LivelySpeaker: Towards Semantic-Aware Co-Speech Gesture Generation

Yihao Zhi[1,*]   Xiaodong Cun[2,*]   Xuelin Chen[2]   Xi Shen[3]
Wen Guo[4]   Shaoli Huang[2]   Shenghua Gao[1,5,6,†]

[1]ShanghaiTech University   [2]Tencent AI Lab   [3]Intellindust   [4]INRIA
[5]Shanghai Engineering Research Center of Intelligent Vision and Imaging
[6]Shanghai Engineering Research Center of Energy Efficient and Custom AI IC
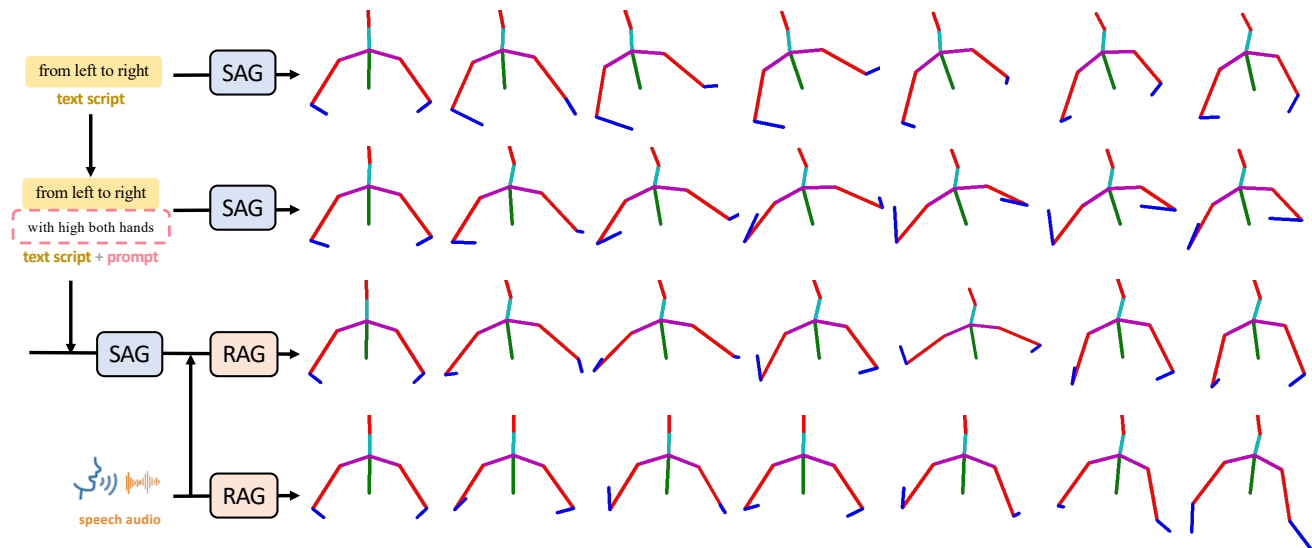
https://github.com/zyhbili/LivelySpeaker

Figure 1. We propose LivelySpeaker, a novel system that decouples the co-speech gesture generation into two stages, namely semantic-aware generator (SAG) and rhythm-aware generator (RAG), respectively. Powered by the proposed two-stage framework, our method can generate semantic-aware gestures (top three rows) other than purely audio-driven results (bottom row). We can also add additional prompts to the text script to specify the gestures (second line). Last, when we integrate the proposed components altogether, our method can still retain the semantic gestures and also appropriate rhythm for a lively speaker (third row).

Abstract

Gestures are non-verbal but important behaviors accompanying people's speech. While previous methods are able to generate speech rhythm-synchronized gestures, the semantic context of the speech is generally lacking in the gesticulations. Although semantic gestures do not occur very regularly in human speech, they are indeed the key for the audience to understand the speech context in a more immersive environment. Hence, we introduce LivelySpeaker, a framework that realizes semantics-aware co-speech gesture generation and offers several control handles. In particular, our method decouples the task into two stages: script-based gesture generation and audio-guided rhythm refinement. Specifically, the script-based gesture generation leverages the pre-trained CLIP text embeddings as the guidance for generating gestures that are highly semantically aligned with the script. Then, we devise a simple but effective diffusion-based gesture generation backbone simply using pure MLPs, that is conditioned on only audio signals and learns to gesticulate with realistic motions. We utilize such powerful prior to rhyme the script-guided gestures with the audio signals, no-

* Equal contribution.
† Corresponding author.

tably in a zero-shot setting. Our novel two-stage generation framework also enables several applications, such as changing the gesticulation style, editing the co-speech gestures via textual prompting, and controlling the semantic awareness and rhythm alignment with guided diffusion. Extensive experiments demonstrate the advantages of the proposed framework over competing methods. In addition, our core diffusion-based generative model also achieves state-of-the-art performance on two benchmarks. The code and model will be released to facilitate future research.
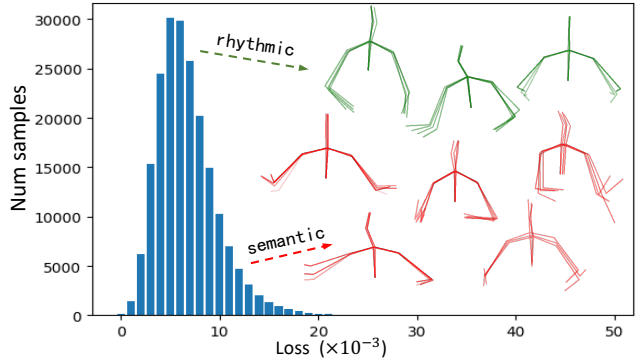

Figure 2. We plot the L_2 loss histogram on the training samples of the pre-trained trimodal method [58]. Although it is learned from multiple conditions (*e.g.*, text, audio), their methods are still dominated by repeated rhythm, and hard to model the rarely appeared diversity gestures, *e.g.*, the semantic-aware motions.

1. Introduction

During human conversation, non-verbal behaviors are typically present and among them, the most significant is gesture language. These non-linguistic gestures serve as an auxiliary but effective means of conveying key messages, enriching the conversation with contextual cues, and facilitating better understanding among participants. [9, 18, 24, 27]. Empowering the digital replicas of humans with the ability to gesticulate has been a long pursuit in the research community, as such ability can benefit many downstream applications, including digital humans in the coming virtual universe, non-player game characters, robot assistants, *etc*.

Given the speech content in the form of texts and/or audio streams, the objective is to generate realistic co-speech gestures. Traditional methods achieve this with hard-coded rules [10,11,25,41], *e.g.*, "*good*" in the speech will be simply represented by the gesture "*thump up*". However, these methods usually produce deterministic results; more importantly, they can not guarantee smooth transitions in the results. Recently, deep learning-based methods have been prevalent in the field of gesture generation from multi-modality inputs. In particular, these methods formulate the problem as conditional motion generation and tackle it via training a conditional generative model that takes as input the speaker identities [17], audio waves [46], speech texts [8], or a combination of these multi-modal signals [3, 40, 58]. Although multiple modalities are incorporated in the formulation, the results are often dominated by the rhythm of the audio signal since it is highly correlated with the performance of gestures during speech. While other works recognize the importance of the semantics conveyed through co-speech gestures, their framework heavily depends on pre-defined gesture types [7, 37] or keywords [57], making it difficult to express more complex intentions effectively.

We begin with insights from the following two perspectives: (*i*) Real-world human conversations contain a limited number of semantic gestures (See Fig. 2), which presents difficulties in learning co-speech gestures that are semantic-sensitive but rhythm-irrelevant. This partially explains why prior approaches have yielded results that heavily rely on the audio rhythm. (*ii*) Most previous methods are built on gen-

erative adversarial networks (GANs), which might be hard to train, especially when learning a many-to-many mapping between the text/audio and the gesture [46].

Following this, we present LivelySpeaker, a simple and effective framework for semantic-aware co-speech gesture generation. In particular, our framework *explicitly* decouples the generation into two stages, namely the *script-based gesture generation*, and the *audio-guided rhythm refinement*. Specifically, the script-based gesture generation leverages the pre-trained CLIP [47] text embeddings as the guidance for generating gestures that are highly semantically correlated with the textual script. In the second stage, we devise a simple but effective diffusion-based gesture generation backbone with pure MLPs, that is conditioned on only audio signals and learns to gesticulate with realistic motions. We utilize such powerful prior to rhyme the script-guided gestures with the audio signals, notably in a zero-shot setting. In detail, we gradually add Gaussian noise for T steps to the motion extracted from the dataset, on which an MLP-based [20] motion denoising model [53] is conditioned on the corresponding audio and predicts the clean motion. We show that this diffusion-based model is effective in rhyming rather smooth gestures produced from the script-based generation module with the audio signals.

Building upon these two powerful modules, our method can generate diverse and high-quality co-speech gestures that are semantically meaningful, given the textual description of the speech and audio. Extensive experiments show that the proposed framework yields state-of-the-art performance in co-speech gesture generation. We also conduct experiments to show the control ability of our method by extending it to a number of scenarios that are not possible with competing methods, including changing the gesticulation style, editing the co-speech gestures via textual prompting, and controlling the semantic awareness and rhythm alignment with guided

diffusion.

The main contribution of this paper is summarised:

- We propose LivelySpeaker, a novel two-stage framework for semantic-aware and rhythm-aware co-speech gesture generation.

- A novel MLP-based diffusion-based backbone is devised, that achieves state-of-the-art performance on the two benchmarks for co-speech generation.

- Our framework enables several new applications in co-speech gesture generation, such as text prompt-based gesture control, balancing the control between two different condition modalities (*i.e*., text and audio).

2. Related work

Co-speech Gesture Generation. As aforementioned, the research of co-speech gesture generation has taken several routes in the past decades, including rule-based [29, 30, 45], machine learning-based [26, 32, 50], and deep learning-based ones [7, 17, 33, 40, 46, 58]. In our coverage, we mainly review the deep learning-based ones, as they have shown better performance and are more relevant to our method. Earlier works along this line consider the problem as an end-to-end regression of 2D keypoints of the human body, with different dedicated designs on the network architectures. For example, Speech2Gesture [17] generates personalized 2D keypoints from audio using a conditional generative adversarial network. Ahuja *et al*. [1] propose a few-shot method for personalized motion transfer. Furthermore, being aware of the many-to-many nature of the co-speech gesture generation, [46] proposes a novel audio template-based method to reduce the uncertainty of the generation. Going beyond simple 2D keypoints, [39] proposes a co-speech generation framework that leverages an unsupervised motion representation instead of a structural human body prior and involves the image-based rendering technique to generate co-speech videos like talking face generation [63]. These methods are limited in their applicability to many real-world scenarios as they generate motions of 2D key points or directly output 2D imagery of the speaker.

To generate motions for 3D avatars, TriModal [58] extracts diverse upper body motions from the TED talks and designs an LSTM-based neural network that is conditioned on the audio, text, and identity and generates co-speech gestures. Speech2AffectiveGesture [7] extends this work for more semantic-aware gesture generation. Nevertheless, they have only demonstrated on five predefined gesture classes. As a follow-up that increases the effectiveness of Speech2AffectiveGesture, HA2G [40] further extracts the hand keypoints on TED datasets and uses a hierarchical GRU [13] network. More recently, Ao *et al*. [3] propose a method using VQ-VAE [55], and SEEG [37] is designed

to generate semantic gestures of several kinds. In summary, most of these methods learn in an end-to-end fashion, where the conditional audio in fact dominates the conditional generation, and only focuses on limited types of semantic gestures [7, 37]. Differently, our framework can produce co-speech gestures that are highly semantically aligned with the textual description of the speech provided by the user.

Conditional Motion Generation. Co-speech generation is also a sub-topic of human motion generation, which aims at generating 3D human motion from various conditions. One hottest topic is text-to-motion. Language2Pose [2] employs a curriculum learning approach to learn a joint embedding space for both text and pose. The decoder can thus take text embedding to generate motion sequences. Ghost *et al*. [16] extend it through manifold representations for the upper body and the lower body movements. Similarly, Motion-CLIP [52] also tends to align text and motion embedding but proposes to utilize CLIP [47] as the text encoder and employ rendered images as extra supervision. It shows the ability to generate out-of-distribution motion and enable latent code editing. As for the generation of the simple action, several methods [5, 43, 44, 60] have been proposed by a pre-defined action class [43], an additional text encoder [44] and temporal motion compositions from a series of natural language descriptions [5]. Guo *et al*. [19] also proposes to incorporate motion length prediction from text to produce motion with reasonable length. Motion can also be generated with music [4, 12, 31, 34–36], which has a similar form as our task but they only focus on the rhythm. *e.g*., [12] produce a graph-based network to optimize the choreography-aware features, Li *et al*. [35] generate the dance motions from a transformer-based network and a high-quality dataset. Li *et al*. [36] train a separate VQ-VAE [55] to model the upper body and lower body individually. Through sharing a similar goal with these works, our work differs from them as we consider both the semantic and rhythm awareness in our unified framework.

Diffusion-based Motion Generation. Very recently, the denoising diffusion-based models have also shown very promising results in generating human motion. Prior works of MDM [53] and MotionDiffuse [62] generate realistic motions from noise inspired by the denoising diffusion model [22]. PhysDiff [59] extend MDM via the physics-aware restriction. EDGE [54] design a stronger dance generation network using the powerful pre-trained audio model, jukebox [14]. These methods only use the diffusion-based model for conditional generation, whereas our method finds more interesting applications via the diffusion-based model in motion synthesis.

3. Method

Given the speech content in the form of audio, its corresponding text script, and the identity information, our system aims to generate 3D skeletal gestures that are semantically
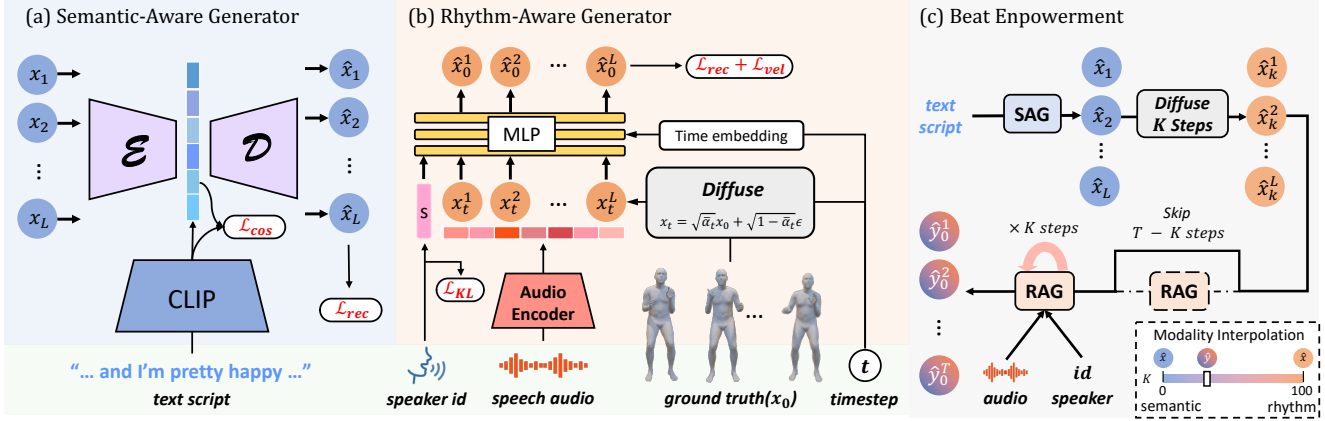
Figure 3. We propose a two-stage framework for semantic-aware gesture generation (a) and rhythm-aware gesture generation (b), respectively. Then, we combine them via a beat empowerment method as in (c) for the full pipeline of the proposed LivelySpeaker.

and rhythmically aligned with the speech.

We tackle this problem with a two-stage framework consisting of a semantic-aware generator (SAG), and a rhythm-aware generator (RAG), as shown in Fig. 3 (a) and (b), respectively. After training each component, we can generate the gestures from the text scripts first, and then leverage the rhythm-aware network as a beat empowerer as in Fig. 3 (c). In the following, we first present the details of the semantic-aware and rhythm-aware generator in Sec. 3.1 and Sec. 3.2, followed by details of the inference pipeline and application of the whole framework in Sec. 3.3.

3.1. Generating Semantic Gestures from Text Script

Current co-speech generation approaches [37, 40, 58] often consider the script and audio features equally with identical timestamps concatenation. In these approaches, the text features aligned with timestamps are more likely to act as alternative beat signals generating the rhythm-dominant gestures, which have a small effect on previous methods. To utilize the semantic information well, in the first stage of our framework, we only train a semantic-aware generator (SAG) to generate the gesture from text scripts.

Inspired by the progress of text to motion [52, 53, 60], we consider the text script as a kind of semantic description to generate the corresponding motion. As shown in Fig. 3 (a), we split the motion sequences into fixed segments and send them into an encoder-decoder-like Transformer [56] for motion generation [7, 58]. Our network contains 3-layer encoders and decoders. Each Transformer layer has a latent dimension of 512, and the dimension of the feed-forward layer equals 1024. To integrate the semantic-aware information, we use a pre-trained CLIP [48] of ViT-B/32 as the text embedding network, getting 512-dimension semantic features of the whole script sequences, other than the frame-wise semantic feature as in previous works [40, 58].

For training, we feed the ground truth pose sequences $x_{1:t}$ to the transformer encoder to generate the motion latent: $z_{emb} = E(x_{1:t}) \in R^{512}$, and a decoder is used to decode this latent code to reconstruct a sequence of poses $\hat{x}_{1:t} = D(E(x_{1:t}))$. Then, we calculate the distance between the semantic embeddings of CLIP z_{CLIP} and the latent code z_{emb} using a cosine similarity loss L_{cos}. We also measure the reconstruction loss L_{rec} between the generated motion and the original one using simple mean square error. The full training objective of SAG is:

$$L_{full} = L_{rec}(x_0, \hat{x}_0) + \lambda L_{cos}(z_{CLIP}, z_{emb}), \quad (1)$$

where we set $\lambda = 1$ empirically. In testing, we generate the motion sequences directly from the CLIP embedding.

3.2. Diffusion-Based Rhythm-Aware Generator

Although our SAG can produce some semantic-aware gestures, the out-sync gestures also restrict the realism of the generated motion. However, it is hard to align the temporal information of the generated motion only and keep other content unchanged. We take advantage of the diffusion-based model for its powerful ability in distribution modeling [22, 53] and editing [42, 51].

The denoising diffusion model is a Markov noising process, which first shows its potential in image generation [22]. Following the human motion diffusion model [53], the input pose sequences can be defined as $\{x_t^{1:N}\}_{t=0}^{T}$, where $x_0^{1:N}$ is sampled from the data distribution and

$$q(x_t^{1:N}|x_{t-1}^{1:N}) = N(\sqrt{\alpha_t}x_{t-1}^{1:N}, (1 - \alpha_t)I). \quad (2)$$

Here, $\alpha_t \in (0, 1)$ are constant numbers. When α_t is sufficiently small, we can make an approximation that $x_t^{1:N}$ follows a normal distribution with mean 0 and variance I. Henceforth, we will refer to the complete sequence at noise

step t as x_t. In our task, we also follow previous work [22] to predict the signal itself, *i.e.*, $\hat{x}^0 = G(x_t, t, c)$, where c is the conditional audio and G is our denoising backbone.

As for the network structure, our audio encoder C is simply constructed using four 1d-convolution blocks activated by leaky ReLU, where we feed in a raw audio waveform and generate a sequence of 256-channel feature vectors. We have also tested other similar audio encoders in previous work [15, 40], however, there is no observed performance gain. As for the denoising diffusion network G, different from the original MDM [53], we use N layers MLP-based network [20] to construct, which generates better rhythm and generate more smooth results. In detail, we first use a linear layer to project input data to a higher-dimension latent space. After applying a series of MLP blocks, a last linear layer is used to project the latent feature back to poses as output. Each MLP block is composed of one FC layer for temporal merging and one FC layer for spatial merging. For each MLP block, we use layer normalization (LN) [6] as pre-normalization, SiLU [49] as activation, and apply skip-connections [21]. As for the additional conditions, we concatenate the audio feature to the sampled pose and add the time step embedding t_{emb} to each MLP block. We also embed the speaker id into the vector and calculate the style embedding s through reparameterization, where s is concatenated along the temporal dimension.

For training the denoising network, we split the long motion sequence to the specific length and calculate the reconstruction loss via the Huber loss L_{huber} [58] of the diffusion model as:

$$L_{rec} = E_{x_0 \sim q(x_0|c), t \sim [1,T]}[L_{huber}(x_0, \hat{x_0})]. \qquad (3)$$

Similarly, we add velocity loss as:

$$L_{vel} = E_{x_0 \sim q(x_0|c), t \sim [1,T]}[L_{huber}(\dot{x_0}, \dot{\hat{x_0}})]. \qquad (4)$$

Besides, since human motion is subject-related [40, 58], the Kullback–Leibler divergence [28] L_{KL} is used to regularize the distribution of all speakers on the s embeddings.

Overall, the loss function of training the rhythm-aware generator can be written as:

$$L_{full} = L_{rec} + \lambda L_{KL} + \beta L_{vel}, \qquad (5)$$

where λ and β equals to $1e^{-2}$ and 1, respectively. We follow the previous works [40, 58] and set the threshold for the L_{huber} to 0.1. To generate longer motions, we concatenate 4 previous frames to achieve visual continuity similar to previous works [17, 40, 58].

3.3. Full LivelySpeaker Pipeline and Applications

After training both the semantic-aware and rhythm-aware generator models, we can utilize the latter to address the rhythm issues with the output of the former. In detail, as shown in Fig. 3 (c), after generating the semantic-aware motions from the SAG, following SDEdit [42], we can invert the generated motion by adding K steps noises, and then, we consider this motion as the generated motion of $T - K$ ($K = 20$ in our cases), denoising it to a new distribution via the guidance of the audio using DDIM ($T = 100$). When inferring a long sequence, we repeat the procedure mentioned above for each motion clip (consisting of 34 frames) sequentially and then concatenate them together. Thanks to the power of the diffusion-based model, this simple beat empowerment step keeps both the diversities from the semantic-aware generator and hugely increases the rhythm's alignment.

Since we learn each stage individually and each stage models a different distribution, our methods enable some interesting applications. Below, we give brief introductions: **Semantic motion generation via new text prompts.** We find that the individual learned semantic-aware generator is also a good controllable gesture generator. We can add some new text prompts to the CLIP encoder of our semantic-aware generator, and our method also generates the corresponding motions. We give a brief visualization in Fig. 1 where the generated gestures are in the new pose. This phenomenon reveals that even if the semantic information rarely appears in the dataset, it is also learned by our network. We give more examples in the supplementary videos to show the effectiveness of the proposed methods.
Interpolating poses between different modalities. Our method enables the applications of generating different gestures of the script-based motion and the rhythm-based refinements by controlling the denoising steps of the diffusion model (see Fig. 3 (c)). We give the comparison and details in the supplementary materials.

4. Experiments

4.1. Datasets

We validate our pipeline using two datasets, including the TED Gesture dataset [58] and BEAT dataset [38]. The TED Gesture dataset [58] contains 1766 videos sourced from online TED speech videos and utilizes three modalities: audio, text, and speaker identity. The human pose is represented by direction vectors of 10 upper body joints.

Besides the body movements, clean finger movements are also essential for a lively speaker's delivery. Therefore, instead of using the noisy TED-Expressive [40] which captures the figure motion by OpenPose, we evaluate the performance of the newly introduced high-quality dataset BEAT [38] for its high fidelity on hand poses. BEAT [38] is constructed using a commercial MOCAP system, including additional annotations for emotion and semantic modalities. It captures the rotation angles of joints that are invariant to body shape. During training, we convert its origin Euler angle to rot6d

representation [64] to ensure better convergence.

Following the previous settings [38, 40], both datasets are resampled into 15 fps and divided into fragments of 34 frames in length with overlapping clips.

4.2. Evaluation Metrics

We adopt three main metrics to evaluate the generation quality, including Frechet Gesture Distance (FGD) [58], Beat Consistency Score (BC [40]), and Diversity [40] as previous methods [40, 58]. FGD is a metric to measure the distribution disparity between generated output and ground truth across the entire dataset, where a pre-trained autoencoder is used to project motion into latent space. On TED [58] dataset, we use the autoencoder provided in [58] for a fair comparison, while on BEAT [38] we re-train the autoencoder using rot6d representation. Beat Consistency Score calculates the average distance between every audio beat and its nearest motion beat. Intuitively, the denser the motion beats are, the better BC. Thus, it would be invalid in cases of anomalous gesture sequences that contain numerous motion beats. Fortunately, we can take FGD as a reference to solve it. Diversity is assessed by measuring the variations in generated gestures, which are also calculated using a pre-trained autoencoder [58]. It is computed by the L_1 distance between randomly sampled motion feature pairs [40].

4.3. implementation Details

The training is composed of two stages. For Semantic Aware Generator, we train it with the Adam optimizer ($lr = 0.0001, \beta = (0.9, 0.99)$) for 400 epochs. As for Rhythm Aware Generator, we train it with the AdamW optimizer ($lr = 0.0001, \beta = (0.9, 0.999)$) for 1200 epochs. The total diffusion steps $T = 1000$ in training, in inference, we generate the motion via 20 steps DDIM sampler. All experiments are conducted with a batch size of 512 on a single NVIDIA A100. When evaluating metrics, we use DDIM [51] with 100 steps for faster sampling.

4.4. Baselines

We compare our method with the following methods. **Speech2Gestures** [17] and **Trimodal** [58] are two representative methods in co-speech gesture generation. Trimodal fuses three-modality information and achieves better performance than S2G. **HA2G** [40], the SOTA model on TED Gesture dataset, implements a coarse-to-fine hierarchical gesture generator and learns a powerful audio extractor through contrastive learning. **CaMN** [38], the SOTA model on the BEAT dataset, designs a cascaded architecture and takes into account all six modalities present in the BEAT dataset. All these methods are learned in an end-to-end fashion where the audios are dominated the gesture generation process. We implement their open-source code on two datasets to conduct a fair comparison.

Methods	FGD↓	BC↑	Diversity↑
Real Video	0	0.697	108.780
S2G [17]	24.887	0.723	97.272
TriModal [58]	4.501	0.659	102.978
HA2G [40]	5.429	0.698	106.290
Ours Rhythm (w=1)	2.152	0.656	107.988
Ours Rhythm (w=1.5)	2.359	0.676	112.327
Ours Rhythm (w=2.2)	6.622	0.699	113.051
Ours Full (w=1)	11.310	0.634	108.663
Ours Full (w=1.5)	9.154	0.664	107.781
Ours Full (w=2.2)	8.446	0.696	109.880

Table 1. Comparison with baselines on TED Gesture dataset. Our method outperforms the three baselines in most cases.

Methods	FGD↓	BC↑	Diversity↑
Real Video	0	0.867	216.541
S2G [17]	24.887	0.872	152.367
TriModal [58]	20.513	0.621	173.214
CaMN [40]	8.169	0.768	183.671
Ours Rhythm (w=1)	7.845	0.886	193.060
Ours Rhythm (w=1.5)	7.561	0.892	206.969
Ours Full (w=1)	10.863	0.886	183.201
Ours Full (w=1.5)	9.269	0.893	194.362

Table 2. Comparison with baselines on BEAT dataset. Our method achieves the best performance in most cases.

Note that several recent works [3, 37] also achieve noticeable performance. We do not compare to SEEG [37] since they utilize additional data annotations (Semantic Prompt Gallery). Rhythmic Gesticulator [3] lacks open-source codes. It employs an elaborately designed rhythm-based segmentation strategy to construct its training data, which is different from previous settings [37, 40, 58].

4.5. Quantitative Evaluation

4.5.1 Rhythm-aware Diffusion Generator

Thanks to the ability of the diffusion-based method, we can generate the co-speech gestures with varying guidance weights w during inference by employing the classifier-free guidance sampler [23]. We exhibit the numerical results in the Table. 1 and Table. 2 marked with *Ours Rhythm*. We can observe that the Beat Consistency Score and Diversity is proportional to the guidance weight. Meanwhile, we achieve our best FGD when $w = 1$. As demonstrated in Table. 1, our RAG is able to beat all baselines in most cases on the TED dataset [58]. S2G [17] achieves the highest Beat Consistency Score. However, our observation reveals that it generates rapid and unnatural body movements regardless of the audio beats, as shown in our supplementary video, resulting in the worst FGD and abnormal Beat Consistency Score, which
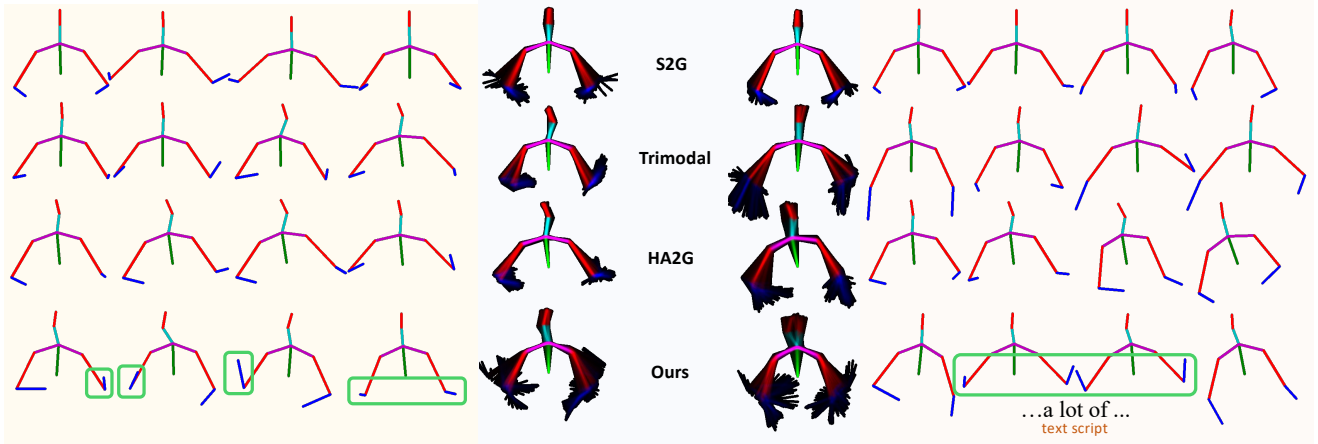
Figure 4. **Visual comparisons with three baselines**. We display the scope of gestures across frames in the blue background area. Larger scopes mean better diversity. Despite the close Beat Consistency Scores compared to the three baselines, our method stands out by clearly greater diversity as shown in both examples. On the left side of the figure, we roughly show the rhythm of the proposed method. According to Table. 1, these methods all achieve high and closely comparable BC scores. However, upon visualizing the results, our result excels in terms of diversity. Specifically, our result changes hands from left to right and then waves both hands. In contrast, the baselines maintain a consistent motion pattern throughout. On the right of the figure, we give the text scripts ("... a lot of..."), where the proposed method can produce semantic-aware gestures whether other methods may fail.

further shows the weakness of this metric as discussed in recent work [54]. Meanwhile, on the BEAT [38] dataset, *Ours Rhythm* and CaMN outperform previous works across all metrics. Despite utilizing only three modalities (audio, emotion, and speaker ID) as input, our results show comparable FGD as the state-of-the-art model that employs all five modalities. Furthermore, our rhythm-aware Diffusion Generator excels in generating diverse and rhythmically complex gestures thanks to our powerful MLP-based diffusion model.

4.5.2 Full System

As mentioned in Sec. 3.3, our RAG has the capability to add beats to any motion sequence in a two-step process of diffusing and denoising through DDIM [51]. We evaluate the whole system as in Fig. 3 (c) with 20 diffuse steps by adding Gaussian noise to the generated motion. The results of our full system are listed in Tab. 1 and 2 marked with *Ours Full*. On the TED dataset, our full system is also competitive when compared to existing methods. As discussed in SEEG [37], the slight downgrade in the FGD can be attributed to the fact that the semantic gestures exhibit worse metrics. A similar observation has been founded in Tab. 2, where our full network still keeps the similar Beat Consistency.

Since human motion is hard to be visualized, we give a simple comparison of the diversity and the semantic-aware gestures in Fig. 3, where the proposed method generates very diverse gestures than the baseline methods. Besides, the proposed method also shows the semantic-aware gestures from this example. This further shows that the proposed SAG

Methods	Natural	Smooth	Diversity	Semantic
S2G [17]	41.6%	36.4%	33.3%	37.5%
TriModal [58]	6.30%	5.20%	10.4%	6.30%
HA2G [40]	9.40%	9.40%	10.4%	11.4%
Ours Full	**42.7%**	**49.0%**	**45.8%**	**44.7%**

Table 3. The percentage of the user's favorite methods in terms of naturalness, smoothness, diversity, and semantics.

learns some out-of-domain knowledge in terms of FGD and Diversity but still keeps the Beat Consistency. We give more examples in the supplementary video for comparison.

4.5.3 User Studies

Since the generated content is very subjective, we conduct a user study to show the effectiveness of the proposed method over the state-of-the-art methods. Specifically, we ask 16 subjects on four different methods (*i.e.*, Speech2Gesture [17], Trimodal [58], HA2G [40] and ours). We provide 12 samples of the results and let them choose the best one in terms of the motion naturalness, the motion smoothness, the diversity of the generated content, and the semantic preservation, yet 768 opinions in total. We then calculate the percentage of each task on each metric. As shown in Tab. 3, the participants like our methods most in terms of four metrics.

4.6. Ablation Studies

We ablation two different designs in our method on the TED dataset [58] for ablation studies. Firstly, we evaluate the effectiveness of the whole system. On the other hand, we

ablate the performance of our rhythm-aware diffusion model. More experiments are provided in the supplementary.

4.6.1 System Overview

Since each stage is considered individually, we can numerically evaluate the performance of each stage. As shown in Table. 1 and Table. Table. 2, we have evaluated that our full pipeline can utilize the rhythm of the rhythm-aware generator. Here, we give more ablation studies. As shown in Table. 4, although our SAG generates a very different distribution than the original model (FGD) and bad beat consistency, our single semantic-aware generator gains much more diverse motions than previous methods. As a combination of our two networks (in the third row), the proposed diffusion model will hugely pull the distribution to the trained one (as represented by FGD and BC) but still has very diverse results. We also try another method that utilizes the fast Fast Fourier Transform (FFT) to remove the high-frequency beat information and synthesize the hand-crafted dataset for training the beat alignment network, as shown in the second row of Table. 4, this beat alignment network is less effective in terms of FGD and only improve the beat consistency a little.

4.6.2 Network Structure Ablation on the Rhythm-Aware Diffusion Model

As for the rhythm-aware diffusion model, we set the guidance weights of classifier-free guidance $w = 1$ in this section. To validate the effectiveness of our MLP-based model, we replace it with a widely-used transformer structure in the recent motion diffusion model [53]. Specifically, we keep others unchanged and utilize the noisy motion sequence as a query and employ the audio features as the key and value to calculate the cross attention. As demonstrated in the second row of Table. 5, our MLP-based model exhibits clear superiority over the model built upon the Transformer Decoder, especially in terms of the Beat Consistency Score. This is crucial for ensuring effective beat empowerment. Meanwhile, we conducted an ablation study on the audio encoder. In this regard, we rebuilt the audio encoder using 2D convolutions and utilized a 128-channel Mel-spectrogram as the audio input. As shown in the third row of Table 5, the result indicates that a relatively simple audio encoder proves to be sufficiently expressive for the task. Additionally, we also evaluate the impact of our loss components. The KL divergence loss item improves FGD and the diversity metrics to a certain extent since it regulates the talking style.

4.7. Limitation

We propose a diffusion-based rhythm-aware generator that acts as a beat empowerment module, allowing for editing given motion in diffusing first and then denoising manner. Therefore, the inversion steps K are of great significance.

Methods	FGD↓	BC ↑	Diversity↑
SAG	56.878	0.388	128.894
SAG + Syn. data	65.718	0.472	133.753
LivelySpeaker (Ours)	**8.446**	**0.696**	109.880

Table 4. Ablation Studies on the whole framework where the proposed rhythm-aware Generator can be considered as a stronger beat empowerer.

Methods	FGD↓	BC ↑	Diversity↑
Ours Rhythm ($w = 1$)	**2.152**	**0.656**	**107.988**
w/ Transformer	6.509	0.418	104.737
w/ Mel-Spectrogram	4.951	0.568	101.952
w/o KL loss	5.256	0.650	105.126

Table 5. Ablation studies on the network structure of the proposed rhythm-aware generator.

For instance, when editing with large diffusing steps K (in extreme cases up to 100), the original motion would be drowned out by the Gaussian noise. If we could get the pair various data of paired sync and out-of-sync data, our results would be further improved via controllable adaptor [61]. Similarly, for long sequence generation, individual guidance weight w should also be taken into consideration. As for our semantic-aware generator, its performance is limited by sentence splitting. Take Fig. 1 as an example, we cannot generate a semantic-aware motion with the bad phrases split, like '... from left' and 'to right...'. Instead of splitting data using a sliding window (as done in most recent methods), we would pursue a better solution, like pre-parsing the sentence during the training and testing.

5. Conclusion

In this paper, we present LivelySpeaker, a novel semantic- and rhythm-aware system for co-speech gesture generation. To achieve this, we first generate the motion from the semantic-aware generator, after that, we train a diffusion-based rhythm-aware generator and utilize it for rhythm-aware refinement. Powered by our decoupled framework, our method enables multiple new applications in co-speech generation, including text-based pose style controlling, and interpolating between the text- and audio-based gestures. Besides, our pure diffusion-based backbone also achieves state-of-the-art performance in co-speech gesture generation.

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] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *International Conference on 3D Vision (3DV)*, 2019. 3

[3] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. In *SIGGRAPH Asia*, 2022. 2, 3, 6

[4] Andreas Aristidou, Anastasios Yiannakidis, Kfir Aberman, Daniel Cohen-Or, Ariel Shamir, and Yiorgos Chrysanthou. Rhythm is a dancer: Music-driven motion synthesis with global structure. *arXiv*, 2021. 3

[5] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. TEACH: Temporal Action Compositions for 3D Humans. In *International Conference on 3D Vision (3DV)*, 2022. 3

[6] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[7] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, New York, NY, USA, 2021. Association for Computing Machinery. 2, 3, 4

[8] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 1–10. IEEE, 2021. 2

[9] Justine Cassell, David McNeill, and Karl-Erik McCullough. Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & cognition*, 7(1):1–34, 1999. 2

[10] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. Animated conversation: Rule-based generation of facial expression, gesture amp; spoken intonation for multiple conversational agents. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '94, page 413–420, New York, NY, USA, 1994. Association for Computing Machinery. 2

[11] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. *BEAT: the Behavior Expression Animation Toolkit*, pages 163–185. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. 2

[12] Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)*, 2021. 3

[13] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 3

[14] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 3

[15] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *CVPR*, pages 18770–18780, 2022. 5

[16] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[17] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. Learning individual styles of conversational gesture. In *CVPR*. IEEE, June 2019. 2, 3, 5, 6, 7

[18] Susan Goldin-Meadow and Martha Wagner Alibali. Gesture's role in speaking, learning, and creating language. *Annual review of psychology*, 64:257–283, 2013. 2

[19] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the Conference on CVPR*, 2022. 3

[20] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *WACV*, pages 4809–4819, 2023. 2, 5

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS (NeurIPS)*, 2020. 3, 4, 5

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[24] Autumn B Hostetter and Martha W Alibali. Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, 15(3):495–514, 2008. 2

[25] Chien-Ming Huang and Bilge Mutlu. Robot behavior toolkit: generating effective social behaviors for robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 25–32, 2012. 2

[26] Chien-Ming Huang and Bilge Mutlu. Learning-based modeling of multimodal behaviors for humanlike robots. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 57–64, 2014. 3

[27] Jana M Iverson and Susan Goldin-Meadow. Why people gesture when they speak. *Nature*, 396(6708):228–228, 1998. 2

[28] James M Joyce. Kullback-leibler divergence. In *International encyclopedia of statistical science*, pages 720–722. Springer, 2011. 5

[29] Michael Kipp. *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers, 2005. 3

[30] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents: 6th International Conference, IVA 2006, Marina Del Rey, CA, USA, August 21-23, 2006. Proceedings 6*, pages 205–217. Springer, 2006. 3

[31] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In *nips*, 2019. 3

[32] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. Gesture controllers. In *Acm siggraph 2010 papers*, pages 1–11. 2010. 3

[33] Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Linchao Bao, and Zhenyu He. Audio2gestures: Generating diverse gestures from audio. *arXiv preprint arXiv:2301.06690*, 2023. 3

[34] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *arXiv*, 2020. 3

[35] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *CVPR*, 2021. 3

[36] Siyao Li, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the Conference on CVPR*, 2022. 3

[37] Yuanzhi Liang, Qianyu Feng, Linchao Zhu, Li Hu, Pan Pan, and Yi Yang. Seeg: Semantic energized co-speech gesture generation. In *CVPR*, 2022. 2, 3, 4, 6, 7

[38] 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 *ECCV*, pages 612–630. Springer, 2022. 5, 6, 7

[39] Xian Liu, Qianyi Wu, Hang Zhou, Yuanqi Du, Wayne Wu, Dahua Lin, and Ziwei Liu. Audio-driven co-speech gesture video generation. *NeurIPS*, 2022. 3

[40] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *CVPR*, pages 10462–10472, 2022. 2, 3, 4, 5, 6, 7

[41] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '13, page 25–35, New York, NY, USA, 2013. Association for Computing Machinery. 2

[42] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 4, 5

[43] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[44] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3

[45] Helmut Prendinger, Sylvain Descamps, and Mitsuru Ishizuka. Mpml: A markup language for controlling the behavior of life-like characters. *Journal of Visual Languages & Computing*, 15(2):183–203, 2004. 3

[46] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. Speech drives templates: Co-speech gesture synthesis with learned templates. In *ICCV*, pages 11077–11086, 2021. 2, 3

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2, 3

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[49] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 5

[50] Mehmet E Sargin, Yucel Yemez, Engin Erzin, and Ahmet M Tekalp. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1330–1345, 2008. 3

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4, 6, 7

[52] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3, 4

[53] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *arXiv*, 2022. 2, 3, 4, 5, 8

[54] Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. Edge: Editable dance generation from music. *arXiv preprint arXiv:2211.10658*, 2022. 3, 7

[55] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017. 3

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS (NeurIPS)*, 2017. 4

[57] Jing Xu, Wei Zhang, Yalong Bai, Qibin Sun, and Tao Mei. Freeform body motion generation from speech. *arXiv preprint arXiv:2203.02291*, 2022. 2

[58] 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. *ACM Transactions on Graphics*, 39(6), 2020. 2, 3, 4, 5, 6, 7

[59] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. *arXiv preprint arXiv:2212.02500*, 2022. 3

[60] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052*, 2023. 3, 4

[61] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 8

[62] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv*, 2022. 3

[63] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. 3

[64] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, pages 5745–5753, 2019. 6