Multi-Modal Rhythmic Generative Model for Chinese Cued Speech Gestures Generation

1st Li Liu*
HKUST(GZ)
Guangzhou, China

2nd Wentao Lei*
HKUST(GZ)
Guangzhou, China

3rd Wenwu Wang
University of Surrey
Guildford, England

***Abstract*—Cued speech (CS) is a novel visual coding system, which combines lip reading with several specific hand codings to help hearing-impaired people to communicate effectively. This work focuses on the audio/text-driven CS gestures (*i.e.*, continuous lip and hand gestures movements) generation. Previous work used template-based statistical methods for the French CS generation. However, these methods are fragile since they need careful hand-crafted pre-processing to fit models, resulting in poor robustness. Furthermore, the natural rhythm in generated CS gesture sequences, which is essential for a coding system of spoken languages, was overlooked in prior studies. To solve the above-mentioned problems, we innovatively propose a two-branched rhythmic CS gesture generation framework, which contains a multi-modal adversarial semantic generator (MASG) to generate accurate multi-modal CS gestures (*i.e.*, lip, hand shape and hand position movements), and an audio-driven rhythm generator (ARG) to extract the rhythm information. Moreover, we design a new Gesture Audio Difference (GAD) metric to evaluate the rhythm coherence considering the issue of asynchrony between CS hand gestures and lip movements. Extensive experimental results are presented on two datasets of two tasks (a CS dataset named MCCS-2024 and a co-speech TED dataset) with comprehensive ablation analysis and user study, demonstrating the effectiveness of our method. The code and dataset with multi-modal annotations were made public at https://mccs-2024.github.io/.**

***Index Terms*—Cued Speech, Speech Gesture Generation, Multi-modality, Rhythm**

I. Introduction

To tackle the disadvantage of lip reading and enhance the reading ability of hearing-impaired people, in 1967, Cornett developed the **Cued Speech (CS)** system [1], which uses hand codings to complement lip reading by providing clear visibility of all phonemes in a spoken language [2], [3]. For example, in Mandarin Chinese CS (MCCS) [4] (see Fig. 1), it uses five hand positions to encode vowel groups and eight hand shapes to encode consonant groups. With CS, hearing-impaired individuals can distinguish sounds that may appear identical on lips, such as [u], [y], by utilizing hand information. Therefore, they can understand spoken languages using solely visual information. Another commonly used communication method is Sign Language (SL) [5]. It is essential to note that CS is not a visual language like SL but a spoken speech coding system. Therefore, it can be learned much quicker than SL, according to studies [6], [7].

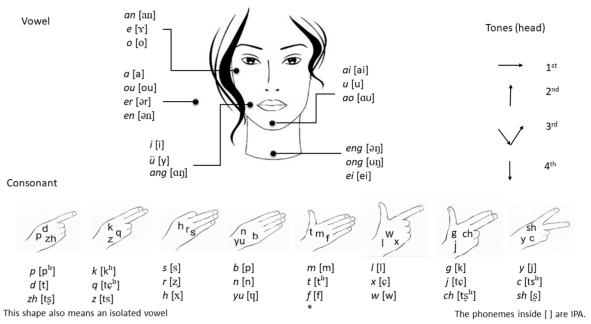

Fig. 1. The chart for the Mandarin Chinese CS (figure from [8]).

While there are numerous research efforts in CS Recognition [9], [10], as far as we know, CS gesture generation is under-explored because of the limited size of CS datasets and the expensive annotation cost of complicated CS multi-modal gestures. In addition, CS gesture is a fine-grained gesture generation task (*e.g.,* as shown in Fig. 1, CS hand position "mouth" and "chin" represent different vowels, but their location is very close), making the task challenging.

Previous work [11], [12] used template-based statistical methods for the French CS generation. However, these methods are fragile since they need careful hand-crafted pre-processing to fit the algorithm, and thus the robustness is poor. Deep Learning (DL)-based methods for CS gesture generation still suffer from the limited availability of the data and multi-modal annotations, leading to low accuracy of the generated CS gestures. Existing research on gesture generation, such as SL [13] and co-speech [14] generation, cannot fully meet the needs of our CS gesture generation.

As a coding system for spoken languages, CS requires natural rhythm dynamics to ensure complete semantic expression [15]. This involves generating multi-modal CS speech gestures—specifically hand shapes and positions that match the rhythm of speech. Unlike previous works, it's crucial to account for the lip-hand asynchrony phenomenon [16], where hand movements precede lip movements. This asynchrony varies depending on the cuer[1] [16]–[18] and must be reflected in the generated CS gestures to align lip and hand movements with human speech rhythm accurately.

* Equal Contribution.
Corresponding Author: Li Liu, avrillliu@hkust-gz.edu.cn.

[1]People who perform CS are called the cuer.

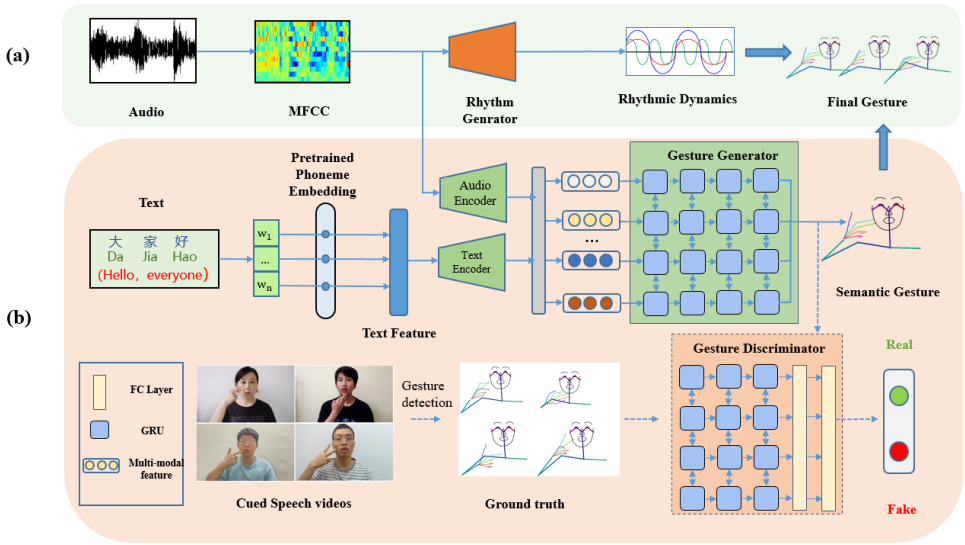

Fig. 2. The framework of our end-to-end multi-modal rhythmic CS hand gesture generation framework. The whole framework consists of two branches: (a) is a rhythmic gesture branch (*i.e.*, ARG) that learns the mapping from speech audio to rhythmic dynamics, and (b) is a semantic gesture generation branch (*i.e.*, MASG) that is constructed as a multi-modal generator based on semantic signals in audio speech and text to generate CS gestures. The content in the dashed line means the discriminator, which is only used in the training process.

To overcome the two above-mentioned challenges (*i.e.*, **low accuracy** and **rhythm scarcity** of the generated CS gestures), this work aims to propose the first end-to-end deep learning-based method for multi-modal CS hand gesture generation. We propose a novel two-branch framework for generating rhythmic CS hand gestures. An overview of the framework can be seen in Fig. 2. It contains a **M**ulti-modal **A**dversarial **S**emantic **G**enerator (**MASG**) that produces accurate multi-modal CS hand gestures, and an **A**udio-driven **R**hythmic **G**enerator **(ARG)** that considers the overall rhythm of the motion to extract rhythm information. Moreover, to better evaluate the rhythmic effect by considering the temporal asynchrony between CS hand gestures and lip reading, we further design a new Gesture Audio Difference (GAD) metric to evaluate the rhythm coherence. It should be noted that the proposed GAD can also be used to measure other multi-modal speech gesture asynchrony problems, such as the alignment between audio speech and co-speech gestures.

It should be noted that, in this work, we propose a GAN-based method [19] rather than the diffusion-based method for gesture generation because diffusion models [20], [21] usually rely heavily on a strong encoder [22], with most works utilizing text encoders like CLIP [23]. However, the gesture-text CS dataset is not enough to train a strong CLIP model for our task. Using the existing pre-trained CLIP may suffer from the domain gap (*i.e.*, the image-text mapping for CS is image-phoneme, which is different from the image-text mapping of the existing CLIP). However, the GAN-based methods can generate a realistic gesture without needing a large amount of data, thus making them more suitable for the CS generation.

In summary, the main contributions of this work can be summarized as follows. **1)** We design a novel multi-modal rhythmic gesture generation framework, which consists of two branches for accurate and rhythmic gesture generation. **2)** Taking into account the temporal asynchrony between CS hand gestures and lip reading, we specifically propose a new metric GAD to evaluate the rhythmic coherence and asynchrony of the generated multi-modal CS gestures. **3)** Extensive experiments are conducted on two datasets (including MCCS-2024 and one general co-speech TED Gesture dataset). Results show that the proposed method achieves state-of-the-art (SOTA) performance on both datasets under different metrics. Ablation studies and user studies further verify the effectiveness of the proposed model.

II. Method

A. *Multi-modal Adversarial Semantic Generator*

1) Text Feature Extraction: We first decompose the text as a sequence of phonemes, where the number of phonemes corresponds to the frame rate of the video. This is the same length as the video, which contains lip and hand movements. The word embedding layer converts all word sequences into 300-dimensional word vectors. The Chinese word embedding layer is trained based on the *Chinese Wikipedia* corpus[2] containing over 9 million sentences, and it is derived from the well-known *FastText* [24], which is a pre-trained word embedding that generates word vectors. Then, we use a temporal convolutional network (TCN) [25] to encode these word vectors for the text modality.

To process the audio modality, we use Mel-frequency cepstral coefficients (MFCC) [26] as the audio feature.

2) Semantic Gesture Generator: Every gesture is comprised of 137 key landmarks, including 70 for the face, 42 for the hands, and 25 for the body's posture. To maintain

[2]https://dumps.wikimedia.org/

consistency in the temporal duration, denoted as N, we synchronize the duration of the input with that of the output gesture. This synchronization allows our designed system to analyze and process data frame by frame, facilitating the CS generation.

Inspired by the non-saturating generative adversarial network (NS-GAN) [19], we introduce a novel approach utilizing a multi-layer bidirectional gated recurrent unit (GRU) network [27] as the foundational architecture for our gesture generation model. To train our gesture generator, we employ the following loss function L_G^{GAN}:

$$L_G^{\mathrm{GAN}} = -\mathbb{E}[\log(D(\hat{M}))], \tag{1}$$

where the discriminator D is used to differentiate between real and generated gestures. Specifically, our model architecture comprises a multilayer bidirectional GRU network, which yields a binary outcome for each individual frame, culminating in a fully connected (FC) layer that aggregates these binary results. For every frame i, we merge the encoded text and audio features into a single composite feature vector. The generator then processes the merged feature to produce the subsequent pose $\hat{M}_{i+1}$ sequentially.

The discriminator is trained using the loss function L_D:

$$L_D = -\mathbb{E}[\log(D(M))] - \mathbb{E}[\log(1 - D(\hat{M}))]. \tag{2}$$

Through the alternating optimization of the generator and discriminator, we aim to enhance the generator's ability to fool the discriminator. This iterative refinement process is designed to yield a gesture generator of superior quality.

B. *Audio-driven Rhythmic Gesture Generator*

In addition to the accurate position of the gesture, the natural temporal rhythm of the gesture motion is also an essential part of CS gesture generation. We believe the corresponding audio speech signal contains the rhythmic dynamics in CS, contributing to visual and auditory coherence. In this work, we introduce a new rhythmic gesture branch, which uses three convolution layers as the rhythmic dynamics generator, to match the dynamics with the CS rhythm further.

Loss for Rhythmic Dynamics Generator. The rhythmic motion branch learns the rhythmic dynamics $\widetilde{M}$ independent of the ground truth gesture (M). The corresponding loss function is defined as $L_r = ||\widetilde{M} - \left(M - \bar{M}\right)||$, where $\bar{M}$ is the arithmetic mean of motions in M. The difference between M and $\bar{M}$ measures the magnitude of hand movement. L_r ensures the generated result $\widetilde{M}$ (based on audio signal) with the proper offset to the mean gesture, which helps generate the motion dynamics without affecting the ground truth gesture.

Total Loss. We apply the reconstruction loss to the final gesture $M^* = \hat{M} + \widetilde{M} : L_{rec} = ||M^* - M||$. The total loss is as follows:

$$\mathcal{L} = \lambda_1 L_{\mathrm{G}}^{Huber} + \lambda_2 L_{\mathrm{G}}^{GAN} + \lambda_3 L_D + \lambda_4 L_r + \lambda_5 L_{\mathrm{rec}}\ , \tag{3}$$

where λ_j $(j = 1, \cdots, 5)$ are the weights for each loss. $L_G^{\mathrm{Huber}} = \frac{1}{N}\sum_{i=1}^{N} H\left(M_i, \hat{M}_i\right)$, where H is the Huber loss and N is the length of frames in a video.

III. Experiments

A. Datasets

In this work, we conducted experiments on the Mandarin Chinese CS dataset (MCCS-2024) [28], [29], which contains 4000 CS videos from four native Chinese CS cuers and another public Co-speech TED Gesture dataset [30]. The TED Gesture dataset is a collection of 2D and 3D upper-body gestures from English TED videos. The dataset includes 253,186 data samples, with 80% for training, 10% for validation, and the remaining 10% for testing.

B. Experimental Settings

The experiments are implemented using PyTorch, with four *A100* GPU cards for model training. We use Adam as our optimizer, which is set to $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and the learning rate was 0.0002. The model was trained for 100 epochs. At the same time, we experimentally adjusted the best hyperparameters ($\lambda_1 = 500$, $\lambda_2 = 5, \lambda_3 = 0.05, \lambda_4 = 0.5, \lambda_5 = 1.0$) for loss function in Eq. (3). To guarantee stable training, we set a 10 epochs warm-up period with ($\beta = 0$). In GAD, we set τ =200ms for MCCS-2024, and τ =500ms for TED and Trinity co-speech datasets.

C. Evaluation Metrics

The evaluation of the generated gestures is conducted quantitatively using four commonly used metrics, Percentage of Correct Keypoint (PCK) [31], Mean Absolute Joint Errors (MAJE) [32], Mean Acceleration Difference (MAD) [32], and Fréchet Gesture Distance (FGD) [32]. Especially, to consider the above-mentioned asynchronous coherence phenomenon for lip and hand movements in CS, a new metric to measure the rhythmic degree of generated gestures should be proposed.

Gesture Audio Difference Metric (GAD). We initially present the rationale behind introducing the proposed gesture audio difference metric. The existing metrics, such as Percentage of Matched Beats (PMB) [33], designed for measuring rhythm in co-speech, are not suitable for CS as they lack the consideration of audio-to-gesture correspondence. Another commonly used metric, GA, fails to capture the coherence between lip and hand movements, which significantly influences the rhythm of CS gestures. Specifically, in order to evaluate whether the generated CS gestures are rhythmic or not, **we propose a new metric GAD**:

$$\mathrm{GAD}(P, A) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[||\boldsymbol{C}_i^P - \boldsymbol{C}_i^A||_1 < \tau], \tag{4}$$

where the gesture and audio speech are denoted as P and A, respectively. As the number of annotated temporal segments of speech and gesture are the same, the number of segments is denoted as N. C_i is the middle time instant of the segment, which means a specific time at which a gesture or speech happens. $\mathbf{1}$ is the indicator function that maps elements of the subset (satisfies $||\boldsymbol{C}_i^P - \boldsymbol{C}_i^A||_1 < \tau$) to one, and all other elements to zero. Considering the asynchrony between audio speech and CS hand movement, we set a threshold τ to ensure

TABLE I
EXPERIMENT RESULTS ON MCCS-2024 DATASET COMPARED WITH SOTA METHODS. W/O TF AND W/O RD REPRESENT WITHOUT TEXT FEATURE AND WITHOUT RHYTHMIC DYNAMIC, RESPECTIVELY. ↑ MEANS THE HIGHER THE BETTER, ↓ MEANS THE LOWER THE BETTER.

Methods	PCK (%)↑	FGD↓	MAJE (mm)↓	MAD (mm/s^2)↓	GAD (%)↑
GES [37]	31.5	36.7	60.87	2.87	67.62
GTC [32]	33.2	34.1	58.46	2.65	71.25
S2AG [36]	37.8	31.9	53.75	1.92	74.57
RG [33]	43.2	29.5	49.53	0.97	74.57
Ours (w/o TF)	36.1	36.9	65.04	1.53	71.45
Ours (w/o RD)	43.9	31.2	53.32	0.95	76.68
Ours	**45.7**	**28.3**	**37.21**	**0.67**	**84.68**

their alignment. The threshold τ is determined by a statistical study of hand preceding time [34], *i.e.*, the mean value of all time differences between the hand target instants and the acoustic instants for all phonemes in the dataset.

IV. RESULT AND ANALYSIS

Our approach is compared with five recent gesture synthesis methods: Style Gesture (SG) [35], Gestures from Trimodal Context (GTC) [32], S2AG [36], Gesticulator (Ges) [37], and Rhythmic Gesticulator (RG) [33].

A. Comparison with SOTA Methods

Results on MCCS-2024 Dataset. Table I summarizes the performance of various methods on the MCCS-2024 Dataset. Our proposed method demonstrates the lowest values for the PCK, MAJE, MAD, and FGD metrics. Notably, our method's FGD values are significantly lower compared to other methods, indicating that the gestures synthesized by our methods possess higher perceptual quality. The rapid increase in FGD values in the absence of the text feature (w/o TF) highlights the crucial role of the text feature and modality confusion in gesture quality. Although the audio-only inference leads to a decrease in generation performance, the generated gestures are still deemed acceptable. In terms of rhythm performance, our methods achieve the highest GAD values on MCCS-2024 datasets. The significant drop in GAD values without the rhythm branch (w/o RD) highlights the crucial role of the proposed rhythmic features.

Results on TED Datasets. To demonstrate the generalization capability of our method to other audio-visual tasks, we conduct experiments on the public TED Gesture dataset. As shown in Table II, our method achieves SOTA results on these two datasets as well. It is worth noting that the performance in different metrics significantly drop when the text feature and rhythmic branch are absent, which aligns with observations on the MCCS-2024 dataset.

B. User Study

In addition, we design a detailed user study to assess the generated CS gesture. We use three metrics, *i.e.*, Accuracy, Rhythm Quality (RQ), and Naturalness of the generated CS gestures. More precisely, Accuracy measures how closely the generated gestures align with the intended target or reference. RQ evaluates the consistency and smoothness of the generated

TABLE II
EXPERIMENT RESULTS ON THREE DATASETS COMPARED WITH SOTA METHODS ON TRINTY DATASETS. W/O TF AND W/O RD REPRESENT WITHOUT TEXT FEATURE AND WITHOUT RHYTHMIC DYNAMIC, RESPECTIVELY.

Methods	PCK (%)↑	FGD↓	MAJE (mm)↓	MAD (mm/s^2)↓	GAD (%)↑
S2AG [36]	43.4	3.73	26.95	3.03	75.57
RG [33]	51.7	2.04	18.13	2.29	89.54
Ours (w/o TF)	36.8	3.68	25.31	2.94	73.48
Ours (w/o RD)	53.2	2.04	16.32	2.58	88.32
Ours	**54.1**	**1.92**	**17.26**	**2.42**	**91.14**

gestures in relation to the rhythm of the accompanying audio. Naturalness assesses the degree to which the generated CS gestures resemble natural and human-like movements, avoiding any signs of artificiality or robotic behavior.

A total of six CS teachers completed the questionnaire. They were asked to rate each video in the questionnaire on a score of 1 (worst) to 10 (best) for the three metrics. To conduct the experiment, we randomly select 30 sets of data, each consisting of a ground truth CS gesture video (*i.e.*, gestures obtained by the Openpose detector based on original CS videos), an audio-driven CS gesture video generated using the baseline method, which corresponds to our approach but does not incorporate text features and the ARG branch, and our method (*i.e.*, audio- and text-driven two branched CS hand gesture generation method shown in Figure 2). We can see that the CS hand gesture generated by our method (green bars) outperforms the one generated by baseline (red bars) in three evaluation metrics and is close to the ground truth (blue bars).

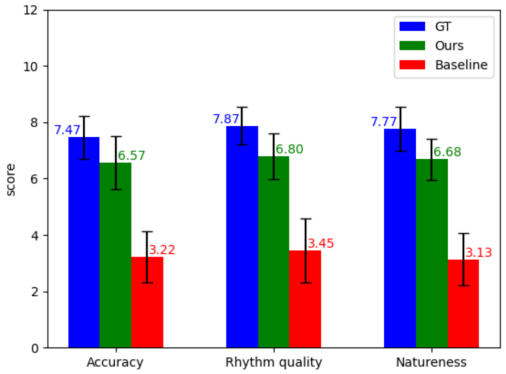

Fig. 3. User study results of the ground truth, baseline and Ours.

V. CONCLUSION

In this work, we propose a novel framework for generating semantic accurate and rhythmic Chinese CS hand gestures. This framework includes two branches specifically designed for generating hand movements in a multi-modal manner. Extensive experiments are conducted to validate the effectiveness of our proposed method. The results, both quantitative and qualitative, demonstrate that our method can generate high-quality CS hand gestures. In the future, we will explore the diffusion model for automatic CS gesture generation in the data limited scenario.

REFERENCES

[1] R. Orin Cornett, "Cued speech," *American Annals of the Deaf*, vol. 112, no. 1, pp. 3–13, 1967.

[2] N. Puviarasan and S. Palanivel, "Lip reading of hearing impaired persons using hmm," *Expert Systems with Applications*, vol. 38, no. 4, pp. 4477–4481, 2011.

[3] A. Fernandez-Lopez, O. Martínez, and M. Sukno, "Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database," in *FG*, 2017.

[4] Li Liu and Gang Feng, "A pilot study on mandarin chinese cued speech," *American Annals of the Deaf*, vol. 164, pp. 496–518, 2019.

[5] Jr. Stokoe, C. William, "Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf," *The Journal of Deaf Studies and Deaf Education*, vol. 10, no. 1, pp. 3–37, 2005.

[6] S. Reynolds, "An examination of cued speech as a tool for language, literacy, and bilingualism for children who are deaf or hard of hearing," *Independent Studies and Capstones. Paper 315.*, 2007.

[7] S. Baber, "Cued speech: Not just for the deaf anymore," in *Senior Honors Theses Projects*, 2007.

[8] Lei Liu and Li Liu, "Cross-modal mutual learning for cued speech recognition," *ICASSP*, 2023.

[9] P. Katerina and P. Gerasimos, "A fully convolutional sequence learning approach for cued speech recognition from videos," in *EUSIPCO*, 2021.

[10] Panikos Heracleous, Denis Beautemps, and Noureddine Aboutabit, "Cued speech automatic recognition in normal-hearing and deaf subjects," *Speech Communication*, vol. 52, no. 6, pp. 504–512, 2010.

[11] P. Duchnowski, Louis D. Braida, D. Lum, M. Sexton, Jean C. Krause, and S. Banthia, "Automatic generation of cued speech for the deaf: Status and outlook," in *AVSP*, 1998.

[12] G. Bailly, Yu Fang, F. Elisei, and D. Beautemps, "Retargeting cued speech hand gestures for different talking heads and speakers," in *AVSP*, 2008.

[13] V. Lucas, D. Amanda, and G. Xavier, "Diffusion models: A comprehensive survey of methods and applications," *arXiv:2012.10941*, 2020.

[14] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao, "Speech drives templates: Co-speech gesture synthesis with learned templates," *ICCV*, 2021.

[15] Li Liu, Lufei Gao, Wentao Lei, Fengji Ma, Xiaotian Lin, and Jinting Wang, "A survey on deep multi-modal learning for body language recognition and generation," *arXiv preprint arXiv:2308.08849*, 2023.

[16] Li Liu, Gang Feng, B. Denis, and Xiao-Ping Zhang, "Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 292–305, 2021.

[17] A. Virginie, B. Denis, C. Marie-Agnès, and O. Matthias, "A pilot study of temporal organization in cued speech production of french syllables: rules for a cued speech synthesizer," *Speech Communication*, vol. 44, no. 1, pp. 197–214, 2004.

[18] A. Virginie, C. Marie-Agnès, and B. Denis, "Temporal measures of hand and speech coordination during french cued speech production," in *International Gesture Workshop*, 2005.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.

[21] Wentao Lei, Jinting Wang, Fengji Ma, Guanjie Huang, and Li Liu, "A comprehensive survey on human video generation: Challenges, methods, and insights," *arXiv preprint arXiv:2407.08428*, 2024.

[22] K. Minguk, Z. Jun-Yan, Z. Richard, P. Jaesik, S. Eli, P. Sylvain, and P. Taesung, "Scaling up gans for text-to-image synthesis," in *CVPR*, 2023.

[23] C. Hallacy A. Ramesh G. Goh S. Agarwal G. Sastry A. Askell P. Mishkin J. Clark G. Krueger A. Radford, Jong W. Kim and I. Sutskever, "Learning transferable visual models from natural language supervision," *ICML*, 2021.

[24] B. Piotr, G. Edouard, J. Armand, and M. Tomas, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2016.

[25] Shaojie Bai, J. Zico Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv:1705.07215*, 2018.

[26] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition," *Speech Communication*, vol. 54, pp. 543–565, 2012.

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

[28] Wentao Lei, Li Liu, and Jun Wang, "Bridge to non-barrier communication: Gloss-prompted fine-grained cued speech gesture generation with diffusion model," *IJCAI*, 2024.

[29] Lei Liu, Li Liu, and Haizhou Li, "Computation and parameter efficient multi-modal fusion transformer for cued speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[30] F. Ylva and M. Rachel, "Investigating the use of recurrent motion modelling for speech gesture generation," in *IVA*, 2018.

[31] Y. Yi and R. Deva, "Articulated human detection with flexible mixtures of parts," *TPAMI*, vol. 35, no. 12, pp. 2878–2890, 2013.

[32] Y. Youngwoo, C. Bok, L. Joo-Haeng, J. Minsu, L. Jaeyeon, K. Jaehong, and L. Geehyuk, "Speech gesture generation from the trimodal context of text, audio, and speaker identity," *ACM Transactions on Graphics*, vol. 39, no. 6, pp. 1–16, 2020.

[33] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu, "Rhythmic gesticulator," *ACM Transactions on Graphics*, vol. 41, no. 6, pp. 1–19, 2022.

[34] Li Liu, Gang Feng, B. Denis, and Xiao-Ping Zhang, "Re-synchronization using the hand preceding model for multi-modal fusion in automatic continuous cued speech recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 292–305, 2020.

[35] A. Simon, H. Gustav Eje, K. Taras, and B. Jonas, "Style-controllable speech-driven gesture synthesis using normalising flows," *Computer Graphics Forum*, vol. 39, no. 2, pp. 487–496, 2020.

[36] U. Bhattacharya, E. Childs, N. Rewkowski, and D. Manocha, "Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning," in *ACM MM*, 2021.

[37] T. Kucherenko, P. Jonell, S. van Waveren, G. Eje Henter, S. Alexandersson, I. Leite, and H. Kjellström, "Gesticulator: A framework for semantically-aware speech-driven gesture generation," in *ICMI*, 2020.