Supplementary Materials for MoDDM: 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

1 Implementation Details

Our codebase is implemented using PyTorch Lightning and Hydra frameworks. All models were trained and evaluated using 2 NVIDIA GeForce GTX 1080 GPUs. Models of both the datasets - HumanML3D and KIT are trained with a batch size of 128. Evaluations are done with a batch size of 32. For training the discrete diffusion model, Adam optimizer is used with a learning rate of $2e^{-3}$.

The pose representation dimensions for HumanML3D and KIT-ML are 263 and 251 respectively. The codebook size for the Vector Quantized Variational Autoencoder is 1024 with the embedding dimension being 64. The content sequence length for diffusion is 49 (number of motion snippets). Condition sequence length is 77 and hidden dimension of the text condition is 512.

1.1 VQ-VAE Details

We briefly experimented with the VQ-VAE architecture as shown in Table 1. This architecture also includes attention blocks coupled with Resnet Blocks which are not presented in the encoder and decoder of TM2T. After our initial experiments, we observed that the TM2T VQ-VAE displayed better motion-to-motion reconstruction. Hence, as VQ-VAE has already been studied extensively before, we decided to proceed with the TM2T architecture with pretrained checkpoints for further experiments and finetuning for both datasets.

2 User Study Details

The user study experiment was conducted on an AWS EC2 instance. Participants were recruited through the Prolific crowd-sourcing platform and were provided with the experiment link through the platform.

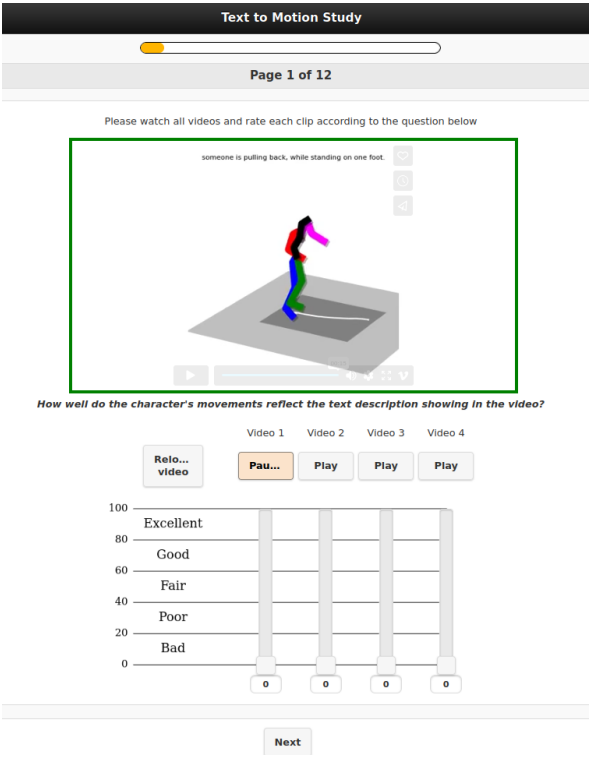

Figure 1: Interface used in the user studies.

For each method (our method, TM2T, and MDM), we randomly selected 10 text sentences from the test set and used them to generate motion videos. These generated videos, along with the corresponding ground truth motion, were randomly shuffled and shown to 40 participants (20 men and 20 women) with a maximum age of 60 and a minimum age of 19. Each participant compared and rated each video on a scale of 0 to 100.

The user interface used in the user studies is shown in Figure. The question posed to participants was "How well do the character's movements reflect the text description showing in the video?" Each participant was shown 10 sets of videos, with 4 videos in each set, generated by different methods using the same conditioned text.

To ensure data validity, we incorporated an attention check process that occasionally presented overlay text on the videos, requiring participants to select a specific value. This helped prevent invalid submissions. Two additional pages with attention checks were included in the study, and the scores from those pages were not considered in the final results. Among the 40 participants, two participants failed the attention check, and their submissions were excluded from the result analysis.

The average study time for each participant was 15 minutes, and participants were compensated with 3 dollars for their time. There was one set of videos where the text was not well related to the ground truth motion, so it was excluded from the results. For each participant, there were 9 valid ratings for each method, which were used in the final analysis. Our method (MoDDM) received the highest average rating of 61.1, while MDM received a rating of 49.1 and TM2T received a rating of 33.3. Figure 2 shows the distributions of participants'

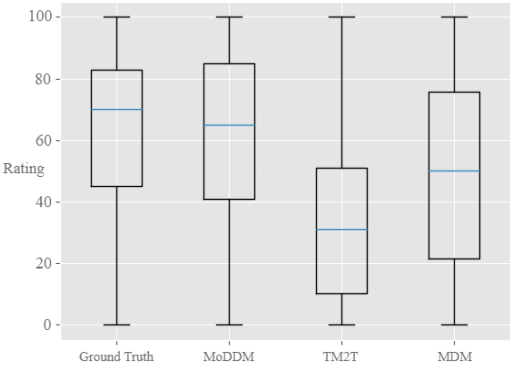

Figure 2: Evaluation results of motion appropriateness to input text. Box plots for ratings of ground truth and 3 motion generation methods. Boxes cover 25th and 75th percentiles, and whiskers represent the 5th and 95th percentiles. Box notches represent median values.

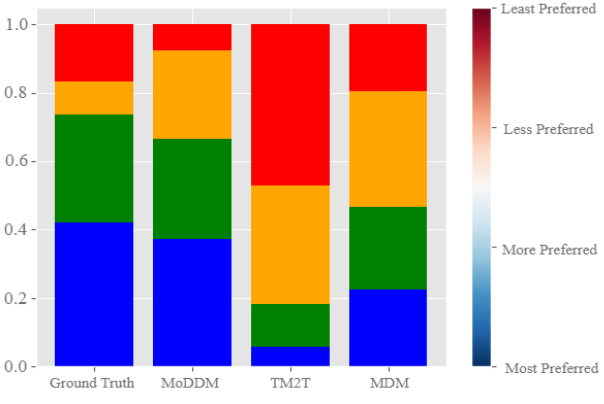

Figure 3: Evaluation results of motion appropriateness to input text. Quantitative evaluation of user preference among the generated motions. For ground truth and each comparison method, a color bar (from blue to red) indicates the percentage of its preference levels (from most to least preferred)

ratings.

When considering the ranking for each video, our method (MoDDM) received the most number of top-1 ratings (128), followed by MDM with 77, TM2T with 20, and the ground truth motion with 144. Entries with the same scores were considered to have the same top-k ranks on the corresponding level. Figure 3 shows the results of participants' preference.

Modules	Architecture
Encoder	Conv1d(pose_dim, 64, kernel size=(3,), stride=(1,), padding=(1,)) (ResBlock 64-64): Sequential((0): Sigmoid() (1): Conv1d(64, 64, kernel size=(3,), stride=(1,), padding=(1,)) (2): Sigmoid() (3): Conv1d(64, 64, kernel size=(3,), stride=(1,), padding=(1,))) (ResBlock 64-64) (Downsample 64): Sequential((0): Conv1d(64, 64, kernel size=(3,), stride=(2,), padding=(0,)) (1): AvgPool1d(kernel size=2, stride=2) (ResBlock 64-128) (ResBlock 128-128) (Downsample 128) (ResBlock 128-256) (AttnBlock 256): AttLayer((W q): Conv1d(256, 256, kernel size=(1,), stride=(1,), padding=(0,)) (W k): Conv1d(256, 256, kernel size=(1,), stride=(1,), padding=(0,)) (W v): Conv1d(256, 256, kernel size=(1,), stride=(1,), padding=(0,)) (softmax): Softmax(dim=2) Conv1d(256, 256, kernel size=(1,), stride=(1,), padding=(0,))) (ResBlock 256-256) (AttnBlock 256) (ResBlock 256-256) (AttnBlock 256) (ResBlock 256-256) Sigmoid() Conv1d(256, 64, kernel size=(3,), stride=(1,), padding=(1,)))
Decoder	Conv1d(64, 256, kernel size=(3,), stride=(1,), padding=(1,)) (ResBlock 256-256) (AttnBlock 256) (ResBlock 256-256) (ResBlock 256-256) (ResBlock 256-256) (ResBlock 256-256) (Upsample 256): Sequential((0): Interpolate(scale factor=2.0, mode=nearest) (1): Conv1d(256, 256, kernel size=(3,), stride=(1,), padding=(1,)) (ResBlock 256-128) (ResBlock 128-128) (ResBlock 128-128) (Upsample 128) (ResBlock 128-64) (AttnBlock 64) (ResBlock 64-64) (AttnBlock 64) (ResBlock 64-64) (AttnBlock 64) Sigmoid() Conv1d(64, pose_dim, kernel size=(3,), stride=(1,), padding=(1,)))

Table 1: Model architecure for VQ-VAE experimentation.