MoMu-Diffusion: On Learning Long-Term Motion-Music Synchronization and Correspondence

Fuming You, Minghui Fang, Li Tang, Rongjie Huang, Zhou Zhao

Zhejiang University

Background

Table 1: Comparison with the state-of-the-art audio-visual generation works, including but not limited to motion-music generation.

Method	Pub.	Joint Generation	Pretrain	Long-Term Synthesis	Latent Space
Diff-Foley	NeurIPS'23	✗	✓	✗	✓
MM-Diffusion	CVPR'23	✓	✗	✗	✗
LORIS	ICML'23	✗	✗	✓	✗
D2M	NeurIPS'19	✗	✓	✗	✓
CDCD	ICLR'23	✗	✗	✓	✓
MoMu-Diffusion		✓	✓	✓	✓

Motion-to-Music and Music-to-Motion generations are separately researched.

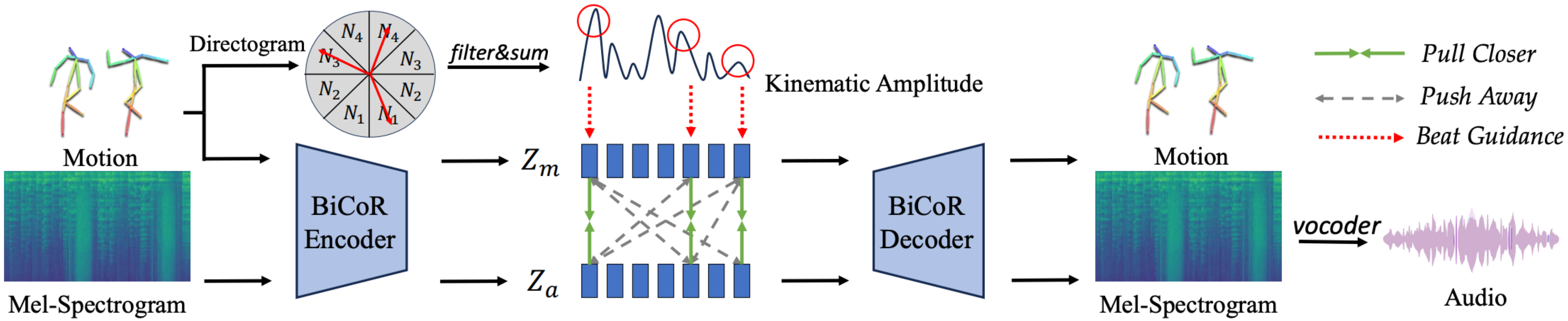
Challenges

Table 1: Comparison with the state-of-the-art audio-visual generation works, including but not limited to motion-music generation.

Method	Pub.	Joint Generation	Pretrain	Long-Term Synthesis	Latent Space
Diff-Foley	NeurIPS'23	✗	✓	✗	✓
MM-Diffusion	CVPR'23	✓	✗	✗	✗
LORIS	ICML'23	✗	✗	✓	✗
D2M	NeurIPS'19	✗	✓	✗	✓
CDCD	ICLR'23	✗	✗	✓	✓
MoMu-Diffusion		✓	✓	✓	✓

1. Maintain long-term coherence in typically lengthy motion-music sequences.
2. Ensure temporal synchronization and rhythmic alignment between motion and music sequences.

BiCoR-VAE



(a) Bidirectional Contrastive Rhythmic VAE (BiCoR-VAE)

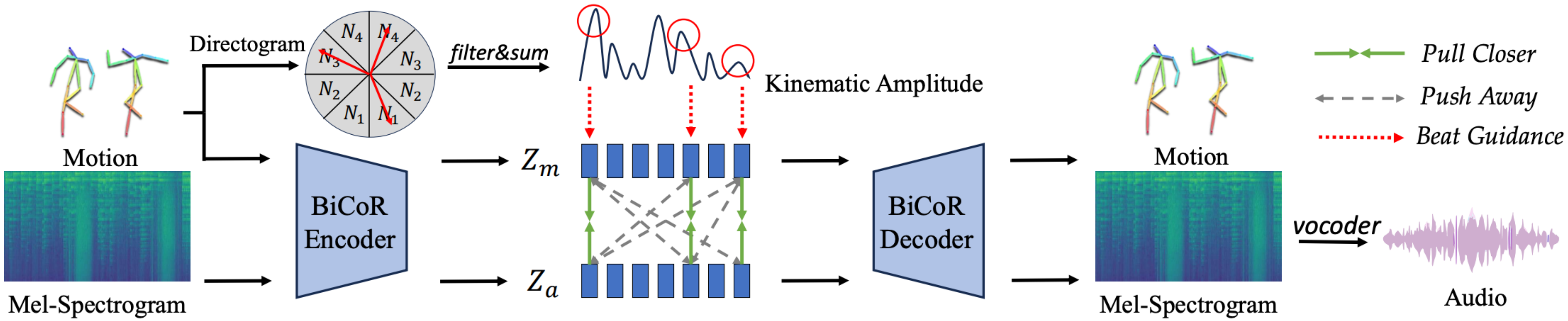
2D Motion Directogram:

$$D(r,\theta) = \sum_{j=1}^{J} ||F(r,j)||_2 \mathbb{1}_\theta(\angle F(r,j)), \quad \text{where} \mathbb{1}_\theta(\phi) := \begin{cases} 1, & |\theta - \phi| \leq 2\pi/K, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Kinematic Amplitude:

$$Q(r) = \sum_{k=1}^{K} \max(0, |D(r,k)| - |D(r-1,k)|), \tag{2}$$

BiCoR-VAE



(a) Bidirectional Contrastive Rhythmic VAE (BiCoR-VAE)

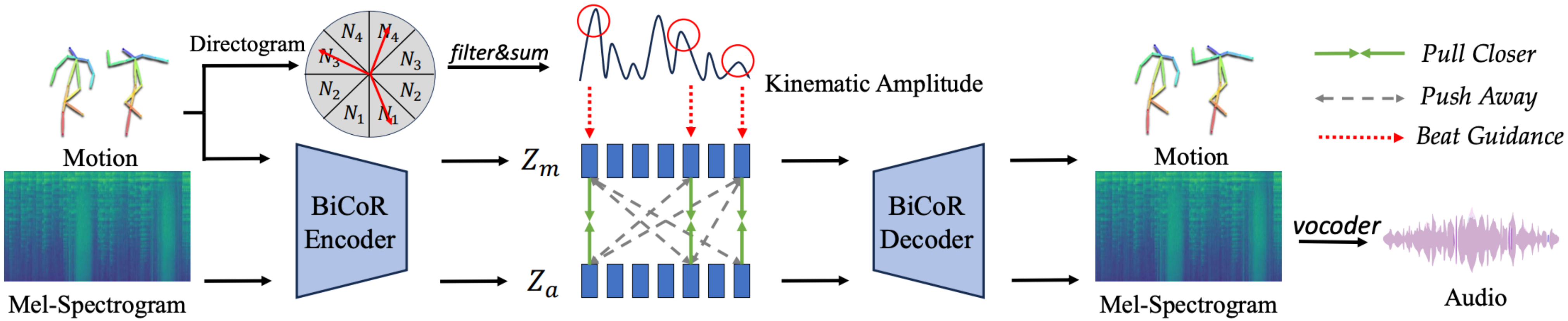
Sampling contrastive pairs:

$$c_a^{r_s:r_e} = P_{\max}(z_a^{r_s} : z_a^{r_e}),\ c_m^{r_s:r_e} = P_{\max}(z_m^{r_s} : z_m^{r_e}),\ Q(r_s : r_e) = \max(Q(r_s) : Q(r_e)), \tag{3}$$

Contrastive Loss for rhythmic alignment:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{2}\log\frac{\exp(sim(c_a^i, c_m^j)/\tau)}{\sum_{c=1}^{N_C}\exp(sim(c_a^i, c_m^c)/\tau)} - \frac{1}{2}\log\frac{\exp(sim(c_a^i, c_m^j)/\tau)}{\sum_{c=1}^{N_C}\exp(sim(c_a^c, c_m^j)/\tau)}. \tag{4}$$

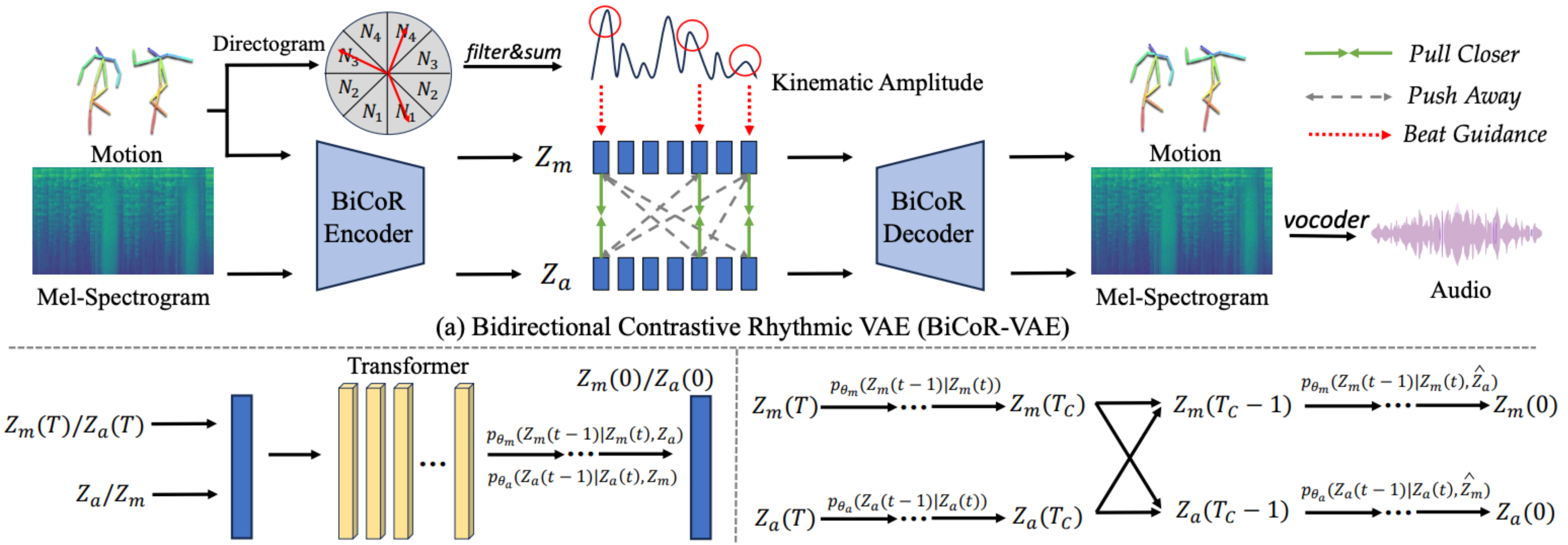
Training Strategy for BiCoR-VAE



(a) Bidirectional Contrastive Rhythmic VAE (BiCoR-VAE)

1. The VAE presents a trade-off between representational fidelity and generative alignment, posing optimization challenges.
2. Initially, we train the music VAE with a reconstruction loss, a KL loss, and a GAN loss to prevent over-smoothing of the mel-spectrogram.
3. Then, we fix the trained music VAE and train the motion VAE with a reconstruction loss, a KL loss, and the proposed contrastive rhythmic loss.

Cross-Modal Generation



(a) Bidirectional Contrastive Rhythmic VAE (BiCoR-VAE)
(b) Cross-Modal Generation
(c) Multi-Modal Joint Generation

Training loss:

$$\mathcal{L}_{\text{m2a}} = ||\epsilon_{\theta_a}(z_a(t), t, z_m) - \epsilon||_2^2, \qquad \mathcal{L}_{\text{a2m}} = ||\epsilon_{\theta_m}(z_m(t), t, z_a) - \epsilon||_2^2, \tag{6}$$

Classifier-free guidance:

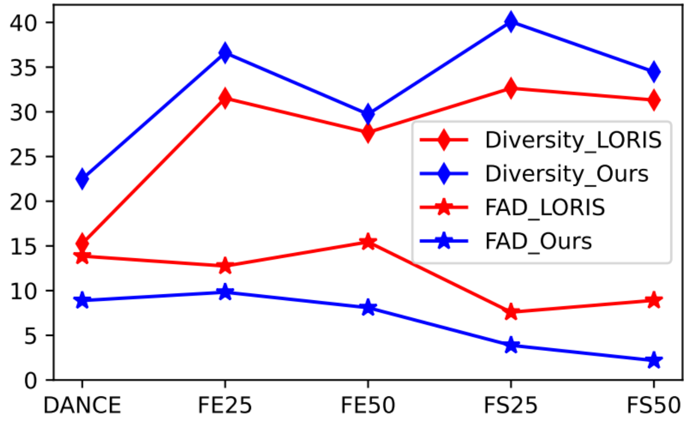
$$\hat{\epsilon}_{\theta_a}(z_a(t), t, z_m) = \epsilon_{\theta_a}(z_a(t), t, \emptyset) + s \cdot (\epsilon_{\theta_a}(z_a(t), t, z_m) - \epsilon_{\theta_a}(z_a(t), t, \emptyset)) \tag{7}$$

Results: Motion-to-Music

Table 2: Motion-to-music with **beat-matching** metrics.

Subset / Metrics	AIST++ Dance BCS↑	CSD↓	BHS↑	HSD↓	F1↑
Foley	96.4	6.9	41.0	15.0	57.5
CMT	97.1	6.4	46.2	18.6	62.6
D2MGAN	95.6	9.4	88.7	19.0	93.1
CDCD	96.5	9.1	89.3	18.1	92.7
LORIS	**98.6**	6.1	90.8	13.9	94.5
Ours	97.5	**5.2**	**98.6**	**2.8**	**98.1**

Figure 3: Motion-to-music with **generation quality** metrics: FAD↓ and Diversity↑.



Subset / Metrics	Floor Exercise-25s BCS↑	CSD↓	BHS↑	HSD↓	F1↑	Floor Exercise-50s BCS↑	CSD↓	BHS↑	HSD↓	F1↑
Foley	36.0	36.2	32.3	30.7	34.1	32.6	38.0	28.4	32.5	30.4
CMT	46.4	30.1	57.4	29.8	51.3	42.3	32.0	53.8	31.7	47.4
D2MGAN	45.3	27.7	58.7	30.1	51.1	41.9	29.2	54.7	32.7	47.5
CDCD	49.0	21.1	61.0	27.0	54.3	45.9	23.8	57.5	29.3	51.0
LORIS	58.8	19.4	67.1	21.1	62.7	54.7	**21.6**	63.8	24.5	58.9
Ours	**66.6**	**14.3**	**76.9**	**19.1**	**71.4**	**62.7**	24.0	**68.1**	**20.2**	**65.3**

Table 3: Results on the Floor Exercise dataset with **beat-matching** metrics.

Results: Music-to-Motion

Subset Metrics	AIST++ Dance BCS↑	CSD↓	BHS↑	HSD↓	F1↑	BHS Dance BCS↑	CSD↓	BHS↑	HSD↓	F1↑
D2M	23.7	13.8	42.8	23.6	30.5	35.1	15.9	57.5	35.0	43.6
DiffGesture	28.5	16.7	40.4	25.7	33.4	42.8	21.3	61.1	23.9	50.3
Ours	**39.2**	**10.2**	**56.3**	**12.0**	**46.2**	**47.9**	**8.4**	**78.5**	**12.1**	**59.5**

Table 5: Results on the AIST++ Dance and BHS Dance datasets with **beat-matching** metrics.

Subset Metrics	AIST++ Dance FID↓	Diversity↑	Mean KLD↓	BHS Dance FID↓	Diversity↑	Mean KLD↓
D2M	17.3	46.2	14.5	11.6	55.9	7.4
DiffGesture	18.6	37.1	12.6	13.8	38.9	7.0
Ours	**7.3**	**52.7**	**4.9**	**6.5**	**67.4**	**4.2**

Table 6: Results on the AIST++ Dance and BHS Dance datasets with **generation quality** metrics.

Results: Ablations

Id	Method	Music Metrics FAD ↓	F1 ↑	Motion Metrics FID ↓	F1↑
#1	Ours w/ Directional Vectors	10.9	91.4	14.7	38.0
#2	Ours w/o Mel-spectrogram	12.8	95.6	9.5	41.6
#3	Ours w/o Rhythmic Contrastive Learning (RCL)	8.5	93.1	8.1	37.9
#4	Ours w/o Diffusion Transformer (DiT)	11.0	95.8	11.6	41.4
#5	Ours (Joint Generation)	8.1	96.5	8.8	45.4
#6	Ours (Joint Generation&Variable Length)	9.1	97.6	8.5	49.6
#7	Ours (Cross Generation)	8.9	98.1	7.3	46.2

Table 7: Ablation study on motion-to-music and music-to-motion generations. We use the FAD/FID as the quality assessment and the F1 score as the beat-matching assessment.

Thanks