Music-Driven Group Choreography

Nhat Le[1], Thang Pham[1], Tuong Do[1], Erman Tjiputra[1], Quang D. Tran[1], Anh Nguyen[2]

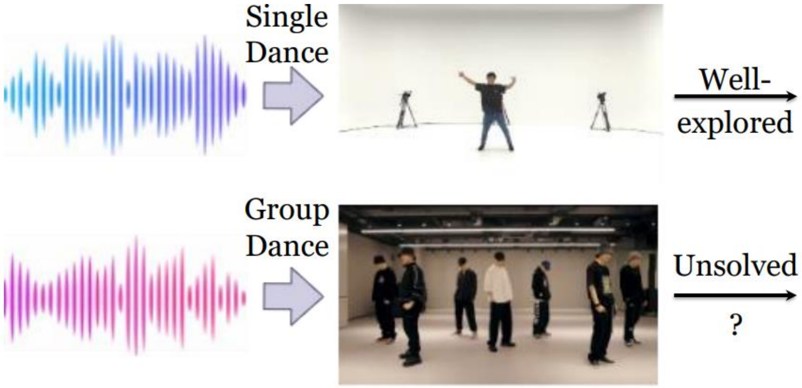
[1]AIOZ, Singapore [2]University of Liverpool, UK

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

WED-AM-043


I Love GAME

Motivation



- **Group dance generation** poses a much more challenging but compelling problem, yet has not been well-investigated.
- Most current music-to-dance datasets are only for solo dance.

Goal: Creating consistent and coherent group dancing motions.

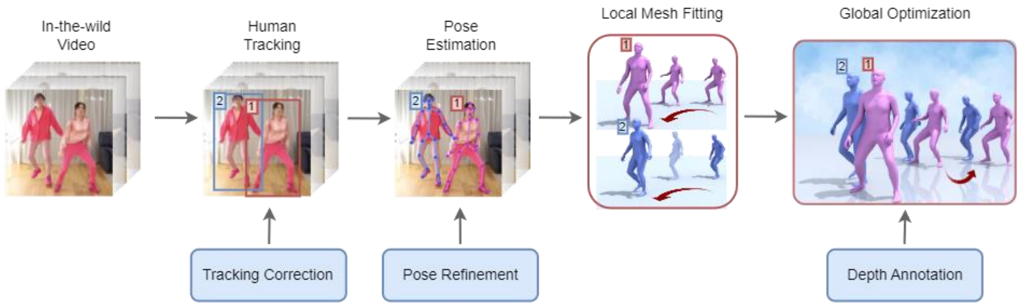
Challenges of Group Dance

- Different choreographies of dancers but semantically unified under a group dance performance.
- Physical interaction/contact between dancers.
- Incorporate many 3D-related problems including multi-person motion, occlusion, and global coherency that single-dance datasets cannot capture.

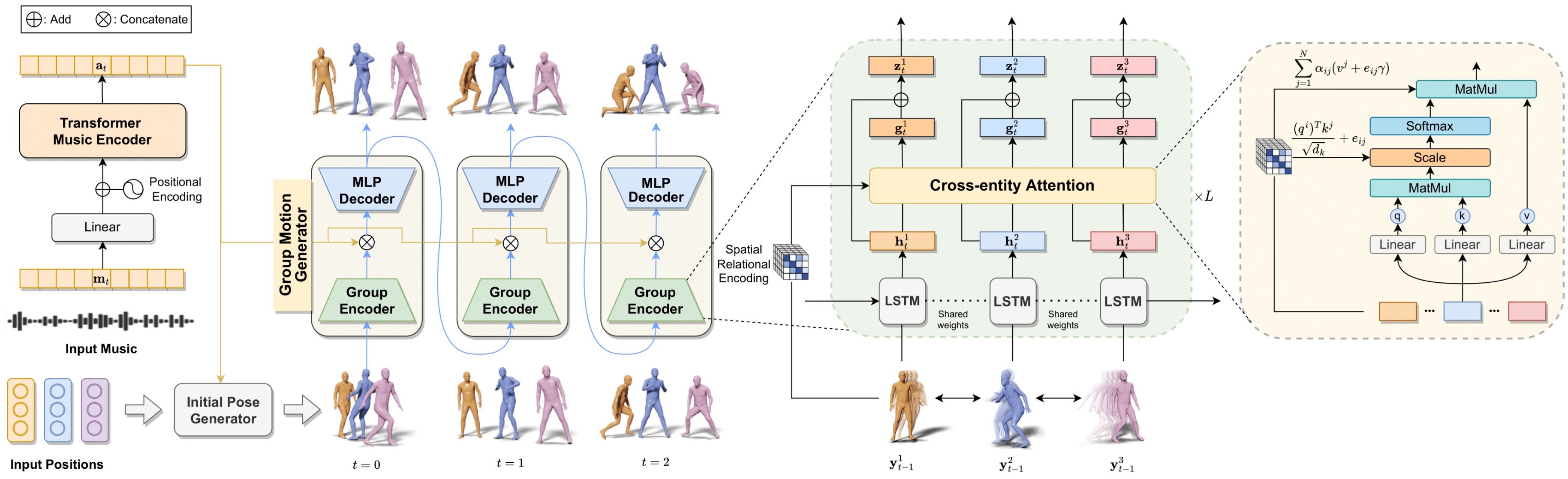
→ These challenges demand for the availability of a large-scale dataset to advance the research of group choreography generation.

AIOZ-GDANCE Dataset

- Total 16.7 hours of in-the-wild videos collected from Youtube, Tiktok, Facebook, at 1920 × 1080 resolution and 30 FPS.
- 16 Music genres: Pop, Electronic, Ballad, Folk, Disco,...
- 7 Dance styles: Modern commercial, Zumba, Rumba,...
- Containing paired music and 3D group dance motions.



Music-driven 3D Group Dance generatoR (GDanceR)



➢ **Input:** Music sequence, a set of initial 3D positions of N dancers.

➢ **Goal:** Maintain consistency between the motions and the music & Motions of dancers should be coherent with each other.

Experiments

- *Three new group dance evaluation metrics: GMR, GMC, TIF*

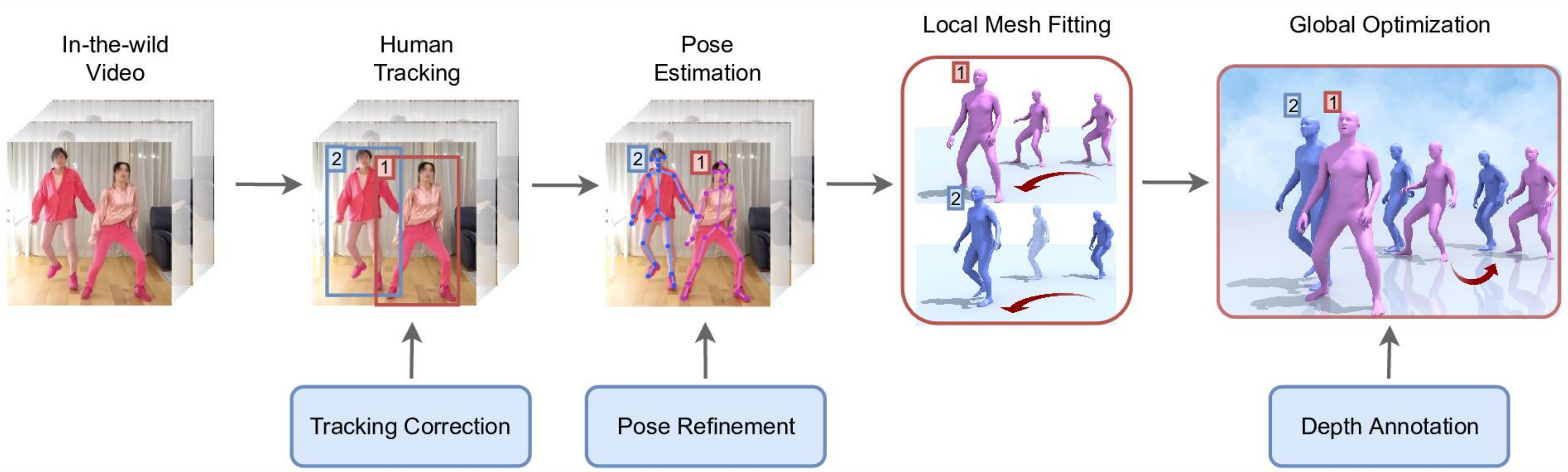
Method		Single-dance Metric			Group-dance Metric		
		FID↓	MMC↑	GenDiv↑	GMR↓	GMC↑	TIF↓
FACT [1]		56.20	0.222	8.64	101.52	62.68	0.321
GDanceR (ours)	*w/o* CA	63.83	0.218	8.99	109.80	68.47	0.379
	***w* CA**	**43.90**	**0.250**	**9.23**	**51.27**	**79.01**	**0.217**

FACT [1] *GDanceR (ours)*

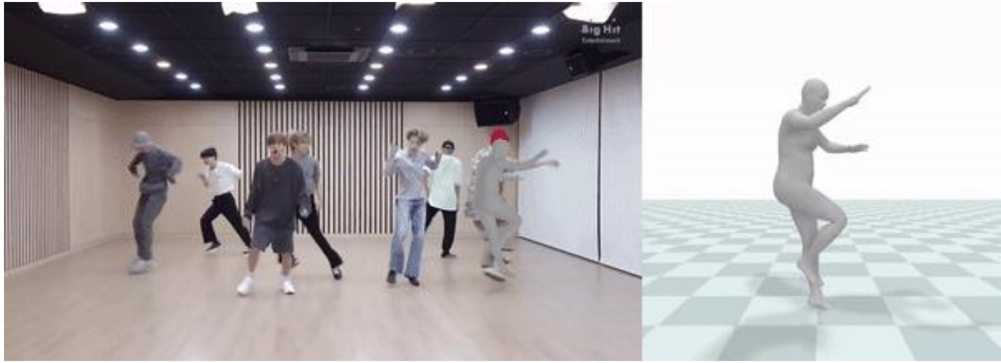
Dance Styles	Single-dance Metric			Group-dance Metric		
	FID↓	MMC↑	GenDiv↑	GMR↓	GMC↑	TIF↓
Zumba	45.86	0.268	9.77	50.97	72.70	0.133
Aerobic	38.68	0.252	6.57	63.62	75.12	0.249
Commercial	46.22	0.232	8.58	51.18	81.02	0.056
Bollywood	81.89	0.211	2.14	101.49	74.00	0.377
Irish	42.02	0.219	8.56	42.73	82.00	0.083
Rumba	69.62	0.273	3.91	68.00	71.85	0.228
Samba	71.00	0.228	7.77	98.83	67.76	0.441

[1] Li et al., Ai choreographer: Music-conditioned 3d dance generation with aist++. In ICCV, 2021.

Data Pipeline



Data Pipeline – Local Mesh Fitting



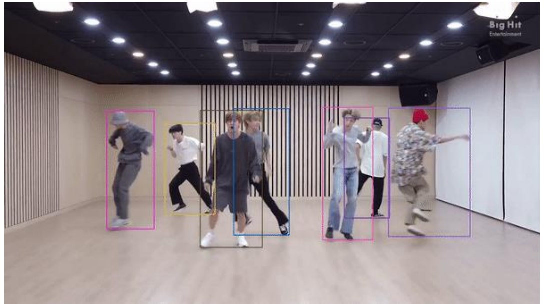
- We use SMPL body model [1] to represent the 3D human
- Optimizing variables: Pose parameter $\boldsymbol{\theta}$, Shape parameter $\boldsymbol{\beta}$, Root translation $\boldsymbol{\tau}$
- Fitting objective:

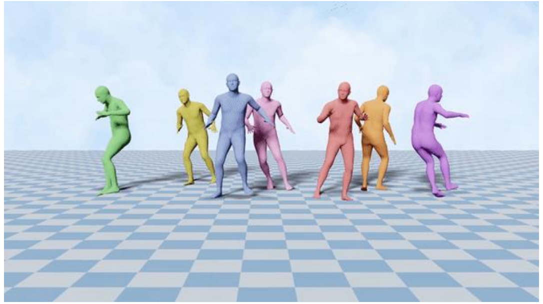
$$E_{\text{local}} = E_J + \overbrace{\lambda_\theta E_\theta + \lambda_\beta E_\beta}^{\text{pose and shape priors [2]}} + \lambda_S E_S + \lambda_F E_F$$

2D reprojection error   smoothness term   contact velocity term

[1] Loper et al., SMPL: A skinned multiperson linear model. ACM Transaction of Graphics, 2015
[2] Pavlakos et al., Expressive body capture: 3d hands, face, and body from a single image. In CVPR, 2019.

Data Pipeline – Global Optimization





- Global Fitting objective:

regularization term

global ground contact term

$$E_{\text{global}} = E_J + \lambda_{\text{pen}} E_{\text{pen}} + \lambda_{\text{reg}} \sum_p E_{\text{reg}}(p) + \lambda_{\text{dep}} \sum_{p,p',t} E_{\text{dep}}(p,p',t) + \lambda_{\text{gc}} \sum_p E_{\text{gc}}(p)$$

2D reprojection error

penetration term

depth relation term

Ordinal depth relation label

$$r_t(p,p') = \begin{cases} 1, & \text{if dancer } p \text{ is closer than } p' \\ -1, & \text{if dancer } p \text{ is farther than } p' \\ 0, & \text{if their depths are roughly equal} \end{cases}$$

$$E_{\text{dep}}(p,p',t) = \begin{cases} \log(1+\exp(z_t^p - z_t^{p'})), & r_t(p,p') = 1 \\ \log(1+\exp(-z_t^p + z_t^{p'})), & r_t(p,p') = -1 \\ (z_t^p - z_t^{p'})^2, & r_t(p,p') = 0 \end{cases}$$

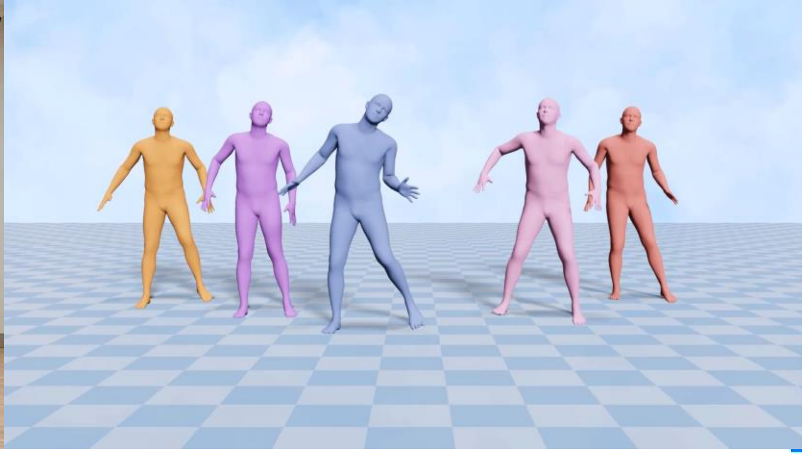
global ground contact term

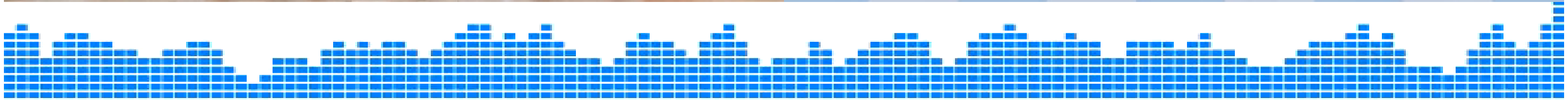
$$E_{\text{gc}}(p) = \sum_{t=1}^{T-1} \sum_{j \in \mathcal{F}} c_{j,t}^p \|\mathbf{X}_{j,t+1}^p - \mathbf{X}_{j,t}^p\|^2 + c_{j,t}^p \|(\mathbf{X}_{j,t}^p - \mathbf{f})^\top \mathbf{n}^*\|^2$$

AIOZ-GDANCE Dataset

Video

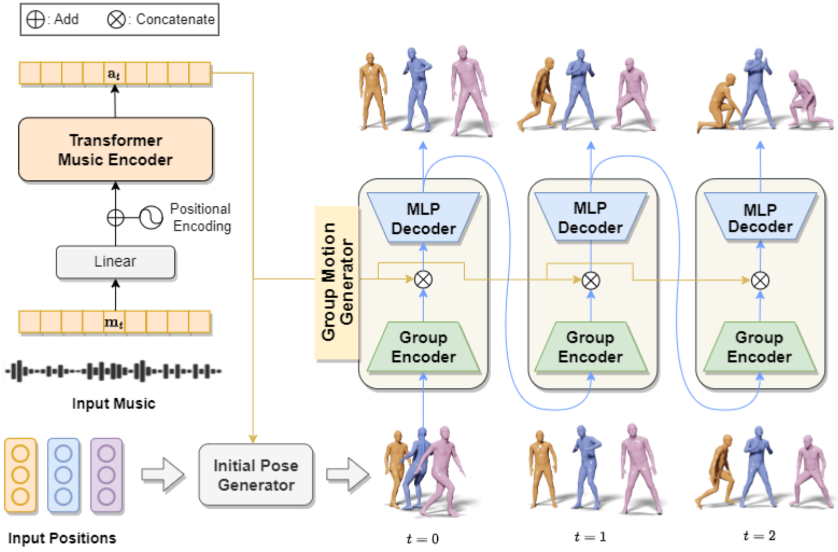
3D Ground truth





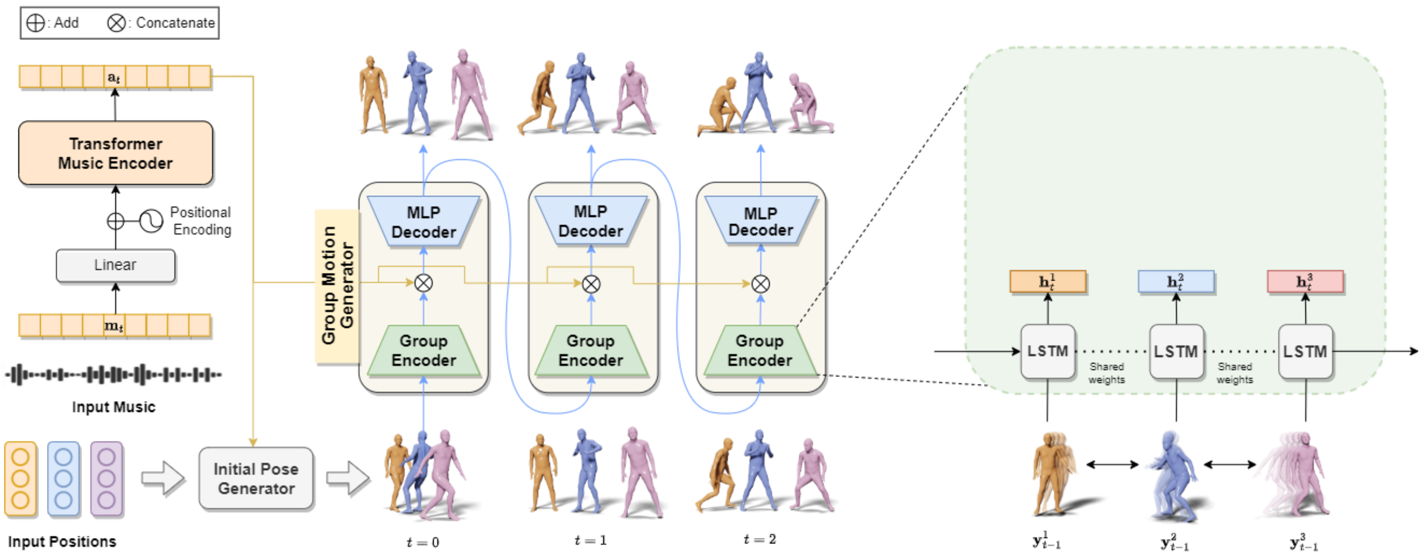
Music

Music-driven 3D Group Dance generatoR (GDanceR)



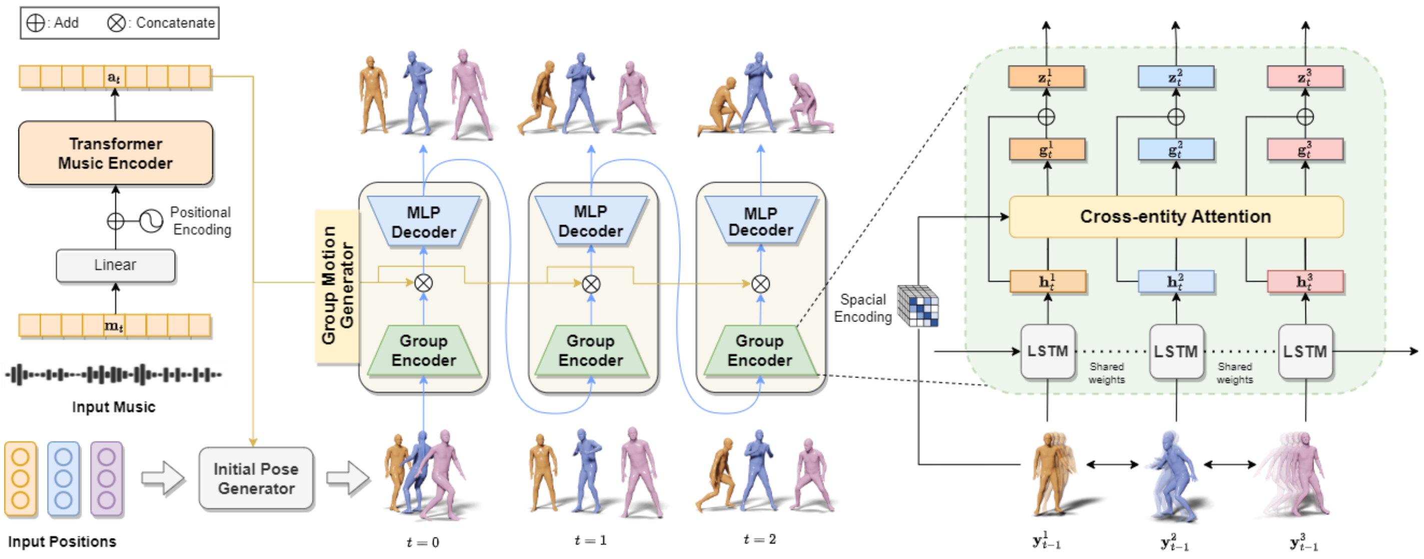
- **Input:** Music sequence, a set of initial 3D positions of N dancers.
- **Goal:** Maintain consistency between the motions and the music & Motions of dancers should be coherent with each other.

Music-driven 3D Group Dance generatoR (GDanceR)



- **Input:** Music sequence, a set of initial 3D positions of N dancers.
- **Goal:** Maintain consistency between the motions and the music & Motions of dancers should be coherent with each other.

Music-driven 3D Group Dance generatoR (GDanceR)

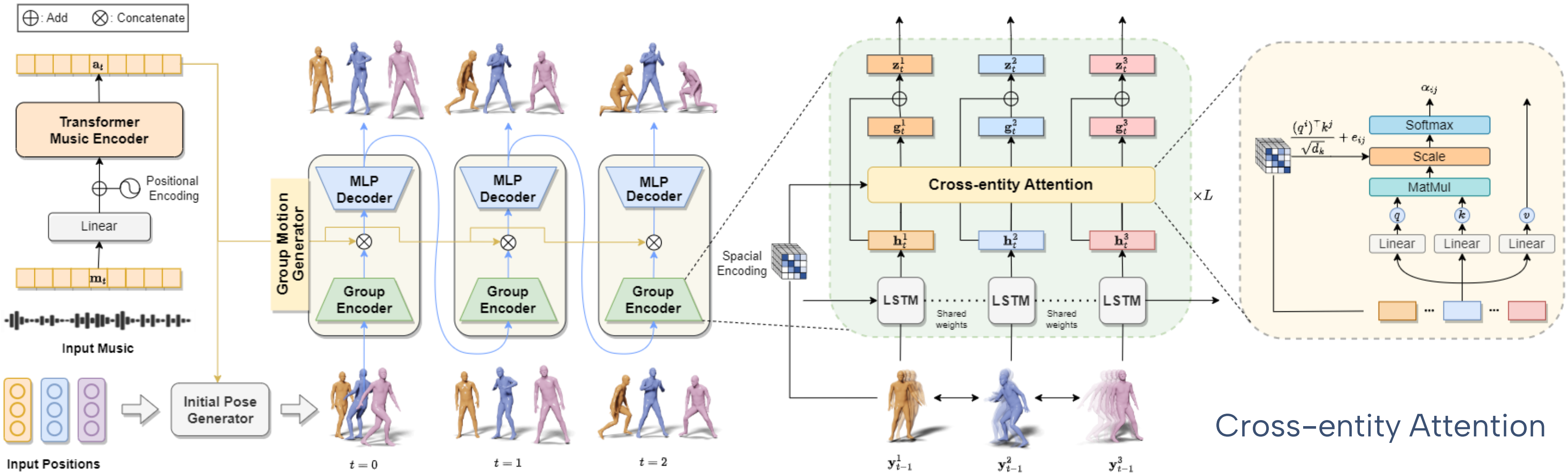


- **Input:** Music sequence, a set of initial 3D positions of N dancers.
- **Goal:** Maintain consistency between the motions and the music & Motions of dancers should be coherent with each other.

Spacial Encoding

$$e_{ij} = \exp\left(-\frac{\|\tau^i - \tau^j\|^2}{\sqrt{d_\tau}}\right)$$

Music-driven 3D Group Dance generatoR (GDanceR)



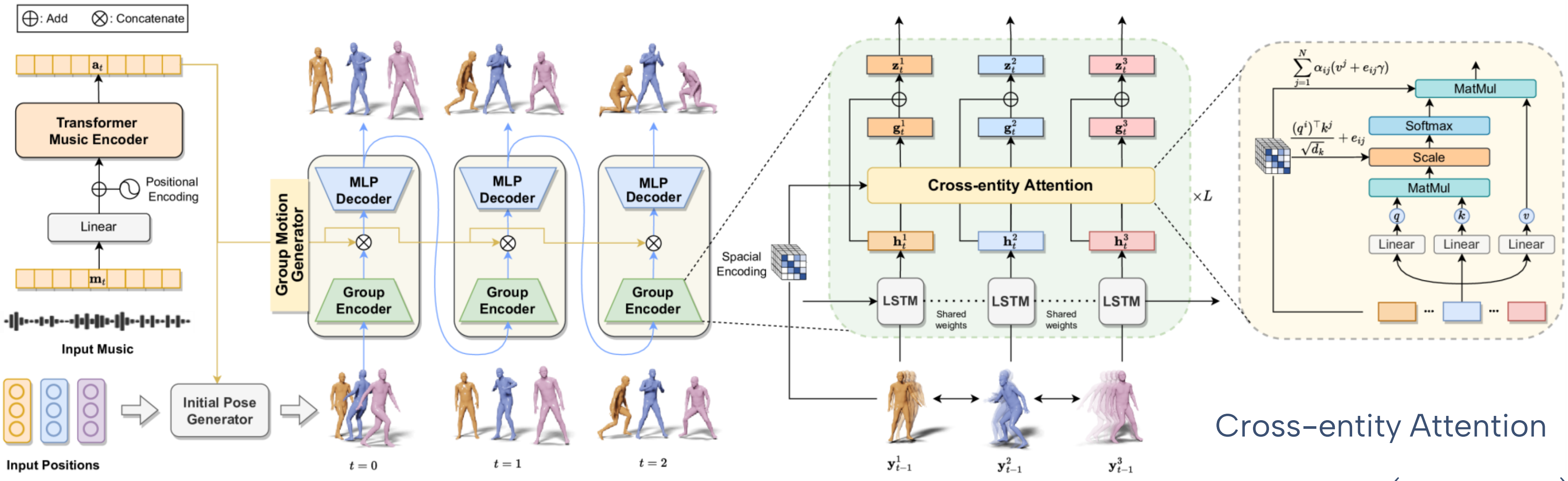
Cross-entity Attention

$$\alpha_{ij} = \text{softmax}\left(\frac{(q^i)^\top k^j}{\sqrt{d_k}} + e_{ij}\right)$$

- **Input:** Music sequence, a set of initial 3D positions of N dancers.

- **Goal:** Maintain consistency between the motions and the music & Motions of dancers should be coherent with each other.

Music-driven 3D Group Dance generatoR (GDanceR)



Cross-entity Attention

$$\alpha_{ij} = \text{softmax}\left(\frac{(q^i)^\top k^j}{\sqrt{d_k}} + e_{ij}\right)$$

$$g^i = \sum_{j=1}^{N} \alpha_{ij}(v^j + e_{ij}\gamma)$$

- **Input:** Music sequence, a set of initial 3D positions of N dancers.
- **Goal:** Maintain consistency between the motions and the music & Motions of dancers should be coherent with each other.

Results

Method		Single-dance Metric			Group-dance Metric		
		FID↓	MMC↑	GenDiv↑	GMR↓	GMC↑	TIF↓
FACT [1]		56.20	0.222	8.64	101.52	62.68	0.321
GDanceR	*w/o* CA	63.83	0.218	8.99	109.80	68.47	0.379
(ours)	***w* CA**	**43.90**	**0.250**	**9.23**	**51.27**	**79.01**	**0.217**

Evaluation results on AIOZ-GDANCE test set

N Generated Dancers	Single-dance Metric			Group-dance Metric		
	FID↓	MMC↑	GenDiv↑	GMR↓	GMC↑	TIF↓
2	48.82	0.248	9.66	53.83	75.44	0.086
3	44.47	0.245	9.46	52.85	74.07	0.104
4	47.32	0.248	9.24	58.79	77.71	0.162
5	44.19	0.249	9.38	55.05	78.72	0.218
6	50.95	0.250	9.25	59.05	75.24	0.319
7	48.86	0.250	9.19	56.23	76.01	0.367

Results with varying number of dancers

Dance Styles	Single-dance Metric			Group-dance Metric		
	FID↓	MMC↑	GenDiv↑	GMR↓	GMC↑	TIF↓
Zumba	45.86	0.268	9.77	50.97	72.70	0.133
Aerobic	38.68	0.252	6.57	63.62	75.12	0.249
Commercial	46.22	0.232	8.58	51.18	81.02	0.056
Bollywood	81.89	0.211	2.14	101.49	74.00	0.377
Irish	42.02	0.219	8.56	42.73	82.00	0.083
Rumba	69.62	0.273	3.91	68.00	71.85	0.228
Samba	71.00	0.228	7.77	98.83	67.76	0.441

Results of different dance styles.

[1] Li et al., Ai choreographer: Music-conditioned 3d dance generation with aist++. In ICCV, 2021.

Results

FACT [1]

GDanceR (Ours)

[1] Li et al., Ai choreographer: Music-conditioned 3d dance generation with aist++. In ICCV, 2021.

Results

Dance Style: Irish
Music Genre: Electronic
Number of dancers: 2

Dance Style: Aerobic
Music Genre: R&B
Number of dancers: 3

Conclusion

- We have introduced AIOZ-GDANCE, the currently largest in-the-wild dataset for audio-driven group dance generation.
- We propose a strong baseline along with evaluation protocols for group dance generation task.
- We hope that the release of our dataset will foster more research on music-driven group choreography.
- Our paper and more results are available at: https://aioz-ai.github.io/AIOZ-GDANCE/

Project Page