Text-to-Motion Synthesis using Discrete Diffusion Model

Ankur Chemburkar achemburkar@ict.usc.edu
Shuhong Lu slu@ict.usc.edu
Andrew Feng feng@ict.usc.edu

Institute for Creative Technologies, University of Southern California, Los Angeles, USA

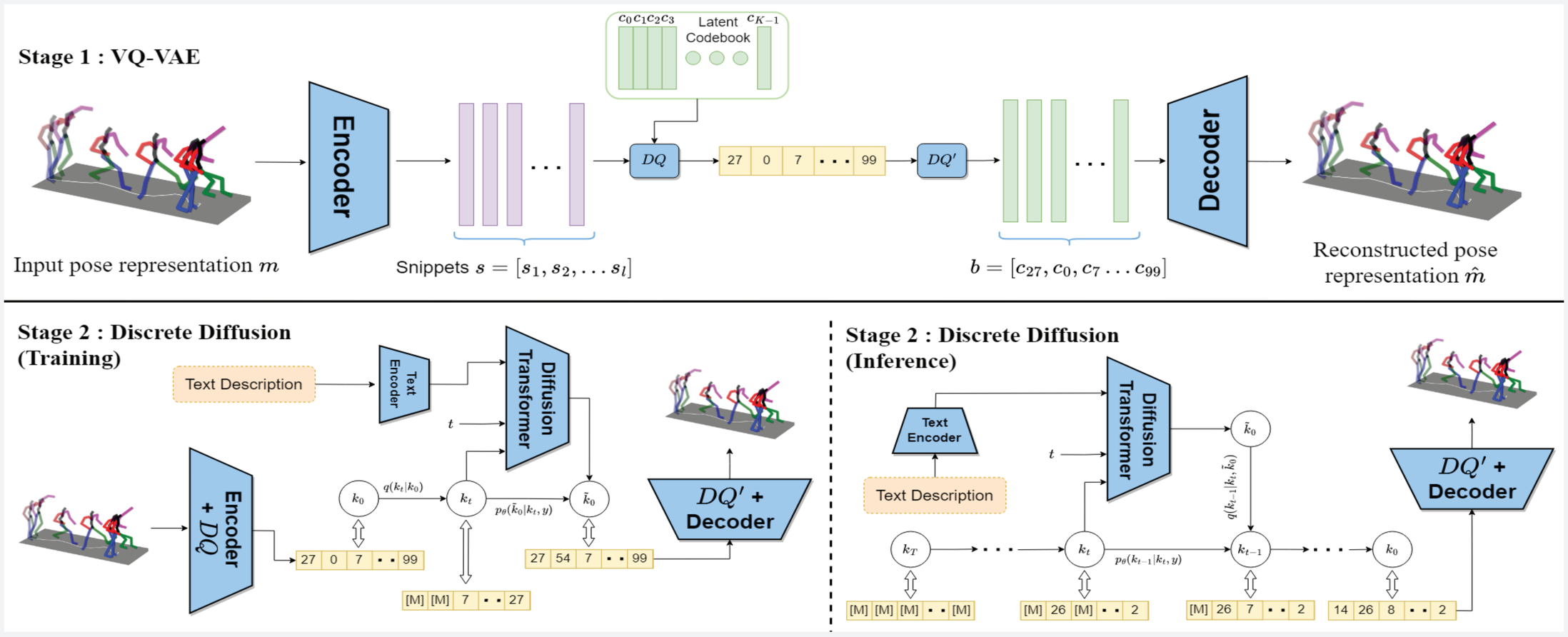
Motivation

- Synthesizing well aligned human motions based on input conditions is an essential task with many applications in gaming, simulation, and virtual reality.
- For a many-to-many task like text-motion synthesis, probabilistic methods such as Diffusion are necessary. Despite, excellent generative abilities, diffusion is expensive on raw motions and so faster alternatives are needed.
- Hence, we use latent space Discrete Diffusion with the assumption that human motion can be efficiently represented by discrete sequence of small motion snippets.

Contributions

- We model the text-to-motion generation task as a discrete denoising diffusion probablistic model, which allows reduced diffusion steps for faster inferences while producing high quality results.
- Evaluated our method (MoDDM) in a comparison with state-of-the-art methods using both objective metrics and subjective user study. The results demonstrated that our method outperforms the previous methods in both motion quality and text-to-motion matching accuracy.

Two-stage Architecture consisting of VQ-VAE and Discrete Diffusion Model

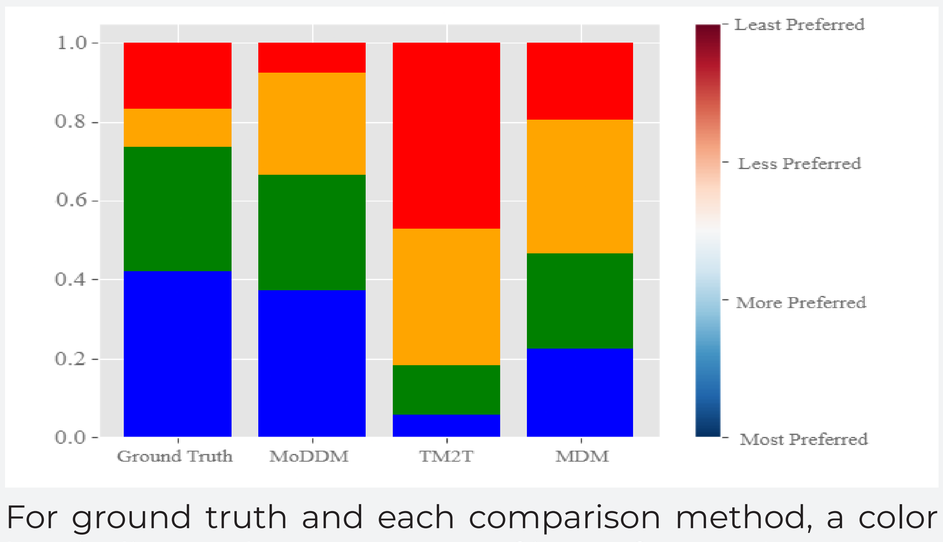


Quantitative Evaluation on HumanML3D Test Set

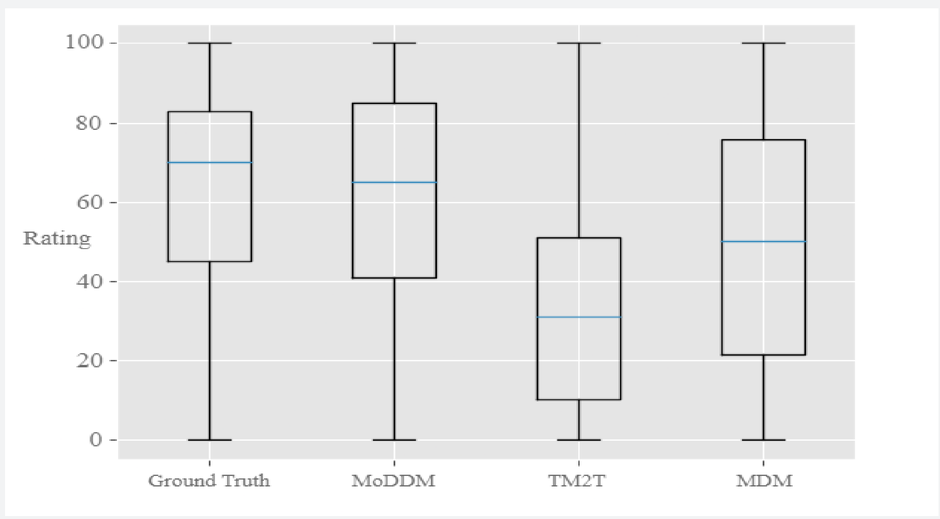
Methods	R Precision ↑			FID ↓	MultiModal Dist ↓	Diversity →
	Top 1	Top 2	Top 3			
Real Motions	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$
Seq2Seq [21]	$0.180^{\pm.002}$	$0.300^{\pm.002}$	$0.396^{\pm.002}$	$11.75^{\pm.035}$	$5.529^{\pm.007}$	$6.223^{\pm.061}$
Language2Pose [9]	$0.246^{\pm.002}$	$0.387^{\pm.002}$	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$
MDM [53]	-	-	$0.611^{\pm.007}$	$\underline{0.544}^{\pm.044}$	$5.566^{\pm.027}$	$\mathbf{9.559}^{\pm.086}$
TM2T [10]	$\underline{0.424}^{\pm.003}$	$\mathbf{0.618}^{\pm.003}$	$\mathbf{0.729}^{\pm.002}$	$1.501^{\pm.017}$	$\mathbf{3.467}^{\pm.011}$	$8.589^{\pm.076}$
MoDDM (Ours)	$\mathbf{0.425}^{\pm.004}$	$\underline{0.615}^{\pm.004}$	$\underline{0.713}^{\pm.003}$	$\mathbf{0.294}^{\pm.006}$	$\underline{3.553}^{\pm.009}$	$\underline{9.178}^{\pm.093}$

± indicates 95% confidence interval, and → means the closer to Real motions the better. Bold face indicates the best result, while underscore refers to the second best.

Evaluation of Motion Alignment to Text



For ground truth and each comparison method, a color bar indicates the percentage of its preference levels.



Boxes cover 25th and 75th percentiles, and whiskers represent the 5th and 95th percentiles. Box notches represent median values.

Qualitative Comparisons on HumanML3D Test Set