BoDiffusion: Diffusing Sparse Observations for Full-Body Human Motion Synthesis

Angela Castillo[1], María Escobar[1], Guillaume Jeanneret[2], Albert Pumarola[3], Pablo Arbeláez[1], Ali Thabet[3], Artsiom Sanakoyeu[3]

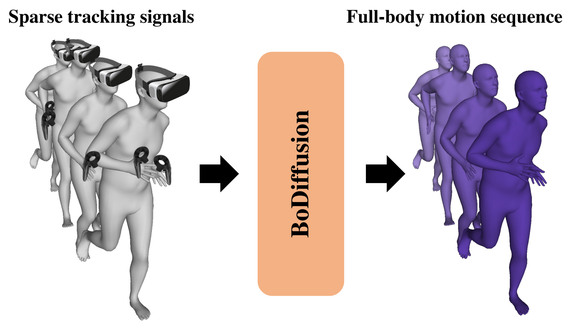
[1]CINFONIA, Universidad de Los Andes, Colombia
[2]Normandy University, ENSICAEN, UNICAEN, CNRS, GREYC, France
[3]Meta AI

cv4metaverse
ICCV23 PARIS

1. INTRODUCTION

Full-body motion capture enables natural interactions between real and virtual worlds for immersive mixed-reality experiences.

However, typical head-mounted devices can only track head and hand movements, leading to a limited reconstruction of full-body motion.



Our goal is to predict full-body pose estimation based on sparse tracking signals.

2. METHOD

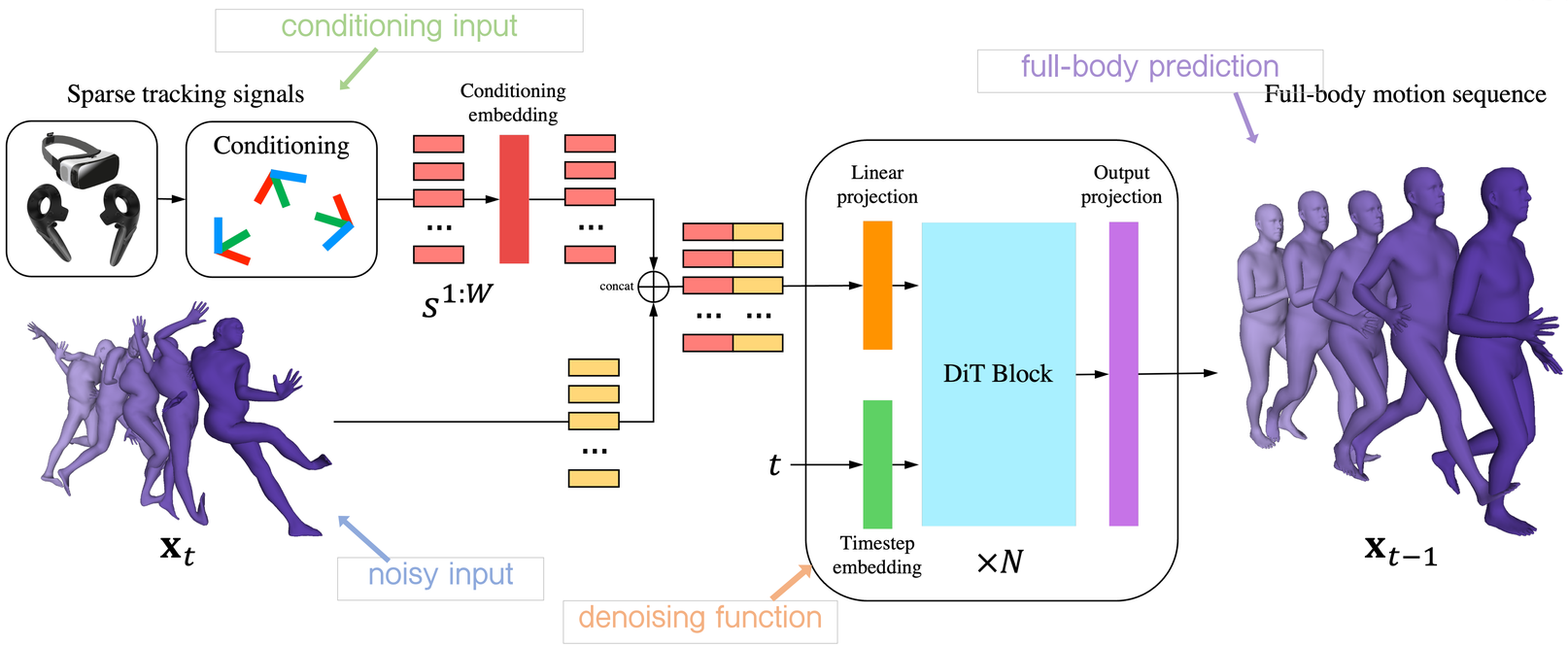
We leverage upon Diffusion Models to reformulate the input conditioning to produce full-body poses.

Let $\mathbf{x}_t = x_t^{1:W}$ be the sequence of size W.
We learn a conditional distribution of the full-body sequences $\mathbf{x_0}$ which is defined as:

$$p_\theta(\mathbf{x}_{0:T}|s^{1:W}) = p(\underbrace{\mathbf{x}_T}_{\text{noisy input}}) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \underbrace{s^{1:W}}_{\text{conditioning input}})$$

noisy input: Gaussian noise
conditioning input: sparse tracking signals

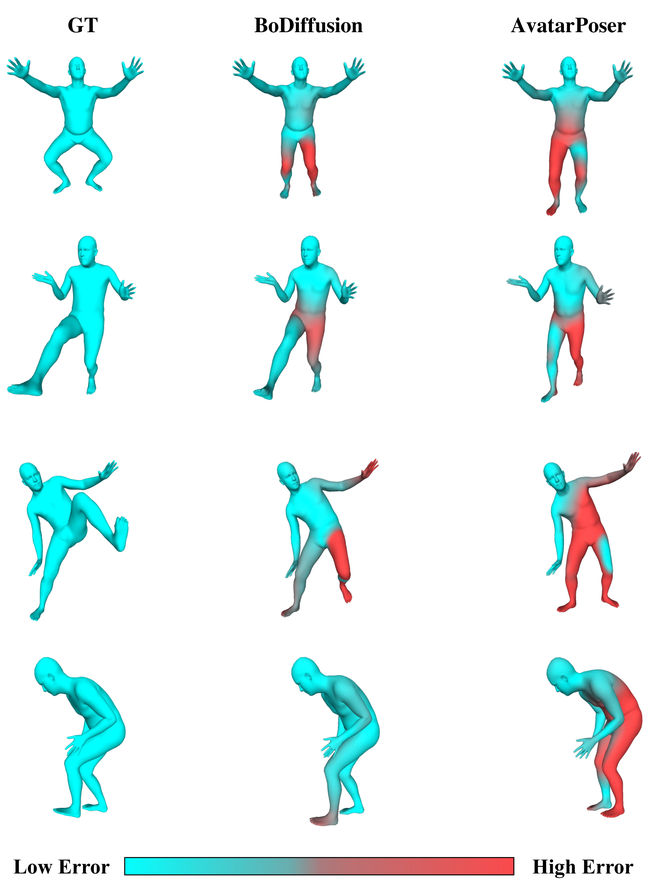
We treat each pose as an individual token and combine the feature and joint dimensions. Thus, we take advantage of the temporal information and efficiently process the motion sequence.

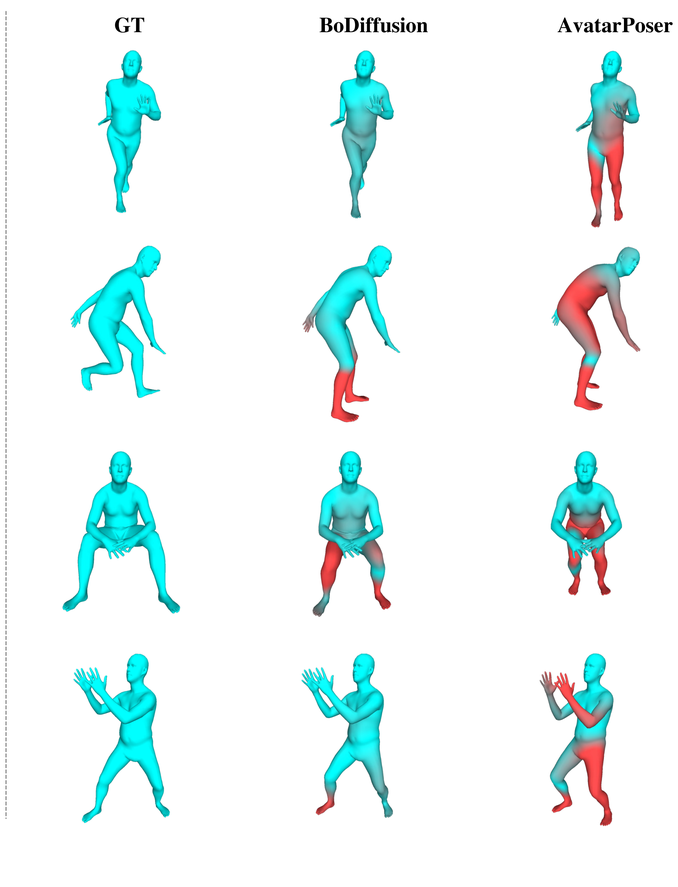


We use the novel Transformer backbone DiT to build our BoDiffusion model because

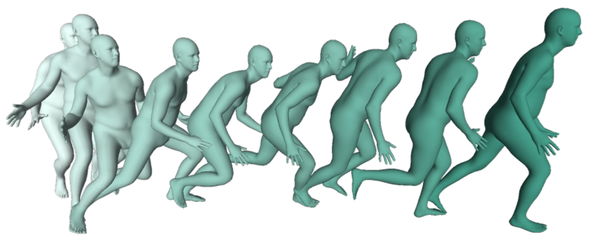
a) it was shown to be superior for image synthesis task compared to the frequently used UNet backbone.
b) it is more naturally suited for modeling heterogeneous motion data.

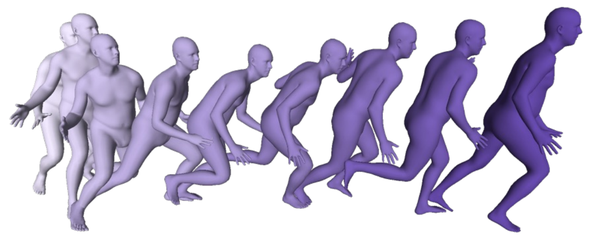
3. RESULTS





- BoDiffusion produces plausible poses for the upper and lower bodies.
- We enforce the temporal consistency by leveraging the novel conditioning scheme and learning to generate sequences of poses instead of individual poses.





Our results support the effectiveness of our conditioning scheme for guiding the generation towards realistic movements that lead to smoother and more accurate motions.

Method	Jitter	MPJVE	MPJPE	Hand PE	Upper PE	Lower PE	MPJRE	FCAcc ↑
Final IK*	-	59.24	18.09	-	-	-	16.77	-
LoBSTr*	-	44.97	9.02	-	-	-	10.69	-
VAE-HMD*	-	37.99	6.83	-	-	-	4.11	-
AvatarPoser [15]	1.53	28.23	4.20	2.34	1.88	8.06	3.08	79.60
AvatarPoser-Large [15]	1.17	23.98	3.71	2.20	1.68	7.09	**2.70**	82.30
BoDiffusion (Ours)	**0.49**	**14.39**	**3.63**	**1.32**	**1.53**	**7.07**	**2.70**	**87.28**

4. CONCLUSION

We propose a state-of-the-art method based on sparse tracking signals of the upper body.

BoDiffusion uses a novel spatio-temporal conditioning scheme and enables motion synthesis with significantly reduced jittering artifacts, especially on lower bodies.

REFERENCES

1. Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In Proceedings of European Conference on Computer Vision. Springer, 2022

2. William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In Proceedings of International Conference on Computer Vision, 2023.

Full project