0.623**	**0.000**	**1.982**	**0.003**	**0.0111**

Table 2. Comparison of text-to-motion with different amounts of noisy motion data training for Morph-MoMask† (combined with MoMask [1] motion generator, without fine-tuning motion generator). N refers to the total number of generated noisy motion data samples, which is three times the amount of the original real training data. D refers to the number of generated motion data used to train the MPR module. We set τ as 0.5 for testing.

Methods	Common Generation Metrics				Physical Plausibility Metrics				
	RTOP-1 ↑	RTOP-3 ↑	FID ↓	Diversity ↑	PFC ↓	Penetrate ↓	Float ↓	Skate ↓	IFR ↓
D=25%N	0.495	0.791	0.087	9.477	0.866	0.120	2.997	0.035	0.0262
D=50%N	0.498	0.795	0.082	9.536	0.815	0.022	2.870	0.023	0.0205
D=75%N	0.512	0.800	0.078	9.569	0.761	0.002	2.429	0.012	0.0178
D=100%N	**0.516**	**0.802**	**0.074**	**9.578**	**0.669**	**0.000**	**2.268**	**0.011**	**0.0153**

Table 3. Comparison of text-to-motion with multi-round optimization of the MPR module and motion generator based on Morph-MoMask. We set τ as 0.5 and use the total number of generated noisy motion data to train.

Methods	Common Generation Metrics				Physical Plausibility Metrics				
	RTOP-1 ↑	RTOP-3 ↑	FID ↓	Diversity ↑	PFC ↓	Penetrate ↓	Float ↓	Skate ↓	IFR ↓
One-Round w/o FT	0.516	0.802	0.074	9.578	0.669	0.000	2.268	0.011	0.0153
One-Round	0.525	0.816	0.041	9.689	0.645	0.000	2.141	0.010	0.0149
Two-Round w/o FT	0.526	0.817	0.041	9.692	0.632	0.000	2.129	0.010	0.0134
Two-Round	0.527	0.818	0.038	9.697	0.625	0.000	2.108	0.007	0.0129
Three-Round w/o FT	0.526	0.821	0.040	9.701	0.621	0.000	2.121	0.009	0.0131
Three-Round	**0.528**	**0.823**	**0.034**	**9.715**	**0.618**	**0.000**	**2.100**	**0.006**	**0.0122**

(combined with MoMask [1] motion generator, without fine-tuning motion generator). When τ is set to 0, the motion refined by the MPR module is not utilized, and Morph directly outputs the results from the motion generator. As τ increases, the physical plausibility metrics improve significantly. However, the generation metrics show a slight decrease due to the inclusion of some incorrectly refined or non-grounded motions at higher thresholds. Larger values of τ incorporate more refined motions, improving the physical plausibility metrics. However, this also increases the acceptance of incorrectly refined motions, leading to a shift in the motion distribution and a corresponding decline in the generation metrics. According to Tab. 1, we observe that $\tau = 0.5$ strikes a balance between generation and physical plausiibility metrics. Therefore, we set τ to 0.5 in this paper.

C. Effect of Varying Amounts of Noisy Motion Data on Morph

In Tab. 2, we investigate the impact of varying amounts of generated motion data on the training of Morph. Different numbers of generated motion data are used to train the MPR module in Morph-MoMask†. As shown in Tab. 2, increasing the amount of training data for the Motion Physics Refinement (MPR) module leads to improvements in both the generation and physical plausibility metrics on the test set. These results indicate that a larger volume of generated motion data enhances the MPR module's ability to better mimic the input motion and produces higher-quality outputs. Conversely, when the MPR module is trained with a smaller dataset, its motion imitation capability diminishes, leading to greater discrepancies between the generated and input motions. This results in a decline in both the generation and physical plausibility metrics. These results further highlight the effective data augmentation capability of our proposed Morph.

Table 4. The win rate of Morph over baselines.

Module	Semantic Consistency	Realism	Physical Plausibility	Fluency
MDM-Morph vs. MDM	84.8%	80.1%	96.6%	85.0%
T2M-GPT-Morph vs. T2M-GPT	87.1%	79.6%	94.4%	81.2%
MoMask-Morph vs. MoMask	90.4%	88.3%	97.5%	80.9%

D. Effect of Multi-Round Optimization of the MPR module and MG on Morph

In Tab. 3, we analyze the effect of multi-round optimization of the Physics Refinement (MPR) module and Motion Generator (MG) on Morph using Morph-MoMask. To further validate the effectiveness of this round-based training approach in enhancing both the MG and the MPR module, we conducted an additional round of training beyond this single-round training described in the main text. This extra round explores the potential for mutual enhancement between the two modules. In Tab. 3, the following terms are defined:

- *One-Round w/o FT*: The first round of training where only the MPR module is trained.
- *One-Round*: The first round of training that includes both training the MPR module and fine-tuning the MG.
- *Two-Round w/o FT*: Training the MPR module again using the motion data generated by the fine-tuned MG from the first round.
- *Two-Round*: Fine-tuning the Motion Generator using the results from *Two-Round w/o FT*.
- *Three-Round w/o FT*: Training the MPR module again using the motion data generated by the fine-tuned MG from the second round.
- *Three-Round*: Fine-tuning the Motion Generator using the results from *Three-Round w/o FT*.

As shown in Tab. 3, in the first round of training, MG improves the performance of MPR module, enhancing the physical quality of its generated motion. The refined motion data from the trained MPR module is then used to fine-tune the MG, boosting its performance further. In the second round, the fine-tuned MG from the first round is used to generate training data for the MPR module (initialized with first-round weights). We observed improvements in *Two-Round w/o FT* compared to *One-Round*, with PFC increasing by 0.013, Float by 0.012, and IFR decreasing, indicating enhanced motion imitation by the MPR module. After fine-tuning the MG once again, *Two-Round* shows improvements in the RTOP-1 and RTOP-3 metrics. The model's generation and physical performance reached their best in Three-Round. These results clearly demonstrate that the MG and MPR modules can mutually enhance each other. Moreover, alternating training between the MG and MPR modules across multiple rounds can further improve the performance of Morph.

E. Semantic Alignment Analysis

As shown in Fig. 2, Morph significantly outperforms MG on alignment metrics. Globally, Morph demonstrates superior realism by closely matching the statistical distribution of real motions—exhibiting similar clustering patterns and range of variation (Left). Crucially, it also achieves stronger semantic alignment with input text features, forming tighter clusters around corresponding text embeddings to better capture intended meanings (Middle). Locally, Morph provides enhanced semantic matching at the segment level, ensuring fine-grained motion elements correspond more accurately to detailed text semantics throughout the sequence (Right). In summary, Morph can generate semantically faithful, realistic motions compared to the MG baseline.

F. Cross-Task Generalization Ability

To evaluate the cross-task generalization of the MPR module, we conducted cross-validation by testing its performance across two distinct tasks: text-to-motion (using the MoMask dataset) and music-to-dance (using Bailando). These tasks differ significantly in input modalities—one driven by linguistic descriptions, the other by rhythmic audio-and in motion characteristics, from daily actions to stylized dance moves. As shown in Tab. 5, the MPR module retains strong performance even when trained without task-specific synthetic data: it not only preserves motion quality (e.g., smooth transitions and natural postures) but also maintains physical plausibility (avoiding joint distortions or gravity-defying movements). This confirms its ability to generalize beyond specific task boundaries.

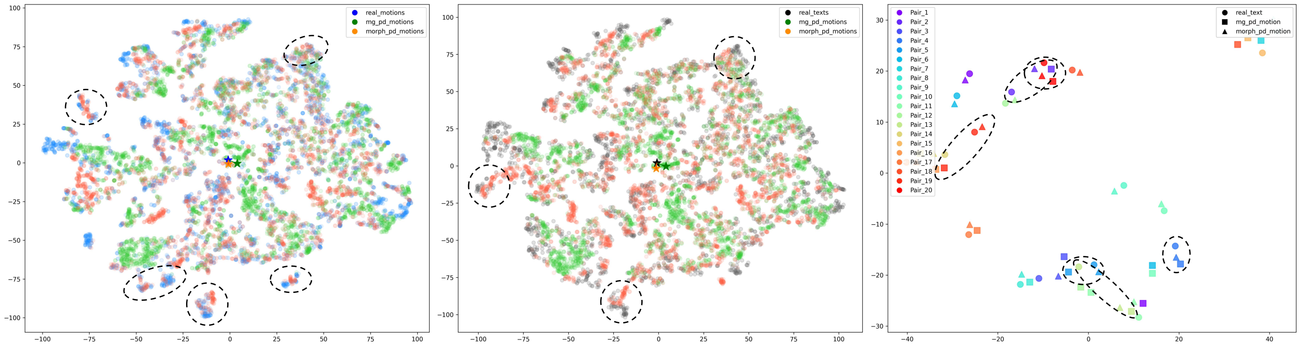


Figure 2. T-sne of motion and text distribution between MG and Morph.

Table 5. Cross-Task generalization results on Music2Dance and Text2Motion

Task	$FID/FID_k \downarrow$	$FID_g \downarrow$	$RTOP3/Div_k \uparrow$	$PFC \downarrow$	$Penetrate \downarrow$	$Float \downarrow$
			Cross-task generalization evaluation			
Text-to-Motion (MPR trained on T2M)	0.074	–	0.802	0.669	0.000	2.268
Text-to-Motion (MPR trained on M2D)	0.116	–	0.795	0.881	0.000	2.436
Music-to-Dance (MPR trained on M2D)	35.48	7.70	12.03	0.044	0.000	2.076
Music-to-Dance (MPR trained on T2M)	37.66	7.38	14.02	0.057	0.000	2.193
			Action-to-motion tests			
Action_Label-to-Motion (MDM-action)	0.497	–	0.396	0.544	15.770	7.467
Action_Label-to-Motion (MDM-action with Morph)	0.424	–	0.416	0.509	0.000	2.115
			GAN-based Transformer tests			
Text-to-Motion (GAN-based Transformer)	0.628	-	0.736	0.933	47.612	21.008
Text-to-Motion (GAN-based Transformer with Morph)	0.606	-	0.755	0.742	0.000	2.637
			Long-duration dance samples tests			
Music-to-Dance (30s long-term dance, Lodge)	45.56	34.29	6.75	0.114	46.772	29.857
Music-to-Dance (30s long-term dance, Lodge-Morph)	43.96	32.88	6.90	0.083	0.000	3.163
			PHC-based baseline			
Text-to-Motion (MoMask+PHC)	0.183	-	0.785	0.749	0.000	2.451
Text-to-Motion (MoMask+Morph)	**0.041**	-	**0.816**	**0.647**	**0.000**	**2.141**

G. User Study

In the user study, we evaluated win rates across four critical dimensions—semantic alignment (matching text descriptions), authenticity (resembling real motions), physical plausibility (avoiding unnatural joint movements), and fluency (smooth temporal transitions). Morph decisively outperformed baseline methods here: as shown in Tab. 4, it achieved significantly higher win rates of 87.4%, 85.2%, 96.2%, and 82.4% respectively. Such consistent leads across all key metrics confirm its comprehensive advantages in motion generation quality.

H. More Qualitative Results

Fig. 3 and Fig. 4 provide the additional qualitative results for the text-to-motion and music-to-dance generation tasks using Morph.

As shown in Fig. 3, in the text-to-motion generation task, floating and penetration are common artifacts in motion generation, often resulting from inaccuracies in the estimation of translation. However, Morph effectively addresses these issues, successfully mimicking the input motion and demonstrating a significant improvement in mitigating these artifacts. The generated motions are both physically plausible and realistic, showcasing Morph's enhanced performance in this task.

As shown in Fig. 4, in the music-to-dance generation task, floating and penetration are the most prominent issues. Due to the faster frequency of dance movements, these artifacts occur more frequently. Morph effectively mitigates these issues, generating motions that are not only physically plausible but also exhibit a higher degree of realism.

In summary, Morph demonstrates significant improvements in both the text-to-motion and music-to-dance tasks. By accurately estimating translational motion, Morph is able to generate motions that are not only physically feasible but also exhibit a higher degree of realism.

References

[1] 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. 2

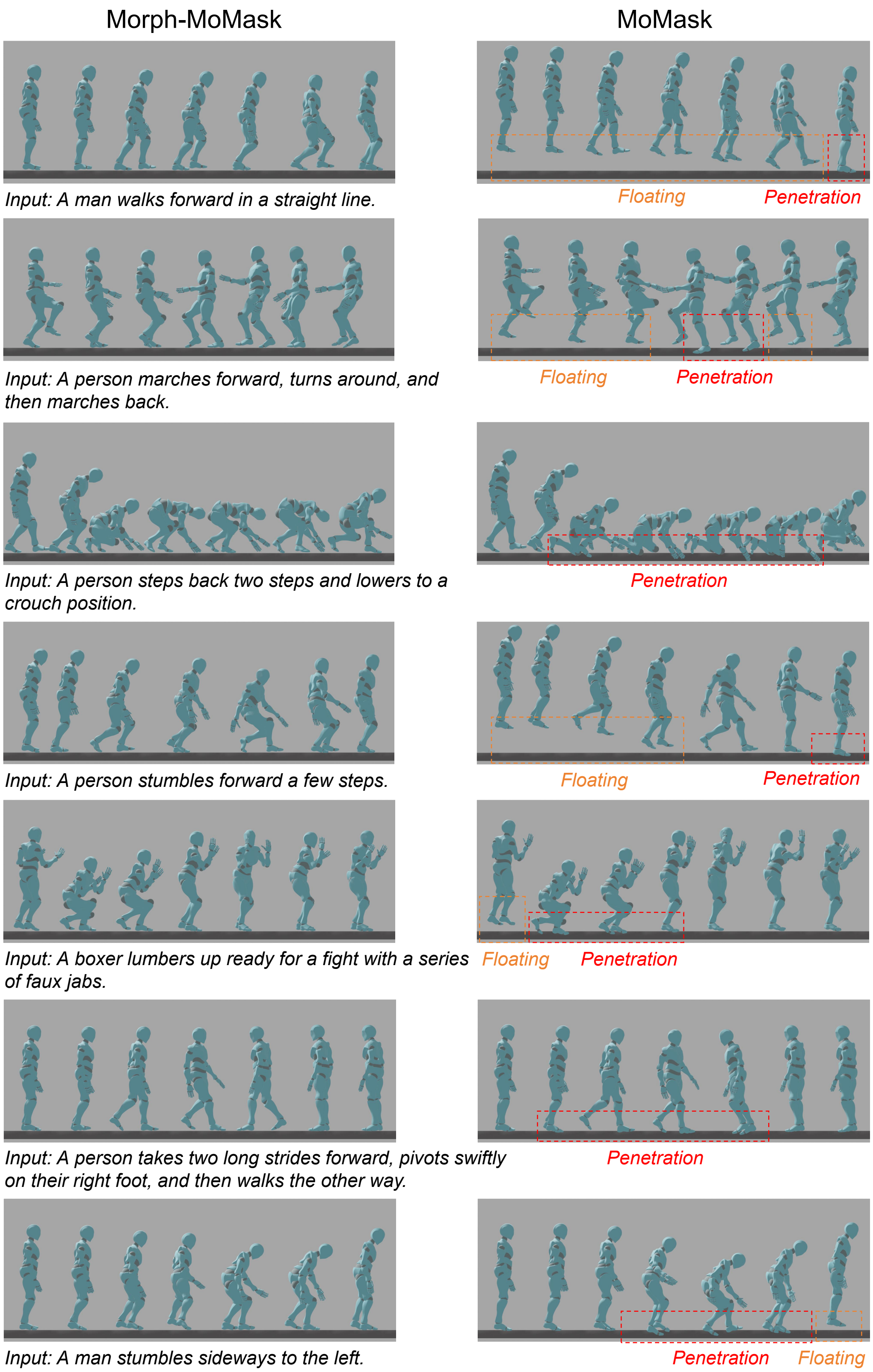

Morph-MoMask
MoMask
Input: A man walks forward in a straight line.
Floating
Penetration
Input: A person marches forward, turns around, and then marches back.
Floating
Penetration
Input: A person steps back two steps and lowers to a crouch position.
Penetration
Input: A person stumbles forward a few steps.
Floating
Penetration
Input: A boxer lumbers up ready for a fight with a series of faux jabs.
Floating
Penetration
Input: A person takes two long strides forward, pivots swiftly on their right foot, and then walks the other way.
Penetration
Input: A man stumbles sideways to the left.
Penetration
Floating

Figure 3. Qualitative comparisons for text-to-motion on HumanML3D test set between Morph-MoMask and MoMask.

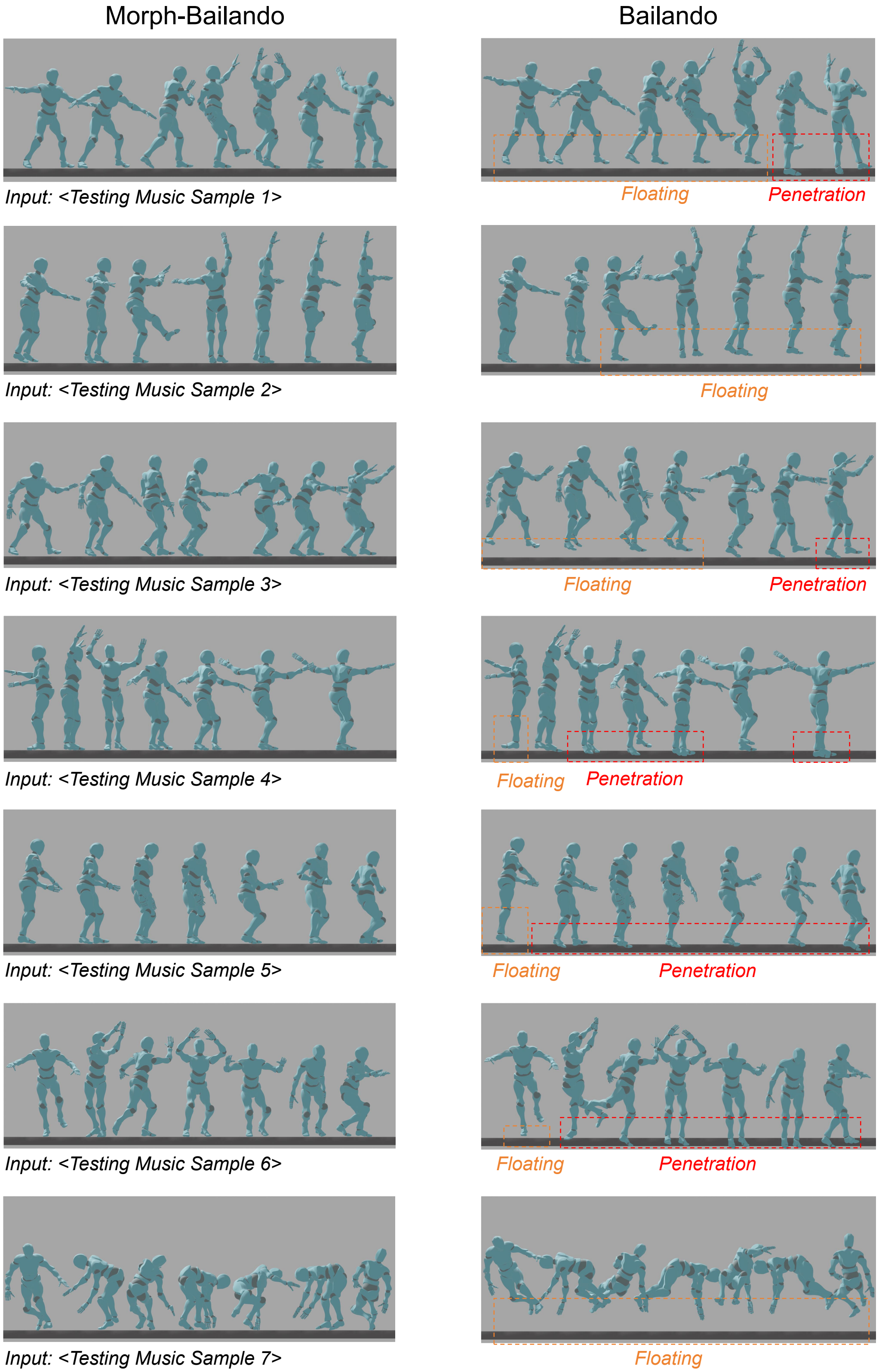

Morph-Bailando
Bailando
Input: <Testing Music Sample 1>
Floating
Penetration
Input: <Testing Music Sample 2>
Floating
Input: <Testing Music Sample 3>
Floating
Penetration
Input: <Testing Music Sample 4>
Floating
Penetration
Input: <Testing Music Sample 5>
Floating
Penetration
Input: <Testing Music Sample 6>
Floating
Penetration
Input: <Testing Music Sample 7>
Floating

Figure 4. Qualitative comparisons for music-to-dance on AIST++ test set between Morph-Bailando and Bailando. For music-to-dance, the testing music samples will be used as inputs.