AnyoneCue: Gloss-Prompted Fine-grained and Personalized Cued Speech Video Generation

Li Liu†,*, Wentao Lei†, Jun Wang, Wenwu Wang, *Senior Member, IEEE*

***Abstract*—Cued Speech (CS) is a visual coding system, which combines lip-reading with several specific hand codings to help hearing-impaired people to communicate effectively. Generating CS videos from audio speech and text can significantly improve accessibility and communication for individuals with hearing impairments. However, existing video generation methods primarily concentrate on general gestures, such as human walking, and hence are not directly suitable for generating CS videos. Moreover, current approaches struggle to produce realistic, fine-grained, personalized videos adhering to specific CS coding rules. To address these challenges, firstly, we propose a Gloss-based Diffusion Pose Generation Model (GlossDiff), where the gloss is a novel CS motion parsing prompt to integrate additional linguistic rules knowledge into the CS pose generation model. The glosses are automatically generated descriptive texts based on Large Language Models (LLMs) to establish a direct and delicate semantic connection between CS gestures and spoken language. Secondly, a Pose-Refined Video Diffusion Model (PRV-DM) is proposed to leverage the generated pose sequences to produce fine-grained and personalized CS videos. Specifically, to address the critical challenges of pose scale mismatches with personalized references, ambiguous lip shape, and hand deformations in generated videos, we introduce a Multi-faceted Pose-Refined Module (MFPR) that contains pose alignment, lip enhancement and hand refinement stages. Furthermore, we record and publish the largest Mandarin Chinese CS dataset (named MCCS-2024), containing seven Chinese CS cuers. Extensive experiments and user studies demonstrate the effectiveness of our method, making it the first diffusion model based approach for generating fine-grained and personalized CS videos. The code and dataset with multi-modal annotations were made public at https://mccs-2024.github.io/.**

***Index Terms*—Cued Speech Video Generation, Human-Computer Interaction for Hearing-Impaired, CS Gloss, Diffusion Model.**

I. Introduction

According to the World Health Organization (WHO), over 5% of the world's population, approximately 466 million people, experience hearing loss. Lip-reading, a primary communication method for individuals with hearing impairments, faces significant challenges due to visual ambiguity [1], [2]. For example, it often fails to distinguish between phonemes with similar lip movements, *e.g.*, [u] and [y], creating barriers for hearing-impaired individuals in accessing spoken language through traditional educational methods.

To address the shortcomings of lip-reading and enhance the literacy skills of those with hearing impairments, Cornett developed the Cued Speech (CS) system in 1967 [4]. This system enhances lip-reading by incorporating hand gestures, including specific finger shapes and hand positions, to provide a precise visual representation of all phonemes in spoken language [1], [2]. For instance, in Mandarin Chinese CS (MCCS) [5] (see Fig. 1), five hand positions are used to encode vowel groups, while eight finger shapes represent consonant groups. By integrating hand cues, CS enables individuals with hearing impairments to distinguish sounds that may appear identical on the lips. It is important to note that while Sign Language (SL) is another widely used communication method [6]–[8], CS differs fundamentally as it is not a visual language but rather a coding system for spoken language [4]. Research has shown that CS can be learned more rapidly than SL [9]. Furthermore, compared to written text, CS is more accessible and easier to adopt for hearing-impaired individuals who may lack literacy skills [10], [11].

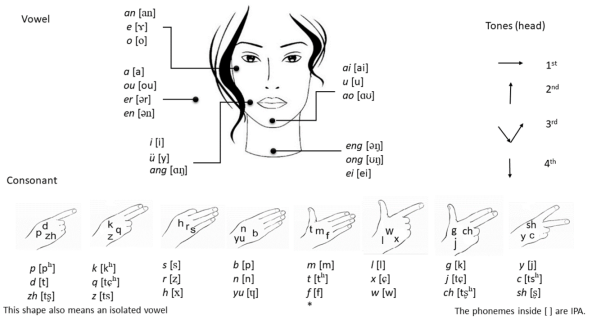

Fig. 1. The encoding rule for Mandarin Chinese Cued Speech (figure from [3]), which utilizes five separate hand positions to represent vowels and eight hand configurations to represent consonants.

Recently, the automatic conversion between multi-modal CS video (*i.e.*, face, lip-reading, hand shape, and hand position movements) and text/audio speech attracts researchers' attention [12]–[14]. It includes CS video-to-text/audio (see direction 1 in Fig. 2) and the inverse text and audio speech to CS video generation (see direction 2 in Fig. 2). This automatic conversion of the CS system can significantly improve the communication efficiency between the hearing-impaired and hearing-impaired/normal hearing. Even though this research topic has been studied for a long time, most of the works focused on the direction 1 (*i.e.*, French/British CS video-to-text recognition), while the multi-modal CS video generation is under-explored because of the following reasons: **1)** high requirement for fine-grained CS video generation, as shown in Fig. 3(a), where nuances in the hand's position and hand

† Equal Contribution.

* Corresponding Author: avrillliu@hkust-gz.edu.cn.

Li Liu and Wentao Lei are with the Hong Kong University of Science and Technology (Guangzhou). Jun Wang is with Tencent. Wenwu Wang is with Surrey University.

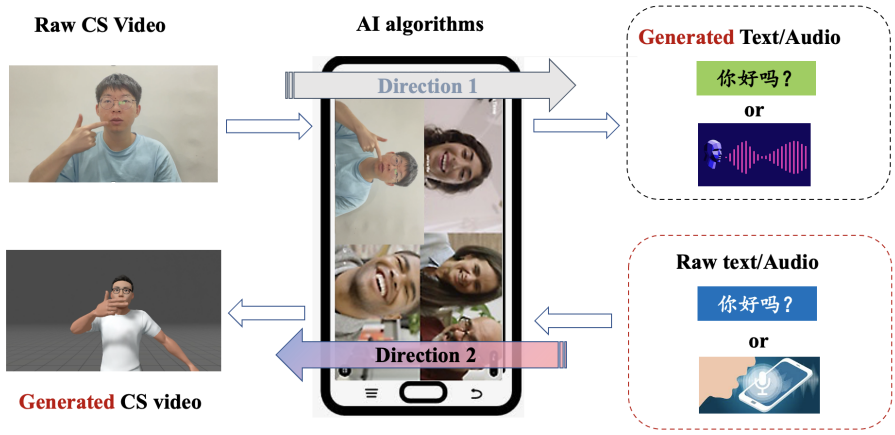

Fig. 2. The overall framework of the conversion between CS and text/speech. Direction 1 means CS to text/speech generation, and direction 2 means text/speech to CS gesture generation. The first direction aims to synthesize text or speech to make normal hearing better understand the hearing-impaired people, and the second direction can help the hearing-impaired to visually understand normal-hearing people. **This work focuses on the multi-modal CS video generation based on the audio speech and text (*i.e.*, Direction 2)**.

shape lead to quite different semantic meanings. The limited size of CS datasets and the high annotation cost of complex fine-grained CS gestures make it very challenging. **2)** When generating the CS video, existing models produce unclear lip movements, leading to poor expressiveness in lip-reading (see Fig. 3(b)). **3)** In the process of generating personalized CS videos, where a personalized reference image controls the character's appearance, there are issues due to the mismatch between the scale of the generated human pose[1] and the size of the given personalized reference image. This mismatch leads to distortions in the facial appearance. Moreover, rapid movements and transformations of the hands result in the generated hand fingers being severely deformed (see examples in Fig. 3(c)).

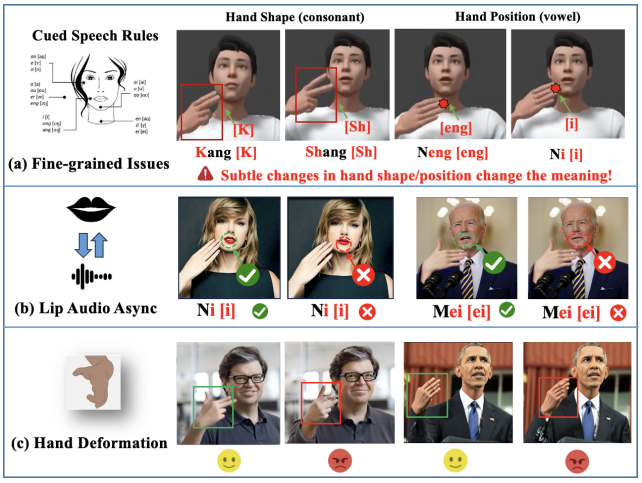

Fig. 3. Three typical challenges in the automatic CS video generation.

To address these challenges, we propose **AnyoneCue**, the first speech and text-driven diffusion-based framework (see Fig. 4) for generating fine-grained and personalized CS videos. To overcome the first challenge mentioned above, a novel CS **gloss**, which is a direct CS motion instruction to bridge the gap between spoken language and CS gestures, is first proposed. The gloss (*i.e.*, intermediate instruction text) describes the process of using CS gestures to express the text phonetically. On this basis, a **Gloss**-based **Diff**usion model (**GlossDiff**) is proposed to produce the CS pose sequence, which can be used to generate fine-grained CS video. The gloss is used as an effective prompt for the GlossDiff.

Furthermore, rhythm plays a crucial role as paralinguistic information in spoken language. We argue that, as a coding system for spoken languages, CS also requires consideration of natural rhythmic dynamics for complete semantic expression. Here, rhythm specifically refers to the generation of synchronized multi-modal CS gesture movements (*i.e.*, hand and finger movements) that align with the phoneme durations and utterance prosody of the speech signal. Notably, prior research has largely overlooked this important aspect. To address this, we introduce an **A**udio-driven **R**hythmic **M**odule (**ARM**) designed to ensure that CS movements are rhythmically aligned with corresponding speech signals.

Lastly, the CS poses generated by GlossDiff are used to generate the CS video. When generating personalized CS videos, where a personalized reference image controls the character's appearance, there are issues due to the mismatch between the scale of the generated human pose and the size of the given personalized reference image. This mismatch leads to distortions in the facial appearance. Moreover, in this process, existing models suffer from issues like lip shape ambiguity and hand deformation (as mentioned in Fig. 3 (b) and (c)). To tackle the above problems (*i.e.*, the second and the third challenges shown in Fig. 3), we design a **Pose-Refined Video Diffusion Module (PRV-DM)**, which includes three-faceted processing: pose alignment between the generated pose and given reference image, lip enhancement and hand refinement.

In summary, the main contributions of this work are as follows:

- We introduce AnyoneCue, the first framework for gener-

[1]Pose means the keypoints sequence of lip and hand in CS videos.

ating fine-grained and personalized CS videos driven by audio speech and text.
- We propose a GlossDiff that generates fine-grained CS pose sequences (hand position, shape, and lip movements) by introducing a CS gloss to link text/speech with precise hand and lip motions. In addition, a rhythm-aware mechanism ensures pose alignment with audio speech signals.
- We develop a PRV-DM to produce high-quality CS videos, effectively tackling three key challenges: inconsistent pose scaling relative to the reference image, lip shape ambiguity, and hand deformations.
- We establish and release the MCCS-2024, the largest Chinese CS dataset with seven CS cuers. Extensive experiments demonstrate state-of-the-art (SOTA) performance, supported by qualitative, ablation, and user studies.

Note that this work is an extension of our previous conference paper presented at the IJCAI[2] 2024 [15] (7 pages), which introduced a diffusion-based framework for fine-grained CS pose generation. In contrast to the prior work, which focused solely on CS pose generation, this study advances the field by addressing the more challenging task of **CS video generation**. The key improvements of this work include: **(i)** a PRV-DM that generates high-quality CS videos from pose sequences, ensuring fine-grained and personalized outputs; **(ii)** a Multi-faceted Pose-Refined Module (MFPR) designed to tackle the challenges of CS video generation by refining pose alignment, enhancing lip synchronization, and correcting hand deformations, significantly improving video quality; and **(iii)** the introduction of the MCCS-2024 dataset, the largest Mandarin Chinese CS dataset to date, which expands the number of cuers from 4 to 7, including one hearing-impaired cuer, providing a more diverse and robust foundation for training and evaluation. Together, these advancements enable AnyoneCue to produce realistic and accurate CS videos, significantly enhancing communication accessibility for the hearing-impaired community.

II. Related Work

A. Automatic Cued Speech Generation

In previous studies, prior work in generation of CS gestures [16], [17] primarily relied on rule-based methods. For example, in [16], researchers manually selected specific keywords and used low-context sentences [18], while predefined templates were created for corresponding hand gestures. This approach involved recognizing CS content and then mapping the recognized text to predefined hand gesture templates. However, this method was heavily dependent on manual design, which limited the expressiveness of CS gestures and required significant human effort. In [17], a post-processing algorithm was proposed to enhance synthesized hand gestures by adjusting hand rotation and translation. Despite this improvement, the method still required prior human knowledge to adapt the algorithm to new images, leading to limited robustness. More recently, in [19], a pre-trained audiovisual text-to-speech model was employed to generate hand and lip poses for CS. This approach, based on a Bi-LSTM architecture, focused on generating individual hand and lip poses using a non-public French CS dataset rather than addressing the complete task of CS video generation. **To the best of our knowledge, there remains a gap in research on diffusion model-based CS gesture and video generation.**

[2]International Joint Conferences on Artificial Intelligence (IJCAI)

B. Co-speech and Sign Language Generation

Co-speech gesture generation involves creating body movements that align with audio speech input, a technique widely applied in virtual character animation, particularly for virtual speaking and advertising. Our focus here is on deep learning-based approaches for co-speech gesture generation. Earlier research primarily focused on developing large-scale speech-gesture datasets to learn the mapping between audio speech and human skeletal movements using deep learning techniques, as seen in [20]. To enhance the expressiveness of gestures, some methods employ Generative Adversarial Networks (GANs) to achieve more realistic results [21], [22]. Recently, diffusion models such as DiffGesture [23] have demonstrated effectiveness in linking speech and gestures while maintaining temporal consistency, enabling the generation of high-quality co-speech gestures. However, while co-speech gesture generation emphasizes fluency and style, it still struggles with fine-grained accuracy, particularly in generating subtle hand gestures and precise lip shapes. This limitation renders existing co-speech generation methods less suitable for CS generation, where precise synchronization of hand gestures and lip movements is critical for effective communication.

In the literature, several methods have been proposed for Sign Language (SL) generation: 1) The Neural Machine Translation approach in [24] treats SL generation as a translation task, utilizing neural models to process SL text. 2) The Motion Graph method in [24] constructs a directed graph from motion capture data to generate SL gestures. 3) Conditional generation methods, such as GANs and Variational Autoencoders (VAEs), have also been applied to SL gesture synthesis. 4) Transformer-based models, as discussed in [25], have shown promise in this domain. Despite these advancements in SL gesture generation, applying these methods to CS still has some limitations. Firstly, CS generation requires fine-grained hand and lip gesture generation, especially in their position and shape, which aligns with specific phonemes. In contrast, SL gesture generation [26] is abstract and not sensitive to the detailed position of the hand. Secondly, CS requires synchronization between hand gestures and lip movements to enhance comprehension, which is a property not present in SL. The SL generation task [27] does not emphasize lip synchronization with spoken speech rhythm and lacks the specific synchrony characteristics observed in CS [14], [28]. Directly employing the current methodologies designed for SL generation does not yield satisfactory results in our CS gesture generation task.

C. General Human Motion Video Generation

In recent developments within the field of 2D human gesture video generation, diffusion models [29] have emerged as a

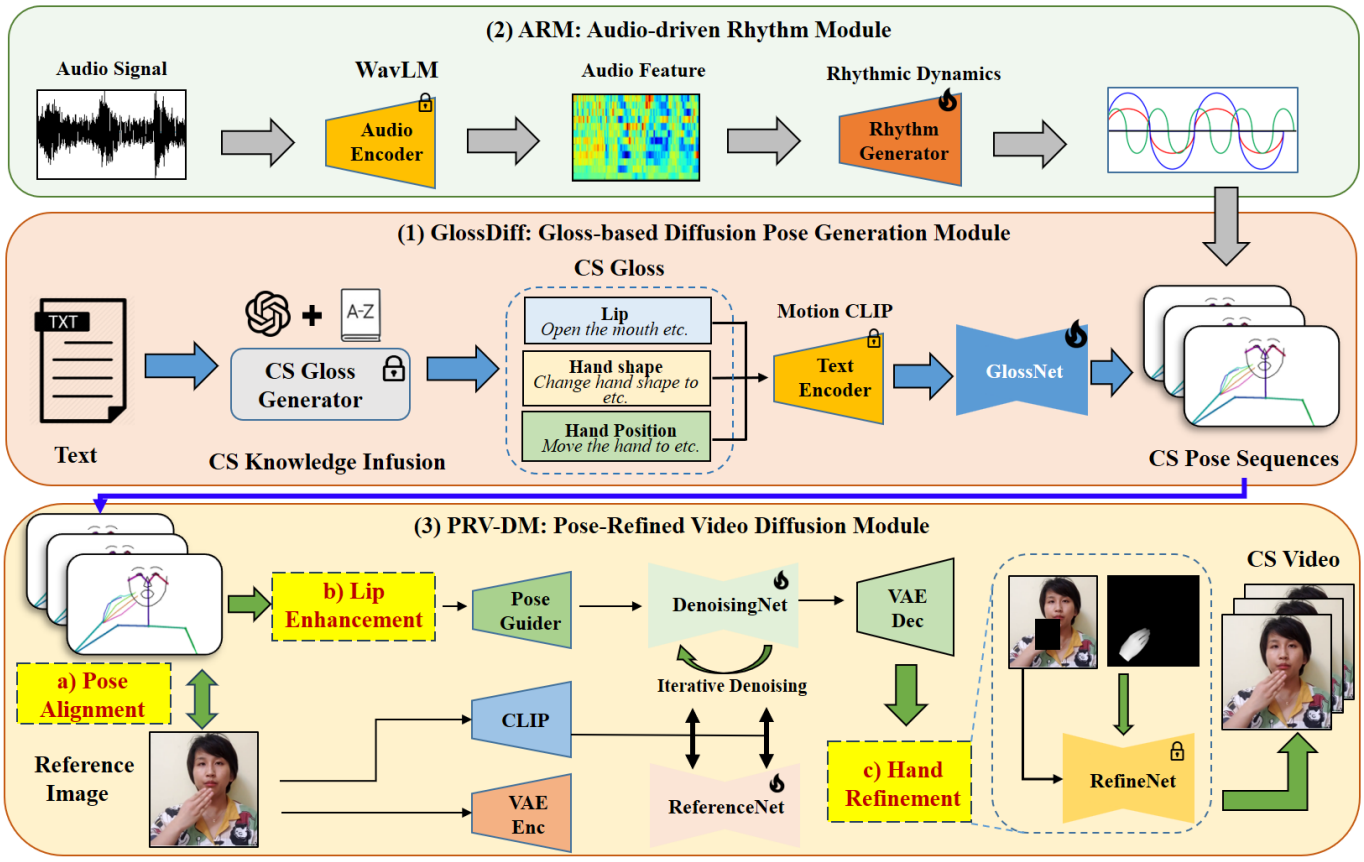

Fig. 4. The overall framework of the proposed AnyoneCue framework, where (1), (2), (3) represent gloss-based diffusion pose generation module, audio-driven rhythm module, and pose-refined video diffusion module, respectively.

leading approach, primarily applied in two key areas: text-driven general body motion gesture generation (*e.g.*, human walking, jumping and kicking) [30], [31] and pose-driven motion video generation [32], [33]. In this task, fine-grained pose generation is a crucial problem, which refers to the controllable generation of face (especially the lips), hands (especially the fingers), and body movement details. As far as we know, it remains a challenging area so far, particularly for CS generation which does not have enough data. Existing research has made some attempts. For instance, talking head methods [34], [35] have achieved lip-syncing with speech but are limited in generating body details. For fine-grained hand generation, [36], [37] have explored hand refinement, but these efforts are limited to the image level. Consequentially, how to achieve fine-grained video generation for human body remains an open research problem.

III. Proposed Method

In this section, we introduce the proposed AnyoneCue framework, which consists of three main components: a Gloss-based diffusion Pose Generation Module (GlossDiff), an Audio Speech-driven Rhythmic Module (ARM) to capture the rhythmic dynamics of CS gestures, and a Fine-grained Pose-Refined Video Diffusion Model (PRV-DM), which contains a Multi-faceted Pose-Refined Module (MFPR). The overall architecture is illustrated in Fig. 4.

A. Problem Formulation

Automatic CS video generation aims to generate the corresponding pose sequence of CS gestures, denoted as M^*, which includes lip movements, hand shapes (fingers), and hand positions, given an input audio speech signal A and the associated text T.

The combined features of A, T, and the generated rhythmic information are fed into the CS gesture generator. The final CS gesture poses (M^*) are obtained by minimizing the following objective function:

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

where K denotes the total number of frames in the current CS video. The ground truth CS gesture keypoints M_i in the i-th frame of the CS video are extracted using the *Expose* method [38]. Here, $M_i^* = \hat{M}_i + \widetilde{M}_i$, where $\hat{M} = \mathcal{G}_{\mathcal{S}}(T, A)$ contains all generated semantic gesture keypoints. Specifically $\hat{M} = \{\hat{M}_i\}_{i=1}^{K}$, where K is the number of frames, and $\hat{M}_i$ contains all the landmarks of human pose in the i-th frame. $\mathcal{G}_{\mathcal{S}}$ is the semantic gesture generator. Additionally, $\widetilde{M} = \mathcal{G}_{\mathcal{R}}(A)$ corresponds to the rhythmic information derived from the input audio speech, with $\mathcal{G}_{\mathcal{R}}$ serving as the rhythm generator.

The generated CS pose sequences $M^* = \{M_i^*\}_{i=1}^{K}$, are then used to synthesize the final CS video, where K is the number of frames, and M_i^* contains all the landmarks of human pose

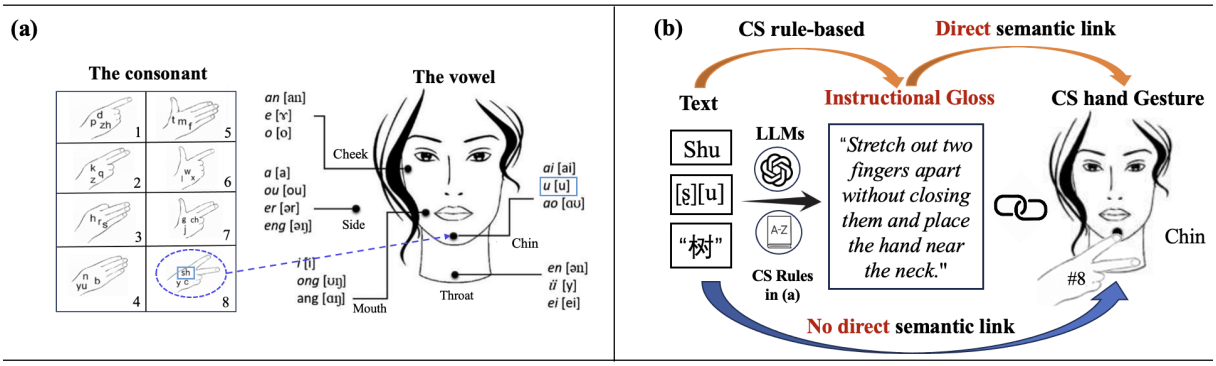

Fig. 5. The illustration of the Cued Speech (CS) gloss generation process is presented. (a) shows the encoding rule for Mandarin Chinese Cued Speech (figure from [3]), where vowels are represented by five unique hand positions and consonants by eight distinct finger shapes. (b) showcases the pipeline of our proposed instructional gloss, designed to directly connect the source text to the corresponding CS movements.

in the i-th frame. We employ a deep generative model $\mathcal{G}_{\mathcal{V}}$ that takes the pose sequence M^* and reference images I_{ref} as input and produces realistic video V.

$$V = \mathcal{G}_{\mathcal{V}}(I_{ref}, M^*), \tag{2}$$

where $\mathcal{G}_{\mathcal{V}}$ is trained to map from pose representations to photorealistic video frames while preserving temporal consistency and fine-grained details. More details about $\mathcal{G}_{\mathcal{S}}$, $\mathcal{G}_{\mathcal{R}}$, $\mathcal{G}_{\mathcal{V}}$ are discussed in the following subsection III-B, III-C, III-D respectively.

B. GlossDiff: Gloss-based diffusion Pose Generation Module

In this section, we introduce a novel Gloss-based diffusion Pose Generation Module (GlossDiff) designed to simultaneously generate fine-grained hand positions, hand shape movements, and lip shapes in CS. This module comprises three main components: the Knowledge Infusion Module, the Gloss-based Motion CLIP Fine-tuning, and the Gloss-Prompted Diffusion Model (named GlossNet).

1) ***Knowledge Infusion Module****:* The main goal of the knowledge infusion module is to convert language text T (such as speech transcription) into gloss (*i.e.*, direct text instructions, see Fig. 5(b)), which phonetically detail the corresponding fine-grained CS motions. To accomplish this, we utilize the LLM, specifically ChatGPT-4 [39], applying a prompt engineering strategy to incorporate the encoding rules of Chinese CS [5] into our framework through the following:

$$g = \text{LLM}(T, P), \tag{3}$$

where P represents our specially crafted prompt, derived from CS domain knowledge (*i.e.*, established transformation rules of CS [5]), and T is the input text. This approach ultimately enables the transformation of text that is indirectly linked to the semantics of CS video into a gloss that is directly aligned with the content of the CS video.

2) ***Gloss-based Motion CLIP Fine-tuning****:* MotionCLIP [40] is a large-scale multimodal model designed specifically for generating a variety of human motion gestures. To achieve an accurate feature embedding for CS gloss, we utilize MotionCLIP as a pre-trained foundation model and fine-tune it with the generated CS gloss (refer to Subsection III-B1) along with the associated CS gestures.

In the fine-tuning phase, we employ a CLIP-style contrastive learning approach [41] to adjust the encoders using CS data. Consider a batch of pairs consisting of CS gesture motion and gloss embeddings, represented as $\mathcal{B} = \{(z_i^m, z_i^g)\}_{i=1}^B$, where B represents the batch size. The latent features $z_i^m, z_i^g \in \mathbb{R}^C$, where C is the dimension of the latent space. $\mathcal{E}_m$ and $\mathcal{E}_g$ are the encoders for the motion sequence and gloss, respectively. The latent features are expressed as $z^m = \mathcal{E}_m(M), \quad z^g = \mathcal{E}_g(g)$. The training aims to enhance the similarity between the paired z_i^m and z_i^g in the batch while reducing the similarity of mismatched pairs $\left(z_i^m, z_j^g\right)_{i \neq j}$. A symmetric cross entropy (CE) loss L_{CE} is optimized based on these similarity scores. The formal expression for the loss is:

$$\begin{aligned} \mathcal{L}_{\text{CLIP}} = \mathbb{E}_{\mathcal{B}\sim\mathcal{D}}[&L_{CE}\left(y\left(z_i^m\right), p_m\left(z_i^m\right)\right) \\ &+L_{CE}\left(y\left(z_j^g\right), p_g\left(z_j^g\right)\right)], \end{aligned} \tag{4}$$

where y is a one-hot encoding that indicates the true relationship between gestures z_i^m and gloss z_j^g within the training batch $\mathcal{B}$. If they form a pair, $y = 1$, otherwise, $y = 0$. The probability p is defined as follows:

$$p_m\left(z_i^m\right) = \frac{\exp\left(z_i^m \cdot z_i^g / \eta\right)}{\sum_{j=1}^B \exp\left(z_i^m \cdot z_j^g / \eta\right)}, \tag{5}$$

where η is the softmax temperature, and $p_g\left(z_j^g\right)$ is computed in a similar manner.

3) ***GlossNet: Gloss-Prompted Diffusion Model****:* To generate CS gesture pose sequences, we introduce GlossNet, a Gloss-Prompted Diffusion Model. Specifically, the semantic hand gesture pose generator $\mathcal{G}_{\mathcal{S}}$ is constructed based on the latent diffusion model [42], which performs diffusion and denoising operations within a pre-trained latent space. The model is trained using the standard noise estimation loss [29], which is defined as follows:

$$\mathcal{L}_{\text{noise}} = ||\epsilon - \epsilon_\theta\left(Z_n, n, g, A\right)||_2^2, \tag{6}$$

where Z_n denotes the latent CS gesture at each time step n. Here, A represents the Mel-spectrogram of audio speech, and g is the generated gloss. The term ϵ is the actual noise, which share the same dimension with the latent feature. ϵ_θ is the

noise predicted by the latent diffusion model, with θ being its parameters.

To incorporate gloss prompt information into the diffusion network, we utilize an adaptive instance normalization (AdaIN) layer [43]. Specifically, we use the fine-tuned MotionCLIP gloss encoder $\mathcal{E}_g$ to transform the gloss prompt into a gloss embedding z^g. A Multilayer Perceptron (MLP) network is then trained to map this gloss embedding z^g to parameters that adjust the per-channel mean and variance of the AdaIN layer.

For training the GlossNet, we apply classifier-free guidance as outlined in [44]. During training, we allow $\mathcal{G}_\mathcal{S}$ to learn both conditional and unconditional semantic distributions by randomly setting gloss g to $\varnothing$, where $\varnothing$ denotes the absence of any gloss condition, effectively deactivating the AdaIN layer during the training process with a probability p, which is set to 10% [30]. During inference, the expected noise is:

$$\epsilon_n^* = p\epsilon_\theta\left(Z_n, n, g, A\right) + (1-p)\epsilon_\theta\left(Z_n, n, \varnothing, A\right). \quad (7)$$

Once the predicted noise ϵ_n^* is obtained, the model proceeds in a reverse step-by-step manner over N time steps, updating a latent gesture sequence Z_n at each time step n. It starts by generating a sequence of latent codes $Z_N \sim \mathcal{N}(0, I)$ and then calculates a series of denoised sequences Z_n by iteratively removing the estimated noise ϵ_n^* from Z_n for $n = N-1, \ldots, 0$. The final generated CS gesture latent embedding Z_0 is obtained through N reverse diffusion steps. This Z_0 is input into a Transformer-based decoder [45] to produce semantic CS gesture motion $\hat{M}$.

*4) **Training of the GlossDiff**:* We utilize a semantic loss function to ensure the semantic accuracy of the final generated CS pose sequences. Specifically,

$$\mathcal{L}_{\text{semantic}} = 1 - \cos\left(Z_0, Z_0^*\right). \quad (8)$$

Here, $\cos(\cdot,\cdot)$ signifies the cosine distance, whereas Z_0 and Z_0^* refer to the final generated CS gesture latent embedding and the actual CS pose gesture motions, respectively.

In line with the conventional training approach for denoising diffusion models, we aim to minimize the following loss function:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{noise}} + \beta\mathcal{L}_{\text{semantic}} + \gamma\mathcal{L}_{\text{rhythm}}, \quad (9)$$

where α represents the weight assigned to $\mathcal{L}_{\text{noise}}$ (Equation (6)), β denotes the weight for $\mathcal{L}_{\text{semantic}}$ (Equation (8)), and γ indicates the weight for $\mathcal{L}_{\text{rhythm}}$ (Equation (10)).

C. Audio Speech-driven Rhythm Module

In generating CS gestures, achieving precise hand gesture generation is not the only concern; the natural rhythm of gesture movements is also important. We believe that the audio speech signal encompasses not only semantic content but also rhythmic dynamics inherent to CS, which are essential for ensuring visual and auditory synchronization.

*1) **Rhythmic CS Pose Modeling**:* To tackle this issue, we present an innovative Audio Speech-driven Rhythmic Module (ARM) that is crafted to capture the rhythmic dynamics of CS gestures. This module makes use of three convolutional layers as a rhythmic dynamics generator $\mathrm{G_R}$, which further aligns the motion dynamics with the CS rhythm.

Prior studies (*e.g.*, WavLM and AudioLDM) [46], [47] have demonstrated that audio features extracted using large pre-trained models possess a more robust expressive capability compared to Mel-Frequency Cepstral Coefficients (MFCC) features, thereby minimizing information loss. Without loss of generality, we employ the encoder from WavLM, denoted as $\mathcal{E}_A$, to derive audio features, thereby preserving richer and higher-dimensional rhythmic information.

To address the lip-hand synchronization challenge [14] in CS, we redefine the task as determining the motion magnitude for each frame in successive motion sequences. Unlike approaches that strive for perfect alignment between generated gestures and speech, our method implicitly learns to produce asynchronous gestures corresponding to input speech. Instead of directly manipulating the gestures of each frame, we concentrate on regulating the overall rhythm of a motion sequence.

The loss function for ARM is as follows:

$$\mathcal{L}_{\text{rhythm}} = ||\widetilde{M} - \left(M - \bar{M}\right)||, \quad (10)$$

where $\bar{M}$ means the average pose motion within the set of generated CS motions M. The discrepancy between M and $\bar{M}$ measures the magnitude of hand position and finger movement. The goal of $\mathcal{L}_{\text{rhythm}}$ is to ensure that the generated $\widetilde{M} = \mathcal{G}_\mathcal{R}(\mathcal{E}_A(A))$ maintains a natural offset relative to the mean gesture. Here, $\mathcal{E}_A$ functions as the encoder of WavLM. This offset is crucial for creating motion dynamics that are natural and non-mechanical, without changing the semantics of the CS pose. We validate the effectiveness concerning rhythm quality and naturalness through quantitative results in Sec. V-A, along with qualitative results in Sec. V-B.

*2) **Novel Quantitative Rhythmic Metrics**:* In this work, we treat rhythm as a critical paralinguistic feature for enhancing the effectiveness of CS communication for the first time. To capture the distinctive asynchronous dynamics between lip and hand movements in CS situations, we introduce an innovative metric, the Gesture Audio Difference (GAD), to assess the rhythmic synchronization of generated gestures.

This metric is defined as follows:

$$\mathrm{GAD}(M, A) = \frac{1}{L}\sum_{i=1}^{L}\mathbf{1}[||\boldsymbol{U}_i^M - \boldsymbol{U}_i^A||_1 < \tau], \quad (11)$$

where M and A denote the CS gesture pose and Mel-spectrogram of audio speech, respectively. L represents the number of annotated temporal segments, which are identical for both speech and gesture. The variable $\boldsymbol{U}_i$ refers to the midpoint of a segment, indicating a specific moment when a gesture or speech occurs. The function $\mathbf{1}$ is an indicator function, assigning a value of one to elements within the subset (meeting the condition$||\boldsymbol{U}_i^M - \boldsymbol{U}_i^A||_1 < \tau$) and zero to all other elements.

Acknowledging the asynchrony between audio speech and CS hand movements, we introduce a threshold τ to ensure their

alignment. This threshold is empirically established based on a statistical analysis of the hand preceding time [48], which refers to the time difference between the hand reaching its target position and the corresponding phoneme being produced by the lips in CS. This time lag occurs because hand and lip movements in CS are not synchronized, with the hand typically moving in advance to provide visual cues before the lips produce the sound.

D. PRV-DM: Pose-Refined Video Diffusion Module

Based on the proposed GlossDiff, we obtain the fine-grained CS pose sequence. To generate personalized CS 2D video, we further propose a Pose-Refined Video Diffusion Module (PRV-DM), which is built on a temporal-aware 3D U-Net structure that incorporates temporal attention layers for modeling frame dependencies. Given a reference image I_{ref} and target pose sequence $M^{1:K}$ with frame length K, the video diffusion module $\mathcal{G}_{\mathcal{V}}$ generates temporally consistent videos while preserving detailed appearance features.

The framework employs a dedicated appearance encoder F_a to extract dense appearance features y_a from the reference image:

$$y_a = F_a(z_n|I_{ref}, n), \tag{12}$$

where z_n is the noise latent, and I_{ref} is the reference image for each denoising step n. These features are injected into the attention blocks through a spatial attention mechanism. The appearance encoder adopts a symmetrical U-Net structure without temporal layers. Each corresponding layer integrates features through spatial attention, enabling comprehensive learning of reference image relationships in a consistent feature space.

A lightweight pose motion guidance network F_m processes pose conditions and obtain the pose motion feature $y_m^{1:K}$:

$$y_m^{1:K} = F_m(z_n|M, n). \tag{13}$$

The network consists of 4 convolution layers with 4×4 kernels and 2×2 strides, with channel progression from 16 to 128. The final projection layer incorporates a zero convolution, and the network is initialized using RandomNormal initialization. The output resolution is aligned with the noise latent space.

The temporal layer processes feature maps $x \in \mathbb{R}^{b\times k\times h\times w\times c}$ by reshaping it to $x \in \mathbb{R}^{(b\times h\times w)\times k\times c}$, performing self-attention along temporal dimension K, and integrating via the residual connection. The full denoising process for a sequence with length K is formulated as follows:

$$\epsilon_\theta^{1:K}(z_n^{1:K}, n, I_{ref}, M^{1:K}) = F_T(z_n^{1:K}|n, y_a, y_m^{1:K}). \tag{14}$$

For long video generation, the sequence $z^{1:K}$ is divided into overlapping segments for generation.

The training consists of two stages. In the first stage, temporal layers are temporarily removed, and the model is trained with single-frame noise input. The appearance encoder and pose guidance network are jointly optimized by:

$$\mathcal{L}_1 = \mathbb{E}_{z,n,I_{ref},M,\epsilon\sim\mathcal{N}(0,1)}[||\epsilon - \epsilon_\theta||_2^2]. \tag{15}$$

The second stage introduces temporal layers, training on video clips with the loss:

$$\mathcal{L}_2 = \mathbb{E}_{z^{1:K},n,I_{ref},M^{1:K},\epsilon^{1:K}\sim\mathcal{N}(0,1)}[||\epsilon^{1:K} - \epsilon_\theta^{1:K}||_2^2]. \tag{16}$$

E. Multi-faceted Pose-Refined Processing

To solve the issues, such as inconsistent scale between the generated pose and the given reference image, lip shape ambiguity, and hand shape deformation, we design a multi-faceted pose refinement module, which contains the following three parts.

(a) FPA: Fine-grained Pose Alignment Module. Current methods typically assume strict alignment between the reference image and driving pose sequence, significantly limiting their applicability in CS gesture generation where precise pose alignment is challenging. To tackle the challenges of CS gesture generation where precise pose alignment is difficult, we propose a novel fine-grained alignment approach that enhances generation quality without extra training. Our method achieves fine-grained alignment accuracy while retaining both motion dynamics and identity features. The FPA contains the following two steps.

• **Structure-guided Transformation**: Given a generated pose sequence M and a reference pose M_{ref}, we first align the skeletal structure to ensure proper proportions. For each pair of connected keypoints (i_k, j_k) in frame k, we compute the scaling factor:

$$s_k = \frac{\|p_{\text{ref},j_k} - p_{\text{ref},i_k}\|}{\|p_{j_k} - p_{i_k}\|}, \tag{17}$$

where p and p_{ref} represent the keypoints from M and M_{ref}, respectively. Next, we scale the keypoints in M to match the proportions of M_{ref}:

$$p_{\text{scaled},i_k} = s_k \cdot p_{i_k}. \tag{18}$$

To ensure proper alignment, we compute the offset Δ between the center points of M and M_{ref}:

$$\Delta = p_{\text{ref},c} - p_c, \tag{19}$$

where $p_{\text{ref},c}$ and p_c are the center points of M_{ref} and M, respectively.

Then we compute the rotation angle θ between M and M_{ref} and apply a rotation matrix R:

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{20}$$

Finally, the aligned keypoints are computed by applying the scaling and offset:

$$p_{\text{align},i_k} = R \cdot p_{\text{scaled},i_k} + \Delta. \tag{21}$$

The aligned pose sequence M_{align} is obtained by applying the above transformations to all keypoints in M, ensuring that the skeletal structure matches the proportions and orientation of M_{ref}.

• **Initial Frame Calibration**: To enhance temporal coherence, we introduce an initial frame calibration mechanism that aligns the reference image with the first pose frame. We leverage pose-guided synthesis to generate a calibrated reference that matches the starting pose while retaining identity features. Our calibration process includes: 1) Extracting structural features from the reference and initial pose, 2)

Applying pose-guided transformation to align the reference with the initial pose, and 3) Maintaining detailed features during the first reference frame generation.

(b) Audio Speech-driven Lip Enhancement Module. To achieve more accurate lip synchronization in CS gesture generation, we introduce an audio speech-driven lip enhancement module. It leverages pre-trained audio processing models to extract lip-related features from speech signals and generate corresponding lip shapes.

Given an audio speech sequence $A = \{a_t\}_{t=1}^{T}$, we employ a pre-trained wav2vec model to extract audio features:

$$f_t = \text{Wav2vec}(a_t). \tag{22}$$

These features are then mapped to lip landmark sequences through a lightweight network:

$$M_{lip}^{t} = \mathcal{G}_{\mathcal{A}}(f_t), \tag{23}$$

where $\mathcal{G}_{\mathcal{A}}(\cdot)$ consists of two fully connected layers that transform audio speech features into lip landmark coordinates. The generated lip landmarks are subsequently integrated into our pose alignment framework.

(c) Hand Refinement Module. To enhance the quality of generated hand regions in CS gesture videos, we propose a post-processing hand refinement approach. This refinement process aims to rectify potential anatomical inconsistencies while preserving the semantic meaning of CS gestures.

In the first stage, we employ a hand mesh reconstruction model to extract structural hand information:

$$H_M = \mathcal{R}(V \odot m_h), \tag{24}$$

where V is the generated CS video and m_h represents the hand region mask in the generated video frames, and $\mathcal{R}$ is a pre-trained hand mesh reconstruction model [36] that provides anatomically plausible hand mesh H_M.

Then the depth map D_h is rendered from the reconstructed mesh H_M with a pretrained model $\mathcal{G}_{\mathcal{D}}$:

$$D_h = \mathcal{G}_{\mathcal{D}}(H_M). \tag{25}$$

The second stage involves conditional inpainting guided by the depth map. The refined hand region is generated by :

$$V_h^* = \mathcal{G}_{\mathcal{H}}(V, D_h, m_h), \tag{26}$$

where $\mathcal{G}_{\mathcal{H}}$ is a ControlNet-based inpainting model, m_h is the hand region mask, and V_h^* is the refined result. To ensure temporal consistency, we blend the refined region with the original video to obtain the final refined CS video frame V_r:

$$V_r = m_h \odot V_h^* + (1 - m_h) \odot V. \tag{27}$$

IV. Experimental Setup

A. Large-scale Mandarin Chinese CS dataset

Previously, only two publicly accessible CS datasets existed: one in French[3] [50], comprising recordings of a single cuer delivering 238 sentences, and another in British English[4] [51], similarly featuring a single cuer reciting 97 sentences. To address the lack of Chinese CS data, we have recorded and constructed, for the first time, a large-scale **M**andarin **C**hinese **CS** dataset, named **MCCS-2024**, which includes contributions from seven CS cuers.

The dataset was built by first selecting 1000 text sentences based on the following principles: (1) The sentences cover common daily scenarios, including colloquial dialogues, formal expressions, and written language. (2) The materials aim to encompass all possible syllable combinations. Overall, our text collection spans 23 main topics, 72 subtopics, and includes the 399 most frequently used Mandarin syllables. The dataset consists of 1000 sentences, totaling 10,482 words, with an average of 10.5 words per sentence. The shortest sentence contains 4 words, while the longest comprises 25 words. Subsequently, we recorded CS videos from each of the seven cuers performing these 1000 sentences, resulting in a total of 7000 sentences.

All videos were captured using either a camera or a mobile phone in landscape mode. The seven cuers underwent systematic training to ensure accurate and fluent performance of Mandarin Chinese CS. It is important to note that the dataset was collected with the explicit consent of all participants and is suitable for open-source distribution.

B. Experimental Setting

During the training stage, we first pre-train the motion clip and subsequently adopt an end-to-end pipeline to train the latent diffusion model. The experiments are conducted using PyTorch, leveraging four *A6000* GPU cards for model training. In the inference phase, the latent diffusion model is utilized to generate CS gestures. The dataset is randomly divided into training and test sets in a $4:1$ ratio. The diffusion process consists of 1000 steps, with a training batch size of 128. The weights for the loss components are set to $\alpha = 1$, $\beta = 0.2$ and γ = 0.1.

C. Evaluation Metrics

• **Evaluation for the Generated CS Pose**. Traditional evaluation metrics for generated CS pose gestures include three categories: Percentage of Correct Keypoint (PCK) [52], Fréchet Gesture Distance (FGD) [22], Mean Absolute Joint Errors (MAJE) [22], and Mean Acceleration Difference (MAD) [22]. Additionally, to further assess the unique asynchronous dynamics between lip and hand movements in CS, we employ a novel metric, GAD, as detailed in Sec. III-C2, to evaluate the rhythmic synchronization of the generated gestures.

• **Evaluation for the Generated CS Video**. For a comprehensive evaluation of our method's effectiveness in CS video generation, we conducted extensive comparisons across multiple established metrics including Peak Signal-to-Noise Ratio (PSNR) [53], Structural Similarity Index Measure (SSIM) [54], L1, Learned Perceptual Image Patch Similarity (LPIPS) [55], Fréchet Inception Distance for Video (FID-VID) [56] and Fréchet Video Distance (FVD) [57]. These metrics collectively assess different aspects of video quality - from image-level quality (PSNR, SSIM, L1) to perceptual quality (LPIPS)

[3]https://zenodo.org/record/5554849\#.ZBBCvOxBx8Y
[4]https://zenodo.org/record/3464212\#.ZBBAJuxBx8Y

Methods	PCK (%)↑	FGD↓	MAJE (mm)↓	MAD (mm/s^2)↓	GAD (%)↑
Speech2Gesture [21]	36.84	19.25	61.26	3.97	66.8
GTC [22]	41.23	6.73	55.43	2.54	66.7
HA2G [49]	43.51	4.07	46.78	2.29	67.2
DiffGesture [23]	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 I

THE EXPERIMENTAL RESULTS ON THE MCCS-2024 DATASET IN COMPARISON TO STATE-OF-THE-ART (SOTA) METHODS. THE LABEL "GLOSS-PROMPT" SIGNIFIES THE INCLUSION OF A GLOSS KNOWLEDGE INFUSION MODULE. "WAVLM" INDICATES THAT MEL-FREQUENCY CEPSTRAL COEFFICIENT (MFCC) FEATURES WERE REPLACED BY REPRESENTATIONS EXTRACTED FROM THE LARGE-SCALE PRE-TRAINED SPEECH MODEL, WAVLM. "GLOSS-CLIP" REPRESENTS THE UTILIZATION OF GLOSS-BASED MOTION CLIP FINE-TUNING.

Method	PSNR ↑	SSIM ↑	L1 ↓	LPIPS ↓	FID-VID ↓	FVD ↓
Disco	16.98	0.732	5.78E-04	0.289	66.28	529.60
MagicAnimate	17.53	0.763	5.41E-04	0.271	52.57	417.09
Moore-AnimateAnyone	17.32	0.774	5.13E-04	0.285	47.16	401.80
Unianimate	17.95	0.791	4.86E-04	0.257	31.72	361.35
MimicMotion	17.87	0.759	4.61E-04	0.243	25.57	257.80
Ours	**18.48**	**0.806**	**3.74E-04**	**0.216**	**13.77**	**208.92**

TABLE II

QUANTITATIVE COMPARISONS WITH SOTAS ON MCCS-2024 FOR VIDEO GENERATION.

and video-level fidelity (FID-VID, FVD). By evaluating our approach against SOTA methods using this diverse set of metrics, we can thoroughly demonstrate its capabilities in generating high-quality CS motion video while maintaining both personalized consistency and fine-grained motion.

D. Compared Methods

Firstly, to evaluate the generated CS pose quality, we compare our approach with four recent gesture synthesis methods, *i.e.*, Speech2Gesture [21], Gestures from Trimodal Context (GTC) [22], HA2G [58], and DiffGesture [23].

Secondly, to evaluate the generated CS video quality, we compared our proposed AnyoneCue with several SOTA methods, including Disco [59], MagicAnimate [60], Animate Anyone [32], and UniAnimate [61]. The compared models were fine-tuned on MCCS-2024 dataset. Extensive experiments on both public human animation datasets and our proposed CS dataset demonstrate that AnyoneCue outperforms these competing methods in terms of personalized consistency and fine-grained details.

V. RESULTS AND ANALYSIS

A. Quantitative Results and Analysis

In this section, we show the comparisons with SOTA for CS pose generation and CS video generation, respectively.

1) Comparison with SOTA for CS Pose Generation: We consider DiffGesture as the SOTA method among these approaches, as it achieves the best performance on the TED Gesture datasets [62].

Table I presents a comprehensive comparison between our method and previous approaches on the MCCS-2024 dataset. Our proposed method, GlossDiff, achieves the best results in PCK, MAJE, MAD, and GAD metrics, with most metrics showing a significant improvement over the reference systems. These results highlight the superior quality of fine-grained gesture generated by our system. The only exception is a slightly lower FGD score than the SOTA method, although it still outperforms all other reference methods. Notably, our method's PCK values are substantially higher than those of other methods, demonstrating its effectiveness in fine-grained generation. Furthermore, our method excels in rhythm performance, achieving the highest GAD values. This superiority of the GAD metrics underscores our method's ability to effectively capture the rhythmic dynamics in CS gestures.

2) Comparison with SOTA for CS Video Generation: A comprehensive quantitative evaluation was conducted to compare our method with several SOTA approaches on the MCCS-2024 dataset, as shown in Table II. Our method consistently outperforms previous approaches in multiple metrics, achieving higher scores in PSNR (18.48), SSIM (0.806) and L1 (3.74E-04). Specifically, compared to the strongest baseline (*i.e.*, Unianimate), our method shows substantial improvements of 0.53dB in PSNR and 0.015 in SSIM. The significant reduction in LPIPS (0.216), FID-VID (13.77), and FVD (208.92) metrics further demonstrates our method's capability to generate high-quality videos with better perceptual similarity and temporal consistency. Notably, our approach achieves a 19% improvement in FVD compared to MimicMotion (257.80), indicating enhanced temporal coherence. The consistent superior performance across both spatial quality metrics (PSNR, SSIM) and temporal metrics (FID-VID, FVD) validates the effectiveness of our proposed method in generating high-fidelity videos while maintaining temporal consistency. These comprehensive results demonstrate that our method can better capture the dynamic nature of video content while preserving frame-level quality, representing a significant advancement in video generation.

3) Ablation Study: We conduct an ablation study for the three modules, as shown in Table I. The term "Gloss-prompt"

represents the integration of the Gloss Knowledge Infusion Module. "WavLM" refers to the use of features extracted from the pre-trained large-scale speech model WavLM, replacing conventional MFCC features. "Gloss-CLIP" indicates the incorporation of Gloss-based Motion CLIP Fine-tuning. The results reveal that removing any module leads to a decrease in performance metrics, underscoring the effectiveness of each module within our framework. Specifically, the absence of the Gloss-prompt and Gloss-CLIP modules reduces PCK by 3.11% and 0.82%, respectively, emphasizing their crucial role in fine-grained generation.

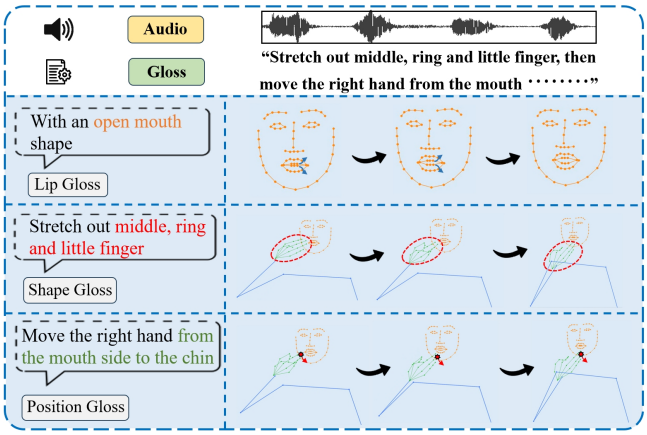

Fig. 6. The visualization of fine- grained Gloss and the corresponding generated gesture.

B. Qualitative Results and Analysis

1) Fine-grained Gesture Generation with Gloss Prompts: Fig. 6 illustrates fine-grained hand gestures generated with gloss prompts, where each row depicts the detailed gloss of different body parts along with their corresponding gesture sequences. Arrows are used to indicate lip movement trends, red circles highlight finger shape transformations, and red stars denote hand position shifts, including their movement directions. The first row demonstrates the expansion of the lips' contour in response to the gloss input. The second row focuses on detailed shape changes synchronized with detailed finger gloss. In the third row, subtle hand position shifts are observed, marked by red stars moving from near the mouth to the chin area, showcasing our method's capability to utilize detailed gloss for guiding CS gesture generation.

2) Distribution of Fine-grained Gesture Features: We employed t-SNE [63] for dimensionality reduction to visualize the generated CS gestures in the feature space. Frames were uniformly sampled from the generated CS sequences, and hand gesture features corresponding to the text were extracted. As illustrated in Fig. 5, the MCCS-2024 dataset utilizes 8 distinct finger shapes to represent the 24 consonants of the Chinese language, along with 5 hand positions to denote the 16 vowels. In the left portion of Fig. 7, eight distinct clusters are observed, with each cluster corresponding to a specific set of finger shapes (where each color represents a different consonant group). Clusters that are closer in distance exhibit similar finger shapes, such as shape 8 and shape 6, as well as shape 2 and shape 7. This visualization confirms the effectiveness of our method in capturing the fine-grained semantics of CS hand and finger shapes. On the right side of Fig. 7, differences in features among hand positions are evident, but the clusters exhibit more overlap, indicating that they are less distinctly separated at the feature level compared with finger shapes.

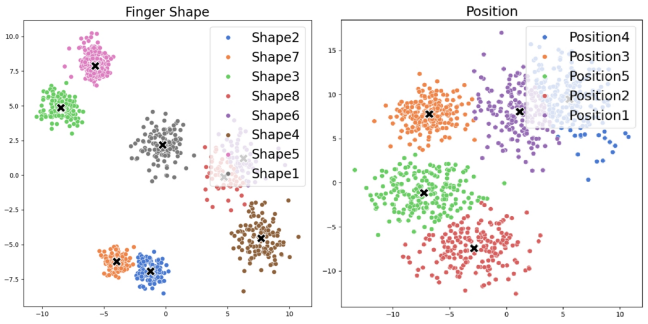

Fig. 7. The t-SNE clustering results are visualized to illustrate the separation of eight consonant groups (based on finger shape) and five vowel groups (based on hand position). Each group is represented by a distinct color.

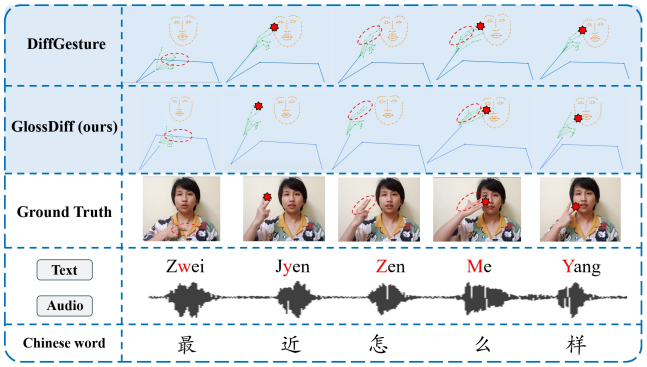

Fig. 8. The visualization result of the generated gestures compared to the SOTA method.

3) Comparative Analysis of CS Gesture Poses with SOTA: Fig. 8 presents a visual comparison between our method and the SOTA method (*i.e.*, DiffGesture). This comparison includes the corresponding audio speech, text, and the ground truth video frames for the gestures. Phonemes are highlighted in red, while red stars and circles are used to mark hand position and finger shapes, respectively.

Our method demonstrates a notable improvement in gesture accuracy, especially in fine-grained details. For instance, in

Fig. 9. Visualization result of generated hand in CS videos. It shows that the hand refinement module can effectively improve the hand details and accuracy of hand shapes.

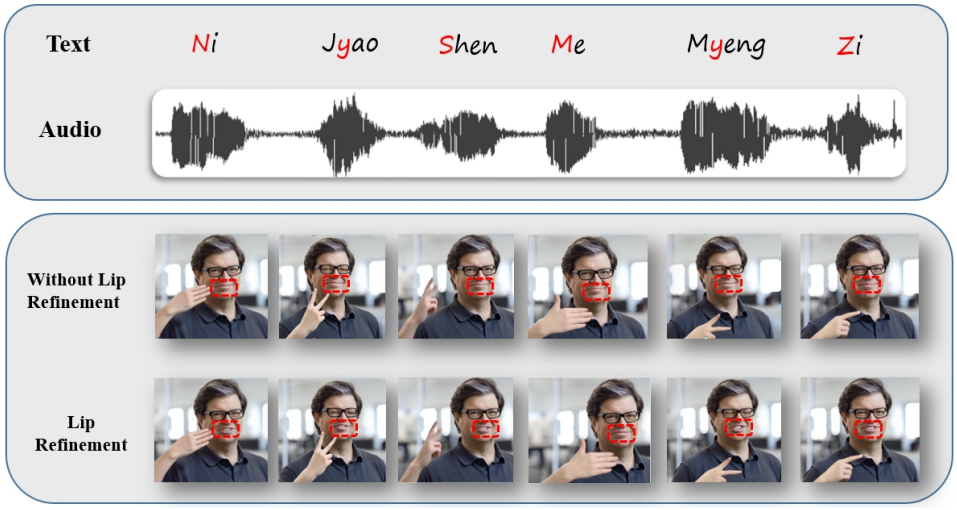

Fig. 10. Visualization result of generated CS video frames. It shows that the lip enhancement module can effectively improve the synchrony of lip shapes with the corresponding audio speech.

the first column, the index finger shape generated by our method is more precise compared to the SOTA method. In the second column, our method correctly positions the hand beside the face, whereas the SOTA method places it near the eye. The fourth column highlights our method's superior precision in thumb positioning and overall gesture alignment with the ground truth, reflecting stronger adherence to CS rules and enhanced accuracy in fine-grained details.

4) Visualization of Generated CS Videos: In Fig. 9, we present a visualization comparison between the baseline and hand-refined results across six paired examples. In the first sample, the refined version shows an improved definition of finger joints and more natural curvature in the open palm gesture, with clearer separation between fingers compared to the baseline's slightly rigid representation. The second frame demonstrates enhanced articulation of the hand position, where the refined model better captures the natural flex of fingers in a similar open-hand pose. In the third example, the peace sign gesture exhibits more precise finger alignment and realistic joint positioning in the refined version, correcting the minor distortions visible in the baseline result. The fourth frame shows a significant improvement in capturing the raised hand position, with more accurate finger spacing and natural hand curvature. In the fifth and sixth examples, the refined model achieves a better definition of finger proportions and realistic hand contours, particularly evident in the angular hand positions. These consistent improvements across all samples demonstrate that our hand refinement module effectively enhances the anatomical accuracy and natural appearance of hand gestures, successfully addressing the limitations in hand detail generation observed in the baseline model.

Fig. 10 demonstrates the effectiveness of our lip enhancement module in improving lip synchronization with audio speech through a comparative analysis using a Chinese phrase "Ni Jyao Shen Me Myeng Zi" in Chinese ("What's your name?" in English). The visualization presents two rows of video frames corresponding to models without and with lip refinement alongside the audio waveform and text representation of each syllable. The refined version exhibits significant improvements in several critical aspects of lip synchronization. Notably, the lip shapes show enhanced precision in matching specific phonemes, particularly evident in the transitions between syllables like "Ni" and "Jyao". The temporal alignment between lip movements and audio waveform peaks demonstrates superior coherence in the refined model, while the articulation detail shows an improved definition of lip contours and mouth shapes during various vowel and consonant combinations. Furthermore, the refined version maintains more consistent and natural lip movements throughout the sequence, effectively eliminating the subtle misalignments observed in the baseline version. These improvements collectively validate that our lip enhancement module successfully achieves more realistic and accurate lip-sync performance, effectively synchronizing visual lip movements with the corresponding audio speech.

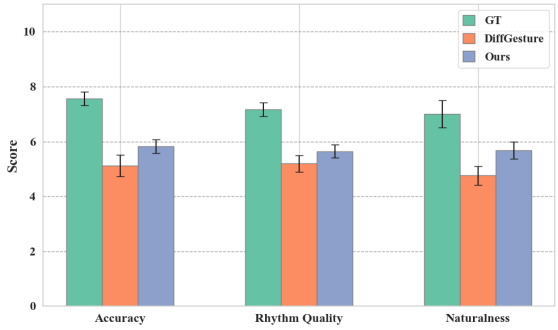

Fig. 11. User study results of the ground truth (GT), current SOTA (DiffGesture) and our method (GlossDiff).

C. User Study

1) User Study on CS Pose Generation: We conducted a user study to assess the quality of CS gestures generated by our method compared to the SOTA method and the ground truth. The study included 10 groups of videos, each containing a

ground truth CS gesture video, videos generated by the current SOTA method (DiffGesture), and videos generated by our method (GlossDiff). All videos were randomly shuffled. Ten participants trained in CS were asked to evaluate CS gesture videos based on three criteria: accuracy, rhythm quality, and naturalness, with each criterion scored on a scale from 0 to 10 (higher scores indicate better performance). The average scores and confidence intervals were calculated for each case.

As illustrated in Fig. 11, our method outperformed the current SOTA DiffGesture across all three metrics, achieving results closer to the ground truth. This highlights our method's capability to generate more accurate and natural CS gestures, particularly in rhythm quality, which is attributed to the proposed ARM. Our approach significantly surpasses DiffGesture in accuracy, demonstrating its effectiveness in fine-grained gesture generation.

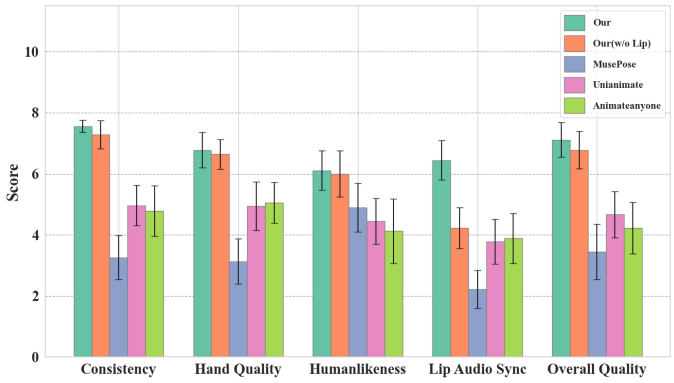

Fig. 12. User study results of the generated videos of our methods and SOTA.

2) User Study on CS Video Generation: To comprehensively evaluate the quality of our generated CS videos, we conducted a second user study comparing our method against UniAnimate (current SOTA), AnimateAnyone, MusePose and ablated versions of our model. We prepared 6 groups of videos, containing the same content generated by UniAnimate, AnimateAnyone, MusePose and our full method (GlossDiff + FPA-Diff) along with our ablated versions: without lip enhancement. The study involved 13 participants who rated each video on five aspects using a scale of 1-10: Consistency, hand integrity (completeness and clarity of finger shapes), humanlikeness and audio speech synchronization (alignment between lip movements and speech), and overall quality (visual quality and consistency).

The user study results summarized in Fig. 12, clearly demonstrate that our full model (incorporating both GlossDiff and FPA-Diff) outperforms all other methods across all evaluation metrics. In particular, we observe a substantial advantage in hand integrity (Hand Quality) and lip-audio synchronization (Lip Audio Sync), which are notably challenging aspects of character-driven video generation. Removing the lip enhancement module led to a noticeable drop in Lip Audio Sync scores, confirming the effectiveness of our approach. The variance analysis further shows that our model provides more stable and consistent results, delivering high-quality, realistic outputs. These results highlight our method's significant advantage in generating consistent and fine-grained CS videos.

VI. Discussion and Conclusion

A. Limitation and Discussion

The proposed AnyoneCue excels in generating fine-grained and personalized CS videos. It was reported that hand gestures and lip-reading are sufficient to accurately convey semantic information [4], [5]. Nevertheless, from a human-centric perspective, we believe that the generation of subtle facial expressions remains an essential component. Future work could integrate emotional features into the framework, potentially leveraging affective computing techniques to enhance the expressiveness of the generated videos. This would further bridge the gap between synthetic and natural human communication, making the system more expressive for users.

Besides, while our methodology is theoretically applicable to CS across different languages, the current scarcity of open-source CS datasets for English and French (*i.e.*, only 238 French CS videos [48] and 97 English CS videos publicly available [51]) severely limits the effectiveness of training GlossDiff and diffusion models proposed in this work. This data insufficiency may lead to suboptimal performance, including imprecise gloss prompts, unrealistic hand articulations, or inconsistent lip synchronization. To address these limitations and fully validate the cross-linguistic robustness of our approach, expanding data collection efforts to include CS datasets with more diverse languages represents a critical future research direction.

B. Conclusion

In this work, the proposed AnyoneCue framework represents a significant advancement in the automated generation of CS videos based on audio speech and text, offering a fine-grained and personalized solution that adheres to specific CS coding rules. By introducing the CS gloss as a novel action parsing prompt, we effectively integrate additional linguistic knowledge, bridging the semantic gap between gestures and spoken language. The proposed Pose-Refined Video Diffusion Model (PRV-DM) enhances the realism and precision of generated videos. Extensive experiments and user studies confirm the efficacy of the AnyoneCue, establishing it as the first end-to-end deep learning approach for CS video generation.

References

[1] 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.

[2] A. Fernandez-Lopez, O. Martínez, and M. Sukno, "Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database," in *International Conference on Automatic Face & Gesture Recognition (FG)*, 2017.

[3] L. Liu and L. Liu, "Cross-modal mutual learning for cued speech recognition," *ICASSP*, 2023.

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

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

[6] J. 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.

[7] K. S. Liddell and E. R. Johnson, "American sign language: The phonological base," *Sign Language Studies*, pp. 195–278, 1989.

[8] R. Timothy, "Linguistics of american sign language: An introduction," *Studies in Second Language Acquisition*, vol. 25, no. 1, p. 157–158, 2003.

[9] 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.

[10] S. Cox, M. Lincoln, J. Tryggvason, M. Nakisa, M. Wells, M. Tutt, and S. Abbott, "Tessa, a system to aid communication with deaf people," in *Proceedings of the Fifth International ACM Conference on Assistive Technologies*, 2002, pp. 205–212.

[11] D. Power, M. R. Power, and B. Rehling, "German deaf people using text communication: Short message service, tty, relay services, fax, and e-mail," *American Annals of the Deaf*, vol. 152, no. 3, pp. 291–301, 2007.

[12] L. Liu, G. Feng, and B. Denis, "Automatic temporal segmentation of hand movements for hand positions recognition in french cued speech," in *ICASSP*, 2018.

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

[14] L. Liu, G. Feng, B. Denis, and X.-P. 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.

[15] W. Lei, L. Liu, and J. Wang, "Bridge to non-barrier communication: Gloss-prompted fine-grained cued speech gesture generation with diffusion model," *arXiv preprint arXiv:2404.19277*, 2024.

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

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

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

[19] S. Sankar, M. Lenglet, G. Bailly, D. Beautemps, and T. Hueber, "Cued speech generation leveraging a pre-trained audiovisual text-to-speech model," *arXiv:2501.04799*, 2025.

[20] T. Ao, Q. Gao, Y. Lou, B. Chen, and L. Liu, "Rhythmic gesticulator," *ACM Transactions on Graphics*, vol. 41, no. 6, pp. 1–19, 2022.

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

[22] 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.

[23] L. Zhu, X. Liu, X. Liu, R. Qian, Z. Liu, and L. Yu, "Taming diffusion models for audio-driven co-speech gesture generation," in *CVPR*, 2023.

[24] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, "Text2Sign: Towards sign language production using neural machine translation and generative adversarial networks," *IJCV*, 2020.

[25] S. Ben, N. Camgoz, and B. Richard, "Progressive transformers for end-to-end sign language production," in *ECCV*, 2020.

[26] S. Fang, L. Wang, C. Zheng, Y. Tian, and C. Chen, "SignLLM: Sign languages production large language models," *arXiv:2405.10718*, 2024.

[27] V. Baltatzis, R. A. Potamias, E. Ververas, G. Sun, J. Deng, and S. Zafeiriou, "Neural sign actors: A diffusion model for 3d sign language production from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1985–1995.

[28] L. Liu and L. Liu, "Cross-modal mutual learning for cued speech recognition," in *ICASSP*, 2023.

[29] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020.

[30] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano, "Human motion diffusion model," in *ICLR*, 2022.

[31] W. Zhao, L. Hu, and S. Zhang, "Diffugesture: Generating human gesture from two-person dialogue with diffusion models," in *ICMI*, 2023.

[32] L. Hu, "Animate anyone: Consistent and controllable image-to-video synthesis for character animation," in *CVPR*, 2024.

[33] S. Tan, B. Gong, X. Wang, S. Zhang, D. Zheng, R. Zheng, K. Zheng, J. Chen, and M. Yang, "Animate-X: Universal character image animation with enhanced motion representation," *arXiv:2410.10306*, 2024.

[34] H. Wei, Z. Yang, and Z. Wang, "Aniportrait: Audio-driven synthesis of photorealistic portrait animation," *arXiv:2403.17694*, 2024.

[35] S. N. Gowda, D. Pandey, and S. N. Gowda, "From pixels to portraits: A comprehensive survey of talking head generation techniques and applications," *arXiv:2308.16041*, 2023.

[36] W. Lu, Y. Xu, J. Zhang, C. Wang, and D. Tao, "Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting," in *ACM MM*, 2024.

[37] S. Narasimhaswamy, U. Bhattacharya, X. Chen, I. Dasgupta, S. Mitra, and M. Hoai, "Handiffuser: Text-to-image generation with realistic hand appearances," in *CVPR*, 2024.

[38] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black, "Monocular expressive body regression through body-driven attention," in *ECCV*, 2020.

[39] OpenAI and J. et.al, "GPT-4 technical report," 2023.

[40] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or, "Motionclip: Exposing human motion generation to clip space," in *ECCV*, 2022.

[41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[42] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[43] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *ICCV*, 2017.

[44] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv:2207.12598*, 2022.

[45] M. Petrovich, M. J. Black, and G. Varol, "Action-conditioned 3d human motion synthesis with transformer vae," in *ICCV*, 2021.

[46] M. Lebourdais, M. Tahon, A. Laurent, and S. Meignier, "Overlapped speech and gender detection with wavlm pre-trained features," *arXiv:2209.04167*, 2022.

[47] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *ICML*, 2023.

[48] L. Liu, G. Feng, D. Beautemps, and X.-P. 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.

[49] X. Liu, Q. Wu, H. Zhou, Y. Xu, R. Qian, X. Lin, X. Zhou, W. Wu, B. Dai, and B. Zhou, "Learning hierarchical cross-modal association for co-speech gesture generation," in *CVPR*, 2022.

[50] L. Liu, H. Thomas, G. Feng, and B. Denis, "Visual recognition of continuous cued speech using a tandem cnn-hmm approach." in *INTERSPEECH*, 2018.

[51] L. Liu, J. Li, G. Feng, and X.-P. S. Zhang, "Automatic detection of the temporal segmentation of hand movements in british english cued speech." in *INTERSPEECH*, 2019.

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

[53] A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *ICPR*. IEEE, 2010.

[54] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[55] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018.

[56] Y. Balaji, M. R. Min, B. Bai, R. Chellappa, and H. P. Graf, "Conditional gan with discriminative filter generation for text-to-video synthesis." in *IJCAI*, 2019.

[57] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv:1812.01717*, 2018.

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

[59] T. Wang, L. Li, K. Lin, Y. Zhai, C.-C. Lin, Z. Yang, H. Zhang, Z. Liu, and L. Wang, "Disco: Disentangled control for realistic human dance generation," in *CVPR*, 2024.

[60] Z. Xu, J. Zhang, J. H. Liew, H. Yan, J.-W. Liu, C. Zhang, J. Feng, and M. Z. Shou, "Magicanimate: Temporally consistent human image animation using diffusion model," in *CVPR*, 2024.

[61] X. Wang, S. Zhang, C. Gao, J. Wang, X. Zhou, Y. Zhang, L. Yan, and N. Sang, "UniAnimate: Taming unified video diffusion models for consistent human image animation," *arXiv:2406.01188*, 2024.

[62] 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.

[63] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.