MDD: A Dataset for Text-and-Music Conditioned Duet Dance Generation

Prerit Gupta, Jason Alexander Fotso-Puepi, Zhengyuan Li, Jay Mehta, Aniket Bera

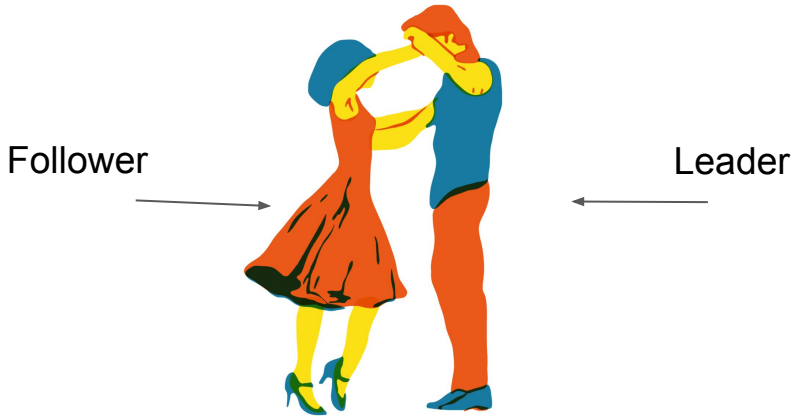
IDEAS Lab, Purdue University, West Lafayette

Background

Computational Challenges

- Complex Spatial relationships
- Real-time partner interactions
- Semantic Understanding of moves
- Multimodal synchronization

Duet Dancing



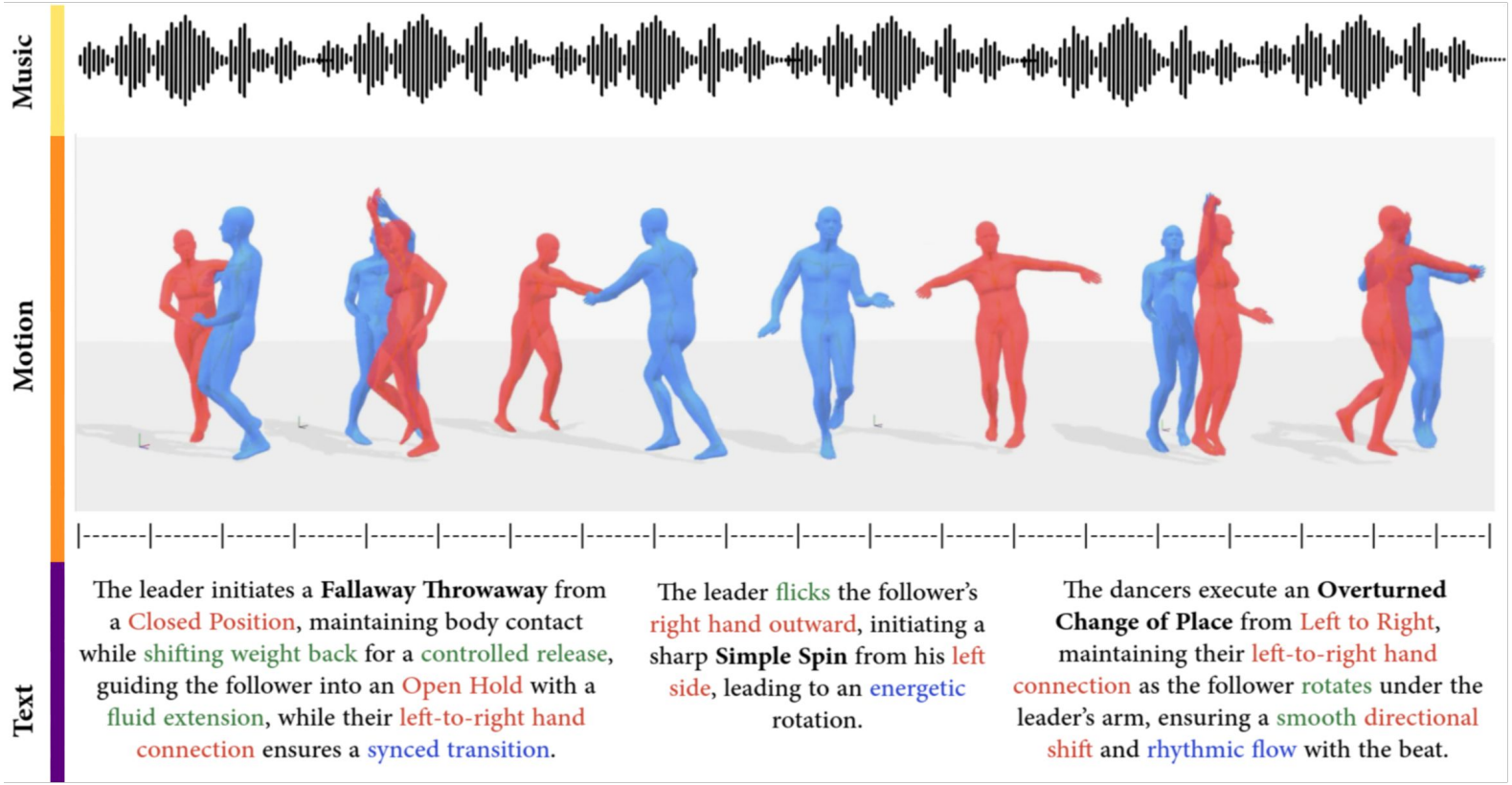
Motivation

Existing interactive motion datasets: music-motion or text-motion only

No large-scale duet dancing dataset

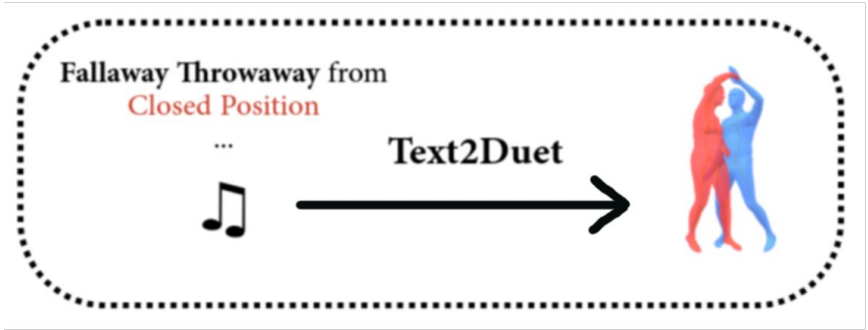
No descriptions using rich dance movement vocabulary

Multi-modal Duet Dance (MDD) Dataset



Tasks

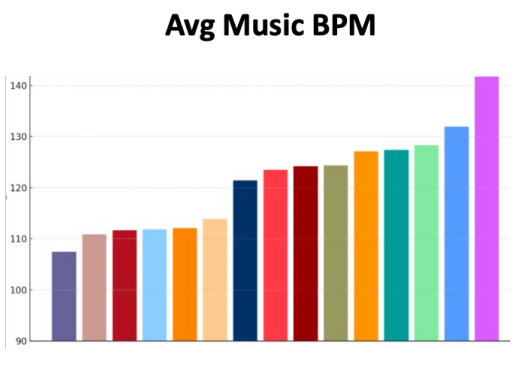
(1) Interactive

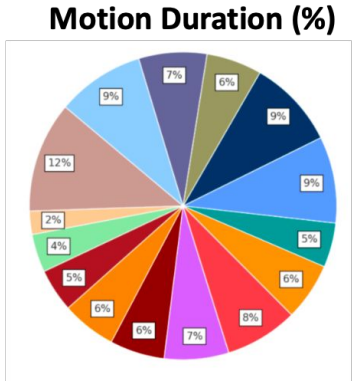


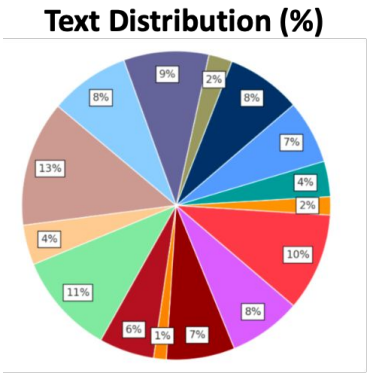
(2) Reactive

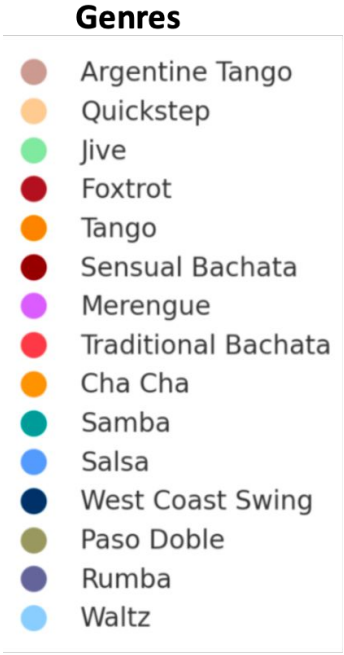
Dataset Statistics

- **Motion:** 10.34h Mocap, 15 genres (Ballroom, Latin, Social)
- **Music:** Diverse genres
- **Text:** 10K+ fine grained descriptions









Dataset Collection

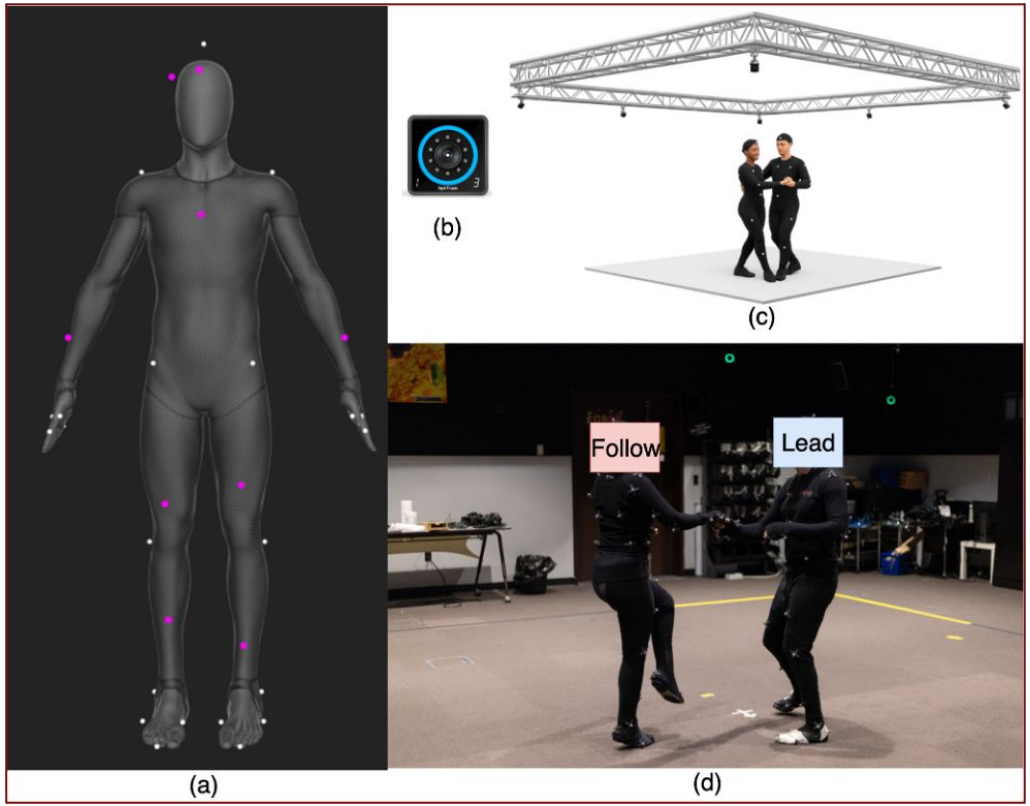
(1) Music Selection:

- 50-60 unique samples per genre
- Mostly copyright-free music

Dataset Collection

(2) Motion Capture:

- 120 fps
- 24 x 24 ft. facility
- OptiTrack MoCap
- 16 infrared cameras
- 53 markers
- 30 subjects (16F/14M)
- ≥3 yr dance exp.

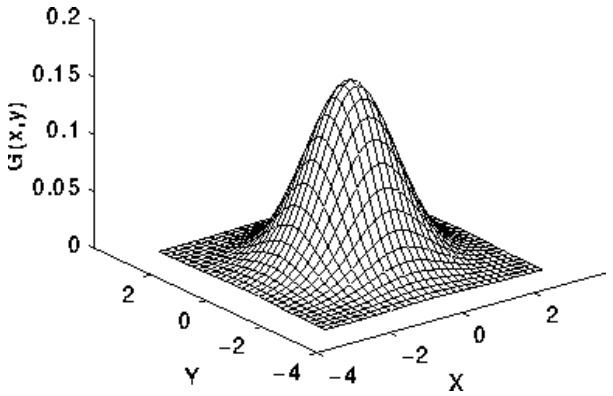


Motion Capture setup

Dataset Collection

(3) Motion post-processing:

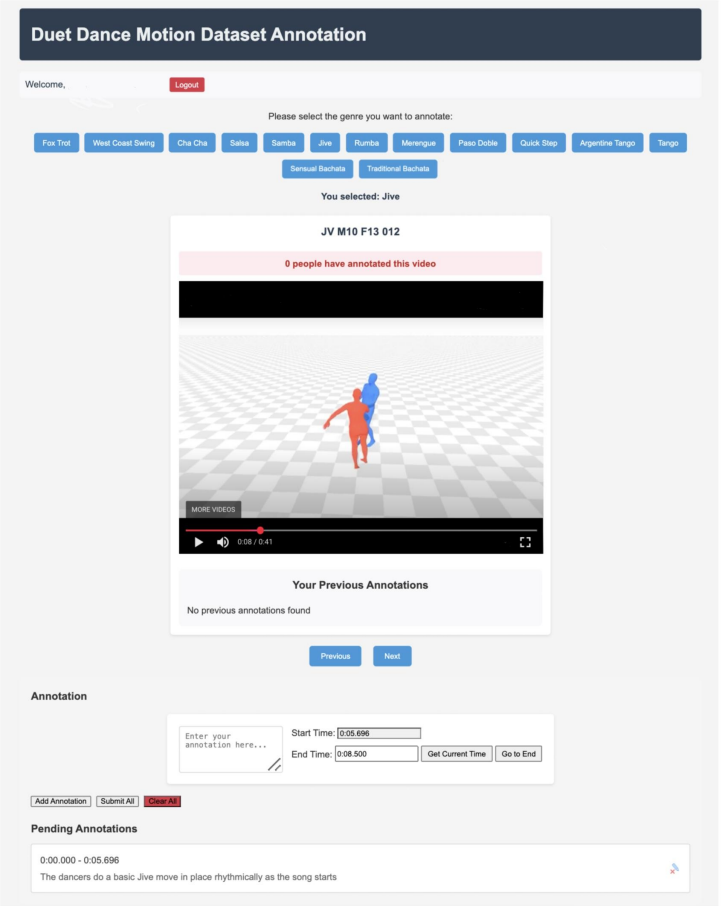
- Outlier removal
- Gaussian filtering
- Zero-pose vector removal using interpolation
- Block-aware blending



Dataset Collection

(4) Annotating Descriptions:

- Designed user-friendly annotation tool
- Professional dancers recruited
- Each segmented sample (< 10s)



Dataset Collection

(5) Annotation Refining

Videos samples for each genre

Argentine Tango

The dancers remain in a closed embrace, where the leader initiates a **Forward-backward swinging** move with the left leg, establishing a subtle yet expressive cadence while the follower attentively mirrors the movement, maintaining a delicate balance between anticipation and responsiveness with the rhythm of the music.

Experimental Results

Adapted Baselines:

- MDM [1]
- InterGen [2]
- DuoLando [3]

Quantitative Evaluation

Text-to-Duet									
Methods	R-Precision ↑			FID ↓	MM Dist ↓	Diversity →	Mmodality ↑	BED ↑	BAS ↑
	Top 1	Top 2	Top 3						
Ground Truth	0.231	0.398	0.522	0.065	0.077	1.387		0.327	0.17
MDM (text-only)	0.082	0.124	0.192	1.42	2.133	1.216	0.811	0.211	0.186
MDM (music-only)	0.041	0.102	0.135	2.241	2.471	1.192	0.411	0.21	0.192
MDM (both)	0.061	0.108	0.163	1.739	2.244	1.235	0.787	0.194	**0.231**
InterGen (text-only)	0.113	0.223	0.305	**0.405**	1.462	1.405	1.231	0.422	0.194
InterGen (music-only)	0.023	0.067	0.088	2.014	2.526	1.3	**1.768**	0.364	0.163
InterGen (both)	0.105	0.206	0.302	0.426	1.532	1.38	1.352	0.385	0.185
InterGen w. Jukebox (both)	**0.138**	**0.245**	**0.341**	0.41	**1.396**	**1.388**	1.33	**0.454**	0.184

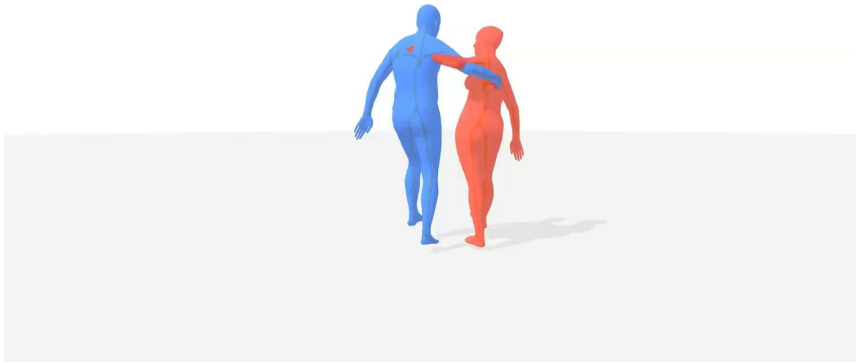
Quantitative Evaluation

Text-to-Dance Accompaniment								
Methods	R-Precision ↑			FID ↓	MM Dist ↓	Diversity →	BED ↑	BAS ↑
	Top 1	Top 2	Top 3					
Ground Truth	0.231	0.398	0.522	0.065	0.077	1.387	0.327	0.17
DuoLando (text-only)	0.047	0.121	0.182	1.538	2.811	1.422	0.311	0.195
DuoLando (music-only)	0.069	0.141	0.202	0.721	2.633	**1.39**	0.305	0.216
DuoLando (both)	**0.078**	**0.156**	**0.219**	**0.698**	**2.113**	1.371	**0.395**	**0.224**

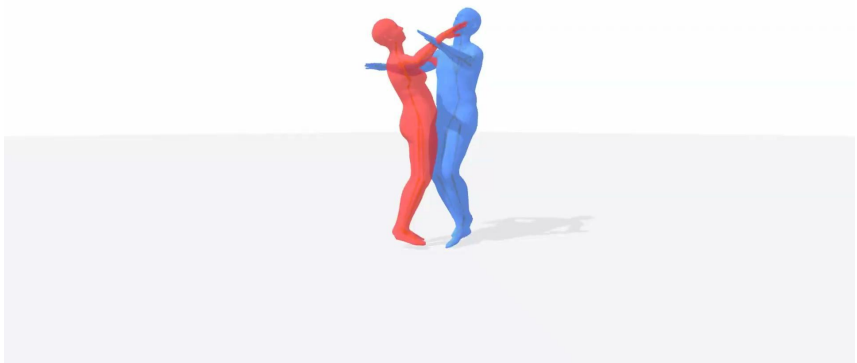
Qualitative Evaluation

Task: Text-to-Duet

Jive



Ground Truth



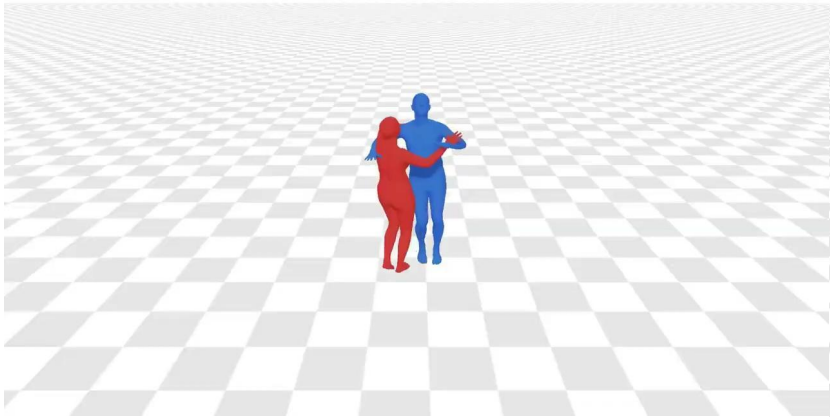
Generated (InterGen)

The dancers are executing a **Jive Mooch** where at first the leader's right hand is connected to the follower's left and afterwards they switch the connection to left to right once they change positions swiftly and travel across the dance floor fluidly. Here in the first part, the dancers are facing away from the camera when they start, and later they turn towards the camera.

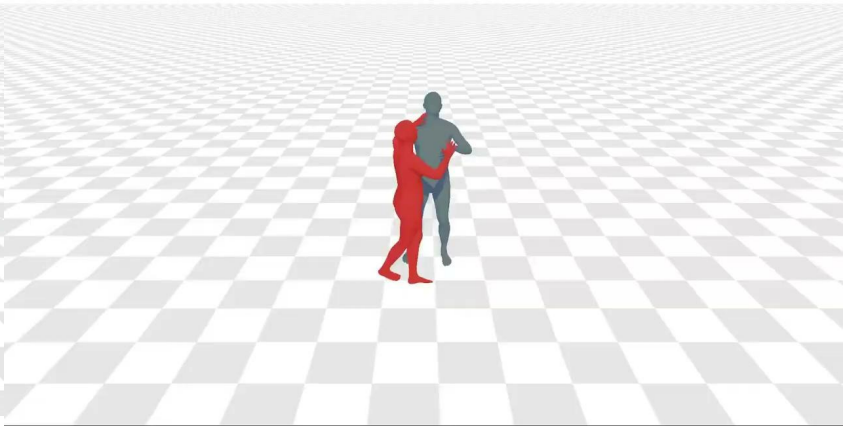
Qualitative Evaluation

Task: Text-to-Dance Accompaniment

Argentine Tango



Ground Truth



Generated (DuoLando)

The leader initiates a **box step** in a cross system, guiding the follower to mirror his movements.
The follower transitions to a parallel system with a cross.

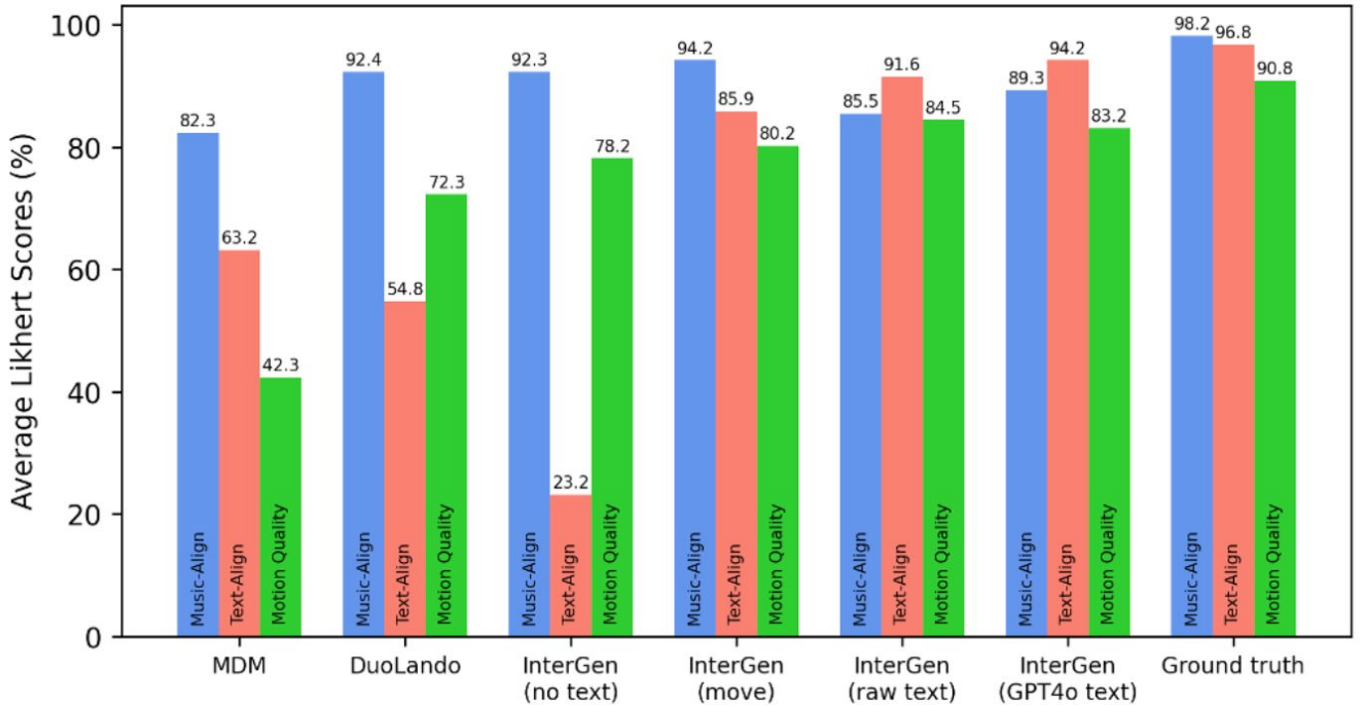
User Study

(Q1) Which motion better aligns semantically with the textual description?

(Q2) Which motion is better synchronized with the musical beats?

(Q3) Which motion has higher overall quality (e.g., naturalness, smoothness)?

User Study Results

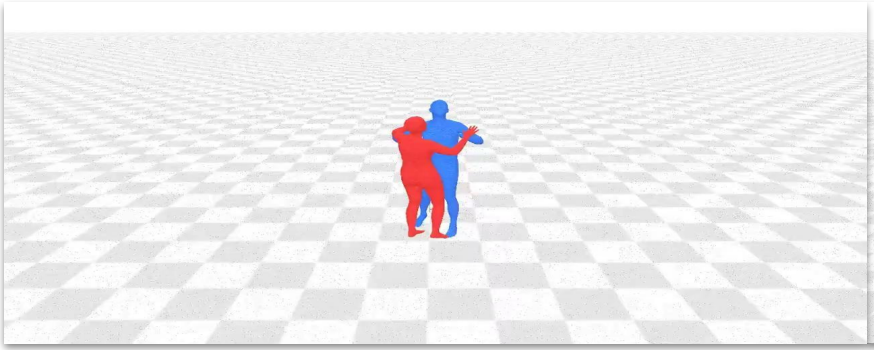


Generalization Experiments w/ Text & Music

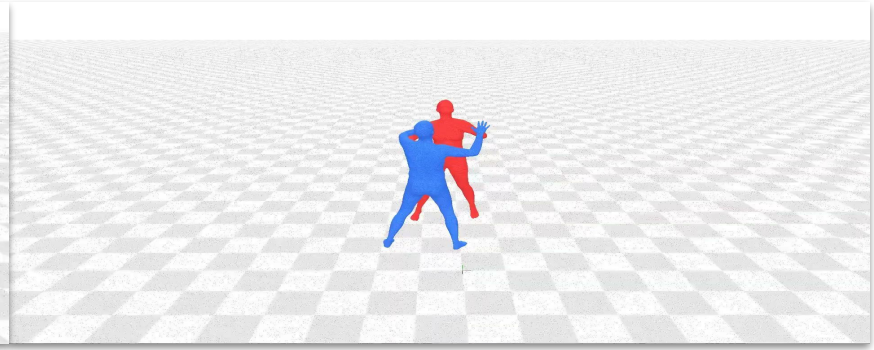
(Intergen)

Same Genre Same Music Different Text

Argentine Tango



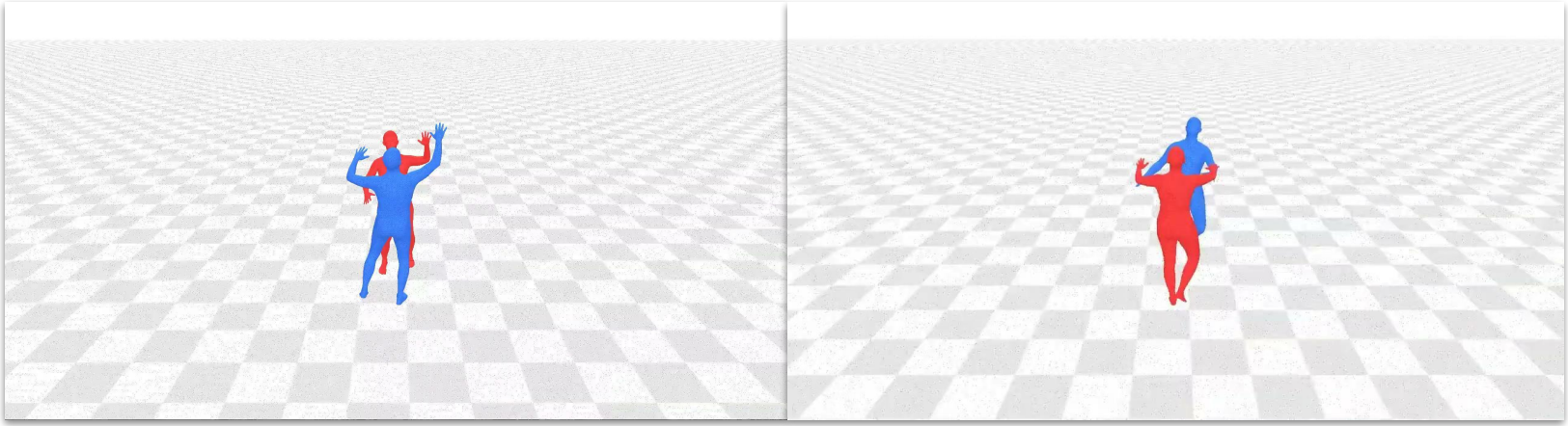
While in open embrace, the leader steps to the side with his left leg and positions his right leg forward in between the followers legs. He then gently rotates the torso of the follower using the open side of the embrace along with the cadence of music, and the follower gives him a **Gancho** with her right leg with all her weight on the left leg. Then she swings the right leg back drawing a half circle and brings it back forward while slightly lifting it.



The dancers initiate a Basic Backward step on the sharp beat, with the leader **stepping backward** with the left foot in an Open Embrace, establishing a strong yet fluid connection while the follower responds by **stepping backward** with the right foot, mirroring the leader's movement while maintaining a tight, responsive frame. The leader takes two steps back and one step to the left side while the follower mirrors him in open embrace.

Same Genre Same Music Different Text

Jive

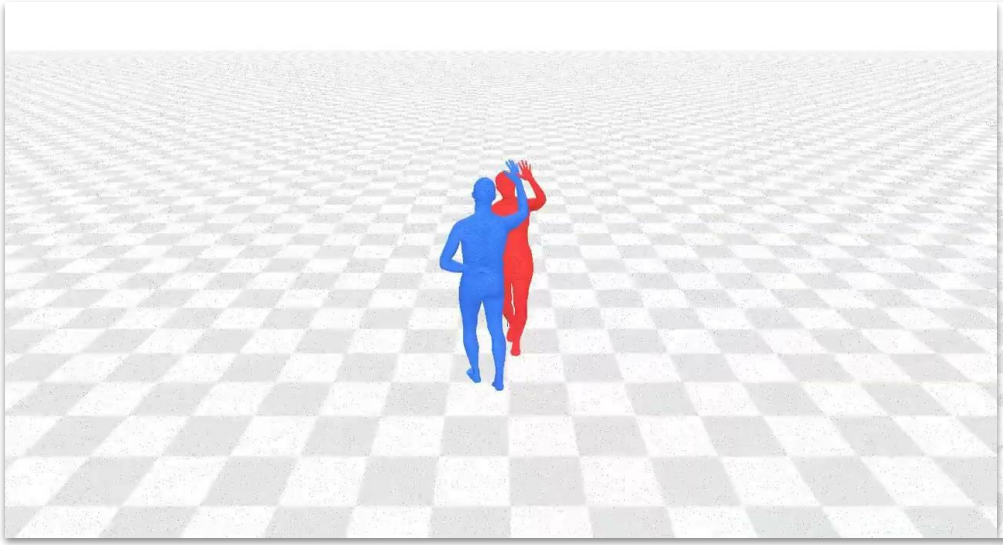


The leader initiates a **Fallaway Throwaway** from a Closed Position, maintaining body contact while shifting weight back for a controlled release, guiding the follower into an Open Hold with a fluid extension, while their left-to-right hand connection ensures a synced transition.

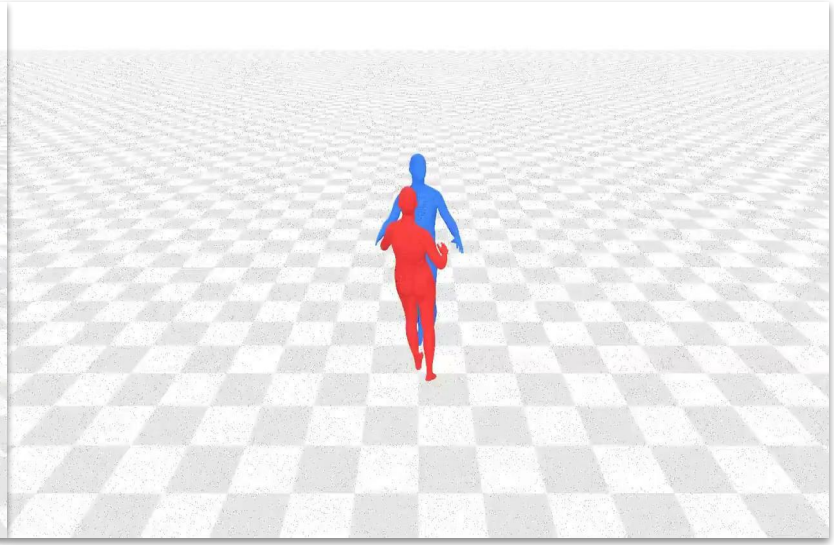
In Closed Position, while both their hands are connected, they execute the **Heel Toe Step** alternating their weight from left to right with the rhythmic flow of the beats.

Same Genre Same Music Subtle Different Text

Salsa



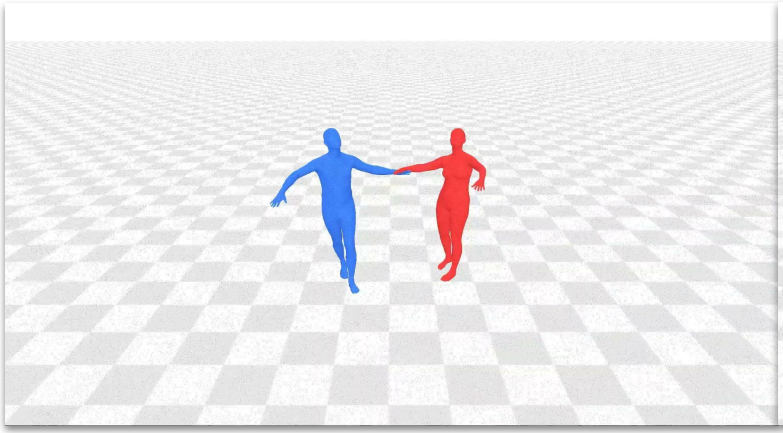
The dancers execute a **Forward Basic** step in the first half of the phrase, maintaining an open hold with hand-to-hand connection while the leader's right hand and the follower's left hand is connected and raised as an indication of preparation of a turn, ensuring a smooth flow.



The dancers execute a **Forward Basic** step in the first half of the phrase, maintaining an open hold with hand-to-hand connection, ensuring a smooth flow

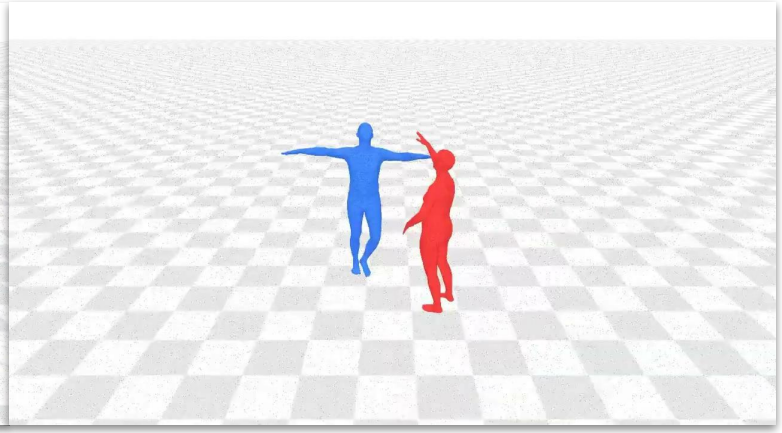
Text From Different Genre Same Music

Music: Jive



Cha Cha

The leader gives the follower a **Simple Spin**, initiating outwards from a left-to-right hand connection, as the follower rotates under the leader's arm, ensuring a smooth directional shift and rhythmic flow with the beat.

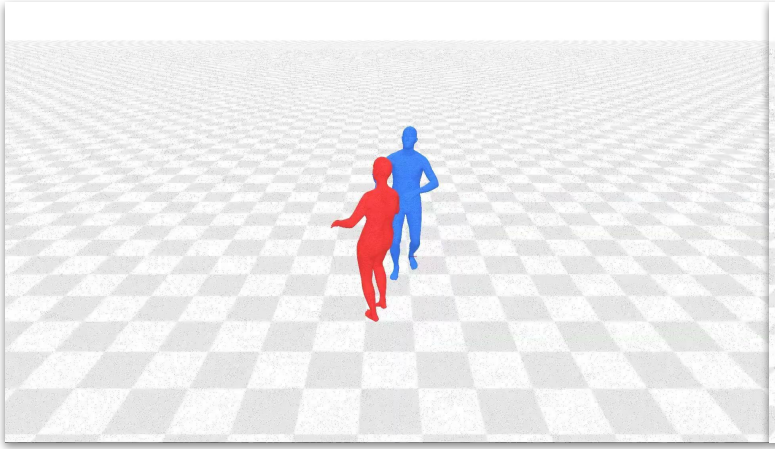


Jive

The dancers execute an **Overturned Change of Place** from Left to Right, maintaining their left-to-right hand connection. as they execute a crisp rapid rotation, and immediately gives her two more energetic inward spins. 8

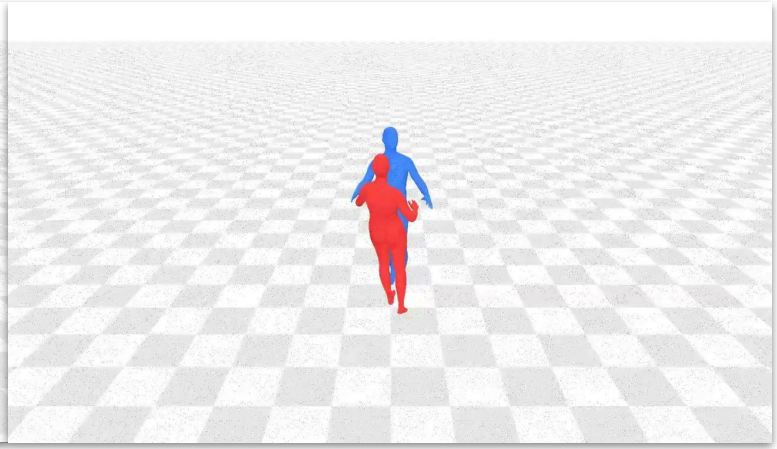
Text From Different Genre Same Music

(Music: Salsa)



West Coast Swing

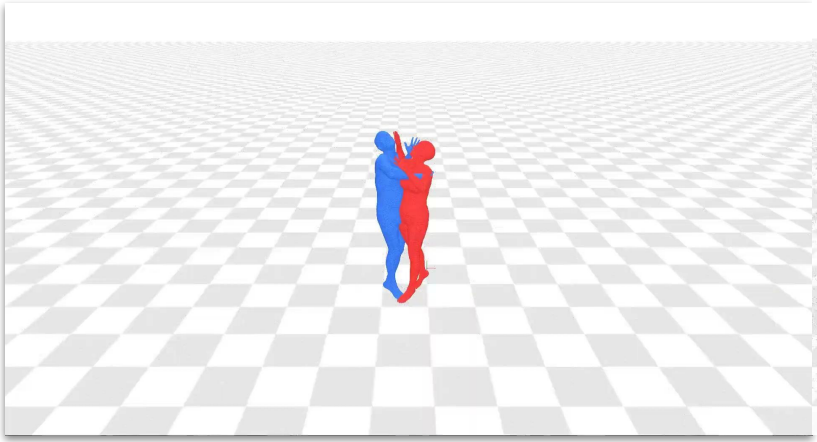
The leader guides the follower into a **crossbody forward walk** along with an **inside turn** which he leads using his right hand which is connected with the follower's right hand during the mambo part of the music.



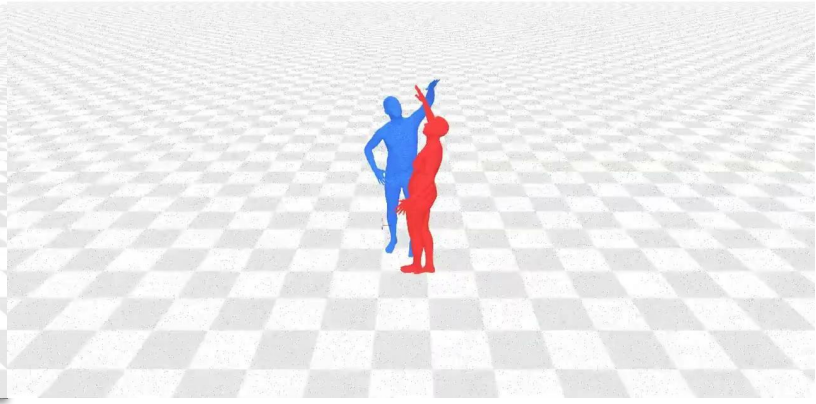
Salsa

The dancers execute a **Forward-backward basic**, maintaining the open hold where the leader's right hand and the followers left hand are connected and extended outward, they are stepping with the fast beats ensuring smooth flow.

Music From Different Genre Same Text



Jive Music



Rumba

Rumba Text for both: The follower executes a simple turn slowly and then another wider turn before reaching the left side of the follower

Conclusion

- **Introduced MDD:** Large-scale, richly annotated dataset for text-driven multi-agent dance generation.
- **620 minutes, 15 genres, and 10K+ detailed descriptions:** Surpassing prior datasets in scale and depth.
- **Multimodality:** Uniquely combines paired motion, music alignment, and fine-grained text, enabling advances in interactive animation, automated choreography, and human interaction modeling.
- **Established two benchmark tasks:** Text-to-Duet and Text-to-Dance Accompaniment to drive future research.

References

[1] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In The Eleventh International Conference on Learning Representations, 2023

[2] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. International Journal of Computer Vision, 132(9):3463–3483, 2024.

[3] Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. arXiv preprint arXiv:2403.18811, 2024

Thank You