Towards realtime co-speech gestures synthesis using STARGATE

Louis Abel, Vincent Colotte, Slim Ouni

Towards realtime co-speech gestures synthesis using STARGATE

Louis ABEL[1]*, Vincent COLOTTE*[1]*, Slim OUNI*[1]

[1]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

louis.abel@loria.fr, vincent.colotte@loria.fr, slim.ouni@loria.fr

Abstract

The field of co-speech gestures synthesis is gaining more and more interest. However, many new systems utilize complex or resource-intensive architectures, making them impractical for integration into Embodied Conversational Agents (ECAs) or for exploration in fields like linguistics, where understanding the connection between speech and gestures is challenging. This paper introduces STARGATE, a novel architecture for Spatio-Temporal Autoregressive Graph from Audio-Text Embeddings. The model leverages autoregression for fast gestures generation, alongside graph convolutions and attention to integrate explicit structural knowledge and facilitate efficient spatial and temporal processing. Through both subjective and objective assessments against state-of-the-art models, our research demonstrates our model capabilities of generating convincing gestures fast. It also achieves slightly better scores in terms of credibility and coherence of generated gestures in relation to speech.

1. Introduction

Co-speech gesture synthesis has rapidly gained traction in recent years. While the exact mechanisms behind human gesture generation and its link to speech remain under investigation, researchers have made significant progress in developing methodologies for generating gestures from speech data, encompassing both spoken transcripts [1] and acoustic signals [2]. The ubiquity of gestures in human communication underscores their importance role in simulating natural human interactions.

To understand and integrate gestures into artificial communication, researchers have explored gesture analysis and classification [3]. Initially, rule-based systems were employed to create Embodied Conversational Agents (ECA) [4], drawing insights from neuroscience and linguistics. However, these early systems were rudimentary and often inconsistent with findings from various literature sources. The absence of a unified classification scheme for gestures (e.g.,[3, 5, 6]) and the varying conclusions regarding the relationship between gestures and speech within these frameworks [7, 8, 9] hindered the development of reliable and consistent rules.

In recent years, data-driven approaches have emerged as a promising avenue for implicitly extracting the intricate patterns and rules governing the relationship between speech and gesture. These approaches encompass a spectrum of architectures, ranging from basic autoencoders [10, 11] to more sophisticated models such as variational autoencoders (VAEs) and conditional VAEs [12, 13], with the aim of capturing a broader array of gestures and enhancing conditioning from speech input.

Significantly, StyleGestures [14] has garnered attention for its pioneering autoregressive architecture integrating normalizing flow techniques [15]. Normalizing flow, a specialized neural network approach, adeptly captures intricate distributions. This model has emerged as a cornerstone for benchmarking gesture synthesis systems, as evidenced by its widespread adoption in subsequent studies [12, 2, 15]. Furthermore, its selection as the baseline model for the GENEA Challenge [16] underscores its lasting impact on the field.

While diffusion models [2, 17, 18, 1] have achieved impressive results in generating high-fidelity gesture sequences, their complex architectures often lead to slower processing times. This trade-off between quality and speed presents a significant challenge in our field. Developing new theoretical frameworks holds potential for furthering our understanding of the speech-gesture relationship. Additionally, faster and more responsive models would significantly benefit gesture-enabled ECAs.

To address this challenge, we propose exploring graph convolutional networks (GCNs) [19] as a lighter and more interpretable alternative. GCNs are a type of neural network specifically designed to work with graph-structured data, where data points (nodes) are connected by edges. This structure inherently aligns well with the modeling of skeletal structures, which can be naturally represented as graphs where body joints are nodes and bones are edges. Inspired by the success of GCNs in locomotion synthesis, a field closely related to gesture generation but without speech input, we believe GCNs hold promise for our work. They offer the potential to create a lightweight and efficient architecture capable of generating realistic gestures while considering anatomical constraints.

Motivated by these advancements, we introduce an innovative network architecture designed to overcome the previously mentioned constraints in gesture synthesis. Our proposed architecture aims to accomplish two primary goals:

- Leveraging graph convolutions to capture explicit gesture structure within a deep neural network, ultimately aiming to generate convincing gestures.
- Efficient design using an autoregressive architecture, to accommodate applications where speed is critical, such as in ECAs.

In the following sections, we present our novel architecture and the methodologies employed, followed by a comprehensive evaluation, using both quantitative metrics and subjective evaluations. We finalize by analyzing the obtained results and considering potential future directions for our model.

2. Methods

We propose a novel architecture named STARGATE (for Spatio-Temporal Auto-Regressive Graph from Audio-Text Embeddings), an overview is depicted in Figure 1. This architecture follows an encoder-decoder structure, employing a

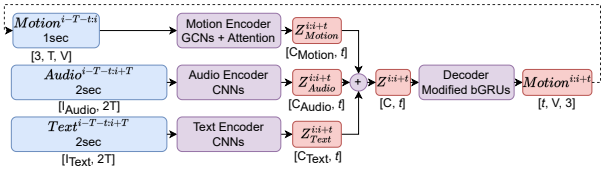

Figure 1: *An overview of the STARGATE network with its encoder-decoder structure. Our network use 3 separate encoders to process all 3 modalities separately and a unique decoder to generate motion from a multimodal latent representation. Numbers between brackets depicts tensor shapes: T being the half-window length, V the number of joints, t the chunk size, C_x the latent feature size, I_x the input feature size.*

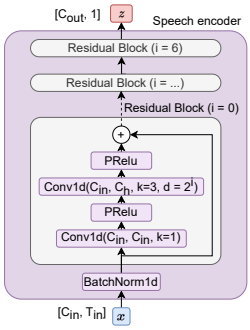

Figure 2: *The speech encoder, used for both audio and text encoding separately. Conv1D parameters have the following meaning : input channels, output channels, kernel size, dilation size (default = 1) which double at each layer.*

chunked-autoregressive approach. This means that the network takes input from three different modalities:

- **Audio**: A window of 1s of past and 1s of future speech.
- **Text**: A window of 1s of past and 1s of future words.
- **Motion**: A history of 1s of past motions.

The choice of such context window is driven by the slow nature of gestures, with an average duration of 1-2 seconds depending on whether the gestures correspond to a single word or an entire sentence [20]. Each modality has a dedicated encoder to generate a specific latent space representation, which are then fused to form a multimodal representation of speech/gestures. This representation is subsequently decoded into a chunk of next gesturer poses spanning t frames. We opt for chunk output instead of frame-by-frame output to allow more flexibility for gesture generation without overly relying on the autoregressive motion history, but also to have more efficient computations. The first second of motion history is a sequence of zeros, to start the autoregression loop.

2.1. Speech encoders

Speech can be categorized into two primary components: acoustic content and linguistic content. The acoustic signal produced during speech carries various pieces of information, such as prosody or emotional state. Meanwhile, the linguistic content, which is also part of the acoustic signal but presents a phonetic representation of what has been spoken, conveys semantic information from the text. Text serves as a crucial source of information for modeling iconic, deictic, and metaphoric gestures, all of which are directly linked to semantic content, while beat gestures, the last category according to [3], are associated with the acoustic signal. Therefore, both modalities (acoustic and textual) are essential for generating dynamic and meaningful gestures. In our architecture, we employ both modalities through two similar but distinct CNN-based encoders, as illustrated in Figure 2. The audio encoder takes 27 mel-frequency cepstrum coefficients (MFCCs) as input, with convolution channels set to: 64, 96, 128, 128, 256, 256. Meanwhile, the text encoder uses BERT embeddings [21], where embeddings are tiled based on word segmentation to obtain frame-level BERT embeddings (similar to the procedure of [22] for their text input). This encoder uses convolution channels configured as follows: 768, 768, 512, 512, 396, 396.

2.2. Motion encoder

Since we are operating within an autoregressive framework, we have the capability to use motion as an input for subsequent predictions. Consequently, our third input modality consists of a history of previous motion, aiming to maintain coherent trajectories in long-term synthesis and to establish a speech-gestures multimodal latent space in the decoder. Our motion encoder is based on the work of [23]. Input motions are represented using exponential maps, which offer numerous advantages, including serving as a continuous representation over Euler angles and being more compact than quaternions.

2.2.1. Graph Neural Network

In order to generate human-like gestures and to delve into how our network generates them, we integrated multiple mechanisms within our motion encoder. The primary mechanism involves employing a graph convolution network (GCN) [19] instead of traditional CNNs. In the context of a graph, convolutions are computed using an adjacency matrix to determine neighboring nodes. In our approach, we adopted the ST-GCN (for Spatio-Temporal GCN) block from [23], which incorporate multiple adjacency matrices, each containing specific links. These are coupled with a temporal convolution network (TCN) to create a network capable of efficiently processing spatio-temporal data such as motion. Further details can be found in the original publication.

To the best of our knowledge, our adaptation is the first work in the field of co-speech gesture synthesis to employ graph convolutions for injecting prior knowledge of gestures and to obtain a more explicit representation of motion.

2.2.2. Attention mechanism

Both in the ST-GCN model by [23] and in our implementation, a self-attention mechanism is applied to the adjacency matrices prior to graph convolutions. The input motion data undergoes a scaled dot-product self-attention process to generate an 'attention matrix' for each adjacency matrix. These 'attention matrices' are then combined with the base adjacency matrices to produce what we refer to as 'dynamic adjacency matrices'. This approach is motivated by the observation that while we allow the network to make minor adjustments to the adjacency matrices during training, they remain static during inference. The incorporation of this attention mechanism enables dynamic modifications during inference, allowing the network to focus more on specific body parts for each chunk of generated frames.

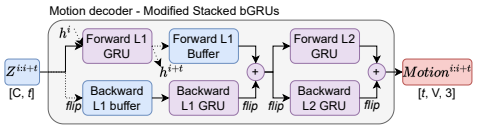

Figure 3: *Our decoder uses a stacked GRUs approach. We implemented bidirectionality using buffers at first layer (L1) GRUs, the forward one storing forward L1 hidden features, the backward one storing longer latent sequence.*

2.3. Motion decoder

The audio, text, and gestures latent spaces are merged to generate a multimodal latent space spanning t frames, which is then fed into the motion decoder as depicted in Figure 3. This decoder, comprising stacked RNNs (in our case, Gated Recurrent Units, GRUs [24]), produces the next chunk of t-frame skeleton poses, which are subsequently used to compute the next batch of frames. One significant limitation of autoregression is to work only with previous information, lacking the ability to analyze complete sequences of gestures. Consequently, bidirectional GRUs could not be used to gain a comprehensive understanding of entire gesture sequences. However, motivated by the potential benefits it could offer and considering that we are generating batches of frames, bidirectional GRUs can be applied to these partial sequences. As continuity cannot be maintained across subsequent layers, only the first forward GRU layer hidden state is tracked. This bidirectional approach aims to enable the network to learn relationships between past and future information present in the multimodal representation.

3. Training

3.1. Dataset : BEAT

We conducted training for all our models using the BEAT dataset [25]. This dataset offers a large volume of high-quality multimodal data, encompassing audio recordings, word and phoneme-level transcriptions, as well as motion capture data for body, hands, and face. In our study, we used the data corresponding to the speaker 1, giving us 4 hours of data, which we split into training, validation, and test sets using a 90/5/5 ratio. All data preprocessing and augmentation follows the protocol and code proposed by StyleGestures [14].

3.2. Loss

During training, our model minimizes a combination of two Huber loss functions [26]. The first targets the exponential map, ensuring overall gesture accuracy. However, minimizing solely the exponential map treats all joints equally. To address this, we include a second Huber loss applied directly to joint positions. This prioritizes precise control over the hips and spine, crucial body parts that significantly influence the movement of all other joints (end effectors). Therefore, the loss function is defined as follows:

$$Loss = \mathcal{H}(r, \hat{r}) + \mathcal{H}(p, \hat{p})$$

With r and p respectively, the positions and exponential map of the reference sample, $\hat{r}$ and $\hat{p}$ respectively the positions and exponential map of the generated sample and $\mathcal{H}$ the Huber loss.

Figure 4: *An example of ECA using motion generated by our STARGATE network. This example highlights capabilities of generating different type of gestures such as iconic gestures.*

4. Evaluation

4.1. Quantitative metrics

This section presents evaluations of our proposed model and a variant, "Audio Only," which excludes the text encoder. Those two models are compared against state-of-the-art model. We compare the performance of these models against a well-established benchmark in the field, StyleGestures [14]. This choice is motivated by StyleGestures' autoregressive architecture and its frequent adoption as a reference model for gesture synthesis research [16].

Frechet Gesture Distance (FGD). The most promising effort to establish an objective quality metric for speech synthesis draws inspiration from the Frechet Inception Distance (FID) used in image synthesis research [27]. This approach was adapted by [28] to introduce the FGD metric. We retrained the proposed inception network because our output differs significantly from the available network. The advantage of this method is that the inception network acts as an unbiased evaluator, which allows to get a metric closer to actual human perception.

As shown in Table 1, both variants of STARGATE outperform StyleGestures, with the Audio Only variant being the best model in terms FGD.

Performance. While prioritizing performance is usually the main focus when designing a model for gesture synthesis, our objective was to create a network capable of operating in scenarios with a need for fast generating models, such as for Embodied Conversational Agents (ECAs), and producing convincing gestures as quickly as feasible. To evaluate performance within this framework, we conducted benchmarks that take into account preprocessing steps, which can notably affect computational load (e.g., BERT embedding computations). Thus, all timing results provided are derived from raw waveform/sentence input with a batch size of 1. Additionally, we present the execution time per frame to enable a fairer comparison.

The Table 1 show us that both STARGATE variants are consistently faster than StyleGestures. We can also observe that StyleGestures' performance does not improve with longer input lengths, whereas STARGATE models exhibit better performance in processing longer sequences.

4.2. Subjective Evaluations

In order to assess the gesture quality of our model more comprehensively, we conducted a Mean Opinion Score (MOS) subjective evaluation. We tailored the evaluation protocol from the GENEA Challenge [16], specifically focusing on refining the questions to gain a clearer understanding of the aspects being evaluated for each question.

	Nb params	Graph?	Audio?	Text?	FGD ↓	Inference time ↓ [Time per frame ↓] 5s	10s	30s	80s
StyleGestures	82M	✗	✓	✗	14.15	7.76s [90ms]	12.90s [70ms]	31.05s [50ms]	80.07s [26ms]
STARGATE	43.5M	✓	✓	✓	10.58	6.51s [37ms]	8.31s [17ms]	13.40s [8ms]	23.78s [5ms]
STARGATE *Audio Only*	30.9M	✓	✓	✗	**8.61**	**3.49s** [**19ms**]	**3.98s** [**8ms**]	**6.13s** [**3ms**]	**10.68s** [**2ms**]

Table 1: *Results of quantitative comparison using FGD and models benchmarks according to the duration of the utterance (5s to 80s). Note that StyleGestures outputs 20fps while our model outputs 60fps. Bold values depict best model. Benchmark used the following hardware configuration: i7-11850H and NVIDIA RTX A3000 Laptop.*

Model	Human-like ↑	Credibility ↑	Consistency ↑
Reference	6.19 ± 0.28	5.27 ± 0.23	5.16 ± 0.23
Mismatch	N/A	4.92 ± 0.20	4.77 ± 0.22
StyleGestures	5.97 ± 0.25	4.87 ± 0.22	4.70 ± 0.23
STARGATE	5.89 ± 0.28	5.0 ± 0.20	4.85 ± 0.22

Table 2: *Results of our MOS evaluation, we report the mean and a 0.95 confidence interval for each aspect.*

The evaluation process consisted of two phases. Initially, participants viewed videos of a 3D avatar gesturing without audio and rated the perceived human-likeness of the gesture motion. In the second phase, participants watched videos with accompanying audio and responded to questions regarding the credibility and consistency of the gestures relative to the speech. Responses were provided on a scale ranging from 1 to 7. The study assessed four systems: Reference (ground truth), Mismatch (synthetic motion with different audio), StyleGestures, and STARGATE. We presented 30 videos for each system, each lasting 9 seconds. We had a total of 25 participants (12 female and 13 male) for this study. Results are summarized in Table 2.

The findings indicate that StyleGestures marginally outperforms STARGATE in terms of human-likeness (without audio), whereas our model demonstrates slightly superior coherence and credibility when audio is present.

4.3. Discussions

Our primary objective was to ensure high-quality gesture generation, as evidenced by the superior performance of our models in terms of FGD, as shown in Table 1. Additionally, our MOS evaluation in Table 2 indicates slightly higher credibility and coherence scores compared to the StyleGestures model when audio is included. However, StyleGestures outperforms our model in the human-likeness aspect when audio is unavailable. We attribute this difference to the generation of semantic gestures in our model, which can occasionally result in unclear gestures, blending iconic and beat gestures and leading to perceived unnaturalness without audio cues. However, a more detailed analysis by gesture specialists is needed to fully substantiate this claim.

The absence of semantic gestures in the "Audio Only" model naturally leads to its lower FGD score. Additionally, semantic gestures present in the StyleGestures model might deviate from the reference data, as a single speech segment can be accompanied by various iconic gestures. Despite these factors, our graph architecture achieves a better FGD score compared to StyleGestures. This suggests that our model may possess a stronger ability to link gesture representations to the underlying anatomy, potentially leading to enhanced motion comprehension and a more robust connection between text and the generated motions.

Table 2 also supports previous research findings [16], where the Mismatch model performs better than StyleGestures and our model. This may be attributed to the prevalence of beat gestures in the dataset. These gestures primarily focus on synchronizing with the audio rhythm, leading to them being perceived as "correct" even if the audio does not match.

Our secondary objective was to develop a model suitable for scenarios where generation speed is crucial. In short 5s sequence generation, our model performs up to 1.4x faster than the input length, while in 80s long sequence generation, our model performs up to 7.5x faster than the input length. In comparison, StyleGestures is 1.5x slower in the first scenario and remains neither faster nor slower in the second scenario. However, this model outputs 20fps gesture sequences, whereas ours output 3 times more frames at 60fps. Thus, when considering the time per frame, our model is 4.7x faster than StyleGestures per frame generated in the short sequence scenario, and in long sequence generation, our model takes advantage of parallel processing of input modalities, becoming 13x faster per frame generated for our Audio Only variant, and 5x faster for our standard model (audio + text).

This demonstrates the capabilities of our architecture for integration into ECAs, facilitating more natural interaction with gestures-enabled avatars with low computation latency, both with audio-only and text-integrated inputs.

5. Conclusion

Our work demonstrates the potential of GCNs for co-speech gesture synthesis. To further explore this direction, future research can delve into several areas. One avenue involves developing techniques for explainable gesture generation would provide valuable insights into the model's decision-making process. By understanding how the GCN arrives at specific gesture outputs, we can potentially refine the system for improved control and overall explainability. These advancements can further solidify the role of GCNs as a powerful tool for generating natural and meaningful gestures.

Our findings highlight also the influence of beat gestures on gesture synthesis models. To encourage more semantically-driven generation, future work can explore several avenues. One approach involves disentangled representation learning, where the model learns separate representations for beat and semantic gestures, allowing for better distinction during generation. Alternatively, a loss function with semantic weighting could penalize the model for generating beat gestures when clear semantic cues are present in the audio and text input. Finally, implementing an attention mechanism on audio features could guide the model to focus on audio aspects that convey semantic meaning, rather than just the rhythm. By incorporating these techniques, we can potentially steer the model towards generating gestures that are more tightly coupled with the intended message.

6. References

[1] A. Deichler, S. Mehta, S. Alexanderson, and J. Beskow, "Diffusion-based co-speech gesture generation using joint text and audio representation," in *Proceedings of the 25th International Conference on Multimodal Interaction*, 2023, pp. 755–762.

[2] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter, "Listen, denoise, action! audio-driven motion synthesis with diffusion models," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–20, 2023.

[3] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*. Chicago and London: The University of Chicago Press, 1992.

[4] J. Cassell, "Embodied conversational agents: representation and intelligence in user interfaces," *AI magazine*, vol. 22, no. 4, pp. 67–67, 2001.

[5] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge ; New York: Cambridge University Press, 2004.

[6] D. Boutet, "Une morphologie de la gestualité : structuration articulaire," *Cahiers de linguistique analogique*, no. 5, pp. 81–115, Dec. 2008. [Online]. Available: https://hal.science/hal-00607593

[7] W. C. So, S. Kita, and S. Goldin-Meadow, "Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand," *Cognitive Science*, vol. 33, no. 1, pp. 115–125, Jan. 2009. [Online]. Available: https://doi.org/10.1111/j.1551-6709.2008.01006.x

[8] J. P. de Ruiter, A. Bangerter, and P. Dings, "The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis," *Topics in Cognitive Science*, vol. 4, no. 2, pp. 232–248, Mar. 2012. [Online]. Available: https://doi.org/10.1111/j.1756-8765.2012.01183.x

[9] R. M. Krauss and U. Hadar, "The role of speech-related arm/hand gestures in word retrieval," in *Gesture, Speech, and Sign*. Oxford University Press, Jul. 1999, pp. 93–116. [Online]. Available: https://doi.org/10.1093/acprof:oso/9780198524519.003.0006

[10] K. Takeuchi, S. Kubota, K. Suzuki, D. Hasegawa, and H. Sakuta, "Creating a gesture-speech dataset for speech-based automatic gesture generation," in *International Conference on Human-Computer Interaction*. Springer, 2017, pp. 198–202.

[11] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström, "Analyzing input and output representations for speech-driven gesture generation," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 97–104.

[12] J. Li, D. Kang, W. Pei, X. Zhe, Y. Zhang, Z. He, and L. Bao, "Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 293–11 302.

[13] S. Lu, Y. Yoon, and A. Feng, "Co-speech gesture synthesis using discrete gesture token learning," *arXiv preprint arXiv:2303.12822*, 2023.

[14] S. Alexanderson, G. E. Henter, T. Kucherenko, and J. Beskow, "Style-controllable speech-driven gesture synthesis using normalising flows," in *Computer Graphics Forum*, vol. 39, no. 2. Wiley Online Library, 2020, pp. 487–496.

[15] T. Ao, Q. Gao, Y. Lou, B. Chen, and L. Liu, "Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 6, pp. 1–19, 2022.

[16] T. Kucherenko, P. Jonell, Y. Yoon, P. Wolfert, and G. E. Henter, "A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020," in *26th international conference on intelligent user interfaces*, 2021, pp. 11–21.

[17] W. Zhao, L. Hu, and S. Zhang, "Diffugesture: Generating human gesture from two-person dialogue with diffusion models," in *Companion Publication of the 25th International Conference on Multimodal Interaction*, 2023, pp. 179–185.

[18] F. Zhang, N. Ji, F. Gao, and Y. Li, "Diffmotion: Speech-driven gesture synthesis using denoising diffusion model," in *International Conference on Multimedia Modeling*. Springer, 2023, pp. 231–242.

[19] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[20] G. Ferré, "Timing relationships between speech and co-verbal gestures in spontaneous french," in *Language Resources and Evaluation, Workshop on Multimodal Corpora*, vol. 6, 2010, pp. 86–91.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[22] T. Kucherenko, P. Jonell, S. Van Waveren, G. E. Henter, S. Alexandersson, I. Leite, and H. Kjellström, "Gesticulator: A framework for semantically-aware speech-driven gesture generation," in *Proceedings of the 2020 international conference on multimodal interaction*, 2020, pp. 242–250.

[23] K. Zhou, Z. Cheng, H. P. Shum, F. W. Li, and X. Liang, "Stgae: Spatial-temporal graph auto-encoder for hand motion denoising," in *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2021, pp. 41–49.

[24] K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *EMNLP*, 2014.

[25] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt, and B. Zheng, "Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis," in *European Conference on Computer Vision*. Springer, 2022, pp. 612–630.

[26] P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 492–518.

[27] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[28] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, and G. Lee, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–16, 2020.