SemTalk: Holistic Co-speech Motion Generation with Frame-level Semantic Emphasis

Supplementary Material

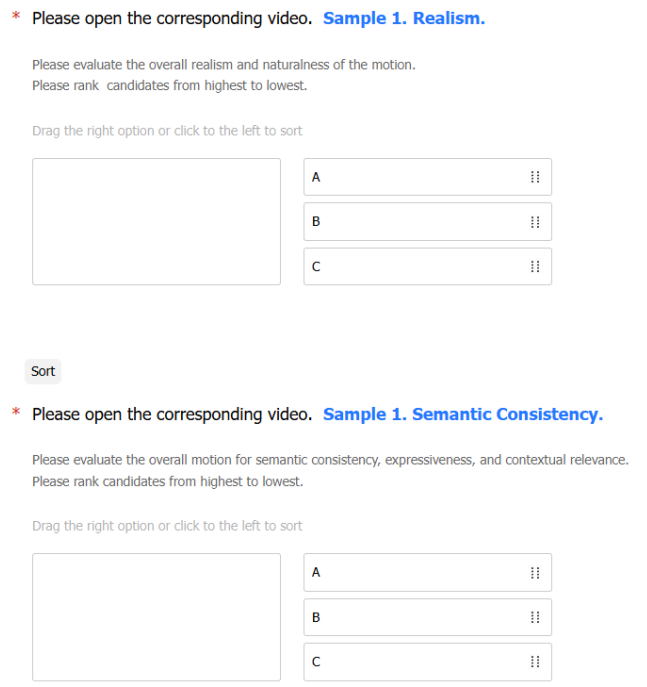

Figure 1. **Part of our user study questionnaire.** Participants are asked to rank the outputs from various methods, assessing them on several criteria to capture their preferences and perceptions of quality.

A. User Study

The user study was conducted using the Tencent Questionnaire platform, as shown in the partial view of the questionnaire in Fig. 1. To provide a seamless viewing experience and avoid potential playback issues such as latency or audio-video misalignment, participants were asked to download the videos to their local devices, following [3]. The video sequence was randomized for each participant to minimize order-related biases, with Fig. 2 illustrating an example of the shuffled video display. This ranking process enables us to gather comparative insights into each method's performance from the user's perspective, providing valuable data for understanding user satisfaction and the perceived effectiveness of each approach.

B. Evaluation Metrics

In line with the metrics established by [5], we use several measures to evaluate the quality and expressiveness of generated gestures. **Fréchet Gesture Distance (FGD)** quantifies the alignment between the distribution of generated gestures and real gestures, with lower FGD values indicating a closer match. Defined as

$$FGD(g, \hat{g}) = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}), \quad (1)$$

FGD uses the mean (μ) and covariance (Σ) of the real (r) and generated (g) gesture distributions, based on latent features extracted from a pre-trained SKCNN encoder [1], chosen for its effectiveness in feature capture. **L1 Diversity** captures the variation in generated gestures by calculating the average L1 distance between pairs of motion clips, focusing on local motion by setting translations to zero. **Beat Constancy (BC)** assesses rhythm synchronization by comparing gesture beats—identified as local minima in joint velocities—with audio beats, where a higher BC score indicates better alignment with the audio rhythm.

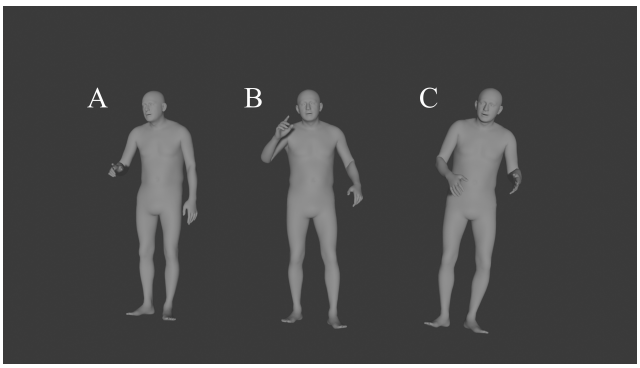

Figure 2. **A screenshot from the video used in our user study.**

C. Analysis on Smoothness and Agility

Metrics for smoothness and agility. Following [3], the comparison of agility and smoothness metrics on BEAT2 [5] 15 test data, as outlined in Table 1, highlights the strengths of our method relative to the baselines. Three key metrics are computed to evaluate these aspects: **AE** (mean acceleration error), **Vel** (mean velocity), and **MLVS** (mean local velocity standard deviation of 3D joint positions). The results reveal several key insights:

- **Smoothness (AE)**: Our method achieves an AE of 7.486, which is the lowest among all methods evaluated, indicating superior smoothness. Lower AE values suggest a closer alignment to ground truth (GT) in terms of motion smoothness. In contrast, EMAGE and DiffSHEG show

Figure 3. **Motion, velocity, and acceleration comparison on BEAT.** The motion is from 2_scott_0_8_8 in the BEAT test set. Smoother graphs reflect smoother motion. (a) Motion: frame-wise mean change in joint positions. (b) Velocity: mean absolute residuals of adjacent frame motions. (c) Acceleration: mean absolute residuals of adjacent frame velocities.

	GT	Ours	EMAGE	DiffSHEG
AE ↓	0	**7.486**	9.991	9.113
Vel §	1.221	**1.325**	1.554	0.998
MLVS §	6.600	**7.466**	8.666	5.731

Table 1. **Smoothness and agility metric results.** ↓ means the lower the better, and § means the closer to GT the better. We report AE$\times 10^-3$, Vel$\times 10^-2$ and MLVS$\times 10^-3$ for simplify

higher AE values of 9.991 and 9.113, respectively, reflecting less smoothness compared to our approach.

- **Agility (Vel and MLVS)**: For Vel and MLVS, which measure motion agility and should ideally be close to GT, our method again performs best, with values of 1.325 for Vel and 7.466 for MLVS closest to GT. While DiffSHEG achieves a Vel value of 0.998, it sacrifices MLVS consistency, with a notably lower value (5.731) that suggests slower and less dynamic motion. EMAGE, on the other hand, exhibits higher values for both metrics (1.554 for Vel and 8.666 for MLVS), indicative of more exaggerated and less realistic movements.

Numerical Observations and Visualizations. The quantitative motion analysis in Fig. 3 highlights each method's ability to replicate realistic human motion. Displacement curves in Fig. 3(a) show that DiffSHEG suffers from erratic fluctuations, causing jitter, while EMAGE overly smooths transitions, losing natural rhythm and expressiveness. Velocity analysis in Fig. 3(b) reveals that our method maintains consistent velocity profiles, while DiffSHEG exhibits excessive spikes, leading to unnatural energy, and EMAGE generates overly dampened motions. Acceleration curves in Fig. 3(c) further show that our approach ensures stable transitions, whereas DiffSHEG introduces abrupt peaks, disrupting coherence, and EMAGE's overly muted acceleration fails to capture natural motion variability. SemTalk achieves a balance between smoothness and dynamism, closely aligning with natural human motion.

K	G	FGD↓	BC↑	DIV↑	MSE↓	LVD↓
512	3	5.521	7.638	11.02	6.851	7.313
512	6	5.022	7.626	11.97	6.829	7.305
256	3	4.835	7.742	11.18	6.563	7.201
256	6	**4.397**	**7.776**	**12.49**	**6.100**	**6.898**

Table 2. **Performance of SemTalk* with different hyper parameters of RVQ-VAE.** K denotes codebook size, and G means the number of codebooks.

$\mathcal{L}_{\text{Rhy}}^{(L)}$	$\mathcal{L}_{\text{Rhy}}^{(G)}$	FGD↓	BC↑	DIV↑	MSE↓	LVD↓
-	-	4.897	7.702	12.42	13.416	15.72
√	-	4.438	7.722	12.03	6.590	7.218
-	√	4.522	7.740	11.86	6.988	7.379
√	√	**4.397**	**7.776**	**12.49**	**6.100**	**6.898**

Table 3. **Ablation study on local-global consistency loss.**

D. Supplementary Ablation Study

Effectiveness of RVQ-VAE. To assess the impact of the Residual Vector Quantization Variational Autoencoder (RVQ-VAE) [2, 4, 6] on SemTalk*'s performance, we conducted experiments on the BEAT2 dataset. The results show that increasing K alone offers limited benefits, while a smaller K with an optimal G significantly enhances RVQ-VAE's representation capacity. This configuration improves alignment, diversity, and accuracy in generated motions, underscoring the importance of RVQ-VAE parameters to capture both rhythmic and semantic nuances for expressive, context-aware motion generation.

Effectiveness of local-global consistency loss. To assess the impact of local-global consistency loss on rhythmic alignment, we performed an ablation study on the BEAT2 datasets using combinations of local frame-level consistency loss $\mathcal{L}_{\text{Rhy}}^{(L)}$ and global sentence-level consistency loss $\mathcal{L}_{\text{Rhy}}^{(G)}$. The results, presented in Tab. 4, highlight the distinct contributions of these components to gesture quality, rhythmic alignment, and temporal coherence. Adding $\mathcal{L}_{\text{Rhy}}^{(L)}$ improves fine-grained synchronization at the frame level, capturing subtle temporal variations. Incorporating $\mathcal{L}_{\text{Rhy}}^{(G)}$ enhances rhythmic coherence across the sequence, ensuring smooth and consistent motions. The combination of both losses achieves the best results, balancing fine-grained alignment with long-term rhythmic coherence. These findings validate the effectiveness of rhythmic consistency loss in addressing both short-term and long-term alignment challenges in co-speech motion generation.

Effectiveness of semantic emphasis learning and separate learning. To assess the impact of semantic emphasis learning and separate learning of semantic and rhythmic features, we conducted an ablation study comparing different integration methods for semantic features f_t with rhythmic features γ. In SemTalk†, semantic features are directly added to rhythmic features without weighting. While this approach introduces semantic information, it is dominated by rhythm-related cues, failing to effectively capture sparse and meaningful gestures, resulting in limited expressiveness and diversity. SemTalk‡ amplifies the loss weight of semantic features based on semantic scores, but this static weighting brings only minor improvements over SemTalk†, showing limited effectiveness in balancing expressiveness and rhythmic alignment. The full SemTalk framework, incorporating semantic emphasis learning and separate modeling of semantic and rhythmic features, achieves the best performance. By dynamically emphasizing semantically relevant frames, SemTalk balances semantic expressiveness with rhythmic alignment, significantly improving diversity, coherence, and synchronization with speech. These findings emphasize the importance of separate learning and dynamic integration in achieving expressive, rhythmically consistent gestures.

Method	FGD↓	BC↑	DIV↑	MSE↓	LVD↓
SemTalk†	4.893	7.740	11.86	6.643	7.200
SemTalk‡	4.891	7.722	11.99	6.988	7.379
SemTalk	**4.278**	**7.770**	**12.91**	**6.153**	**6.938**

Table 4. **Ablation study on semantic emphasis learning.** † denotes the direct addition of semantic features f_t to rhythmic features γ, while ‡ represents the same operation, followed by amplifying the loss weight based on the semantic score.

β	0.3	0.5	0.7	0.9
FGD	4.703	**4.278**	4.351	4.379

Table 5. **Ablation study on β.**

Semantic Threshold Beta. To balance base and sparse motions, we set β to 0.5, as Tab. 5 show this optimally preserves smoothness while enhancing semantic emphasis. User studies and demo videos confirm natural, seamless motion transitions.

E. Limitations

While SemTalk achieves state-of-the-art results in co-speech motion generation, it has certain limitations. First, the framework relies on high-quality, well-aligned training data to effectively capture rhythmic and semantic features. In real-world scenarios with noisy or misaligned data, performance may degrade, reducing robustness in practical applications. Second, the computational complexity of separate learning processes and adaptive semantic emphasis increases training time and resource demands. Future improvements, such as robust noise handling, data augmentation, or more efficient model architectures, could enhance the practicality and scalability of SemTalk.

References

[1] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 39(4):62–1, 2020. 1

[2] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023. 3

[3] Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7361, 2024. 1

[4] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 3

[5] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1154, 2024. 1

[6] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021. 3