Beat-It

Beat-Synchronized Multi-Condition 3D Dance Generation

Zikai Huang[1] Xuemiao Xu[1,3,4†] Cheng Xu[2†] Huaidong Zhang[1,3] Chenxi Zheng[1] Jing Qin[2] Shengfeng He[5]

[1]South China University of Technology [2]The Hong Kong Polytechnic University

[3]Guangdong Engineering Center for Large Model and GenAI Technology

[4]Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information

[5]Singapore Management University

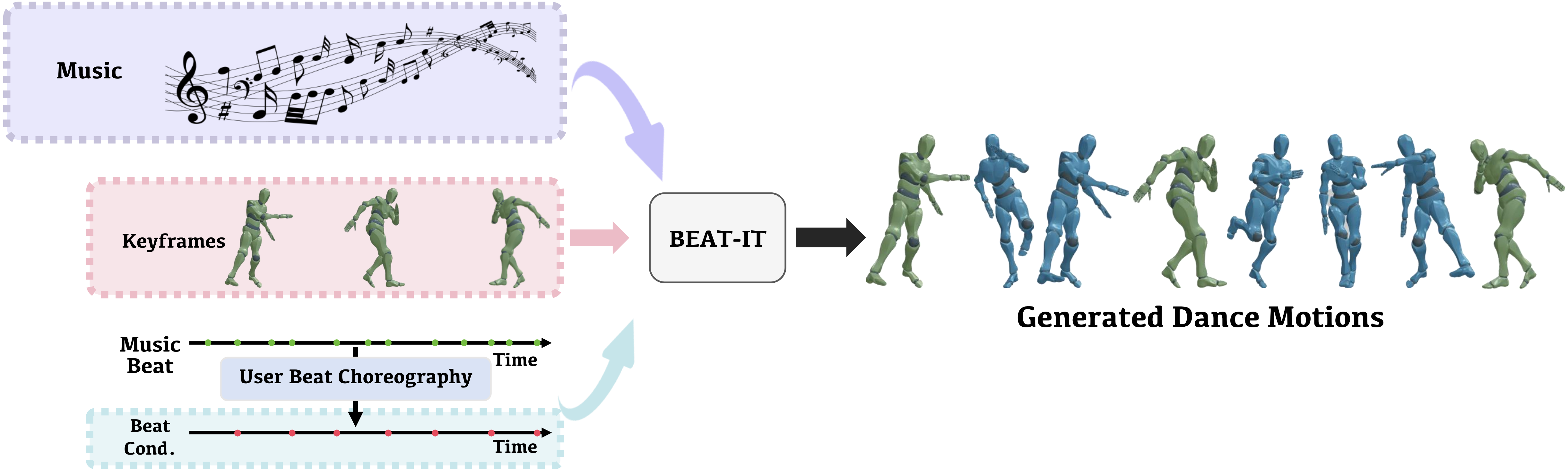
ECCV

EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO

2024

Goal



Music

Keyframes

Music Beat

Time

User Beat Choreography

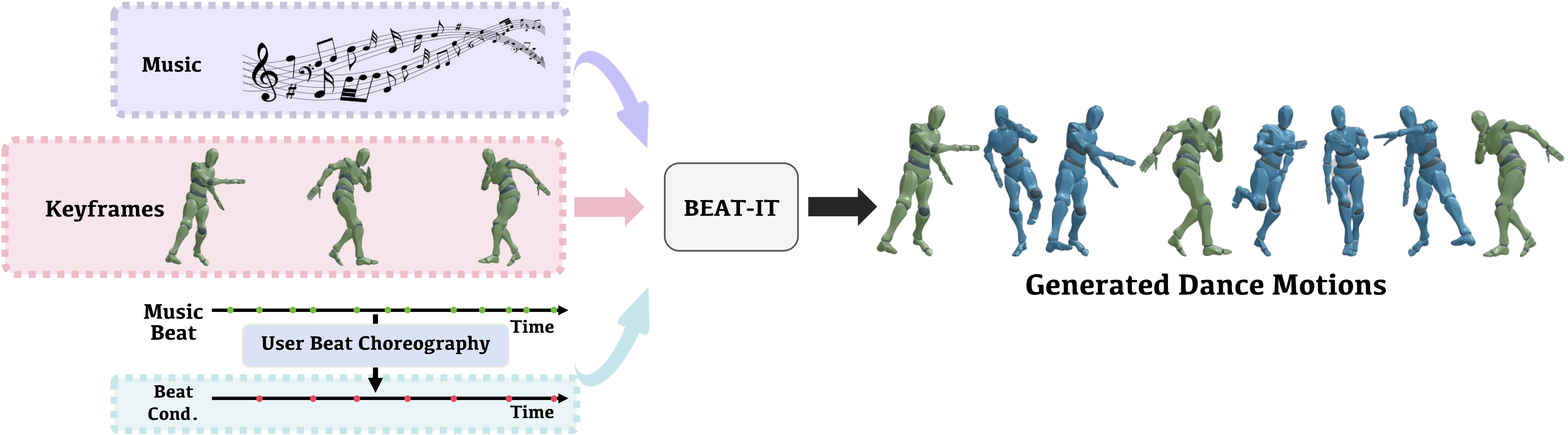
Beat Cond.

Time

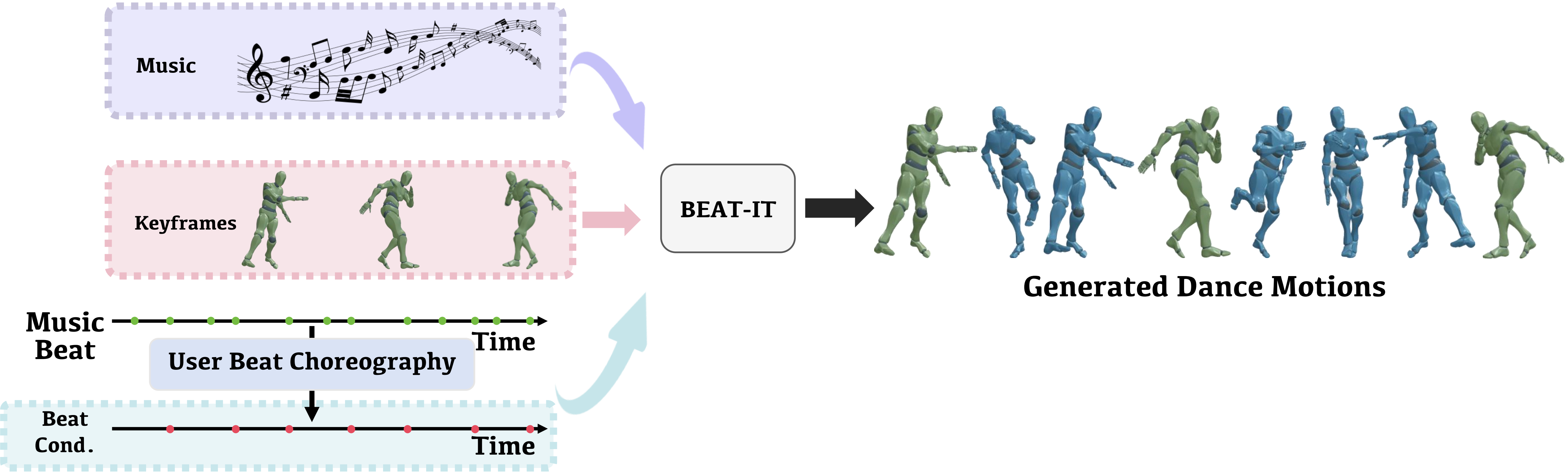
BEAT-IT

Generated Dance Motions

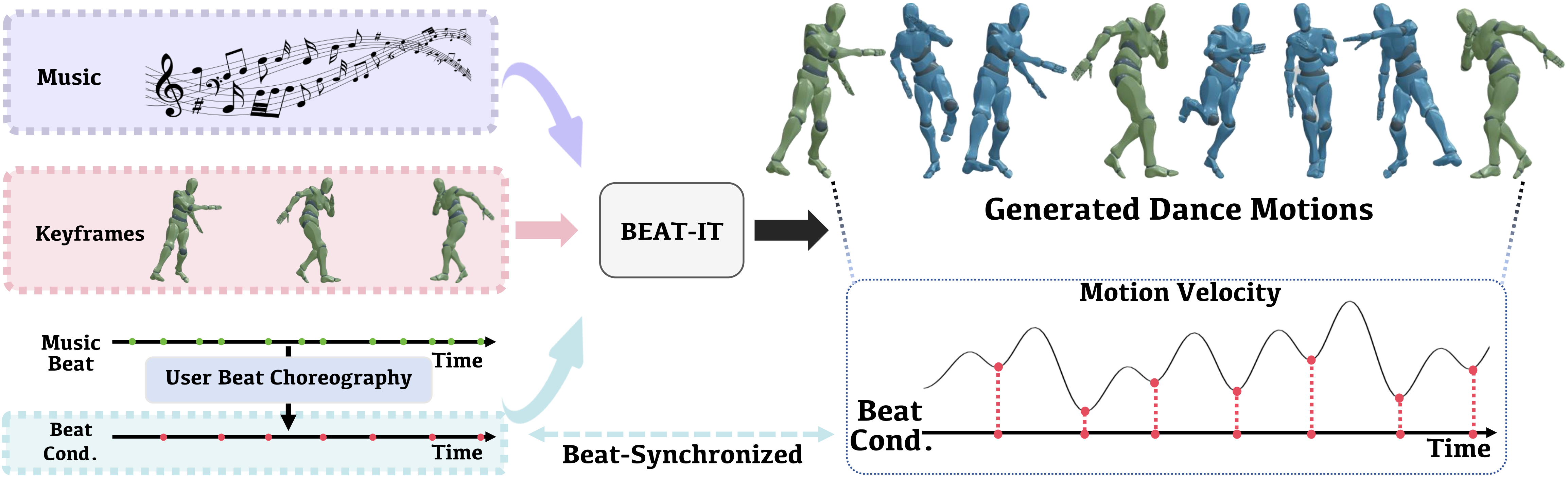
Goal


Music
Keyframes
Music Beat
Time
User Beat Choreography
Beat Cond.
Time
BEAT-IT
Generated Dance Motions

Goal


Music
Keyframes
Music Beat
Time
User Beat Choreography
Beat Cond.
Time
BEAT-IT
Generated Dance Motions

Goal


Music
Keyframes
Music Beat
Time
User Beat Choreography
Beat Cond.
Time
BEAT-IT
Beat-Synchronized
Generated Dance Motions
Motion Velocity
Beat Cond.
Time

Challenges

😣 Inability to ensure alignment between generated dance and music.

😣 Lack of an effective beat representation and constraint for precise control over beat conditions.

😣 Difficulties in fusing multiple conditions with distinct information densities.

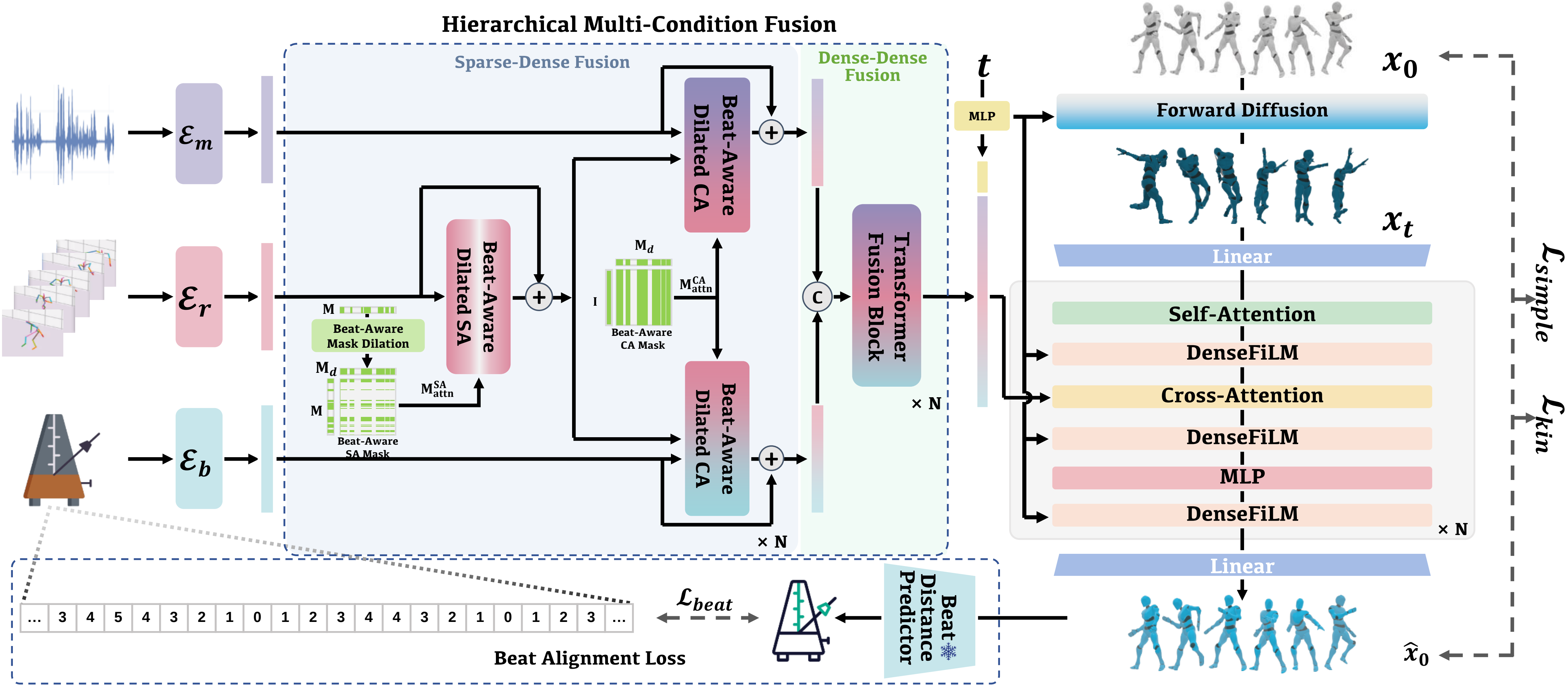
Contributions

💃 Introduce a **multi-conditional** dance generation framework that achieves beat synchronization and enhanced motion controllability.

💃 Present a **hierarchical multi-condition fusion mechanism** to effectively suppress the conflicts and fully exploit the complementary information among different conditions.

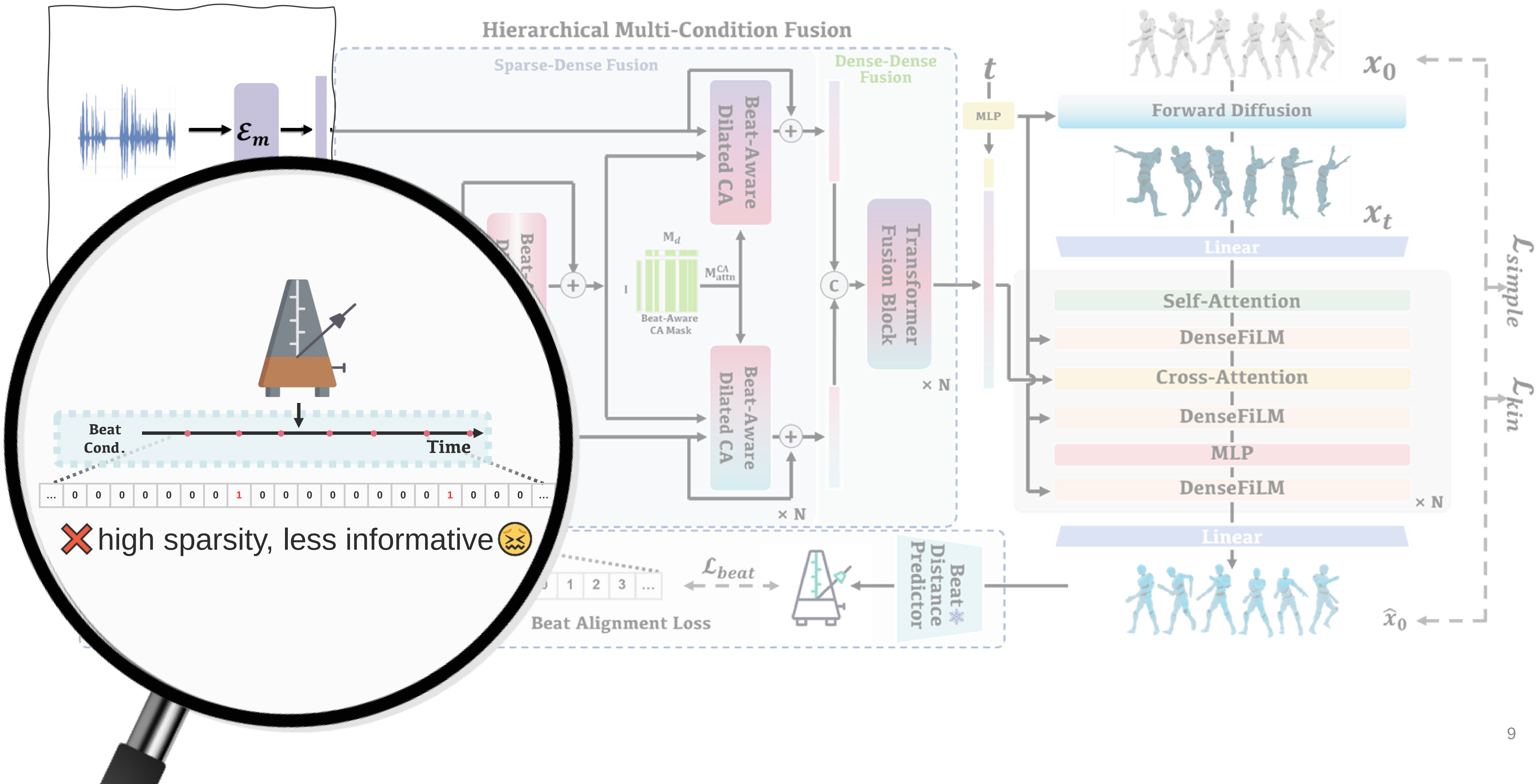
💃 Design a tailored **beat alignment loss** to explicitly guide the synchronization between generated dance motions and given beat conditions.

Method



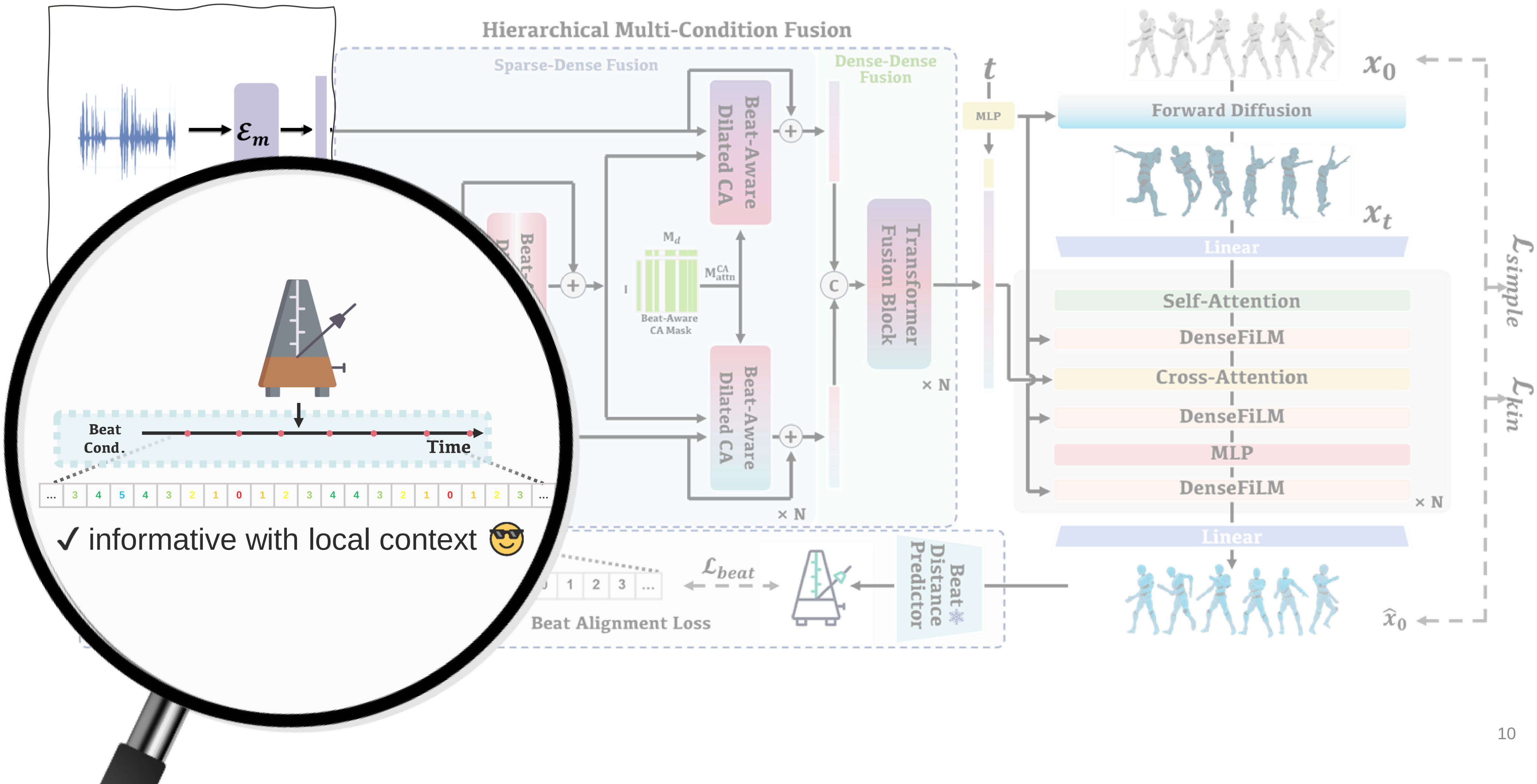
Hierarchical Multi-Condition Fusion
Sparse-Dense Fusion
Dense-Dense Fusion
$\mathcal{E}_m$
$\mathcal{E}_r$
$\mathcal{E}_b$
Beat-Aware Dilated CA
Beat-Aware Dilated SA
Beat-Aware Dilated CA
M
Beat-Aware Mask Dilation
$\mathbf{M}_d$
M
Beat-Aware SA Mask
$\mathbf{M}^{SA}_{attn}$
$\mathbf{M}_d$
I
$\mathbf{M}^{CA}_{attn}$
Beat-Aware CA Mask
× N
C
Transformer Fusion Block
× N
t
MLP
x_0
Forward Diffusion
x_t
Linear
Self-Attention
DenseFiLM
Cross-Attention
DenseFiLM
MLP
DenseFiLM
× N
Linear
$\hat{x}_0$
$\mathcal{L}_{simple}$
$\mathcal{L}_{kin}$
Beat Distance Predictor
$\mathcal{L}_{beat}$
... 3 4 5 4 3 2 1 0 1 2 3 4 4 3 2 1 0 1 2 3 ...
Beat Alignment Loss

Method



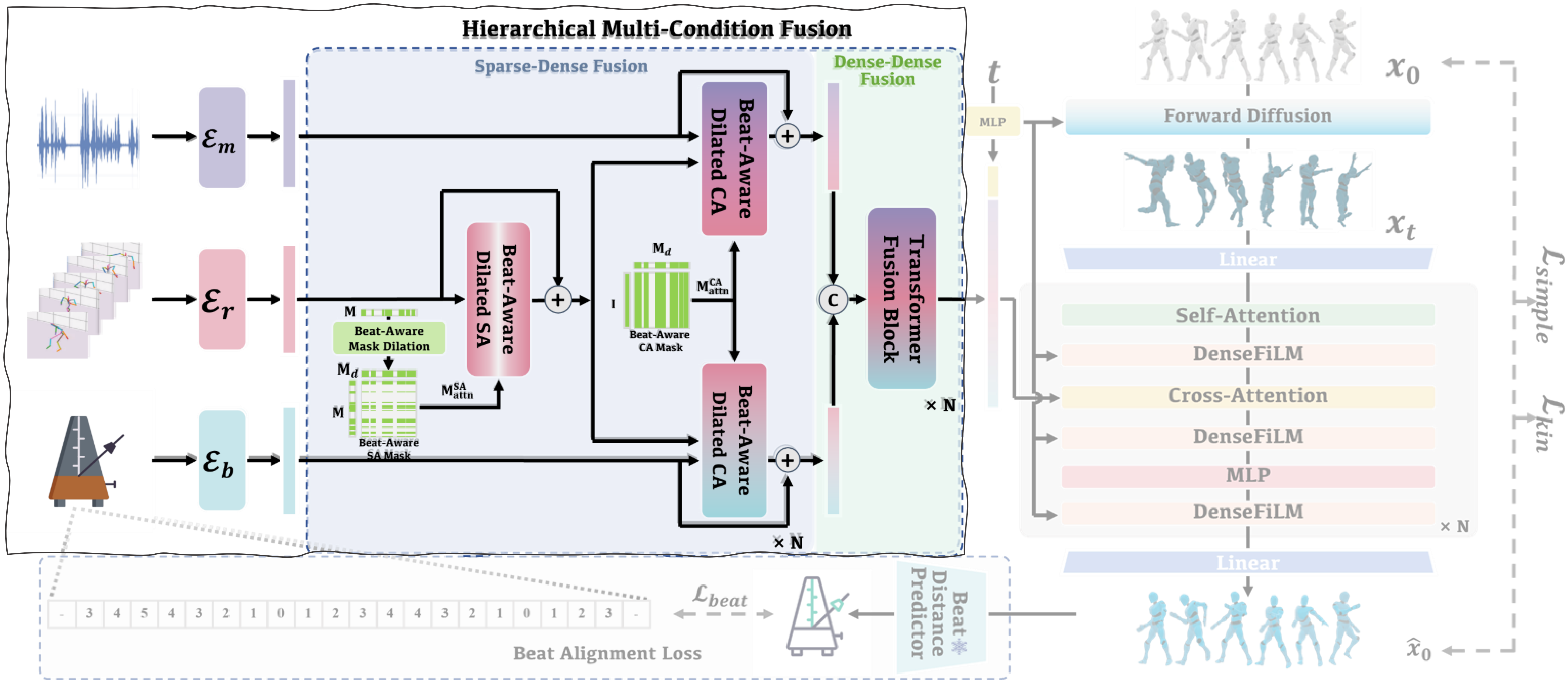
❌high sparsity, less informative😖

Method

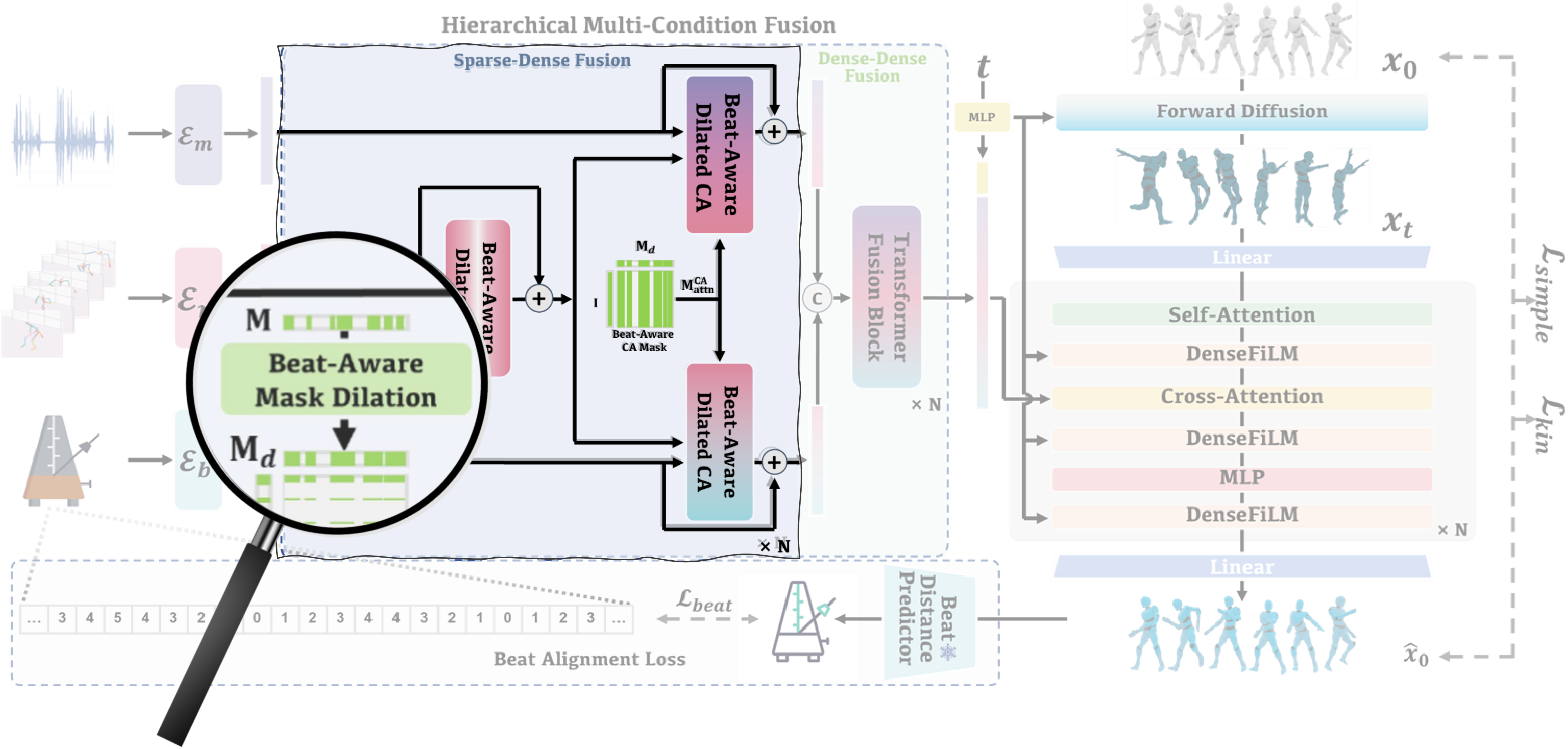


✓ informative with local context 😎

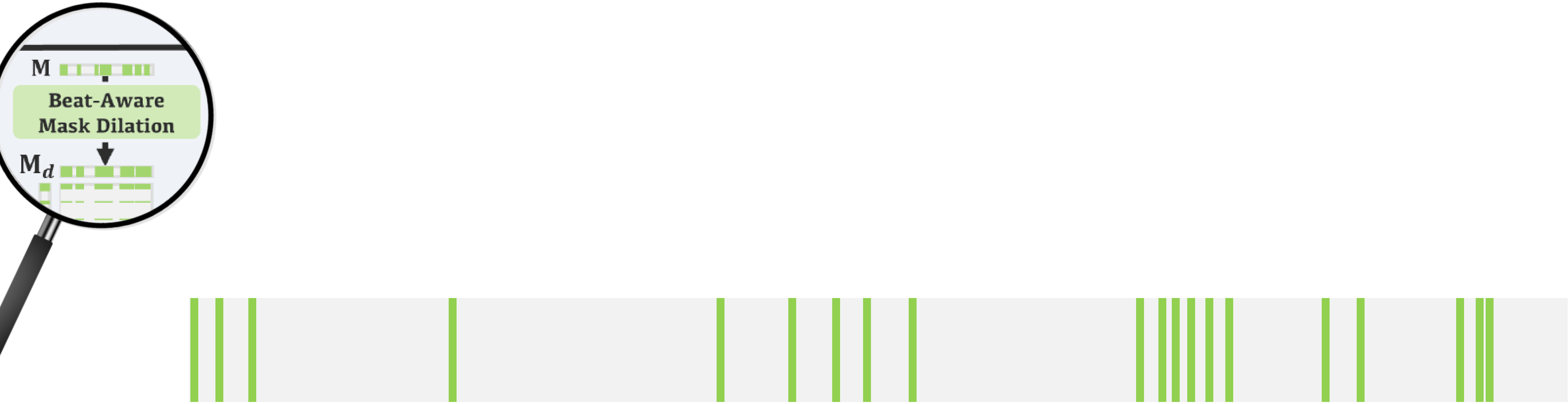
Method


Hierarchical Multi-Condition Fusion
Sparse-Dense Fusion
Dense-Dense Fusion
$\mathcal{E}_m$
$\mathcal{E}_r$
$\mathcal{E}_b$
Beat-Aware Dilated SA
Beat-Aware Dilated CA
Beat-Aware Dilated CA
M
Beat-Aware Mask Dilation
M_d
M
M^{SA}_{attn}
Beat-Aware SA Mask
M_d
I
M^{CA}_{attn}
Beat-Aware CA Mask
× N
C
Transformer Fusion Block
× N
t
MLP
x_0
Forward Diffusion
x_t
Linear
Self-Attention
DenseFiLM
Cross-Attention
DenseFiLM
MLP
DenseFiLM
× N
Linear
$\hat{x}_0$
$\mathcal{L}_{simple}$
$\mathcal{L}_{kin}$
Beat Distance Predictor
$\mathcal{L}_{beat}$
- 3 4 5 4 3 2 1 0 1 2 3 4 4 3 2 1 0 1 2 3 -
Beat Alignment Loss

Method

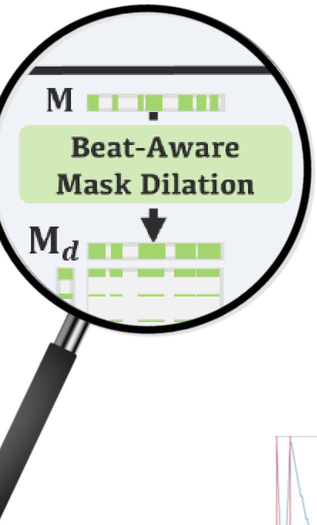

Hierarchical Multi-Condition Fusion
Sparse-Dense Fusion
Dense-Dense Fusion
$\mathcal{E}_m$
$\mathcal{E}_p$
$\mathcal{E}_b$
Beat-Aware Dilated CA
Beat-Aware Dilated CA
Beat-Aware Dilated CA
M_d
M_{attn}^{CA}
I
Beat-Aware CA Mask
× N
M
Beat-Aware Mask Dilation
M_d
C
Transformer Fusion Block
× N
t
MLP
x_0
Forward Diffusion
x_t
Linear
Self-Attention
DenseFiLM
Cross-Attention
DenseFiLM
MLP
DenseFiLM
× N
Linear
$\hat{x}_0$
$\mathcal{L}_{simple}$
$\mathcal{L}_{kin}$
$\mathcal{L}_{beat}$
Beat Distance Predictor
... 3 4 5 4 3 2 0 1 2 3 4 4 3 2 1 0 1 2 3 ...
Beat Alignment Loss

Method



Keyframe Mask

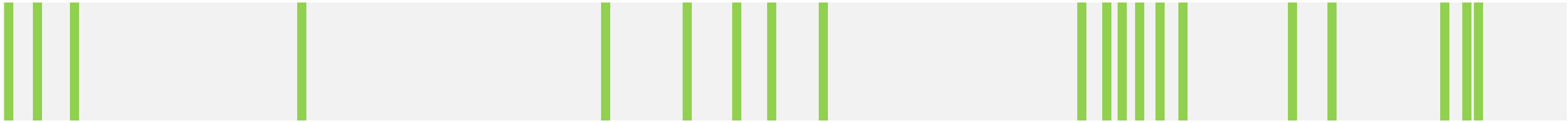
Method



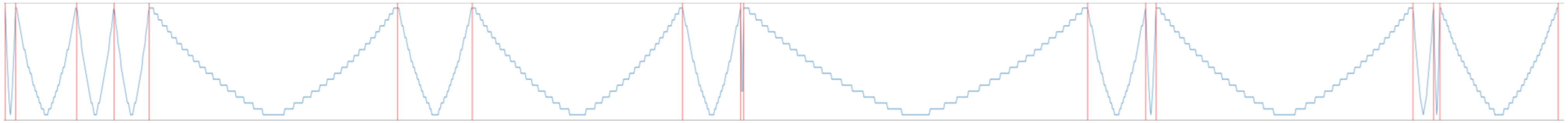
M

Beat-Aware Mask Dilation

M_d

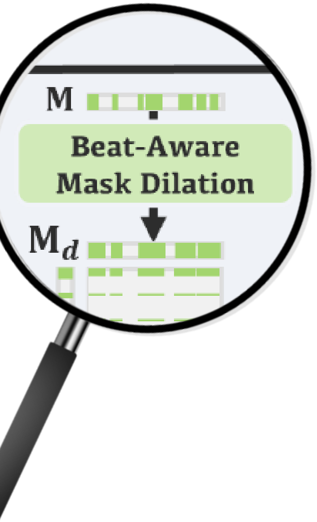


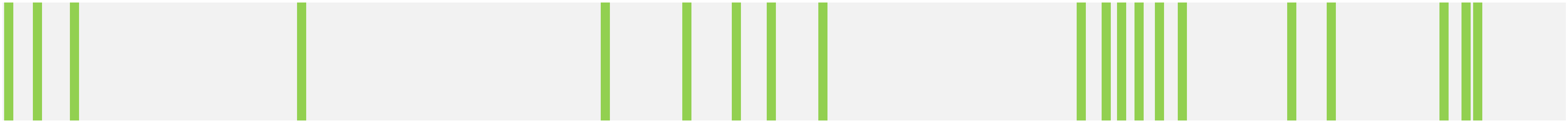
Keyframe Mask



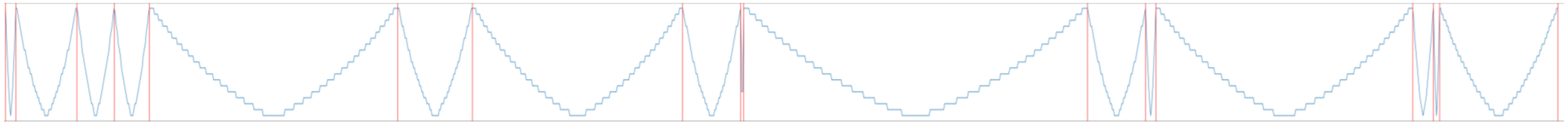
Beats

Method

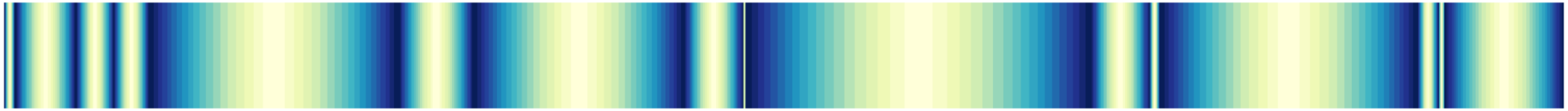




Keyframe Mask



Beats



Heatmap of Dilation Step $\boldsymbol{n}$

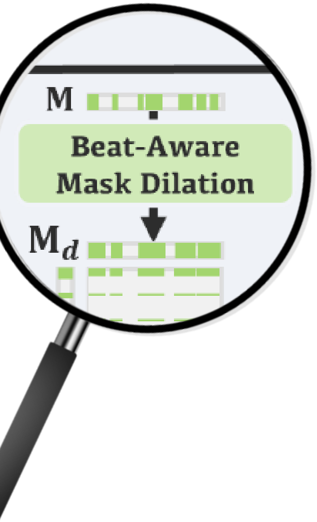
$$\boldsymbol{n} = \left\lceil \boldsymbol{s} \cdot \boldsymbol{e}^{-2\frac{\boldsymbol{b}^{i}}{\boldsymbol{d}^{i}}} \right\rceil$$

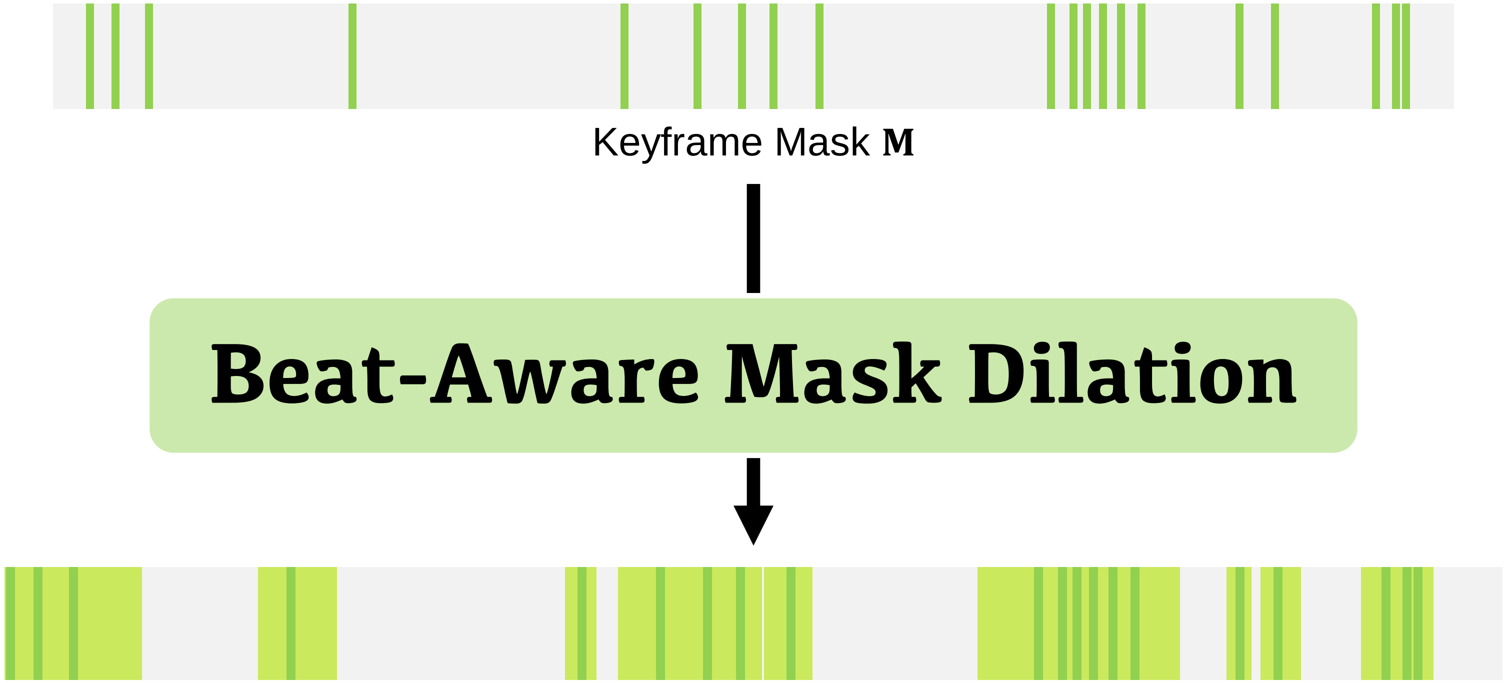
s - base dilation step,

b^i - beat distance at frame i, d^i - distance between frame i's adjacent beat frames

The closer to the beat, **the larger** the dilation step

Method





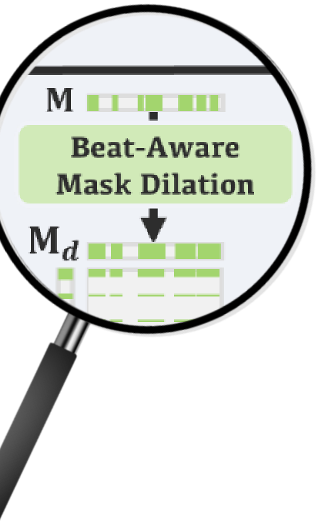
Keyframe Mask $\mathbf{M}$

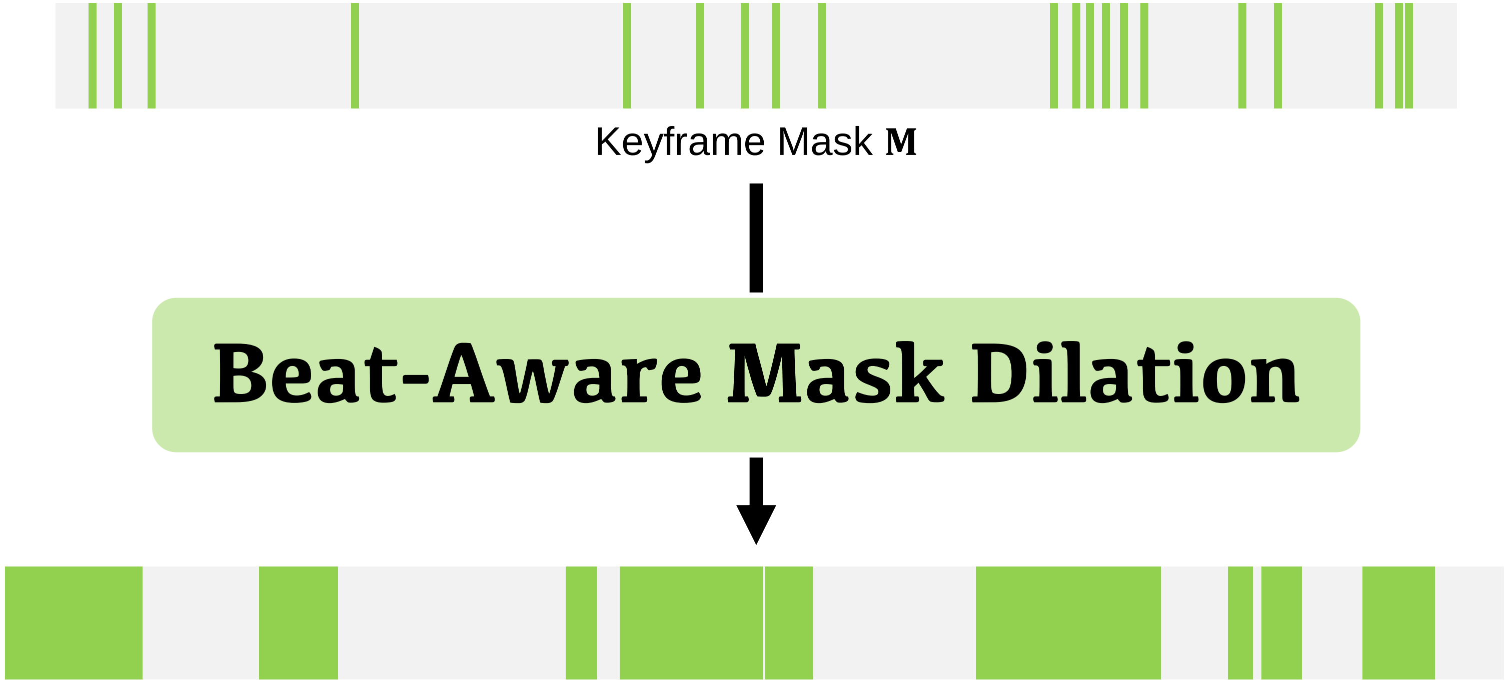
Beat-Aware Mask Dilation

Dilated Keyframe Mask $\mathbf{M}_d$

$$\mathbf{M}_d[i] = \begin{cases} \max_j \mathbf{M}[i-j], j \in \{-n, n+1, \ldots, n-1, n\} & \text{if } \mathbf{M}_i = 1, \\ \mathbf{M}_i, & \text{otherwise.} \end{cases}$$

Method

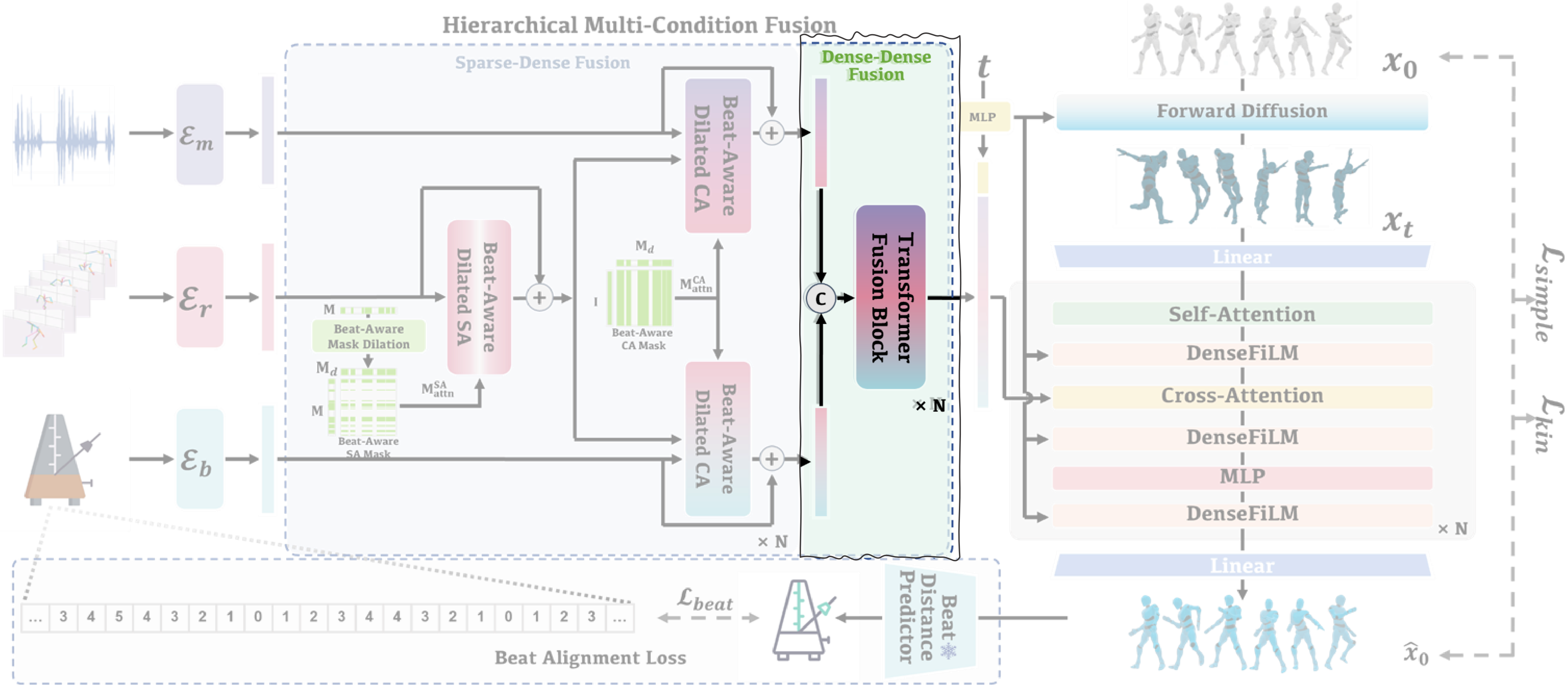




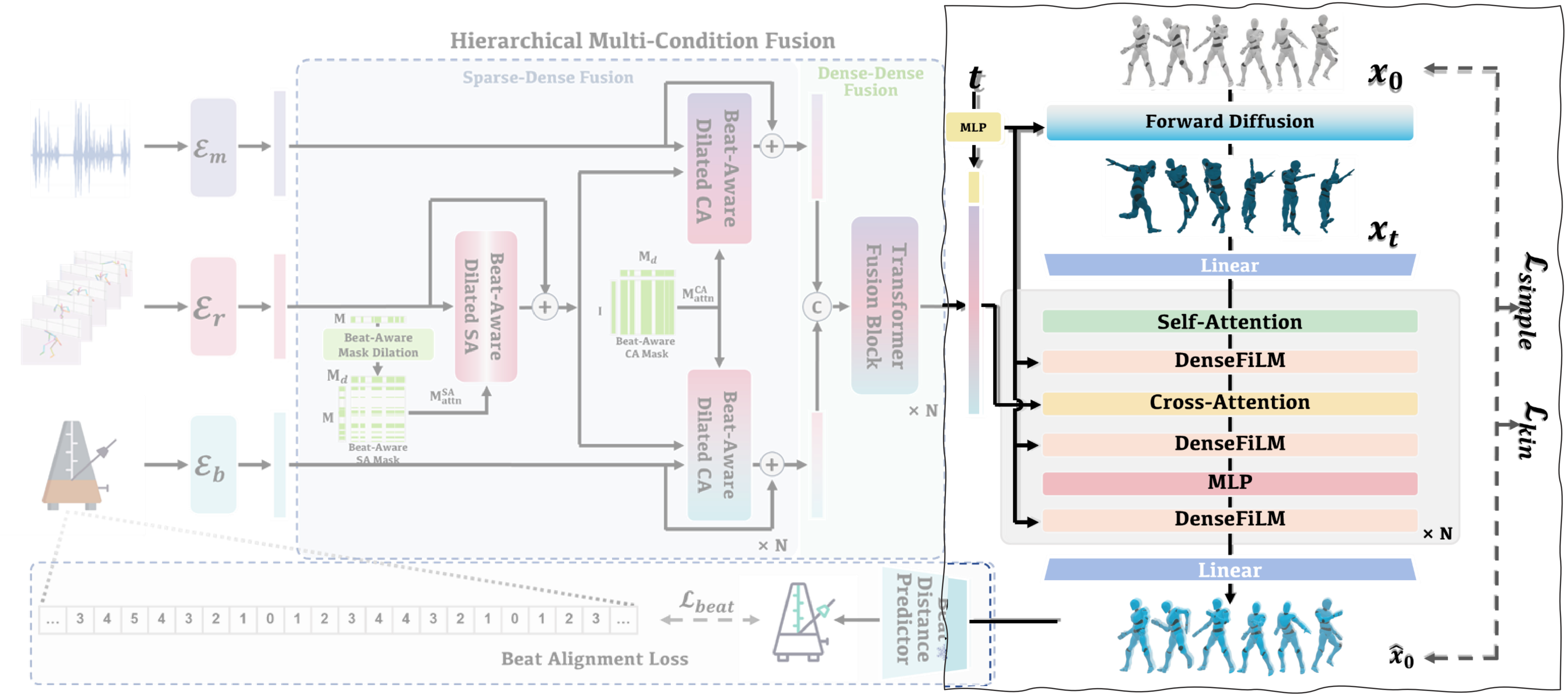
Dilated Keyframe Mask $\mathbf{M}_d$

$$\mathbf{M}_d[i] = \begin{cases} \max_j \mathbf{M}[i-j], j \in \{-n, n+1, \dots, n-1, n\} & \text{if } \mathbf{M}_i = 1, \\ \mathbf{M}_i, & \text{otherwise.} \end{cases}$$

Method

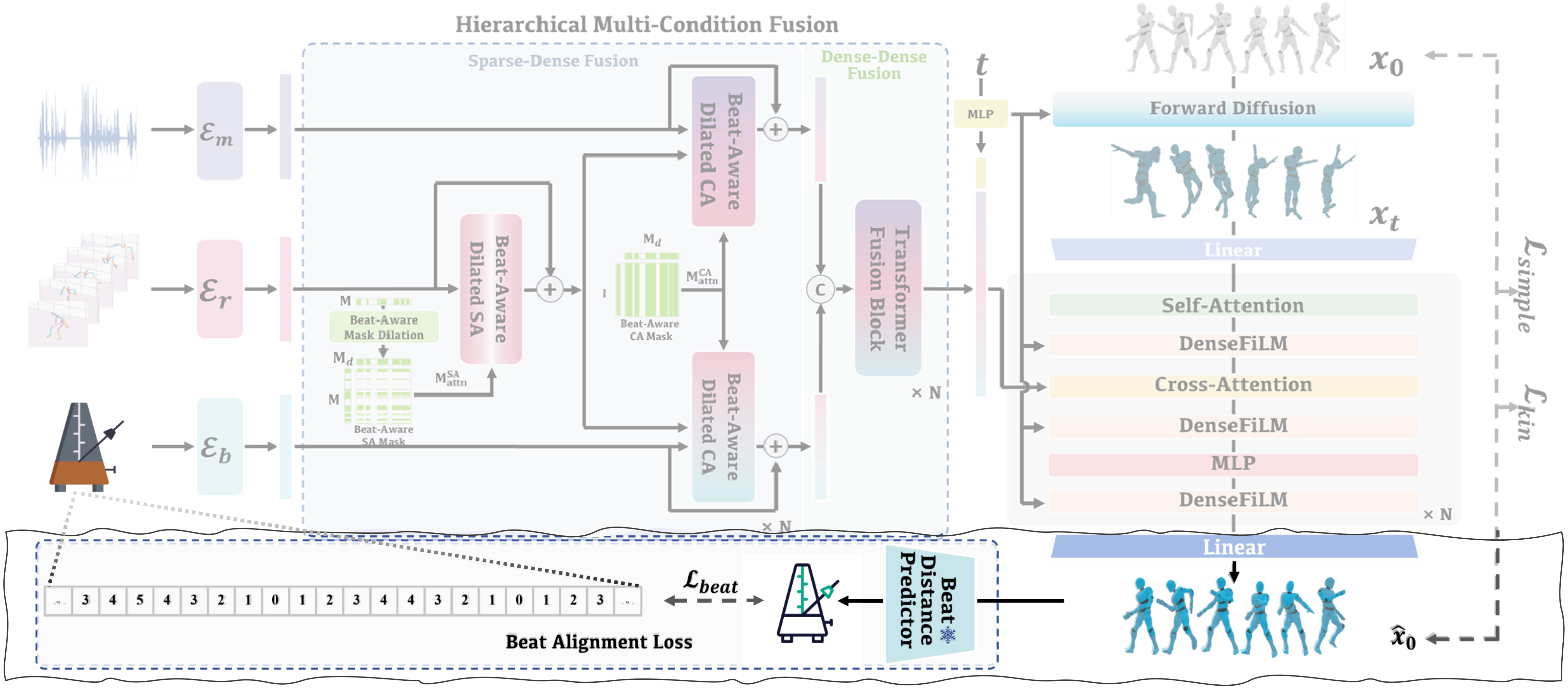

Hierarchical Multi-Condition Fusion
Sparse-Dense Fusion
Dense-Dense Fusion
$\mathcal{E}_m$
$\mathcal{E}_r$
$\mathcal{E}_b$
Beat-Aware Dilated CA
Beat-Aware Dilated SA
Beat-Aware Dilated CA
Beat-Aware Mask Dilation
Beat-Aware SA Mask
Beat-Aware CA Mask
M
M_d
M_{attn}^{SA}
M_{attn}^{CA}
C
Transformer Fusion Block
× N
× N
t
MLP
Forward Diffusion
x_0
x_t
Linear
Self-Attention
DenseFiLM
Cross-Attention
DenseFiLM
MLP
DenseFiLM
× N
Linear
$\hat{x}_0$
$\mathcal{L}_{simple}$
$\mathcal{L}_{kin}$
Beat Distance Predictor
$\mathcal{L}_{beat}$
... 3 4 5 4 3 2 1 0 1 2 3 4 4 3 2 1 0 1 2 3 ...
Beat Alignment Loss

Method



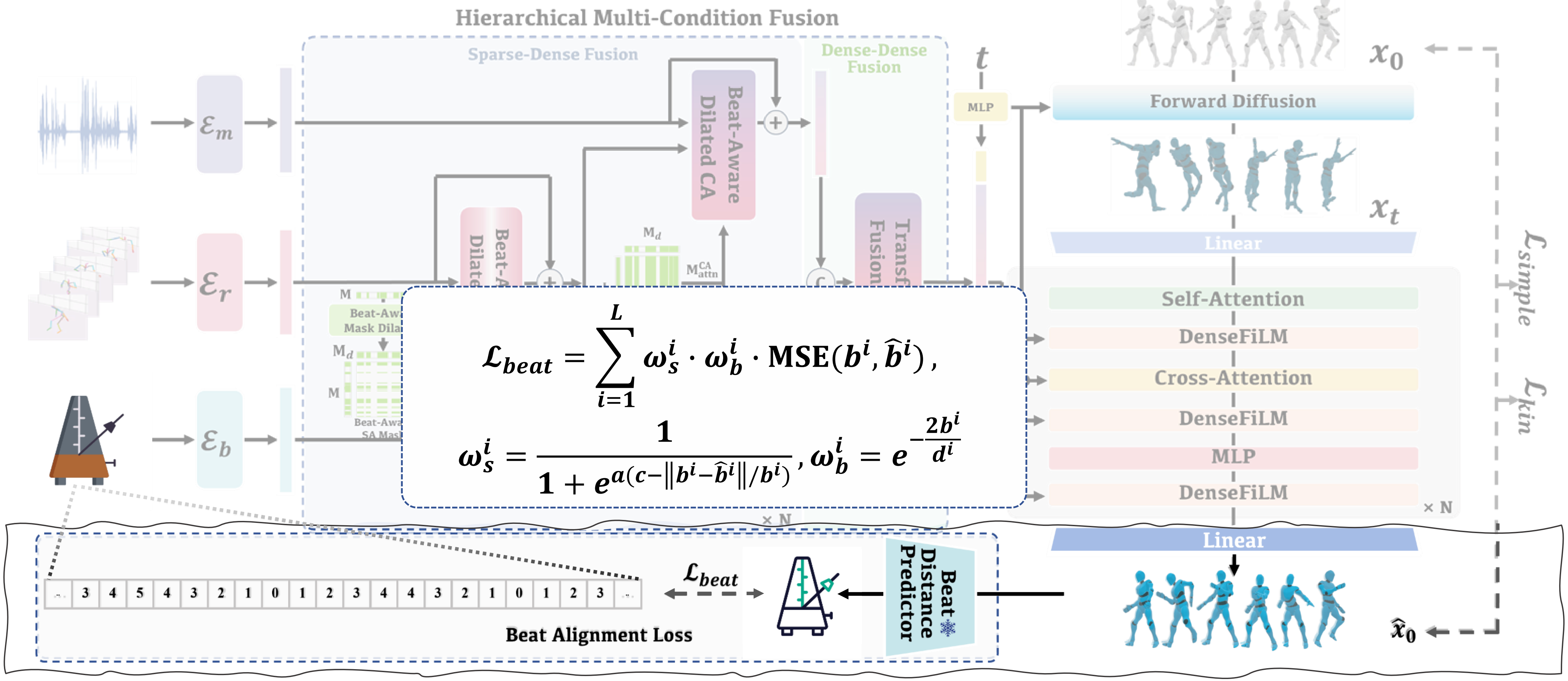
Backbone Model: Tseng, J., Castellon, R., Liu, K. “Edge: Editable dance generation from music.” CVPR. 2023.

Method



Hierarchical Multi-Condition Fusion
Sparse-Dense Fusion
Dense-Dense Fusion
$\mathcal{E}_m$
$\mathcal{E}_r$
$\mathcal{E}_b$
Beat-Aware Dilated CA
Beat-Aware Dilated SA
Beat-Aware Mask Dilation
Beat-Aware SA Mask
Beat-Aware CA Mask
M
M_d
M_{attn}^{SA}
M_{attn}^{CA}
C
Transformer Fusion Block
× N
t
MLP
Forward Diffusion
x_0
x_t
Linear
Self-Attention
DenseFiLM
Cross-Attention
DenseFiLM
MLP
DenseFiLM
× N
Linear
$\hat{x}_0$
$\mathcal{L}_{simple}$
$\mathcal{L}_{kin}$
Beat Distance Predictor
$\mathcal{L}_{beat}$
Beat Alignment Loss
.. 3 4 5 4 3 2 1 0 1 2 3 4 4 3 2 1 0 1 2 3 ..

Method

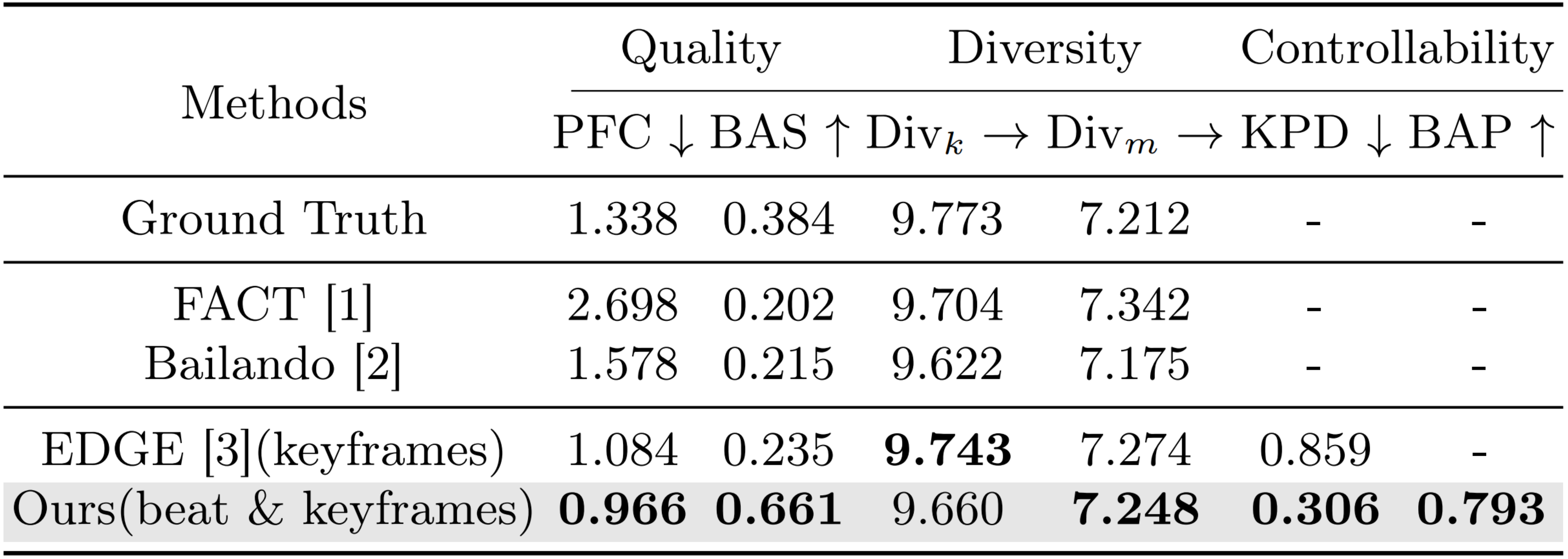


$$\mathcal{L}_{beat} = \sum_{i=1}^{L} \omega_s^i \cdot \omega_b^i \cdot \mathrm{MSE}(b^i, \hat{b}^i)\,,$$

$$\omega_s^i = \frac{1}{1 + e^{a(c - \|b^i - \hat{b}^i\|/b^i)}}, \omega_b^i = e^{-\frac{2b^i}{d^i}}$$

Experiments

Table 1: Quantitative comparisons among different methods on AIST++.



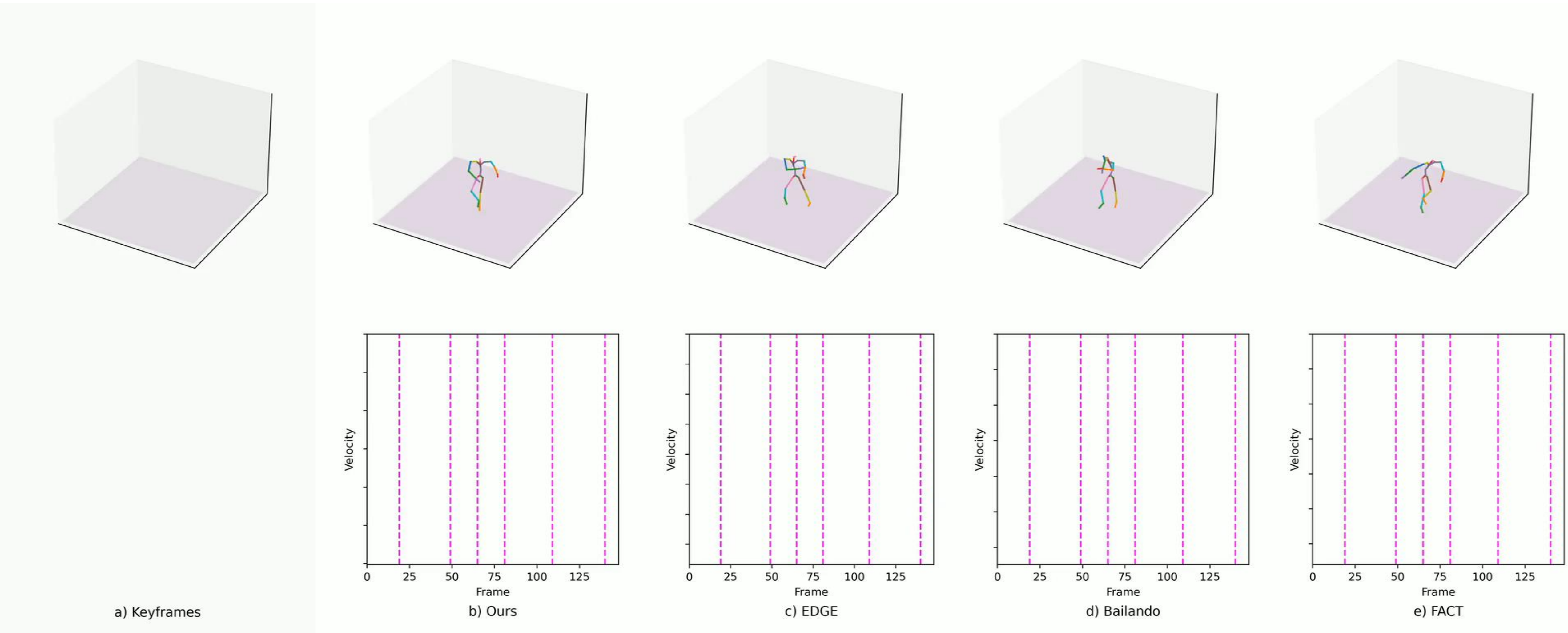
Methods	Quality		Diversity		Controllability	
	PFC ↓	BAS ↑	Div_k →	Div_m →	KPD ↓	BAP ↑
Ground Truth	1.338	0.384	9.773	7.212	-	-
FACT [1]	2.698	0.202	9.704	7.342	-	-
Bailando [2]	1.578	0.215	9.622	7.175	-	-
EDGE [3](keyframes)	1.084	0.235	**9.743**	7.274	0.859	-
Ours(beat & keyframes)	**0.966**	**0.661**	9.660	**7.248**	**0.306**	**0.793**

[1] Li, R., Yang, S., Ross, D.A., Kanazawa, A. "Ai choreographer: Music conditioned 3d dance generation with aist++." ICCV. 2021.

[2] Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C.C., Liu, Z. "Bailando: 3d dance generation by actor-critic gpt with choreographic memory." CVPR. 2022.

[3] Tseng, J., Castellon, R., Liu, K. "Edge: Editable dance generation from music." CVPR. 2023.

Experiments



a) Keyframes

b) Ours

c) EDGE

d) Bailando

e) FACT

***Dashed vertical lines denote the input beats**

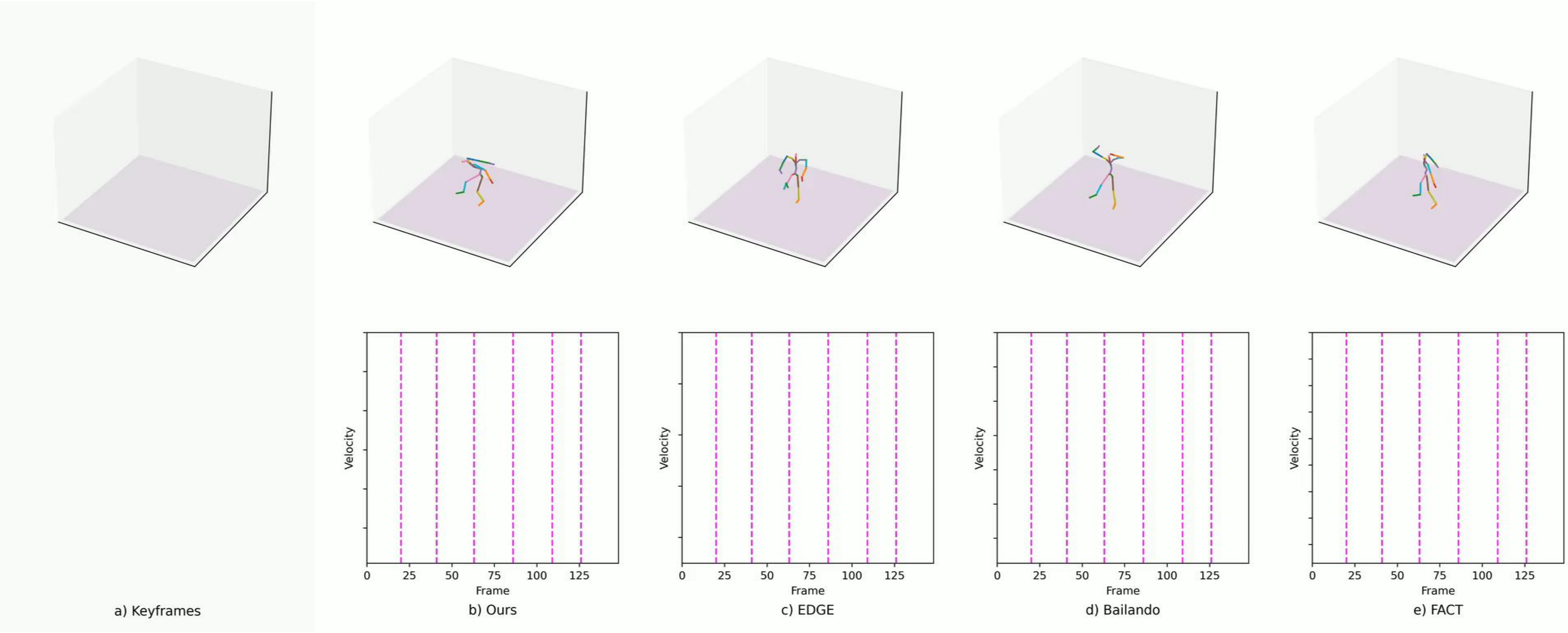
***Local minimum velocity points indicate motion beats**

[1] Li, R., Yang, S., Ross, D.A., Kanazawa, A. "Ai choreographer: Music conditioned 3d dance generation with aist++." ICCV. 2021.

[2] Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C.C., Liu, Z. "Bailando: 3d dance generation by actor-critic gpt with choreographic memory." CVPR. 2022.

[3] Tseng, J., Castellon, R., Liu, K. "Edge: Editable dance generation from music." CVPR. 2023.

Experiments



a) Keyframes b) Ours c) EDGE d) Bailando e) FACT

Dashed vertical lines denote the input beats

Local minimum velocity points indicate motion beats

[1] Li, R., Yang, S., Ross, D.A., Kanazawa, A. "Ai choreographer: Music conditioned 3d dance generation with aist++." ICCV. 2021.

[2] Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C.C., Liu, Z. "Bailando: 3d dance generation by actor-critic gpt with choreographic memory." CVPR. 2022.

[3] Tseng, J., Castellon, R., Liu, K. "Edge: Editable dance generation from music." CVPR. 2023.

Experiments

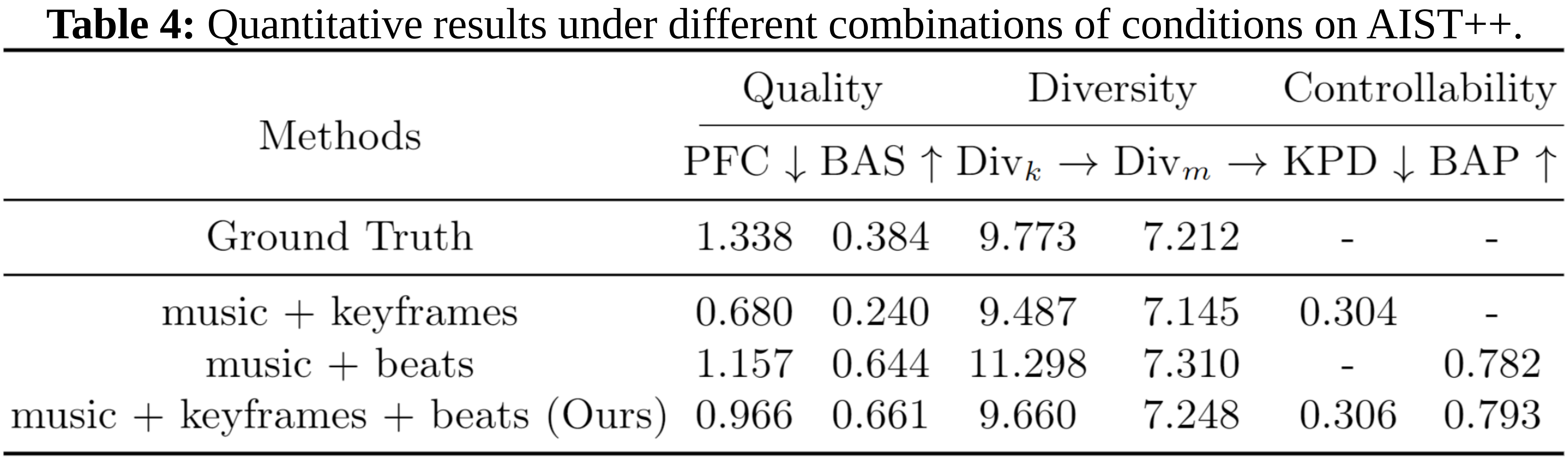


Table 4: Quantitative results under different combinations of conditions on AIST++.

Methods	Quality		Diversity		Controllability	
	PFC ↓	BAS ↑	Div_k →	Div_m →	KPD ↓	BAP ↑
Ground Truth	1.338	0.384	9.773	7.212	-	-
music + keyframes	0.680	0.240	9.487	7.145	0.304	-
music + beats	1.157	0.644	11.298	7.310	-	0.782
music + keyframes + beats (Ours)	0.966	0.661	9.660	7.248	0.306	0.793

Experiments

music + beat

music + keyframes

music + keyframes + beat

keyframes

Experiments

In the Wild

generated

keyframes

Experiments

In the Wild

generated

keyframes

Takeaways

✨ A novel multi-condition diffusion-based framework, Beat-It, for beat-synchronized and key pose-guided dance generation.

✨ A hierarchical multi-condition fusion mechanism equipped with a beat-aware dilation scheme to integrate conditions with different information sparsity.

✨ A specifically designed beat alignment loss to provide explicit guidance and supervision on motion beats.

Thank you

More demos can be found on our project page:
https://zikaihuangscut.github.io/Beat-It/