MoMask: Generative Masked Modeling of 3D Human Motions

Supplementary Material

Methods	R Precision↑		FID↓	MMDist↓
	Top 1	Top 3		
Ours (B)	$\mathbf{0.521}^{\pm.002}$	$\mathbf{0.807}^{\pm.002}$	$\mathbf{0.045}^{\pm.002}$	$\mathbf{2.958}^{\pm.008}$
Ours (U)	$0.514^{\pm.003}$	$0.805^{\pm.003}$	$0.210^{\pm.008}$	$3.002^{\pm.009}$

Table 3. **Ablation of bidiretionality on HumanML3D.** We conduct an ablation study by employing a *bidirectional* Transformer encoder (B) and a *unidirectional* Transformer decoder (U) seperately as the backbone in our MoMask framework.

Ablation on bidirectionality. In Table 3, we assess the performance of MoMask using a unidirectional Transformer Decoder as the backbone (Ours (U)). In this configuration, causal attention is employed, and the model can only attend to previous positions in the sequence. This stands in contrast to the design presented in the paper, which utilizes a bidirectional Transformer Encoder (Ours (B)). The experiment results highlight the significance of bidirectionality in the MoMask framework. Additionally, it's worth noting that the baseline of T2M-GPT [49] represents a state-of-the-art unidirectional approach.

Which components are trained together? RVQ-VAE, M-Transformer, and R-Transformer are trained independently. RVQ-VAE is trained first and then kept frozen during training of the other two generative models.

Shared parameters in R-transformer. In the R-Transformer, codebook embeddings (in Fig 1.a) are no longer utilized due to observed suboptimal performance. With V+1 (0:V) vector quantization (VQ) level of tokens, R-Transformer includes V input token embedding layers for the 0:V-1 VQ levels, and V output prediction layers for the 1:V VQ levels. The tokens predicted by the j-th output prediction layer become the input for the $(j+1)$-th token embedding layer, which proceeds to predict the next-level tokens. Because both the j-th prediction layer and the $(j+1)$-th embedding layer handle the same token level, their linear projection weights can be shared for efficient learning, akin to BERT [10] and VALL-E [45].

Learning R-transformer layers. In the training process, we independently sample a random residual quantization (RQ) layer for each data point within a batch. Given our batch size of 64, it's highly likely that the sampled RQ layers in a batch cover all RQ levels. Although Eq.4 for a single data point only involves n terms each time, the optimization is actually conducted across all RQ levels with batch data.

Metric of diversity. Evaluation results for the "diversity" metric are provided in Tab. 4. It's essential to note that the "diversity" metric assesses overall diversity across the entire test set, specifically designed to detect generation of similar content regardless of text inputs.

Numbers mismatch. All results in Table 2 (in paper) are based on the default setting of **18** inference steps. However, in our subsequent sweeping experiment on inference steps (see Figure 7), we discovered that **10** steps of inference performed slightly better. Consequently, we report these final results in Table 1 (in paper).

Datasets	Real.	T2M [15]	MDM [42]	T2M-GPT [49]	MotionGPT [21]	Ours
HML3D	$9.503^{\pm.07}$	$9.175^{\pm.08}$	$9.559^{\pm.09}$	$9.722^{\pm.08}$	$\mathbf{9.528}^{\pm.07}$	$9.624^{\pm.08}$
KITML	$11.08^{\pm.10}$	$10.72^{\pm.14}$	$10.84^{\pm.10}$	$\mathbf{10.92}^{\pm.10}$	-	$10.78^{\pm.08}$

Table 4. **Comparison of diversity.**

Discussion of higher VQ layer (codebook size, learned contents). Table 2 indicates a constant increase in reconstruction quality with more quantization levels, though with diminishing impact. We acknowledge that beyond a certain point, additional quantization layers may introduce quite minor information. The number of quantization layers should be determined cautiously. Variable codebook sizes for different quantization levels are worth future exploring.

Thresholding minimum confidence. Thresholding can be an alternative approach; however, it heavily relies on the choice of the threshold. It would be interesting to explore dynamic threshold in the future.

Limitations. We acknowledge certain limitations of MoMask. Firstly, while MoMask excels in fidelity and faithfulness for text-to-motion synthesis, its diversity is relatively limited. We plan to delve into the underlying causes of this limitation in future work. Secondly, MoMask requires the target length as input for motion generation. This could be properly addressed by applying the text2length sampling [15] beforehand. Thirdly, akin to most VQ-based methods, MoMask may face challenges when generating motions with fast-changing root motions, such as *spinning*. Exemplar cases are presented in the supplementary videos.