dfki Deutsches Forschungszentrum für Künstliche Intelligenz German Research Center for Artificial Intelligence
SIC Saarland Informatics Campus
MAX PLANCK INSTITUTE FOR INFORMATICS
Snap Inc.

DuetGen: Music Driven Two-Person Dance Generation via Hierarchical Masked Modeling

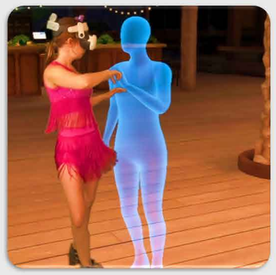
Anindita Ghosh, Bing Zhou, Rishabh Dabral, Jian Wang, Vladislav Golyanik, Christian Theobalt, Philipp Slusallek, Chuan Guo

Main Contributions

- The first framework to generate sequences of **two-person dance motion with close interactions**, conditioned on music signals
- A technique to model two-person dance motions using
  - **a unified two-person representation**
  - **multi-scale motion quantization**
  - **two-stage generative masked transformers**

Applications

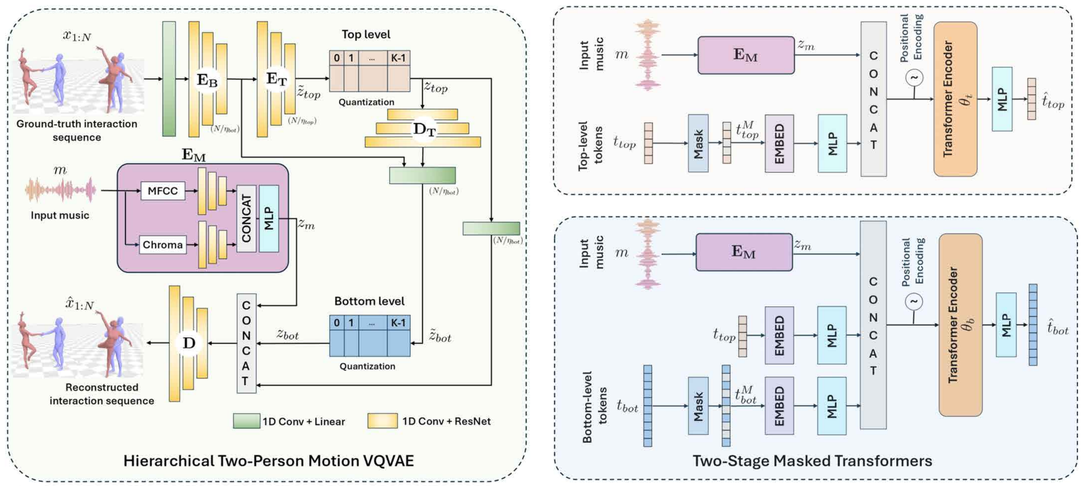
- **Games and Animation**
- **Virtual Reality**
- **Interactive Education**



Acknowledgements. *This research was supported by Snap Inc., the EU Horizon 2020 grant Carousel+ (101017779), and the BMBF grant MOMENTUM (01IW22001)*

Training Framework

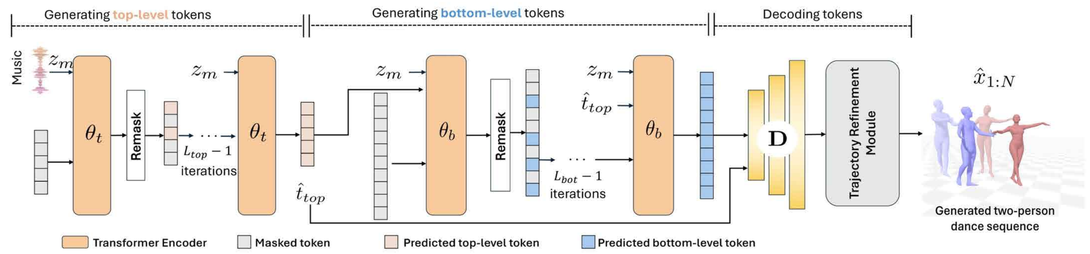
- Encode movements into **two levels of tokens** representing global semantics and finer details
- The first transformer generates **top-level tokens from music**
- The second transformer, **conditioned on both music and the top-level tokens**, learns to **generate bottom-level tokens**



DuetGen Training Framework. Left: Hierarchical two-person motion VQ-VAE. Right: Two-stage masked transformer

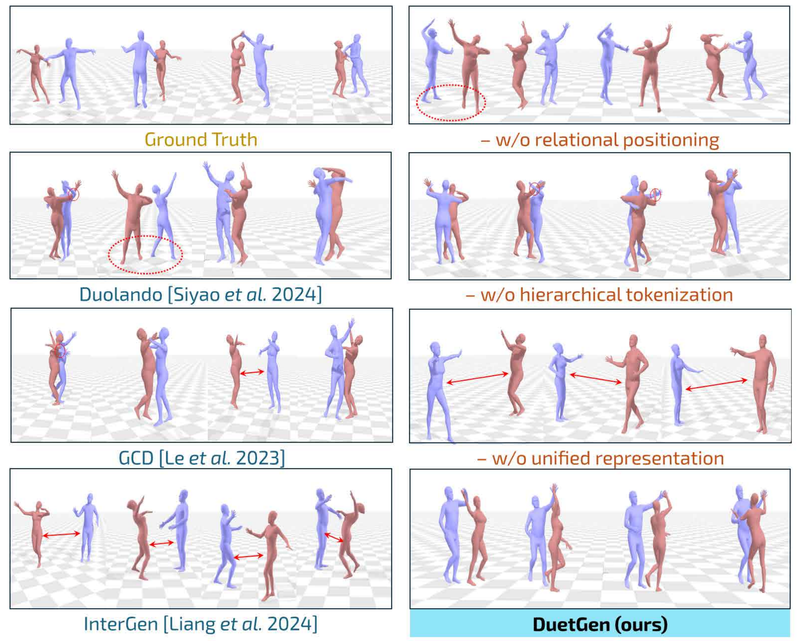
Inference Pipeline

Iteratively predict a token sequence, then decode it into motion samples



Inference Pipeline. Given a music signal, the masked transformer iteratively predicts the complete token sequence

Qualitative Comparison



Quantitative Performance

Method	FID ↓	Div →	PFID ↓	PFC ↓	CF (%) →	BAS ↑
Ground Truth	–	15.67	–	0.36	82.6	0.213
Duolando [Siyao *et al.* 2024]	12.21	14.72	13.18	16.22	74.7	0.202
GCD [Le *et al.* 2023a]	9.71	15.03	12.03	8.11	78.1	0.203
InterGen [Liang *et al.* 2024]	13.77	15.01	14.11	12.40	60.1	0.172
MoFusion [Dabral *et al.* 2023]	21.20	15.60	23.09	7.50	21.1	0.202
– w/o relational positioning	5.03	14.34	14.97	4.83	79.5	0.203
– w/o hier. tokenization	4.77	14.45	15.44	5.88	80.2	0.197
– w/o unified representation	5.65	14.02	18.99	5.66	75.7	0.204
– w/o trajectory refinement	2.62	14.11	2.81	5.31	78.2	0.211
– with 438-D music rep.	2.45	14.99	2.87	3.45	72.1	0.193
– with 3 levels of hierarchy	1.55	15.62	2.95	5.45	70.1	0.210
DuetGen (ours)	**1.31**	**15.71**	**2.54**	**1.47**	**83.2**	**0.215**