Diffusion-Based Human Motion Generation

MVA
Mathis Wauquiez
Félix Fourreau

January 2025

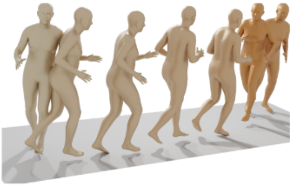
Diffusion-Based Human Motion Generation

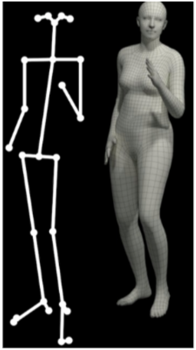
MVA
Mathis Wauquiez
Félix Fourreau

Introduction

Useful for:

- Animation and Film Production
- Video Games
- Virtual Reality or Augmented Reality





$$x \in \mathbb{R}^{T \times J \times D}$$

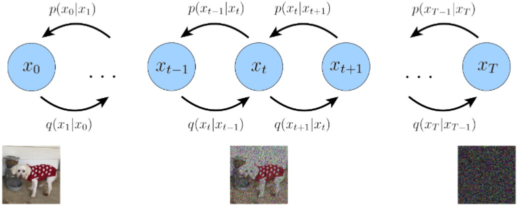
How to generate natural human motion ?

Predictive approaches : Predict $x = F_\theta(c)$, with c a textual embedding.
Solution : $\mathbb{E}[x \mid c]$. Doomed to generate an unnatural motion.

Generative approaches :

- VAEs -> TEMOS, ACTOR
- GANs (too difficult to train)
- Normalizing flows
- Diffusion models (most popular) $\rightarrow$ MDM, MotionDiffuse, FLAME, Motion Latent-based Diffusion model (MLD)

Diffusion Models



Key idea :

- **Forward process :** Gradually noise the sample :
  $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, $\bar{\alpha}_t = \prod_{s=0}^{T} \alpha_s$, $\bar{\alpha}_T \approx 0$
- **Reverse process :** Optimize the network to eliminate the noise : $\min_\theta \mathbb{E}_{t\sim\mathcal{U}([0,T]),(x_0,c)\sim p_{\text{data}}}[\|\mathcal{G}(x_t,t,c)-x_0\|^2]$
- **Inference process :** Sample $x_T \sim \mathcal{N}(0,1)$, then $\forall t \in [1, T-1]$, predict $\widehat{x}_0 = \mathcal{G}(x_t, t, c)$, and add noise to obtain x_{t-1}. Repeat until x_0 is reached.

Key idea : Diffusion Model with Transformer-encoder-based Denoiser

Enables different reconstruction losses :

- $\mathcal{L}_{\text{simple}} = E_{x_0 \sim q(x_0|c), t \sim [1,T]} \left[\| x_0 - G(x_t, t, c) \|_2^2 \right]$. Variant of the variational bound, proposed by Denoising Diffusion Probabilistic Models.
- $\mathcal{L}_{\text{pos}} = \frac{1}{N} \sum_{i=1}^{N} \| FK(x_0^i) - FK(\hat{x}_0^i) \|_2^2$
- $\mathcal{L}_{\text{foot}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left\| \left(FK\left(\hat{x}_0^{i+1}\right) - FK\left(\hat{x}_0^i\right) \right) \cdot f_i \right\|_2^2$
- $\mathcal{L}_{\text{vel}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \left\| \left(x_0^{i+1} - x_0^i \right) - \left(\hat{x}_0^{i+1} - \hat{x}_0^i \right) \right\|_2^2$

$FK(.)$: function that converts joint rotations to joint positions
$f_i \in \{0, 1\}^J$: binary foot contact mask for each frame i

MDM : Motion Diffusion Model

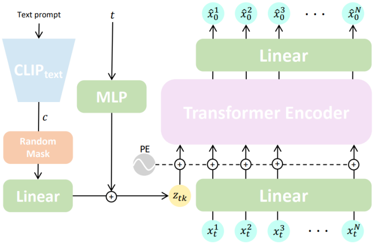


Figure – Denoiser architecture - Used with a cosine scheduler

Diversity-Fidelity trade-off : by interpolating conditional and unconditional generation :

$$G_s(x_t, t, c) = G(x_t, t, \emptyset) + s \cdot (G(x_t, t, c) - G(x_t, t, \emptyset))$$

Editing : Motion in-betweening and body-parts re-synthesizing by overwriting $\widehat{x}_0$ at each timestep, as in diffusion image inpainting.

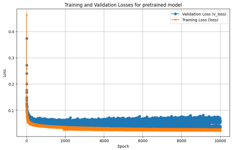
Pretrained MDM-SMPL - Evaluation

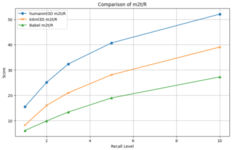
- Trained on Amass dataset (Archive of Motion Capture As Surface Shapes)

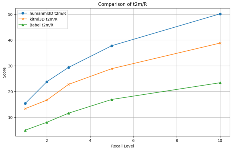
-Associated labels :

1. kitml : mostly locomotive motions, starts by 'A person is...
2. humanml3d : more verbose, covers more motions
3. babel : single word description

- Pretrained on humanml3d with 10000 epochs







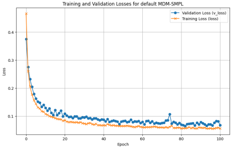
MATHÉMATIQUES VISION

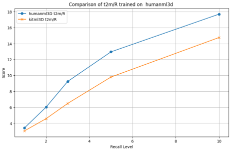
Improvement with text augmentation - Baseline

Diffusion-Based Human Motion Generation

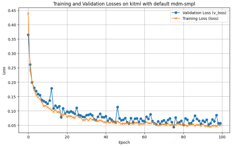
MVA
Mathis Wauquiez
Félix Fourreau

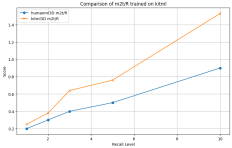
- Train on humanml3d - 100 epochs

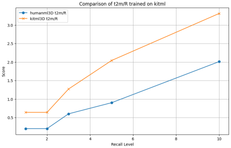




-Train on kitml - 100 epochs





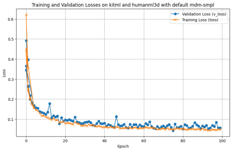


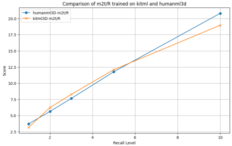
Improvement with text augmentation

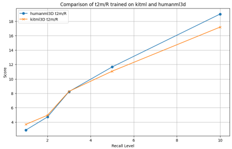
Diffusion-Based Human Motion Generation

MVA
Mathis Wauquiez
Félix Fourreau

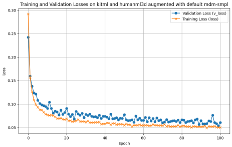
-Train on kitml + humanml3d : 100 epochs

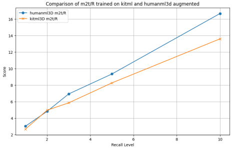


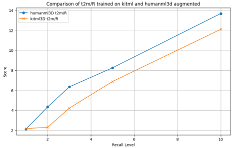




-Train on kitml + humanml3d data augmentation : 100 epochs

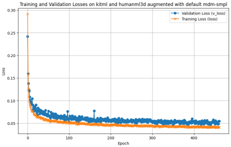


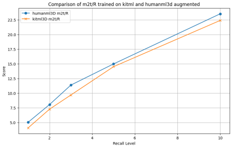


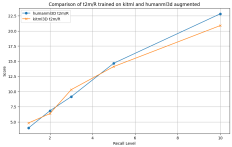


Improvement with text augmentation

-Train on kitml + humanml3d data augmentation : 500 epochs







Possible improvements

Use domain losses : MDM-SMPL is not trained using MDM losses.
Improve the denoiser : StableMoFusion claims significant improvements by changing the denoiser to a modified Conv1D UNet (Cond1DUnet). **Boost could not be replicated**.
Change sampler : Use DPM-Solver++ instead of DDPM during inference.

Dataset	Network	FID ↓	R Precision (top3) ↑
HumanML3D	Conv1D UNet basline	0.245	0.780
	+ cross-attention	0.074	0.821
	+ GroupNorm Tweak	0.089	0.840
	DiT baseline	0.884	0.711
	+ cross-attention	0.113	0.787
	RetNet baseline	1.673	0.740
	+ cross-attention	0.147	0.853
KIT-ML	Conv1D UNet+ cross-attention	0.658	0.756
	+ GroupNorm Tweak	0.237	0.780

Cond1DUnet

Diffusion-Based Human Motion Generation

MVA
Mathis Wauquiez
Félix Fourreau

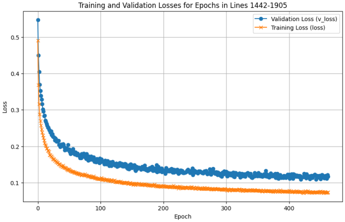
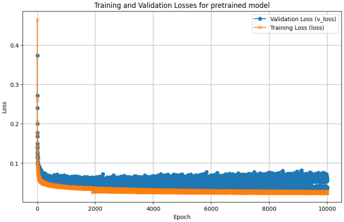


Figure – ConvUNet1D with attention



s

Figure – Baseline

Figure – Comparison between the ConvUNet1D and the baseline. The baseline is a better denoiser.

Conclusions

Diffusion-Based Human Motion Generation

MVA
Mathis Wauquiez
Félix Fourreau

- Struggle with original prompt

Original input - dab movement :

```
- bending the left arm across the face ## 0.0 ## 5.0 ## left arm
- raising the right arm straight diagonally upward ## 0.0 ## 5.0 ##
right arm
- tilting the head toward the left elbow ## 0.0 ## 5.0 ## head
- standing with a slight bend in the knees ## 0.0 ## 5.0 ## legs
```

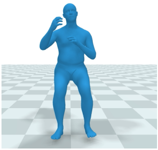


Figure – Visualization of the Dab Movement