ReactDance: Hierarchical Representation for High-Fidelity and Coherent Long-Form Reactive Dance Generation

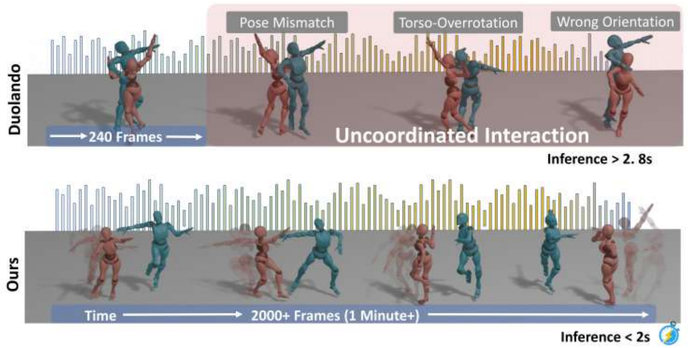
Jingzhong Lin[1] , Xinru Li[1] , Yuanyuan Qi[1] , Bohao Zhang[1] , Wenxiang Liu[1] , Kecheng Tang[1] ,Wenxuan Huang[1] , Xiangfeng Xu[1] , Bangyan Li[1] , Changbo Wang[1] , Gaoqi He[1,✉]

[1]East China Normal University

ICLR

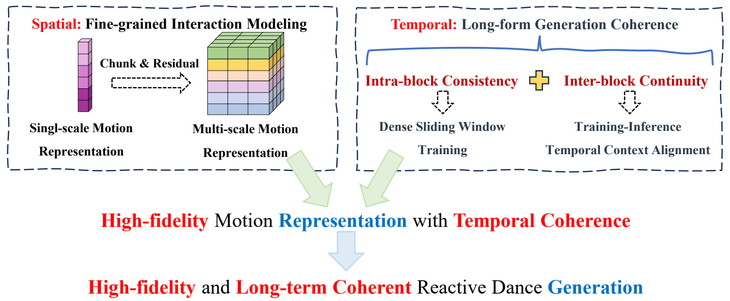
Motivation

- **Spatial Incoherence:** Single-scale representations (like standard VQ-VAE) fail to distinguish macro-postures from high-frequency joint details, causing uncoordinated interactions and inter-penetrations.
- **Temporal Drift:** Autoregressive sampling (serial clip-by-clip stitching) suffers from error accumulation at boundaries. When generating long sequences, this drift inevitably leads to jitter collapse.

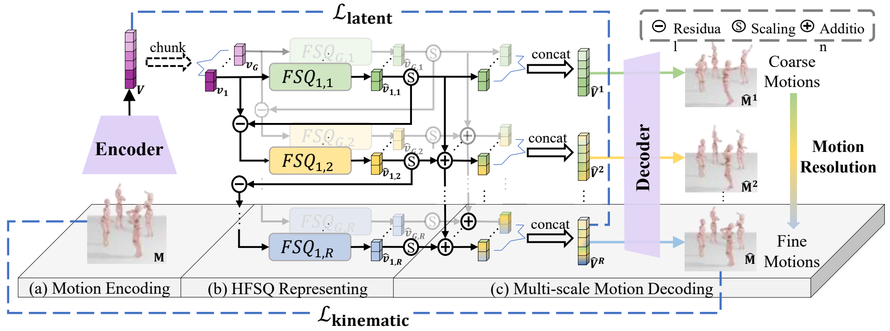


Main Idea

- **Spatial:** A **Hierarchical Latent Space (HFSQ)** explicitly disentangles coarse body rhythms from fine-grained details for precise interaction.
- **Temporal: Blockwise Local Context (BLC)** structurally eliminates drift by strictly aligning train-inference temporal contexts, enabling robust, **parallel synthesis** of long sequences.

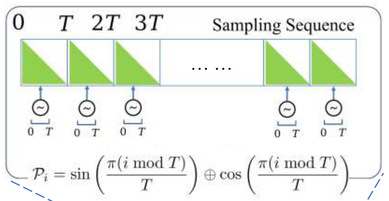


HFSQ (Hierarchical Finite Scalar Quantization)

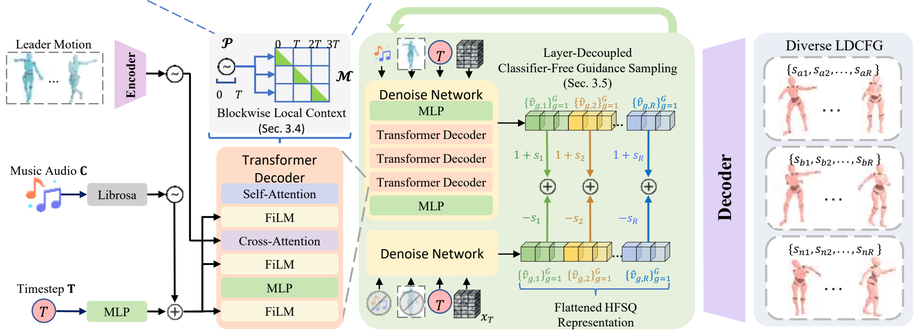


- **Mechanism:** Projects motion into a **continuous scalar grid** over residual stages.
- **The Intuition:** The base layer captures low-frequency global posture; residual layers encode high-frequency details. This "diffusion-friendly" continuous manifold avoids the codebook collapse inherent in VQ-VAEs.

BLC (Blockwise Local Context) Sampling



- **Training Prior:** Dense Sliding Window (DSW) forces the decoder to learn phase-agnostic, smooth boundary transitions.
- **Inference Execution:** Replaces autoregressive generation. Uses Periodic Causal Attention Masking (PCAM) and Phase-aligned Positional Encodings (PPE) to perfectly match the training context, synthesizing all blocks in parallel.



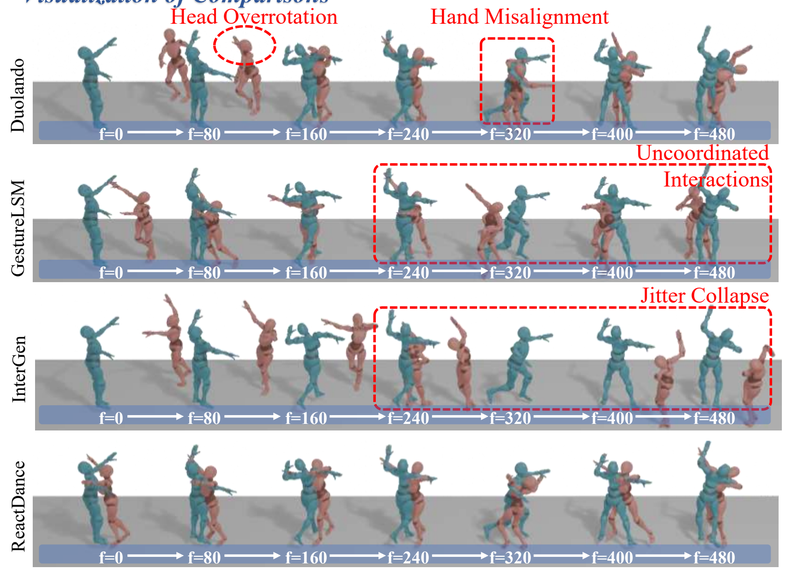
Experiments

Evaluation on DD100 benchmark

Table 1: ReactDance consistently outperforms SOTA methods across both single-person (above dotted line) and duet/multi-person (below dotted line) generation tasks.

Method	Solo Metrics								Interactive Metrics				
	$FID_k(\downarrow)$	$FID_g(\downarrow)$	$Div_k(\rightarrow)$	$Div_g(\rightarrow)$	MPJPE(↓)	MPJVE(↓)	PFC(↓)	BAS(→)	$FID_{cd}(\downarrow)$	$Div_{cd}(\rightarrow)$	BED(→)	IPR(↓)	AITS(↓)
Ground Truth	-	-	10.86	7.82	-	-	-	0.1791	-	12.53	0.5308	-	-
GestureLSM	14.65	34.23	12.20	9.00	171.37	19.01	0.7903	**0.1734**	40.09	9.45	0.1903	19.01	3.13
EDGE	68.68	113.25	5.62	10.05	235.52	20.76	0.6226	0.1854	1523.27	**12.37**	0.1904	8.44	2.91
TCDiff	105.97	251.35	14.23	24.54	182.64	22.91	1.1165	0.2270	1472.00	11.31	0.2304	**7.58**	3.34
InterGen	35.89	47.28	12.24	9.13	210.66	18.69	0.9842	0.1634	176.19	18.10	0.2746	17.58	5.22
Duolando	27.68	35.01	10.95	8.70	174.54	18.72	0.9276	0.2086	17.49	14.73	0.3285	17.42	4.41
ReactDance	**5.57**	**7.63**	**10.82**	**7.76**	**132.99**	**15.68**	**0.6039**	0.2031	**14.17**	10.58	**0.3863**	7.84	**1.75**

Visualization of Comparisons



Paper & Project & Code

Our complete codebase—including training, inference, and visualization—is publicly available on GitHub.

Paper Project Code