Conditional GAN for Enhancing Diffusion Models in Efficient and Authentic Global Gesture Generation from Audios

Yongkang Cheng[*]
Tencent AI Lab
cyk19990422@gmail.com

Mingjiang Liang[*]
University of Technology Sydney
breeliang@gmail.com

Shaoli Huang[†]
Tencent AI Lab
shaol.huang@gmail.com

Gaoge Han
Tencent AI Lab
hangaoge@nwafu.edu.cn

Jifeng Ning
Northwest A&F University
njf@nwsuaf.edu.cn

Wei Liu
University of Technology Sydney
wei.liu@uts.edu.au

Abstract

Audio-driven simultaneous gesture generation is vital for human-computer communication, AI games, and film production. While previous research has shown promise, there are still limitations. Methods based on VAEs are accompanied by issues of local jitter and global instability, whereas methods based on diffusion models are hampered by low generation efficiency. This is because the denoising process of DDPM in the latter relies on the assumption that the noise added at each step is sampled from a unimodal distribution, and the noise values are small. DDIM borrows the idea from the Euler method for solving differential equations, disrupts the Markov chain process, and increases the noise step size to reduce the number of denoising steps, thereby accelerating generation. However, simply increasing the step size during the step-by-step denoising process causes the results to gradually deviate from the original data distribution, leading to a significant drop in the quality of the generated actions and the emergence of unnatural artifacts. In this paper, we break the assumptions of DDPM and achieves breakthrough progress in denoising speed and fidelity. Specifically, we introduce a conditional GAN to capture audio control signals and implicitly match the multimodal denoising distribution between the diffusion and denoising steps within the same sampling step, aiming to sample larger noise values and apply fewer denoising steps for high-speed generation. In addition, to enable the model to generate high-fidelity global gestures and avoid artifacts, we introduce an explicit motion geometric loss to enhance the quality and global stability of the generated gestures. Numerous qualitative and quantitative experiments show that compared to contemporary diffusion-based methods, our method offers faster generation speed and higher fidelity, and compared to non-diffusion methods, it provides a more stable global effect and a more natural user experience.

1. Introduction

Body movements and gesture sequences play a pivotal role in human communication by conveying non-verbal cues to accurately express intent. The objective of simultaneous gesture generation is to create diverse and dynamic human postures synchronized with audio cues. Generating spontaneous co-speech gestures for non-player characters (NPCs) in virtual reality, gaming, and film can offer a more immersive user experience, a topic that has recently garnered substantial research interest. Some generation methods based on GANs and VAEs show promise but are limited by strict motion representation requirements and jittery arm posture generation, respectively. Diffusion-based generation methods, such as DiffGesture, are at the cutting edge in terms of generation quality, as they can effectively capture target distributions of various gestures. However, these methods suffer from low generation efficiency; for instance, DiffGesture requires 6 seconds to generate an 88-frame gesture sequence slice. The inability to achieve rapid generation while maintaining high gesture quality hinders their effectiveness in real-world applications, such as real-time conversational virtual humans.

With concerted efforts, strategies to mitigate the low efficiency of diffusion-based generation methods have bifurcated into two primary approaches: 1) latent space diffusion, and 2) the DPM-Solver technique. The former, initially proposed by MLD and validated in GestureClip, learns the latent space of global gestures and subsequently performs diffusion on low-dimensional, high-density latent variables. However, this two-stage method's performance

---

[*]Yongkang Cheng and Mingjiang Liang are co-first authors.
[†]Corresponding author.

(a) ZeroEGGs

Inference Time: 0.37s/80frames

(b) BEAT

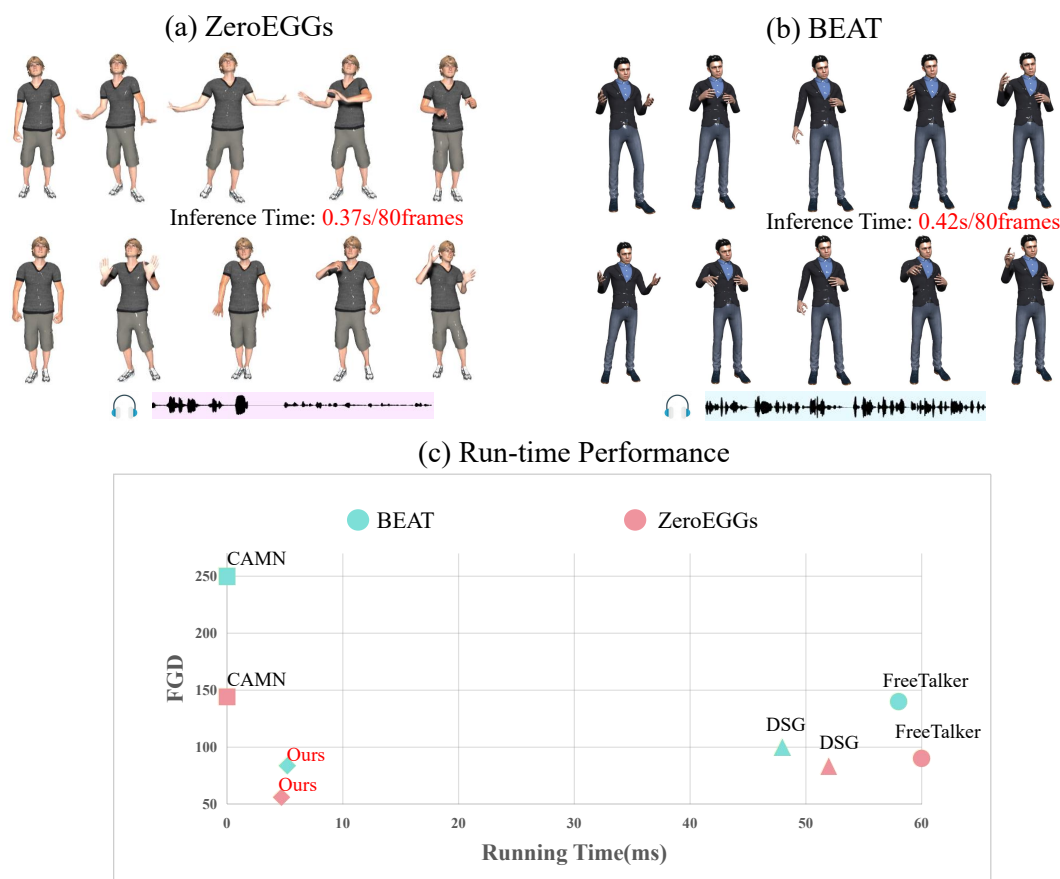Inference Time: 0.42s/80frames

(c) Run-time Performance

Figure 1. Our method generates global gesture motion highly matched with input audio melody in a short runtime. On the (a) BEAT dataset and (b) ZeroEGGs dataset, our method's average runtime is 5ms and 4.6ms per frame, respectively. For reference, DSG's corresponding times are 5.2 and 4.8s. (c) The overall comparison of inference time cost and FGD on both datasets. In the figure, we compare the per-frame runtime with the state-of-the-art methods' FGD.

is contingent on the gesture embedding space learned in the initial stage, making the learning of an effective embedding space a challenging task. The expressiveness of this space often dictates the outcome of downstream gesture generation. Conversely, the DPM-Solver approach, tailored explicitly for diffusion models, commences with the standard 1000-step DDPM denoising method. It first analyzes linear terms in diffusion steps, substitutes complex integral terms using the logarithm of the signal-to-noise ratio, and finally obtains a noise estimate via Taylor expansion of the exponential integral of the noise prediction model. Despite the accurate computation of known terms, approximation estimation of the neural network part still introduces discretization errors. Moreover, the first-order expansion form of DPM-Solver, the DDIM sampling strategy, expedites generation by increasing step size and decreasing step count. However, it overlooks that DDPM's denoising is predicated on a Gaussian distribution assumption with minuscule noise values at each step. This oversight leads to the DDIM sam-

pling strategy skipping many reverse steps, resulting in sampled noise often emanating from a denoising distribution more intricate than the Gaussian distribution, thereby compromising the generation effect. Higher-order expansion methods, due to the presence of higher-order derivatives, necessitate multiple invocations of the denoising function, thereby prolonging the time spent per step, as demonstrated in the quantitative experiments presented in Section 4. Consequently, accurately capturing more complex multimodal distributions becomes an imperative for several step sampling methods.

Compared to DDIM's strategy of directly stacking denoising steps, we have astutely pinpointed the root of the problem. Upon reevaluating the denoising process of DiffGesture, we discern that the method hinges on the assumption that the posterior noise added at each denoising step is sampled from a unimodal distribution and is relatively small. With a sufficient number of steps, it can denoise the original motion sequence from the standard Gaussian distri-

bution. Consequently, the inefficiency of previous methods stems from small step noise sampling and a large number of sampling steps. Motivated by this challenge, and drawing inspiration from the latest work in the fast text-to-image domain [23], we introduce a framework for swiftly generating expressive speech gesture sequences. To the best of our knowledge, this is the inaugural work in the speech gesture generation domain to consider real-time performance, thereby breaking free from the constraints of prior methods and paving the way for advancements in this field. Our objective is to increase step size and diminish generation steps, essentially modeling the intricate multi-step sampling process. Unlike the fully implicit matching constraint method of DDGAN [21] in the image generation domain, we incorporate an explicit constraint on motion geometric loss to circumvent the introduction of detrimental motion noise during the training process.

Specifically, we guide the GAN to model the distribution by conditioning on the time step t, enabling the model to capture complex denoising distributions with a specified number of steps (or step sizes). Simultaneously, we use audio signals (style labels and spectral features) as control conditions to guide the model in capturing the distribution of heterogeneous motions. This implicit matching adversarial learning strategy employs a conditional GAN to match the conditional distribution between the diffusion and denoising processes, allowing for the addition of large random noise between adjacent diffusion steps, and achieving denoising in just a few steps. Moreover, we introduce explicit motion constraints in the form of motion geometric loss to alleviate motion jitter issues. Notably, our framework is an easy-to-train end-to-end model.

Our method's capability to swiftly generate high-quality co-speech gesture sequences is showcased in Figure 1, achieving a time cost reduction of approximately 12.35 times compared to DiffuseStyleGesture. Additionally, we carried out extensive experimental assessments, substantiating that our approach considerably surpasses contemporary diffusion-based methods in terms of generation efficiency and greatly outperforms non-diffusion methods in quality.

2. Related Work

Co-speech gesture generation is an exceptionally intricate task that necessitates understanding speech melody, semantics, gesture motion, and their mutual relationships. Early data-driven approaches (e.g., those proposed in [8, 16]) strive to learn gesture matching from human demonstrations but often yield limited motion diversity. Subsequent research (such as [8, 22, 27]) bolsters models' capacity to generate diverse outcomes and introduces the notion of generating unique and expressive gesture results. Some studies, like [1,2,24,25], train a unified model for multiple speakers, embedding each speaker's style in the space or incorporat-

ing style transfer techniques. Other works [7, 31] employ motion matching to generate gesture sequences, but these methods often demand intricate matching rules. Despite the challenges, audio-driven animation has garnered considerable attention.

Diffusion generation models have garnered remarkable achievements across multiple domains [3, 17, 18], particularly in simultaneous gesture generation. Specifically, DiffGesture [32] pioneers the integration of diffusion models into audio-controlled gesture generation methods, recovering gesture actions through stepwise noise prediction. DiffStyleGes [25] is a landmark work that introduces motion diffusion models to reconstruct original motion representations by learning the relationship between motion representations and input audio control conditions, while also providing stylized hard labels for controlling the generated gesture style. Recent research, FreeTalker [26] and ExpGest [5], further concentrates on global gesture generation methods under the joint control of text and audio. In our work, we leverage attention mechanisms to capture the relationship between gesture sequences and speech, generating results that closely match speech. However, due to the high dimensionality and iterative nature of diffusion models, runtime generation based on the original diffusion model DDPM [13] has been beset by time overhead. GestureClip [2] introduces the concept of latent diffusion into runtime generation, reducing computational resource requirements by diffusing from low-dimensional, high-density embedding spaces through latent variable sampling. Its core idea entails first training a VAE for gesture motion embedding and subsequently applying latent diffusion in the learned latent space. However, this two-stage method is a non-end-to-end training approach. Moreover, merely stacking noise steps and discarding redundant denoising steps result in artifacts that are unacceptable to humans. In contrast, our method is an end-to-end model. In this paper, we propose a conditional GAN for implicit matching constraints, modeling complex distributions (potentially multimodal distributions that deviate from the DDPM [10] assumption) in multi-step processes through contrastive learning strategies. Simultaneously, we explicitly constrain the human body geometric loss, maintaining high fidelity of human motion while enhancing generation speed.
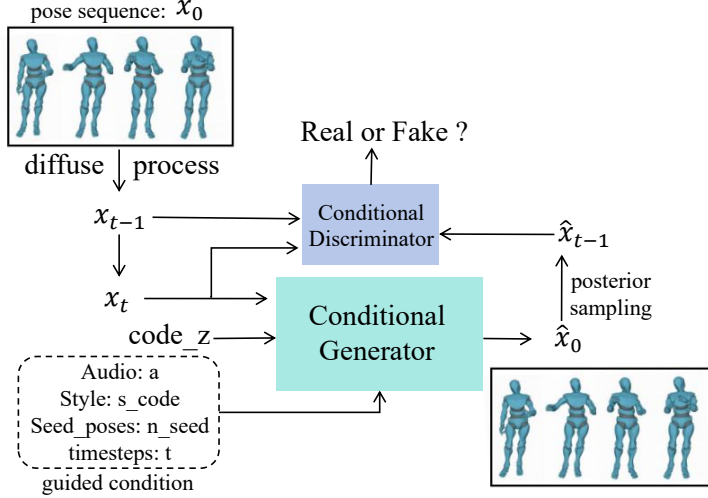
3. Method

Our objective is to accelerate the denoising process with fewer steps and larger step sizes, enabling our model to efficiently generate high-fidelity and diverse co-speech gesture sequences based on the input audio signal in a shorter time.

3.1. Diffusion Model for Generating Gestures

Gesture sequences are represented as $x^{1:N}$ and are associated with corresponding conditional control signals c,
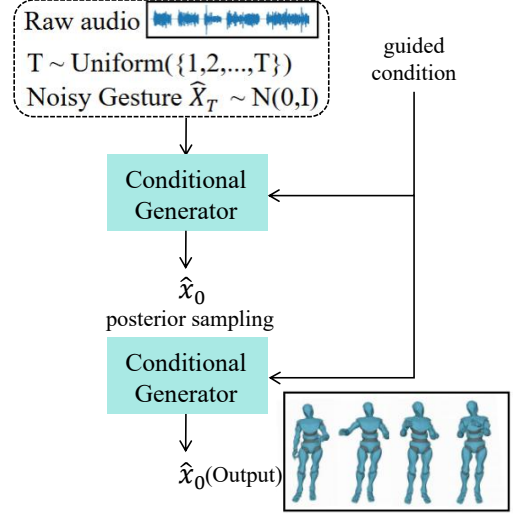
2166

Figure 2. **Network architecture.** During training, we introduce a GAN structure based on conditional denoising diffusion to capture the complex distribution of gesture sequences in a multi-step process, enabling larger sampling step sizes. During inference, we use large step sizes and fewer steps for sampling, according to the input audio control signal, to achieve fast, high-quality gesture sequences, thus supporting real-time tasks.

such as audio signals [25], where N denotes the number of frames in the gesture sequence. Unconditional gesture generation can also be achieved by setting the conditional signal to $\emptyset$. In this paper, we use a diffusion probabilistic model [10] to generate co-speech gestures, where the diffusion model gradually anneals pure Gaussian noise into a gesture distribution $p(x)$. Consequently, the model can predict noise from a T-timestep 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:

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

where $\{\beta_t\}_{t=1}^T$ is the variance schedule and $\alpha_t = \prod_{s=1}^t (1-\beta_s)$. Then the reverse process becomes $p_\theta(x_{0:T}) := p(x_T)\prod_{t=1}^T p_\theta(x_{t-1}|x_t)$, starting from $x_T \sim \mathcal{N}(0, I)$ with noise predictor ϵ_t^θ:

$$x_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_t^\theta(x_t)) + \sigma_t z_t, \quad (2)$$

where $z_t \sim \mathcal{N}(0, I)$ and $\sigma_t^2 = \beta_t$ means the variance schedule stays constant.

However, unlike the original DDPM [10], we consider the extreme physical constraints inherent in 3D human bodies and deviate from image generation by replacing predicted noise with original human representations. Consequently, at each step of the denoising process, we reconstruct the original representation from pure Gaussian noise, ultimately producing the final generation results through a cyclic process of noise addition and denoising:

$$\hat{x}_0 = \epsilon_t^\theta(x_t|c), x_{t-1} = \frac{1-\alpha_{t-1}+\sqrt{\alpha_{t-1}}\hat{x}_0}{1-\alpha_t\hat{x}_0}+\sigma_t z_t, \quad (3)$$

where c is the conditions.

3.2. Implicit Joint Distribution Matching

Although the conditional DDPM-based methods can generate satisfactory results, the 1000-step denoising process renders these impressive outcomes inapplicable to real-time tasks in practice. If we aim to achieve fast sampling by reducing the denoising steps to a few dozen, the Gaussian assumption followed by DDPM's $p_\theta(x_{t-1}|x_t)$ process (noise sampling from a unimodal distribution with a small value at each step) would no longer hold, and L2 reconstruction could not be applied to model the KL divergence.

To address this issue, we introduce a GAN structure based on conditional denoising diffusion, which inherits the Conditional Gesture Generator and Conditional Gesture Discriminator, as shown in Figure 2. We consider the time step t, audio control signal c, and seed gesture in conditional control simultaneously to implement a divide-and-conquer approach to effectively help the model learn the conditional denoising distribution. Additionally, we propose an

adversarial learning strategy, using a conditional GAN to match the conditional distribution between $q(x_{t-1}|x_t)$ and $p_\theta(x_{t-1}|x_t)$, and adding random large noise between adjacent diffusion steps to achieve rapid denoising. The formulation can be summarized as follows:

$$\min_\theta \max_{D_{adv}} \sum_{t>0} \mathbb{E}_{q(x_t)} D_{adv}(q(x_{t-1}|x_t)||p_\theta(x_{t-1}|x_t)), \quad (4)$$

More specifically, we will denote the Conditional Gesture Discriminator relying on time steps and control signals as $D_\emptyset(x_{t-1}, x_t, c, t)$. Here, x_{t-1} and x_t represent the noise at time steps $t-1$ and t, respectively, while c denotes the joint of audio control signals and seed gestures. The conditional GAN's structure is shown in Figure 3. In our adversarial learning strategy, fake samples from distribution $p_\theta(x_{t-1}|x_t)$ will be pitted against real samples from distribution $q(x_{t-1}|x_t)$. Given the joint distribution formula $q(x_t, x_{t-1}) = \int dx_0 q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})$, our discriminator can be trained as follows:

$$\min_\theta \sum_{t>0} \mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})}[-log(D_\emptyset(x_{t-1}, x_t, c, t))]$$
$$+\mathbb{E}_{p_\theta(x_{t-1}|x_t)}[-log(1 - D_\emptyset(x_{t-1}, x_t, c, t))], \quad (5)$$

This rewritten formulation (D-real) of conditional GAN allows us to model $p_\theta(x_{t-1}|x_t)$ more flexibly. Given the training goal of the Conditional Gesture Discriminator, we also need to train the Conditional Gesture Generator, whose formula (G-fake) is defined as follows:

$$\max_\theta \mathbb{E}_{t\ [1,T], q(x_t)} \mathbb{E}_{p_\theta(x_{t-1}|x_t)}[-log(-D_\emptyset(x_{t-1}, x_t, c, t))], \quad (6)$$

3.3. Explicit Geometric Constraints

Although the conditional GAN based on control signals seems to reasonably address the large stride and small step count of high-speed sampling, the purely implicit adversarial learning for the connected x_{t-1} and x_t is statistically inefficient, particularly when x_t is a high-dimensional redundant representation (which is often the case in human body sequences). This might also be the reason for the sharp decline in the performance of DDGAN on complex data distributions. Furthermore, we observe that after the adversarial process of the conditional GAN, the generator often produces gestures with artifacts. We speculate that this is because the purely implicit matching constraint on the joint distribution cannot provide geometric constraints for the gestures. Therefore, we propose an explicit geometric constraint for training the generator to improve the quality of the generated gestures. More specifically, following the approach of MDM [20], we directly predict clean gesture sequences with $\hat{x}_0 = G(x_t, z, c, t)$, and we employ a Huber loss [11] to constrain its reconstruction loss:

$$\mathcal{L}_{recon} = E_{x_0\ q(x_0|c), t\ [1,T]}[HuberLoss(x_0 - \hat{x}_0)], \quad (7)$$
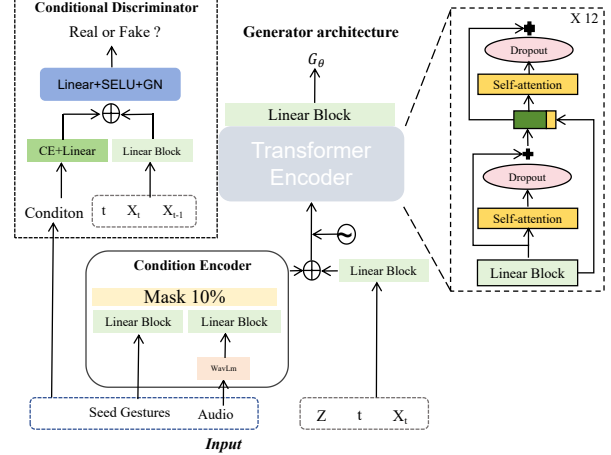

(c) Structure Details

Figure 3. **Conditional GAN architecture.** The Conditional Gesture Generator is employed for the audio-to-gesture task, where z is the latent variable sampled from a Gaussian distribution. It consists of a Transformer Encoder, Linear, and Mask components. The Conditional Gesture Discriminator is utilized to distinguish between genuine and counterfeit gesture sequences, comprising Linear, SELU, and Group Norm modules.

Lastly, we combine the implicit joint distribution matching constraint with the explicit gesture geometric constraint to train the Conditional Gesture Generator (also referred to as the denoiser).

4. Experiments

In this section, we will delve into the technical complexity of our model, comparisons with other frameworks, potential applications, and the validation of design choices for each module through ablation studies.

Implementation Details. As depicted in the right portion of Figure 3, we construct a Conditional Gesture Generator based on the Transformer architecture, comprising a 12-layer, 8-head encoder with default skip connections. The conditional control signal encoder consists of linear layers and WavLM. The linear block maps the seed pose and style label linearly. The audio input is initially encoded by WavLM to obtain spectral information, then linearly mapped to a space sharing the same dimensions as the former. As shown in the top-left corner of Figure 3, the Conditional Gesture Discriminator is a 7-layer MLP network, composed of linear layers, SELU activations, and Group-Norm layers. All models are trained employing the AdamW optimizer with a fixed learning rate of 1e-5. During training, we apply EMA decay to the optimizer. In the diffusion training phase, our mini-batch size is set at 64. Based on empirical values for GAN network training, we establish the generator's learning rate at 3e-5 and the discriminator's

5

2*Method	BEAT				ZeroEGGs			
	FGD↓	BA↓	DIV↓	$ms/frame$	FGD↓	BA↓	DIV↓	$ms/frame$
GroundTruth	-	0.93	0.88	-	-	0.95	0.81	-
CAMN [15]	258.4	0.73	0.61	-	137.4	0.69	0.70	
Trimodal(re-train) [28]	222.3	0.77	0.68	-	183.1	0.72	0.69	-
FreeTalker [26]	147.2	0.84	0.53	55.0	82.7	0.79	0.63	49.7
DiffStyleGesture(re-train) [25]	100.3	**0.89**	0.71	75.0	78.4	0.82	0.80	77.4
DiffGes(re-train) [32]	247.7	0.66	**0.82**	42.3	157.4	0.68	0.88	40.0
Ours	**84.0**	0.87	0.76	**3.2**	**54.7**	**0.83**	**0.88**	**4.1**

Table 1. **Comparison.** Bold indicates the best-performing method for each evaluation criterion. All experiments are conducted and evaluated on the BEAT and ZeroEgg dataset. The results for CAMN [15] are reproduced using the official code, while the results for DiffStyleGesture [25], DiffGes [32] and Trimodal [28] are obtained by retraining the official code. Dashes ("–") indicate that specific metrics are not available.

learning rate at 1.25e-4 for stable adversarial training. For the ZeroEGGs dataset training task, conditioning a denoising diffusion GAN requires approximately 80 hours on a single A100 GPU. During inference evaluation, we execute all experiments five times on a single V100 GPU and calculate the average.

4.1. Datasets and Evaluation Metrics

Datasets. Our audio data is derived from the large-scale, high-quality **BEAT** dataset [15] and the **ZeroEGGs** [6] dataset. The BEAT dataset encompasses 76 hours of multimodal speech data from 30 speakers, featuring 8 emotions and 4 languages. We selected the **English** data from demonstrations and conversations for training. On the other hand, the ZeroEGGs dataset is a highly expressive collection of high-quality gesture sequences, containing nearly 2 hours of full-body motion capture and monologue videos performed by a female actor in 19 distinct styles.

Evaluation Metrics. We concurrently assess the quality (FGD), beat alignment (BA), and diversity (DIV) of gestures generated by our method. Gesture quality is evaluated using Fréchet Gesture Distance (FGD) [28], which computes the distance between the latent feature distributions of generated and real gestures, thereby gauging gesture quality. Lower FGD values signify higher motion quality. Beat Alignment (BA) [12] represents the inverse angular distance between audio and gesture beats, and is employed to appraise the similarity of gesture audio beats; the higher the value, the better the alignment with audio beats. DIV measures the L1 distance among multiple body gestures produced under the same control signal, with larger values indicating greater diversity. We adhere to the same evaluation implementation as in the BEAT paper.

4.2. Comparison to Existing Methods

We juxtapose our approach with existing diffusion-based and non-diffusion-based methods. On the BEAT dataset,
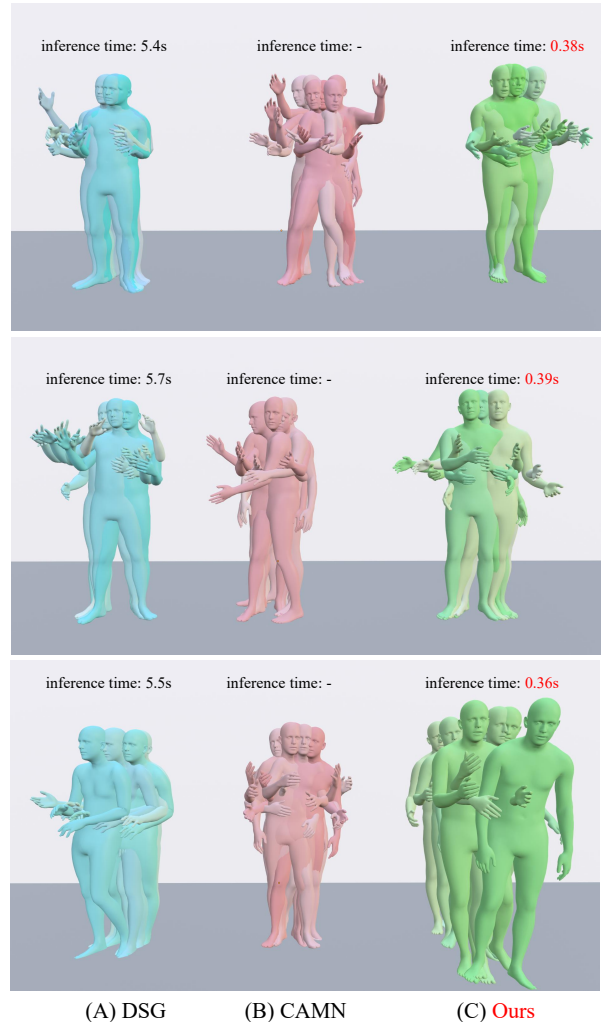

Figure 4. In comparison to contemporary diffusion-based methods such as DSG and non-diffusion methods like CAMN, our approach achieves high-speed generation while ensuring the utmost generation quality.

as illustrated in Table 2, non-diffusion methods frequently grapple with jitter issues in generated results, resulting in highly unnatural movements and inferior FGD scores. Conversely, diffusion-based methods are encumbered by the high generation time costs; for instance, DiffStyleGesture attains respectable FGD scores, but its generation speed approaches 75 milliseconds per frame, constraining its potential for deployment in real-time tasks. In contrast, our large-step, fewer-step method achieves a breakthrough in this domain, enhancing generation speed by over seven times to a mere 4.2 milliseconds per frame. This improvement allows a 4-second slice to be generated within 0.3 seconds, rendering it highly suitable for real-time tasks without any compromise in generation quality compared to DSG.

Qualitative results, as displayed in Figure 4, indicate that our method outperforms the previously best diffusion-based generation method, DSG, in terms of gesture effects while maintaining high-speed generation capabilities. The generation speed during the inference phase has improved nearly 13 times compared to DSG, necessitating only about 0.4 seconds to generate 80 frames (4 seconds) of gesture outcomes. This advancement renders our method suitable for real-time tasks, such as virtual human communication and real-time human-machine interaction. In contrast to non-diffusion methods like CAMN, our approach showcases more stable joint movements and avoids unnatural global jitter issues.

4.3. Contrasting with Other Speedup Approaches

In the table, we juxtapose our approach with other acceleration strategies tailored for diffusion-based generation methods. Specifically, our experimental results are contrasted with strategies employing DPM-Solver and the original DDGAN method. The experiments reveal that for the DPM-Solver strategy, its first-order Taylor expansion form corresponds to the well-known DDIM sampling strategy. Accelerating sampling by merely reducing the sampling step size often results in an imprecise approximation of complex multimodal distributions, causing a drastic deterioration in quality. The second-order Taylor expansion, due to the presence of the second-order derivative, necessitates invoking the denoising function twice at the midpoint position during sampling, consequently diminishing generation speed while providing limited enhancement in generation quality. Higher-order Taylor expansions demand even more frequent calls to the denoising function, which contradicts the primary objective of acceleration.

Furthermore, we directly integrate the recent work, DDGAN, with DiffGes for comparison. Specifically, we train an unconditional discriminator for DiffGes on the BEAT dataset, eliminate explicit geometric loss, and compare it with our method. The results, as displayed in the table, indicate that the DiffGes+DDGAN configuration performs exceedingly poorly concerning the generated global gesture quality. This is attributable to the fact that, unlike images, human representations often possess more stringent geometric conditions and necessitate more specific constraints.

4.4. Ablation Study

In this section, we verify the efficacy of our designed key modules. All experiments are carried out on the BEAT dataset. Given that the BEAT dataset encompasses data from 30 English speakers, resulting in substantial training costs, we opt for sequences from three speakers for validation purposes. All modules employ identical experimental configurations to guarantee fairness.

Structural Ablation. As illustrated in Table 3, concerning audio control information, when substituting WavLM features with the first 13 coefficients of traditional MFCC, the generated results' performance fails to meet users' subjective perception due to the absence of further extraction of semantic and emotional information from audio input latencies. As a result, the scores markedly decline. Upon replacing our Transformer-Encoder with a GRU structure, user scores experience a significant reduction. This is attributable to the inherent asynchrony between audio and gestures, where the attention mechanism proves more reliable in capturing such asynchronous relationships compared to GRU. Simultaneously, given that in actual communication, arms generally express emotional styles while fingers convey semantics, we devise a decoupled structure for both, employing learnable weights for balance. This approach allows our method to surpass DSG. It is worth noting that during structural ablation, we consistently utilize the model comprising 20 diffusion steps.

For the ablation experiment evaluation, we primarily assess human subjective perception consistency metrics, evaluating from two dimensions. The first two dimensions follow GENEA [29], assessing human likeness (HL) and gesture-audio appropriateness (GA). The third dimension is gesture style matching (SA). To understand our method's true visual performance, we conducted user studies comparing generated gesture sequences with real motion capture data. Evaluation clips range from 16 to 60 seconds, averaging 38.1 seconds. We recruited 60 participants in three groups, with scores ranging from 5 to 1, labeled as "excellent," "good," "average," "below average," and "poor."

Sampling Step Ablation. In this experiment, we studied the impact of different sampling steps on model performance on the BEAT dataset. Specifically, we trained models with the same structure using 1, 5, 10, 20, 30, and 50 steps, respectively. The final results, as shown in Table 4, indicate that fewer steps result in a higher sampling speed. However, an excessively large step size causes a sharp decline in quality. When the number of steps is 1, our struc-

7

2*Method	BEAT				ZeroEGGs			
	FGD↓	BA↑	DIV↑	$ms/frame$	FGD↓	BA↑	DIV↑	$ms/frame$↓
DPM-Solver-1(DDIM)	151.2	0.74	0.63	3.6	99.8	0.76	0.70	4.5
DPM-Solver-2	144.7	0.70	0.68	5.8	119.6	0.72	0.67	6.9
DiffGes+DDGAN	374.8	0.53	-	3.5	352.8	0.55	-	**3.9**
Ours	**84.0**	0.87	0.76	**3.2**	**54.7**	0.83	0.88	4.1

Table 2. Comparative results with contemporary accelerated diffusion methods are presented. Ensuring fairness, all methods employ 10-step denoising. The original DDGAN, lacking geometric constraints, produces near-static outcomes, while DPM-Solver's higher-order components necessitate multiple denoising function calls, impacting generation speed. Our approach strikes a delicate balance between speed and quality.

1*Method	$SA \uparrow$	$HL \uparrow$	$GA \uparrow$	$ms/frame \downarrow$
Ground Truth	0.86	4.41	4.72	-
DiffStyleGesture	0.80	3.97	3.74	75.0
Ours	0.76	4.03	3.82	3.2
w/o WavLM	0.67	3.79	3.77	4.01
w/ GRU	0.72	3.71	3.76	3.08

Table 3. **Ablation study on structure.** 'w/' means with and 'w/o' means without.

1*Weights	$FGD \downarrow$	$BA \uparrow$	$SA \uparrow$
0	988.7	0.52	-
1	88.2	0.85	-
10	**84.6**	**0.87**	0.71
100	86.7	0.83	**0.74**

Table 5. Ablation study on Geometric Constraints. All models adopt 20 diffusion steps.

1*Steps	$FGD \downarrow$	$BA \uparrow$	$SA \uparrow$	$ms/frame \downarrow$
1	1174.4	0.61	-	2.2
5	298.17	0.74	0.61	2.8
10	84.0	**0.87**	**0.76**	3.6
20	84.6	0.87	0.71	5.1
30	83.7	0.85	0.63	7.7
50	**83.5**	0.86	0.65	9.3

Table 4. Ablation study on sampling steps.

ture degenerates into a traditional GAN model. From the results, we observe that the improvement in FGD becomes more gradual after 10 steps, but an increase in the number of steps also leads to slower speed. Therefore, we compromise and choose the result with 20 steps as the final demo model.

Influence of Geometric Constraints. In this experiment, we validate the role of the explicit gesture geometric constraints mentioned in Section 3.3 on the BEAT dataset. When the geometric constraint weight is set to 0, our method degenerates into a purely implicit joint distribution matching constraint for x_t and x_{t-1}, which is a statistically inefficient and original DDGAN incapable of handling high-dimensional feature representations, resulting in a significant decline in gesture quality, as shown in Table 5. With the addition of explicit geometric constraints, the gesture quality is noticeably improved. Based on our experience, we set the weights to 1, 10, and 100 for the experiments and found that the results did not differ significantly. In this paper, we use the model with a weight of 10 as the final demo.

5. Conclusions and Discussions

In conclusion, this paper has presented an innovative approach to address the limitations of diffusion models in the generation of co-speech gestures. The focus has been on enhancing the generation speed and modeling complex distributions between multiple sampling steps. Unlike the acceleration strategies of previous diffusion-based text-to-motion methodologies, such as those utilized in DDIM, our method avoids the stacking of small-step noise. Instead, we have introduced a conditional GAN to model the complex denoising distribution under audio control signals, enabling our model to denoise with larger and fewer steps, thus improving efficiency.

Furthermore, our approach addresses the statistical inefficiency of DDGAN in image generation and its inability to accommodate high-dimensional feature representations. By moving away from the original purely implicit matching constraint on joint distribution and incorporating explicit gesture geometric constraints, our method achieves significant acceleration while preserving high-fidelity generation results. This approach provides a promising direction for future real-time gesture generation tasks. We pledge to make our code repository publicly available, allowing future research to focus on improving the proposed method [9], exploring its applicability in other domains (such as providing more diverse generated data for motion capture [4,14] and whole body generations [30]), and investigating the integration of additional constraints to enhance the quality and realism of generated gestures. In addition, some target detection methods [19] are used to extract details of hand and facial information to optimize the generated results.

8

References

[1] Chaitanya Ahuja, Dong Won Lee, Yukiko I Nakano, and Louis-Philippe Morency. 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*, pages 248–265. Springer, 2020. 3

[2] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. 2023. 3

[3] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 3

[4] Yongkang Cheng, Shaoli Huang, Jifeng Ning, and Ying Shan. Bopr: Body-aware part regressor for human shape and pose estimation. *arXiv preprint arXiv:2303.11675*, 2023. 8

[5] Yongkang Cheng, Mingjiang Liang, Shaoli Huang, Jifeng Ning, and Wei Liu. Expgest: Expressive speaker generation using diffusion model and hybrid audio-text guidance. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 3

[6] Saeed Ghorbani, Ylva Ferstl, Daniel Holden, Nikolaus F Troje, and Marc-André Carbonneau. Zeroeggs: Zero-shot example-based gesture generation from speech. In *Computer Graphics Forum*, volume 42, pages 206–216. Wiley Online Library, 2023. 6

[7] Ikhsanul Habibie, Mohamed Elgharib, Kripasindhu Sarkar, Ahsan Abdullah, Simbarashe Nyatsanga, Michael Neff, and Christian Theobalt. A motion matching-based framework for controllable gesture synthesis from speech. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 3

[8] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 101–108, 2021. 3

[9] Gaoge Han, Mingjiang Liang, Jinglei Tang, Yongkang Cheng, Wei Liu, and Shaoli Huang. Reindiffuse: Crafting physically plausible motions with reinforced diffusion model. *arXiv preprint arXiv:2410.07296*, 2024. 8

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. volume 33, pages 6840–6851, 2020. 3, 4

[11] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992. 5

[12] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 6

[13] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. 2022. 3

[14] Mingjiang Liang, Yongkang Cheng, Hualin Liang, Shaoli Huang, and Wei Liu. Ropetp: Global human motion recovery via integrating robust pose estimation with diffusion trajectory prior. *arXiv preprint arXiv:2410.20358*, 2024. 8

[15] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European Conference on Computer Vision*, pages 612–630. Springer, 2022. 6

[16] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-driven co-speech gesture video generation. volume 35, pages 21386–21399, 2022. 3

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[18] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 3

[19] Baozhen Sun, Zhenhua Wang, Shilei Wang, Yongkang Cheng, and Jifeng Ning. Bidirectional interaction of cnn and transformer feature for visual tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 8

[20] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. 2022. 5

[21] Hao Xie, Jiangqun Ni, and Yun-Qing Shi. Dual-domain generative adversarial network for digital image operation anti-forensics. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1701–1706, 2021. 3

[22] Pan Xie, Qipeng Zhang, Zexian Li, Hao Tang, Yao Du, and Xiaohui Hu. Vector quantized diffusion model with codeunet for text-to-sign pose sequences generation. *arXiv preprint arXiv:2208.09141*, 2022. 3

[23] Yanwu Xu, Mingming Gong, Shaoan Xie, Wei Wei, Matthias Grundmann, Tingbo Hou, et al. Semi-implicit denoising diffusion models (siddms). *arXiv preprint arXiv:2306.12511*, 2023. 3

[24] Sicheng Yang, Zilin Wang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Qiaochu Huang, Lei Hao, Songcen Xu, Xiaofei Wu, Zonghong Dai, et al. Unifiedgesture: A unified gesture synthesis model for multiple skeletons. 2023. 3

[25] Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Ming Cheng, and Long Xiao. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. 2023. 3, 4, 6

[26] Sicheng Yang, Zunnan Xu, Haiwei Xue, Yongkang Cheng, Shaoli Huang, Mingming Gong, and Zhiyong Wu. Freetalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness. *arXiv preprint arXiv:2401.03476*, 2024. 3, 6

[27] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black.

Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023. 3

[28] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. volume 39, pages 1–16. ACM New York, NY, USA, 2020. 6

[29] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 736–747, 2022. 7

[30] Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark. *arXiv preprint arXiv:2310.20436*, 2023. 8

[31] Chi Zhou, Tengyue Bian, and Kang Chen. Gesturemaster: Graph-based speech-driven gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 764–770, 2022. 3

[32] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10544–10553, 2023. 3, 6