

Contributions: -





MDMP: Multi-modal Diffusion for supervised Motion Predictions with uncertainty

Leo Bringer, Joey Wilson, Kira Barton, Maani Ghaffari University of Michigan





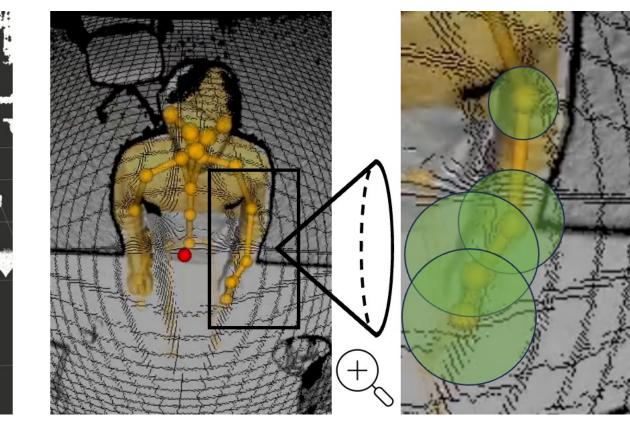
[3] Chuan Guo, et al. Momask: Generative masked modeling of 3d human motions. CVPR, 2024

Motivation:

 Diffusion model guided by text & skeletal data for accurate long-term motion prediction.

 Associated spatial uncertainty estimation per joint and timestep.

 Graph-based transformer capturing spatial-temporal motion dynamics efficiently.



In many human-robot collaborative settings such as industrial robotics or self-driving cars estimating uncertainty could significantly improve spatial awareness and safety by incorporating zones of presence with varying confidence levels. **Quantitative Results:** MotionGPT [2] → MoMask [3] 300 5.0 Time (s)

Method: Architecture Overview: Summary **Denoised sequence + Variance** Our method integrates **both skeletal motion & text** to generate long-term motion predictions with uncertainty: The whole Diffusion process: X_0, V_0 MDMP Input Text A person picked up a tool and is mounting an object Transformer **Encoder** $X_{\mathrm{T}} \sim \mathcal{N}(0, I)$ **Input Motion** Generated Motion **Noising Process** "a person is walking and starts running" **Noised sequence Text Prompt**

Some visual results: "A man postures his arms like "A person puts their hands up while holding a dance partner and walking and then puts them down." dances the waltz." "He grabs something with his right hand and then walks to his left and grabs "The person was pushed while walking something with his left hand." but did not fall." Comparative trajectory study: "a person walks in a small circle to the right."

MDM [1]

MDMP (ours)

