Qeios

Peer Review

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

Téo Guichoux[1]

1. Pierre and Marie Curie University - Paris 6, Paris, France

Global assessment

HoloGest is an ambitious diffusion-based system for audio-driven co-speech gesture generation. The authors combine a divide-and-conquer diffusion pipeline with motion-prior networks and a semi-implicit adversarial loss. Although the architecture is interesting and the quantitative results on BEAT-X are promising, several key claims (music-to-dance generalisation, faster inference, finger-motion improvement) are **not demonstrated**, and important baselines and details are missing.

Strengths

Clear problem focus – tackles the weak audio-gesture correlation and the difficulty of modelling expressive fingers.

Divide-and-conquer architecture – four specialised diffusion streams (fingers, global trajectory, upper- and lower-body) are an intuitive way to handle different motion scales.

Motion-prior idea – pre-training on non-speech (100-Style + AMASS) and sign-language data is well motivated.

Additional objective metrics – Skating, Floating, and (claimed) Semantic-Align supplement FGD/BA/Diversity and target common artefacts.

Comprehensive ablation – semi-implicit diffusion, semantic alignment, and both motion priors are individually evaluated.

doi.org/10.32388/CKO11G

Qualitative material – videos illustrate reduced foot-sliding and smoother trajectories.

Weaknesses

Unverified claims – no experiment for music-to-dance or for faster inference.

Missing baseline – GestureDiffuCLIP[1] (diffusion + semantic alignment) is not compared.

Finger-specific evaluation absent – improvements on fingers are asserted but not measured.

Semantic-Alignment metric nearly identical to previous work – originality is unclear, and cross-paper comparison is hindered.

Related-work section hard to follow – key techniques (e.g., EMAGE's body-part VAEs, Listen, Denoise, Action [2]) are not fully discussed.

Equation details scarce – variables in Eqs. 1–4 and 8 are undefined.

Only one evaluation dataset – despite stating "multiple datasets," all experiments use BEAT-X.

Detailed review

Introduction

Good overview of VAE smoothness vs. diffusion cost, but the claim that VAEs lack motion priors is debatable—many VAEs are trained *only* on motion.

The finger-monotony argument needs references; SMPL-X provides 30 finger DoF vs. 25 for the rest of the body, so joint training can, in fact, over-emphasise fingers.

The promised music-to-dance generalisation is not evaluated.

Related Work

The section is difficult to navigate and rarely links methods to HoloGest's stated issues.

Techniques in EMAGE (body-focused VAEs, external translation predictor, mixed speech/non-speech training) are germane but not analysed.

Key diffusion papers such as Listen-Denoise-Action and GestureDiffuCLIP are absent.

Diffusion generative models

Additional diffusion references should be cited, especially Listen-Denoise-Action (also applied to music-to-dance).

Method

Sub-section	Notes
System overview	Built on MDM; two blocks: audio-conditioned decoupled diffusion and motion-prior refiners. JEPA is used for semantic cues, meaning the model is also text-conditioned. Figure 2 is clear but sub-module labels (AFD, D) would help.
Decoupled diffusion denoiser	Explanation of MDM is fine, but Eq. 1 lacks definitions for α, σ, ε, z. The "holistic modelling hurts finger expressiveness" claim needs evidence or a citation. Decoupling into three parts plus a global token is a strong design choice.
Semi-implicit matching constraint	Eqs. 2–4 omit definitions for q, p, Dø. The GAN-style objective is interesting, but the distributions must be clarified.
Motion-prior optimiser	Global prior trained on non-speech AMASS + 100-Style; finger prior on SignAvatar then fine-tuned on BEAT-X. An ablation on training data sources would show the true benefit of non-speech motion.
Semantic alignment	Contrastive BERT-motion module is almost the same as GestureDiffuCLIP's; the paper should acknowledge this and justify the average-pool modification.

Experiments

Metrics – FGD, BA, Diversity plus Skating, Floating, SA. SA is too similar to GestureDiffuCLIP's semantic score (SC); explain why SC was not reused. Eq. 8 does not define Vg, Vs.

Baselines – Three diffusion and three VAE models are solid, but omitting GestureDiffuCLIP is problematic.

Results – HoloGest outperforms on all metrics, particularly Skating/Floating. Yet without a finger-centric metric, finger claims cannot be judged.

Qualitative comparison

Examples match the quantitative findings and the supplementary videos are helpful.

Ablation study

Shows clear impact of each component (semi-implicit diffusion, semantic alignment, priors). Again, add a finger metric and a dataset-ablation for priors.

User study

Protocol is described, but the number of stimuli per participant is missing. Authors mention a subjective test without motion prior, but the corresponding results are absent from Table 4.

Overall consistency issues

Only BEAT-X is used, despite "multiple datasets" claim.

No music-to-dance demo contradicts the generalisation claim.

Faster inference is claimed but never measured.

Final recommendation

Revise and resubmit after major revision.

The architecture and results are promising, but key evidence is missing. Adding a music-to-dance evaluation, inference-time benchmarks, a finger-specific metric and a comparison with GestureDiffuCLIP—along with clearer equations and a cleaned-up related-work section—would substantially improve the paper.

References

1. ^ *Tenglong Ao, Zeyi Zhang, Libin Liu. (2023). GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. ACM Trans. Graph., vol. 42 (4), 1-18. doi:10.1145/3592097.*

2. ^ *Simon Alexanderson, Rajmund Nagy, Jonas Beskow, Gustav Eje Henter. (2023). Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models. ACM Trans. Graph., vol. 42 (4), 1-20. doi:10.1145/3592458.*

Declarations

Potential competing interests: No potential competing interests to declare.