Article

DiT-Gesture: A Speech-Only Approach to Stylized Gesture Generation

Fan Zhang [1,2,†], Zhaohan Wang [3,†], Xin Lyu [3,†], Naye Ji [2,*,†], Siyuan Zhao [1] and Fuxing Gao [2]

[1] The Faculty of Humanities and Arts, Macau University of Science and Technology, Macau 999078, China; fanzhang@cuz.edu.cn (F.Z.); 2109853jai30001@student.must.edu.mo (S.Z.)

[2] The College of Media Engineering, Communication University of Zhejiang, Hangzhou 310018, China; fuxing@cuz.edu.cn

[3] The School of Animation and Digital Arts, Communication University of China, Beijing 100024, China; 2022201305j6018@cuc.edu.cn (Z.W.); lvxinlx@cuc.edu.cn (X.L.)

[*] Correspondence: jinaye@cuz.edu.cn

[†] These authors contributed equally to this work.

Abstract: The generation of co-speech gestures for digital humans is an emerging area in the field of virtual human creation. Prior research has progressed by using acoustic and semantic information as input and adopting a classification method to identify the person's ID and emotion for driving co-speech gesture generation. However, this endeavor still faces significant challenges. These challenges go beyond the intricate interplay among co-speech gestures, speech acoustic, and semantics; they also encompass the complexities associated with personality, emotion, and other obscure but important factors. This paper introduces "DiT-Gestures", a speech-conditional diffusion-based and non-autoregressive transformer-based generative model with the WavLM pre-trained model and a dynamic mask attention network (DMAN). It can produce individual and stylized full-body co-speech gestures by only using raw speech audio, eliminating the need for complex multimodal processing and manual annotation. Firstly, considering that speech audio contains acoustic and semantic features and conveys personality traits, emotions, and more subtle information related to accompanying gestures, we pioneer the adaptation of WavLM, a large-scale pre-trained model, to extract the style from raw audio information. Secondly, we replace the causal mask by introducing a learnable dynamic mask for better local modeling in the neighborhood of the target frames. Extensive subjective evaluation experiments are conducted on the Trinity, ZEGGS, and BEAT datasets to confirm WavLM's and the model's ability to synthesize natural co-speech gestures with various styles.

Keywords: gesture generation; gesture synthesis; cross-modal; speech driven; diffusion model; transformer; DMAN

check for updates

Citation: Zhang, F.; Wang, Z.; Lyu, X.; Ji, N.; Zhao, S.; Gao, F. DiT-Gesture: A Speech-Only Approach to Stylized Gesture Generation. *Electronics* **2024**, *13*, 1702. https://doi.org/10.3390/electronics13091702

Academic Editors: Kibum Kim and Huawei Tu

Received: 12 April 2024
Revised: 24 April 2024
Accepted: 26 April 2024
Published: 27 April 2024

1. Introduction

Recently, The utilization of 3D virtual human technology has witnessed a notable surge in popularity, paralleling the emergence of the metaverse. This technology finds extensive applications in various domains of real-world society, encompassing animation, gaming, human–computer interaction, VTuber platforms, virtual guidance systems, digital receptionists, presenters, and various other areas.

To create realistic and engaging virtual humans, a crucial objective is the integration of non-verbal (co-speech) gestures that appear natural and align with human communication patterns. Although motion capture systems have been developed to fulfill this requirement, their implementation necessitates specialized hardware, dedicated space, and trained actors, resulting in significant expenses. As an alternative, automatic gesture generation presents a cost-effective approach that eliminates the need for human intervention during the production phase. Among the potential solutions, speech-driven gesture generation emerges as a viable option. Nevertheless, a major challenge in this endeavor lies in effectively matching and synchronizing relevant gestures with the input speech, given the inherent complexities of cross-modal mapping, many-to-many relationships, and the diverse and ambiguous nature of gesture patterns. Furthermore, the same utterance often elicits distinct gestures at different temporal instances, even when uttered by the same or different individuals [1].

The close relationship between gestures and the acoustic signals of speech is widely acknowledged in scholarly discourse. Consequently, considerable research [2–5] has been dedicated to extracting pertinent features from speech audio signals, such as mel-frequency cepstrum coefficients (MFCCs). These extracted features serve as input for neural networks, thereby facilitating the generation of corresponding co-speech gestures. However, it is important to recognize that gestures are not exclusively tied to speech acoustic features. Rather, they exhibit intricate associations with various other factors, including well-established aspects such as personalities, emotions, and speech context, among others, as well as a multitude of unknown variables. This intricate interplay presents significant challenges in the pursuit of generating co-speech gestures that possess enhanced naturalness and realism.

Prior research [6–10] has explored the use of manual labels and diverse feature inputs to facilitate the synthesis of personalized gestures. However, these methodologies exhibit a pronounced reliance on various unstructured feature inputs and necessitate complex multimodal processing. This dependency poses a significant barrier to the practical implementation and broader adoption of virtual human technologies.

Due to the multifaceted nature of speech signals, encompassing aspects such as speaker personality, acoustic, etc., our objective is to exclusively extract the signal originating from raw speech audio while abstaining from processing various modalities concurrently.

The recent emergence of expansive pre-training models offers a promising opportunity to significantly improve pre-training outcomes. These models can potentially be effectively transferred to various subsequent tasks. In our exploration, we have identified WavLM, a noteworthy system that learns from a massive number of unlabeled speech data to acquire universal speech representations. WavLM demonstrates remarkable adaptability across diverse speech-processing tasks, from Automatic Speech Recognition (ASR) to non-ASR, further validating its efficacy and potential for practical applications.

A further challenge within this domain pertains to achieving a high degree of gesture-speech synchronization while maintaining naturalness in gestures. Recent advancements have centered on employing generative models, marking a pivotal shift in methodologies that has significantly enhanced the efficiency and flexibility of gesture generation technologies. Prominent examples of such innovative efforts include Style Gesture (SG) [5], Diffuse Style Gesture (DSG) [7], Diffuse Style Gesture+ (DSG+) [8], and Diffmotion [2]. Nevertheless, these approaches often grapple with challenges related to achieving insufficient or excessive correlation between gesture and speech, compromising the naturalness of the generated gestures.

This paper presents DiT-Gestures, a novel transformer, and diffusion-based probabilistic architecture specifically developed to generate speech-driven gestures. The core innovation of DiT-Gestures lies in its ability to automatically extract these stylistic attributes directly from speech audio. This capability ensures that the generated gestures are synchronized with speech in terms of timing and rhythm and rich in contextual and emotional subtleties.

Furthermore, we integrate the Dense Motion Attention Network (DMAN) module into the transformer architecture to enhance its ability to capture local gesture relationships. This integration allows the transformer to focus more precisely on short-term dependencies within the gesture data, improving the accuracy and realism of the generated movements. This enhancement is crucial to generating fluid and lifelike animations that accurately reflect the subtleties of human motion in response to speech.

Our contributions can be summarized as follows:

- We are involved in pioneering the use of the WavLM generative pre-trained transformer large model for extracting style features from raw speech audio features without any manual annotation.
- We extended our previous autoregressive Diffmotion model to a non-autoregressive variant known as DiT-Gestures. This extension encompasses a novel diffusion model, which adopts a transformer-based architecture that replaces the causal mask with a dynamic mask attention network (DMAN). The DMAN effectively enhances the adaptive modeling of local frames.
- Extensive subjective and objective evaluations reveal that our model outperforms the current state-of-the-art approaches. These results show the remarkable capability of our method in generating natural, speech-appropriate, and stylized gestures.

This research builds upon our prior architecture, Diffmotion [2]. However, this work extends the previous paper significantly by introducing novel features and improvements. Firstly, we introduce a non-autoregressive generative model that employs a transformer-based architecture. Unlike its predecessor, DiT-Gestures generates the entire sequence of full-body gestures instead of generating them frame by frame, resulting in more coherent and holistic gesture synthesis. Secondly, we enhance the feature extraction process by replacing the traditional mel-frequency cepstral coefficients (MFCCs) with the WavLM generative pre-trained transformer large model. Lastly, unlike Diffmotion, which generates redundancy gestures and necessitates post-processing techniques to mitigate jitter-induced inconsistencies in the gesture sequence, DiT-Gestures overcomes this limitation and produces more stable and refined gesture sequences.

The remainder of this paper is organized as follows: The related works about co-speech gesture generation are described in Section 2. Then, we elaborate on the DiT-Gestures schedule in Section 3. The experimental results of three baseline detection algorithms on our dataset are presented in Section 4. We conclude this paper in Section 5.

2. Related Work

The alignment of non-verbal communication, specifically co-speech gestures, with the communicative intent of virtual agents requires the establishment of a meaningful correspondence between the two modalities. The investigation of automated co-speech gesture generation, relying on speech information, can be broadly categorized into two primary domains: rule-based methods and data-driven approaches. In light of the notable success of deep learning techniques in various computer tasks, synthesizing co-speech gestures has shifted from rule-based approaches (extensively reviewed by Wanger et al. [11]) to data-driven approaches, particularly with the introduction of deep learning methodologies. Moreover, within the realm of deep learning, a distinction exists between deterministic and generative models. This discussion will briefly focus on generative models for speech-driven gesture generation.

2.1. Data-Driven Generative Approaches

In real-life scenarios, the same utterance can be accompanied by varying gestures, even when repeated by the same speaker at different time points, highlighting the lack of coherence in gesture production. This presents a significant challenge for deterministic models, which struggle to capture the extensive variation between speech and gestures. Consequently, there has been a shift in research focus from deterministic models to probabilistic generative models. Generative adversarial networks (GANs) have shown promise in generating persuasive random samples. Accordingly, Ylva et al. [12] attempted to explore GANs [13] with multiple discriminators to convert speech into 3D gesture motion. However, this approach requires manual dataset annotation, and the results still lack realism. In contrast, Wu et al. [14] verified the effectiveness of conditional and unrolled GANs, showing that they outperformed existing deterministic models.

Normalizing flows [15], built on unsupervised learning algorithms such as NICE [16] and RealNVP [17], are capable of constructing complex distributions and approximating the true posterior distribution. Impressively, Alexanderson et al. [5] demonstrated the effectiveness of a network called MoGlow, based on normalizing flows, in generating a diverse set of plausible gestures given the same input speech signal, without the need for manual annotation. Li et al. [18] employed a conditional variational autoencoder (VAE) model to capture the strong correlation between audio and motion, enabling the random generation of diverse motions. Taylor et al. [3] extended normalizing flows by combining them with a variational autoencoder called Flow-VAE. Their evaluation demonstrated that this approach produces expressive body motion close to the ground truth while utilizing fewer trainable parameters. However, it should be noted that normalizing flows requires the imposition of topological constraints on the transformation [16,17]. Furthermore, the MoGlow method employs an LSTM architecture, necessitating the generation of the entire sequence of gestures frame by frame in an autoregressive manner. This approach inevitably results in an obvious increase in the overall generation time.

Diffusion models [19,20] represent an alternative class of generative models that leverage a Markov chain to transform a simple distribution into a complex data distribution gradually. These models can be efficiently trained by optimizing the variational lower bound (ELBO). They have been successfully applied in image synthesis [19] and multivariate probabilistic time-series forecasting [21], with connections to denoising score matching [22]. Our previous work proposed DiffMotion [2], a diffusion model-based framework with an LSTM architecture that generates co-speech gestures frame by frame [2]. Furthermore, Alexanderson et al. [4] adapted the DiffWave architecture, replacing dilated convolutions with Conformers [23] to enhance the modeling power and incorporating classifier-free guidance to adjust the strength of stylistic expression. Another diffusion model-based framework, called GestureDiffuCLIP [24], learns a latent diffusion model to generate high-quality gestures and incorporates large-scale Contrastive Language–Image Pre-training (CLIP) representations for style control. However, this system requires learning a joint embedding space between corresponding gestures and transcripts by using contrastive learning, which provides semantic cues for the generator and effective semantic loss during training.

2.2. Condition Encoding Strategy

Co-speech gesture generation systems have recently incorporated conditional information as input, including audio, transcripts, style labels, and other relevant factors. This approach allows the system to consider additional contextual information during the gesture generation process. By incorporating these conditional inputs, the system can generate gestures more aligned with the speech content, style, and other specified conditions, leading to more contextually appropriate and expressive gestures.

2.2.1. Audio Representation

The most suitable audio representation is an open research question [25]. One of the most common audio speech representations chosen in previous work is mel-frequency cepstral coefficients (MFCCs) [2,3,5], which better approximates how humans perceive sounds. Another approach, ZeroEGGS [10], combines the log amplitude of the spectrogram, the mel-frequency scale, and the energy of the audio as speech audio features. While in GestureDiffuCLIP [24], the speech audio ($A = [a_i]_{i=1}^{L}$) is parameterized as a sequence of acoustic features, where each a_i encodes the onsets and amplitude envelopes that reflect the beat and volume of speech, respectively. Although these approaches have provided impressive results, these approaches only represent acoustic information; there is scope for more descriptive features.

2.2.2. Style Control

For the creation of style-specific gestures, Diffuse Style Gesture [7] and Diffuse Style Gesture+ [8] utilize discrete labels to direct the stylistic attributes of the gestures produced. Considering that human emotions are more accurately depicted on a continuous spectrum [26,27] and arise from a complex interplay of fuzzy factors, reliance on discrete emotion labels may oversimplify the gesture generation process, potentially curtailing the expressiveness and subtlety of the resulting gestures. To overcome these constraints, Ghrobani et al. [10] developed ZeroEGGS, a model that employs example motion clips to influence the style of gestures. Although this approach allows for zero-shot capabilities, it still requires the use of sample animation clips.

To address the challenges posed by the reliance on structured feature inputs and complex multimodal processing in previous gesture generation models, we turned to WavLM [28], a cutting-edge pre-trained model. WavLM leverages a vast dataset of unlabeled raw speech audio to learn comprehensive speech representations. Its effectiveness has been extensively validated across a wide array of speech-processing tasks in Automatic Speech Recognition (ASR) and non-ASR applications such as speaker diarization, speech separation, speech recognition, and emotion recognition. The versatility and robust performance of WavLM make it an ideal candidate for enhancing the synthesis of speech-driven gestures.

This study investigates the potential of leveraging WavLM to generate stylized gestures driven by raw speech audio inputs without any discrete style labels. By harnessing the detailed speech representations learned by WavLM, we aim to create a gesture synthesis model that can interpret and convert spoken language into personalized gestures.

Furthermore, to bolster the model's capability to capture the nuanced dynamics between speech and gestures, we integrate a DMAN transformer architecture. This architecture enhances the model's ability to process and interpret the complex temporal and spatial relationships inherent in speech-driven gesture generation. The DMAN transformer, by handling sequential data more effectively, contributes significantly to improving the fidelity and naturalness of the generated gestures, thus promising a more realistic and responsive virtual human interaction experience.

3. Proposed Approach

The task in this paper is to generate a sequence of human poses $x_{1:T}$ given a raw speech audio waveform $a_{1:T}$ for the same time instances.

3.1. Problem Formulation

First, we define the co-speech gesture generation problem. We denote the gesture features and the acoustic signal by $x^0 = x^0_{1:T} \in [x^0_1, \ldots, x^0_t, \ldots, x^0_T] \in \mathbb{R}^{T \times D}$ and $a = a_{1:T} \in [a_1, \ldots, a_t, \ldots, a_T] \in \mathbb{R}^T$, where $x^0_t = \mathbb{R}^D$ is the angle of the 3D skeleton joints at frame t, D indicates the number of channels of the skeleton joints, the superscript represents the diffusion time step, a_t is the current subsequence audio waveform signal at frame t, and T is the sequence length. Let $p_\theta(\cdot)$ denote the Probability Density Function (PDF), which aims to approximate the actual gesture data distribution $p(\cdot)$ and allows for easy sampling. We are tasked with generating the whole sequence of pose $x \sim p_\theta(\cdot)$ in a non-autoregressive manner according to its conditional probability distribution given acoustic signal a as the covariate:

$$x^0 \sim p_\theta\left(x^0|a\right) \approx p(\cdot) := p\left(x^0|a\right) \tag{1}$$

where $p_\theta(\cdot)$ aims to approximate $p(\cdot)$ trained by the denoising diffusion model. We discuss these two modules in detail in Section 3.5.

3.2. Model Architecture

We propose extending the diffusion models by utilizing a transformer architecture as the backbone. The architecture referred to as DiT-Gestures is depicted in Figure 1. The architecture comprises three main components: (1) a condition encoder, (2) a gesture encoder and a gesture decoder, (3) stacks of multi-head attention blocks with dynamic mask attention networks (DMANs), and (3) a final layer.

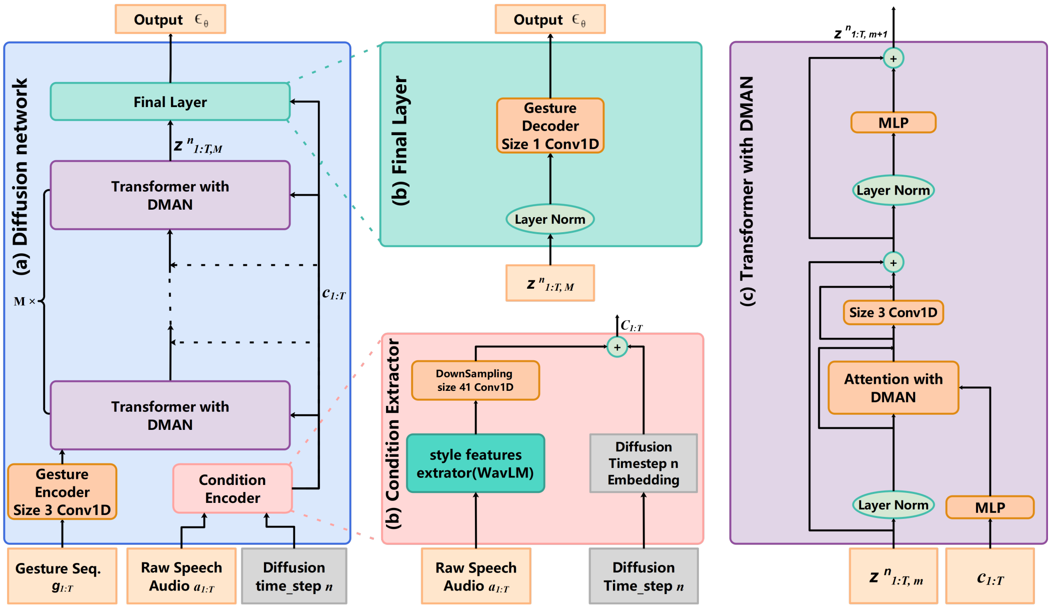

Figure 1. The architecture of DiT-Gestures. The model is a multi-block causal attention structure with a dynamic mask attention network (DMAN), a learnable mask matrix that can model localness adaptively. The condition encoder takes the raw speech audio features extracted by WavLM as input. It feeds them to the multi-head causal attention blocks to learn the relation between the co-speech gestures and the audio features, estimating the diffusion noise. (**a**) The whole architecture. (**b**) The condition encoder and final layer. (**c**) Multi-head attention with the DMAN.

The condition encoder is vital to extracting and embedding speech audio features by using the WavLM model while incorporating the embedded diffusion time step n. Simultaneously, the gesture encoder processes the input gesture sequence and transforms it into a latent representation. To capture the complex relationship between speech and gestures, the model employs stacks of multi-head attention blocks with the DMAN, which are stacked M times. This architecture enables the model to capture the dependencies and interactions between speech and gestures effectively. To ensure precise generation of the gesture sequence, the final layer incorporates a size 1 conv 1D for output noise prediction. This architectural design enhances the generation process, enabling the production of realistic and diverse gestures within the context of co-speech communication.

3.2.1. Condition Encoder

The condition encoder, illustrated in Figure 1b, converts raw audio input into a sequence of speech embedding space by the WavLM large-scale pre-trained model [19]. In our study, we integrated the WavLM model due to its ability to effectively handle the intricate nature of speech audio signals and its capacity to discern diverse speech styles, encompassing emotions, personalities, and other related aspects. The WavLM model has been extensively trained on a large-scale dataset consisting of unlabeled speech audio data, covering a wide range of tasks, Automatic Speech Recognition (ASR) and non-ASR tasks, such as speaker verification, speech recognition, paralinguistics, spoken content, and emotion recognition. The model was trained on a substantial amount of English audio, totaling 94k hours, featuring diverse speakers, topics, speaking styles, and scenarios. We believe that the WavLM model exhibits enhanced robustness and can extract various features from the speech audio data, including acoustic characteristics, speaker personalities, affective information, and more. The pre-training process equips the model with the ability to capture universal latent representations, denoted by Z_a, which encapsulate the essential information contained within the speech signals.

By leveraging the capabilities of the WavLM model, we aim to enhance the performance of co-speech stylized gesture generation tasks. This approach differs from the conventional methodology that relies solely on mel-frequency cepstral coefficients (MFCCs) for audio feature extraction, as observed in our previous Diffmotion model and other related studies. The WavLM model offers promising prospects because it can go beyond acoustic information and incorporate knowledge from various tasks, leading to more comprehensive and contextually relevant gesture generation.

A downsampling module is seamlessly integrated into the architecture to ensure alignment between each latent representation and the corresponding sequence of poses. This module takes the form of a Conv1D layer with a kernel size of 41, which means that every 41 lengths of target label output of WavLM is mapped to one frame of the gesture sequence. Its primary objective is to facilitate the synchronization of latent representations with the gesture sequence, enabling the generation of coherent and contextually relevant gestures.

3.2.2. Gesture Encoder and Decoder

We employ convolution 1D with a kernel size of 3 to embed the sequence of gestures from sequential data. convolution 1D operates by sliding the kernel across the input sequence and performing element-wise multiplication and summation to generate feature maps [14,23].

The selection of a kernel size of 3 is driven by its efficacy in capturing local patterns and dependencies within the sequence. It enables the model to consider neighboring elements and effectively capture short-term temporal dependencies [29]. This is particularly advantageous in gesture sequences, where adjacent frames often exhibit specific patterns or transitions contributing to overall motion dynamics.

By utilizing a kernel size of 3, we balance capturing fine-grained details and avoiding excessive parameterization. Smaller kernel sizes may overlook important contextual information, while larger ones can introduce more parameters and increase computational complexity [30]. Our experimental analysis found that using a kernel size of 1 resulted in animation jitter, underscoring the importance of an appropriate kernel size for gesture sequence extraction.

We employ convolution with a kernel size of 1 instead of a fully connected layer for several reasons in the context of gesture decoding. Convolution 1D with a kernel size of 1 enables us to capture local dependencies and interactions within the sequence while preserving the spatial dimensionality of the data. By convolving a 1D kernel with each position in the input sequence, the model can extract meaningful features and relationships between adjacent elements [31,32]. In contrast, a fully connected layer would necessitate connecting each input element to every output neuron, resulting in a significantly larger number of parameters and loss of the spatial structure of the data. Furthermore, using convolution with a kernel size of 1 provides flexibility in capturing local patterns and fine-grained details within the sequence. This enables the model to learn non-linear relationships between neighboring elements, which is particularly crucial in tasks such as gesture decoding, where short-term dependencies significantly contribute to understanding motion dynamics [33].

3.3. Transformer with DMAN

We present a novel transformer model designed to elucidate the intricate relationship between speech and gestures. This transformer is structured around multiple blocks of multi-head attention, integrated with the dynamic mask attention network (DMAN), as shown in Figure 2. A cross-attention mechanism is employed to forge a direct correlation between speech and gesture modalities. The introduction of the DMAN [34] augments the transformer's capacity for localness modeling, thereby enhancing its ability to focus on specific input segments for more precise gesture generation.

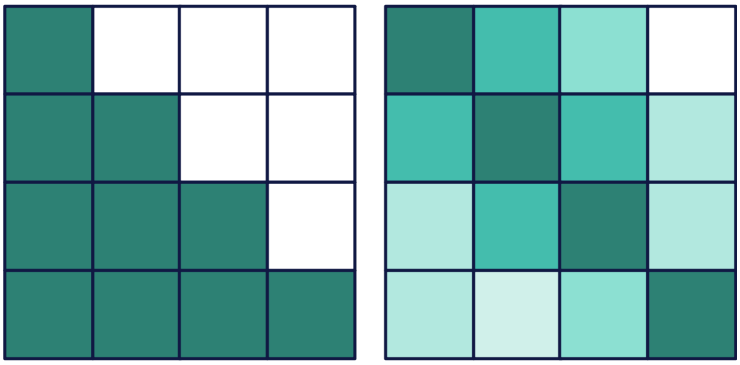

Figure 2. The masks of the causal mask (**left**) and the DMAN (**right**). Color that fades from deep cyan to white means that the values in the mask matrices decrease from 1 to 0.

In this study, we construct a distance-dependent mask matrix, denoted by SM. This matrix is designed to encapsulate the modeling of relationships between the frames of features, conditional upon the constraint that each frame of features only interacts with other frames within a bounded distance of b units. To achieve this, we define a specific function as follows:

$$\mathbf{SM}[t,s] = \begin{cases} 0, & |t-s| > b, \\ 1, & |t-s| \leqslant b, \end{cases} \tag{2}$$

where t and s represent the positions of the query and key, respectively. The value of $\mathbf{SM}[t,s]$ is then defined as the entry located at the intersection of the t-th row and s-th column within matrix $\mathbf{SM}$.

By leveraging the mask matrix (SM), our approach selectively considers frame features that are within b units of proximity and disregards those beyond this range. Although this static mask effectively prioritizes frames within a specified neighborhood, it notably lacks adaptability. This is because the optimal neighborhood size is not uniform across all query frames; different frames may derive varying degrees of benefit from their local semantic contexts. Additionally, the requirement for mask matrices to align with distinct attention heads and layers within mask attention networks (MANs) further complicates this issue. Drawing inspiration from the work of [34], we propose the dynamic mask attention network (DMAN), which supersedes the static mask matrix with a more flexible approach. By integrating considerations of query tokens, relative distances, attention heads, and network layers, we develop a dynamic mask function. This function transitions from the binary gating mechanism, traditionally represented by hard 0/1 values, to a graduated gating mechanism employing a sigmoid activation function, as delineated in Equation (3). This innovative modification facilitates a more nuanced and adaptable attention mechanism within MANs.

$$\mathbf{DM}_i^l[t,s] = \sigma(h_t^l W^l + P_{t-s}^l + U_i^l) \tag{3}$$

where s and t denote the positions of the key and query, respectively, while i represents the attention head, and l denotes the layer within the network. The term P_{t-s}^l is introduced as a parameterized scalar that accounts for the relative position between t and s. Similarly, U_i^l is designated for the i-th attention head. Furthermore, we introduce W^l as a matrix belonging to the real space $\mathbb{R}^{d\times 1}$, where d represents the dimensionality of the input features. Importantly, W^l, P_{t-s}^l, and U_i^l are identified as trainable parameters within the model. This parameterization enables the dynamic adjustment of the mask function, tailoring the model's focus to specific interactions between query and key positions, attention heads, and layers, thus enhancing the adaptability and effectiveness of the dynamic mask attention network (DMAN).

3.4. Final Layer

The architecture culminates in a gesture decoding layer, which incorporates 1 × 1 convolution (conv 1 × 1). This terminal layer is responsible for outputting the predicted noise.

3.5. Training and Inference with Denoising Diffusion Probabilistic Model

We have introduced the Denoising Diffusion Probabilistic Model (DDPM), a specific variant within the broader category of diffusion models [20], formalized by the equation $p_\theta := \int p_\theta(x^{0:N})dx^{1:N}$, where $x^1, \ldots, x^N$ denote the latent variables of identical dimensionality to data x^n at the n-th diffusion time stage. This model comprises two primary processes: the diffusion process and the generation process. During training, the diffusion process incrementally transitions the original data (x^0) into white noise (x^N) by optimizing a variational bound on the data likelihood. Conversely, during inference, the generation process reconstructs the data by inversely navigating this noise introduction via a Markov chain that employs Langevin sampling [35]. This methodology enables the generation of an entire gesture sequence in a non-autoregressive manner by drawing samples from the conditional data distribution. The Markov chains integral to both the diffusion and generation processes are delineated as follows:

$$\begin{aligned} p\left(x^n|x^0\right) &= \mathcal{N}\left(x^n; \sqrt{\bar{\alpha}^n}x^0, (1-\bar{\alpha}^n)I\right) \quad \textit{and} \\ p_\theta\left(x^{n-1}|x^n, x^0\right) &= \mathcal{N}\left(x^{n-1}; \tilde{\mu}^n\left(x^n, x^0\right), \tilde{\beta}^n I\right), \end{aligned} \tag{4}$$

where $\alpha^n := 1 - \beta^n$ and $\bar{\alpha}^n := \prod_{i=1}^n \alpha^i$. As shown by [19], β^n is a increasing variance schedule $\beta^1, \ldots, \beta^N$ with $\beta^n \in (0,1)$, and $\tilde{\beta}^n := \frac{1-\bar{\alpha}^{n-1}}{1-\bar{\alpha}^n}\beta^n$.

The training objective is to optimize the parameters θ that minimize the NLL via Mean Squared Error (MSE) loss between the true noise ($\epsilon \sim \mathcal{N}(0, I)$) and the predicted noise (ϵ_θ):

$$\mathbb{E}_{x^0_{1:T},\epsilon,n}[||\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}^n x^0} + \sqrt{1-\bar{\alpha}^n}\epsilon, a_{1:T}, n\right)||^2], \tag{5}$$

where ϵ_θ is a neural network, which uses input x^{n-1} , a, and n to predict ϵ, and the architecture is shown in Figure 1a. The complete training procedure is outlined in Algorithm 1.

Algorithm 1: Training for the whole sequence gestures.

```
Input: data $x^0_{1:T} \sim p(x^0|a_{1:T})$ and $a_{1:T}$
repeat
    Initialize $n \sim \mathrm{Uniform}(1, \ldots, N)$ and $\epsilon \sim \mathcal{N}(0, I)$
    Take gradient step on
        $\nabla_\theta||\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_n}x^0_{1:T} + \sqrt{1-\bar{\alpha}_n}\epsilon, a_{1:T}, n\right)||^2$
until converged;
```

After training, we expect to use variational inference to generate new gestures matching the original data distribution ($x^0 \sim p_\theta(x^0, a)$). We follow the sampling procedure in Algorithm 2 to obtain a sample x^0_t of the current frame. σ_θ is the standard deviation of $p_\theta(x^{n-1}|x^n)$. We choose $\sigma_\theta := \tilde{\beta}^n$.

In the inference phase, our approach has evolved from the methodology detailed in our previous work [2], where concatenated data comprising past poses x^0 and acoustic features a were utilized as inputs. Instead, we now exclusively input raw audio into the condition encoder, specifically the WavLM model. The output from WavLM is then channeled directly into the diffusion model, which is tasked with generating the entire sequence of the accompanying gesture (x^0). This refined method underscores a streamlined process that leverages the sophisticated audio processing capabilities of WavLM to enhance the generation of gesture sequences, simplifying the input requirements and potentially improving the efficiency and accuracy of gesture synthesis.

Algorithm 2: Sampling $x^0_{1:T}$ via annealed Langevin dynamics.

```
Input: noise $x^N_{1:T} \sim \mathcal{N}(0, I)$ and raw audio waveform $a_{1:T}$
for $n = N$ to 1 do
    if $n > 1$ then
        $z \sim \mathcal{N}(0, I)$
    else
        $z = 0$
    end if
    $x^{n-1}_{1:T} = \frac{1}{\sqrt{\alpha^n}}\left(x^n_{1:T} - \frac{\beta^n}{\sqrt{1-\bar{\alpha}^n}}\epsilon_\theta\left(x^n_{1:T}, a_{1:T}, n\right)\right) + \sqrt{\sigma_\theta}z$
end for
Return: $x^0_{1:T}$
```

4. Experiments

To demonstrate the efficacy of our approach, we utilized three co-speech gesture datasets for training and inference with our model. All experiments in this study exclusively focused on generating 3D gestures involving the full body. This deliberate selection presented a more demanding task than generating solely upper- body motions, as it entails higher dimensionality in the output space and introduces significant visual challenges, such as addressing artifacts like foot skating and ground penetration.

4.1. Dataset and Data Processing

4.1.1. Datasets

Our system underwent both the training and evaluation phases on three distinguished speech–gesture datasets: Trinity [36], ZEGGS [10], and BEAT [37]. Each dataset is characterized by a unique focus: the Trinity dataset is dedicated to individual spontaneous speech, the ZEGGS dataset captures a broad spectrum of emotional expressions, and the BEAT dataset comprises personalized movements demonstrated by diverse individuals, as shown in Table 1.

Table 1 presents an overview of the three datasets (Trinity, ZEGGS, and BEAT).

Table 1. Overview of the three datasets.

Dataset	Total Time	fps	Rate	Audio Sample Character	Content
Trinity	244 min	60	44 kHz	1 male	Spontaneous speech on different topics
ZEGGS	135 min	60	48 kHz	1 female	19 different motion styles
BEAT	35 h	120	48 kHz	30 speakers	Speech on diverse content

4.1.2. Speech Audio Data Process

Within the Trinity dataset, audio recordings were initially captured at a sampling rate of 44 kHz. Conversely, for both the ZEGGS and BEAT datasets, audio was recorded at a higher sampling rate of 48 kHz. Given the foundational pre-training of the WavLM large model on speech audio specifically sampled at 16 kHz, we opted to uniformly resample all audio data across these datasets to align them with this lower frequency. This standardization facilitates compatibility with the WavLM model's parameters and optimizes our system's performance by ensuring consistent input data characteristics.

4.1.3. Gesture Data Process

We concentrated exclusively on full-body gestures, employing the data processing methods delineated by Alexanderson et al. [5] (https://github.com/simonalexanderson/StyleGestures (accessed on 25 April 2024)). Due to data quality and structure variability across motion datasets, we adapted our approach by selectively analyzing specific joints in each dataset. We excluded hand skeleton data from the Trinity gesture dataset because of their lower quality. Conversely, our ZEGGS and BEAT datasets analysis encompassed finger joints, which were standardized to the same set of joints as those in the Trinity dataset. All datasets capture both translational and rotational velocities to record the root's trajectory and orientation accurately. The data were uniformly downsampled to a frame rate of 20 fps to maintain consistency. To ensure a precise and continuous representation of joint angles, we employed the exponential map technique [38]. For training and validation, data were segmented into 20-s clips. We segmented the generated gesture sequences into 10-s clips for user evaluation to enhance the evaluation process's efficiency.

To streamline the processes of training, evaluation, and testing, we structured the dataset into segments, each with a fixed duration of 20 s. This methodology guarantees a uniform and manageable input size for the model, significantly enhancing the learning phase's efficiency and performance evaluation's accuracy. This uniform segmentation simplifies data handling and ensures that the model receives consistently sized inputs, facilitating a more structured and effective learning environment.

4.2. Model Settings

Our experimental framework was anchored on a configuration comprising 12 transformer blocks, each containing 16 attention heads, a structural detail illustrated in Figure 1. This architecture processes each frame of the gesture sequence through an encoding mechanism, transforming it into hidden states denoted by $h \in \mathbb{R}^{1280}$. In conjunction with this, we incorporated the WavLM Base+ model, accessed from the pre-trained repository (https://huggingface.co/microsoft/wavlm-base-plus (accessed on 25 April 2024)). This model is characterized by 12 transformer encoder layers, each producing hidden states with a dimensionality of 768, and is equipped with 8 attention heads. A translation-invariant self-attention (TISA) mechanism [39] was employed to achieve temporal translation invariance within our model. The bounded distance, b, of the DMAN was set to 80. All the settings can be found in our open code at https://github.com/zf223669/DiT-Gestures (accessed on 25 April 2024).

For the diffusion process, we adopted a quaternary variance schedule beginning from $\beta_1 = 1 \times 10^{-4}$ and concluding at $\beta_N = 5 \times 10^{-5}$, following a linear beat schedule. The diffusion sequence was set to $N = 500$ steps. The model's training protocol specified a batch size of 32 per GPU, leveraging an AdamW optimizer with a learning rate of 20×10^{-5} and employing a LinearScheduler with 10^3 warm-up steps.

The entire model architecture was developed by using the PyTorch Lightning framework to streamline the construction and scalability of our experiments. Computational resources included an Intel i9 processor and a solitary A100 GPU. The duration of model training varied across datasets, amounting to approximately 4 h for Trinity, 4 h for ZEGGS, and an extended 21 h for the BEAT dataset.

4.3. Visualization Results

The visual outcomes of our system, designed to generate lifelike gestures from training across three distinct datasets, are shown in the figures below for each dataset—Trinity (referenced in Figure 3), ZeroEGGS (referenced in Figure 4), and BEAT (referenced in Figure 5). Our model demonstrates high proficiency in creating gestures that exhibit a lifelike quality and synchronize precisely with the corresponding speech audio. Furthermore, it adeptly encapsulates and reflects the acoustic properties, semantic content, emotions, and personality traits inherent in speech, thereby offering a holistic representation of gestural expression.

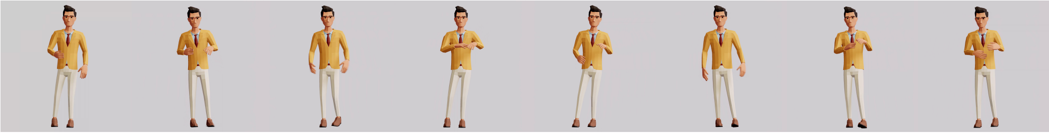

Figure 3. The full-body gestures generated in response to the audio from Record_008.wav in the Trinity dataset. The results suggest that the proposed architecture can generate gestures that accompany audio and produce more diverse, relaxed, and non-hyperactive gestures.

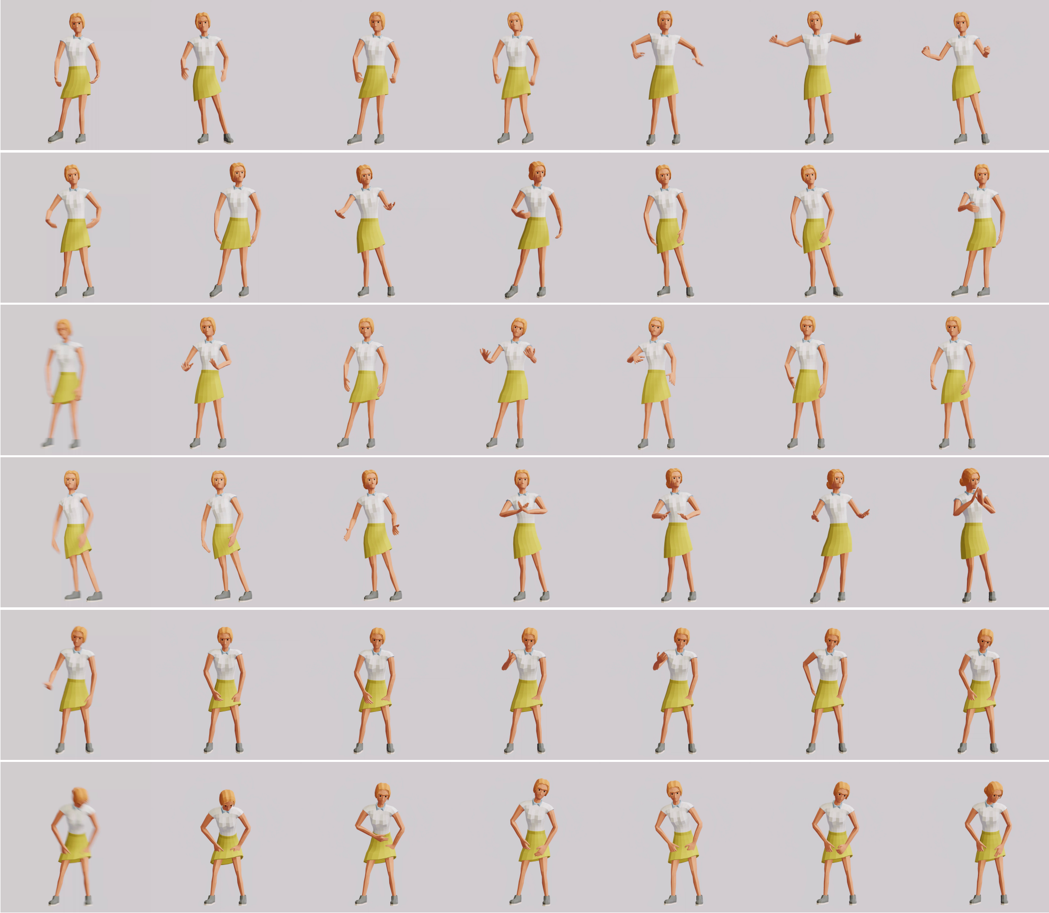

Figure 4. The results obtained by utilizing various audio styles in the ZEGGS dataset. It illustrates the model's ability to generate diverse motion styles, such as emotions, and age-related nuances. Remarkably, this variability is achieved solely by leveraging the speech audio information extracted by WavLM without requiring any manual annotation. Each row represents a distinct emotional gesture, arranged in the following order from top to bottom (the audio corresponding to the gesture is indicated within parentheses): happiness (011_Happy_0_x_1_0.wav), anger (026_Angry_0_x_0_9.wav), sadness (006_Sad_0_x_0_9.wav), threatening (051_Threatening_0_x_1_0.wav), old (022_Old_1_x_0_9.wav), and tired (063_Tired_1_x_0_9.wav).

Figure 6 (left) employs the t-SNE method to depict the distribution of generated gestures linked to various emotional states, as further detailed in Figure 6 (right) for personalities. This visual representation reveals distinct clusters for certain emotional states or personality types, whereas others show similarities yet are sufficiently distinct to be distinguishable. Such delineation underscores our approach's efficacy in generating nuanced and identifiable gestures solely based on raw speech audio without requiring explicit labels or manual annotation.

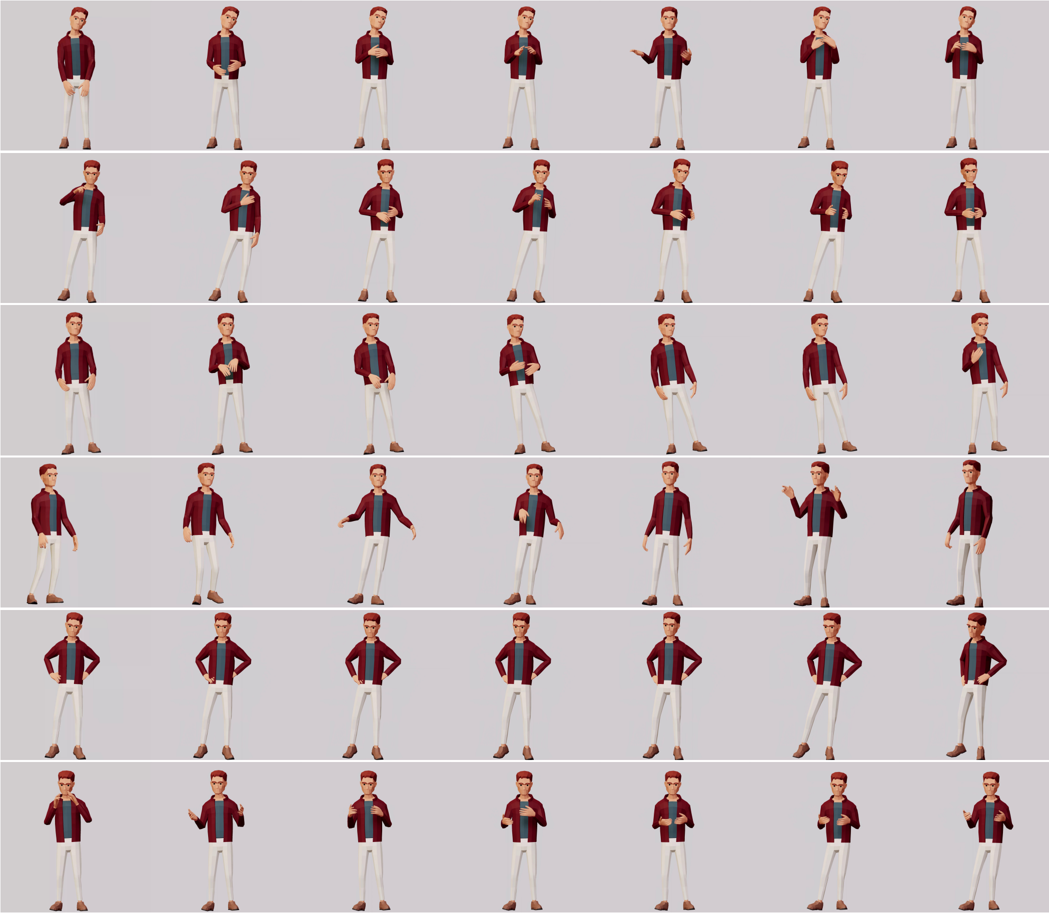

Figure 5. Extrapolating results on the BEAT dataset confirmed the ability of the proposed method to distinguish between different personalities of different people based solely on speech audio. The method effectively generates co-speech gestures that mirror the distinguishing features of the respective persons. Each row represents different personalities. Each row represents a distinct personalized gesture, arranged in the following order from top to bottom (the audio corresponding to the gesture is indicated within parentheses): Kieks (10_kieks_0_9_9.wav), Lu (13_lu_0_9_9.wav), Carlos (15_carlos_0_2_2.wav), Zhao (12_zhao_0_88_88.wav), Zhang (14_zhang_1_3_3.wav), and Wayne (1_wayne_0_39_39.wav).

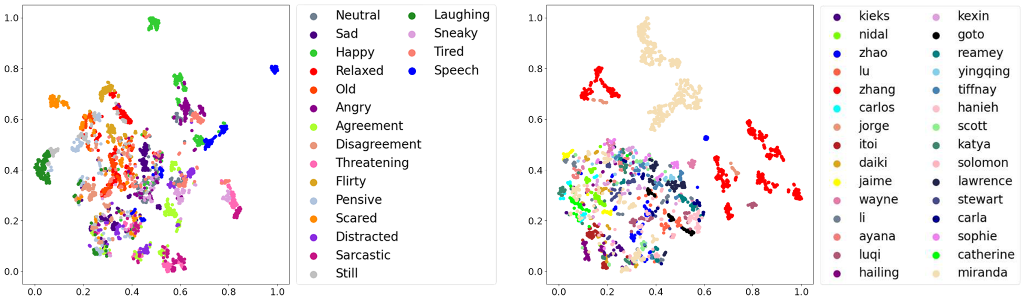

Figure 6. The visualization of gesture differentiation via T-SNE clustering reveals a compelling narrative about the capabilities of our method. Distinct colors signify varying emotions or personality gestures.

4.4. Comparison

Consistently with the prevailing practices in gesture generation research, we conducted a series of subjective and objective evaluations to evaluate the co-speech gestures generated by our DiT-Gestures (DG) model. For fairness and comparability, we selected baselines followed by (1) utilization of the same dataset, (2) availability as open-source resources, (3) incorporation of 3D full-body gestures, and (4) capability to generate sequences of at least 20 s.

For the Trinity dataset, we integrated insights from our prior work, Diffmotion (DMV1) [2], alongside StyleGestures (SG) [5], which utilizes a flow-based mechanism combined with LSTM technology. In the case of the ZEGGS dataset, our approach included the application of Zero-EGGS [10] and DiffuseStyleGesture (DSG) [7], noting that DSG is designed to accommodate only six distinct styles: happy, sad, neutral, old, relaxed, and angry. Additionally, for the BEAT dataset, we employed Cascaded Motion Network (CaMN) [37] and also DSG+ [8], albeit with its limitation in terms of speaker diversity [8]. These baseline models were meticulously selected to facilitate a thorough evaluation of our methodology across various datasets. Given the constraints of DSG's style classifications in both the ZEGGS and BEAT dataset experiments, our evaluation of this model is confined to the outcomes relevant to these predefined styles. Conversely, for the remaining models, our assessment spans all accessible classifications, ensuring a comprehensive analysis. We summary all the selected methods in Table 2.

Table 2. The summary information of all chosen approaches.

Method	Condition Encoding Strategy	Style Control	Architecture
SG [5]	MFCCs	Hand height, hand speed, and gesticulation radius	LSTM architecture combined with normalizing flows
ZeroEGGS [10]	Log amplitude of spectrogram, mel-frequency scale, and energy of audio	Employs example motion clips to influence style in a zero-shot learning framework	Uses a variational framework to learn style embedding
CAMN [37]	Raw audio and text	Emotion and Speaker ID label	LSTM-based structure
DSG [7]	Raw audio with WavLM and linear component	Uses classifier-free guidance to adjust stylistic expression based on discrete style labels	Diffusion model with self-cross local attention
DSG+ [8]	Similar to DSG but augmented with text semantic information	Similar to DSG; however, it employs categorical labels for the representation of distinct personality roles	Similar to DSG
Diffmotion [2]	MFCCs	No	LSTM-based diffusion model
DiT-Gestures (ours)	Raw audio with WavLM and Conv component	Raw audio	Diffusion + DMAN transformer

Our ablation study aimed to assess the comparative impact of incorporating the WavLM encoder (DG-W) versus mel-frequency cepstral coefficients (MFCCs) (denoted as DG-M) within our model's architectural framework. Further, we scrutinized the effect of employing the dynamic mask attention network (DMAN) by juxtaposing it with the alternative of integrating a causal mask (designated as DG-CM). This experimental setup allowed us to methodically evaluate the contribution of each component—WavLM and MFCC for audio encoding and DMAN versus causal mask for attention modulation—to the model's overall performance in generating gestures.

4.4.1. User Study

The primary intent of speech-driven gesture generation is to generate gestures that exhibit naturalness and persuasiveness. However, relying solely on objective measures for assessing gesture synthesis may not fully capture the subjective perception of quality by humans [5,40,41]. This study primarily emphasizes conducting subjective evaluations to assess human perception. Furthermore, we include supplementary objective evaluations, which are discussed in Section 4.4.2. For the subjective evaluation, we utilized a five-point Likert scale to rate three evaluation metrics: (1) human likeness, (2) appropriateness, and (3) style appropriateness.

The aspect of human likeness evaluates the naturalness and resemblance of the generated gestures to those made by an actual human, regardless of accompanying speech. Conversely, the aspect of appropriateness assesses the temporal consistency of the generated gestures, particularly in terms of their alignment with speech rhythm. Lastly, the aspect of style appropriateness measures the degree of similarity between the generated and original gestures. By incorporating these evaluation aspects, we aimed to assess the quality of the generated gestures comprehensively.

Each model was trained, and three gesture clips, each lasting 20 s, were generated with the same speech audio as input. The experiments were chosen separately for each emotional state audio in the ZEGGS dataset and per-person ID audio in the BEAT dataset. In the Trinity dataset, Record_008.wav and Record_015.wav were chosen.

A total of 30 volunteers were recruited, including 18 males and 12 females (aged 19–23). All of them were from China (22 from China and 8 international students from the USA, UK, etc.). They were asked to rate the scale for the evaluation aspects. The scores were assigned from 1 to 5, representing worst to best.

To initiate the formal experimental phase, we first acquainted participants with the methodology, showcasing example clips external to the evaluation dataset. Subsequently, participants were instructed to don headphones and position themselves in a serene environment, free from distractions, facing a computer screen. The identity of the method associated with each video remained undisclosed to the participants throughout the duration of the experiment. To ensure a randomized yet comprehensive exposure, the sequence of video presentations was randomized; however, each video was assured to be displayed thrice, with participants rendering their assessments immediately following each viewing.

The evaluation of human likeness involved the sequential presentation of individual videos devoid of any accompanying speech audio, to focus solely on the gestures. In contrast, the assessment of appropriateness included the additional element of the corresponding speech audio, mirroring the protocol of the human likeness evaluation save for this inclusion. For the evaluation of style appropriateness, the display was bifurcated into two segments: the left section featured ground truth (GT) gestures as a referential benchmark, whereas the right section exhibited a randomly chosen assortment of co-speech gestures synthesized by various models, enabling a comparative analysis of gesture style against the reference material.

One-way Analysis of Variance (ANOVA) was performed to assess the presence of statistical differences among the models' scores across the three distinct evaluation criteria. The outcomes are summarized in Table 3 for a tabulated overview. This analytical approach enabled us to quantitatively compare the effectiveness of each model in generating human-like gestures, assessing appropriateness in relation to accompanying speech, and evaluating style appropriateness, thereby providing a comprehensive understanding of each model's performance in these key areas.

The analysis yielded results indicating a statistically significant variation in the human likeness ratings between the GT gestures and those synthesized by the models across experiments conducted on the three datasets. Observationally, it was noted that the GT gestures encompass a relatively limited assortment of diverse movements, each marked by unique traits that contribute to the overall realism and dynamism of the depicted actions. Nevertheless, these particular gestures are characterized by their rarity within the dataset, falling into the long-tail distribution, which inherently complicates the models' ability to learn and replicate them accurately. Furthermore, the presence of these distinct gestures not only affects the scores related to human likeness but also has a discernible impact on the evaluations of appropriateness and style appropriateness, suggesting their influential role in the comprehensive assessment of gesture synthesis models.

Table 3. Subject mean perceptual rating scores. Bold fonts are utilized to emphasize the best results for each metric among the different methods, except for the GT (ground truth). The upward arrow (↑) signifies that higher scores indicate superior performance, while the downward arrow (↓) denotes that lower scores are associated with better outcomes.

Methods		**Subject Evaluation Metric**			**Objective Evaluation Metric**		
Dataset	**Model**	**Human ↑ Likeness**	**Appropriateness ↑**	**Style ↑ Appropriateness**	**FGD ↓ on Feature Space**	**FGD ↓ on Raw Data Space**	**BeatAlign ↑**
Trinity	GT	4.32 ± 0.35	4.53 ± 0.42	/	/	/	0.76
	SG [5]	2.11 ± 1.52	2.71 ± 0.88	2.77 ± 1.15	187.32	21,568.25	0.43
	DMV1 [2]	2.9 ± 0.67	2.37 ± 1.26	2.61 ± 1.21	179.52	21,356.86	0.50
	(Ours) DG-W	**4.30 ± 0.26**	**4.31 ± 0.13**	**4.19 ± 0.82**	**43.52**	**3358.18**	**0.67**
	(Ours) DG-CM	4.01 ± 0.70	4.12 ± 0.82	4.00 ± 0.25	46.45	3652.12	0.60
	(Ours) DG-M	4.22 ± 0.50	4.22 ± 1.18	4.02 ± 0.75	53.56	3925.66	0.61
ZEGGS	GT	4.50 ± 0.50	4.51 ± 0.50	/	/	/	0.81
	Zero-EGGS [10]	4.29 ± 0.77	4.26 ± 0.78	4.11 ± 0.23	32.05	2886.56	0.62
	DSG [7]	4.18 ± 0.84	4.15 ± 0.92	4.02 ± 0.25	33.26	3011.22	0.63
	(Ours) DG-W	**4.30 ± 0.72**	**4.27 ± 0.81**	**4.82 ± 0.32**	**31.96**	**2864.70**	**0.68**
	(Ours) DG-CM	3.00 ± 1.42	2.95 ± 1.41	4.11 ± 1.22	36.15	3021.53	0.62
	(Ours) DG-M	2.96 ± 1.40	2.95 ± 1.41	3.02 ± 1.28	47.24	3681.95	0.61
BEATS	GT	4.51 ± 0.50	4.50 ± 0.50	/	/	/	0.83
	CaMN [37]	3.49 ± 1.13	3.48 ± 1.12	3.48 ± 1.12	123.63	16,873.89	0.63
	DSG+ [8]	4.25 ± 0.75	4.24 ± 0.80	4.32 ± 0.73	**18.04**	1495.65	0.59
	(Ours) DG-W	**4.31 ± 0.73**	**4.30 ± 0.76**	**4.37 ± 0.70**	**18.04**	**1490.70**	**0.66**
	(Ours) DG-CM	4.24 ± 0.43	4.16 ± 0.47	4.00 ± 0.71	38.69	2597.23	0.61
	(Ours) DG-M	4.23 ± 0.42	4.02 ± 0.70	3.99 ± 0.70	38.78	2619.85	0.62

The findings from our experiments with the Trinity dataset demonstrate that our developed model, designated as DG-W, manifests statistically significant performance enhancements, outperforming the DMV1 and SG architectures in terms of all three evaluated metrics. This superiority is largely ascribable to the observed inconsistencies, such as the jitter present in full-body gestures generated by both DMV1 and SG. Specifically, DMV1 is characterized by a tendency towards hyperactivity in its generated gestures, whereas SG is prone to a gradual deceleration of motion over time. In contrast, the gestures synthesized by DG-W are characterized by their natural flow and relaxed dynamics, which markedly bolsters its performance across the evaluated metrics.

During the evaluations carried out on the ZEGGS dataset, our model showcased statistically significant differences in terms of human likeness, appropriateness, and style appropriateness metrics in comparison to both the Zero-EGGS and DSG models. These findings indicate that our model either outperforms or is outperformed by Zero-EGGS and DSG in these critical dimensions of gesture synthesis, underlining the distinct capabilities and limitations of each approach. The identification of statistically significant disparities suggests that our model offers a unique contribution to the field, either by advancing the realism and contextual alignment of the generated gestures or by highlighting areas for further refinement.

Within the experimental framework utilizing the BEAT dataset, our model (DG-W) manifested statistically significant enhancements across three evaluative metrics when benchmarked against CaMN and DSG. This outcome underscores the inherent benefits associated with integrating transformer and diffusion model architectures within our approach. The observed distinctions particularly underscore DG-W's capability to generate gestures of greater stability, circumventing the foot-skating problems that were evident with the DSG method. These results affirm the effectiveness of DG-W in addressing and mitigating specific challenges encountered in gesture synthesis, thereby evidencing its superior performance in creating more realistic and contextually appropriate gestures.

The outcomes derived from the investigations carried out on the ZEGGS and BEAT datasets underscore the efficacy of our methodology in producing a broad array of gestures. A pivotal aspect of our approach is its ability to accomplish this feat autonomously, without the dependency on preliminary seed sequences or the necessity for classification labels. This capability not only showcases the method's adaptability but also solidifies its standing as a significant advancement in the realm of gesture generation. By enabling the creation of a diverse spectrum of gestures tailored to various speech contexts, our model presents a compelling solution to the challenge of enhancing the naturalness and expressiveness of virtual or robotic entities.

4.4.2. Objective Evaluation

In our study, we implemented three objective evaluations to assess gesture generation quality rigorously: Fréchet Gesture Distance (FGD) evaluated in two dimensions—FGD in the feature space and FGD in the raw data space [42]—along with the BeatAlign metric [43]. FGD draws inspiration from the Fréchet Inception Distance (FID) [44], a renowned measure designed to quantify the quality of visual content generated by models. FGD has demonstrated a moderate albeit nonzero correlation with human assessments of gesture likeness, positioning it as a valuable tool for objective quality evaluation compared with other metrics [41]. Additionally, BeatAlign measures synchrony by quantifying the gesture–audio beat alignment by computing the Chamfer Distance between audio beats and gesture markers. This multifaceted evaluation framework enabled a comprehensive assessment of the generated gestures, focusing on both their qualitative likeness to human gestures and their temporal alignment with the corresponding audio cues.

The final results are shown in Table 3. Our methods demonstrated SoTA performance based on objective evaluations using the FGD and BeatAlign metrics. Our model outperformed other architectures in both FGD metrics, demonstrating its capability to generate distributions closely resembling the ground truth. Furthermore, it achieved a superior BeatAlign score compared with alternative approaches, highlighting its proficiency in synthesizing co-speech that closely aligns with the rhythm of the audio.

4.4.3. Ablation Studies

Our research included detailed ablation studies designed to elucidate the performance impact of distinct components integrated within our model. Specifically, we focused on evaluating the effectiveness of two key aspects: (1) the employment of WavLM features as opposed to MFCCs and (2) the incorporation of the DMAN versus substituting the DMAN with a causal mask. To ensure a comprehensive understanding of each component's contribution, user studies were meticulously conducted independently for each aspect. The outcomes of these evaluations were systematically compiled and are accessible in Table 3. This structured approach allowed us to discern these components' individual and combined effects on our model's overall performance, facilitating targeted improvements and optimization.

After replacing the WavLM audio feature extraction method with MFCCs, the scores in all three metrics decreased, particularly in the style appropriateness metrics. These results suggest that utilizing WavLM as a feature extractor better facilitates the synthesis of style-corresponding gestures than using MFCC features. The aforementioned observation can be ascribed to the pre-trained WavLM model's marvelous capacity for extracting more comprehensive information from speech audio.

Upon substituting the dynamic mask attention network (DMAN) with a causal mask within our model, there was a discernible decrease in the scores across all three evaluated metrics. This outcome suggests that the DMAN plays a crucial role in enhancing the model's performance by contributing to the precision and quality of gesture generation. The observed decline across the metrics indicates the DMAN's effectiveness in facilitating more nuanced attention mechanisms, which the causal mask, by comparison, does not equally support. This finding underscores the importance of the DMAN component in the architectural design of gesture synthesis models for achieving superior results.

5. Conclusions

This study introduced "DiT-Gestures", a novel approach to generating co-speech gestures utilizing a speech-conditional diffusion-based and non-autoregressive transformer-based generative model. Our methodology distinctively leverages the WavLM pre-trained model and a dynamic mask attention network (DMAN) to synthesize individualized and stylized full-body gestures from raw speech audio. This advancement represents a significant leap in the virtual human technology domain, notably enhancing the naturalness and realism of virtual human interactions without requiring complex multimodal processing or manual annotation.

Our approach offers several advantages: (1) It relies solely on raw speech audio to generate stylized gestures, eliminating the need for additional inputs and thus enhancing user friendliness. (2) It ensures superior synchronization of full-body gestures with speech, adeptly capturing rhythm, intonation, and certain semantics without sacrificing naturalness.

The extensive subjective and objective evaluations conducted across the Trinity, ZEGGS, and BEAT datasets demonstrated our model's superior performance in generating co-speech gestures that exhibit high degrees of naturalness, human likeness, and synchrony with the accompanying speech. Our approach has shown exceptional capability in capturing and expressing the nuanced variations attributable to different speakers' emotions, personality traits, and styles solely from the speech audio. These outcomes underscore the efficacy of integrating the WavLM model for speech feature extraction, which enables the extraction of a richer and more nuanced feature set that goes beyond basic acoustic properties to include semantic cues, emotional valence, and personality indicators embedded within the speech signal.

Our findings also highlight the critical role of the dynamic mask attention network (DMAN) in our model's architecture. The DMAN significantly contributes to the model's ability to generate more contextually aligned and expressive gestures by facilitating a more flexible and adaptive attention mechanism. This feature is particularly beneficial in addressing the inherent variability and complexity of human gesture–speech interactions, where the same speech content can elicit a wide range of gesture responses based on the speaker′s emotional state, personality, and interaction context.

Our study identifies several key areas for improvement: Firstly, the model's exclusive reliance on speech audio may limit its capability to capture style features in segments characterized by minimal speech. Secondly, using the diffusion model (DDPM) extends the generation times. Thirdly, our model might not effectively replicate certain gestures that are essential to expressing specific emotional or contextual states. Fourthly, we will investigate the utilization of large-scale models based on different languages, such as Chinese. These observations underscore the necessity for further enhancements to broaden the model's ability to convey a wider spectrum of human gestures accurately.

Author Contributions: Conceptualization, F.Z. and N.J.; methodology, F.Z.; software, Z.W.; validation, Z.W., S.Z. and F.G.; writing—original draft preparation, F.Z. and Z.W.; writing—review and editing, F.Z., N.J. and X.L.; visualization, F.Z. and F.G.; supervision, N.J.; funding acquisition, F.Z. and N.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the "Pioneer" and the "Leading Goose" R&D Programs of Zhejiang (2023C01212) and the National Key Research and Development Program of China (2022YFF0902305).

Data Availability Statement: The data presented in this study are available at https://github.com/zf223669/DiT-Gestures (accessed on 25 April 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Matthew, B. Voice puppetry. In Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, 8–13 August 1999; pp. 21–28.

2. Zhang, F.; Ji, N.; Gao, F.; Li, Y. DiffMotion: Speech-Driven Gesture Synthesis Using Denoising Diffusion Model. In Proceedings of the MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, 9–12 January 2023; Proceedings, Part I; Springer: Berlin/Heidelberg, Germany, 2023; pp. 231–242.
3. Sarah, T.; Jonathan, W.; David, G.; Iain, M. Speech-Driven Conversational Agents Using Conditional Flow-VAEs. In Proceedings of the European Conference on Visual Media Production, London, UK, 6–7 December 2021; pp. 1–9.
4. Alexanderson, S.; Nagy, R.; Beskow, J.; Henter, G.E. Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models. *arXiv* **2022**, arXiv:2211.09707.
5. Simon, A.; Eje, H.G.; Taras, K.; Jonas, B. Style-Controllable Speech-Driven Gesture Synthesis Using Normalising Flows. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2020; Volume 39, pp. 487–496.
6. Bhattacharya, U.; Childs, E.; Rewkowski, N.; Manocha, D. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, China, 20–24 October 2021; pp. 2027–2036.
7. Yang, S.; Wu, Z.; Li, M.; Zhang, Z.; Hao, L.; Bao, W.; Cheng, M.; Xiao, L. DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models. *arXiv* **2023**, arXiv:2305.04919.
8. Yang, S.; Xue, H.; Zhang, Z.; Li, M.; Wu, Z.; Wu, X.; Xu, S.; Dai, Z. The DiffuseStyleGesture+ entry to the GENEA Challenge 2023. *arXiv* **2023**, arXiv:2308.13879.
9. Li, J.; Kang, D.; Pei, W.; Zhe, X.; Zhang, Y.; Bao, L.; He, Z. Audio2Gestures: Generating Diverse Gestures From Audio. *IEEE Trans. Vis. Comput. Graph.* **2023**, *14*, 1–15. [CrossRef] [PubMed]
10. Ghorbani, S.; Ferstl, Y.; Holden, D.; Troje, N.F.; Carbonneau, M.A. ZeroEGGS: Zero-Shot Example-Based Gesture Generation from Speech. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2023; Volume 42, pp. 206–216.
11. Wagner, P.; Malisz, Z.; Kopp, S. Gesture and speech in interaction: An overview. *Speech Commun.* **2014**, *57*, 209–232. [CrossRef]
12. Ylva, F.; Michael, N.; Rachel, M. Multi-objective adversarial gesture generation. In Proceedings of the Motion, Interaction and Games, Newcastle upon Tyne, UK, 28–30 October 2019; pp. 1–10.
13. Ian, G.; Jean, P.A.; Mehdi, M.; Bing, X.; David, W.F.; Sherjil, O.; Aaron, C.; Yoshua, B. Generative Adversarial Nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9.
14. Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; Catanzaro, B. Diffwave: A Versatile Diffusion Model for Audio Synthesis. *arXiv* **2020**, arXiv:2009.09761.
15. Rezende, D.; Mohamed, S. Variational Inference with Normalizing Flows. In Proceedings of the International Conference on Machine Learning (PMLR), Lille, France, 7–9 July 2015; pp. 1530–1538.
16. Dinh, L.; Krueger, D.; Bengio, Y. Nice: Non-linear Independent Components Estimation. *arXiv* **2014**, arXiv:1410.8516.
17. Dinh, L.; Sohl-Dickstein, J.; Bengio, S. Density Estimation Using Real Nvp. *arXiv* **2016**, arXiv:1605.08803.
18. Jing, L.; Di, K.; Wenjie, P.; Xuefei, Z.; Ying, Z.; Zhenyu, H.; Linchao, B. Audio2Gestures: Generating Diverse Gestures from Speech Audio with Conditional Variational Autoencoders. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 11293–11302.
19. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
20. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International Conference on Machine Learning (PMLR), Lille, France, 7–9 July 2015; pp. 2256–2265.
21. Rasul, K.; Sheikh, A.S.; Schuster, I.; Bergmann, U.; Vollgraf, R. Multivariate probabilistic time series forecasting via conditioned normalizing flows. *arXiv* **2020**, arXiv:2002.06103.
22. Song, Y.; Ermon, S. Generative modeling by estimating gradients of the data distribution. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–13.
23. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y. Conformer: Convolution-Augmented Transformer for Speech Recognition. *arXiv* **2020**, arXiv:2005.08100.
24. Ao, T.; Zhang, Z.; Liu, L. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. *arXiv* **2023**, arXiv:2303.14613.
25. Windle, J.; Greenwood, D.; Taylor, S. UEA Digital Humans Entry to the GENEA Challenge 2022. In Proceedings of the GENEA: Generation and Evaluation of Non-Verbal Behaviour for Embodied Agents Challenge, Bengaluru, India, 7–11 November 2022.
26. Cambria, E.; Livingstone, A.; Hussain, A. The hourglass of emotions. In Proceedings of the Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, 21–26 February 2011; Revised Selected Papers; Springer: Berlin/Heidelberg, Germany, 2012; pp. 144–157.
27. Russell, J. A Circumplex Model of Affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
28. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X. Wavlm: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. [CrossRef]
29. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [CrossRef]
30. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
31. Chen, Y. Convolutional Neural Network for Sentence Classification. Master's Thesis, University of Waterloo, Waterloo, ON, Canada, 2015.

32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
33. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
34. Fan, Z.; Gong, Y.; Liu, D.; Wei, Z.; Wang, S.; Jiao, J.; Duan, N.; Zhang, R.; Huang, X. Mask Attention Networks: Rethinking and Strengthen Transformer. *arXiv* **2022**, arXiv:2203.05297.
35. Paul, L. Sur la Théorie du Mouvement Brownien. *C. R. Acad. Sci.* **1908**, *65*, 530–533+146.
36. Ylva, F.; Rachel, M. Investigating the Use of Recurrent Motion Modelling for Speech Gesture Generation. In Proceedings of the 18th International Conference on Intelligent Virtual Agents, Sydney, NSW, Australia, 5–8 November 2018; pp. 93–98.
37. Liu, H.; Zhu, Z.; Iwamoto, N.; Peng, Y.; Li, Z.; Zhou, Y.; Bozkurt, E.; Zheng, B. BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis. *arXiv* **2022**, arXiv:2203.05297.
38. Grassia F. Sebastian. Practical Parameterization of Rotations Using the Exponential Map. *J. Graph. Tools* **1998**, *3*, 29–48. [CrossRef]
39. Wennberg, U.; Henter, G.E. The Case for Translation-Invariant Self-Attention in Transformer-Based Language Models. *arXiv* **2021**, arXiv:2106.01950.
40. Wolfert, P.; Robinson, N.; Belpaeme, T. A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents. *IEEE Trans. Hum.-Mach. Syst.* **2022**, *52*, 379–389. [CrossRef]
41. Kucherenko, T.; Wolfert, P.; Yoon, Y.; Viegas, C.; Nikolov, T.; Tsakov, M.; Henter, G.E. Evaluating Gesture-Generation in a Large-Scale Open Challenge: The GENEA Challenge 2022. *arXiv* **2023**, arXiv:2303.08737.
42. Youngwoo, Y.; Bok, C.; Joo-Haeng, L.; Minsu, J.; Jaeyeon, L.; Jaehong, K.; Geehyuk, L. Speech Gesture Generation from the Trimodal Context of Text, Audio, and Speaker Identity. *ACM Trans. Graph.* **2020**, *39*, 1–16.
43. Li, R.; Yang, S.; Ross, D.A.; Kanazawa, A. Ai choreographer: Music conditioned 3d dance generation with aist++. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13401–13412.
44. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–12.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.