Uniform Text-Motion Generation and Editing via Diffusion Model

Ruoyu Wang, Xiang Li, Tengjiao Sun, Yangfan He, TIANYU SHI, yitingxie
Genfun.ai https://genlab3d.genfun.ai/
ty.shi@mail.utoronto.ca

Motivation

Limited to unimodal inputs and outputs

Insufficient Guidance by Textual Instruction

- Rely solely on textual instructions for guidance, lacking the capability to process motion or multimodal inputs.
- Textual instructions are often brief and ambiguous, making them insufficient to achieve the desired outcomes in many scenarios.

Restricted to motion generation

- Prevent them from performing related tasks such as motion annotation or generating text-based descriptions of motions.

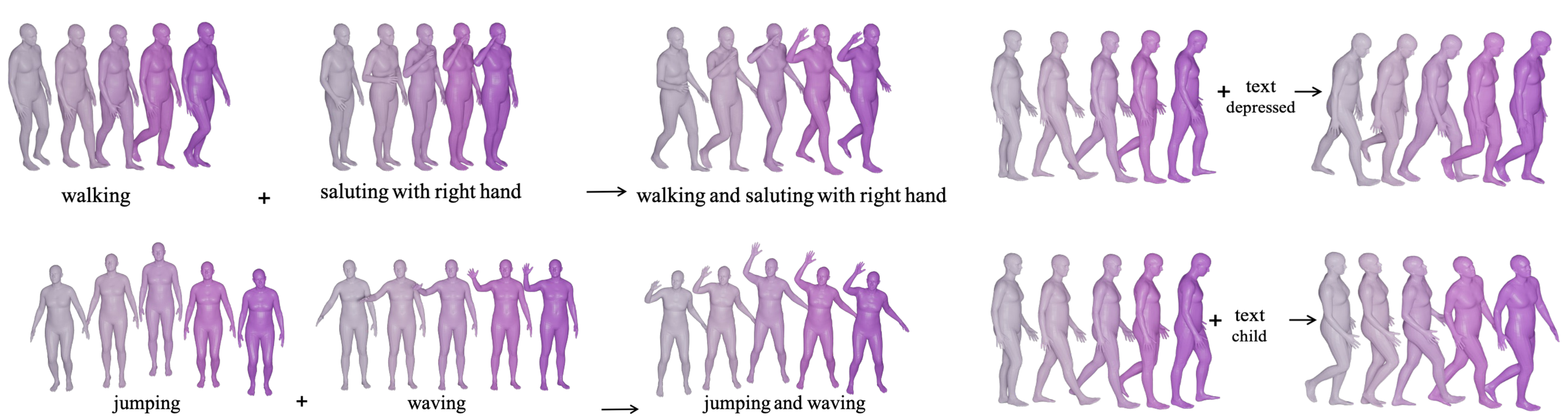
Conclusion

Overcome the limitation of single-modal inputs and outputs

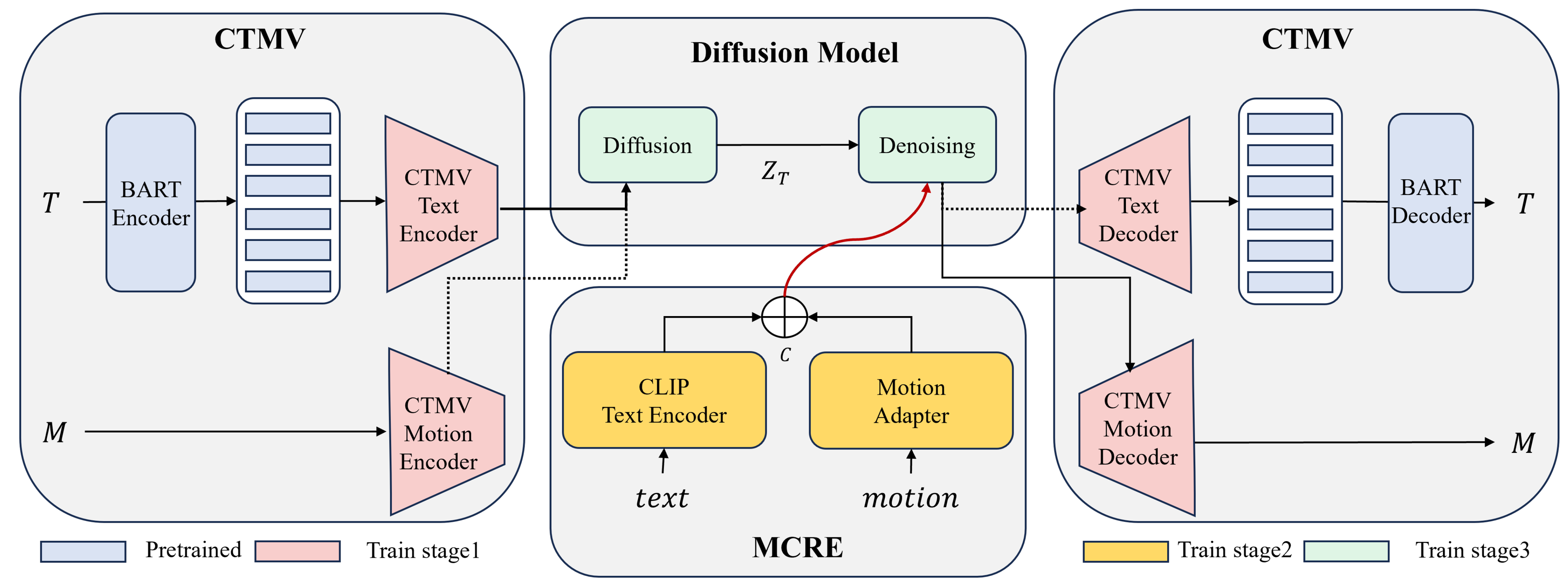
- Demonstrate advanced effectiveness and generalization across multiple tasks, including text-driven motion generation, motion captioning, motion completion, and multimodal motion editing.
- Text-Driven Motion Generation

a man is throwing
a man is shooting basketball

- Multimodal Motion Editing

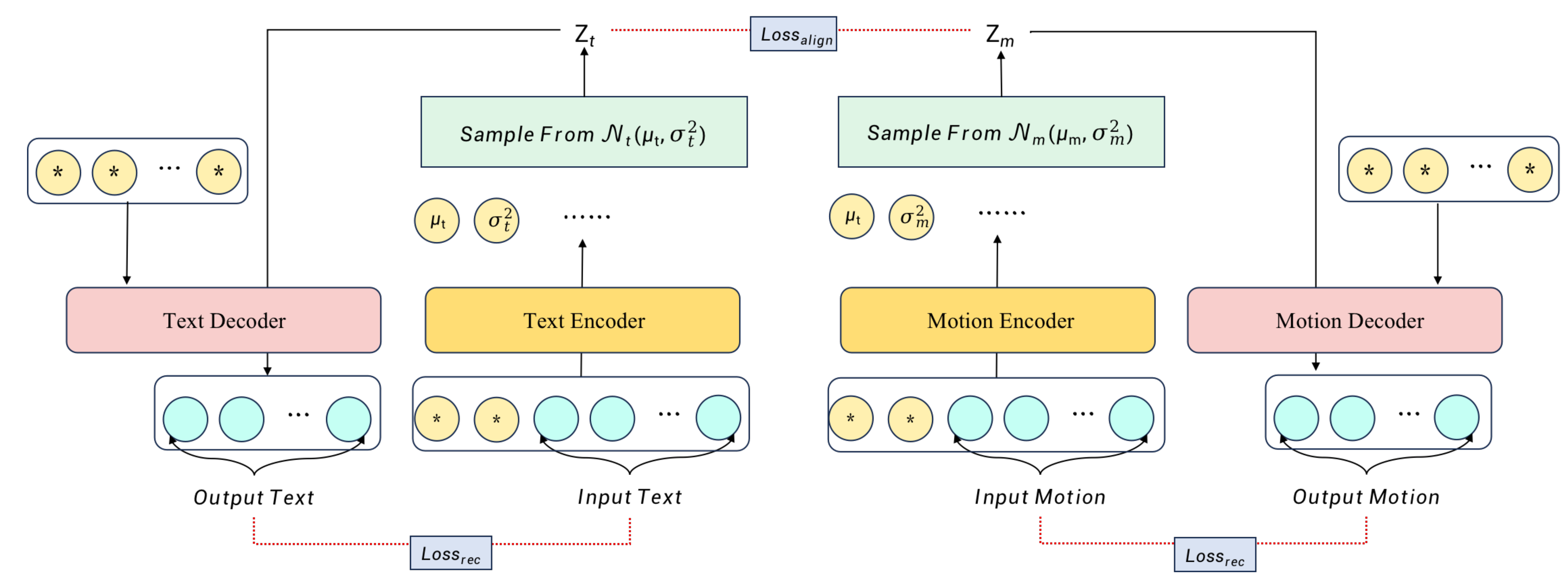


Method



Multimodal Conditional Representation and Editing (MCRE)

- Designs a motion adapter to align motion with CLIP's text representations, leveraging its rich semantic understanding.



Contrastive Text-Motion Variational Autoencoder (CTMV)

- Aligns text-motion pairs into a shared latent space via transformer-based encoders leveraging contrastive learning.
- Offloads the task of generating high-frequency details of different modalities to the autoencoder
- Enables the diffusion to focus on the high-level semantics generation.

Uniform Text-Motion Generation and Editing via Diffusion Model

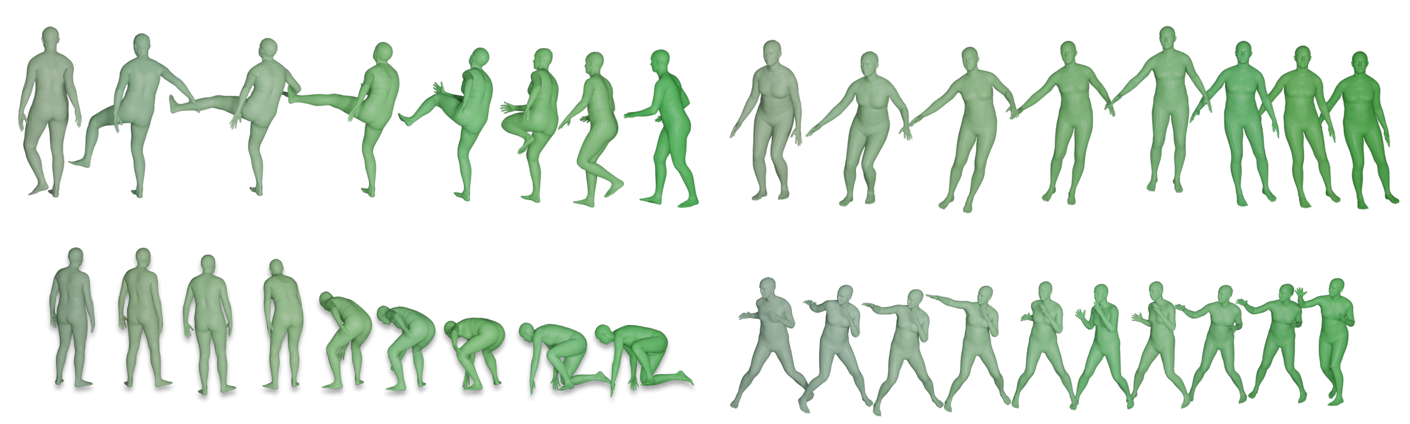
Experiments

Training Data--HumanML3D dataset

•Comprising 14,616 motions and 44,970 descriptions.

•Data span various domains daily activities, exercise, and artistic performances.

•With an average duration of 7.1 seconds per action and an average description length of 12 words.



Training Strategy

- Jointly training by training losses, which is composed of reconstruction and alignment losses.

Reconstruction Loss:

- Cross-Entropy Loss for Text
- L2 loss for Motion

Alignment Losses

- Cosine loss
- KL Loss.

Results

Demonstrate advanced effectiveness and generalization across multiple tasks

- Text-to-Motion

Table 1. Quantitative results of text-to-motion task on the HumanML3D and Motion-X test set

Method	R-Precision↑ Top1	Top2	Top3	FID↓	MM-Dist↓	Diversity→	MModality↑
"HumanML3D Test set"							
Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
MDM	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$\underline{9.559}^{\pm.086}$	$\mathbf{2.799}^{\pm.072}$
MotionDiffuse	$\underline{0.491}^{\pm.001}$	$\underline{0.681}^{\pm.001}$	$\mathbf{0.782}^{\pm.001}$	$0.630^{\pm.001}$	$\underline{3.113}^{\pm.001}$	$9.410^{\pm.049}$	$1.553^{\pm.042}$
MLD	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$\underline{0.473}^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
Ours	$\mathbf{0.499}^{\pm.003}$	$\mathbf{0.683}^{\pm.003}$	$\underline{0.780}^{\pm.002}$	$\mathbf{0.339}^{\pm.009}$	$\mathbf{3.087}^{\pm.008}$	$\mathbf{9.527}^{\pm.053}$	$\underline{2.500}^{\pm.083}$
"Motion-X Test set"							
Real	$0.509^{\pm.004}$	$0.702^{\pm.004}$	$0.794^{\pm.003}$	$0.003^{\pm.001}$	$2.995^{\pm.009}$	$9.508^{\pm.064}$	-
MDM	$0.301^{\pm.006}$	$0.477^{\pm.005}$	$0.590^{\pm.008}$	$0.645^{\pm.045}$	$5.587^{\pm.028}$	$\mathbf{9.558}^{\pm.087}$	$\mathbf{2.788}^{\pm.073}$
MotionDiffuse	$0.470^{\pm.002}$	$0.660^{\pm.002}$	$0.757^{\pm.002}$	$0.638^{\pm.002}$	$\underline{3.232}^{\pm.002}$	$9.411^{\pm.050}$	$1.524^{\pm.043}$
MLD	$\underline{0.475}^{\pm.004}$	$\underline{0.665}^{\pm.004}$	$\underline{0.764}^{\pm.003}$	$\underline{0.484}^{\pm.014}$	$3.255^{\pm.011}$	$9.723^{\pm.083}$	$2.312^{\pm.080}$
Ours	$\mathbf{0.490}^{\pm.004}$	$\mathbf{0.677}^{\pm.004}$	$\mathbf{0.772}^{\pm.003}$	$\mathbf{0.355}^{\pm.010}$	$\mathbf{3.098}^{\pm.009}$	$\underline{9.560}^{\pm.054}$	$\underline{2.401}^{\pm.084}$

- Motion-to-Text

Table 2. Quantitative results of motion-to-text task on the HumanML3D and Motion-X test set

Method	R-Precision↑ Top1	Top3	MM-Dist↓	$Length_{avg}$ ↑	Bleu@1↑	Bleu@4↑	Rouge↑	Cider↑	BertScore↑
"HumanML3D Test set"									
Real	0.523	0.828	2.901	12.750	-	-	-	-	-
TM2T	0.516	0.823	2.935	10.670	<u>48.900</u>	7.000	**38.100**	16.800	32.200
MotionGPT	**0.543**	**0.827**	**2.821**	**13.040**	48.200	**12.470**	<u>37.400</u>	<u>29.200</u>	<u>32.400</u>
Ours	<u>0.520</u>	<u>0.825</u>	<u>2.878</u>	<u>13.000</u>	**49.300**	<u>11.700</u>	35.000	**30.300**	**32.800**
"Motion-X Test set"									
Real	0.520	0.821	2.892	17.250	-	-	-	-	-
TM2T	0.484	0.803	2.975	11.901	46.500	6.781	35.903	15.201	29.909
MotionGPT	**0.518**	<u>0.817</u>	**2.858**	**14.340**	<u>47.800</u>	**11.890**	**36.202**	<u>27.601</u>	<u>31.009</u>
Ours	<u>0.510</u>	**0.820**	<u>2.886</u>	<u>13.890</u>	**48.302**	<u>11.025</u>	<u>34.800</u>	**28.998**	**31.198**

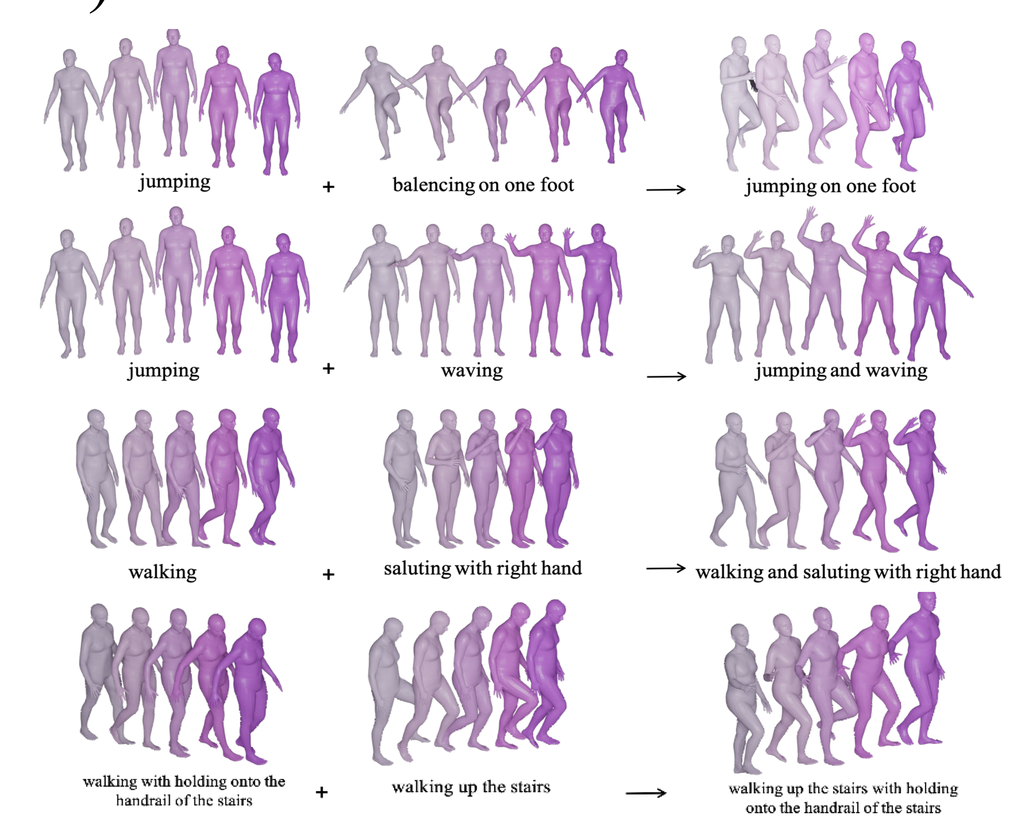
- Motion Completion

Table 4. Quantitative results of motion completion task on the HumanML3D and Motion-X test set

Method	Motion Prediction FID↓	Diversity→	ADE↓	FDE↓	Motion-In-between FID↓	Diversity↑	ADE↓
"HumanML3D Test set"							
Real	0.002	9.503	-	-	0.002	9.503	-
MDM	6.031	7.813	5.446	8.561	2.698	8.420	3.787
Ours	**1.702**	**9.001**	**4.740**	**6.670**	**1.203**	**9.600**	**3.669**
"Motion-X Test set"							
Real	0.003	9.508	-	-	0.003	9.508	-
MDM	8.931	7.783	7.846	10.021	4.398	8.150	2.987
Ours	**2.102**	**8.901**	**5.940**	**7.970**	**1.583**	**9.230**	**3.042**

- Multimodal Motion Editing

1) motion based



2) multimodal based