

Rethinking Diffusion for Text-Driven Human Motion Generation: Redundant Representations, Evaluation, and Masked Autoregression

Supplementary Material

We further discuss our proposed approach with the following supplementary materials:

- Appendix A: Detailed Deduction
- Appendix B: Detailed Related Works
- Appendix C: Implementation Details
- Appendix D: Additional Quantitative Results
- Appendix E: Temporal Editing
- Appendix F: Additional Qualitative Results
- Appendix G: Limitations

A. Detailed Deduction

A.1. Detailed Deduction for Eq. (4)

In paper, we define $\delta_{\mathbf{x}_0}$ and δ_ϵ to be:

$$\delta_\epsilon = \|\epsilon_\theta(\mathbf{x}_t, t) - \epsilon\|_2^2 \quad (20)$$

and

$$\delta_{\mathbf{x}_0} = \|\mathbf{x}'_0 - \mathbf{x}_0\|_2^2 \quad (21)$$

Since in diffusion-based methods, in each step, diffusion-based methods reconstruct the original motion by:

$$\mathbf{x}'_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t)) \quad (22)$$

where $\epsilon_\theta(\mathbf{x}_t, t)$ is the model's prediction of the noise ϵ . Then we have:

$$\delta_{\mathbf{x}_0} = \left\| \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t)) - \mathbf{x}_0 \right\|_2^2$$

If we substitute \mathbf{x}_0 from Eq. (1):

$$\begin{aligned} \delta_{\mathbf{x}_0} &= \left\| \frac{1}{\sqrt{\bar{\alpha}_t}}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon) - \frac{1}{\sqrt{\bar{\alpha}_t}}(\sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(\mathbf{x}_t, t)) - \mathbf{x}_0 \right\|_2^2 \\ &= \left\| \mathbf{x}_0 + \frac{\sqrt{1 - \bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}(\epsilon - \epsilon_\theta(\mathbf{x}_t, t)) - \mathbf{x}_0 \right\|_2^2 \\ &= \left\| \frac{\sqrt{1 - \bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}(\epsilon - \epsilon_\theta(\mathbf{x}_t, t)) \right\|_2^2 \\ &= \left\| \frac{\sqrt{1 - \bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} \right\|_2^2 \delta_\epsilon \end{aligned} \quad (23)$$

an standard error relation $\delta_\epsilon \rightarrow \delta_{\mathbf{x}_0}$ if \mathbf{x}_0 is processed correctly which should only responds to time coefficient $\bar{\alpha}$.

B. Detailed Related Works

Human Motion Generation. Early text-to-motion approaches [2, 24, 69, 70, 87, 106] attempt to align the latent spaces of text and motion. However, this strategy encounters significant challenges in generating high-fidelity motions due to the inherent difficulty of seamlessly aligning these fundamentally distinct latent spaces. Consequently, recent advancements in human motion generation have shifted focus toward diffusion-based and VQ-based methods, as discussed below.

Diffusion-Based Human Motion Generation. Inspired by the success of denoising diffusion models in the image generation domain [32, 86], several pioneering works [44, 88, 113] have adapted denoising diffusion processes to human motion generation. Building on these works, MLD [12] further optimized the denoising process in latent space to improve training and sampling efficiency. PhysDiff [110] added the physical constraints in the motion generation. And a lot of following works [4, 5, 18, 33, 40, 49, 54, 60, 73, 95, 112, 121] keep exploring diffusion-based human motion generation from different perspectives. In this paper, we thoroughly investigate the limitations of diffusion-based methods and propose a novel approach to address them.

VQ-Based Human Motion Generation. TM2T [25] first introduces Vector Quantization (VQ) to text-to-human motion generation, enabling discrete motion token modeling. A lot of the following works [8, 26, 53, 75, 76, 109, 111, 119] improved the VQ-based methods. T2M-GPT [111] extended this by leveraging a GPT [6] to motion autoregressive generation. Subsequent methods have sought to integrate a larger model [37, 115] (e.g. large language models), or manipulate attention mechanisms [119]. Most recently, MMM [76] and MoMask [26] revisit generation methodology by employing bidirectional attention-based masked generation techniques inspired by MaskGIT [8]. BAMB [75] introduced a dual-iteration framework that combines unidirectional generation with bidirectional refinement to enhance the coherence of generated motions. The concurrent work ScaMo [61] explored the scaling law in human motion generation by training the model with large-scale data. In this paper, we examine the strengths of these approaches and improve a diffusion model inspired by these insights.

Autoregressive Generation with Continuous Data. In motion synthesis, recent works [11, 22, 85, 89, 116] have started to explore integrating autoregressive structures into diffusion-based frameworks. However, due to the chal-

Table A1. **Reconstruction Results** of latent encoders in our method vs baseline methods on HumanML3D [24] data. The AutoEncoder in our method exhibits better reconstruction results.

Methods	FID ↓	MPJPE ↓	R-Precision ↑		
			Top 1	Top 2	Top 3
VQ-VAE [111]	0.081±.001	72.6±.001	0.483±.003	0.680±.003	0.780±.002
RVQ-VAE [26]	0.029±.001	31.5±.001	0.497±.002	0.693±.003	0.791±.002
VAE [12]	0.023±.001	13.7±.001	0.499±.002	0.695±.003	0.791±.003
AE (Ours)	0.004±.001	1.0±.001	0.502±.003	0.696±.002	0.793±.002

Table A2. Further Ablation Study and Optimization Routine.

Approach	FID↓	R-Precision↑		
		Top-1	Top-2	Top-3
MDM [88]-50Step- ϵ	31.265	0.054	0.103	0.147
+Masked AR	2.196	0.387	0.595	0.703
++Essential Only	0.657	0.475	0.668	0.774
+++AE (Ours)	0.116	0.492	0.690	0.790
++++ X_0 -Pred	0.135	0.485	0.686	0.784
++++Velocity (Ours)	0.114	0.500	0.695	0.795

lenges of performing direct causal next motion prediction with MSE loss (as done in discrete token settings), these methods typically only use previously generated motion as a prefix condition, rather than modeling the next step motion directly using previous motion as input. In contrast, recent image generation methods have explored tighter coupling between autoregression and diffusion. GIVT introduced the idea of giving previous generation as input, using outputs from an autoregressive model as parameters for a Gaussian Mixture Model to enable probabilistic chaining of autoregressive generation. MAR further refined this by feeding logits from a masked autoregressive model into a small diffusion branch, producing more fine-grained generation. Inspired by these approaches, we propose to integrate diffusion-based motion generation with masked autoregression, enabling a more direct autoregressive technique beyond simple prefix conditioning to achieve improved generative performance.

Human Motion Generation and Beyond. Recent methods have diversified their focus, exploring retrieval-augmentation [114], controllable generation [15, 41, 43, 74, 79, 94, 101], human-scene/object interactions [10, 14, 17, 20, 34, 38, 46, 48, 50, 59, 63, 67, 72, 96, 97, 100, 104, 105, 108, 118], human-human interaction [7, 19, 36, 56, 99, 103], stylized human motion generation [27, 52, 120], more datasets [58, 102], long-motion generation [71, 122], voice-conditioned motion generation [9], unified motion generation and understanding [47], shape-aware motion generation [90], fine-grained text controlled generation [35, 39, 84, 107, 123], fine-tuning pretrained motion generation model as priors [42, 83], as well as investigating advanced architectures [98, 117] such as Mamba [21].

Table A3. Training Baseline Methods with Reformed Motion Data Representation and Distribution, Linear schedule and ϵ -prediction

Approach	FID↓	R-Precision↑		
		Top-1	Top-2	Top-3
MDM [88]	1.574	0.279	0.336	0.415
MDM [88]-Essential	0.753	0.436	0.627	0.732
MotionDiffuse [113]	0.778	0.450	0.641	0.753
MotionDiffuse [113]-Essential	0.533	0.459	0.650	0.757
MDM [88]-Latent	0.327	0.475	0.663	0.768

Table A4. Original Evaluator Results on HumanML3D.

Approach	FID↓	R-Precision↑		
		Top-1	Top-2	Top-3
GT	0.002	0.511	0.703	0.797
GT→Joints→HumanML3D	0.015	0.503	0.697	0.789
MDM [88]-50Step	0.489	0.455	0.645	0.749
MDM [88]-50Step-Reproduce	0.481	0.459	0.651	0.753
T2M-GPT [111]	0.141	0.492	0.679	0.775
T2M-GPT [111]-Reproduce	0.115	0.497	0.685	0.779
MMM [76]	0.089	0.515	0.708	0.804
MMM [76]-Reproduce	0.071	0.517	0.711	0.805
MoMask [26]	0.045	0.521	0.713	0.807
MoMask [26]-Reproduce	0.093	0.508	0.701	0.796
Ours	0.061	0.523	0.715	0.810

Table A5. **Model Scaling** results of our model. Increasing model size results in better overall performance on HumanML3D.

Size	Transformer	MLP	FID ↓	R-Precision ↑		
				Top 1	Top 2	Top 3
S	6 head 384 dim	3 layers 1024 dim	0.278	0.481	0.676	0.779
	12 head 768 dim	8 layers 1280 dim	0.189	0.479	0.676	0.779
M	6 head 384 dim	3 layers 1024 dim	0.278	0.481	0.676	0.779
	12 head 768 dim	8 layers 1280 dim	0.173	0.477	0.679	0.780
L	6 head 384 dim	12 layers 1536 dim	0.137	0.485	0.683	0.785
	12 head 768 dim	12 layers 1536 dim	0.125	0.487	0.685	0.785
XL	16 head 1024 dim	16 layers 1792 dim	0.116	0.492	0.690	0.790

C. Implementation Details

For our method, the AutoEncoder is a 3-layer ResNet-based encoder-decoder with a hidden dimension of 512 and a total downsampling rate of 4. For the generation branch, we utilize a single-layer AdaLN-Zero transformer encoder with a hidden dimension of 1024 and 16 heads as our masked autoregressive transformer. The diffusion MLPs consist of 16 layers with a hidden dimension of 1792. We also present the model scalability results in Appendix D.5.

During training, we use the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. Following prior works [24, 26, 76, 111], the batch size is set to 256 and 512 for training the AutoEncoder on the HumanML3D and KIT-ML datasets, respectively, with each sample containing 64 frames. For training the generation branch, the batch size is set to 64 for HumanML3D and 16 for KIT-ML, with a maximum sequence length of 196 frames. The learning rate is set at 2×10^{-4} with a linear warmup of 2000 steps. We

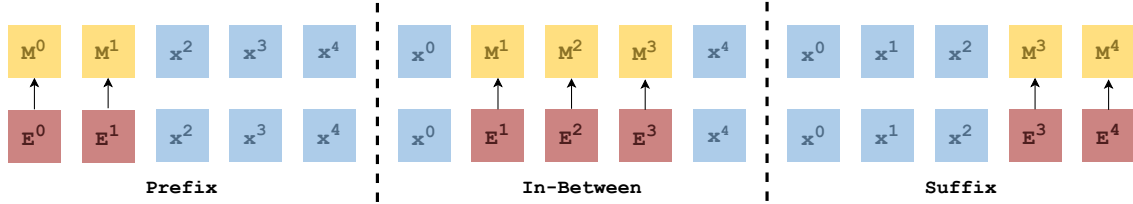


Figure A1. **Our Method’s Temporal Editing** process, including prefix, in-between, and suffix editing. The editing latents (red color) are treated as masked latents (yellow color). The sequence is then input into the generation branch in Fig. 3 to generate edited latents conditioned on the editing textual instruction and non-edit latents (blue color).

Table A6. **Average Inference Time Results Comparison** between our method and baseline methods.

Methods	MDM [88]	MotionDiffuse [113]	T2M-GPT [111]	MLD [12]	MMM [76]	MoMask [26]	Ours
AIT	14.31s	7.35s	0.32s	0.21s	0.06s	0.04s	2.4s

train the AutoEncoder for 50 epochs and modify the learning rate to decay by a factor of 20 or 10 at milestones of 150,000 and 250,000 iterations for HumanML3D and KIT-ML datasets, respectively. For the generation branch, the learning rate decays by a factor of 0.1 at 50,000 iterations for HumanML3D and 20,000 iterations for KIT-ML during a 500-epoch training process. Following image diffusion works [62, 66, 80], we also incorporate exponential moving average (EMA) when updating the model parameters to achieve more stable performance. In the generation process, for HumanML3D, the CFG [31] scale is set to 4.5 and for KIT, the conditioning scale is set to 2.5.

D. Additional Quantitative Results

D.1. AutoEncoder Reconstruction Results

In Tab. A1, we present the reconstruction results of VQ-VAE from T2M-GPT [111], RVQ-VAE from MoMask [26], VAE from MLD [12], and the AutoEncoder (AE) in our method. Our AutoEncoder has much better reconstruction capability than baseline methods, which ultimately benefits both diffusion model training and sampling.

D.2. Baseline Methods Training With Reformed Data Representation and Distribution

In Tab. A3, we demonstrate that training baseline methods using only essential dimensions can already lead to significant improvements, and processing into latent space may further enhance results.

D.3. Original Evaluation Results

The original evaluator is flawed due to the unnecessary focus on redundant motion representations and the new evaluators are proposed to deal with this issue. Therefore, we strongly discourage utilizing the original evaluation method to access all methods. Also, using the original evaluator requires additional processing to convert our outputs to joints and back to the redundant representations. This inevitably introduces errors, and loses one motion frame

(from joints to HumanML3D representations), and thus unfairly penalizes our method. Nevertheless, we include results in Tab. A4, where our method can still achieve superior performance on R-precision metrics. Notably, our reproduced MoMask exhibits worse results, similar to issues reported in their GitHub (Issues 27, 43, 99), and even ground truth motions were penalized due to the additional operations.

D.4. Further Ablation Study and Optimize Routine

In Tab. A2, we provide a further ablation study and an optimization routine starting from an MDM-based cosine schedule, ϵ prediction approach to our approach. The results demonstrate the advantage of masked regression over original diffusion and the importance of our further optimization (motion representation reformation) over pure adoption of image MAR.

D.5. Model Scalability

We train six versions of our proposed model (DDPM approach), varying three transformer sizes and four diffusion MLP sizes (S, M, L, XL). These models range in size from around 30M, 100M, 180M, to 290M parameters. The performance results are summarized in Tab. A5. We observe that increasing the model size, particularly the diffusion MLPs size, improves overall generation performance, especially in terms of FID.

E. Temporal Editing

Our method is capable of performing temporal editing in a zero-shot manner (*i.e.* utilizing the model trained for text-to-motion generation without any editing-specific fine-tuning). In our method, temporal motion editing is easily achieved by treating the latents that need to be edited as masked latents and then generating motions following our standard generation procedure in Sec. 3.2 which is conditions on the unmasked tokens (*i.e.* non-edit latents) and the editing textual instructions. We visually illustrate this process in Fig. A1 and we also include temporal editing results

in the locally-run, anonymous HTML file referenced in Appendix F.

F. Additional Qualitative Results

Beyond the qualitative results presented in the main paper, we also provide comprehensive video visualizations hosted on a locally-run, anonymous HTML webpage to further demonstrate the effectiveness of our approach. These visualizations include additional comparisons with state-of-the-art baseline methods, showcasing that our method generates more realistic motions and adheres more closely to textual instructions. We also present motion videos from our ablation studies to highlight the significance of each component. For example, omitting motion representation reformulation results in noticeable shaking and poses inaccuracies, while excluding the autoregressive modeling approach leads to worse textual instructions following. Furthermore, we also demonstrate our method’s capability for temporal editing with prefix, in-between, and suffix editing results. Finally, we provide additional visualizations to illustrate that our method can generate a wide range of diverse and contextually appropriate motions.

G. Limitations

Since our method incorporates both standard reverse-diffusion processes (over T time steps) and autoregressive generation within each step to produce high-quality and diverse motion, it inherently requires more time for motion generation compared to some baseline methods (*e.g.*, MoMask, MMM). To provide a clear comparison, in Tab. A6, we report the efficiency of motion generation in terms of average inference time (AIT) over 100 samples on a single Nvidia 4090 device. Notably, our method still outperforms several diffusion-based methods, *e.g.* MDM and Motion-Diffuse, in generation speed by a significant margin. For future work, we aim to explore strategies to optimize and accelerate both standard reverse-diffusion and autoregressive generation processes.