DIDiffGes: Decoupled Semi-Implicit Diffusion Models for Real-time Gesture Generation from Speech

Yongkang Cheng[1, 3], **Shaoli Huang**[1*], **Xuelin Chen**[1], **Jifeng Ning**[3], **Mingming Gong**[2,4]
[1]Tencent AI Lab
[2]School of Mathematics and Statistics, The University of Melbourne
[3]College of Information Engineering, Northwest A&F University
[4]Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates
cyk19990422@gmail.com, shaol.huang@gmail.com, xuelin.chen.3d@gmail.com, njf@nwsuaf.edu.cn, mingming.gong@unimelb.edu.au

Abstract

Diffusion models have demonstrated remarkable synthesis quality and diversity in generating co-speech gestures. However, the computationally intensive sampling steps associated with diffusion models hinder their practicality in real-world applications. Hence, we present DIDiffGes, for a Decoupled Semi-Implicit Diffusion model-based framework, that can synthesize high-quality, expressive gestures from speech using only a few sampling steps. Our approach leverages Generative Adversarial Networks (GANs) to enable large-step sampling for diffusion model. We decouple gesture data into body and hands distributions and further decompose them into marginal and conditional distributions. GANs model the marginal distribution implicitly, while L2 reconstruction loss learns the conditional distributions exciplictly. This strategy enhances GAN training stability and ensures expressiveness of generated full-body gestures. Our framework also learns to denoise root noise conditioned on local body representation, guaranteeing stability and realism. DIDiffGes can generate gestures from speech with just 10 sampling steps, without compromising quality and expressiveness, reducing the number of sampling steps by a factor of 100 compared to existing methods. Our user study reveals that our method outperforms state-of-the-art approaches in human likeness, appropriateness, and style correctness.

Introduction

The integration of large language models like ChatGPT (Schulman et al. 2022) into our digital lives has heralded a significant evolution in human-computer interaction. These models facilitate more natural and engaging conversations. However, to fully exploit the potential of these interactions, integrating real-time, realistic non-verbal elements, such as gestures, in harmony with the advanced verbal capabilities of these models, is crucial.

The realm of gesture synthesis faces the challenge of precisely and swiftly generating complex human gestures, encompassing hand and body movements. GAN-based methods, like those in (Liu et al. 2022a), although promising, struggle with motion representation and training complexities. VAE-based approaches, as seen in (Yoon et al. 2020), often result in less fluid arm poses, diminishing the realism of the generated gestures. Diffusion model-based methods, highlighted in studies such as (Ao, Zhang, and Liu 2023; Yang et al. 2023b,a), have emerged as promising alternatives due to their comprehensive gesture movement capture. However, their slow generation speed is a substantial barrier, especially for real-time interactions with virtual agents.

Recent advancements like the MLD approach from (Chen et al. 2023), which introduces a latent space structure for encoding body movements (Kingma and Welling 2013) and adapts the latent diffusion model (Rombach et al. 2022), along with the DDIM sampling strategy (Song, Meng, and Ermon 2020), attempt to address these challenges. Nevertheless, these methods often prioritize increased computation speed at the expense of generating overly smooth motions, thereby lacking the nuanced complexity observed in natural human movement. In this paper, we propose DIDiffGes, a Decoupled Semi-Implicit Diffusion model framework that aims to resolve the speed and fidelity issues in gesture generation. Our model is inspired by the recent Semi-Implicit Denoising Diffusion model (SIDDM) (Xu et al. 2023), which demonstrates unconditional high-fidelity image generation in a few steps. "Semi-implicit" denotes the training of the large-step denoiser using adversarial learning as the implicit objective, complemented by L2 loss as an explicit objective for better convergence. DIDiffGes generalizes SIDDM to human gesture generation by further decoupling the denoiser's output into distinct representations for hand and body noise, ensuring high-quality gesture generation at a few diffusion steps.

To be more specific, DIDiffGes's innovation lies in adversarially training separate noise components for body and hand movements at each diffusion step, better capturing their varying complexities and dynamics. This ensures a balanced and accurate representation of both movement types, crucial for generating realistic, expressive, and diverse gestures. The introduction of semi-implicit objectives into the diffusion framework is transformative, facilitating significant data distribution changes at each step and producing outputs closely resembling real gesture data. This capability is particularly

*Corresponding author: Shaoli Huang.

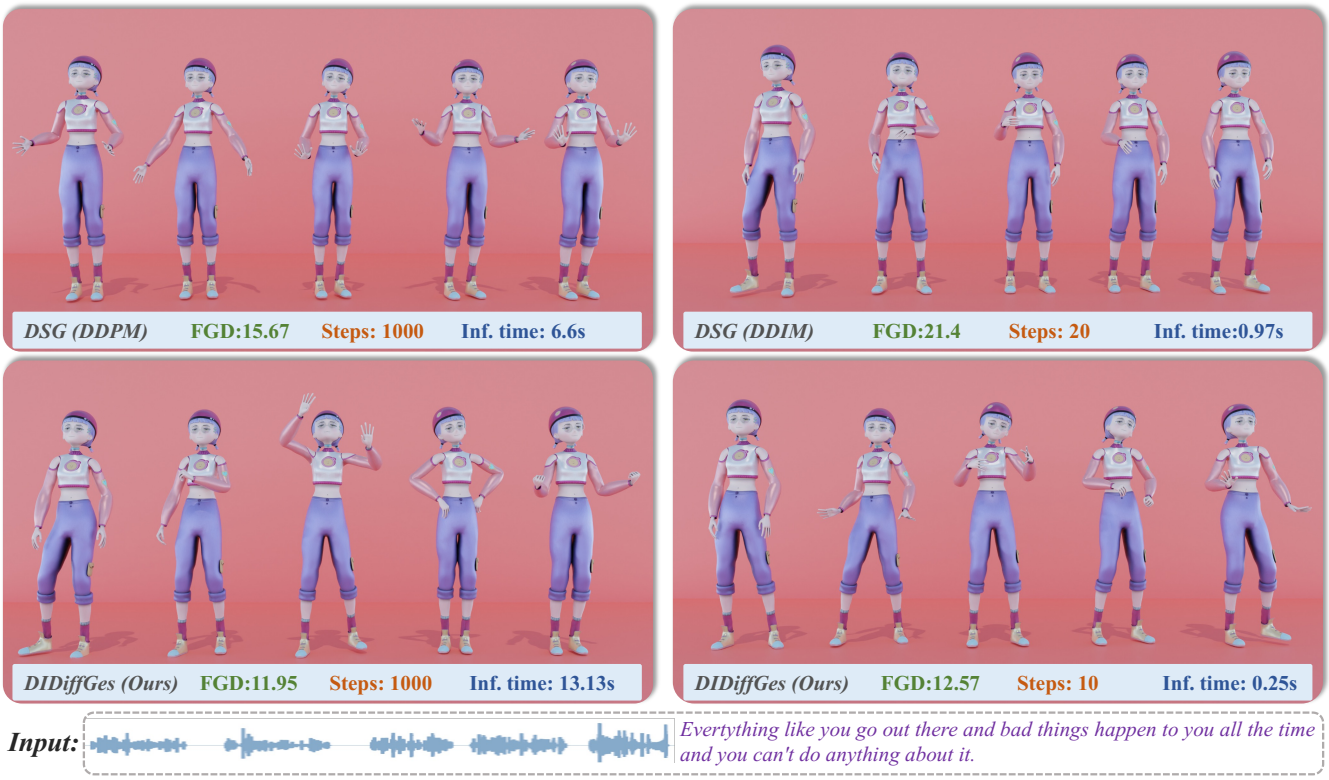

Figure 1: Comparison of four different sampling methods: DSG (Yang et al. 2023b) with DDPM, DSG with DDIM, our method with 1000-step sampling, and our method with 10-step sampling.

valuable in real-time applications where fast, accurate, and natural-looking gesture generation is essential.

Moreover, DIDiffGes employ a sequential diffusion denoiser strategy that recovers root noise conditioning on local body representations. By focusing on the interplay between root motion and local body motion, this strategy improves coordination helps reduce unnatural phenomena, such as foot sliding and body drifting, which can occur during gesture generation.

We validate our approach on both the BEATs (Liu et al. 2022a) and ZeroEGGs (Ghorbani et al. 2023) datasets. Experimental results demonstrate that our method can generate high-quality, realistic, and natural gesture motions with fewer sampling steps. Notably, with only **10** sampling steps, our method surpasses existing open-source diffusion model-based approaches in the FGD metric and is nearly 15 times faster, inferring 88 frames of gesture motion in just 0.4 seconds. Our user study also confirms the superior quality of our rapidly generated results compared to methods requiring more sampling steps. Furthermore, we can quickly generate dance motions in the music-to-dance task, achieving performance comparable to existing methods.

In summary, DIDiffGes represents a significant advancement in the field of gesture synthesis, addressing critical needs for efficient, high-quality, and realistic gesture generation. It enhances the capabilities of large language models in real-time applications, paving the way for more natural and expressive human-AI interactions, and setting a new standard in motion synthesis across various domains.

Related Work

Audio-to-Gesture Generation leverages rich multimodal inputs such as speech audio, music files, and frame-missing action sequences to generate action sequences. The task of co-speech gesture generation is particularly complex, requiring understanding of speech melody, semantics, gesture action, and their interrelationships. Early data-driven methods, such as those proposed by (Liu et al. 2022b; Habibie et al. 2021), attempted to learn gesture matching from human demonstrations but often resulted in less diverse actions. Later works, like (Habibie et al. 2021; Yi et al. 2023; Xie et al. 2022), enhanced the model's ability to generate diverse results and introduced the concept of generating distinctive and expressive gesture results. Some studies, such as (Yang et al. 2023b,a; Ahuja et al. 2020; Ao, Zhang, and Liu 2023), trained a unified model for multiple speakers, embedded each speaker's style in space, or introduced style transfer technology. Other works (Zhou, Bian, and Chen 2022; Habibie et al. 2022) used motion matching to generate gesture sequences, although this method often requires complex matching rules. Despite its challenges, audio-driven animation has attracted widespread attention and has been significantly improved by the recent emergence of the distinctive, high-quality dataset ZeroEGGs (Ghorbani

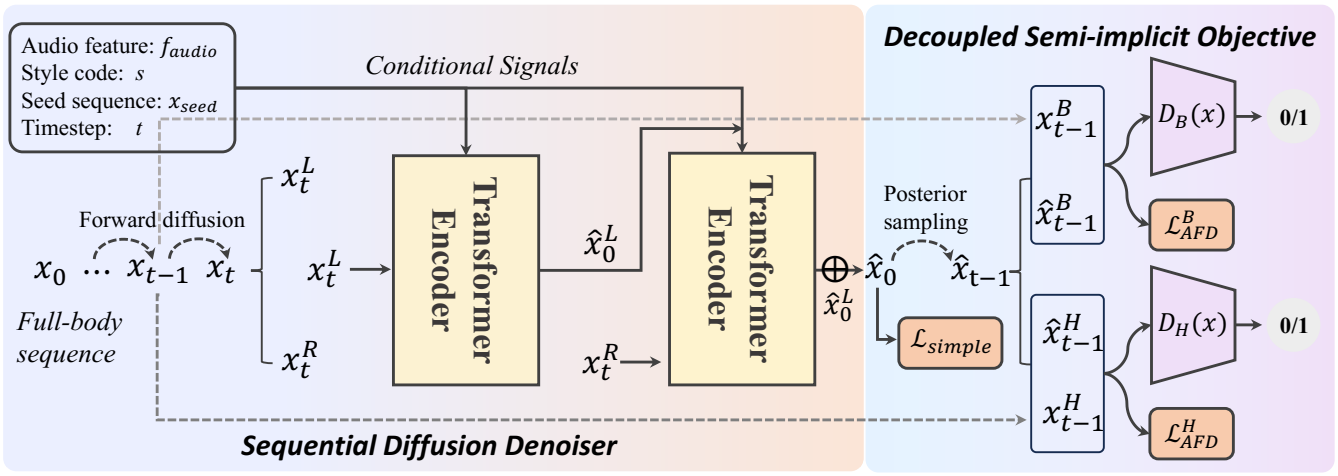

Figure 2: Our learning framework integrates a Sequential Diffusion Denoiser with two transformer encoders and a Decoupled Semi-implicit Objective. The first encoder denoises local motion and provides a conditional signal for the second encoder, which denoises root noise. The final result, a combination of local motion and root result, is added with t-1 step noise via posterior sampling and then decoupled into body and hand noise. These noise undergo adversarial training against prior sampled noise, supervised by Auxiliary Forward Diffusion Loss. For a detailed description of the network architecture, please refer to our supplementary materials.

et al. 2023). Our DIDiffGes stands out in key ways. It's the first in gesture gen. to use GANs for accelerating diffusion-based methods. Also, it denoises local motion noise before root noise, enhancing stability. Finally, by interfacing and separately supervising body and hands with GANs, it ensures finger movement, a unique feature.

Diffusion Models have shown great results in many fields (Rombach et al. 2022; Chen et al. 2023; Saharia et al. 2022), esp. in human motion gen. Motion Diffuse (Zhang et al. 2022) was the 1st to use diffusion models for text-conditioned human motion gen., giving fine-grained instr. for 2 body parts. MDM (Tevet et al. 2022) is a key work, introducing motion diffusion model to handle the relationship between motion rep. and text control cond. Recent work has focused on motion trajectory and joint control. In our work, we capture the relationship between gesture seqs. and speech by an attention mech. to gen. highly matched results. But due to the high dim. and iterative nature of diffusion models, motion gen. based on DDPM has time overhead issues. MLD (Chen et al. 2023) brought latent diffusion into motion gen. to enhance quality and reduce resource req., by training a VAE for motion embedding 1st and then applying latent diffusion in the latent space. However, it's a non-end-to-end method and may produce artifacts. Our DIDiffGes focuses on efficiently gen. high-quality motion. Unlike methods using DDIM or latent reps., we use GANs for large-step sampling in diffusion models.

Method

Our goal is to swiftly create high-fidelity and richly expressive co-speech gesture sequences derived from input audio signals using diffusion generative models. We ultimately aim to integrate these high-quality models into real-time applications for everyday use. To achieve this, we endeavor to expedite the denoising process by implementing a diminished number of steps and increasing step sizes. In the following sections, we elucidate our Overall Structure, Sequential Diffusion Denoiser, and Decoupled Semi-implicit Objective.

Overall Structure

Our framework is capable of receiving real-time audio signals of speech or pure music and guiding the generation of highly expressive full-body gestures or dance sequences. Additionally, we allow for the input of other control signals, such as style labels to enrich the emotional content of the generated gesture sequences, or seed sequences for producing smooth long-frame sequences. As illustrated in Figure 2, our overall architecture comprises two core components: the Sequential Diffusion Denoiser and the Decoupled Semi-implicit Objective. The former mitigates the unnaturalness of denoised motion sequences in non-physics-based environments, while the latter breaks the assumption dependency of DDPM by modeling large-stride denoising distributions to achieve fewer-step sampling, thereby enabling high-speed generation. Furthermore, decoupling the body and hand parts for independent modeling of their respective distributions significantly enhances the expressiveness of the generated sequences, providing an improved user experience. To the best of our knowledge, DIDiffGes is the first framework attempting real-time high-fidelity co-speech gesture generation, offering the possibility of practical real-time applications for current audio-to-gesture research.

Sequential Diffusion Denoiser

Traditional diffusion model-based methods directly add noise to the global gesture representation, and then denoise through a simple Transformer-Based network to obtain the

final generation result. However, these methods train their models in a non-physical environment, making issues such as foot sliding and global jitter inevitable. Incorporating a simulation environment during the sampling process would compromise our real-time requirements. Nevertheless, we observe that these physical sliding issues often occur in large-scale gestures. For instance, when swinging the arm, the human legs remain stationary, but our torso bends with the arm movement, often leading to incorrect root joint displacements and resulting in a sliding effect. Therefore, we further decouple the limbs, extracting the root joint's representation from the limb representation, and use the denoised local motion as a new conditional control signal to guide the generation of root motion that conforms to the local motion. This approach alleviates the physical sliding problem and enhances the visual effect without significantly increasing the time overhead, maintaining the original training stability.

Preliminary. The gesture sequence is represented as $x^{1:N}$, where N denotes the number of frames in the gesture sequence. Subsequently, we employ the diffusion probability model (Ho, Jain, and Abbeel 2020) to generate co-speech gestures, in which the diffusion model gradually anneals pure Gaussian noise into the gesture distribution $p(x)$. As a result, the model can predict noise from the T-step Markov noise process $\{x_t^{1:N}\}_t^T$, where $x_0^{1:N}$ is directly sampled from the original data distribution. The diffusion process is as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{\frac{\alpha_t}{\alpha_{t-1}}}x_{t-1}, (1-\frac{\alpha_t}{\alpha_{t-1}})I), \quad (1)$$

where $\{\beta_t\}_{t=1}^T$ is the variance schedule and $\alpha_t = \prod_{s=1}^t(1-\beta_s)$. Subsequently, following the idea in MDM (Tevet et al. 2022), we retain the denoising result as $\hat{x_0} = \epsilon_t^\theta(x_t)$. However, due to the widespread use of the acceleration strategy DDIM, we present the description in the form of DDIM to cater to a broader audience in the diffusion model domain. To avoid further confusion, we preserve the description of $\hat{x_0}$ and ϵ from the DDIM paper, defining the model output as predicted noise. We rewrite the reverse process as follows:

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\alpha_t}\epsilon_t^\theta(x_t)}{\sqrt{\alpha_t}}. \quad (2)$$

Upon obtaining $\hat{x}_0$ and the known forward diffusion result x_t, we can calculate the posterior distribution and sample $\hat{x}_{t-1}$. The formula representation is as follows:

$$\hat{x}_{t-1} = \sqrt{\alpha_{t-1}}\hat{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_t^\theta(x_t) + \sigma_t z_t. \quad (3)$$

Decoupled Denoiser Structure. Our structure is illustrated in Figure 2. We first forward-add noise to the original motion sequence $X_0^{1:N}$, sampled from $p(x_0)$, to pure Gaussian noise $x_t^{1:N}$ using Equation (1). Subsequently, we decouple the root joint's noise representation X_t^R as the noise input for the next stage, along with the audio, seed sequence, and style code, which are input into the denoiser. Our conditional denoiser contains a Transformer-based encoder and a conditional control signal encoder. The former consists of 12 layers and 8-head self-attention modules with default skip connections, while the latter is composed of Linear layers and WavLM modules. The Linear Block linearly maps the seed sequence and style code, while the audio input is encoded through WavLM to capture spectral information and is linearly mapped to the same dimensional space. We then concatenate all control signals as conditional features input into the Encoder. As shown in Equation (3) of Section 3.1, our denoiser directly predicts the clean local gesture sequence, $\hat{x}_0^L = G(x_t, t, c)$, where c is the conditional signals and t is timestep. Next, we encode $\hat{x}_0^L$ as a local motion condition and additionally concatenate it to the control signals as a new guiding condition to generate the global motion sequence $\hat{x}_0^R$. Finally, we concatenate the local sequence $\hat{x}_0^L$ and global sequence $\hat{x}_0^R$ into the overall motion $\hat{x}_0$, and constrain it with the following reconstruction loss and physical constraints:

$$\begin{aligned}\mathcal{L}_{simple} &= E_{x_0\ q(x_0|c),t\ [1,T]}[HuberLoss(x_0 - \hat{x}_0)],\\ \mathcal{L}_{foot} &= \frac{1}{N-1}\sum_{i=1}^{N-1}\|\hat{x}_0^{i+1} - \hat{x}_0^i\| f_i,\end{aligned} \quad (4)$$

where f_i is determined by calculating the rate of change in the y-axis position of the footstep to judge whether the foot is in contact with the ground. Notably, the root joint regeneration part proposed in our method effectively optimizes the human trajectory, thereby avoiding the unnatural state of both feet being suspended simultaneously.

Decoupled Semi-implicit Objective

Problem Description. The key to implementing a real-time denoising diffusion model lies in increasing the noise addition step size to reduce the number of denoising steps, thereby achieving the goal of high-speed generation. As described in Section Introduction, DDPM is based on the assumption that the noise added at each step is small and sampled from a unimodal distribution. Hence, we can parameterize $p_\theta(x_{t-1}|x_t)$ as a Gaussian Distribution, using the $L2$ loss to model the KL divergence between $p_\theta(x_{t-1}|x_t)$ and $q(x_{t-1}|x_t, x_0)$ at the same t. However, when we directly increase the noise step size based on this, $p_\theta(x_{t-1}|x_t)$ no longer conforms to a Gaussian Distribution. Simple explicit $L2$ loss cannot model such a complex motion distribution, leading to the generation of unnatural jittery results.

To model complex distributions, we naturally consider the GAN model. To address this issue, some researchers (Wang et al. 2022) proposes conditional generators and conditional discriminators, introducing adversarial learning strategies to model the complex motion distribution between multiple sampling steps:

$$\min_\theta \max_{D_{adv}} \sum_{t>0} \mathbb{E}_{q(x_t)} D_{adv}(q(x_{t-1}|x_t)||p_\theta(x_{t-1}|x_t)), \quad (5)$$

Within this framework, the conditional discriminator endeavors to differentiate between the predicted denoising distribution and the original motion distribution, while the conditional generator aspires to render them indistinguishable. Nevertheless, adversarial learning constitutes a purely implicit strategy; during the training process, factors like data volume and distribution complexity may impact training sta-

bility, requiring repeated hyperparameter searches to facilitate training. When attempting to model human motion distributions with higher physical geometric constraint requirements, this purely implicit learning strategy is often proven to be statistically inefficient.

Semi-implicit Matching Constraints. Upon examining the implementation of Equation(5), it becomes evident that, during the adversarial phase, the method indirectly matches the conditional distribution by aligning with the joint distribution:

$$\min_{\theta}\max_{D_{adv}}\mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})}D_{adv}(q(x_{t-1},x_t)||p_\theta(x_{t-1},x_t)), \tag{6}$$

This approach requires connecting large-stride denoising distributions between two adjacent time steps in each discrimination phase of adversarial learning. However, the large-step noise distribution is often a complex multimodal distribution rather than the unimodal distribution assumed by DDPM. This undoubtedly makes GAN-based frameworks difficult to train and requires a reasonable design of the sampling step t, with the cost of searching parameters during the adversarial training phase being too high. However, adversarial training is a purely implicit matching process, typically used to constrain distributions that cannot be explicitly represented. We consider using a simpler marginal distribution to replace the joint distribution in Equation(6). That is, we directly compute the posterior distribution and subsequently perform adversarial learning with the forward process to model the large-step denoising distribution. The formula representation is as follows:

$$\min_{\theta}\max_{D_{adv}}\mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})}[-log(D_{adv}(x_{t-1},c,t))]+[-log(1-D_{adv}(\hat{x}_{t-1},c,t))], \tag{7}$$

Although we have simplified the implicit matching process, making adversarial training more stable, we have encountered a new issue. Since the large-step denoising distribution is often a complex multimodal distribution, the posterior sampling $p_\theta(\hat{x}_{t-1}|x_t,\hat{x}_0)$ result still exhibits significant differences from the forward process, rendering our denoiser unable to successfully reverse from the pure noise distribution to the original distribution. Based on this, we introduce a regularization term, Auxiliary Forward Diffusion Constraint (AFD), for explicitly constraining the similarity between posterior sampling results and forward diffusion results at the same time step. Its representation is as follows:

$$\mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})}\frac{(1-\beta_t)||\hat{x}_{t-1}-x_{t-1}||^2}{\beta_t}, \tag{8}$$

where $\sqrt{1-\beta_t}x_{t-1}$ represents the mean of the forward process $q(x_t|x_{t-1})$, and β_t represents its variance.

Component Decoupling Structure. Although our proposed semi-implicit method can better adapt to the complex human motion distribution modeling, thus achieving high-quality rapid generation, the adversarial learning strategy of GAN is characterized by learning the associations of all components within the distribution. It is evident that the motion distribution of fingers is distinct from that of limbs. Limb motion has a larger amplitude and aligns more closely with the melody, while finger motion is smaller and more precise, leaning towards semantic matching. Holistic modeling would lead GAN to fit body data more closely while neglecting finger motion, thereby reducing the expressiveness of the overall gesture. Based on this, we decouple the hands x_{t-1}^H and body x_{t-1}^B to independently learn their denoising distributions. We describe the conditional discriminator D, which depends on the time step and audio control signal, as accepting the noise sequence at step $t-1$, the noise step t, and the conditional feature c. In our adversarial learning strategy, fake samples from distribution $p_\theta(x_{t-1}|x_t)$ will compete with real samples from distribution $q(x_{t-1}|x_t)$. D is a 7-layer MLP network, composed of Linear layers, SELU activation, and GroupNorm layers. All models are trained using the AdamW optimizer with a fixed learning rate L. We employ EMA decay for the optimizer during training. For more training settings, please see the implementation details in the experiment. Finally, our final training objective of proposed method is:

$$\min_{\theta}\max_{D_{adv}}\mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})}[-log(D_{adv}(x_{t-1},c,t))]+[-log(1-D_{adv}(\hat{x}_{t-1},c,t))]+\lambda_{recon}(\mathcal{L}_{simple}+\mathcal{L}_{foot})+\lambda_{AFD}\frac{(1-\beta_t)||\hat{x}_{t-1}-x_{t-1}||^2}{\beta_t}, \tag{9}$$

where λ_{recon} represents the reconstruction weight of the denoiser, and λ_{AFD} represents the weight of the regularization term.

Experiments

In this section, we evaluated the effectiveness of our proposed method in concurrent gesture generation. To verify the generalizability of the approach, we also supplemented the experiments with an extension in the music-driven dance domain. We compared the performance and computational efficiency of our approach with other state-of-the-art techniques in these domains. By applying our method to real-time tasks, we showcased its robust functionality and promising potential. We highly recommend readers to consult the supplementary video material to gain further insights into the qualitative outcomes.

Data and Representation. Our experiments employed three distinct high-quality 3D motion capture datasets: BEATs (Liu et al. 2022a), ZeroEGGs (Ghorbani et al. 2023), and AIST++ (Li et al. 2021). The first two were used for concurrent gesture generation, and the latter for real-time dance synthesis. Each dataset contained style labels. For the BEATs dataset, we selected the corresponding English audio data for model training, while the entirety of the training sets was used for the other datasets. Given that our experiments considered the motion of all joints (body and fingers), the dimension of our motion representation significantly increased. This, coupled with physical challenges such as foot sliding and global displacement jitter, made the training more challenging when compared to half-body methods (Yoon et al. 2020; Ao et al. 2022; Yoon et al. 2022;

	ZeroEGG					BEAT				
	FGD↓	BA↑	DIV↑	Inf. time↓	steps	FGD↓	BA↑	DIV↑	Inf. time↓	steps
DSG (Yang et al. 2023b)	15.67	0.81	0.63	6.01	1000	113.7	**0.89**	0.71	5.92	1000
FreeTalker (Yang et al. 2024)	17.12	0.72	**0.84**	5.5	1000	147.2	0.84	**0.77**	5.5	1000
DiffGesture(re-train) (Zhu et al. 2023)	25.7	0.58	-	4.72	1000	382.6	0.61	-	5.02	1000
Ours	**12.57**	**0.87**	0.76	**0.29**	10	**98.6**	0.88	0.74	**0.38**	10
Ours(DDPM)	11.95	0.92	0.85	11.9	1000	91.7	0.91	0.76	11.94	1000
Trimodal(re-train) (Yoon et al. 2022)	22.4	0.72	0.80	-	-	222.3	0.77	0.68	-	-
HA2G(re-train) (Liu et al. 2022b)	18.8	0.81	0.73	-	-	156.8	0.82	0.70	-	-
CAMN (Liu et al. 2022a)	15.52	0.83	0.77	-	-	258.4	0.73	0.61	-	-

Table 1: Objective Metrics. The FGD evaluation model is trained across the entire training set, with lower evaluation values indicating closer adherence to the original motion distribution. The 'Inf.time' metric is computed by statistically inferring the time taken to generate each frame (measured in milliseconds); lower values signify faster generation speeds.

Liu et al. 2022b; Zhu et al. 2023), but concurrently improved the visual outcomes. We uniformly downsampled the original data to 20FPS for concurrent gesture training and to 30FPS for dance data. Since ZeroEGGs is not in Vicon standard and cannot be directly used to drive the SMPLX standard character model, we converted it to the Vicon standard using MotionBuilder, ensuring consistency with BEATs. Regarding the motion representation for each individual joint, we employed a 6D rotation representation $r \in \mathbb{R}^6$, 3D keypoints $l \in \mathbb{R}^3$, angular velocity $l \in \mathbb{R}^6$, and linear velocity $l \in \mathbb{R}^3$ to describe motion information, while also incorporating gaze direction $z \in \mathbb{R}^3$ to depict head movements during speech. All gesture data was segmented into 80-frame training clips, and dance data into 150-frame segments. To facilitate the generation of smooth, long-frame motions, we appended a tenth of the seed poses to each segment. Given that all datasets provided discrete style labels, we uniformly transformed them into one-hot encodings.

Implementation Details. Our real-time diffusion generation framework is end-to-end, implemented exclusively using Pytorch. Additionally, we developed a script for Blender (Community 2018) that can accept generated motion sequences in real-time, driving character feedback for users and achieving state-of-the-art visual effects in human-computer interaction. In our implementation, using a V100 GPU, we generated 80 frames of concurrent gestures in 0.4 seconds and 150 frames of dance motions in 0.88 seconds. For a fair comparison with contemporary methods, we conducted experiments on a single V100 GPU. By default, we utilized a single A100 GPU for model training. For the concurrent gesture model, we trained the generator and discriminator for 80 hours using a batch size of 128 and learning rates of 3e-5 and 1.25e-4, respectively. The dance model was trained for 48 hours with a batch size of 128 and learning rates of 5e-5 and 1.5e-4. We set the default diffusion steps to 20 (although similar results can be achieved with 10 steps, offering faster speed) for evaluating all metrics. Moreover, in CFG (Ho and Salimans 2022), we set the conditional weight to 3.5. In the denoising model, we set the weights of the KL loss and Geo Loss to 0.5 and 10, respectively. These hyperparameters were found to yield the best empirical results.

Comparison with Existing Methods

We first evaluated the efficiency and generation quality of our method and compared it with other contemporary open-source diffusion and non-diffusion generation techniques. For the ZeroEGGs dataset (Ghorbani et al. 2023), we assessed various styles, including happiness, sadness, anger, aging, neutral, fatigue, and relaxation. For the BEATs dataset (Liu et al. 2022a), we selected four speaker sequences as the data used in this study. All comparison methods employed WavLM for extracting audio features. We split the original dataset into training, validation, and test sets with proportions of 0.8, 0.1, and 0.1, respectively, and trained on the entire training set. During the training process, we also train our DDPM implementation with the same structure. For the BEATs dataset, we retrained DSG, DiffGesture (Zhu et al. 2023) and Trimodal (Yoon et al. 2022) using open-source code.

Quantitative Comparison. It is well-known that evaluating a model's generative capability based solely on a limited number of generated examples is challenging; therefore, we introduce several metrics. (i) Fréchet Gesture Distance (FGD)(Yoon et al. 2020) calculates the distance between the latent feature distributions of generated gestures and actual gestures, thereby assessing gesture quality. Lower FGD values indicate higher motion quality. (ii) We compute the number of frames generated per second during the inference stage to demonstrate our outstanding generation efficiency. (iii) We also compared the Beats Alignment (BA) and Diversity (DIV) on the BEAT dataset using the same evaluation method as described in the original paper. Moreover, as shown in Table 1, it is noteworthy that our semi-implicit structure effectively maintains alignment with the audio control task, preserving the optimal FGD score while achieving nearly a 15-fold speed improvement.

Comparison with Contemporary Acceleration Strategies

In the Table 2, we compared our method with other acceleration strategies tailored for diffusion-based generation methods. Specifically, our experimental results were compared with strategies employing DPM-Solver and the original DDGAN (Wang et al. 2022) and SIDDMs (Xu et al. 2023) acceleration. The experiments showed that for the DPM-Solver strategy, its first-order Taylor expansion form corresponds to the well-known DDIM sampling strategy. Accelerating sampling merely by reducing the sampling step size often leads to inaccurate approximations for complex

	ZeroEGG					BEAT				
	FGD↓	BA↑	DIV↑	Inf. time↓	steps	FGD↓	BA↑	DIV↑	Inf. time↓	steps
DPM-Solver-1(DDIM) (Lu et al. 2022)	21.7	0.76	0.60	0.27	10	203.7	0.71	0.66	0.17	10
DPM-Solver-2 (Lu et al. 2022)	19.92	0.70	0.54	0.39	10	148.0	0.68	0.71	0.47	10
Naive DDGAN (Wang et al. 2022)	44.7	0.47	-	0.22	10	334.2	-	-	0.25	10
Naive SIDDMs (Xu et al. 2023)	49.2	0.49	-	0.27	10	309.0	-	-	0.28	10
Ours	**12.57**	**0.87**	0.76	**0.29**	10	**98.6**	0.88	0.74	**0.38**	10

Table 2: Comparative results with contemporary accelerated diffusion methods are presented. Ensuring fairness, all methods employ 10-step denoising.

multi-modal distributions, resulting in a sharp decline in quality. Due to the presence of second-order derivatives, the second-order Taylor expansion requires calling the denoising function twice at midpoint positions during sampling, thus reducing generation speed while providing limited improvement in generation quality. Higher-order Taylor expansions require more frequent calls to the denoising function, which contradicts the primary goal of acceleration.

Moreover, we also compared the original DDGAN with SIDDMs acceleration strategy. Specifically, we trained an unconditional discriminator for DiffGes on the BEAT dataset, eliminating explicit geometric loss (which is different from our method), and compared it with our approach. The results (as shown in the table) indicate that the original configuration of implicit and semi-implicit strategies perform extremely poorly in terms of generated global gesture quality. This is because, unlike images, human representations typically have more stringent geometric conditions that require more specific constraints.

Ablation Studies

In this section, we investigated the impact of diffusion steps, reconstruction loss weight, and auxiliary forward loss on model performance. All ablation studies were conducted on the ZeroEGGs dataset, and the results are presented in Table 3.

Sampling Step Ablation. In this experiment, we study the effect of different sampling steps on model performance. Specifically, we train models with the same structure using 1, 5, 10, 20, 30, and 50 steps, respectively. The final results, as shown in Table 3, indicate that the FGD value stabilizes after 10 steps, but an increase in the number of steps also leads to slower speed. When the number of steps is 1, our structure reverts to a traditional GAN model, and the generated gesture quality declines sharply.

Reconstruction Loss Impact. When the reconstruction loss weight is set to 0, the gesture quality degrades significantly. The introduction of explicit reconstruction loss notably improves the gesture quality. Based on empirical evidence, we set the weight for the experiment to 1, 10, 100 and observe that the weight magnitude does not affect the FGD metric and generation quality. The results without reconstruction loss can be seen in the accompanying video.

Forward Noise Constraint Impact. The term "w/o" indicates that we eliminate the forward loss and train the model solely based on adversarial learning. It is worth noting that pure adversarial training, although capable of helping the model learn the marginal distribution, still results in a large gap between the posterior sampling outcome and the forward process noise at the same time step due to the complexity of the large-step distribution, leading to a significant decline in FGD.

Steps			Recon loss		AFD loss	
num	FGD↓	Inf. time↓	weight	FGD ↓		FGD↓
1	85.44	0.03	0	93.10	w/o	26.9
5	20.91	0.23	1	12.74	w/	**12.32**
10	12.57	**0.29**	10	**12.32**		
20	12.32	0.4	100	12.38		
30	12.71	0.64	-	-		
50	**12.02**	0.99	-	-		

Table 3: Ablation Experiments. We find that when the diffusion steps reach 10, the FGD metric stabilizes, while it degrades sharply when reduced to a traditional GAN network with only one step. The absence of body reconstruction loss severely impacts the generated quality, but the weight of this constraint has little influence on model learning. AFD can explicitly constrain the difference between forward noise and noise sampled from the denoising distribution at the same time step, further enhancing the quality of the generated gesture sequences.

Conclusion

In this paper, we address the speed limitations of diffusion models in generating collaborative speech gestures and explore the challenge of modeling complex denoising distributions across multiple sampling steps. Unlike previous text-to-motion approaches, we introduce implicit marginal constraints based on audio control signals and explicit auxiliary forward diffusion regularization. This improves the model's ability to fit audio control, perform denoising with larger step sizes, and reduce the number of steps, resulting in faster inference. Additionally, we decouple body and finger movements and independently model finger and body distributions. Our approach generates more diverse finger movements while maintaining stability compared to non-physically-based training methods, enhancing the user's viewing experience. These groundbreaking attempts have significantly accelerated DIDiffGes while maintaining high-fidelity generation results, providing new insights for future real-time simultaneous gesture generation tasks.

References

Ahuja, C.; Lee, D. W.; Nakano, Y. I.; and Morency, L.-P. 2020. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, 248–265. Springer.

Ao, T.; Gao, Q.; Lou, Y.; Chen, B.; and Liu, L. 2022. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. volume 41, 1–19. ACM New York, NY, USA.

Ao, T.; Zhang, Z.; and Liu, L. 2023. GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents. *ACM Trans. Graph.*, 42(4): 42:1–42:18.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18000–18010.

Community, B. O. 2018. Blender—a 3D modelling and rendering package. *Amsterdam: Blender Foundation, Stichting Blender Foundation*.

Ghorbani, S.; Ferstl, Y.; Holden, D.; Troje, N. F.; and Carbonneau, M.-A. 2023. ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech. In *Computer Graphics Forum*, volume 42, 206–216. Wiley Online Library.

Habibie, I.; Elgharib, M.; Sarkar, K.; Abdullah, A.; Nyatsanga, S.; Neff, M.; and Theobalt, C. 2022. A Motion Matching-based Framework for Controllable Gesture Synthesis from Speech. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–9.

Habibie, I.; Xu, W.; Mehta, D.; Liu, L.; Seidel, H.-P.; Pons-Moll, G.; Elgharib, M.; and Theobalt, C. 2021. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, 101–108.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. volume 33, 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13401–13412.

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

Liu, X.; Wu, Q.; Zhou, H.; Du, Y.; Wu, W.; Lin, D.; and Liu, Z. 2022b. Audio-Driven Co-Speech Gesture Video Generation. volume 35, 21386–21399.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

Schulman, J.; Zoph, B.; Kim, C.; Hilton, J.; Menick, J.; Weng, J.; Uribe, J. F. C.; Fedus, L.; Metz, L.; Pokorny, M.; et al. 2022. ChatGPT: Optimizing language models for dialogue. *OpenAI blog*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; and Bermano, A. H. 2022. Human motion diffusion model.

Wang, Z.; Zheng, H.; He, P.; Chen, W.; and Zhou, M. 2022. Diffusion-gan: Training gans with diffusion. *arXiv preprint arXiv:2206.02262*.

Xie, P.; Zhang, Q.; Li, Z.; Tang, H.; Du, Y.; and Hu, X. 2022. Vector quantized diffusion model with codeunet for text-to-sign pose sequences generation. *arXiv preprint arXiv:2208.09141*.

Xu, Y.; Gong, M.; Xie, S.; Wei, W.; Grundmann, M.; Hou, T.; et al. 2023. Semi-Implicit Denoising Diffusion Models (SIDDMs). *arXiv preprint arXiv:2306.12511*.

Yang, S.; Wang, Z.; Wu, Z.; Li, M.; Zhang, Z.; Huang, Q.; Hao, L.; Xu, S.; Wu, X.; Dai, Z.; et al. 2023a. UnifiedGesture: A Unified Gesture Synthesis Model for Multiple Skeletons.

Yang, S.; Wu, Z.; Li, M.; Zhang, Z.; Hao, L.; Bao, W.; Cheng, M.; and Xiao, L. 2023b. DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models.

Yang, S.; Xu, Z.; Xue, H.; Cheng, Y.; Huang, S.; Gong, M.; and Wu, Z. 2024. Freetalker: Controllable Speech and Text-Driven Gesture Generation Based on Diffusion Models for Enhanced Speaker Naturalness. *arXiv preprint arXiv:2401.03476*.

Yi, H.; Liang, H.; Liu, Y.; Cao, Q.; Wen, Y.; Bolkart, T.; Tao, D.; and Black, M. J. 2023. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 469–480.

Yoon, Y.; Cha, B.; Lee, J.-H.; Jang, M.; Lee, J.; Kim, J.; and Lee, G. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. volume 39, 1–16. ACM New York, NY, USA.

Yoon, Y.; Wolfert, P.; Kucherenko, T.; Viegas, C.; Nikolov, T.; Tsakov, M.; and Henter, G. E. 2022. The GENEA Challenge 2022: A large evaluation of data-driven co-speech ges-

ture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, 736–747.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*.

Zhou, C.; Bian, T.; and Chen, K. 2022. Gesturemaster: Graph-based speech-driven gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, 764–770.

Zhu, L.; Liu, X.; Liu, X.; Qian, R.; Liu, Z.; and Yu, L. 2023. Taming Diffusion Models for Audio-Driven Co-Speech Gesture Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10544–10553.