Self-Intersection-Aware 3D Human Motion Generation Using an Efficient Human Sphere Proxy

Pascal Herrmann[1], Maarten Bieshaar[1], Dennis Mack[1], Robert Herzog[1], Juergen Gall[2,3]

[1]Bosch Research [2]University of Bonn [3]Lamarr Institue for Machine Learning and Artificial Intelligence

Motivation

Text-conditioned 3D human motion generation aims to generate realistic 3D motions of humans, which adhere to a given textual description. This field has made tremendous progress in recent years with state-of-the-art approaches surpassing ground truth data in leading evaluation benchmarks. However, visual inspection of generated motions reveals that current approaches do not consider physical plausibility. Generated motions frequently contain self-intersections, which are strong artifacts, severely limiting the perceived motion quality. Additionally, current evaluation metrics fail to capture these self-intersections, hence, insufficiently measure the quality of generated motions.

Contributions

We introduce a fast and memory-efficient self-intersection loss based on a sphere approximation of human geometries. This generic loss can be used with existing 3D human motion generation methods, reducing generated self-intersections by up to 49% even for the most recent state-of-the-art. Additionally we propose a novel voxel-based metric, called Self-Intersect (SI), which approximates the self-intersection volume, making generated self-intersections measurable.

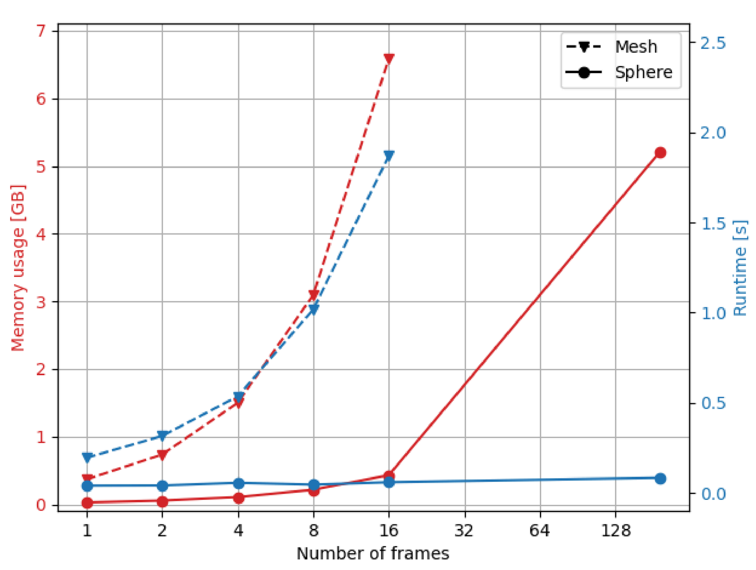


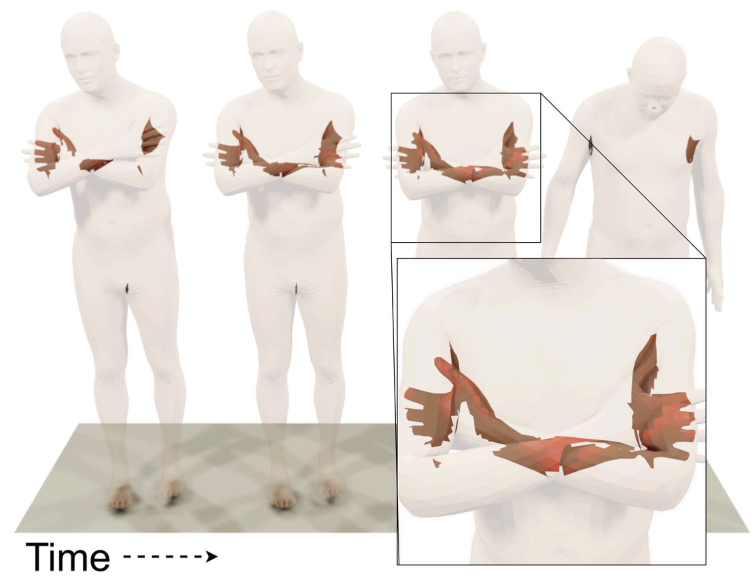
Fig. 1: Comparison of computational efficiency of the self-intersection loss between meshes and the sphere proxy.

Results

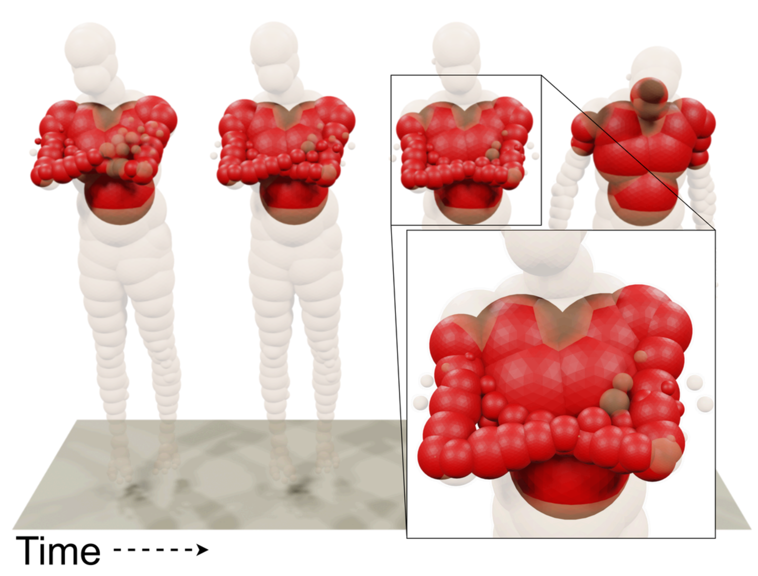
We evaluate our approach on the 3D human motion dataset HumanML3D [4] and show the results in Tab. 1. Incorporating our novel self-intersection loss into the training of MDM [2] and MoMask [1] significantly reduces the amount of generated self-intersections. Additionally, most other metrics improve, highlighting the benefit of focusing on the physical plausibility of generated motions. Fig 2. compares motion generated by MDM [3] and our approach, which further confirms the improved physical plausibility of generated motions. Fig. 1 shows the computational superiority of our novel loss compared to mesh-based approaches, making a self-intersection loss applicable to 3D human motion generation in the first place.

Method

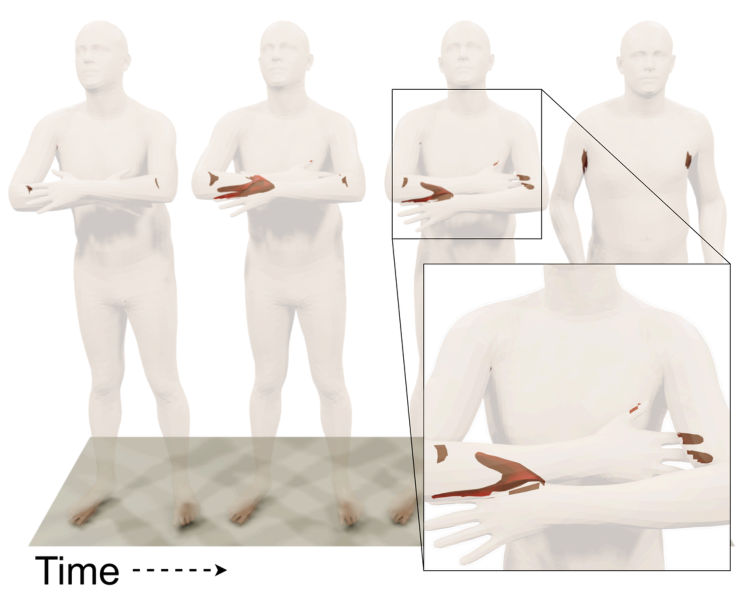
Our goal is to generate physically plausible 3D human motions by explicitly penalizing self-intersections during the model training. Existing approaches [3] calculate a self-intersection loss based on triangular meshes for individual poses. However, these approaches do not scale to 3D human motion generation as shown in Fig. 1. Instead, we approximate the human geometry using a set of spheres. Using spheres reduces the memory requirements to save the human representation. Additionally, computing the intersection of two spheres is faster than computing the intersection of two triangles. This allows an efficient calculation of our loss, making it applicable to all 3D human motion generation methods with limited computational overhead.



a) Motion generated by MDM [2]



b) Our novel sphere proxy



c) Motion generated by SIA-MDM (Ours)

Fig. 2: Qualitative comparison between MDM [2] and SIA-MDM (Ours). Red indicates self-intersections. Text prompt for generation: "Tightly cross arms and open them again".

Method	R Precision top 3↑	FID↓	Multimodal Dist↓	Diversity→	MultiModality↑	SI↓
Real	$0.797^{\pm 0.002}$	$0.002^{\pm 0.000}$	$2.974^{\pm 0.008}$	$9.503^{\pm 0.065}$	-	$447^{\pm 00}$
MDM [2]	$0.707^{\pm 0.004}$	$0.489^{\pm 0.025}$	$3.631^{\pm 0.023}$	$\mathbf{9.449^{\pm 0.066}}$	$\mathbf{2.973^{\pm 0.111}}$	$619^{\pm 11}$
MoMask [1]	$\underline{0.807^{\pm 0.002}}$	$\mathbf{0.045^{\pm 0.002}}$	$\underline{2.958^{\pm 0.008}}$	$9.624^{\pm 0.080}$	$1.241^{\pm 0.040}$	$\underline{316^{\pm 02}}$
SIA-MDM (Ours)	$0.731^{\pm 0.006}$	$0.265^{\pm 0.024}$	$3.462^{\pm 0.026}$	$\underline{9.568^{\pm 0.086}}$	$\underline{2.893^{\pm 0.075}}$	$382^{\pm 09}$
SIA-MoMask (Ours)	$\mathbf{0.813^{\pm 0.002}}$	$\underline{0.068^{\pm 0.002}}$	$\mathbf{2.933^{\pm 0.006}}$	$9.691^{\pm 0.092}$	$1.198^{\pm 0.041}$	$\mathbf{290^{\pm 02}}$

Tab. 1: Qualitative results on the HumanML3D [4] dataset.

References & Code

[1] Guo et al., 2024, Momask: Generative masked ...
[2] Tevet et al., 2023, Human motion diffusion model
[3] Pavlakos et al., 2019, Expressive body capture: ...
[4] Guo et al., 2022, Generating diverse and natural ...

github.com/boschresearch/humanspherepoxy