SemGes: Semantics-aware Co-Speech Gesture Generation using Semantic Coherence and Relevance Learning

Supplementary Material

1. Objective Metrics

FGD evaluates the distributional alignment between generated outputs and ground truth motions over an entire dataset. This is achieved by using a pre-trained autoencoder, which embeds motion sequences into a latent space, enabling a fine-grained comparison of their statistical properties,

$$\begin{aligned}\text{FGD}(\mathbf{g},\hat{\mathbf{g}}) = \|\mu_r - \mu_g\|^2 + \\ \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right)\end{aligned} \tag{1}$$

where μ_r and Σ_r denote the first and second moments of the latent feature distribution z_r, representing real human gestures $\mathbf{g}$. Similarly, μ_g and Σ_g correspond to the first and second moments of the latent feature distribution z_g, representing the generated gestures $\hat{\mathbf{g}}$. These moments capture the statistical characteristics of the respective distributions, enabling a robust evaluation of the alignment between real and synthesized gesture dynamics.

BC Score[2] evaluates the temporal alignment between the rhythmic patterns of gestures and speech beats. This is achieved by defining the onset of speech as the audio beat and identifying motion beats as the local minima in the velocity of upper body joints (excluding finger movements). The degree of synchronization between speech and gesture rhythms is quantified as follows:

$$BA = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-\frac{\min_{\forall b_j^a \in B^a} \left|b_i^m - b_j^a\right|^2}{2\sigma^2}\right) \tag{2}$$

where $B^m = \{b_i^m\}$ represents the set of motion beats, while $B^a = \{b_j^a\}$ corresponds to the set of audio beats.

Diversity [1] quantifies the variablity of generated motions. We compute the average $L1$ distance across pairs of N motion clips. The diversity metric is then defined as:

$$\text{Diversity} = \frac{1}{2N(N-1)}\sum_{t=1}^{N}\sum_{j=1}^{N} \|p_t^i - \tilde{p}_t^j\|_1, \tag{3}$$

where p_t denotes the joint positions at frame t. This evaluation is conducted on the entire test dataset. To ensure the metric captures local motion dynamics, translations are normalized to zero, isolating the diversity of joint movements independent of global positional shifts.

SRGR [3] evaluates the semantic relevance of generated gestures, indicating how they align with annotated ground truth gestures. The metric employs the Probability of Correct Keypoint (PCK), weighted by semantic scores, to measure the percentage of joints recalled within a specified threshold δ. SRGR is computed as:

$$D_{\text{SRGR}} = \lambda \frac{1}{T \times J}\sum_{t=1}^{T}\sum_{j=1}^{J} \nVdash \left(\|\mathbf{p}_{t,j} - \hat{\mathbf{p}}_{t,j}\|_2 < \delta\right), \tag{4}$$

where T is the number of frames, J the number of joints, $\nVdash$ the indicator function, and λ a weighting parameter.

2. Models and Implementation Details

Implementation Details We implement all models using PyTorch and a single NVIDIA Tesla A100 GPU. Training is in two stages. For Beat, we train the first-stage VQVAE generator with the Adam optimizer (lr = 0.0015) for 500 epochs, using a batch size of 256. The second stage is then trained with the AdamW optimizer (lr = 0.0003) for 40 epochs at a batch size of 128. For Ted Expressive, we train the first-stage VQVAE generator with the Adam optimizer (lr = 0.002) for 300 epochs, using a batch size of 256, followed by training the second stage with the AdamW optimizer (lr = 0.0002) for 30 epochs at a batch size of 128.

VQ-VAEs Details The VQ-VAE model encodes body or hand motion inputs into a discrete latent space before reconstructing them through a decoder. Its architecture follows a symmetric design, including an encoder, a vector quantization module, and a decoder.

The encoder consists of $n_{\text{down}} = 3$ convolutional layers with residual blocks to extract hierarchical motion features. The first layer applies a 1D convolution with a kernel size of 3, a stride of 1, and padding of 1, followed by batch normalization and the Leaky ReLU[4] activation function (negative slope = 0.2). Each subsequent convolutional layer keeps these parameters.

The vector quantization (VQ) module discretizes continuous latent representations into a finite codebook. The codebook dimension is $e_{\text{dim}} = 256$. Given an encoded feature z, the quantization module assigns each feature vector to the closest codebook entry by minimizing the Euclidean distance. The discrete representation is then passed to the decoder for reconstruction. Training is guided by a commitment loss, weighted by $\beta = 0.25$, to enforce alignment between the encoder output and the codebook entries.

The decoder mirrors the encoder structure, consisting of $n_{\text{up}} = 3$ convolutional layers with residual blocks. It progressively reconstructs motion features from the quantized latent space. Similar to the encoder, batch normalization and Leaky ReLU activation are applied to stabilize training. The final decoder layer restores the original feature dimensionality.

Semantic Gesture Generator is a transformer-based model designed to generate gesture representations from multimodal inputs, including audio embeddings, speaker identity features, and word embeddings. The model architecture is based on a transformer encoder with cross-modal attention mechanisms, integrating self-attention and cross-modal attention layers to capture dependencies between modalities.

The Transformer consists of $n_{\text{layers}} = 8$ transformer layers, each containing a multi-head self-attention module (heads = 8), a cross-modal attention module, and a feed-forward multi-layer perceptron (MLP). The self-attention mechanism utilizes query-key-value projections through a linear layer with an output dimension of hidden size (768). Attention scores are computed via scaled dot-product attention and normalized using softmax. To preserve temporal structure, we employ a Positional Encoding module that encodes sequence order using sine and cosine functions, computed for 5000 time steps. Input embeddings are projected via a Linear Embedding layer.

Speaker identity embeddings are learned using an embedding layer (embedding dimension = 32) for the Beat dataset, 2048 for TED Expressive (embedding dimension = 256), followed by a Leaky ReLU activation[5] (negative slope = 0.1). The model also incorporates Instance Normalization within its convolutional layers to normalize feature distributions across speakers.

For word embeddings, the model utilizes a Temporal Convolutional Network (TCN) with a kernel size of 2 and multiple residual blocks. Each TemporalBlock consists of two weight-normalized 1D convolution layers with ReLU activations, dropout (0.2), and residual connections.

3. Gesture Generation in Difficult Cases

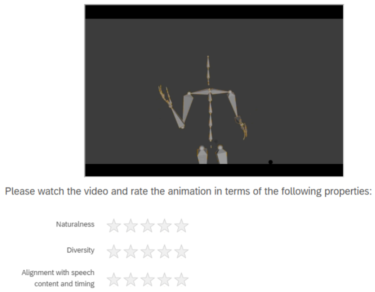
We evaluate how our speech-driven gesture generation model performs under challenging conditions, e.g., when speech is noisy or when speech is not aligned with gestures. We evaluate the model's robustness in two ways. (i) We introduce semantically misaligned text to each gesture. The table below shows that with compromised semantic alignment, the model remains functional but shifts toward rhythm-oriented behaviour with high BC scores. Beat gestures maintain robust temporal synchronization even under degraded semantic conditions due to their dependency on rhythm rather than semantics. (ii) We added Gaussian noise to the audio, which yielded reasonable and generic gestures but diminished output quality (See Table 1).

Table 1. Objective scores of the generation model when using audio with Gaussian noise.

Dataset	FGD ↓	BC ↑	Diversity ↑	SRGR ↑
BEAT Dataset	8.421	0.630	278.539	0.156
TED-Expressive	7.984	0.629	108.892	–

4. Interface for User Study

Fig. 1 presents a screenshot from the user-rating study questionnaire, which we administer through Qualtrics. It contains 24 questions. We integrated this questionnaire into the Prolific platform to recruit participants from English-speaking countries. The screenshot Fig. 2 also includes an attention-checking question. Only responses that answer the attention questions correctly and have a completion time exceeding 20 minutes are considered valid. Additionally, the questionnaire survey video was rendered using Blender to motion capture dataset type. The video is presented in a randomized order to mitigate potential response bias.



Please watch the video and rate the animation in terms of the following properties:

Naturalness ☆☆☆☆☆

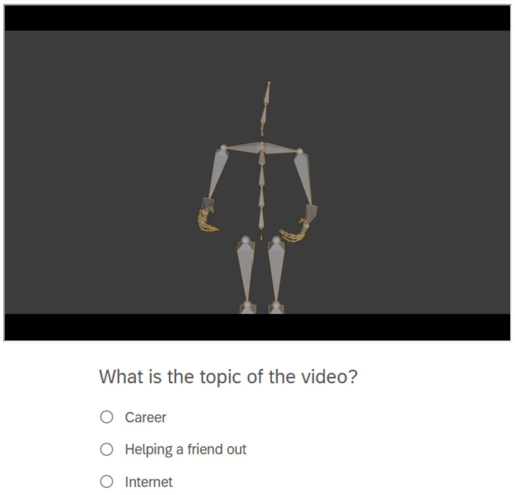
Diversity ☆☆☆☆☆

Alignment with speech content and timing ☆☆☆☆☆

Figure 1. Questionnaire Interface

5. Limitations and Future Directions

While our approach introduces a step forward towards semantics-aware gesture generation and improves the exploitation of speech semantics, it has a few limitations. First, our method is data-driven and, therefore, depends on the limited semantic annotations available in current datasets; future work could leverage richer annotations or employ large language models (LLMs) to supplement this data informed by linguistic research. Second, our model does not yet incorporate facial expressions, which are crucial for conveying semantic content, a natural next step for our work. Finally, our approach



What is the topic of the video?

○ Career

○ Helping a friend out

○ Internet

Figure 2. Attention Check Question Interface

might benefit from more powerful semantic and speech encoders, which we leave for future exploration.

References

[1] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11293–11302, 2021. 1

[2] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 1

[3] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Conference on Computer Vision*, pages 612–630. Springer, 2022. 1

[4] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, page 3. Atlanta, GA, 2013. 1

[5] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015. 2