

T2LM: Long-Term 3D Human Motion Generation from Multiple Sentences

Supplementary Material

1. Qualitative result

In this section, we provide qualitative comparisons to previous methods and the visualized videos of the generated long-term motions from our model.

Qualitative comparison. We first present the side-by-side qualitative comparison with TEACH [1] and Double-Take [2]. Generated videos are stored in name format of “*qualitative_comparison_*.mp4*” The motions generated from the previous methods are acquired from their product page, and we append the motions we generated to the bottom. The qualitative comparisons clearly show that our T2LM generates a higher-quality long-term motion. Specifically, actions from our method are clearer, better follow the given description, and the feet are more stable on the ground. Moreover, our transitions are smoother without an unrealistic gap and realistically connect between consecutive actions. Our T2LM generates continuous and realistic walking movement from sequential inputs “*walk*” and “*walk straight*” in “*qualitative_comparison_2.mp4*”, while the previous methods generate unrealistic stopping in between walking. Remarkably, our method generates fine transitions between semantically and physically far actions such as “*lift up*” and “*walk*”.

Long-term motion from T2LM. Generated videos are stored in name format of “*ours_*.mp4*” Visualized long-term motions are generated from input text sequences of length 6. Note that our method can deal with *arbitrarily* long input sequences. Each action is rendered in different colors, and the input sentences are displayed in the bottom-left of the video. The visualization shows that our T2LM generates realistic long-term motion with high-quality actions and realistic transitions. Remarkably, the actions are of high quality and precisely follow the descriptions even for complex inputs such as “*the toon has both arms raised at an angle above their head, as to be in the squatting motion for exercise.*” and “*a person jogs around in a semi-circle and then back, before walking.*”.

2. Ablation study

This section provides an extra ablation study on learning rates and the number of layers in ResNet1D and Text Encoder.

Learning rates. Tab. S1 presents the experiment on the learning rate for the training. For both VQVAE (1st block) and Text Encoder (2nd block), our observation indicates that the training of our model is sensitive to the decision of the learning rate. More specifically, FID, TS-FID, and Diversity are highly degraded when using inappropriate learning

LR	R-Prec.↑	FID↓	Diversity↑	TS-FID↓
1e-4	0.454	0.594	9.759	1.807
2e-4 (Ours)	0.445	0.457	10.537	1.389
3e-4	0.452	0.634	9.560	1.842
4e-4	0.451	0.700	9.620	1.770
1e-4	0.435	0.815	9.992	1.886
2e-4	0.437	0.706	9.741	1.829
3e-4 (Ours)	0.445	0.457	10.537	1.389
4e-4	0.449	0.502	10.110	1.615

Table S1. **Ablation study on learning rates.** We ablate the performance with respect to the learning rates on VQVAE and Text Encoder. The first block shows the result of alternative learning rates on VQVAE, and the second block presents the experiment on Text Encoder.

# Layers	R-Prec.↑	FID↓	Diversity↑	TS-FID↓
VQ 2	0.442	0.629	9.993	1.412
VQ 3	0.450	0.586	10.039	1.458
VQ 4 (Ours)	0.445	0.457	10.537	1.389
TE 2	0.442	0.473	9.960	1.483
TE 3	0.452	0.458	9.762	1.481
TE 4 (Ours)	0.445	0.457	10.537	1.389

Table S2. **Ablation study on a number of layers.** We ablate the performance with respect to the number of layers in VQVAE and Text Encoder. The first block shows the result of the experiment on a number of 1D convolutional layers in ResNet1D. The second block presents the number of layers in the Text Encoder.

rates for both VQVAE and Text Encoder. Given these results, we chose to train our model with the learning rates of 2e-4 and 3e-4 for VQVAE and Text Encoder, respectively.

Number of layers. Tab. S2 shows the quantitative result of the experiment on the number of layers in ResNet1D and Text Encoder. We have observed that increasing the number of layers leads to better performance. Specifically, while R-Precision does not show noticeable change, the performance measured by FID, Diversity, and TS-FID gets better with more layers. Given these results, we chose to form our model with 4 layers in ResNet1D, and 4 layers in the Transformer-based Text Encoder, respectively.

References

- [1] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. TEACH: Temporal Action Compositions for 3D Humans. In *3DV*, 2022. 1
- [2] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 1