Bridge to Non-Barrier Communication: Gloss-Prompted Fine-grained Cued Speech Gesture Generation with Diffusion Model

Wentao Lei[1,2*], **Li Liu**[1,3†], **Jun Wang**[2]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Tencent AI Lab
[3]The Hong Kong University of Science and Technology

Abstract

Cued Speech (CS) is an advanced visual phonetic encoding system that integrates lip reading with hand codings, enabling people with hearing impairments to communicate efficiently. CS video generation aims to produce specific lip and gesture movements of CS from audio or text inputs. The main challenge is that given limited CS data, we strive to simultaneously generate fine-grained hand and finger movements, as well as lip movements, meanwhile the two kinds of movements need to be asynchronously aligned. Existing CS generation methods are fragile and prone to poor performance due to template-based statistical models and careful hand-crafted pre-processing to fit the models. Therefore, we propose a novel **Gloss**-prompted **Diff**usion-based CS Gesture generation framework (called **GlossDiff**). Specifically, to integrate additional linguistic rules knowledge into the model. we first introduce a bridging instruction called **Gloss**, which is an automatically generated descriptive text to establish a direct and more delicate semantic connection between spoken language and CS gestures. Moreover, we first suggest rhythm is an important paralinguistic feature for CS to improve the communication efficacy. Therefore, we propose a novel Audio-driven Rhythmic Module (ARM) to learn rhythm that matches audio speech. Moreover, in this work, we design, record, and publish the first Chinese CS dataset with four CS cuers. Extensive experiments demonstrate that our method quantitatively and qualitatively outperforms current state-of-the-art (SOTA) methods. We release the code and data at https://glossdiff.github.io/.

1 Introduction

According to the World Health Organization (WHO), more than 5% of the global population (466 million) suffers from the hearing loss. As a predominant communication method for hearing-impaired people, Lip reading [Puviarasan and Palanivel, 2011; Fernandez-Lopez *et al.*, 2017] has a major defect of visual confusion. For instance, it struggles to differentiate pronunciations with similar labial shapes, such as [u] and [y], posing challenges for hearing-impaired individuals in accessing spoken language through conventional education.

To tackle the limitations of lip reading, and to improve the reading skills of individuals with hearing impairments, in 1967, Cornett introduced the **Cued Speech (CS)** system [Cornett, 1967], which employs several hand codings ($i.e.$, finger shapes and hand positions) to complement lip reading, providing a clear visual representation of all phonemes in spoken language [Puviarasan and Palanivel, 2011; Fernandez-Lopez *et al.*, 2017]. For instance, in Mandarin Chinese CS (MCCS) [Liu and Feng, 2019] (see Fig. 1(a)), it utilizes five hand positions for encoding vowel groups and eight finger shapes for encoding consonant groups. With CS, individuals with hearing impairments can differentiate sounds that might appear similar when observed on lips by incorporating hand information. Another widely adopted communication method is Sign Language (SL) [Stokoe, 2005; Liddell and Johnson, 1989; Timothy, 2003]. It is crucial to emphasize that CS is not a visual language like SL; instead, it is a coding system of spoken language [Cornett, 1967]. In addition, studies indicate that CS can be learned much more quickly than SL [Reynolds, 2007]. Given that CS can effectively promote non-barrier communication, audio/text to CS gestures video generation draws researchers' attention. It should be noted that comparing to text, CS is more friendly and more easily adopted by the hearing impaired who are illiterate [Cox *et al.*, 2002; Power *et al.*, 2007].

The multi-modal CS gesture generation is a challenging task for the following reasons: 1) high requirement for fine-grained and accurate gesture generation, as shown in Figure 1(a), where nuances in the hand's position and fingers' shape lead to quite different semantic meanings; 2) the limited size of CS datasets and expensive annotation cost of complicated fine-grained CS gestures. To address these challenges, we design a novel **Gloss**-Prompted **Diff**usion-based CS Gesture generation framework (**GlossDiff**). Specifically, we first propose a CS gloss, which is a direct motion instruction for bridging the gap between spoken language and CS gestures. It is automatically generated by LLM based on the encoding rule of CS in Figure 1(a). As shown in Figure 1(b), when ex-

*Work done during internship at Tencent AI Lab.
†Corresponding Author: avrillliu@hkust-gz.edu.cn.

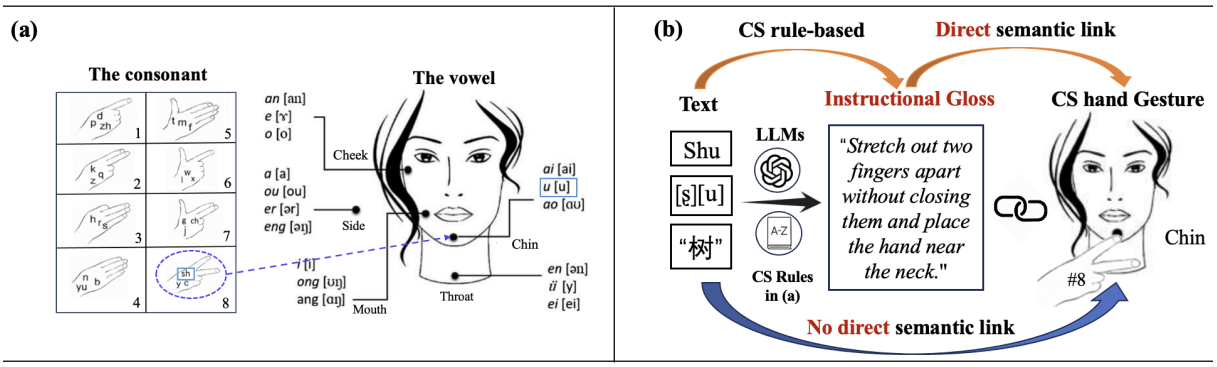

Figure 1: The details of CS rules and conversion process. (a) is the chart for the Mandarin Chinese Cued Speech (figure from [3]), where five different hand positions are used to code vowels, and eight finger shapes are used to code consonants in Mandarin Chinese. (b) shows the proposed instructional gloss, which directly links the text to the CS movements.

pressing the word "tree", which is pronounced as "/ʂ/ /u/" in Chinese, we generate the intermediate instruction text ($i.e.$, gloss) to describe the process of using CS gestures to express this word, $i.e.$, *Stretch out two fingers apart without closing them and place the hand near the neck.* Besides, we design a Gloss-Prompted Diffusion Model that can generate accurate hand and finger movements.

Moreover, rhythm is a critical paralinguistic information in spoken language. As a coding system for spoken languages, we suggest natural rhythm dynamics should also be considered as a very important feature for CS's complete semantic expression. More specifically, the rhythm here refers to the ability to generate multi-modal CS speech gesture movements ($i.e.$, hand and finger movements), which match the phoneme durations and utterance prosody of speech. Unfortunately, previous works have not pay enough attention to this. To this end, we propose an **A**udio-driven **R**hythmic **M**odule (**ARM**) that considers the overall rhythm of the CS movements aligning with speech signals. We leverage the large-scale WavLM [Chen *et al.*, 2022] to extract audio features, which we demonstrate outperforming traditional MFCC features.

We summarized our contributions as follows: **1)** A novel GlossDiff framework that simultaneously generates fine-grained hand position, finger movements, and lip reading in CS. Specifically, we introduce a CS **gloss**, which establishes a direct link between text/audio and CS hand movements, enabling more specific prompts for an accurate fine-grained CS gesture generation. **2)** A new module **ARM** that improves the overall rhythm of the CS movements. **3)** Publication of the first multi-cuer large-scale Mandarin Chinese CS (**MCCS**) dataset, which contains four cuers[1] and 4000 CS videos. **4)** Extensive experiments conducted on the MCCS dataset show the proposed GlossDiff achieves SOTA performance under different metrics. The qualitative and ablation studies, as well as user studies, further verify the effectiveness of the proposed model.

[1]The people who perform CS are called the cuer

2 Related Work

2.1 Cued Speech Generation

In prior work, early attempts at CS gesture generation [Duchnowski *et al.*, 1998; Bailly *et al.*, 2008] are mainly rule-based. Notably, in [Duchnowski *et al.*, 1998], specific keywords were manually selected, along with low-context sentences [Rothauser, 1969], and manual templates for corresponding hand gestures were predefined. This processing involved CS recognition, followed by mapping the recognized text to the hand templates. However, this method relied heavily on hand-crafted designs, which constrained the expressiveness of CS gestures and increased the required manual effort. In [Bailly *et al.*, 2008], a post-processing algorithm was introduced to refine synthesized hand gestures, including adjustments for hand rotation and translation. However, this approach required prior human knowledge to adapt the algorithm to new images, resulting in limited robustness. To the best of our knowledge, there is still a gap in research about end-to-end deep learning-based CS gesture generation.

2.2 Co-speech and Sign Language Generation

The generation of Co-speech gestures involves generating body movements corresponding to audio input. Previous studies mainly developed large speech-gesture datasets to learn how speech audio maps to human skeletons using deep learning, as in [Ao *et al.*, 2022]. To make gestures more expressive, some methods use Generative Adversarial Networks (GANs) for more realistic results [Ginosar *et al.*, 2019; Youngwoo *et al.*, 2020]. Recently, diffusion models like DiffGesture [Zhu *et al.*, 2023], effectively links audio and gestures while keeping time consistency, allowing for high-quality Co-speech gestures. However, Co-speech gesture generation focuses on fluency and style rather than gesture fine-grained accuracy. Existing methods cannot generate accurate subtle CS hand gestures.

In the literature, there are several Sign Language (SL) generation methods: 1) The Neural Machine Translation approach from [Stoll *et al.*, 2020] sees SL generation as translation, using neural models to process SL text. 2) The Motion Graph method in [Stoll *et al.*, 2020] uses motion graphics to make a directed graph from motion capture data for

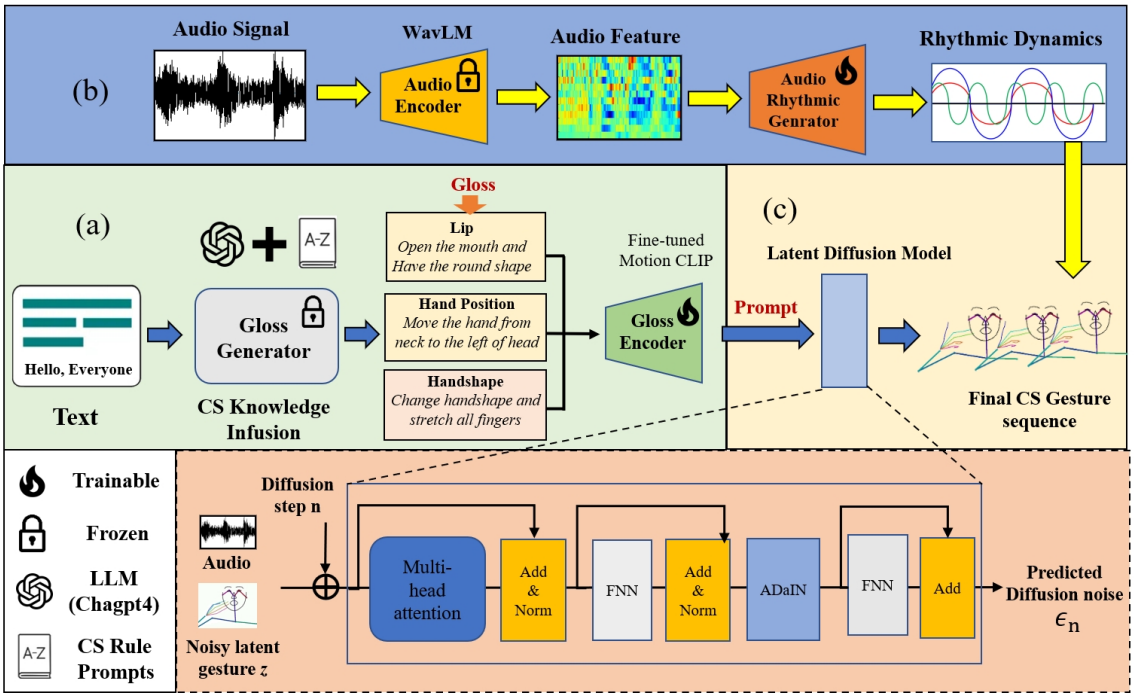

Figure 2: The overall framework of the proposed GlossDiff, where (a), (b), (c) represent the Knowledge Infusion Module, Audio Rhythmic Module and Diffusion-based generation module, respectively.

SL creation. 3) Conditional generation methods, like GANs and VAEs, are also used for SL gestures. 4) Some researches have introduced transformer-based models for SL, as mentioned in [Ben *et al.*, 2020]. Despite these advancements in SL gesture generation, applying these methods to CS gestures has limitations. Firstly, CS gesture generation necessitates more precise methods to achieve complex fine-grained gesture generation, while SL gesture generation is more coarse-grained. Secondly, SL gesture is not related with lip-reading, thus cannot match the speech rhythm and gesture-speech asynchrony characteristics [Liu *et al.*, 2021; Liu and Liu, 2023] in CS gesture generation.

2.3 NeRF and Diffusion-based Gesture Generation

NeRF is a novel technique in 3D modeling, which effectively creates highly detailed and photo-realistic static scenes from 2D images. Its application has been extended to generating life-like talking head models [Guo *et al.*, 2021], demonstrating NeRF's capability in handling subtle facial movements and expressions. However, the application of NeRF in full-body gesture generation is relatively limited. Additionally, the high requirements for data and computation further constrain its use in CS gesture generation.

Currently, in the field of human gesture generation, diffusion models [Ho *et al.*, 2020] are predominantly used in two main applications: generating comparably large body movements (*e.g.*, human walking) [Tevet *et al.*, 2022b; Zhang *et al.*, 2022; Zhao *et al.*, 2023; Ao *et al.*, 2023] and generating poses in Co-speech scenarios [Ji *et al.*, 2023; Zhi *et al.*, 2023; Yang *et al.*, 2023]. However, the existing approaches lack the capability to tackle fine-grained gesture generation. Additionally, they primarily focus on body poses without lip movements. Lastly, their diffusion models require extensive training data, which is not feasible given the limited dataset in our CS scenarios.

3 Method

In this section, we provide a comprehensive description of our proposed method, GlossDiff, designed for rhythm-aware CS gesture generation, which seamlessly integrates domain-specific knowledge for CS generation. As shown in Figure 2, our GlossDiff framework consists of three primary components: the knowledge infusion module, the rhythmic module, and the Diffusion-based generation module.

3.1 Problem Formulation

Automatic CS gesture generation involves generating the corresponding landmarks sequence of CS gesture M^*, given an audio signal A and the text T. In the task of automatic multimodal CS gesture generation, the combined features of A, T, and the generated rhythmic information are input into the CS gesture generator. The final CS gestures (M^*) including lips, fingers, and hand positions are obtained by minimizing:

$$\sum_{i=1}^{L} ||M_i^* - M_i||, \tag{1}$$

where L represents the frame count of the current CS video. The ground truth CS gesture landmarks M_i in the i-th frame of the CS video is obtained by the *Expose* method [Choutas *et*

al., 2020]. $M_i^* = \hat{M}_i + \widetilde{M}_i$, where $\hat{M} = \mathrm{G_D}(T, A)$ are generated semantic gesture landmarks representing the corresponding generated gesture in the i-th frame. $\mathrm{G_D}$ is the diffusion-based semantic gesture generator. Additionally, $\widetilde{M} = \mathrm{G_R}(A)$ is the rhythmic information derived from the correspondinng audio speech, with $\mathrm{G_R}$ as the rhythm generator.

3.2 Knowledge Infusion Module

The primary objective of the knowledge infusion module is to transform spoken language text T ($i.e.$, the speech transcription) into direct text instructions ($i.e.$, gloss, see Figure 1(b)), which describe the corresponding fine-grained CS motions. To achieve this, we leverage the LLM, $i.e.$, ChatGPT4 [OpenAI and et.al, 2023], the prompt engineering approach to infuse the encoding rules of Chinese CS [Liu and Feng, 2019] into our framework by the following:

$$g = \mathrm{LLM}(T, P), \tag{2}$$

where P is our designed prompt based on CS domain knowledge ($i.e.$, prior transformation rules of CS based on [Liu and Feng, 2019]), and T is the input text. Ultimately, this process enables the transformation of our indirectly semantic-related text into directly semantic-related gloss.

3.3 Diffusion-based Generation Module

Gloss-based Motion CLIP Fine-tuning

MotionCLIP [Tevet *et al.*, 2022a] is a multimodal large-scale model specifically designed for generating general motion gestures. To obtain an accurate feature embedding of CS gloss, we leverage the MotionCLIP as our pre-trained model, and fine-tune it using the generated CS gloss (introduced in Subsection 3.2) and the paired CS gestures.

As for the fine-tuning stage, we adopt CLIP-style contrastive learning [Radford *et al.*, 2021] to fine-tune the encoders with CS data. Given a batch of pairs containing CS gesture motion and gloss embeddings, denoted as $\mathcal{B} = \{(z_i^m, z_i^g)\}_{i=1}^B$, where B is the batch size. $\mathcal{E}_m$ and $\mathcal{E}_g$ are the corresponding MotionCLIP encoder for both motion sequence and gloss. $Z^m = \mathcal{E}_m(M), \quad Z^g = \mathcal{E}_g(g)$. The goal of the training is to maximize the similarity between paired z_i^m and z_i^g of in the batch while minimizing the similarity of the incorrect pairs $\left(z_i^m, z_j^g\right)_{i \neq j}$. A symmetric cross entropy (CE) loss L_{CE} is optimized over these similarity scores. Formally, the loss is:

$$\begin{aligned} \mathcal{L}_{\text{CLIP}} = \mathbb{E}_{\mathcal{B}\sim\mathcal{D}}[&L_{CE}\left(\boldsymbol{y}\left(z_i^m\right), \boldsymbol{p}_m\left(z_i^m\right)\right) \\ &+ L_{CE}\left(\boldsymbol{y}\left(\boldsymbol{z}_j^g\right), \boldsymbol{p}_g\left(z_j^g\right)\right)], \end{aligned} \tag{3}$$

where $\boldsymbol{y}$ specifies the true correspondence between the gestures z_i^m and gloss z_j^g in the training batch $\mathcal{B}$. If they are paired, $\boldsymbol{y} = 1$, otherwise, $\boldsymbol{y} = 0$. $\boldsymbol{p}$ is defined as:

$$\boldsymbol{p}_m\left(z_i^m\right) = \frac{\exp\left(z_i^m \cdot z_i^g / \eta\right)}{\sum_{j=1}^B \exp\left(z_i^m \cdot z_j^g / \eta\right)}, \tag{4}$$

where η is the temperature of softmax, and $\boldsymbol{p}_g\left(z_j^g\right)$ follow the same computations.

Gloss-Prompted Diffusion Model

To generate CS gesture video, we propose a Gloss-Prompted Diffusion Model. More precisely, the semantic hand gesture generator $\mathrm{G_D}$ is designed based on the latent diffusion model [Rombach *et al.*, 2022], which applys diffusion and denoising steps in a pre-trained latent space. The latent diffusion model is trained with the standard noise estimation loss [Ho *et al.*, 2020] defined as:

$$\mathcal{L}_{\text{noise}} = ||\epsilon - \epsilon_\theta\left(Z_n, n, g, A\right)||_2^2, \tag{5}$$

where Z_n is the latent CS gesture at each time step n. A is audio speech, and g is the generated gloss. ϵ is the ground truth noise and ϵ_θ is the noise predicted by latent diffusion model, where θ is the parameters of latent diffusion model.

To inject the information of the gloss prompts into the diffusion network, we employ an adaptive instance normalization (AdaIN) layer [Huang and Belongie, 2017]. Specifically, we leverage the fine-tuned MotionCLIP gloss encoder $\mathcal{E}_g$ to convert the gloss prompt into a gloss embedding z^g. Then, we learn a MLP network to map the gloss embedding z^g to parameters that modify the per-channel mean and variance of the AdaIn layer.

To train our Gloss-Prompted Diffusion Model, we employ classifier-free guidance as detailed in [Ho and Salimans, 2022]. Specifically, during training, we enable the diffusion model $\boldsymbol{G}_D$ to master both the semantic conditional and unconditional distributions by randomly configuring $\boldsymbol{g} = \varnothing$. This action effectively deactivates the AdaIN layer with a probability of p during the training phase, which is set to 10% [Tevet *et al.*, 2022b]. During inference, the anticipated noise is calculated using:

$$\epsilon_n^* = p\epsilon_\theta\left(Z_n, n, g, A\right) + (1-p)\epsilon_\theta\left(Z_n, n, \varnothing, A\right). \tag{6}$$

After obtaining the predicted noise ϵ_n^*, the model operates in a reverse step-wise manner over N time steps, updating a latent gesture sequence Z_n at each time step n. It begins by generating a sequence of latent codes $Z_N \sim \mathcal{N}(0, I)$ and subsequently calculates a series of denoised sequences Z_n through the iterative removal of the estimated noise ϵ_n^* from Z_n $(n = N-1, \ldots, 0)$. Z_0 is the final generated CS gesture latent embedding through N reverse diffusion steps. Z_0 is fed into a Transformer-based decoder [Petrovich *et al.*, 2021] to generate semantic CS gesture motion $\hat{M}$.

3.4 Audio-driven Rhythmic Module

In CS gesture generation, it's not just the accurate positioning of the gesture that matters; the natural rhythm of gesture motion plays a crucial role. We believe that the audio speech signal contains not only the semantic information but also the rhythmic dynamics of CS, which significantly contributes to achieving visual and auditory coherence.

To address this, we introduce a novel Audio-driven Rhythmic Module (ARM), designed to capture the rhythmic dynamics of gestures. This module employs three convolution layers as a rhythmic dynamics generator $\mathrm{G_R}$, further aligning the motion dynamics with the CS rhythm.

Existing research ($e.g.$, WavLm and AudioLDM) [Lebourdais *et al.*, 2022; Liu *et al.*, 2023] have shown that compared with MFCC features, audio features extracted by the large

pre-trained model have a stronger expressive capability and can avoid information loss. Without loss of generality, in this work, we use the encoder of WavLM, denoted as $\mathcal{E}_A$ to extract audio features to prevent information loss, thereby preserving richer and higher-dimensional rhythmic information.

To handle the lip-hand synchronization issue [Liu *et al.*, 2021] in CS, we reformulate the task as one of determining the motion magnitude for each frame within consecutive motion sequences. Unlike methods that attempt to enforce perfect alignment between generated gestures and speech, our approach implicitly learns how to produce asynchronous gestures that correspond to the input speech. Rather than directly controlling the gestures of each individual frame, we focus on regulating the overall rhythm of a motion sequence.

The loss function for the ARM is defined as:

$$\mathcal{L}_{\text{rhythm}} = ||\widetilde{M} - \left(M - \bar{M}\right)||, \tag{7}$$

where $\bar{M}$ represents the average motion within the set of generated motions M. The difference between M and $\bar{M}$ quantifies the magnitude of hand and finger movement. The purpose of $\mathcal{L}_{\text{rhythm}}$ is to ensure that the generated $\widetilde{M} = \mathrm{G_R}(\mathcal{E}_\mathrm{A}(\mathrm{A}))$ maintains the natural offset relative to the mean gesture. $\mathcal{E}_A$ is the encoder of WavLM. This offset helps in generating motion dynamics for a natural, non-mechanical movement without disrupting the semantics of the CS gesture. We demonstrate the efficacy regarding rhythm quality and naturalness with quantitative result in Sec. 4.3, as well as qualitative result of in Sec. 4.4.

Novel Quantitative Rhythmic Metrics

In this work, for the first time, rhythm is investigated as an important paralinguistic feature to improve CS' communication efficacy. To capture the unique asynchronous dynamics between lip and hand movements in CS scenarios, we propose a novel metric, Gesture Audio Difference (GAD), to evaluate the rhythmic synchronization of the generated gestures. This metric is defined as follows:

$$\mathrm{GAD}(M, A) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[||\boldsymbol{U}_i^M - \boldsymbol{U}_i^A||_1 < \tau], \tag{8}$$

where M and A represent the CS gesture and audio speech, respectively. The term N denotes the number of annotated temporal segments, which are equal for both speech and gesture. The variable $\boldsymbol{U}_i$ refers to the middle time instant of a segment, indicating a specific moment when a gesture or speech occurs. The function $\mathbf{1}$ is an indicator function, mapping elements within the subset (satisfying $||\boldsymbol{U}_i^M - \boldsymbol{U}_i^A||_1 < \tau$) to one, and all other elements to zero.

Taking the asynchrony between audio speech and CS hand movements into consideration, we introduce a threshold τ, which ensures their alignment and is empirically determined based on a statistical study of the hand preceding time [Liu *et al.*, 2020].

3.5 Training of GlossDiff Framework

We employ a semantic loss to ascertain the semantic accuracy of the final generated gestures. To be specific,

$$\mathcal{L}_{\text{semantic}} = 1 - \cos\left(Z_0, Z_0^*\right), \tag{9}$$

where $\cos(\cdot,\cdot)$ represents the cosine distance, while Z_0 and Z_0^* denote the final generated CS gesture latent embedding and the ground truth CS gesture motions, respectively.

Following the existing training procedure for denoising diffusion models, we optimize the following loss:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{noise}} + \beta\mathcal{L}_{\text{semantic}} + \gamma\mathcal{L}_{\text{rhythm}}, \tag{10}$$

where α is the weight of $\mathcal{L}_{\text{noise}}$ (in Equation (5)), β is the weight of $\mathcal{L}_{\text{semantic}}$ (in Equation (9)), and γ is the weight of $\mathcal{L}_{\text{rhythm}}$ (in Equation (7)).

4 Experiments

4.1 MCCS Datatset

Previously, only two CS datasets were available for public access: One was in French[2] [Liu *et al.*, 2018], consisting of recordings of a single cuer delivering 238 sentences; The other was in British English[3] [Liu *et al.*, 2019], similarly featuring a single cuer reciting 97 sentences. To remedy the scarcity of Chinese CS data, we built in this work, for the first time, a large-scale **M**andarin **C**hinese **CS** dataset that includes contributions from four CS cuers, called **MCCS**.

We first select 1000 text sentences following the below principles: (1) They cover common scenarios in daily life, including colloquial dialogues, more formal words, as well as written words. (2) The materials aim to cover possible syllable combinations. All in all, our text album covers 23 main topics, 72 subtopics, and the most commonly used 399 Mandarin syllables. It comprises a total of 1000 sentences, 10,482 words with an average of 10.5 words per sentence. The shortest sentence contains 4 words, while the longest has 25 words. Then, we recorded CS videos for each of the four cuers performing the 1000 sentences, resulting 4000 sentences in total.

All videos are recorded using either a camera or a mobile phone in landscape mode, The four cuers have received systematic training to ensure they can perform Mandarin Chinese CS smoothly and accurately. Note that our dataset has been collected with the explicit consent of the individuals involved and is eligible for open source.

4.2 Experimental Setup

During the training phase, we pre-train the motion clip first and then follow an end-to-end pipeline to train the latent diffusion model. The experiments are implemented using PyTorch, with four *A6000* GPU cards for model training. During the inference phase, we use the latent diffusion model to generate CS gestures. The training and test data are randomly split as $4:1$. The number of diffusion steps is 1000, and the training batch size is 128. The weight of loss items is set to $\alpha = 1$, $\beta = 0.2$ and γ = 0.1.

Evaluation Metrics

The conventional evaluation metrics of the generated gestures contain three classes: Percentage of Correct Keypoint (PCK) [Yi and Deva, 2013], Fréchet Gesture Distance (FGD) [Youngwoo *et al.*, 2020], Mean Absolute Joint Errors

[2]https://zenodo.org/record/5554849#.ZBBCvOxBx8Y

[3]https://zenodo.org/record/3464212#.ZBBAJuxBx8Y

Methods	PCK (%)↑	FGD↓	MAJE (mm)↓	MAD (mm/s^2)↓	GAD (%)↑
Speech2Gesture [Ginosar *et al.*, 2019]	36.84	19.25	61.26	3.97	66.8
GTC [Youngwoo *et al.*, 2020]	41.23	6.73	55.43	2.54	66.7
HA2G [Liu *et al.*, 2022]	43.51	4.07	46.78	2.29	67.2
DiffGesture [Zhu *et al.*, 2023]	47.58	**3.50**	48.52	2.12	69.9
Our GlossDiff (w/o Gloss-prompt)	51.12	4.72	45.68	1.28	75.6
Our GlossDiff (w/o WavLM)	52.97	4.54	42.31	0.71	78.3
Our GlossDiff (w/o Gloss-CLIP)	53.41	4.31	43.52	0.65	79.1
Our GlossDiff	**54.23**	3.92	**39.28**	**0.52**	**79.4**

Table 1: Experiment results on MCCS Dataset compared with SOTA methods. "Gloss-Prompt" indicates the integration of a Gloss Knowledge Infusion Module. The term "WavLM" refers to the substitution of MFCC features with features from the pre-trained large-scale speech model, wavLM. "Gloss-CLIP" denotes the incorporation of Gloss-based Motion CLIP Fine-tuning.

(MAJE) [Youngwoo *et al.*, 2020], and Mean Acceleration Difference (MAD) [Youngwoo *et al.*, 2020]. In addition, to further measure the unique asynchronous dynamics between lip and hand movements in CS scenarios, we use the novel metric GAD as described in Sec.3.4 to evaluate the rhythmic synchronization of the generated gestures.

4.3 Quantitative Result and Analysis

Comparison with SOTA

We compare our approach with four recent gesture synthesis methods, *i.e.*, Speech2Gesture [Ginosar *et al.*, 2019], Gestures from Trimodal Context (GTC) [Youngwoo *et al.*, 2020], HA2G [Bhattacharya *et al.*, 2021], DiffGesture [Zhu *et al.*, 2023]. We take DiffGesture as the SOTA method among these approaches, since it achieves the best result on the TED Gesture datasets [Youngwoo *et al.*, 2019].

Table 1 provides a detailed comparison among our methods and the previous methods on the MCCS datasets. Our method GlossDiff gives the best results in PCK, MAJE, MAD, and GAD metrics, most of which have a wide superiority leap comparing to the reference systems. The results demonstrate a higher quality of fine-grained gesture generation by our proposed system. The only exception is one FGD score that slightly trails the SOTA method, while it surpasses all other reference methods. Notably, our method's PCK values are significantly higher than other methods, showing its effectiveness in fine-grained generation. Moreover, our method excels in rhythm performance, achieving the highest GAD values. This superiority on GAD metrics demonstrates that our method can effectively capture the rhythm in CS gesture.

Ablation Study

We provide the ablation study for three modules in Table 1. The term "Gloss-prompt" indicates the integration of a Gloss Knowledge Infusion Module. "WavLM" refers to using features extracted from the pre-trained large-scale speech model wavLM instead of conventional MFCC. "Gloss-CLIP" denotes the incorporation of Gloss-based Motion CLIP Fine-tuning. We can observe that the absence of any module leads to a decline in performance metrics, demonstrating the efficacy of each module in our framework. Specially, the absence of the Gloss-prompt and Gloss-CLIP modules results in a decrease in PCK by 1.85% and 1.26%, respectively, highlighting their critical role in fine-grained generation.

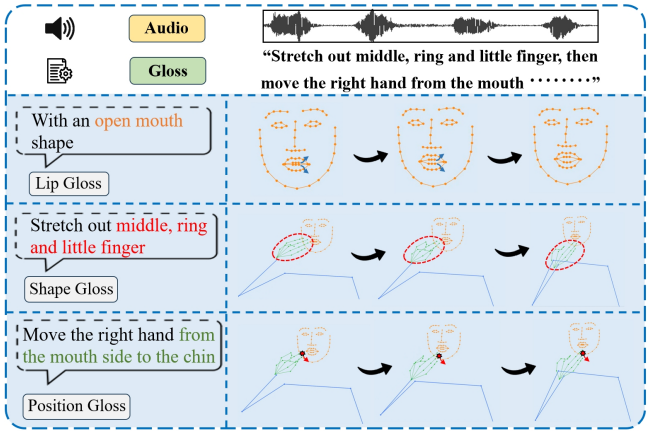

Figure 3: The visualization result of the generated gesture according to fine-grained Gloss. Better view by zooming in.

4.4 Qualitative Result and Analysis

Visualization of Generated Fine-grained CS Gesture

Figure 3 shows fine-grained hand gestures generated with gloss prompts, where each row shows the detailed gloss of different body parts and their gesture sequences. We used arrows to indicate lip movement trends, red circles for finger shape transformations, and red stars for hand position shifts, including their movement directions. The first row in the figure shows the lips' contour expanding as the gloss input. The second row emphasizes finger shape changing aligned with detailed finger gloss. In the third row, there are subtle hand position shifts, marked by red stars moving from near the mouth to the chin area, showing our method's effectiveness in using detailed gloss to guide CS gesture generation.

Distribution of Fine-grained Gesture Feature

To visualize the generated CS gesture in the feature space, we used t-SNE [van der Maaten and Hinton, 2008] for dimension reduction. We uniformly select frames from the generated CS sequences and extract the hand gesture features corresponding to the text. Recall that, as depicted in Figure 1, the MCCS incorporates 8 distinct finger shapes to signify the 24 consonants of the Chinese language, along with 5 hand positions to denote the 16 vowels. In the left part of Figure 4, the 8 distinct clusters are separate, with each cluster corresponding to

a set of finger shapes (where each color represents a different consonant group). Some clusters that are very close in distance have similar finger shapes, such as shape8 and shape6, as well as shape2 and shape7. This visualization validates the effectiveness of our method in capturing the fine-grained semantics of CS hand and finger shapes. On the right side of Figure 4, We can find different hand positions have differences in features, but there is more overlap among the clusters, which means they are not as distinctly differentiated in feature-level as finger shapes.

Visualization of Generated CS Gestures

Figure 5 compares the visualization results of our method with the SOTA method, DiffGesture. This comparison includes the gestures' corresponding audio, text, and ground truth video frames. We highlighted corresponding phonemes in red and used red stars and circles to indicate hand locations and finger shapes, respectively.

Our method shows a noticeable improvement in gesture accuracy, particularly in fine-grained details. For example, our index finger shape is more precise than the SOTA method, as seen in the first column. In the second column, our method accurately places the hand beside the face, unlike the SOTA method's placement beside the eye. The fourth column illustrates our method's superior precision in thumb position and overall gesture alignment with the ground truth, showing greater adherence to CS rules and enhanced detail accuracy.

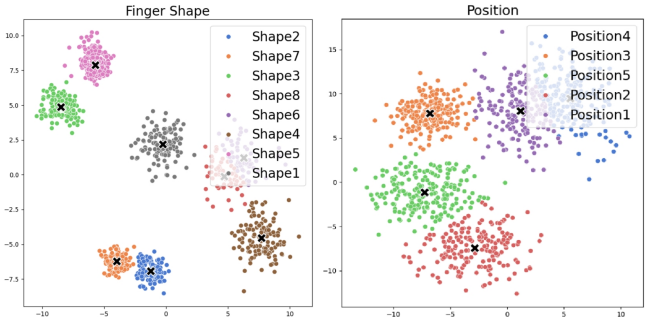

Figure 4: The visualization of t-SNE clustering for eight groups of consonants corresponding to finger shapes, and five groups of vowels corresponding to hand position. Each color represents a group of consonants or vowels.

User Study

We conduct a user study to evaluate CS gestures generated by our method compared with SOTA and the ground truth. This study involved 10 groups of videos, each with a ground-truth CS gesture video, videos generated by the current SOTA method (DiffGesture) and our method (GlossDiff). All videos were randomly shuffled. Ten subjects trained in CS were asked to rate the CS gesture videos from three perspectives: accuracy, rhythm quality, and naturalness, each with a score ranging from 0 to 10 (the higher the better). We calculated average scores and confidence intervals for each case.

It is shown in Figure 6 that our method surpassed the current SOTA DiffGesture in all three metrics, getting closer

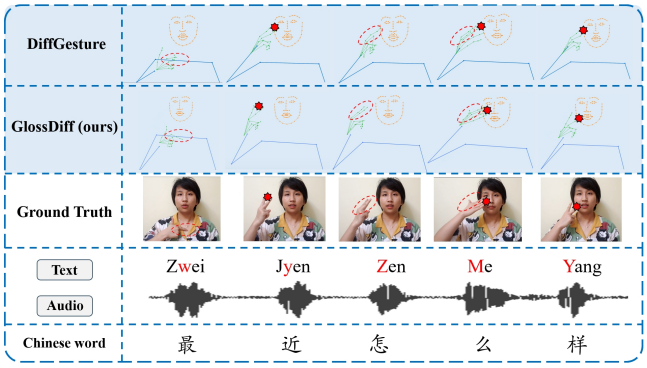

Figure 5: The visualization result of the generated gestures compared to SOTA method. Better view by zooming in.

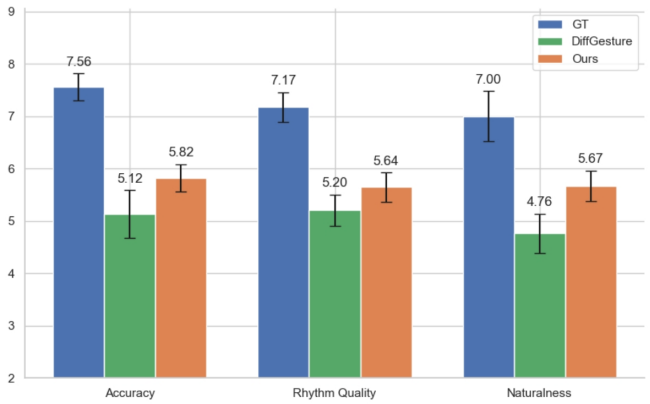

Figure 6: User study results of the ground truth (GT)), current SOTA (DiffGesture) and our method (GlossDiff).

to the ground truth. This demonstrates our method's ability to produce more accurate and natural CS gestures, especially in rhythm quality, attributed to the proposed ARM. Our approach notably outperforms the DiffGesture in accuracy, proving its effectiveness in fine-grained gesture generation.

5 Conclusion

We introduced a novel GlossDiff framework that effectively generates fine-grained CS gesture sequences. We have proposed a gloss knowledge infusion module and an audio rhythm module for an accurate and natural CS gesture video generation. Additionally, we contributed the first large-scale MCCS dataset. Extensive experiments on MCCS demonstrate our approach's efficacy, surpassing current SOTA methods. Qualitative experiments and ablation studies validated our system's overall effectiveness as well as each individual module's. Future work aims to infuse CS video generation with prosody and emotion. The Automatic Prompt Engineering (APE) is also a promising direction to improve gloss quality.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62101351), Guangzhou Municipal

Science and Technology Project: Basic and Applied Basic research projects (No. 2024A04J4232) and Tencent AI Lab Rhino-Bird Program (No. RBFR2023014).

References

[Ao *et al.*, 2022] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator. *ACM Transactions on Graphics*, 41(6):1–19, 2022.

[Ao *et al.*, 2023] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *arXiv:2303.14613*, 2023.

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

[Ben *et al.*, 2020] S. Ben, N. Camgoz, and B. Richard. Progressive transformers for end-to-end sign language production. In *ECCV*, 2020.

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

[Chen *et al.*, 2022] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[Choutas *et al.*, 2020] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *ECCV*, 2020.

[Cornett, 1967] R. Orin Cornett. Cued speech. *American Annals of the Deaf*, 112(1):3–13, 1967.

[Cox *et al.*, 2002] Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. Tessa, a system to aid communication with deaf people. In *Proceedings of the fifth international ACM conference on Assistive technologies*, pages 205–212, 2002.

[Duchnowski *et al.*, 1998] 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.

[Fernandez-Lopez *et al.*, 2017] 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.

[Ginosar *et al.*, 2019] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. Learning individual styles of conversational gesture. In *CVPR*, 2019.

[Guo *et al.*, 2021] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *ICCV*, 2021.

[Ho and Salimans, 2022] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022.

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

[Huang and Belongie, 2017] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[Ji *et al.*, 2023] Longbin Ji, Pengfei Wei, Yi Ren, Jinglin Liu, Chen Zhang, and Xiang Yin. C2g2: Controllable co-speech gesture generation with latent diffusion model. *arXiv:2308.15016*, 2023.

[Lebourdais *et al.*, 2022] Martin Lebourdais, Marie Tahon, Antoine Laurent, and Sylvain Meignier. Overlapped speech and gender detection with wavlm pre-trained features. *arXiv preprint arXiv:2209.04167*, 2022.

[Liddell and Johnson, 1989] K. Scott Liddell and E. Robert Johnson. American sign language: The phonological base. *Sign Language Studies*, pages 195–278, 1989.

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

[Liu and Liu, 2023] Lei Liu and Li Liu. Cross-modal mutual learning for cued speech recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Liu *et al.*, 2018] Li Liu, H. Thomas, Gang Feng, and B. Denis. Visual recognition of continuous cued speech using a tandem cnn-hmm approach. In *INTERSPEECH*, 2018.

[Liu *et al.*, 2019] Li Liu, Jianze Li, Gang Feng, and Xiao-Ping Steven Zhang. Automatic detection of the temporal segmentation of hand movements in british english cued speech. In *INTERSPEECH*, 2019.

[Liu *et al.*, 2020] Li Liu, Gang Feng, Denis Beautemps, 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*, 23:292–305, 2020.

[Liu *et al.*, 2021] 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*, 23:292–305, 2021.

[Liu *et al.*, 2022] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *CVPR*, 2022.

[Liu *et al.*, 2023] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

[OpenAI and et.al, 2023] OpenAI and Josh et.al. Gpt-4 technical report, 2023.

[Petrovich *et al.*, 2021] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *ICCV*, 2021.

[Power *et al.*, 2007] Des Power, Mary R Power, and Bernd Rehling. German deaf people using text communication: Short message service, tty, relay services, fax, and e-mail. *American Annals of the Deaf*, 152(3):291–301, 2007.

[Puviarasan and Palanivel, 2011] N. Puviarasan and S. Palanivel. Lip reading of hearing impaired persons using hmm. *Expert Systems with Applications*, 38(4):4477–4481, 2011.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Reynolds, 2007] 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.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[Rothauser, 1969] E. H. Rothauser. Ieee recommended practice for speech quality measurements. In *Technical Report No. 297*, 1969.

[Stokoe, 2005] 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*, 10(1):3–37, 2005.

[Stoll *et al.*, 2020] S. Stoll, Necati C. Camgoz, S. Hadfield, and R. Bowden. Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *IJCV*, 2020.

[Tevet *et al.*, 2022a] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *ECCV*. Springer, 2022.

[Tevet *et al.*, 2022b] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *ICLR*, 2022.

[Timothy, 2003] R. Timothy. Linguistics of american sign language: An introduction. *Studies in Second Language Acquisition*, 25(1):157–158, 2003.

[van der Maaten and Hinton, 2008] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[Yang *et al.*, 2023] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. *arXiv:2305.04919*, 2023.

[Yi and Deva, 2013] Y. Yi and R. Deva. Articulated human detection with flexible mixtures of parts. *TPAMI*, 35(12):2878–2890, 2013.

[Youngwoo *et al.*, 2019] Y. Youngwoo, K. Woo-Ri, J. Minsu, L. Jaeyeon, K. Jaehong, and L. Geehyuk. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *ICRA*, 2019.

[Youngwoo *et al.*, 2020] 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*, 39(6):1–16, 2020.

[Zhang *et al.*, 2022] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv:2208.15001*, 2022.

[Zhao *et al.*, 2023] Weiyu Zhao, Liangxiao Hu, and Shengping Zhang. Diffugesture: Generating human gesture from two-person dialogue with diffusion models. In *ICMI*. 2023.

[Zhi *et al.*, 2023] Yihao Zhi, Xiaodong Cun, Xuelin Chen, Xi Shen, Wen Guo, Shaoli Huang, and Shenghua Gao. Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *ICCV*, 2023.

[Zhu *et al.*, 2023] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *CVPR*, 2023.