

Peer Review

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

Gaurav Kumar Yadav¹

1. Universitat Rovira i Virgili, Spain

The paper proposes the integration of decoupled diffusion models and motion priors to address significant challenges such as unnatural jittering, floating, and sliding, which is a well-motivated approach to improving co-speech gesture generation. However, there are some issues that need to be addressed:

1. The proposed Semantic Alignment (SA) metric (Eq. 8) uses cosine similarity between gesture and text embeddings. How was this metric validated to ensure it correlates with *human-perceived* alignment?
2. What specific aspects of DSG and EMAGE should be highlighted to justify their use as baseline methods?
3. In complex speech-gesture mappings, how does the current architecture ensure local semantic coherence without explicit cross-model interaction?
4. Experimental results are shown only for one dataset, BEATX. Can you add the experimental results for other data as well?
5. The key contributions should be listed pointwise in the introduction.
6. Equation 1 in the method section is not discussed in the text.
7. Limitations of this work are overlooked in the current manuscript; key limitations should be explicitly stated.

Declarations

Potential competing interests: No potential competing interests to declare.