BridgeDiffusion: Latent Space Optimization for Independent Body-Part Generation with Motion Consistency Bridges in Interactive Dance

Supplementary Material

A. Ablation of CFG

In Table 4, we demonstrate the impact of different classifier-free guidance (CFG) settings on the final generation results. Dance accompaniment is a task involving two distinct modalities—music and motion—as control conditions, requiring the model to balance their respective influences. Using only music as control yields strong alignment with musical beats but results in poor motion quality. Conversely, relying solely on motion control enables the model to effectively capture interactive dynamics but leads to weak synchronization with the musical rhythm. Our approach, as presented in Eq. **??**, effectively combines both modalities, allowing the model to balance their contributions and achieve optimal performance.

B. Computational and Efficiency Comparison

Table 1 presents a comparison of computational resource consumption between our method and Duolando [**?**]. BridgeDiffusion demonstrates significantly lower computational demands across all metrics: it requires only 35.1% of the FLOPs (13.36G vs. 38.1G), 18.7% of the parameters (38.58M vs. 205.98M), 55.6% of the training time (40h vs. 72h), and 32.7% of the inference time (176.08s vs. 537.73s) compared to Duolando. The inference time refers to the duration required to generate a 2040-frame two-person dance sequence. This indicates substantially improved efficiency and reduced resource requirements.

Table 1. Computational and efficiency comparison.

Model	**FLOPs (G)**	**Params (M)**	**Training Time (h)**	**Inference Time (s)**
Duolando	38.1	205.98	72	537.73
BridgeDiffusion	13.36	38.58	40	176.08

C. LSO Stability

Our proposed LSO algorithm depends on the manual setting of hyperparameters τ and σ. In Sec **??**, we state that both τ and σ must be greater than zero to ensure the model can determine the optimization direction based on the sign of the score. In practice, the LSO algorithm exhibits strong robustness to hyperparameter settings and does not require specific values to achieve excellent performance. As shown in Table 2, the model consistently converges to good performance when both parameters are positive. Convergence difficulties occur only when the values differ significantly (e.g., τ=1, σ=100). Furthermore, Table 2 demonstrates that our LSO algorithm requires minimal convergence time. This efficiency arises because the algorithm fine-tunes pre-trained model parameters rather than training from scratch.

Table 2. Experimental results for LSO stability.

(τ, σ)	$\text{FID}_k(\downarrow)$	$\text{FID}_g(\downarrow)$	$\text{FID}_{cd}(\downarrow)$	**Convergence Time**
(1, 1.5)	**15.23**	27.37	**3.82**	1.649h
(1, 3.0)	16.89	28.35	4.91	1.424h
(2, 1.5)	15.78	**26.70**	3.91	2.951h
(2, 3.0)	17.23	27.31	4.91	1.874h

D. Body Part Segmentation

Our BridgeDiffusion partitions the human body into four segments—upper body, lower body, left hand, and right hand—for motion generation. This division aligns with natural motion patterns: joints within each segment typically exhibit coherent movement dynamics, while cross-segment pairs (e.g., left hand and right leg) often display divergent motion behaviors. Empirical results further validate this design choice, as demonstrated in Table 3, where this partitioning strategy achieves optimal performance across all evaluation metrics.

Table 3. Body part segmentation performance comparison.

Body Part Segmentation	$\text{FID}_k(\downarrow)$	$\text{FID}_g(\downarrow)$	$\text{FID}_{cd}(\downarrow)$
One part	60.88	49.54	142.3
Two part (up, down)	38.75	31.92	57.01
Four part (up, down, lhand, rhand)	**15.23**	**27.37**	**3.82**

E. User Study

we created a user study where 24 participants watched 68 pairs of videos. Each pair of videos contains two dance sequences: one generated by our method DanceRLDiff, and the other generated by Duolando. Participants were asked to judge which video appeared more impressive, without knowing the source of either video. In the 72.9% scenario, our method is considered to have a better visual effect than Duolando. This capability stems from our design, which enables BridgeDiffusion to generate movements with diverse motion patterns and anatomically plausible body structures. These qualities make our results more impressive.

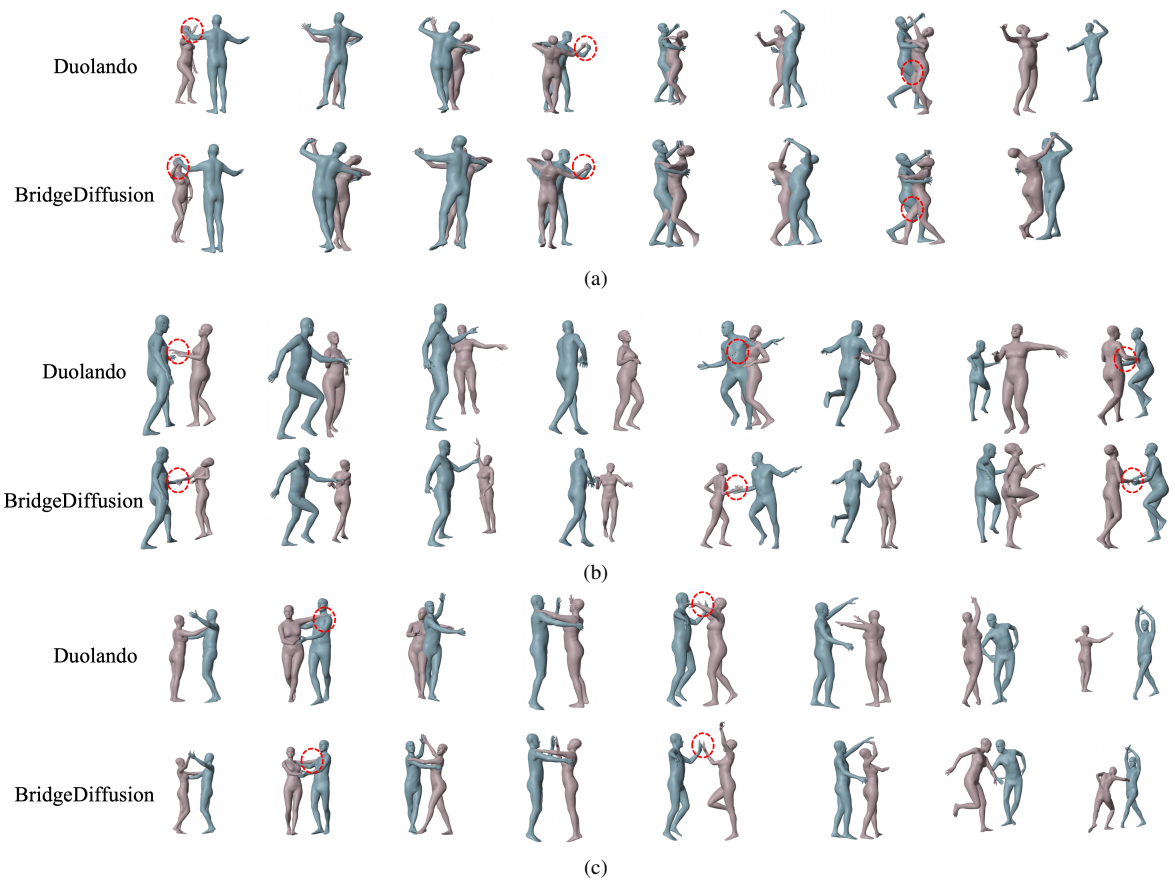

Figure 1. Qualitative comparison with Duolando [**?**] in different types of dance. (a) Foxtrot (b) Jive (c) Pasodoble

Table 4. Ablation study for different Classifier-Free Guidance Setting. **Bold** and underline indicate the best and the second best result.

Method	Solo Metrics				Interactive Metrics				Rhythmic
	$FID_k(\downarrow)$	$FID_g(\downarrow)$	$Div_k(\uparrow)$	$Div_g(\uparrow)$	$FID_{cd}(\downarrow)$	$Div_{cd}(\uparrow)$	CF(%)	BED(↑)	BAS(↑)
Ground Truth	6.56	6.37	11.31	7.61	3.41	12.35	74.25	0.5308	0.1839
BridgeDiffusion w. CFG-all	**15.23**	**27.37**	**11.67**	**7.65**	**3.82**	**9.54**	**54.69**	**0.3191**	0.2072
full-diffusion w. CFG-music only	40.17	30.55	7.94	6.39	16.63	6.13	43.08	0.2554	**0.2241**
full-diffusion w. CFG-leader only	23.49	28.96	10.11	6.90	13.98	7.77	46.82	0.2859	0.2200
full-diffusion w/o. CFG	56.67	28.44	6.83	6.27	28.71	7.33	41.23	0.2555	0.2185