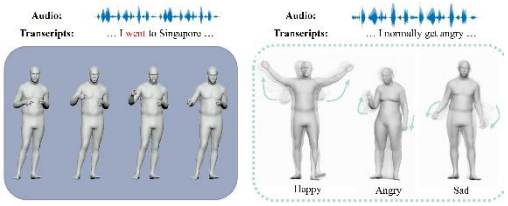
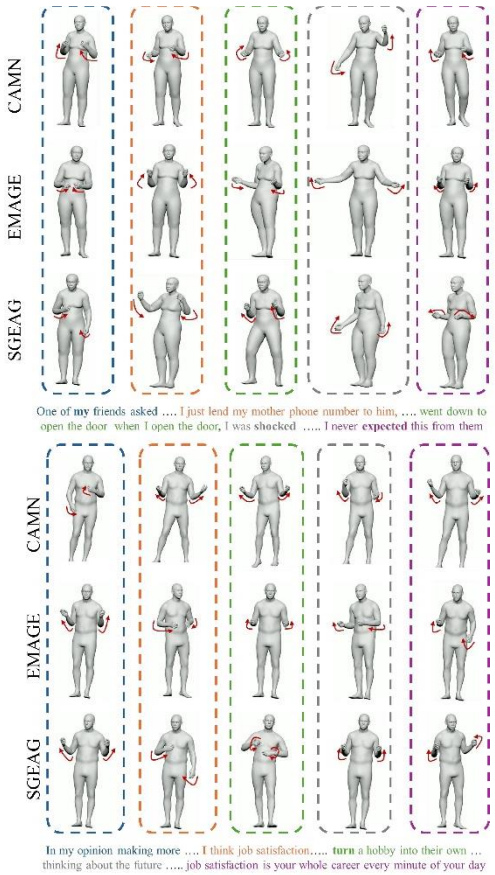
SGEAG: Semantic-guided emotional-aware gesture generation from audio

Gestures and facial expressions are fundamental to natural human communication, yet synthesizing them from speech remains a complex challenge due to the many-to-many mapping between audio and motion, the scarcity of semantic gestures, and the difficulty of capturing emotional nuance. While facial expressions tend to align with phonetic cues, whole-body gestures are more strongly driven by rhythm, semantics, and emotional context. Prior works have made progress in rhythm-aligned gesture generation, but they often neglect the semantic intent of speech [1] and the emotional tone of the speaker [2], resulting in gestures that appear generic or lack expressiveness. To address this gap, we propose SGEAG, a framework that generates whole-body co-speech gestures, including face, hands, upper and lower body, directly from audio. SGEAG introduces a two-module architecture that separates gesture generation into (1) an audio2face module for mapping speech content to facial expressions, and (2) an audio2body module that leverages rhythm and semantic cues to drive natural body movements. A key point is the Semantic-Guided Mechanism (SGM), which dynamically regulates the importance of rhythm versus semantic features, enabling the model to capture rare but meaningful gestures. To ensure expressivity, we further incorporate style and emotion adaptation applied separately to the face and body, allowing the generated gestures to reflect speaker individuality.

Experiments on the BEAT2 dataset show that SGEAG consistently outperforms state-of-the-art methods such as EMAGE [3] and CAMN [4]. Quantitatively, our model reduces the Fréchet Gesture Distance (FGD) to 0.609, improves Beat Alignment (BA) to 0.811, and achieves superior emotion classification accuracy at 64.75%, a gain of more than 13% compared to EMAGE. Qualitative evaluations further confirm that SGEAG produces motion sequences that are semantically relevant and emotionally expressive. Ablation studies highlight the crucial role of both the semantic-guided mechanism and the style-emotion adaptation modules, showing significant performance degradation when these components are removed.



Semantic and emotional influenced gestures



Qualitative comparison of the proposed method (SGEAG) with CAMN, EMAGE over BEAT2 dataset for two different audios. Red arrows show the motion sequences direction.

Beyond numerical performance, a user study with 20 participants demonstrated that gestures generated by SGEAG were perceived as more natural, diverse, emotionally aligned, and semantically appropriate than those of competing methods. Participants rated our model highest across all categories, with improvements in perceived naturalness (0.82) and emotional alignment (0.79). These findings suggest that the benefits of SGEAG extend beyond computational metrics to the level of human perception, an essential benchmark for real-world deployment.

The innovation of SGEAG lies in its joint modeling of semantics, rhythm, and emotion for co-speech gesture generation, enabled through modular design, cross-attention between face and body, and semantic-guided regulation of feature importance. This contribution has significant implications for virtual humans, digital assistants, educational platforms, and immersive entertainment, where natural and expressive gestures enhance interaction quality. Future research may expand SGEAG by incorporating multilingual and culturally diverse datasets, broadening its generalizability across communication contexts.

References

[1] Z. Zhang et al., "Semantic Gesticulator: Semantics-Aware Co-Speech Gesture Synthesis," *ACM TOG*, 43(4), 2024.
[2] K. Chhatre et al., "Emotional speech-driven 3D body animation via disentangled latent diffusion," *CVPR*, pp. 1942–1953, 2024.
[3] H. Liu et al., "EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Expressive Masked Audio Gesture Modeling," *CVPR*, pp. 1144–1154, 2024.
[4] H. Liu et al., "BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis," *ECCV*, pp. 612–630, 2022.
[5] T. Ao et al., "GesturediffuCLIP: Gesture diffusion model with CLIP latents," *ACM TOG*, 42(4), pp. 1–18, 2023