Multi-Modal Controlled Coherent Motion Generation

Yifei Liu[1,2], Qiong Cao[2], Hongwei Yi[3], Huaiguang Jiang[1], and Changxing Ding[1] *

[1] South China University of Technology
ft_lyf@mail.scut.edu.cn, hihuagong2021@scut.edu.cn, chxding@scut.edu.cn
[2] Joy Future Academy
mathqiong2012@gmail.com
[3] Peking University
hongweiyi@pku.edu.cn

Abstract. It is natural for us to walk and talk simultaneously. This paper tackles the challenge of replicating such natural behaviors in 3D avatar motion generation driven by concurrent multi-modal inputs, *e.g.*, a text description "a man is walking" alongside a speech audio. Existing methods, constrained by the scarcity of aligned multi-modal data, typically combine motions from individual modalities sequentially or through weighted sum. However, they often result in mismatched or unrealistic movements. To overcome these limitations, we propose **MOCO**, a novel diffusion-based framework capable of processing multiple simultaneous inputs—including speech audio, text descriptions, and trajectory data—to generate coherent and lifelike motions without requiring aligned multi-modal data. Our key innovation lies in decoupling the motion generation process. During each denoising step, the diffusion model independently generates motions for each modality from the input noise and assembles the body parts according to predefined spatial rules. The resulting combined motion is then diffused and serves as the input noise for the subsequent denoising step. This iterative approach enables each modality to refine its contribution within the context of the overall motion, progressively harmonizing movements across modalities. Consequently, the generated motions become increasingly natural and fluid with each iteration, achieving coherent and synchronized behaviors. We evaluate our approach using a purpose-built multi-modal benchmark. Experimental results demonstrate that **MOCO** outperforms existing baselines, advancing the field of multi-modal motion generation for 3D avatars. The code will be released on https://feifeifeiliu.github.io/MOCO.

Keywords: Human Motion Generation · Diffusion Models · Multi-Modal

1 Introduction

Imagine watching a virtual talk show where the host delivers engaging dialogue complemented by expressive gestures, natural body movements, and precise

* Corresponding author.

movement paths. The host walks across the stage following a scripted trajectory, uses hand gestures to emphasize points based on their speech, and shifts posture in response to both the conversation's flow and predefined text instructions—all occurring in perfect harmony. This level of realism transforms the viewing experience, making interactions feel genuine and immersive. Achieving such lifelike behavior is no small feat, yet it is essential for enhancing user engagement in applications ranging from virtual reality to interactive gaming and beyond.

Driving a 3D avatar to perform such lifelike motions involves managing multiple control signals, such as text descriptions, speech audio, and trajectory data. Multi-modal signals may be provided concurrently, for instance, a text prompt like "a man is walking" alongside a speech audio clip. However, most prior works primarily focus on single-modality control, such as text-to-motion [18, 53] or speech-to-gesture [17, 62]. Recent studies [66, 71, 72] have explored designing unified models capable of addressing multi-modal signals by leveraging datasets from different generation tasks. Nevertheless, these models typically process only one modality at a time, combining motions conditioned on different inputs in a limited and sequential manner when multiple control signals are present.

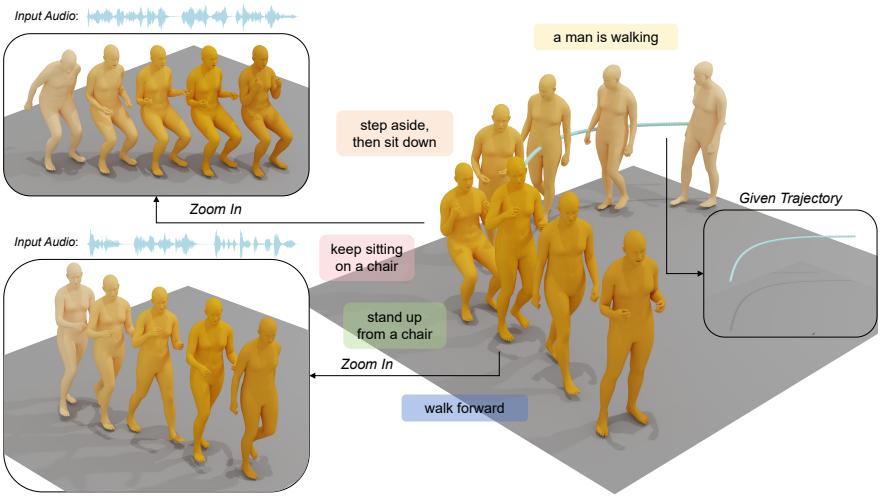

Fig. 1: Examples of Multi-Modal Controlled Motion Generation. Given multiple control signals of different modalities—including text descriptions, speech audio, and trajectory data—our MOCO framework generates realistic and coherent holistic body motion. This includes both body movements and detailed features such as facial expressions and hand gestures, all closely aligned with the provided conditions. To clearly illustrate this, we highlight two clips with temporal zoom, showcasing the natural integration of speech gestures and lower-body movements in our generated motions. Additional visual results are provided in Fig. 4 and the **supplementary materials**.

The primary challenge in achieving simultaneous multi-modal control of motion generation is the scarcity of aligned multi-modal data. While collecting additional multi-modal data could help, it requires significant resources. In addition, the activity regions in speech-to-gesture datasets are often limited, making it hard to train models that can generate trajectory-controlled speech gestures. Some efforts address this issue by combining the predictions of text-driven model and audio-driven model through weighted sum [12, 61]. [34] suggested using speech scripts as pseudo text labels to create aligned text-audio-motion datasets, replacing scripts with movement descriptions during inference. However, these approaches face inherent limitations: for weighted sum, the predominance of standing poses in speech-to-gesture datasets creates imbalanced model corrections; for pseudo-labels, the use of speech scripts as training labels limits generalization to diverse motion descriptions during inference. To overcome these challenges, we propose a novel diffusion-based framework, named **M**ulti-M**O**dal **C**ontrolled C**O**herent Motion Synthesis (MOCO). MOCO exploits a natural spatial decomposition: speech audio primarily drives upper-body dynamics——gestures and facial expressions——while text descriptions mainly influence lower-body movements such as walking and stance shifts (see Appendix A for supporting statistics). To achieve this, MOCO is first trained on multiple datasets, ensuring that the model can independently generate motions conditioned on either text or speech inputs. At each denoising step, the model generates motions for each modality separately from the input noise and assembles the body parts according to predefined spatial rules, i.e. combining audio-driven upper-body motion with text-driven lower-body motion. This combined motion is then diffused and used as the input noise for the next denoising step. The separation ensures that each body part's motion is highly aligned with its corresponding input condition, while the iterative process conditions each generation step on the current state of the combined motion. This allows each modality to refine its contribution within the context of the overall movement. Consequently, with each iteration, the motions generated for different body parts become increasingly harmonized, resulting in natural movements that exhibit coherent and synchronized behaviors. Furthermore, this decoupled generation process enables our framework to incorporate trajectory control into co-speech motion generation. We can leverage trajectory data to generate text-driven motion and combine it with audio-driven motion, producing speech gestures that closely align with the given trajectory.

To facilitate the evaluation of this novel task, **we develop a multi-modal benchmark** comprising 1,000 test clips which are generated from 40 fundamental text descriptions of body movements (*e.g.*, "walk forwards" and "step back and sit down") and 694 audio clips from eight different speakers. Each test clip integrates two text prompts describing a movement with two speech audio clips. We rigorously evaluate our approach against baseline methods using both text-to-motion and speech-to-gesture metrics. Experimental results demonstrate that our method outperforms existing baselines, advancing the field of multi-modal controlled motion generation for 3D avatars.

2 Related Work

2.1 Multi-Modal Conditioned Motion Generation

In recent years, human motion generation has received significant attention, largely driven by advancements in dataset collection. Various scenarios have been explored depending on the input conditions, including action labels [19], text descriptions [6, 14, 18, 21, 24, 25, 32, 53, 56, 65, 67], speech audio [17, 28, 36–38, 44, 47, 62, 69], music [31, 49, 54], scene context [22, 41], spatial signal [52, 59], and even the motion of another person [11, 20, 40, 42, 51, 55]. Beyond single-modality control, research efforts have been extended to handle integrating multi-modal signals. For instance, some works considered speaker identity, speech audio, and transcripts to generate conversational gestures [2, 3, 60, 64], while other works jointly leverage text and scene inputs for motion generation [10, 15, 57, 63]. Moreover, some works focus on integrating various datasets to train unified motion models that enhance scalability and applicability across multiple scenarios [13, 26, 66, 70, 72].

A critical limitation persists, however: current models depend heavily on aligned multi-modal training data, limiting their ability to handle novel input combinations (*e.g.*, text with audio, audio with trajectory) unseen in training. To address this, FreeTalker [61] and SynTalker [12]—the latter concurrent with the first version of this paper [39]—propose combining predictions from text-driven and audio-driven models through weighted sum. Liu et al. [34] suggested using speech scripts as pseudo text labels (*e.g.*, "A male speaker is saying: 'I am shocked by what you have done.' ") to create aligned text-audio-motion datasets for training and replace these scripts with movement descriptions during inference. However, the weighted sum method poses challenges, as the speech-driven model exerts significantly stronger correction forces than the text-driven model. This imbalance often generates motions that overly favor speech conditions while suppressing text conditions (Appendix E). Regarding pseudo labels, the reliance on speech transcripts as supervisory signals creates a domain gap between the transcript distribution and the motion-description distribution, thereby constraining generalization to diverse motion descriptions.

2.2 Diffusion-based Methods in Motion Generation

The diffusion model has been recognized as one of the most advanced generative paradigms and has also gained significant attraction in human motion generation. Its applications can be categorized into two primary streams according to the space where the denoising process is implemented. One representative method in the first stream is the Human Motion Diffusion Model (MDM) [53], which operates directly in the original motion space to perform denoising [1, 27, 29, 33, 73]. This approach offers several advantages, including high editability and strong controllability, enabling operations such as concatenation and combination within the original motion space during the denoising process. The second stream is exemplified by the Motion Latent Diffusion model (MLD) [14], which conducts denoising in the latent space of a Variational Autoencoder (VAE) [8, 16, 46].

Utilizing the VAE latent space improves efficiency in representing complex motion data and significantly reduces the computational load required for diffusion processes, resulting in faster inference speeds. In this work, we select MDM as the foundational model due to its superior editability [5, 43, 48].

2.3 Compositional Motion Generation

A complementary line of work generates complex motion by *composing* multiple actions or sub-motions. Spatial composition assembles motions of different body parts: SINC [5] combines multiple text-driven actions through body-part masks in a VAE-based, offline manner. Temporal composition instead sequences multiple actions over time and focuses on seamless transitions between them, as in TEACH [4], T2LM [30], and FlowMDM [9]. STMC [43] unifies both axes, bringing per-denoising-step body-part composition into the diffusion framework with a DiffCollage-style [68] mechanism for temporally overlapping actions. All these methods compose homogeneous, text-only conditions; we instead build on the per-step composition paradigm and extend it to heterogeneous audio, text, and trajectory control, placing MOCO at the intersection of compositional and multi-modal motion generation.

3 Method

In this section, we begin with a brief introduction to the Motion Diffusion Model (MDM) [53], which serves as our foundational model (Sec. 3.1). Then, we describe our framework's data representations and important model components (Sec. 3.2). Next, we explain our multi-modal decoupled denoising strategy for holistic motion generation (Sec. 3.3). Finally, we address a more complex scenario where the trajectory condition is included and utilize a Large Language Model for motion planning (Sec. 3.4). An overview of our framework can be found in Fig. 2.

3.1 Preliminary: Motion Diffusion Model

MDM is a diffusion-based motion generation model [53]. The same as other diffusion models, it regards diffusion as a Markov noising process $(x_t)_{t=0}^{T}$ that starts from a training sample x_0. As the time step t increases, the distribution of x_T approaches a standard normal distribution. Besides, MDM models the conditional distribution $p(x_0 \mid c)$ by reversing the diffusion process through iterative denoising on x_T. To achieve this, it minimizes a loss function that penalizes the difference between the original and denoised samples. Moreover, it performs sampling from $p(x_0 \mid c)$ iteratively, predicting x_0 at each timestep t and computing x_{t-1} until t reaches 0.

MDM adopts the classifier-free guidance [23] to adjust its adherence to the conditioning signal c. Specially, its denoiser G is trained on both conditioned and unconditioned data by randomly setting c as $\varnothing$, which allows $G(x_t, t, \varnothing)$ to

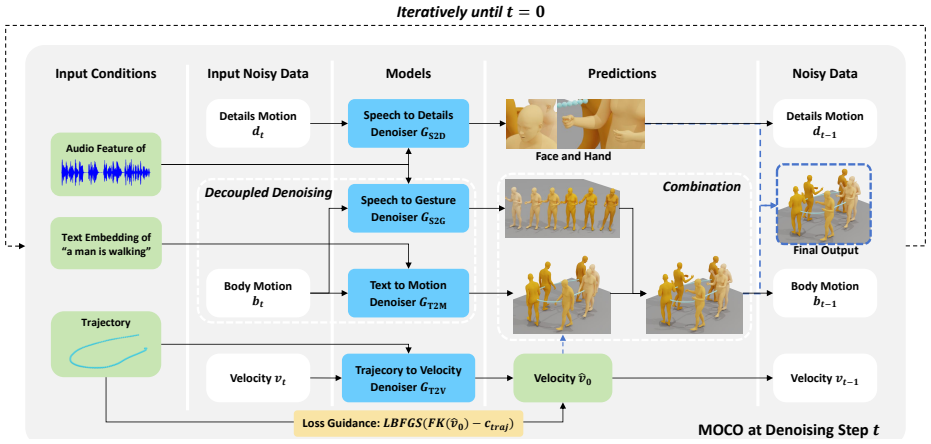

Fig. 2: Overview of MOCO. At each denoising step t, the input conditions and noisy data are fed into their respective denoisers to predict clean motion, which is then diffused for the next iteration. Specifically, the upper-body motion conditioned on speech audio and the lower-body motion conditioned on text description are combined to form the overall body motion. The blue arrows in the figure highlight two key points. One indicates that the denoising process of v_0 is completed before body motion denoising. The other shows that after the denoising process, the detailed facial and hand movements, and the combined body motion are integrated together to produce the final holistic motion.

approximate the unconditional data distribution. During sampling, it adjusts the strength of c using a scaling factor s:

$$G^s\left(x_t, t, c\right) = G\left(x_t, t, \varnothing\right) + s \cdot \left(G\left(x_t, t, c\right) - G\left(x_t, t, \varnothing\right)\right), \tag{1}$$

where G^s denotes the denoiser with the classifier-free guidance.

3.2 Data Representation and Model Architecture

Data Representation. Our framework incorporates four data modalities: motion, text, audio, and trajectory. The motion data is represented as $m = \{m^n\}|_{n=1}^{N} \in \mathcal{R}^{N\times 491}$, where N is the number of frames. Specifically, the motion data for each frame is denoted as $m^n = \{b^n, d^n\}$, with $b \in \mathcal{R}^{205}$ representing the body pose [43], and $d \in \mathcal{R}^{286}$ capturing detailed facial expression and hand movements. The text embeddings are encoded using the CLIP model [45] and are denoted as $c_{text} \in \mathcal{R}^{512}$. Audio features are extracted via the Wav2Vec2 model [7] and represented as $c_{audio} \in \mathcal{R}^{N\times 768}$. Finally, the trajectory data is encoded as $c_{traj} \in \mathcal{R}^{N\times 2}$, representing the position on the XY-plane for each frame.

Model Design. Our framework includes four transformer-based denoisers: one for text-to-motion (T2M), one for speech-to-gesture (S2G), one for trajectory-

to-velocity (T2V), and one for speech-to-details (S2D) that synthesizes facial expressions and hand poses:

$$\hat{b}_0 = G_{\text{T2M}}(b_t, t, c_{text}), \quad \hat{b}_0 = G_{\text{S2G}}(b_t, t, c_{audio}), \tag{2}$$

$$\hat{v}_0 = G_{\text{T2V}}(v_t, t, c_{traj}), \quad \hat{d}_0 = G_{\text{S2D}}(d_t, t, c_{audio}). \tag{3}$$

We denote the sampling with classifier-free guidance for each denoiser as G_u^s, where $u \in \{\text{T2M, S2G, T2V, S2D}\}$.

For the T2M denoiser conditioned on a text embedding c_{text}, we follow prior works [14, 53] by treating c_{text} as a token and applying self-attention to incorporate the semantic information in c_{text} into the motion generation process. In contrast, the other denoisers are conditioned on sequential data; therefore, we utilize cross-attention to model the relationships between the input sequences and the generated motion. Additionally, the T2M and T2V denoisers are trained on HumanML3D [18], while the S2G and S2D denoisers are trained on the BEAT2 [36] dataset. All denoisers adhere to the objective function and diffusion paradigm described in Sec. 3.1. The computational complexity of each component is reported in Appendix F.

Fig. 3: Examples of synchronous and asynchronous conditions. Synchronous conditions occur when all condition signals are provided within the same time interval. In contrast, asynchronous conditions involve multiple conditions, each corresponding to different time intervals.

3.3 Multi-Modal Decoupled Denoising

In this section, we introduce our multi-modal decoupled denoising strategy for generating motion in scenarios where text and speech audio are provided synchronously—that is, within the same time interval—as shown in Fig. 3 (a).

Our key observation is that the speech audio naturally guides upper-body gestures (including head and arm poses), while text descriptions mainly influence lower-body movements (including spine and leg poses) like walking or shifting stance. In a diffusion context, this implies that the joint conditional probability $p(b_{t-1} \mid c_{\text{text}}, c_{\text{audio}}, b_t)$ can be approximated by the product of two probabilities: $p(b_{t-1,\text{lower}} \mid c_{\text{text}}, b_t)$ and $p(b_{t-1,\text{upper}} \mid c_{\text{audio}}, b_t)$. Mathematically, this can be represented as follows:

$$p(b_{t-1} \mid c_{\text{text}}, c_{\text{audio}}, b_t) \approx p(b_{t-1,\text{lower}} \mid c_{\text{text}}, b_t) \cdot p(b_{t-1,\text{upper}} \mid c_{\text{audio}}, b_t). \tag{4}$$

Here, b_t represents the motion at denoising step t, which is composed of the upper-body motion $b_{t,\text{upper}}$ and the lower-body motion $b_{t,\text{lower}}$. A detailed derivation of Eq. (4) is provided in Appendix B.

Based on this observation, we can decouple the multi-modal controlled denoising process into distinct streams. The text-driven stream $p(b_{t-1} \mid c_{\text{text}}, b_t)$ can be formulated as follows:

$$\hat{b}_0^{\text{text}} = G^s_{\text{T2M}}(b_t, t, c_{\text{text}}), \quad b_{t-1}^{\text{text}} = \sqrt{\alpha_{t-1}}\,\hat{b}_0^{\text{text}} + \sqrt{1-\alpha_{t-1}}\,\epsilon, \tag{5}$$

where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$ represents the cumulative product of $(1-\beta_s)$ up to timestep t, and β_s is the variance schedule controlling the amount of noise added at each timestep. The variable $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is Gaussian noise sampled from a standard normal distribution [53]. Similarly, the audio-driven stream $p(b_{t-1} \mid c_{\text{audio}}, b_t)$ is defined as:

$$\hat{b}_0^{\text{audio}} = G^s_{\text{S2G}}(b_t, t, c_{\text{audio}}), \quad b_{t-1}^{\text{audio}} = \sqrt{\alpha_{t-1}}\,\hat{b}_0^{\text{audio}} + \sqrt{1-\alpha_{t-1}}\,\epsilon. \tag{6}$$

Finally, the body motion b_{t-1} at step t is computed as follows:

$$b_{t-1} = I \odot b_{t-1}^{\text{text}} + (1-I) \odot b_{t-1}^{\text{audio}}, \tag{7}$$

where $I \in \mathbb{R}^{205}$ is a binary vector with entries set to 1 for the lower body and 0 for the upper body; $\odot$ denotes element-wise multiplication.

The decoupled denoising allows each body part's motion to be precisely guided by its corresponding input condition, ensuring high fidelity to the control signals. Furthermore, each stream adjusts its output based on the current overall motion, promoting coordination among the body parts. As this process continues, the motions generated for different body parts become increasingly synchronized, resulting in natural and coherent full-body movements.

3.4 Trajectory and Motion Planning

The above subsection handles the multi-modal decoupled denoising for synchronous conditions. We now extend our MOCO framework to tackle more complex, challenging yet more realistic scenarios, such as incorporating trajectory control and leveraging Large Language Model (LLM) for motion planning.

Trajectory Control. Following the approach of Petrovich *et al.* [43], we represent the global transition of body pose using the velocity vector $v = [\dot{r}_x, \dot{r}_y, \dot{\theta}]$, where $\dot{r}_x$ and $\dot{r}_y$ are the linear velocities of the pelvis in the x and y directions, respectively, and $\dot{\theta}$ is the angular velocity about the body's vertical (Z) axis. Given the trajectory data c_{traj}, we first predict $\hat{v}_0$ using Eq. (3). To enhance prediction accuracy, we incorporate loss guidance into our method. During each denoising step for predicting the velocity vector, we compute $\hat{v}_0$ using Eq. (3) and apply loss guidance as follows:

$$L_{\text{guidance}} = FK(\hat{v}_0) - c_{\text{traj}}, \tag{8}$$

where FK represents the differentiable Forward Kinematics function [58] that converts linear and angular velocities into the trajectory. We follow the methodology of InterControl [58] and optimize L_{guidance} with respect to $\hat{v}_0$ using the second-order LBFGS optimizer [35]. This optimization ensures that the predicted global transitions closely match the provided trajectory data.

Once $\hat{v}_0$ is predicted, we substitute the velocity component in $\hat{b}_0$ with $\hat{v}_0$ during each iteration of its generation. This substitution guides the generation process to adapt the remaining elements of $\hat{b}_0$ to align with $\hat{v}_0$, thereby ensuring consistency with the provided trajectory data.

Table 1: Example of an LLM-generated motion timeline.

Condition	Start Time	End Time	Body Part
c_1 = kneel down	f_1^s	f_1^e	spine, legs
c_2 = stand up from ground	f_2^s	f_2^e	spine, legs
c_3 = run forward	f_3^s	f_3^e	spine, legs
c_4 = audio clip 1	f_4^s	f_4^e	head, arms
c_5 = audio clip 2	f_5^s	f_5^e	head, arms

Motion Planning. In real-world applications, models could face audio conditions paired with complex text descriptions, which can challenge their ability to handle intricate semantics. One approach to address this challenge is decomposing complex textual descriptions into basic motion units and generating them separately [50]. However, this decomposition often introduces more complex conditions—such as asynchronous conditions, as illustrated in Fig. 3.

To address these challenges, we propose an LLM-based motion planning pipeline. Our pipeline begins by using an LLM to generate a "motion timeline" from the input text and audio conditions. Particularly, we carefully design a prompt that instructs the LLM to decompose complex text into elementary motion units, insert appropriate pose transitions, and specify the start and end times and the body parts involved for each condition. More details are provided in Appendix C. For example, when given the text condition "kneel down then walk forward" along with two audio clips, the LLM produces a motion timeline comprising multiple intervals, as detailed in Tab. 1.

Notably, the transition condition c_2 is automatically generated by the LLM, which ensures a smooth stance changes and avoids abrupt movements. Users, however, have the flexibility to customize the timeline. We then adopt the strategy of STMC [43], generating body masks according to a predefined timeline. At each denoising step t, the body pose across the entire timeline is generated as follows:

$$\hat{b}_0 = \sum_{j=1}^{J} I_j \odot G_j^s \left(b_{t, f_j^s : f_j^e}, t, c_j \right), \tag{9}$$

where I_j represents a binary mask corresponding to the motion generated by the j-th condition, and $G_j^s \in \{G_{\text{T2M}}^s, G_{\text{S2G}}^s\}$ denotes the denoiser utilized for the j-th condition. The operator $\odot$ stands for element-wise multiplication.

Similarly, the facial and hand movements at step t are computed by:

$$\hat{d}_0 = \sum_{j=1}^{J} G_{\text{S2D}} \left(d_{t, f_j^s : f_j^e}, t, c_j \right). \tag{10}$$

Since the text-to-motion dataset, HumanML3D, lacks facial and hand movement data, the denoiser G_{S2D} is only trained on the speech-to-gesture dataset. Therefore, in intervals without audio clips, we maintain the model's temporal continuity by setting the conditional input c_j to the empty condition $\varnothing$. Finally, we employ DiffCollage [68] to ensure smoother transitions at the interval boundaries.

4 Experiments

4.1 Datasets

Task-Specific Datasets. The HumanML3D dataset is a large **Text-to-Motion** dataset created by amalgamating motion sequences from the HumanAct12 and AMASS datasets [18]. It consists of 14,616 motions and 44,970 descriptions composed of 5,371 distinct words, totaling 28.59 hours of motion data. To align the data representation—specifically, to use SMPL-X parameters for representing joint rotations—we utilize only the AMASS portion of HumanML3D because it has an official SMPL-X version. The BEAT2 dataset is a large-scale **Speech-to-Gesture** dataset specifically designed for research in speech-to-gesture generation [36]. It contains synchronized recordings of speech audio and corresponding 3D motion capture data of human gestures, totaling 60 hours of data from 25 speakers. In addition to audio and motion data, the dataset includes valuable annotations such as text transcriptions and emotional states. Here, we select speakers with IDs lower than 10, resulting in a total of 10.15 hours of training data.

Multi-Modal Benchmark. We create a multi-modal benchmark of 1,000 test clips to evaluate our proposed task effectively. Each test clip is automatically constructed and contains two text descriptions and two audio clips. To create these clips, we manually collect 40 texts focusing on lower-body movements that commonly occur during speech delivery or conversation. We then split the audio from the BEAT2 test set into clips using a Voice Activity Detector (VAD). To serve as ground truth for computing evaluation metrics (Sec. 4.2), we selected motion samples from AMASS and BEAT2 corresponding to each text and audio clip. This structured construction enables controlled and repeatable comparisons under concurrent multi-modal conditions and isolates the challenge of resolving conflicting cross-modal signals. Accordingly, the benchmark is intended to probe compositional consistency rather than broad coverage of open-ended semantics. More details about the benchmark are provided in Appendix H.

Table 2: Comparison with baselines. We highlight the best performance in **bold** and use underlining to indicate the second-best performance.

	Text2Motion				Speech2Gesture		
	FID+ ↓	R1 ↑	M2T ↑	M2M ↑	FID-A ↓	BC ↑	L1div ↑
GT (Ground Truth)	0.000	40.0	0.781	1.000	-	-	-
Weighted Sum [61]	1.335	6.8	0.546	0.537	**2.17**	2.20	4.08
Pseudo-Text [34]	1.593	2.2	0.511	0.503	2.22	2.55	6.43
SynTalker [12]	0.985	9.8	0.601	0.603	6.60	**2.95**	**9.12**
MOCO	**0.862**	**24.6**	**0.649**	**0.639**	3.83	2.72	8.62

4.2 Metrics

We evaluate our method using two categories of metrics: text-to-motion and speech-to-gesture [18, 36, 43]. For text-to-motion, *FID+* assesses realism by measuring the distribution difference between real and generated motions using five random 5-second clips per test sample. *R1* evaluates alignment by recording the frequency of correct text prompts appearing in the top-1 retrieved text. *M2T* (motion-to-text) and *M2M* (motion-to-motion) measure alignment through cosine similarity between embeddings of generated motions and ground truth texts or motions. In the speech-to-gesture category, *FID-A* similarly measures the realism of motion generated based on speech audio. *Beat Consistency (BC)* evaluates how well gestures synchronize with the rhythm and beats of the speech, while *L1 Diversity (L1Div)* quantifies gesture diversity by calculating the average L1 distance between multiple gesture clips. This comprehensive set of metrics thoroughly evaluates our method across key dimensions.

4.3 Comparison with Baselines

Baseline Setting. To effectively evaluate the performance of our method, we propose a series of baselines inspired by reasonable settings and prior works. These include: *Weighted Sum*, a method that follows [61] by combining the predictions of text- and audio-conditioned models through weighted sum; *Pseudo-Text*, a method that follows [34] by using pseudo text descriptions of a speaker's speech as the text condition during training; and *SynTalker* [12], an open-source multi-modal motion generation method performing weighted sum in body-part-specific latent space. Notably, because these baselines cannot handle asynchronous conditions, we align the text time intervals with the audio time intervals to construct synchronous conditions for all baseline methods. The results are reported in Tab. 2.

Result Analysis. As shown in the table, our proposed method, MOCO, exhibits robust performance across both sets of metrics, delivering competitive results in both text-to-motion and speech-to-gesture tasks simultaneously. This underscores the effectiveness of MOCO in generating condition-aligned motions when multi-modal conditions are provided concurrently.

Weighted Sum and *Pseudo-Text* achieves good results on speech-to-gesture metrics but poor performance on text-to-motion metrics, indicating their limited ability to handle multi-modal data. We explain this further in Appendix E.

SynTalker also notices the importance of separating different body parts, resulting in an improvement in text-to-motion metrics compared to *Weighted Sum* and *Pseudo-Text*. However, its insufficient decoupling of conditions and suboptimal joint latent space—negatively impacted by repetitive patterns in the speech-to-gesture data—results in limited expressiveness and unstable movements. This is reflected in its less competitive performance in R1 and FID-A.

Beyond the baselines above, STMC [43] is particularly relevant, as its per-step composition paradigm inspired our multi-modal decoupled denoising. Since STMC is text-only and does not natively accept speech audio, we adapt it to our concurrent setting and conduct a dedicated comparison, including a user study (Appendix G.1 and Appendix G.5). There, STMC attains competitive text-to-motion scores, yet MOCO produces more natural and better-synchronized motion, as confirmed by quantitative metrics, additional naturalness metrics, and the subjective user study.

4.4 Ablation Study

To assess the impact of key designs within MOCO, we conduct an ablation study presented in Tab. 3. This study systematically examines the effects of body masking (*Body Mask*), weight sharing (*Share Weight*), and combination at each step (*Per-Step Comb.*) on the model's performance across text-to-motion and speech-to-gesture metrics. In addition, we introduce the *Transition Smoothness Ratio (TSR)* to measure temporal coherence at segment boundaries, defined as the ratio between the mean frame-to-frame speed within transition windows and the mean speed in non-transition regions.

Body Masking. In Variants 1 and 2, we test our hypothesis that speech audio guides upper-body motion (head and arms) while text descriptions influence lower-body movements (spine and legs). In Variant 1, we expand the body mask to include the spine along with the head and arms (*Body Mask* = head, arms, spine). This modification results in a worse FID+ and a slight decrease in R1, indicating a decline in text-to-motion performance. Moreover, it does not produce significant improvements in speech-to-gesture metrics, suggesting that including the spine in the body mask fails to enhance gesture generation and instead compromises text-driven motion performance.

Variant 2 further adjusts the body mask to include the legs and spine (*Body Mask* = legs, spine), which significantly degrades text-to-motion metrics and Beat Consistency. This mainly stems from a mismatch: the texts in our multimodal benchmark describe diverse actions such as "walk," "sit," and "turn right," whereas the speech-to-gesture data are mostly standing gestures. Controlling lower-body motion with audio makes it hard to align with these texts, while controlling upper-body motion with text complicates beat alignment. Although Variant 2 yields a notable improvement in FID-A—reflecting a bias in the speech-to-gesture data toward standing-in-place motions—its overall performance still degrades.

Table 3: Ablation study on key designs within MOCO. We highlight the best performance in **bold** and use underlining to indicate the second-best performance achieved by our method.

Method	Share Weight	Body Mask $1-I$	Per-Step Comb.	Text2Motion			Speech2Gesture			TSR ↓
				FID+ ↓	R1 ↑	M2T ↑	FID-A ↓	BC ↑	L1div ↑	
GT	-	-	-	0.000	40.0	0.781	-	-	-	-
MOCO	✗	head, arms	✓	0.862	**24.6**	0.649	3.83	2.72	8.62	0.90
Variant 1	✗	head, arms, spine	✓	0.921	22.4	0.634	3.86	**2.81**	8.35	0.90
Variant 2	✗	spine, legs	✓	1.234	7.9	0.554	**2.75**	2.19	5.11	0.95
Variant 3	✓	head, arms	✓	0.866	22.1	**0.656**	4.24	2.68	**8.95**	**0.85**
Variant 4	✗	head, arms	✗	**0.832**	24.4	0.645	3.99	2.27	8.60	1.17

In contrast, our original method (*Body Mask* = head, arms) effectively balances the influences of both text and audio inputs. By assigning the upper body to be guided by audio and the lower body by text, we achieve superior results across both text-to-motion and speech-to-gesture metrics. This demonstrates the advantage of our approach in producing coherent and contextually appropriate motions that align well with the provided conditions.

Weight Sharing. In Variant 3, we enable weight sharing (*Share Weight* = ✓), following previous multi-modal methods [12, 34, 61], while keeping the body mask and transition method unchanged. Compared to the full MOCO model (without weight sharing), enabling weight sharing results in worse performance across several metrics, including R1 and FID-A. This decline suggests that sharing weights between modalities may limit the model's ability to capture modality-specific nuances, thereby reducing its effectiveness in generating accurate and realistic motions for both text-to-motion and speech-to-gesture tasks.

Combination at Each Step. Variant 4 performs body-part combination only at the final denoising step $t = 0$ (*Per-Step Comp.* = ✗). Lacking iterative coordination to aggregate holistic motion information, this baseline exhibits temporal discontinuities (reflected by high TSR) and reduced body coherence. A user study (Sec. 4.6) and qualitative results in the **supplementary videos** further confirm that MOCO produces more natural and synchronized motions.

4.5 Qualitative Analysis

To clearly illustrate the overall performance of MOCO, we visualize four generated samples and their corresponding conditions in Fig. 4. The lighter color of the mesh and the text description background indicates the start of each sequence, while the darker color indicates the end. These results exhibit natural speech-driven upper-body gestures synchronized with diverse lower-body motions such as jogging, walking, and sitting, demonstrating MOCO's ability to generate coherent and realistic motions that closely align with the given multimodal signals. A detailed analysis of these examples is provided in the figure caption. Additional qualitative results, including comparisons with baselines, are included in the supplementary materials.

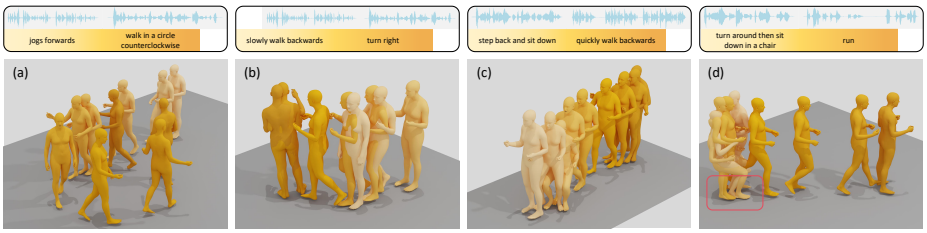

Fig. 4: Qualitative results. We visualize four samples generated by MOCO. Darker colors represent later points in time. The results demonstrate that MOCO is capable of generating coherent and realistic motions that highly align with the given multi-modal control signals. Figures (a), (b), and (d) present natural speech gestures coordinated with various lower-body movements as specified by the text inputs, such as jogging, walking in a circle, turning right, running, and so on. Figure (c) displays natural speech movements while sitting down. Figure (d) reveals a limitation of MOCO. When standing up or sitting down, the foot should remain stationary. However, the foot highlighted in the red box slides, leading to unrealistic results.

Table 4: User study results comparing MOCO against baseline methods.

Baseline	Better Text Following (%)			Better Beat Synchronization (%)		
	Neither	Baseline	MOCO (Ours)	Neither	Baseline	MOCO (Ours)
Pseudo-Text	1.0	0.0	**99.0**	13.0	12.5	**74.5**
Weighted Sum	0.0	0.0	**100.0**	13.5	3.5	**83.0**
Baseline	**Better Body Coherence (%)**			**Better Temporal Fluidity (%)**		
	Neither	Baseline	MOCO (Ours)	Neither	Baseline	MOCO (Ours)
Combine Once (Variant 4)	12.0	17.0	**71.0**	12.6	42.6	**44.8**

4.6 User Study

Tab. 4 presents the results of a user study comparing MOCO with three methods. Specifically, we evaluate MOCO against *Pseudo-Text* and *Weighted Sum*, introduced in Sec. 4.3, and *Combine Once* (Variant 4 in Sec. 4.4), to assess overall performance in text following and audio beat synchronization. Additional experimental details are provided in Appendix G.4.

As shown in the table, MOCO achieves significant advantages over both *Pseudo-Text* and *Weighted Sum*, demonstrating the effectiveness of our decoupled denoising strategy in generating motion aligned with multi-modal conditions. Furthermore, when compared to *Combine Once*, our method was rated significantly higher in both body coherence and temporal fluidity. This indicates that MOCO does more than merely combine different body parts controlled by separate conditions; it ensures that each body part aligns with its corresponding condition while enhancing coordination among all body parts. To further strengthen these findings, we additionally conduct a larger-scale pairwise preference study with 51 participants comparing MOCO against STMC, where MOCO is preferred with

statistical significance on audio synchronization and naturalness ($p < 0.05$); full protocol and results are reported in the supplementary material.

5 Discussion and Limitations

MOCO achieves effective multi-modal motion generation, but its fixed body-part assignment can be suboptimal when control signals conflict. For instance, when text descriptions specify arm movements (*e.g.*, "wave hands"), this rigid separation leaves the upper body solely governed by audio, preventing the model from following text instructions involving arm motions. To mitigate this limitation, MOCO allows users to configure body-part control through the motion timeline (see Appendix D), enabling text to drive upper-body motion when needed.

Our multi-modal benchmark is also deliberately focused on compositional consistency under controlled settings rather than open-ended semantic coverage, enabling precise and repeatable evaluation of cross-modal integration but not claiming generalization to arbitrary text–audio pairs. Future work may explore adaptive body-part assignment or residual blending that dynamically handles spatially overlapping actions, moving beyond fixed spatial rules toward context-aware fusion while keeping the interpretability of decoupled generation.

6 Conclusion

This study presents **MOCO**, a novel diffusion-based framework to generate realistic and coherent holistic body motions from multi-modal control signals, including text descriptions, speech audio, and trajectory data. Our key innovation lies in a decoupled denoising process where, during each denoising step, the model independently generates motions for each modality and assembles them according to predefined spatial rules. This approach ensures that the generated motion is closely aligned with each condition while producing realistic and coherent whole-body movements. Experimental results demonstrate that our approach delivers state-of-the-art performance both qualitatively and quantitatively, advancing the field of multi-modal controlled motion generation.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grants 62476099 and 62076101, Guangdong Basic and Applied Basic Research Foundation under Grants 2024B1515020082 and 2023A1515010007, and the TCL Young Scholars Program.

References

1. Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–20 (2023)
2. Ao, T., Gao, Q., Lou, Y., Chen, B., Liu, L.: Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. ACM Transactions on Graphics (TOG) **41**(6), 1–19 (2022)

3. Ao, T., Zhang, Z., Liu, L.: Gesturediffuclip: Gesture diffusion model with clip latents. ACM Transactions on Graphics (TOG) **42**(4), 1–18 (2023)
4. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Teach: Temporal action composition for 3d humans. In: 2022 International Conference on 3D Vision (3DV). pp. 414–423. IEEE (2022)
5. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Sinc: Spatial composition of 3d human motions for simultaneous action generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9984–9995 (2023)
6. Bae, J., Hwang, I., Lee, Y.Y., Guo, Z., Liu, J., Ben-Shabat, Y., Kim, Y.M., Kapadia, M.: Less is more: Improving motion diffusion models with sparse keyframes. In: International Conference on Computer Vision (ICCV) (2025)
7. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
8. Barquero, G., Escalera, S., Palmero, C.: Belfusion: Latent diffusion for behavior-driven human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2317–2327 (2023)
9. Barquero, G., Escalera, S., Palmero, C.: Seamless human motion composition with blended positional encodings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 457–469 (2024)
10. Cen, Z., Pi, H., Peng, S., Shen, Z., Yang, M., Zhu, S., Bao, H., Zhou, X.: Generating human motion in 3d scenes from text descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1855–1866 (2024)
11. Cen, Z., Pi, H., Peng, S., Shuai, Q., Shen, Y., Bao, H., Zhou, X., Hu, R.: Ready-to-react: Online reaction policy for two-character interaction generation. In: International Conference on Learning Representations (ICLR) (2025)
12. Chen, B., Li, Y., Ding, Y.X., Shao, T., Zhou, K.: Enabling synergistic full-body control in prompt-based co-speech motion generation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 6774–6783 (2024)
13. Chen, C., Zhang, J., Lakshmikanth, S.K., Fang, Y., Shao, R., Wetzstein, G., Fei-Fei, L., Adeli, E.: The language of motion: Unifying verbal and non-verbal language of 3d human motion. In: Computer Vision and Pattern Recognition (CVPR) (2025)
14. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023)
15. Cong, P., Wang, Z., Dou, Z., Ren, Y., Yin, W., Cheng, K., Sun, Y., Long, X., Zhu, X., Ma, Y.: Laserhuman: language-guided scene-aware human motion generation in free environment. arXiv preprint arXiv:2403.13307 (2024)
16. Dai, W., Chen, L.H., Wang, J., Liu, J., Dai, B., Tang, Y.: Motionlcm: Real-time controllable motion generation via latent consistency model. In: European Conference on Computer Vision. pp. 390–408. Springer (2024)
17. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3497–3506 (2019)
18. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022)
19. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)

20. Gupta, P., Verma, S., Grama, A., Bera, A.: Unified multi-modal interactive & reactive 3d motion generation via rectified flow. In: International Conference on Learning Representations (ICLR) (2026)
21. Harithas, S., Sridhar, S.: Motionglot: A multi-embodied motion generation model. In: International Conference on Learning Representations (ICLR) (2025)
22. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2282–2292 (2019)
23. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
24. Hong, S., Kim, C., Yoon, S., Nam, J., Cha, S., Noh, J.: Salad: Skeleton-aware latent diffusion for text-driven motion generation and editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7158–7168 (2025)
25. Hosseyni, S.R., Rahmani, A.A., Seyedmohammadi, S.J., Seyedin, S., Mohammadi, A.: Bad: Bidirectional auto-regressive diffusion for text-to-motion generation. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
26. Hu, L., Ye, Y., Xia, S.: Hmvlm: Human motion-vision-lanuage model via moe lora. arXiv preprint arXiv:2511.01463 (2025)
27. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16750–16761 (2023)
28. Jiang, Y., Liao, Q., Lu, Z.: Smoothsync: Dual-stream diffusion transformers for jitter-robust beat-synchronized gesture generation from quantized audio. arXiv preprint arXiv:2601.04236 (2026)
29. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2151–2162 (2023)
30. Lee, T., Baradel, F., Lucas, T., Lee, K.M., Rogez, G.: T2lm: Long-term 3d human motion generation from multiple sentences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1867–1876 (2024)
31. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13401–13412 (2021)
32. Li, Z., Yuan, W., He, Y., Qiu, L., Zhu, S., Gu, X., Shen, W., Dong, Y., Dong, Z., Yang, L.T.: Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. In: International Conference on Learning Representations (ICLR) (2025)
33. Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. International Journal of Computer Vision pp. 1–21 (2024)
34. Ling, Z., Han, B., Wong, Y., Kangkanhalli, M., Geng, W.: Mcm: Multi-condition motion synthesis framework for multi-scenario. arXiv preprint arXiv:2309.03031 (2023)
35. Liu, D.C., Nocedal, J.: On the limited memory bfgs method for large scale optimization. Mathematical programming **45**(1), 503–528 (1989)
36. Liu, H., Zhu, Z., Becherini, G., Peng, Y., Su, M., Zhou, Y., Iwamoto, N., Zheng, B., Black, M.J.: Emage: Towards unified holistic co-speech gesture generation via masked audio gesture modeling. arXiv preprint arXiv:2401.00374 (2023)

37. Liu, P., Song, L., Huang, J., Liu, H., Xu, C.: Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. In: International Conference on Computer Vision (ICCV) (2025)
38. Liu, Y., Cao, Q., Wen, Y., Jiang, H., Ding, C.: Towards variable and coordinated holistic co-speech motion generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1566–1576 (2024)
39. Liu, Y., Cao, Q., Yi, H., Jiang, H., Ding, C.: Multi-modal controlled coherent motion synthesis. OpenReview (2024), https://openreview.net/forum?id=i5Gxilzk0u
40. Liu, Y., Chen, C., Yi, L.: Interactive humanoid: Online full-body motion reaction synthesis with social affordance canonicalization and forecasting. ArXiv **abs/2312.08983** (2023), https://api.semanticscholar.org/CorpusID:266209846
41. Ma, S., Cao, Q., Zhang, J., Tao, D.: Contact-aware human motion generation from textual descriptions. arXiv preprint arXiv:2403.15709 (2024)
42. Peng, Y., Song, J.T., Jung, S., Liu, R., Liu, H., Chu, X., Liu, R., Wu, E., Koike, H., Kitani, K.: Dyadit: A multi-modal diffusion transformer for socially favorable dyadic gesture generation. arXiv preprint arXiv:2602.23165 (2026)
43. Petrovich, M., Litany, O., Iqbal, U., Black, M.J., Varol, G., Bin Peng, X., Rempe, D.: Multi-track timeline control for text-driven 3d human motion generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1911–1921 (2024)
44. Qi, X., Zhang, H., Wang, Y., Pan, J., Liu, C., Sun, M., Xue, W., Zhang, S., Han, S., Liu, Q., et al.: Cocogesture: Towards coherent co-speech 3d gesture generation in the wild. Information Fusion p. 103613 (2025)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
46. Sampieri, A., Palma, A., Spinelli, I., Galasso, F.: Length-aware motion synthesis via latent diffusion. In: European Conference on Computer Vision. pp. 107–124. Springer (2024)
47. Sha, X., Zhang, L., Mashita, T., Chiba, N., Uranishi, Y.: 3dgespolicy: Phoneme-aware holistic co-speech gesture generation based on action control. arXiv preprint arXiv:2601.18451 (2026)
48. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023)
49. Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C.C., Liu, Z.: Bailando: 3D dance generation by actor-critic GPT with choreographic memory. In: CVPR. pp. 11050–11059 (2022)
50. Sun, J., Zhang, Q., Duan, Y., Jiang, X., Cheng, C., Xu, R.: Prompt, plan, perform: Llm-based humanoid control via quantized imitation learning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 16236–16242. IEEE (2024)
51. Sun, M., Xu, C., Jiang, X., Liu, Y., Sun, B., Huang, R.: Beyond talking–generating holistic 3d human dyadic motion for communication. International Journal of Computer Vision **133**(5), 2910–2926 (2025)
52. Tevet, G., Raab, S., Cohan, S., Reda, D., Luo, Z., Peng, X.B., Bermano, A.H., van de Panne, M.: Closd: Closing the loop between simulation and diffusion for multi-task character control. In: International Conference on Learning Representations (ICLR) (2025)

53. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
54. Tseng, J., Castellon, R., Liu, K.: Edge: Editable dance generation from music. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 448–458 (2023)
55. Wang, Y., Wang, S., Zhang, J., Fan, K., Wu, J., Xue, Z., Liu, Y.: Timotion: Temporal and interactive framework for efficient human-human motion generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7169–7178 (2025)
56. Wang, Y., Li, M., Liu, J., Leng, Z., Li, F.W., Zhang, Z., Liang, X.: Fg-t2m++: Llms-augmented fine-grained text driven human motion generation. International Journal of Computer Vision pp. 1–17 (2025)
57. Wang, Z., Chen, Y., Jia, B., Li, P., Zhang, J., Zhang, J., Liu, T., Zhu, Y., Liang, W., Huang, S.: Move as you say interact as you can: Language-guided human motion generation with scene affordance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 433–444 (2024)
58. Wang, Z., Wang, J., Lin, D., Dai, B.: Intercontrol: Generate human motion interactions by controlling every joint. arXiv preprint arXiv:2311.15864 (2023)
59. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. arXiv preprint arXiv:2310.08580 (2023)
60. Xu, Z., Lin, Y., Han, H., Yang, S., Li, R., Zhang, Y., Li, X.: Mambatalk: Efficient holistic gesture synthesis with selective state space models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
61. Yang, S., Xu, Z., Xue, H., Cheng, Y., Huang, S., Gong, M., Wu, Z.: Freetalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7945–7949. IEEE (2024)
62. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 469–480 (2023)
63. Yi, H., Thies, J., Black, M.J., Peng, X.B., Rempe, D.: Generating human interaction motions in scenes with text control. arXiv preprint arXiv:2404.10685 (2024)
64. Yoon, Y., Cha, B., Lee, J.H., Jang, M., Lee, J., Kim, J., Lee, G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics (TOG) **39**(6), 1–16 (2020)
65. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. arXiv preprint arXiv:2208.15001 (2022)
66. Zhang, M., Jin, D., Gu, C., Hong, F., Cai, Z., Huang, J., Zhang, C., Guo, X., Yang, L., He, Y., et al.: Large motion model for unified multi-modal motion generation pp. 397–421 (2024)
67. Zhang, P., Liu, P., Kim, H., Garrido, P., Chaudhuri, B.: Kinmo: Kinematic-aware human motion understanding and generation. In: International Conference on Computer Vision (ICCV) (2025)
68. Zhang, Q., Song, J., Huang, X., Chen, Y., Liu, M.Y.: Diffcollage: Parallel generation of large content with diffusion models. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10188–10198. IEEE (2023)
69. Zhang, X., Li, J., Zhang, J., Dang, Z., Ren, J., Bo, L., Tu, Z.: Semtalk: Holistic co-speech motion generation with frame-level semantic emphasis. In: International Conference on Computer Vision (ICCV) (2025)

70. Zhang, Z., Wang, Y., Mao, W., Li, D., Zhao, R., Wu, B., Song, Z., Zhuang, B., Reid, I., Hartley, R.: Motion anything: Any to motion generation. arXiv preprint arXiv:2503.06955 (2025)
71. Zhou, Z., Wan, Y., Wang, B.: A unified framework for multimodal, multi-part human motion synthesis. ArXiv **abs/2311.16471** (2023), `https://api.semanticscholar.org/CorpusID:265466120`
72. Zhou, Z., Wang, B.: Ude: A unified driving engine for human motion generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5632–5641 (2023)
73. Zhu, L., Liu, X., Liu, X., Qian, R., Liu, Z., Yu, L.: Taming diffusion models for audio-driven co-speech gesture generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10544–10553 (2023)