

Peer Review

Review of: "HoloGest: Decoupled Diffusion and Motion Priors for Generating Holisticly Expressive Co-speech Gestures"

Anish Monsley Kirupakaran¹

1. Indian Institute of Technology, Madras, India

This article introduces a novel system to synthesize realistic, physically natural gestures in virtual agents purely from speech audio. This research addresses shortcomings in existing gesture generation models by introducing motion priors and decoupling human motion into subcomponents. Some observations and a few suggestions:

- The system relies on multiple large, high-quality datasets (e.g., AMASS, BEATX, SignAvatars). Would the authors consider releasing a reduced version of the pipeline for broader adoption?
- Can the authors clarify how the system may handle prosodic or emotional cues not captured in the transcription? Is there potential to incorporate raw speech features or emotion embeddings to improve alignment?
- Were there any challenges in training the hybrid GAN-diffusion model? What techniques were used to ensure convergence and avoid mode collapse or instability?
- To what extent does the quality of gesture generation depend on the quality and coverage of these priors? Can these priors be fine-tuned jointly with the rest of the pipeline in a low-resource setting?
- What are the practical considerations for deploying HoloGest in interactive applications (e.g., VR, virtual avatars)? It may be helpful to evaluate or discuss the system's feasibility on resource-limited platforms like mobile or embedded devices.
- Typo - The authors have misspelled HoloGest as HoleGest in the Abstract.

Declarations

Potential competing interests: No potential competing interests to declare.