3 April 2025, Preprint v1 · CC-BY 4.0

Qeios

PREPRINT

Research Article

HoloGest: Decoupled Diffusion and Motion Priors for Generating Holisticly Expressive Co-speech Gestures

Yongkang Cheng[1], **Shaoli Huang**[1]

1. Tencent AILab

Animating virtual characters with holistic co-speech gestures is a challenging but critical task. Previous systems have primarily focused on the weak correlation between audio and gestures, leading to physically unnatural outcomes that degrade the user experience. To address this problem, we introduce HoleGest, a novel neural network framework based on decoupled diffusion and motion priors for the automatic generation of high-quality, expressive co-speech gestures. Our system leverages large-scale human motion datasets to learn a robust prior with low audio dependency and high motion reliance, enabling stable global motion and detailed finger movements. To improve the generation efficiency of diffusion-based models, we integrate implicit joint constraints with explicit geometric and conditional constraints, capturing complex motion distributions between large strides. This integration significantly enhances generation speed while maintaining high-quality motion. Furthermore, we design a shared embedding space for gesture-transcription text alignment, enabling the generation of semantically correct gesture actions. Extensive experiments and user feedback demonstrate the effectiveness and potential applications of our model, with our method achieving a level of realism close to the ground truth, providing an immersive user experience. Our code, model, and demo are are available at https://cyk990422.github.io/HoloGest.github.io/.

Corresponding author: Shaoli Huang, shaol.huang@gmail.com

doi.org/10.32388/L4S0VW

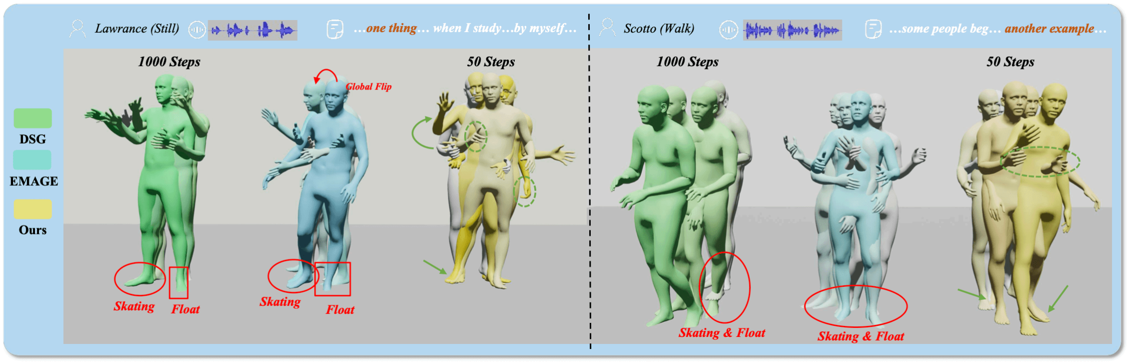

Figure 1. A comparison of three methods: DSG, a diffusion-based co-speech gesture generation method using DDPM (stiff limbs, slow inference, physically unnatural); EMAGE, an autoregressive generation method using VAE (motion artifacts, global flipping, physically unnatural); and our proposed generation method (rich movements, lively fingers, physically natural). The transition from past frames to the current frame (every 10 frames) is represented by the gradient in virtual human color, from light to dark.

1. Introduction

Animating virtual characters with holistic co-speech gestures is a challenging but critical task in various fields such as entertainment, education, and telecommunication. These gestures play a vital role in enhancing the naturalness and appeal of virtual characters, as they convey non-verbal information and improve the overall communication experience. However, generating holistic co-speech gestures that accurately represent the complex interplay between audio and body motion remains a challenging task, primarily due to the weak correlation between audio and both global motion trajectory and finger movements. This weak correlation often leads to physically unnatural outcomes, such as jittering in the global motion trajectory and poor expressiveness of finger motions, significantly reducing the effect of virtual characters.

Previous co-speech gesture generation methods can be divided into two categories: VAE-based or VQ-VAE-based generation systems[1][2] and diffusion-based generation systems[3][4]. The former maps weakly correlated gesture-audio pairs to a low-dimensional latent space and learns a continuous probability distribution, from which new latent vectors are sampled and decoded to obtain co-speech gestures. However, due to the VAE's reconstruction loss[5] based on joint-level errors and the ambiguity in the latent space, the generated gestures often appear overly smooth or unnatural. In contrast, diffusion-based methods model gesture generation as a gradual diffusion process, where the mapping relationship between audio and gestures is gradually established through a series of noise diffusion steps. Compared to VAE, this method can generate gesture sequences with rich details while maintaining audio synchrony. However, the high computational density and the resulting time cost limit the further development of diffusion methods. Furthermore, both approaches lack consideration of prior knowledge in motion, focusing only on the weakly correlated mapping between audio and upper-body gestures, and neglecting the physical laws of overall movement, such as continuity, stability, and rationality. Therefore, the generated holistic gestures may exhibit unnatural sliding, hovering phenomena, and monotonous finger movement, leading to a lack of overall naturalness and expressiveness, as shown in Figure 1.

To address these challenges, we introduce HoloGest, a novel diffusion-based framework for automatically synthesizing high-fidelity holistic co-speech gesture sequences from audio. Our system posits that in holistic co-speech gestures, limb movements are correlated with audio, global trajectories are related to limbs and independent of audio, and fingers are associated with both arm movements and audio information. Based on this, we learn a global trajectory diffusion generative prior model guided by limb movements on a large-scale human motion dataset. Simultaneously, we learn a finger diffusion generative prior model guided by the arms on a mixed sign language and gesture dataset, leaving audio and semantic features blank for subsequent fine-tuning. The former provides our system with strong locomotion prior, overcoming long-standing issues of unnatural sliding and jittering, while the latter offers more diverse finger movements, assisting in generating more vivid and high-fidelity gesture results.

Unlike previous methods that model the denoising process of the whole body as a single distribution, our system decouples the upper limbs, lower body, and fingers into three smaller and simpler subproblems, breaking down holistic co-speech gestures. During the denoising process, each sub-model focuses more on the distribution of specific body parts, thereby improving the generation quality of each part. However, the parallel diffusion denoising processes for the three parts further reduce the generation efficiency. To break free from this constraint, we employ a semi-implicit constraint[6], modeling large-stride complex multimodal distributions between adjacent denoising steps to significantly reduce the required number of denoising steps, thus achieving acceleration.

Predicting gestures from speech is a challenging multimodal mapping task. A single speech segment can correspond to multiple gestures, making the association between the semantics intended to be conveyed in speech and gestures non-intuitive. Our system adopts the JEPA strategy[7][8] to learn a gesture-speech joint embedding space. We first introduce wav2vec2[9] for text transcription, then extract textual features and map them with gestures to a shared low-dimensional space based on a variational autoencoder. Finally, we introduce a predictor layer to further extract semantic features, aligning these abstract semantic features with the low-dimensional latent variables of gestures in this space. This approach maintains semantic alignment while generating natural gestures closely related to speech.

To demonstrate the inspirational value of motion priors in our system for the human motion generation domain, we further fine-tune our framework on the music-to-dance task, addressing the physical naturalness of generated results and showcasing its powerful generalization capabilities. To the best of our knowledge, our system represents the first audio-whole body gesture generation model considering motion priors, capable of generating high-fidelity, diverse, and physically natural holistic co-speech gesture sequences based on arbitrary user-provided audio (speech or music). We showcase our approach on multiple publicly available audio-motion datasets, and extensive experimental results indicate that, compared to VAE systems, our method generates more diverse and higher-quality results, while maintaining the naturalness of overall motion compared to diffusion systems, significantly reducing time costs and providing users with a novel experience. The importance of algorithmic design is also validated through ablation experiments.

2. Related Work

Audio-to-motion Generation. Initial data-driven methods ([2][10]) aimed to learn gesture matching from human demonstrations but lacked diversity. With the increasing interest in these action reconstruction methods[11][12][13], the training datasets have gradually become more abundant. Subsequent works ([10][14][15][16][17]) improved model diversity and introduced unique, expressive gestures. Some studies ([18][19][20]) trained unified models for multiple speakers, embedding styles or applying style transfer techniques. Other research ([21][22]) utilized motion matching for gesture sequences, despite requiring complex rules. Audio-driven animation has gained attention, with virtual speaker animation advancements attributed to high-quality gesture datasets ZeroEGGs[23] and BEATX[24]. Talkshow[14] and EMAGE[24] improved user experience by incorporating facial and expression parameters in virtual speaker generation. However, these methods face unnatural holistic gesture issues, and models lacking human motion knowledge struggle with physical problems like jittering, foot sliding, and floating, hindering the field's development. Some works are limited to the human upper body. HoloGest innovatively introduces a motion prior model to address these physical unnatural problems in gesture generation, providing a more engaging user experience.

Diffusion Generative Models have achieved remarkable results across various domains[25][26][27][28][29], especially in human motion generation. Motion Diffuse[30] first applied diffusion models to text-conditioned human motion generation, offering fine-grained control of body parts. MDM[31] is a milestone work using a motion diffusion model to manipulate motion representation based on input text control conditions. DSG[32] generates well-matched results with speech using an attention mechanism. However, due to the high dimensionality and interactivity of diffusion models, motion generation based on the original diffusion model DDPM[33] suffers from time overhead. MLD[26] introduces latent diffusion to motion generation, reducing computational resources and employing DDIM[34] to enhance inference speed. Nevertheless, this two-stage method is non-end-to-end, and DDIM's noise step stacking and denoising step discarding result in artifacts. HoloGest addresses these issues by being the first method in gesture generation to use GAN[35] for accelerating diffusion model inference speed. By increasing denoising step size and reducing denoising steps, it maintains high-quality diffusion model advantages while enabling rapid generation.

3. Method

3.1. System Overview

Our system synthesizes vivid, physically natural, and holistic co-speech gestures using only audio input. It is built on the human motion diffusion model (MDM) framework, employing the Diffusion Model to model adjacent denoising step distributions and supervising human geometric constraints for motion quality. The system structure, shown in Figure 2, comprises two core components: (a) an end-to-end decoupled diffusion generative model that accepts audio input and denoises human joint sequences in parallel, and (b) a motion prior optimizer pre-trained on a large-scale human dataset, re-optimizing global motion and finger actions based on generated joint cues for natural and vivid virtual speakers. We also transcribe text and utilize the JEPA[36] strategy to extract semantic cues, enhancing result richness. To address DDPM denoising inefficiency, a semi-implicit denoising process is introduced for faster generation. In subsequent sections, we detail the system's key components.

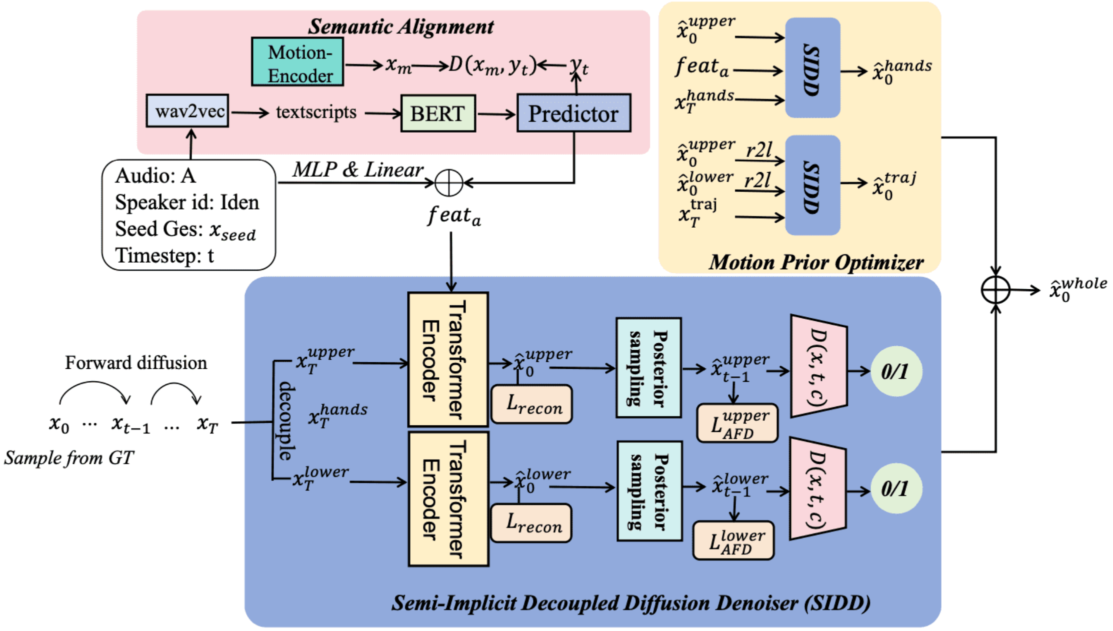

Figure 2. Our system comprises a semantic alignment module and two core components: (a) The semantic alignment module maps both the transcribed text and gesture sequence into the latent space simultaneously, further abstracting the semantic latent variables and aligning them with the gesture latent variables in a higher-level abstract space, serving as independent guiding tokens. (b) The semi-implicit decoupled denoiser, by introducing GAN and semi-implicit constraints, models the complex denoising distribution between adjacent large strides, accelerating generation by reducing the number of steps. (c) The motion prior optimization takes the denoised initial local gesture sequence as a condition, and in conjunction with the audio guiding signal, generates global motion and finger actions for the second time. This system requires no additional input and has no time constraints; any pure audio file can generate a set of vivid, natural, and high-quality holistic co-speech gesture sequences. 'r2l' represents converting the rotation representation to the coordinate representation using the SMPL model.

3.2. Decoupled Diffusion Denoiser

Brief overview of MDM. Unlike traditional diffusion model-based methods, MDM[31] considers the inherent physical constraints of the three-dimensional human body by predicting the original human motion representation instead of predicting noise, deviating from the DDPM process in conventional image generation. Therefore, at each step of the denoising process, MDM reconstructs the original representation from pure Gaussian noise, and ultimately generate the final result through the iterative process of noise addition and denoising:

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

where c is the control signal.

Decoupled Denoiser Structure. We construct our denoiser framework on MDM, tailoring it for the audio-to-gesture task with conditions including noise step , seed pose, audio information, and semantic latent code. The noise step and seed pose are projected to the same dimension via MLP and linear layers, respectively, and subsequently added together. Audio is encoded using WavLM and time-dimension interpolated to align with gesture frames.

Although the denoising probability model generates satisfactory gestures, finger motion differs from limb motion. Limb movements exhibit larger amplitudes and correlate with melody, while finger movements are smaller, more precise, and semantically matched. Holistic modeling prioritizes body data matching over finger movements, reducing overall gesture expressiveness.

To address this, we decouple the human body into upper limbs, lower limbs, and fingers, denoising these parts in parallel. However, the absence of global associations can result in unnatural motion, such as sudden orientation flips. *To alleviate this, we concatenate the three-part features and map them to an independent conditional token, providing a global constraint for generating coherent results.*

3.3. Semi-implicit Matching Constraint

We've improved the network structure for better results, but DDPM's high computational complexity still limits diffusion generative methods' potential. This issue arises from DDPM's assumption that small, unimodal noise is added at each step, requiring many steps for denoising. Increasing noise step size disrupts the Gaussian distribution, making a simple $L2$ loss inadequate for modeling complex motion distribution and causing unnatural jittering.

To address this, we incorporate a GAN structure inspired by SiDDMs[6] as an implicit objective to learn the denoiser. The GAN's conditional discriminator differentiates between the predicted denoising and original motion distributions, while the conditional denoiser aims to make them indistinguishable. The process is described by equation (2).

$$\min_{\theta}\max_{D_{adv}}\sum_{t>0}\mathbb{E}_{q(x_t)}D_{adv}(q(x_{t-1}|x_t)||p_\theta(x_{t-1}|x_t)), \tag{2}$$

By examining the implementation of equation (2), it is clear that, during the adversarial stage, the method indirectly matches the conditional distribution by aligning with the joint distribution:

$$\min_{\theta}\max_{D_{adv}}\mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})}\,D_{adv}(q(x_{t-1},x_t)||p_\theta(x_{t-1},x_t)), \tag{3}$$

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 (3). That is, we directly compute the posterior distribution, and then use the forward process for adversarial learning to model the large-step denoising distribution. The equation is represented as follows:

$$\begin{aligned}\min_{\theta}\max_{D_{\varnothing}}\mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})}&\,[-\log(D_{\varnothing}(x_{t-1},c,t))]\\&+[-\log(1-D_{\varnothing}(\hat{x}_{t-1},c,t))],\end{aligned} \tag{4}$$

Although we have simplified the implicit matching process, making adversarial training more stable, we have also encountered a new problem. Since the large-step denoising distribution is typically a complex multimodal distribution, the posterior sampling $p_\theta(\hat{x}_{t-1}|x_t,\hat{x}_0)$ result still has a significant difference from the forward sampling process, preventing our denoiser from successfully reversing from the pure noise distribution to the original distribution. Based on this, we employ the regularization term, Auxiliary Forward Diffusion Constraint (AFD), to explicitly constrain the similarity between the backward sampling results and forward diffusion results at the same time step. Its expression is as follows:

$$\mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})}\,\frac{(1-\beta_t)||\hat{x}_{t-1}-x_{t-1}||^2}{\beta_t}, \tag{5}$$

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 the variance table within the interval (0,1]. All models are trained using the AdamW optimizer with a fixed learning rate l. We apply EMA decay to the optimizer during the training process. The final training objective is:

$$\begin{aligned}\min_{\theta}\max_{D_{adv}}\mathbb{E}_{q(x_0)q(x_{t-1}|x_0)q(x_t|x_{t-1})}&\,[-\log(D_{\varnothing}(x_{t-1},c,t))]\\&+[-\log(1-D_{\varnothing}(\hat{x}_{t-1},c,t))]+\lambda_{recon}\mathcal{L}_{Recon}\\&+\lambda_{AFD}\,\frac{(1-\beta_t)||\hat{x}_{t-1}-x_{t-1}||^2}{\beta_t},\end{aligned} \tag{6}$$

where λ_{recon} represents the reconstruction weight of the denoiser, and λ_{AFD} represents the weight of the regularization term. $\mathcal{L}_{Recon}$ represents the Mean Squared Error Loss between the denoised $\hat{x}_0$ and the original data x_0.

3.4. Motion Prior Optimizer

Despite the semi-implicit decoupled denoiser's ability to recover detailed and expressive gesture sequences, the correlation between audio and motion remains weak. This results in issues like foot sliding, jittering, and unnatural movements in previous methods. We believe the problem stems from the fact that not all aspects of co-speech gestures depend on audio. For instance, trajectories may not be related to audio beats but are closely connected to limb movements. To address these issues, we designed a motion prior optimizer.

Trajectory Prior Optimizer. In our system, the trajectory is more closely linked to limb posture than to audio. Therefore, we use limb posture as the guiding condition to recreate the global motion trajectory. Thanks to large-scale public human motion datasets, our model can acquire extensive motion prior knowledge. Taking a cue from GLAMR[37], we define the trajectory as a 9-dimensional parameter $G = (\Delta x, \Delta y, \Delta z, \mathrm{rot6d} \in \mathbb{R}^6)$, where the last six dimensions represent global rotation and the first three dimensions represent displacement increments along the XYZ axes, ensuring smoother results.

The trajectory prior model continues to use the semi-implicit diffusion approach, with the 3D coordinates of the 21 human joint parts (excluding the root joint) and the time step t as conditional guidance. These are independently mapped to conditional tokens and input into the Transformer-Encoder-based denoiser. The denoising process follows equation (6).

Finger Prior Optimizer. We've observed that finger movements guided solely by audio or semantic signals often lack dynamism and expressiveness. We believe finger movements correlate with forearm movements, such as a person pointing in a certain direction when raising their arm forward. Hence, we train finger priors on large-scale sign language and gesture datasets. During finger prior training, the guiding condition only uses the 6D rotation representation of the human forearm, leaving semantic and audio features empty for subsequent fine-tuning. The finger results are denoised using equation (3).

When fine-tuning on the BEATX dataset, we incorporate the audio signal into the finger prior's conditional guidance and link the entire system together for fine-tuning. During inference, the motion prior model serves as an optimizer, using the body generated by the semi-implicit decoupled denoiser as a condition, and generates the global trajectory and finger rotations as the final output.

3.5. Semantic Alignment

The existence of many-to-many mapping relationships between audio content and gesture sequences poses a significant challenge for generating semantically aligned actions accurately. To address this issue, we learn a joint embedding space for gestures and audio transcriptions, allowing them to align in an abstract space and reveal the semantic associations between the two modalities. Inspired by I-JEPA[36], we initially train gesture and transcription encoders separately using a motion VAE structure and a BERT tokenizer[38]. We parameterize the transcriptions as tokenized word embedding sequences and linearly map them to a space with the same dimensions as the gesture latent codes x_m. Finally, we introduce a Predictor to further abstract semantic features y_t from the latent space and fine-tune the encoders using CLIP-style contrastive learning $D(x_m, y_t)$. Both the motion VAE and Predictor structures adopt the traditional Transformer architecture. The NT-Xent[39] loss is used in contrastive learning, with the goal of maximizing the similarity of transcription-gesture matched pairs in the latent space while minimizing the similarity of non-matched pairs. Formally, the loss function is as follows:

$$\mathcal{L}(t, m) = -\log \frac{\exp(\mathrm{sim}(x_t, y_m)/\tau)}{\sum_{k \in K} \exp(\mathrm{sim}(x_t, y_m)/\tau)}, \tag{7}$$

where, x_t and y_m are the latent space representations of a matching transcription-gesture pair. sim is similarity score between two latent codes, K is a set containing one positive sample transcription and a group of negative sample gestures, and τ is the temperature parameter used to adjust the sensitivity of the function. Finally, we freeze the trained semantic alignment module and deploy only the transcription encoder into the system, ensuring that the final generated results accurately capture the semantic content.

4. Experiments

In this section, we evaluate the effectiveness of the proposed system in generating holistic co-speech gestures from audio and compare it with contemporary holistic gesture generation methods to demonstrate the superiority of our system. Ablation studies further validate the roles of essential modules and design choices within the system. Generalization experiments showcase the potential value and application prospects of our proposed method in this domain. Considering the subtle nature of human gestures for evaluation, we conduct extensive user studies to substantiate the superior performance of the proposed system. ***We strongly encourage readers to refer to the accompanying video for additional qualitative evaluations and application results.***

4.1. Experiment Design

Datasets. For the audio-independent global trajectory motion prior module, we train on the 100-STYLE[40] and AMASS[41] datasets. Both are large-scale publicly available Mocap datasets, with the former containing over 4 million frames of 100 different locomotion styles, and the latter being a large-scale human motion dataset, both represented with 55 joints in SMPLX[42] format. The trajectory prior is trained using all datasets. For the finger prior module associated with arm movements, we train on the SignAvatars[43] sign language dataset and the audio-removed BEATX[24] dataset. The former contains SMPLX representations of multiple sign language videos shared with us by the authors, and the latter is a publicly available large-scale gesture dataset, uniformly represented in SMPLX format, containing 24 English speakers. When training the finger prior, we mix all speaker data and sign language data for training. Finally, we train the audio-to-gesture model on the BEATX dataset, and during the fine-tuning of fingers, we release the audio features and semantic alignment as additional guiding signals to generate natural and rich finger movements. We evaluate the model's effectiveness on the BEATX test set.

Evaluation Metrics. To evaluate the effectiveness of our proposed system, in addition to focusing on the common Frechet Gesture Distance (FGD)[44], Beat Alignment (BA)[24], and Diversity (DIV)[1] metrics, we also introduce physical naturalness evaluation metrics, including Skating (Skate)[45] and Floating (Float) [46], and define a Semantic Alignment score (SA) to validate the performance of the semantic alignment module. The first three are used to evaluate the quality of generated gestures: (1) **FGD** is a common metric in generative models, used to evaluate the difference between the distribution of generated movements and the original training distribution, providing insights into the fidelity and similarity between generated data and real data. (2) **BA** is used to evaluate the synchronicity of speech and movement, with higher values indicating better alignment with the audio beat. (3) **DIV** measures the L1 distance between multiple body gestures generated under the same control signal, with larger values indicating greater diversity.

To evaluate the physicality of holistic co-speech gestures, we use (4) **Skate** to quantify the displacement distance of the virtual character's toes when their feet are in contact with the ground (determined by setting a toe acceleration threshold). This is crucial for the naturalness and authenticity of overall motion, as realistic motion results can provide users with an immersive experience. (5) **Float** is used to assess the floating distance of the virtual character along the y-axis. We assume the ground level to be the lowest point of the sequence plus 0.5 cm, and when the character has at least one foot in contact with the ground, we calculate the distance between the toes and the ground to quantify the degree of floating.

To evaluate the semantic consistency between speech and generated gestures, we define a new metric called Semantic Alignment (SA)[20]. It assesses the degree of semantic alignment by calculating the similarity between the latent gesture representation in the low-dimensional space and the real text representation in the abstract space. The calculation formula is as follows:

$$SA = \cos(\text{avg\_pool}(V_g(G_{\text{pred}})), \text{avg\_pool}(V_s(S))), \tag{8}$$

where *G* represents the gestures generated by the model, and *S* denotes the hidden states encoded by the BERT[38] model after tokenizing the transcribed text, serving as a representation of the semantics.

Implementation Details. Our system was trained on PyTorch with a denoiser learning rate of 3e-5 and a discriminator learning rate of 1.25e-4. The discriminator's gradient penalty term was set to 0.02, in line with DDGAN[47], and the CFG weight was set to 3.5. All models were trained on an A100 GPU for a uniform 1.3 million-step iteration, taking a total of 5 days. During evaluation, all methods were tested on a single V100 GPU for fairness.

4.2. Comparison with Contemporary Methods

We present the quantitative results for speaker 2's test sequences in the BEATX dataset using the audio-to-gesture method in Table 1. The purpose is to provide a fair comparison with the values reported in the original EMAGE paper. ***For a more comprehensive view of the quantitative experiments, we provide the quantitative results for the entire dataset in parentheses.*** Our findings demonstrate that, in comparison with diffusion-based methods such as DSG[32], FreeTalker[3], and DiffGesture[4], our approach outperforms them in terms of gesture matching, even with a 20-fold reduction in denoising steps. Moreover, our method surpasses VAE and VQ-VAE-based approaches like EMAGE[24], TalkShow[14], and CAMN[1] in terms of beat alignment and diversity.

doi.org/10.32388/L4S0VW

Method	BEATX				
	FGD↓	SA↑	BA↑	DIV↑	steps
HA2G[2]	12.32	0.13	6.77	8.626	-
DisCo[48]	9.417	0.09	6.439	9.912	-
CaMN[1]	6.644	0.22	6.769	10.86	-
TalkShow[14]	6.209	0.22	6.947	13.47	-
EMAGE[24]	5.512(7.305)	0.17	7.724(7.709)	10.88(10.948)	-
DiffGesture(re-train)[4]	12.8	0.07	7.08	11.30	1000
DSG(re-train)[32]	8.811(11.742)	0.08	7.241(7.3368)	11.49(11.121)	1000
FreeTalker(re-train)[3]	7.712	0.19	7.73	10.62	1000
HoloGest(Ours)	**5.3407**(6.457)	**0.66**	**7.957**(8.0281)	**14.15**(13.525)	50

Table 1. Objective metrics on BEATX. EMAGE provides the FGD evaluation model, where a lower value indicates a closer approximation to the original motion distribution. The calculation methods for BA and DIV are consistent with EMAGE. Steps represent the denoising steps in diffusion-based generation methods. ***The values in parentheses represent the evaluation results for the entire BEATX dataset.***

However, gesture evaluation is subtle, and the FGD metric only reflects the similarity between generated results and the original distribution, not the actual effect of the virtual speaker or the trajectory and global rotation of holistic co-speech gestures. While EMAGE has metrics close to our system, it lacks prior knowledge of the entire motion sequence, leading to discord in its generated results, including unnatural global flips and severe skating phenomena. Its fingers also lack rich movements due to the absence of finger priors.

In contrast, our system, which introduces motion priors, generates reliable global movements without affecting vivid gestures, provides stable locomotion without unnatural flips or severe arm jittering, and offers users a more natural and harmonious experience. Table 2 presents the physical metrics and semantic alignment scores, consistent with our observed phenomena.

Method	**BEATX**			
	FGD↓	**Skate↓**	**Float↓**	**SA↑**
Real	1.7e-4	0.0866	8.8015	0.82
EMAGE	5.51	0.7904	34.6534	0.17
DSG	8.811	0.4192	22.7526	0.08
HoloGest(Ours)	5.34	0.1068	9.6317	0.66

Table 2. Objective Metrics. Skate represents the skating metric when in contact with the ground, with values closer to Real being better. Float indicates the floating error during ground contact, with values closer to GT being better.

4.3. Qualitative Comparison

We present the results generated by DSG, EMAGE, and HoloGest on the BEATX test set. As seen in Figure 3, DSG's gesture generation lacks expressiveness, showing little movement during flat speech, resulting in a stiff appearance with unnatural phenomena like sliding and floating. The 1000-step DDPM sampling strategy also leads to inefficient generation. EMAGE, using VAE for direct regression, is fast but prone to motion artifacts and global flipping, affecting user experience.

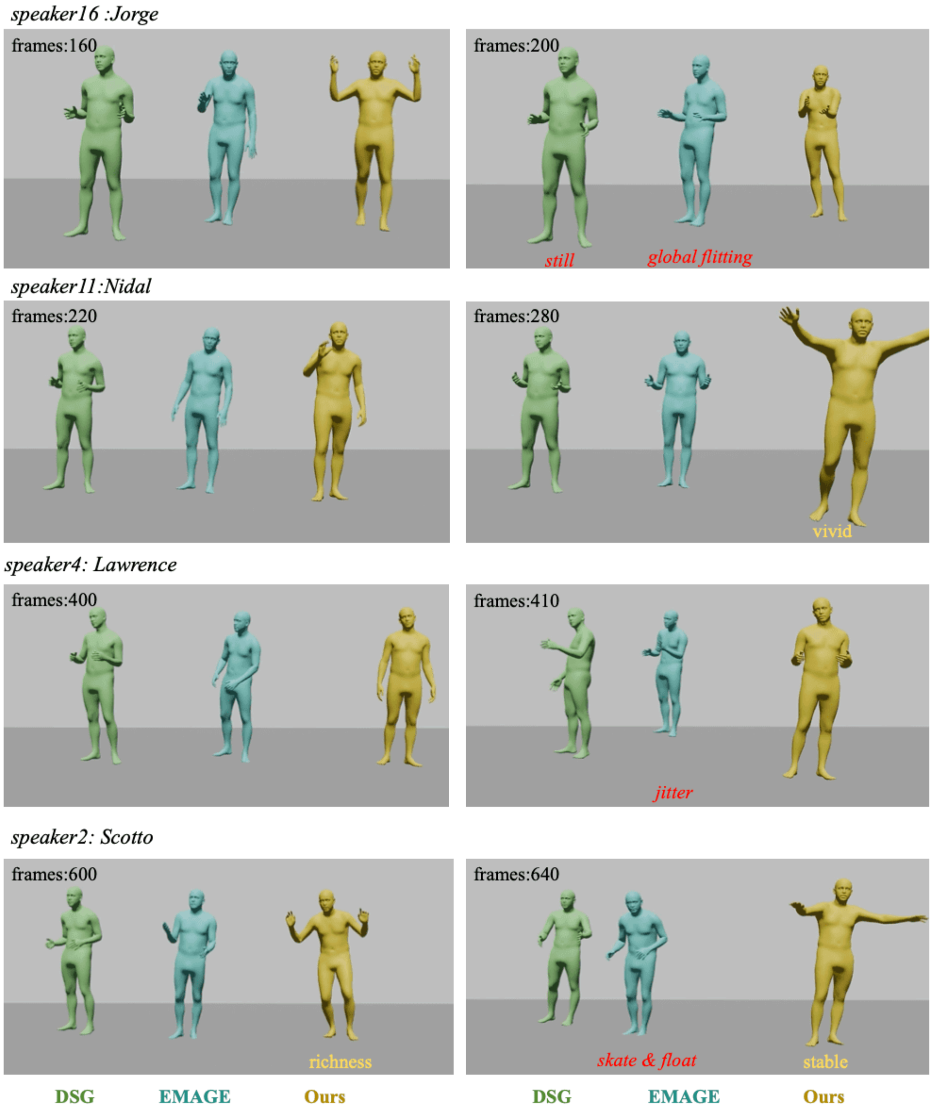

Figure 3. A comparison of three methods: DSG, a diffusion-based co-speech gesture generation method using DDPM (stiff limbs, slow inference, physically unnatural); EMAGE, an autoregressive generation method using VAE (motion artifacts, global flipping, physically unnatural); and our proposed generation method (rich movements, lively fingers, physically natural). We test on a sequence of an English-speaking presenter selected from BEATX. Red annotations indicate defects, while yellow annotations highlight advantages.

In contrast, our model achieves large strides with fewer denoising steps, enhancing generation speed while maintaining high fidelity, making it suitable for real-time applications. The introduction of motion priors improves global motion and physical naturalness. Thanks to the semantic alignment module in the abstract space, our method is highly expressive, with large, lively upper limb movements and natural, rich lower limb movements. Our divide-and-conquer approach enhances the richness of finger movements and the stability of global actions.

4.4. Ablation Study

To validate the importance of each module in the system, we compare various variants obtained from the complete method:

- ***Baseline***, directly decouples three body parts: the upper body, lower body, and fingers, as three parallel sub-models for denoising distribution modeling, while maintaining the DDPM denoising process, using only audio features as guiding signals, as done in DSG.
- ***+ SIDD***, to alleviate the inefficiency in generation caused by body decoupling, we introduce a semi-implicit denoising process that directly models the complex large-step denoising distribution, achieving acceleration by reducing the number of denoising steps.
- ***+ SA***, by employing the JEPA strategy, we use semantic features that are aligned with the real gesture sequences in the abstract space as additional guiding conditions, and independently conditionally tokenize them, similar to what is done in locomotion.
- ***+ Global***, in order to establish connections between the decoupled parts, we associate the features of the three parts and further map them to a single global token, serving as an additional global perceptual information.
- ***+ Prior***, incorporate global trajectory motion priors and finger priors as pre-trained models for secondary generation.

Method	BEATX						steps
	FGD↓	Skate↓	Float↓	SA↑	BA↑	DIV↑	
Real	1.7e - 4	0.0866	8.8015	0.91	-	-	-
DSG	8.811	0.4192	22.7526	0.08	7.241	11.49	1000
EMAGE	5.512	0.7904	34.6534	0.17	7.724	10.88	-
Baseline	7.718	0.3922	19.7831	0.20	7.432	12.83	1000
+ SIDD	7.016	0.5567	25.1263	0.22	7.135	14.12	50
+ SA	6.351	0.5239	17.6612	0.60	7.946	14.26	50
+ Global	5.86	0.3396	19.023	0.66	7.953	**14.29**	50
+ Prior	**5.3407**	**0.1068**	**9.6317**	0.66	**7.957**	14.15	50

Table 3. Ablation study results on the module design in the system.

Table 3 shows the ablation study results, with the complete system outperforming all ablation versions. The diffusion generative model with only decoupled structure shows some improvement compared to DSG but has a noticeable disadvantage in metrics compared to VAE-based methods. This is due to the lack of connections between parts, causing uncoordinated overall gestures when directly merged. Introducing global associations and semi-implicit denoising process alleviates this issue and improves generation efficiency. The introduction of semantic alignment features significantly enhances the richness of generated actions. Despite the improvements in metrics and gesture quality, the lack of global motion prior knowledge still leads to physically unnatural factors like skating and global jitter, impacting user experience. By introducing global trajectory priors and finger priors as pre-trained models for secondary generation, we achieve physically plausible results and provide users with a better experience.

4.5. User Study

We used four human perceptual consistency scoring metrics as described in[49]. These metrics evaluate human likeness (HL), speech-gesture appropriateness (SGA), gesture richness, and whole-body stability.

doi.org/10.32388/L4S0VW

To assess our method's visual performance, we conducted a user study on gesture sequences generated by each method. Evaluation segments varied from 16 to 40 seconds in length, averaging 26.2 seconds. We engaged 30 participants and used a scoring range of 1 to 5, with labels from ”poor” to ”excellent”. Table 5 shows the average user opinion scores. We compared the results generated by the original DSG, EMAGE, our system without Prior, and the complete system. As per user feedback 4, our method generates high-quality co-speech gesture sequences comparable to, or better than, real data, and does so faster than traditional DDPM diffusion generative methods. Notably, our method takes only 0.88 seconds to generate a 2-second gesture sequence, compared to approximately 7 seconds using 1000-step DDPM, making it suitable for real-time applications like human-computer communication.

Method	**BEATX**			
	HL↑	**SGA↑**	**R↑**	**Stable↑**
Real	4.61 ± 0.17	4.72 ± 0.20	4.66 ± 0.07	4.89 ± 0.02
DSG	3.70 ± 0.12	3.91 ± 0.14	4.27 ± 0.15	3.12 ± 0.12
EMAGE	3.44 ± 0.18	4.11 ± 0.14	3.56 ± 0.09	2.87 ± 0.22
HoloGest(Ours)	4.47 ± 0.09	4.51 ± 0.19	4.82 ± 0.1	4.71 ± 0.11

Table 4. 95% Confidence Interval for User Study Average Score.

5. Conclusions

In this study, we tackled challenges in generating holistic co-speech gestures. By innovating upon diffusion-based methods with implicit marginal constraints and explicit auxiliary forward diffusion regularization, our model enabled faster inference and mitigated generation speed inefficiencies. Additionally, we considered motion prior and introduced a pre-trained model on extensive human motion data, generating physically accurate gesture sequences and enhancing user experience. Our approach significantly accelerated HoloGest's generation while maintaining high fidelity, paving the way for future real-time synchronous gesture generation tasks.

References

1. a, b, c, d 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." In: European Conference on Computer Vision. Springer; 2022. p. 612–630.
2. a, b, c Liu X, Wu Q, Zhou H, Du Y, Wu W, Lin D, Liu Z (2022). "Audio-Driven Co-Speech Gesture Video Generation". Advances in Neural Information Processing Systems. **35**: 21386–21399.
3. a, b, c Yang S, Xu Z, Xue H, Cheng Y, Huang S, Gong M, Wu Z (2024). "Freetalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness". arXiv preprint arXiv:2401.03476.
4. a, b, c Zhu L, Liu X, Liu X, Qian R, Liu Z, Yu L (2023). "Taming Diffusion Models for Audio-Driven Co-Speech Gesture Generation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 10544–10553.
5. ^ Kingma DP, Welling M (2013). "Auto-encoding variational bayes". arXiv preprint arXiv:1312.6114. arXiv:1312.6114.
6. a, b Xu Y, Gong M, Xie S, Wei W, Grundmann M, Hou T, et al. Semi-Implicit Denoising Diffusion Models (SIDDMs). arXiv preprint arXiv:2306.12511. 2023.
7. ^ Garrido Q, Assran M, Ballas N, Bardes A, Najman L, LeCun Y (2024). "Learning and Leveraging World Models in Visual Representation Learning". arXiv preprint arXiv:2403.00504.
8. ^ Bardes A, Garrido Q, Ponce J, Chen X, Rabbat M, LeCun Y, Assran M, Ballas N (2023). "V-JEPA: Latent Video Prediction for Visual Representation Learning".
9. ^ Baevski A, Zhou Y, Mohamed A, Auli M (2020). "wav2vec 2.0: A framework for self-supervised learning of speech representations". Advances in Neural Information Processing Systems. **33**: 12449–12460.
10. a, b Habibie I, Xu W, Mehta D, Liu L, Seidel HP, Pons-Moll G, Elgharib M, Theobalt C (2021). "Learning speech-driven 3d conversational gestures from video". In: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents. pp. 101–108.
11. ^ Cheng Y, Huang S, Ning J, Shan Y (2023). "Bopr: Body-aware part regressor for human shape and pose estimation". arXiv preprint arXiv:2303.11675. Available from: https://arxiv.org/abs/2303.11675.
12. ^ Liang M, Cheng Y, Liang H, Huang S, Liu W (2024). "RopeTP: Global Human Motion Recovery via Integrating Robust Pose Estimation with Diffusion Trajectory Prior". arXiv preprint arXiv:2410.20358.

13. ^Yu Z, Huang S, Cheng Y, Birdal T. "Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark." In: European Conference on Computer Vision. Springer; 2024. p. 1-19.

14. a, b, c, dYi H, Liang H, Liu Y, Cao Q, Wen Y, Bolkart T, Tao D, Black MJ. "Generating holistic 3d human motion from speech." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. p. 469-480.

15. ^Xie P, Zhang Q, Li Z, Tang H, Du Y, Hu X (2022). "Vector quantized diffusion model with codeunet for text-to-sign pose sequences generation". arXiv preprint arXiv:2208.09141. Available from: https://arxiv.org/abs/2208.09141.

16. ^Cheng Y, Liang M, Huang S, Han G, Ning J, Liu W (2024). "Conditional GAN for Enhancing Diffusion Models in Efficient and Authentic Global Gesture Generation from Audios". arXiv preprint arXiv:2410.20359. Available from: https://arxiv.org/abs/2410.20359.

17. ^Cheng Y, Liang M, Huang S, Ning J, Liu W. Expgest: Expressive speaker generation using diffusion model and hybrid audio-text guidance. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). IEEE; 2024. p. 1-6.

18. ^Yang S, Wang Z, Wu Z, Li M, Zhang Z, Huang Q, Hao L, Xu S, Wu X, Dai Z, et al. UnifiedGesture: A Unified Gesture Synthesis Model for Multiple Skeletons. arXiv preprint arXiv:2309.07051. 2023.

19. ^Ahuja C, Lee DW, Nakano YI, Morency L-P. "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. Springer; 2020. p. 248--265.

20. a, bAo T, Zhang Z, Liu L (2023). "GestureDiffuCLIP: Gesture Diffusion Model with CLIP Latents". ACM Trans. Graph.. **42** (4): 42:1–42:18. doi:10.1145/3592097. Source.

21. ^Zhou C, Bian T, Chen K (2022). "Gesturemaster: Graph-based speech-driven gesture generation". Proceedings of the 2022 International Conference on Multimodal Interaction. pp. 764–770.

22. ^Habibie I, Elgharib M, Sarkar K, Abdullah A, Nyatsanga S, Neff M, Theobalt C. A motion matching-based framework for controllable gesture synthesis from speech. In: ACM SIGGRAPH 2022 Conference Proceedings. 2022. p. 1-9.

23. ^Ghorbani S, Ferstl Y, Holden D, Troje NF, Carbonneau MA (2023). "ZeroEGGS: Zero-shot Example-based Gesture Generation from Speech". Computer Graphics Forum. Wiley Online Library. **42** (1): 206–216.

24. a, b, c, d, e, fLiu H, Zhu Z, Becherini G, Peng Y, Su M, Zhou Y, Iwamoto N, Zheng B, Black MJ (2023). "EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Masked Audio Gesture Modeling". arXiv preprint arXiv:2401.00374. arXiv:2401.00374.

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

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

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

28. ^*Han G, Liang M, Tang J, Cheng Y, Liu W, Huang S (2024). "Reindiffuse: Crafting physically plausible motions with reinforced diffusion model". arXiv preprint arXiv:2410.07296. Available from: https://arxiv.org/abs/2410.07296.*

29. ^*Han G, Huang S, Gong M, Tang J (2024). "Hutumotion: Human-tuned navigation of latent motion diffusion models with minimal feedback". Proceedings of the AAAI Conference on Artificial Intelligence.* ***38*** *(3): 2031–2039.*

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

31. a, b*Tevet G, Raab S, Gordon B, Shafir Y, Cohen-Or D, Bermano AH (2022). "Human motion diffusion model". arXiv preprint arXiv:2209.14916. arXiv:2209.14916.*

32. a, b, c*Yang S, Wu Z, Li M, Zhang Z, Hao L, Bao W, Cheng M, Xiao L (2023). "DiffuseStyleGesture: Stylized Audio-Driven Co-Speech Gesture Generation with Diffusion Models". arXiv preprint arXiv:2305.04919. Available from: https://arxiv.org/abs/2305.04919.*

33. ^*Ho J, Jain A, Abbeel P (2020). "Denoising diffusion probabilistic models". Advances in Neural Information Processing Systems.* ***33****: 6840–6851.*

34. ^*Song J, Meng C, Ermon S (2020). "Denoising diffusion implicit models". arXiv preprint arXiv:2010.02502. Available from: arXiv:2010.02502.*

35. ^*Xie H, Ni J, Shi YQ (2021). "Dual-domain generative adversarial network for digital image operation anti-forensics". IEEE Transactions on Circuits and Systems for Video Technology.* ***32*** *(3): 1701–1706.*

36. a, b*Assran M, Duval Q, Misra I, Bojanowski P, Vincent P, Rabbat M, LeCun Y, Ballas N. Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. p. 15619–15629.*

37. ^*Yuan Y, Iqbal U, Molchanov P, Kitani K, Kautz J (2022). "GLAMR: Global occlusion-aware human mesh recovery with dynamic cameras". Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pages 11038–11049.*

38. a, b*Devlin J, Chang MW, Lee K, Toutanova K (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding". arXiv preprint arXiv:1810.04805.*

39. ^*Chen T, Kornblith S, Norouzi M, Hinton G (2020). "A simple framework for contrastive learning of visual representations". In: International conference on machine learning. PMLR. pp. 1597–1607.*

40. ^*Mason I, Starke S, Zhang H, Bilen H, Komura T (2018). "Few-shot learning of homogeneous human locomotion styles". Computer Graphics Forum.* **37** *(7): 143–153. Wiley.*

41. ^*Mahmood N, Ghorbani N, Troje NF, Pons-Moll G, Black MJ. "AMASS: Archive of motion capture as surface shapes." In: Proceedings of the IEEE/CVF international conference on computer vision. 2019. p. 5442–5451.*

42. ^*Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ. "SMPL: A skinned multi-person linear model". ACM Transactions on Graphics (TOG).* **34** *(6): 1–16. ACM New York, NY, USA, 2015.*

43. ^*Yu Z, Huang S, Cheng Y, Birdal T (2023). "SignAvatars: A Large-scale 3D Sign Language Holistic Motion Dataset and Benchmark". arXiv preprint arXiv:2310.20436.* arXiv:2310.20436.

44. ^*Yoon Y, Wolfert P, Kucherenko T, Viegas C, Nikolov T, Tsakov M, Henter GE (2022). "The GENEA Challenge 2022: A large evaluation of data-driven co-speech gesture generation." In: Proceedings of the 2022 International Conference on Multimodal Interaction. pp. 736--747.*

45. ^*Karunratanakul K, Preechakul K, Suwajanakorn S, Tang S (2023). "Guided motion diffusion for controllable human motion synthesis." In: Proceedings of the IEEE/CVF International Conference on Computer Vision. p. 2151–2162.*

46. ^*Yuan Y, Song J, Iqbal U, Vahdat A, Kautz J. "Physdiff: Physics-guided human motion diffusion model." In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023. p. 16010-16021.*

47. ^*Xiao Z, Kreis K, Vahdat A (2021). "Tackling the generative learning trilemma with denoising diffusion gans". arXiv preprint arXiv:2112.07804. Available from:* https://arxiv.org/abs/2112.07804.

48. ^*Liu H, Iwamoto N, Zhu Z, Li Z, Zhou Y, Bozkurt E, Zheng B (2022). "Disco: Disentangled implicit content and rhythm learning for diverse co-speech gestures synthesis". Proceedings of the 30th ACM International Conference on Multimedia. pages 3764–3773.*

49. ^*Alexanderson S, Nagy R, Beskow J, Henter GE (2023). "Listen, denoise, action! audio-driven motion synthesis with diffusion models". ACM Transactions on Graphics (TOG).* **42** *(4): 1–20.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

doi.org/10.32388/L4S0VW