Autoregressive Diffusion for Long-Term Controllable Motion Synthesis

William Huang [1], Yifeng Jiang [1], Tom Van Wouwe [1], C. Karen Liu [1]

[1]The Movement Lab, Stanford Computer Science

Background

Long-term generation of realistic human motion is a difficult problem, with applications in human-computer interaction, computer animation, robotics, and more. Most approaches run the following problems (Zhang et al., 2023):

- Unrealistic physical interactions (e.g. foot sliding)
- Repetitive or degenerate outputs (i.e. frozen motion)
- Limited interactive controllability

The goal of this research was to generate *realistic, arbitrary length human motions*, capable of control on trajectories and other aspects of motion.

Diffusion models. Diffusion models have the potential to solve this long-standing problem. They consist of a forward process and a backward process; the forward process adds Gaussian noise from timesteps 0 to T, and the backward process de-noises a sample from timesteps T to 0.

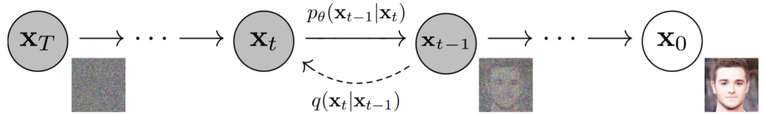


Figure 1. Graphical model of diffusion models, taken from Ho et al., 2020.

Data. We use the Archive of Motion Capture as Surface Shapes (AMASS) dataset, which contains 40+ hours of motion data at 20 FPS. Each frame contains information in the 24-joint SMPL format.

Model Architecture

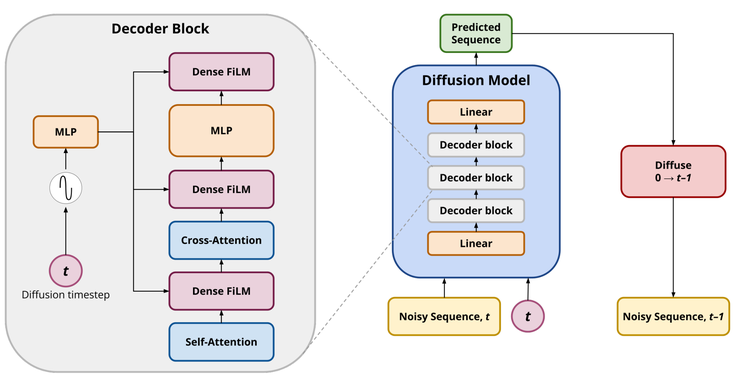
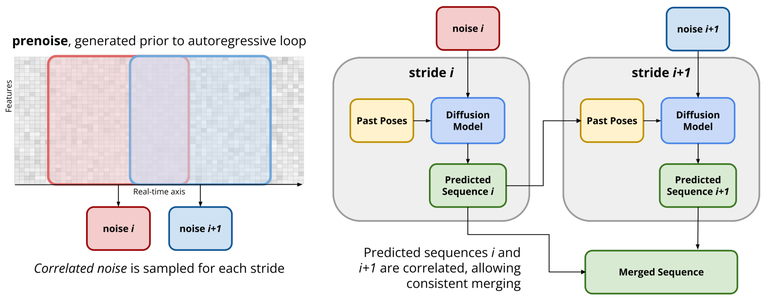


Figure 2. Diffusion model pipeline used for this research. Timestep information is provided via feature-wise linear modulation (FiLM), and sequential nature of motion is processed by attention blocks. Heavily inspired by Tseng et al. 2022.

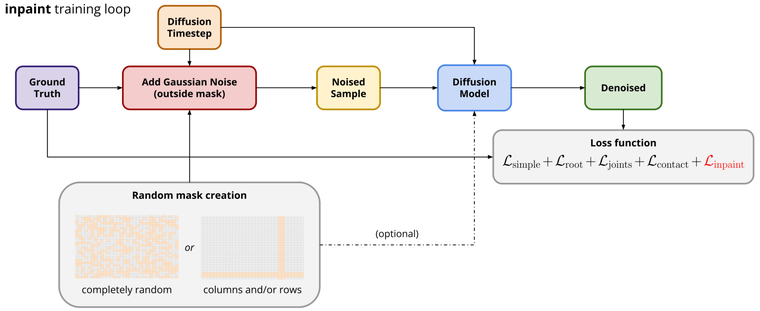
Methods

Three techniques improved long-term human motion generation:

- **Prenoising.** Replacing independent noise with *prenoise* helps create consistent motions across different strides.



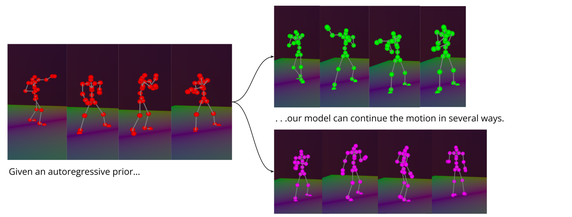
- **Inpaint training.** Using random masks, the diffusion model learns inpainting instead of pure denoising during training.



- **Guided inference.** To better enforce specific constraints on motion, *guided inference* uses the gradient of an analytical loss function to nudge generated samples in the right direction (Janner et al. 2022, Rempe et al. 2023).

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t \in [T, T-1, \ldots, 1]$ **do**
  $\hat{\mathbf{x}}_0 \leftarrow \mathbf{x}_{0,\theta}(\mathbf{x}_t)$ `// (predicted t = 0 product)`
  $\tilde{\mathbf{x}}_0 \leftarrow \hat{\mathbf{x}}_0 - \alpha \Sigma_t \nabla_{\mathbf{x}_t} \mathcal{J}(\hat{\mathbf{x}}_0)$ `// guidance`
  $\boldsymbol{\mu} \leftarrow \boldsymbol{\mu}_\theta(\tilde{\mathbf{x}}_0, \mathbf{x}_t)$
  $x_{t-1} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma_t)$
end
return $\mathbf{x}_0$

Results



Method / Metric	Foot Loss, 3	Foot Loss, 1	APD, 3	APD, 1
Prenoising	0.0186	0.0202	0.130	0.129
Regular	0.0232	0.0390	0.119	0.121

Figure 3. Our autoregressive model is capable of multiple diverse motion continuations of the same prior. Prenoising increases diversity (higher APD) and quality (lower foot loss).

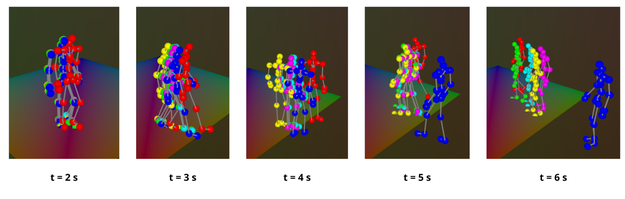


Figure 4. Our model is capable of generating motion that follows certain trajectory constraints. Here, the red body is the trajectory to be followed; the lower body motion is provided, and the upper body is inpainted.

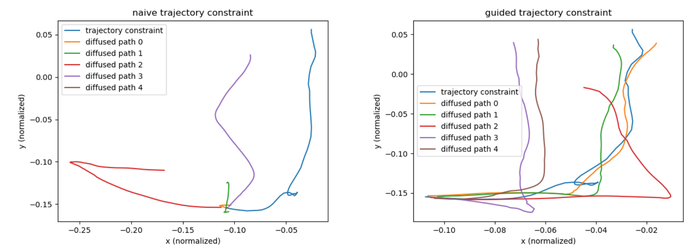


Figure 5. Guided inference helps with trajectory control immensely over existing methods.

Discussion

- The main difficulty in the autoregressive mode is preserving the data's distribution when sampling multiple times across multiple strides.
- Prenoising imposes an extra constraint across strides that helps make more consistent motions across strides.
- Inpaint training and guided inference help make the training and inference tasks more similar, improving performance.
- There remains more to be desired with the strength of trajectory control and real-time diffusion.