Qeios

Peer Review

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

Elakkiya Rajasekar[1]

1. Birla Institute of Technology and Science Pilani, Pilāni, India

1. The paper introduces HoloGest, which decouples motion generation into limb, global, and finger streams using diffusion priors and GAN-based acceleration. The decoupling is a known concept, and the contribution is in the system integration rather than in any new theory or mechanism.
2. The semi-implicit GAN-based denoising strategy is adapted from existing frameworks (e.g., SiDDMs), and the authors do not justify why it outperforms standard latent diffusion in gesture generation. No comparative study with other acceleration strategies such as DDIM or Score Distillation Sampling is provided.
3. The claim that “this is the first audio-whole body gesture generation model with motion priors” is overstated. While it may be one of the first to combine this configuration, motion priors for trajectory stabilization and finger modeling have appeared earlier, though in different architectures.
4. The system uses large-scale datasets (AMASS, BEATX, SignAvatars), but no ablation is performed to check dependence on or contribution from individual datasets. It's unclear how much BEATX alone suffices in fine-tuning.
5. Semantic alignment via JEPA and contrastive loss is not analyzed qualitatively. No retrieval-based evaluation, latent visualization, or linguistic error analysis is provided. This makes it hard to judge the benefit of semantic grounding.
6. Evaluation metrics (FGD, BA, DIV, Skate, Float, SA) are appropriate, but results like those in Table 1 show marginal improvements in SA compared to baselines (e.g., 0.66 vs 0.60 or 0.22). Gains in FGD (5.34 vs 5.51) are not groundbreaking considering the architectural overhead.

doi.org/10.32388/L4X34R

7. Finger priors are learned from sign language data, but no comparison is made to recent sign language generation models, even though the authors use SignAvatars. There is no baseline for finger expressiveness beyond visual observation.
8. The system speeds up inference (50-step diffusion), but this is only compared against EMAGE (VAE-based) and a 1000-step DDPM. No latency comparison with DDIM, RePaint, or distillation-based models is shown.
9. GAN-based denoising is known to introduce instability or mode collapse; the paper doesn't show training stability plots, gradient norm tracking, or discriminator performance to validate this component.
10. No comparative results are given on generalization across languages or accents, despite claiming real-time application. No analysis of robustness under noisy audio or low-quality speech is provided.
11. The user study reports high average scores (e.g., HL: 4.89), but no statistical significance testing (e.g., p-values or effect size) or inter-annotator agreement is shown. The 95% CI is provided, but the interpretation lacks depth.
12. The code, model, and demo are shared via a GitHub Pages site, which is good for reproducibility. However, the paper does not document training seed settings, computational budget beyond GPU type, or memory usage.

Declarations

Potential competing interests: No potential competing interests to declare.