Semantics-Aware Human Motion Generation from Audio Instructions

Zi-An Wang[1,2], Shihao Zou[2*], Shiyao Yu[2], Mingyuan Zhang[3], Chao Dong[2,4]
[1]University of Chinese Academy of Sciences
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3]Nanyang Technological University
[4]Shenzhen University of Advanced Technology
za.wang@siat.ac.cn, sh.zou@siat.ac.cn

Abstract

Recent advancements in interactive technologies, such as GPT-4o, have highlighted the growing prominence of using audio signals to encode semantics. Building on this trend, this paper explores a relatively new task: human motion generation from audio instructions, where audio signals are used as conditioning inputs to generate motions that align with the semantics of the audio. Compared to text-based user interactions, audio provides a more convenient and natural mode of communication in real-world scenarios. However, existing audio-conditioned methods primarily focus on aligning movements with the rhythm of music or speech, which often results in a weak connection between the semantics of the audio and the generated motions. In this work, we propose an end-to-end framework based on masked generative transformer, incorporating a memory-retrieval based attention module to address the challenges posed by sparse and lengthy audio signals. Additionally, we augment existing datasets by rephrasing textual descriptions into a more conversational, spoken language style, and synthesizing corresponding audio with various speaker identities. Experimental results demonstrate both the effectiveness and efficiency of our framework for generating human motion from audio instructions. Notably, audio exhibits a comparable ability to text in representing semantics, while offering more efficient and user-friendly interactions in practical applications.

Keywords: human motion generation, generative models, multi-modality learning.

1. Introduction

Language-guided human motion generation [47] has gained increasing attention due to its wide range of applications, such as in the metaverse, video games, and virtual reality. Many existing works focus on text-conditioned motion generation [34, 20, 39, 43] by leveraging large language pre-trained models [12, 35] to encode text semantics. However, text-based user interaction poses significant challenges in real-world applications, as it often requires users to articulate precise and detailed descriptions, which can be time-consuming and unintuitive. This complexity makes it less practical for dynamic or fast-paced environments where quick, natural, and seamless interaction is essential for user engagement.

In contrast, audio provides a more direct and natural channel for communication and semantic understanding, offering a more intuitive and seamless interaction experience. Recent advances in interactive technologies have further emphasized the potential of audio for encoding semantics. For example, GPT-4o [31] incorporates audio as an interaction modality, greatly enhancing user convenience and immersion. This underscores the unique advantage of audio in improving both the ease and depth of human-computer interaction. In the context of human motion generation, audio-conditioned tasks such as music-to-dance [8, 29, 11, 16] and speech-to-gesture [3, 2, 28, 46] have become prominent, where motions are generated to align with the rhythm of music or speech. However, these conditioning signals often lack explicit motion descriptions, leading to a weak connection between semantics and movement. As a result, this limitation reduces the controllability and precision of the generated motions.

Building on these observations, we introduce a relatively new task in this paper: human motion generation from audio instructions, where audio signals are directly used as conditions to generate motions that align with the semantics of audio instructions. A straightforward approach might involve using a speech recognition algorithm, such as [36], to first convert the audio into text and then apply an existing text-to-motion model, like [19], to generate the motion. However, such a cascaded approach often incurs high computational overhead and latency. Moreover, human brain typically interprets audio signals directly without convert-

*Corresponding author.

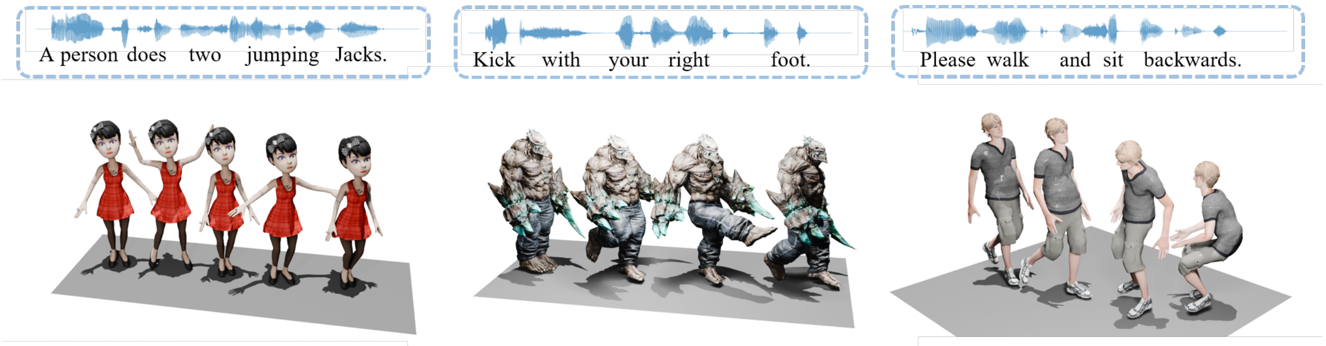


Figure 1. **Overview of Our Work**. Given an audio instruction as the conditional signal (with **text included for reference purposes only**), our generative model is able to produce high-quality human motion sequences that accurately align with the semantics of the audio input.

ing them into text, suggesting that a similar direct processing pathway should be pursued for artificial systems. Therefore, to achieve a more efficient and natural interaction, an approach—capable of generating motion directly from audio instructions—is essential for advancing toward artificial general intelligence.

In this paper, we propose an end-to-end framework based on the masked generative transformer. Specifically, to address the challenges of sparse and lengthy raw audio signals, we leverage WavLM [9] to encode the audio into meaningful features. To further improve efficiency, we introduce a memory-retrieval based attention module that compresses these audio features into more compact, model-friendly representations. These compact features will serve as conditioning inputs for generating corresponding motions. Our generative framework comprises three key components, inspired by [6, 19]: Residual VQ (RVQ), Masked Transformer, and Residual Transformer. RVQ performs multi-layer residual quantization on motion sequences, converting them into multi-level latent motion tokens. The Masked Transformer, conditioned on the encoded audio features, leverages the masked generative paradigm [7] to generate the base-level quantization codes for these tokens. Then the Residual Transformer refines the motion representation by iteratively generating residual-level quantization codes, layer by layer, progressively adding more detailed information. Finally, the base and residual codes are combined and passed through the RVQ motion decoder, which reconstructs the complete motion sequence. This multi-step process ensures that both the broad structure and finer details of the motion are captured, resulting in more accurate and realistic motion generation.

Another challenge is the lack of datasets that pair audio instructions with corresponding human motions. Existing large-scale datasets, such as KIT [34] and HumanML3D [20], only provide paired text and motion data, denoted as *original datasets*. Moreover, the style of the textual descriptions in these datasets does not reflect the natural expression of oral instructions. To address these limitations, we develop a custom prompt for ChatGPT [31] to rephrase text descriptions in [34, 20] into a more conversational, oral language style. We then use the speech synthesis tool Tortoise [4] to synthesize corresponding audios with various speaker identities. This process allows us to create augmented datasets, denoted as *oral datasets*, tailored for the task of human motion generation from audio instructions.

We evaluate our proposed method on both original and oral datasets, comparing it to the latest text-based approaches using a variety of metrics. The results show that our audio-based method performs competitively with text-based methods, demonstrating the effectiveness of audio signals for semantic understanding in practical applications. Furthermore, our end-to-end generative framework is over 50% faster than the cascaded approach, highlighting the importance of directly encoding audio semantics for a more efficient, user-friendly system. Additionally, models trained on the oral dataset significantly outperform those trained on the original dataset in audio-instructed motion generation, showcasing the value of our newly introduced dataset.

In summary, the primary contributions of this paper are as follows:

- We introduce a relatively new task in this paper: human motion generation from audio instructions for a more convenient user interaction system. To achieve this, we design an end-to-end framework based on the masked generative transformer, incorporating a memory-retrieval based attention module to handle the challenges posed by sparse and lengthy audio signals.
- We also augment existing text-based motion generation datasets for this task by rephrasing textual descriptions into a more conversational, spoken language style, and synthesizing corresponding audio with various speaker identities.
- The experimental results demonstrate the effectiveness and efficiency of our proposed end-to-end framework

for generating human motion directly from audio instructions. Notably, the findings indicate that audio offers a comparable ability to text in representing semantics, making it a strong alternative for conditioning signals in developing more efficient and user-friendly systems.

2. Related Work

Audio-based Motion Generation has seen significant progress, with tasks such as music-to-dance and speech-to-gesture gaining popularity. The music-to-dance task aims to generate dance sequences synchronized with musical beats and styles. Tang *et al*. [37] framed this problem as a sequence-to-sequence task, employing an LSTM [18] autoencoder to map music features to dance motions. Lee *et al*. [27] adopted a different approach, using convolutional neural networks (CNNs) to tackle the problem, while ChoreoMaster [8] introduced motion graphs to provide a more structured and flexible method for dance motion generation. In contrast, speech-to-gesture tasks aim to generate gestures that align with the rhythm and semantics of speech. Ginosar *et al*. [14] used pseudo-labeled data from a motion detection system to train their model, enabling the generation of personalized gesture styles. Aud2Repr2Pose [26] developed a joint motion autoencoder and speech encoder, allowing for a more direct mapping between speech features and gesture motion. Audio2Gestures [28] advanced this further by splitting the latent representation into shared and motion-specific components, using carefully designed loss functions to address the one-to-many mapping problem between speech and gestures. Recent works [3, 2, 46] have leveraged emerging generative models, such as diffusion models [22] and VQ-VAE [40], to produce rhythm-aware and semantics-aware co-speech gestures and also apply multi-modal data for stylistic variations. Despite these advances, both music and speech signals inherently lack explicit descriptions of the associated human motions [47], which limits the interpretability and precision of the generated results. To address this limitation, we propose a novel task centered on generating human motion from explicit audio instructions. By leveraging the richness of verbal descriptions, this approach allows for greater control and precision in the generation process, offering a more intuitive and convenient way of creating human motion.

Text-based Motion Generation focuses on synthesizing human motion sequences that correspond to textual descriptions. Early approaches in this field predominantly used deterministic models to directly map textual inputs to motion outputs. For instance, JL2P [1] introduced a joint embedding space that aligned motion and text data, ensuring consistency between the two modalities through a reconstruction task. Similarly, Ghosh *et al*. [13] employed a Gated Recurrent Unit (GRU)-based model to capture fine-grained motion details, integrating a discriminator to enhance the realism of generated motions. Angela *et al*. [30] added a trajectory prediction module, improving both complexity and accuracy in motion generation. More recent work has leveraged large pre-trained models to boost performance. MotionCLIP [38], for example, used the language-image pre-trained model CLIP [35], tapping into its prior knowledge to improve zero-shot motion generation capabilities. However, these deterministic methods have been limited in producing diverse outputs, as they tend to generate a single plausible motion sequence for each input. To overcome this limitation, stochastic modeling techniques have been introduced to account for the inherently many-to-many nature of the text-to-motion task. TEMOS [33], for instance, replaced traditional deterministic autoencoders with a Variational Autoencoder (VAE), enabling the generation of diverse motion sequences. Text2Action employed a Generative Adversarial Networks (GANs) framework to produce a variety of motions based on textual descriptions. Guo *et al*. [20] introduced a length predictor alongside a VAE-based generator, allowing the generation of motions with varying lengths that still align with the input descriptions. Building on the success of stochastic models in text-to-image tasks, techniques such as VQ-VAE [40] and diffusion models [22] have been applied to human motion generation. T2M-GPT [43] used VQ-VAE to learn discrete latent representations of motion sequences, combining this with a GPT-like autoregressive model to predict subsequent motion tokens. Diffusion models, known for their ability to iteratively refine outputs, have also gained attraction. FLAME [24] was the first model to apply diffusion techniques to human motion generation, using free-form language descriptions as input. MDM [39] further advanced this approach by employing a transformer-based diffusion model that predicts the final motion sequence through iterative denoising steps, producing more precise and realistic motions. MoMask [19] further improved efficiency by using a RVQ model for motion representation and masked generative transformers, allowing for faster and more flexible motion generation. However, text-based interaction proves to be inefficient in many practical scenarios, as it demands users to provide exact and comprehensive input, which can be cumbersome and slow. This limitation reduces its suitability for fast-paced applications where intuitive and rapid communication is critical for maintaining a smooth user experience.

Large Audio Model has experienced significant advancements, reshaping audio processing across various applications. WavLM [9] employs a self-supervised masked speech denoising framework to learn universal speech representations from extensive unlabeled datasets, demonstrating adaptability across diverse speech processing tasks. Similarly, Encodec [5] utilizes RVQ for high-fidelity neural audio compression, effectively modeling original audio and enhancing detail recovery in both speech and music through a combination of reconstruction loss and discriminator mechanisms. In the realm of speech recognition, Whisper [36] excels by extracting features from the Log-Mel Spectrogram of audio using an encoder-decoder framework, achieving remarkable multilingual performance supported by extensive training data. Furthermore, ImageBind [15] innovatively addresses multi-modal alignment by leveraging image-paired data to create a unified representation space for audio, images, and videos, thereby enhancing retrieval and classification tasks in zero-shot and few-shot scenarios. Collectively, these developments underscore the dynamic evolution of audio encoding techniques, enabling higher fidelity, efficiency, and versatility in managing audio data.

Generative Models have become a central focus in recent years, offering powerful methods to model complex data distributions. Goodfellow *et al*. [17] introduced GANs, which utilize a generator-discriminator framework where the two networks compete, enabling the generation of highly realistic data. In parallel, Kingma and Welling's Variational Autoencoders (VAEs) [25] introduced a probabilistic latent space to traditional autoencoders, facilitating the creation of diverse samples by sampling latent variables. Flow-based models[32], on the other hand, leverage invertible transformations to map simple distributions into complex data, enabling efficient sampling with exact likelihood estimation. Diffusion models [22] approach data generation through an iterative process of denoising Gaussian noise, excelling particularly in tasks like image generation due to their high-quality outputs. Building on the success of these models, masked generative transformers, such as MaskGIT [7], have emerged, employing an iterative token generation method that balances efficiency with the gradual refinement of generated content. Collectively, these approaches represent significant strides in generative modeling, each bringing unique strengths to different types of data and tasks.

3. Our Approach

In this paper, we propose an end-to-end framework using a masked generative transformer for human motion generation from audio instructions. The process begins by extracting audio features with the pre-trained WavLM model [9]. These features are then compressed into more compact, model-efficient representations using a memory-retrieval based attention module, as depicted in Fig. 3. Conditioned on these encoded audio features, we train a Masked Transformer to generate the base-level quantization codes for latent motion tokens. This is followed by a Residual Transformer that iteratively generates residual-level quantization codes layer by layer. Finally, the base and residual

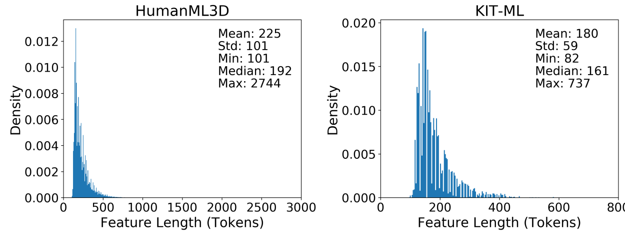


Figure 2. **Distribution of Audio Feature Lengths.** WavLM [9] is utilized to extract audio features from the augmented Oral Datasets derived from HumanML3D [20] and KIT [34]. A statistical analysis of these feature lengths reveals significant variability, with some features exhibiting notably long lengths. This variability presents challenges for processing conditional signals, as it complicates the integration of audio data into subsequent stages of our framework.

codes are combined and decoded by the RVQ motion decoder to reconstruct the full motion sequence, as shown in Fig. 4. This pipeline ensures efficient and accurate motion generation by progressively capturing both high-level structure and detailed nuances.

3.1. Audio Features Extraction

To effectively generate human motion based on audio instructions, we employ a speech vectorization module that converts raw audio signals into suitable feature representations. This module transforms the input audio into vector sequences, ensuring consistent dimensionality for each vector, while allowing the sequence length to vary depending on the input duration. This approach standardizes the feature dimensions across different audio samples, making them compatible with downstream tasks, despite variations in sequence length.

However, the vectorization process for audio signals cannot be directly applied to text due to significant differences in the inherent characteristics of these modalities. Audio signals, as naturally collected data, typically have lower information density, represent lower-level semantics, and consist of much longer sequences of sampled vibration points. In contrast, text is a more abstract representation with higher information density and a clearer, more explicit semantic structure. Additionally, audio data typically involves smaller volumes compared to text, largely due to the complexity and challenges in collection and processing, which require more spatial resources to convey equivalent semantic content. Therefore, selecting an appropriate audio vectorization scheme is critical for effective feature extraction in audio-to-motion generation, ensuring that the extracted features are semantically rich and fully utilized for downstream motion synthesis.

To advance our approach, it is important to review prior work in audio processing to uncover key paradigms and insights that can inform our method. Several pre-trained au-

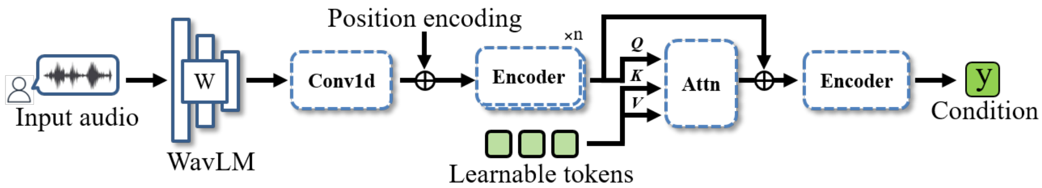


Figure 3. **Audio Conditions Processing Pipeline.** The audio features extracted by WavLM [9] are processed through a memory-retrieval based module, which standardizes the varying lengths of the input audio conditions. This module ensures that all audio signals are converted into a consistent length, facilitating smooth integration with subsequent components in the pipeline.

dio encoders, such as Encodec [5], Whisper [36], and ImageBind [15], have been developed, each tailored to specific tasks like speech synthesis, video retrieval, or speech recognition. However, these task-specific designs often introduce biases into the latent representation of audio signals, limiting their generalizability. In contrast, WavLM [9], a deep learning model developed by Microsoft and pre-trained on over 90,000 hours of human speech data, has demonstrated strong generalization ability across a wide range of downstream tasks. This versatility makes WavLM particularly suitable for the requirements of this study, providing robust and unbiased feature representations for our audio-based tasks.

Consequently, the audio vectorization module in our project is built upon the WavLM pre-trained model. Specifically, the output from the final layer of WavLM is used as the vectorized representation of the audio. For a segment of monaural audio sampled at 16 kHz, the resulting vector sequence length corresponds to the audio duration, while the vector dimensions remain constant. This approach ensures efficient and accurate extraction of audio features necessary for further processing.

3.2. Memory-Retrieval Based Audio Encoding

The distribution of audio feature lengths across samples in the datasets [20, 34] is shown in Fig. 2. Audio features extracted using WavLM vary significantly in sequence length, with some sequences being notably long. Directly feeding these lengthy features into the generative model as conditioning signals could lead to high computational complexity, while also impeding model convergence due to the overwhelming size of the input. To mitigate these challenges, it is crucial to encode the audio features into more compact representations that not only reduce computational burden but also extract rich motion-related semantic information for effective downstream processing.

A straightforward approach is to adopt the text encoding strategy used by CLIP [35], which appends a special token to each sequence of audio features and employs a Transformer Encoder [41] to aggregate the sequence's features through a self-attention mechanism. Given the significant variation in sequence lengths across different samples, this special token can capture information from any token within the sequence, regardless of its distance, allowing it to effectively model varying-range dependencies. Additionally, a CNN can be applied for downsampling before the features are passed into the Transformer Encoder. This process not only reduces sequence length but also compresses the sparse audio representation into a more compact and semantically enriched form, enhancing both efficiency and feature quality.

While this approach offers greater computational efficiency, it yielded suboptimal performance in our experiments. We attribute this to the fact that WavLM's pretraining tasks were not specifically tailored for motion generation, and relying on a single token to represent the rich audio features might create a bottleneck. As a result, the audio features extracted still contain information that is weakly related to motion description, such as speaker identity. This irrelevant information introduces noise, complicating the extraction of motion-relevant semantics. To address this limitation, we draw inspiration from [23] and design a memory-retrieval based module that transforms the audio features into representations more closely aligned with motion control. Specifically, this module leverages a set of learnable tokens stored in memory, which are mapped to key-value pairs in the attention mechanism, denoted as K_m and V_m. For each input x, a fully connected network generates a query vector Q_x, and the attention mechanism computes a weighted sum of the learnable tokens to produce a refined feature representation. After positional encodings are added, this refined feature is passed to the next layer of the encoder, allowing for more targeted and relevant semantic extraction that improves motion generation quality. This memory-retrieval process can be described as follows:

$$\text{Attention}(Q_x, K_m, V_m) = \text{softmax}\left(\frac{Q_x K_m^\top}{\sqrt{d}}\right) V_m, \quad (1)$$

where d stands for the dimension of the key and value vectors. Through memory-retrieval based audio encoding, we compress variable-length audio features into a compact, fixed-size condition signal, denoted as $y \in \mathbb{R}^{d_y}$, where d_y denotes the dimension of the condition signal. This compact representation serves as the conditioning input for human motion generation, ensuring consistent and efficient use of audio data regardless of its original length.

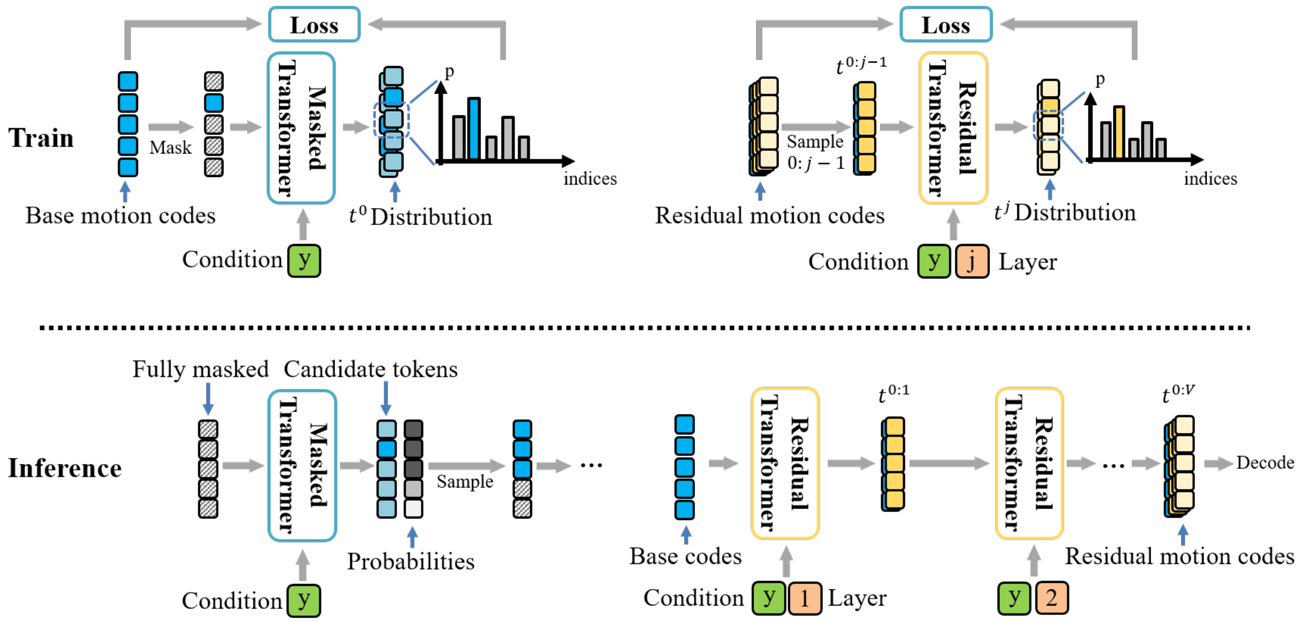


Figure 4. **Overview of Our Generative Framework in Training and Inference.** The framework consists of two key components: The **Masked Transformer** is designed to model the relationship between audio conditions and the base motion codes, which capture the principal components of the motion. The **Residual Transformer** establishes the connection between the audio conditions and the residual motion codes, which represent the finer, detailed aspects of the motion. During inference, these transformers work in sequential stages, progressively generating multi-layer latent motion codes that are then decoded to reconstruct the full motion sequence.

3.3. Masked Generative Model

Latent Motion Representation is handled by the RVQ-VAE framework [5, 19], which quantizes motion sequences into multi-layer token sequences within a discrete latent space, referred to as motion codes. Specifically, a motion sequence $m_{1:N} \in \mathbb{R}^{N\times D}$ is encoded into a latent representation $\tilde{c}_{1:n} \in \mathbb{R}^{n\times d}$, where n/N denotes the downsampling ratio and d denotes the latent dimension. Starting from $r^0 = \tilde{c}_{1:n}$, the latent representation is then decomposed into an ordered $V+1$ layers of motion codes $[c^j_{1:n}]^V_{j=0}$ through the following process:

$$c^j = \mathrm{Q}\left(r^j, C^j\right), \quad r^{j+1} = r^j - c^j, \quad j \in [0, V]. \quad (2)$$

Here, $\mathrm{Q}(\cdot)$ represents the standard quantization operation which finds the closet code c^j in the codebook C^j as the quantization of the input r^j. The codebook $C^j \in \mathbb{R}^{k\times d}$ consists of k codes with each code of dimension d. Given the multi-layer motion codes as the latent motion representation, they are summed together as the input to the motion decoder to recover the input motion, denoted as $\hat{m}_{1:N}$. The overall loss function for the RVQ-VAE consists of two terms: the motion reconstruction loss and the latent embedding loss for each layer:

$$\mathcal{L}_{rvq} = ||m - \hat{m}||_1 + \beta \sum_{j=1}^{V} \|r^j - \mathrm{sg}[c^j]\|_2^2, \quad (3)$$

where $\mathrm{sg}[\cdot]$ is the stop-gradient operation and β is a weighting factor that regulates the latent codes.

After quantizing the motion sequence using RVQ-VAE, we derive multi-layer motion codes represented as $[c^j_{1:n}, t^j_{1:n}]^V_{j=0}$, where $t^j_{1:n}$ denotes the codebook indices of the quantization code $c^j_{1:n}$. Build upon the latent motion representation, our generative model is trained to construct the relationship between audio conditions and base motion code (layer 0) or residual motion codes (layers 1 to V). The overall framework of the model is illustrated in Fig. 4, showcasing the training and inference (generation) process.

Masked Transformer, inspired by MaskGIT [7], is designed to capture the relationship between the audio conditions and the base motion codes $t^0_{1:n} \in \mathbb{R}^n$ ($j = 0$). During training, a random fraction of sequence elements is masked, replacing the corresponding tokens with a special [MASK] token. Letting $\tilde{t}^0$ represent the masked tokens, the model's task is to predict the masked tokens given the condition y and the partially masked tokens $\tilde{t}^0$. Mathematically, the masked transformer p_θ is trained by minimizing the negative log-likelihood of the target predictions:

$$\mathcal{L}_{mask} = \sum_{\tilde{t}^0_k = [\mathrm{MASK}]} -\log p_\theta(t^0_k | \tilde{t}^0, y). \quad (4)$$

Residual Transformer is designed to establish the relationship between the audio conditions and the residual

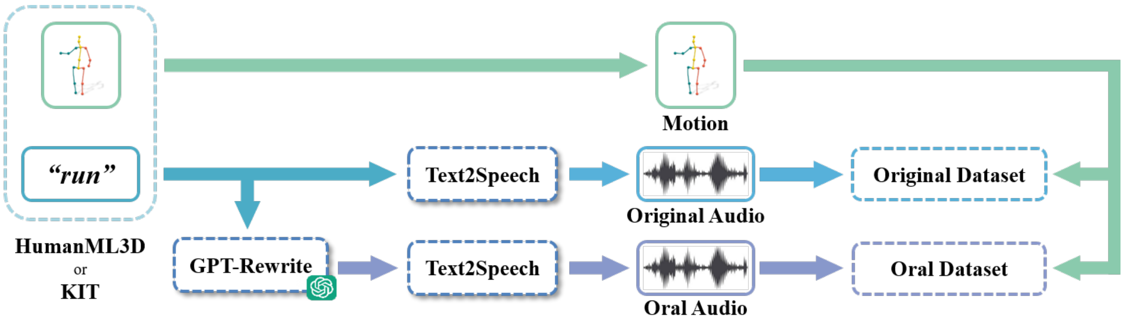


Figure 5. **Overview of the Audio-Motion Dataset Augmentation Process.** The texts from existing datasets, KIT [34] and HumanML3D [20], are fed into the text2speech model, Tortoise [4], generating audio signals with random speaker identities to create the **Original Dataset**. Additionally, the large language model ChatGPT-3.5 [31] is employed to rewrite the original texts in a more conversational, spoken language style. These rewritten texts are then used to generate corresponding audio signals, forming the **Oral Dataset**.

motion codes $t_{1:n}^j \in \mathbb{R}^n$ for $j \in [1, V]$. During training, a quantization layer j is randomly selected. The motion tokens from preceding layers, $t^{0:j-1}$, are collected and summed to create the preceding motion representation as input. Along with the audio conditions y and an indicator embedding for layer j, the residual transformer p_ϕ is trained to predict the tokens for the j-th layer. The training objective for the residual transformer is defined as follows:

$$\mathcal{L}_{res} = \sum_{j=1}^{V} \sum_{i=1}^{n} -\log p_\phi(t_i^j | t_i^{1:j-1}, y, j) \tag{5}$$

Motion Generation is carried out in two stages, as illustrated in Fig. 4. In the first stage, the goal is to iteratively sample the base motion codes over L iterations. The process begins with fully masked codes, which are fed into the masked transformer. The transformer generates a probability distribution over the base-layer codebook C^0, from which candidate motion codes are sampled. At each iteration l, the $\lceil \gamma(\frac{l}{L})n \rceil$ tokens with the lowest probabilities are re-masked with the $[\text{MASK}]$ token and regenerated in subsequent iterations. This process continues until all base motion codes are generated by the end of L iterations. The masking ratio $\gamma(\cdot)$ is dynamically adjusted using a cosine scheduling strategy:

$$\gamma(x) = \cos\left(\frac{\pi x}{2}\right) \in [0, 1] \tag{6}$$

Once the base motion codes are generated, the second stage involves progressively generating the residual motion codes using the residual transformer. Over V iterations, the j-th iteration takes all previously generated motion codes $t^{1:j-1}$ and the layer number j as inputs to predict the residual motion codes t^j for layer j. This process repeats for each subsequent layer, ultimately producing motion codes across all $V + 1$ layers. These multi-layer motion codes are then passed into the motion RVQ-VAE decoder, which reconstructs the final generated motion.

During inference, Classifier-Free Guidance (CFG) is applied at the final linear projection layer before the softmax operation. Here, the final logits w_g are computed by adjusting the conditional logits w_c in relation to the unconditional logits w_u, using a guidance scale s. Specifically, this is done through the formula $w_g = (1+s) \cdot w_c - s \cdot w_u$, which moves the conditional logits further from the unconditional ones.

4. Dataset Augmentation

The dataset augmentation process is illustrated in Fig. 5. First, we apply a text-to-audio synthesis algorithm to transform traditional text-to-motion datasets (see Sec. 4.1) into what we refer to as the **Original Dataset**. Next, we create an oral version of the dataset by rewriting the text descriptions into a more conversational style using ChatGPT [31] (see Sec. 4.2), followed by synthesizing corresponding audio files. This results in the formation of the **Oral Dataset**, which better reflects natural spoken language patterns.

4.1. Original Dataset

For effective human motion generation based on instructive audio signals, an appropriate audio-to-motion dataset is necessary. The most commonly used datasets for motion generation under explicit semantic constraints—such as KIT [34] and HumanML3D [20]—are text-based. The KIT dataset contains 3,911 action sequences paired with 6,278 text descriptions, while HumanML3D, on a larger scale, includes 14,616 action sequences covering a wide variety of behaviors such as dancing and exercising, paired with 44,970 descriptive texts. Despite the wealth of motion data, both datasets rely on textual conditional signals, which makes them unsuitable for tasks in instructive audio-based motion generation. To address this gap, we leverage existing text-to-speech algorithms to synthesize an audio-motion paired dataset from the original text-motion datasets, providing a foundation for subsequent work in audio-conditioned motion generation.

Recent speech synthesis technologies have also seen significant progress. While deep learning based speech synthesis still falls under the category of traditional statistical parametric synthesis, the powerful fitting capabilities of modern neural networks enable highly natural speech generation. As a result, most mainstream text-to-speech synthesis algorithms are now deep learning-based, trained on large-scale text-to-speech datasets to achieve synthesis that closely mirrors real human voices. Among these models, Tortoise [4] stands out. Based on a denoising diffusion probabilistic model [22], which has found significant success in image generation, Tortoise sacrifices some efficiency due to its iterative process in favor of producing highly realistic speech.

For a given text input, Tortoise synthesizes audio that not only matches human voices in timbre and quality but also captures natural features like pauses, emphasis, and tonal variation, reflecting the semantic intricacies of the text. This attention to detail enhances the naturalness of the generated speech, making it suitable for practical applications. Additionally, Tortoise's probabilistic generation framework allows for the synthesis of voices from multiple speakers, enriching its outputs and making it adaptable to diverse scenarios. Since Tortoise does not support specifying speakers directly, we generate data using randomly selected voices, adding variability to the dataset.

To support downstream audio processing algorithms, all synthesized audio is saved in mono WAV format with a fixed sampling rate of 16 kHz. The synthesis process is accelerated using GPUs to expedite the generation of large-scale data. The steps involved in this data synthesis process are illustrated in the accompanying figure.

4.2. Oral Dataset

Considering the practical usage scenarios for this task, users' phrasing and syntactic patterns during voice input often differ significantly from those used in text input. To better align the dataset with real-world applications, this study aims to modify the descriptive texts to reflect a more conversational tone, thereby reducing the gap between the dataset's written descriptions and potential spoken language input. Among the large language models available, OpenAI's ChatGPT [31] has demonstrated exceptional performance in English natural language understanding and instruction following. Therefore, we utilize ChatGPT 3.5 to rewrite the motion descriptions.

By crafting prompts that guide the model to maintain alignment with the original content, we ensure that the generated text retains the core meaning while incorporating more varied sentence structures and vocabulary, better reflecting conversational habits. Tab. 1 displays some examples of motion descriptions rewritten using ChatGPT, showing greater syntactic and lexical diversity without altering the original meaning. Based on these revised descriptions, the Tortoise model is then applied to generate the corresponding audio. Ultimately, this process yields 12,696 oral audio-motion pairs for the KIT-ML dataset and 87,384 for HumanML3D.

Base Datasets	**Oral Datasets**
a person running in circles	(1) Run in circles. (2) Move around in circular motions while running. (3) Jog in circular patterns.
a person kicks in the air, runs forward and then trots backward to their original location	(1) I want you to kick in the air, move forward, and then trot back to your starting point. (2) Kick in the air, run forward, and trot backward to where you started. (3) Can you kick in the air, run ahead, and then trot back to your original spot?
a man squats extraordinarily low then bolts up in an unsatisfactory jump.	(1) Squat down very low and then jump up in an unsatisfactory manner. (2) Go into an extremely low squat, then jump up in a disappointing leap. (3) Squat down remarkably low and then spring up in an unsatisfactory jump.

Table 1. **Examples of Rewritten Descriptions.** The original text descriptions from the Base Datasets were rewritten into a more conversational style, forming the augmented Oral Datasets. This transformation aims to better align the dataset with natural spoken language, enhancing its applicability to real-world voice-based interaction scenarios.

5. Experiment

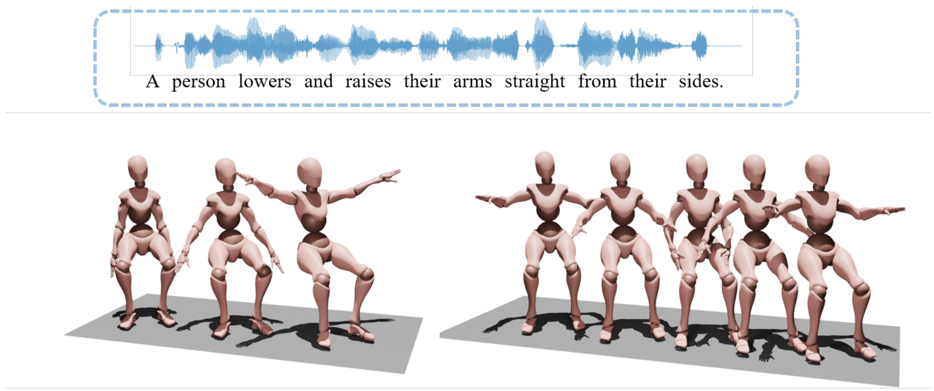
Empirical evaluations will be conducted using our augmented Oral Datasets and Base Datasets from HumanML3D [20] and KIT [34], with pose representations adopted from T2M [20]. These datasets serve as the foundation for assessing the performance of our model in generating motion sequences from audio instructions.

In our study, we adopt a set of evaluation metrics derived from both T2M [20] and MotionDiffuse [44] to ensure a comprehensive quantitative assessment of our results. These metrics collectively provide insights into the generation quality, semantic coherence, and diversity of our output, including:

- **Frechet Inception Distance (FID)**: Measures the quality of generated motions by comparing the distributional differences between the high-level features of the generated and real motion sequences.

Datasets	Method	Training Data	R Precision↑			FID↓	MM Dist↓	MultiModality↑
			Top 1	Top 2	Top 3			
Human ML3D	Momask [19] (Text-Based)	Original	$0.400^{\pm.002}$	$0.583^{\pm.003}$	$0.689^{\pm.002}$	$0.251^{\pm.009}$	$3.808^{\pm.009}$	$\mathbf{1.407}^{\pm.049}$
		Oral	$\mathbf{0.431}^{\pm.002}$	$\mathbf{0.626}^{\pm.002}$	$\mathbf{0.735}^{\pm.002}$	$\mathbf{0.091}^{\pm.003}$	$\mathbf{3.551}^{\pm.006}$	$1.284^{\pm.042}$
	Ours (Audio-Based)	Original	$0.417^{\pm.004}$	$0.603^{\pm.003}$	$0.708^{\pm.002}$	$0.202^{\pm.011}$	$3.697^{\pm.018}$	$\mathbf{1.272}^{\pm.041}$
		Oral	$\mathbf{0.426}^{\pm.007}$	$\mathbf{0.619}^{\pm.009}$	$\mathbf{0.727}^{\pm.009}$	$\mathbf{0.126}^{\pm.005}$	$\mathbf{3.577}^{\pm.037}$	$1.222^{\pm.036}$
KIT-ML	Momask [19] (Text-Based)	Original	$0.309^{\pm.008}$	$0.513^{\pm.009}$	$0.652^{\pm.009}$	$\mathbf{0.258}^{\pm.017}$	$4.017^{\pm.064}$	$\mathbf{1.562}^{\pm.064}$
		Oral	$\mathbf{0.330}^{\pm.006}$	$\mathbf{0.536}^{\pm.007}$	$\mathbf{0.667}^{\pm.008}$	$0.261^{\pm.014}$	$\mathbf{3.840}^{\pm.033}$	$1.166^{\pm.037}$
	Ours (Audio-Based)	Original	$0.314^{\pm.008}$	$0.508^{\pm.012}$	$0.636^{\pm.017}$	$0.487^{\pm.046}$	$4.186^{\pm.140}$	$\mathbf{1.399}^{\pm.052}$
		Oral	$\mathbf{0.324}^{\pm.005}$	$\mathbf{0.529}^{\pm.006}$	$\mathbf{0.667}^{\pm.006}$	$\mathbf{0.207}^{\pm.010}$	$\mathbf{3.899}^{\pm.027}$	$1.312^{\pm.048}$

Table 2. **Quantitative Evaluation on Oral Datasets.** Both the text-based method, Momask, and our audio-based method are trained on either the original or oral datasets and tested on the oral datasets. The results underscore the importance of our oral-style datasets as training data in better aligning with conversational applications. Furthermore, the evaluation highlights the effectiveness of audio signals for semantic conditioning, showing similar performance to text-based conditions. The notation ± indicates a 95% confidence interval.



Train on Original Dataset　　Train on Oral Dataset

Figure 6. **Qualitative Comparison between Models Trained on Original and Oral Datasets**. As illustrated in the figure, a comparison of the performance of models trained on the Oral and Original datasets reveals that the model trained on the Oral dataset demonstrates greater robustness to audio instructions. Which can be attributed to the higher diversity of the training data.

- **R-precision**: Evaluates the semantic consistency between the input conditions (audio/text) and the generated motion.
- **Diversity**: Assesses the variability and richness of the generated action sequences, ensuring the model does not produce overly similar outputs.
- **Multi-modal Distance (MM Dist)**: Computes the average Euclidean distance between the motion features and their corresponding conditioning features, capturing how well the generated motion aligns with the input conditions.
- **MultiModality**: Evaluates the range of distinct motion sequences generated from a single input condition, emphasizing the model's ability to generate multiple plausible outputs for the same input.

Implementation Details. The number of learnable tokens in our memory-retrieval based audio encoding is 256, with the dimension of 512. During training, we set the drop rate of the audio condition to 0.2 and adopt a linear warmup schedule, converging to a learning rate of 2e-4 after 2000 iterations. Both the masked transformer and residual transformer are composed of six transformer layers, each with six attention heads and a latent dimension of 384, and are applied to the HumanML3D and KIT-ML datasets. During generation process, we set the classifier-free guidance scale to 4 and 5 for the masked and residual transformer on HumanML3D, and (2, 5) on KIT-ML, with a fixed sequence length of 10 for both datasets.

5.1. Results on Oral Dataset

As discussed in Sec. 4.2, texts in the original datasets often differ significantly from the phrasing and syntactic patterns users employ during audio input. To highlight this, we present the performance of the most recent text-based model, Momask[19], and our audio-based method, both trained on either the original or oral datasets, and tested on

Datasets	Methods	R Precision↑			FID↓	MM Dist↓	MultiModality↑
		Top 1	Top 2	Top 3			
Human ML3D	TM2T [21]	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.002}$	$1.501^{\pm.017}$	$3.467^{\pm.011}$	$\underline{2.424}^{\pm.093}$
	T2M [20]	$0.455^{\pm.003}$	$0.636^{\pm.003}$	$0.736^{\pm.002}$	$1.087^{\pm.021}$	$3.347^{\pm.008}$	$2.219^{\pm.074}$
	MDM [39]	-	-	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$\mathbf{2.799}^{\pm.072}$
	MLD [10]	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$2.413^{\pm.079}$
	MotionDiffuse [44]	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$1.553^{\pm.042}$
	T2M-GPT [43]	$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$	$0.141^{\pm.005}$	$3.121^{\pm.009}$	$1.831^{\pm.048}$
	ReMoDiffuse [45]	$0.510^{\pm.005}$	$\underline{0.698}^{\pm.006}$	$0.795^{\pm.004}$	$\underline{0.103}^{\pm.004}$	$2.974^{\pm.016}$	$1.795^{\pm.043}$
	Momask [19]	$\mathbf{0.521}^{\pm.002}$	$\mathbf{0.713}^{\pm.002}$	$\underline{0.807}^{\pm.002}$	$\mathbf{0.045}^{\pm.002}$	$\underline{2.958}^{\pm.008}$	$1.241^{\pm.040}$
	Ours*	$\underline{0.519}^{\pm.004}$	$\mathbf{0.713}^{\pm.005}$	$\mathbf{0.808}^{\pm.005}$	$0.121^{\pm.004}$	$\mathbf{2.955}^{\pm.011}$	$1.221^{\pm.032}$
KIT-ML	TM2T [21]	$0.280^{\pm.005}$	$0.463^{\pm.006}$	$0.587^{\pm.005}$	$3.599^{\pm.153}$	$4.591^{\pm.026}$	$\mathbf{3.292}^{\pm.081}$
	T2M [20]	$0.361^{\pm.005}$	$0.559^{\pm.007}$	$0.681^{\pm.007}$	$3.022^{\pm.107}$	$3.488^{\pm028}$	$2.052^{\pm.107}$
	MDM [39]	-	-	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.191^{\pm.022}$	$1.907^{\pm.214}$
	MLD [10]	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$0.404^{\pm.027}$	$3.204^{\pm.027}$	$\underline{2.192}^{\pm.071}$
	MotionDiffuse [44]	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$1.954^{\pm.062}$	$2.958^{\pm.005}$	$0.730^{\pm.013}$
	T2M-GPT [43]	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$3.007^{\pm.023}$	$1.570^{\pm.039}$
	ReMoDiffuse [45]	$\underline{0.427}^{\pm.014}$	$0.641^{\pm.004}$	$0.765^{\pm.055}$	$\underline{0.155}^{\pm.006}$	$\underline{2.814}^{\pm.012}$	$1.239^{\pm.028}$
	Momask [19]	$\mathbf{0.433}^{\pm.007}$	$\mathbf{0.656}^{\pm.005}$	$\mathbf{0.781}^{\pm.005}$	$0.204^{\pm.011}$	$\mathbf{2.779}^{\pm.022}$	$1.131^{\pm.043}$
	Ours*	$0.426^{\pm.006}$	$\underline{0.645}^{\pm.007}$	$\underline{0.771}^{\pm.005}$	$\mathbf{0.113}^{\pm.004}$	$2.817^{\pm.021}$	$1.152^{\pm.048}$

Table 3. **Quantitative Evaluation on the Base Datasets.** Previous baseline models are conditioned on text descriptions from the HumanML3D and KIT-ML test sets. In contrast, our method (denoted as Ours*) is conditioned on audio from the Base Datasets. The symbol ± represents the 95% confidence interval. Results highlighted in **bold** indicate the best performance, while those underlined represent the second-best. These results underscore the effectiveness of using audio signals as a conditioning input, showing that audio can be as applicable as text in human motion generation tasks.

the oral datasets. In terms of quantitative results, following previous conventions [20, 42], each experiment is repeated 20 times, and the reported metric values represent the mean along with a 95% statistical confidence interval.

The quantitative results, shown in Tab. 2, reveal two key findings: (1) Models trained on the original datasets—whether conditioned on text or audio—perform significantly worse when tested on oral datasets whose the semantics representation is more conversational. This underscores the need for our proposed oral datasets to better align with practical applications. As shown in Fig. 6, it is evident that for some specific audio instructions, the models trained on the Oral dataset yield more instruction-congruent generation outcomes. We attribute this to the higher diversity inherent in the Oral dataset, which enhances the robustness of the trained models to handle various forms of instructions. (2) Compared to the text-based method, our audio-based approach presents similar performance regardless of being trained on the original or oral datasets, supporting our hypothesis that audio signals can serve as an effective alternative to text for semantic conditioning. Moreover, audio signals offer more convenient and efficient interactions in practical applications.

5.2. Comparison to Text-Based Methods

To further assess the effectiveness of audio signals in human motion generation, we compare our approach with several state-of-the-art text-based methods. The experiments are conducted using the original datasets, where our method utilizes audio signals as the input condition, while the compared methods are conditioned on text descriptions. Quantitative results are shown in Tab. 3. Our model achieves competitive performance across key metrics, including FID, R-Precision, and multimodal distance, when compared to the latest text-based approaches. These results indicate that instructive audio can serve as an equally effective condition as text, offering a comparable representational capacity for describing human movements explicitly, and also providing more convenient user interactions.

5.3. Ablation Study

In this section, we undertake an ablation study to examine the influence of the key components of our model on its overall performance.

Comparison to Cascaded Method. The previous experiments have demonstrated that audio signals have equivalent semantic representation ability with text for human motion generation. In many real-world applications, a cascaded approach is commonly used, where audio is first converted into text before being processed by downstream tasks. However, when the semantic encoding of audio is sufficiently robust, our proposed end-to-end framework can achieve significantly higher efficiency than such cascaded methods. To validate this, we used the small version of

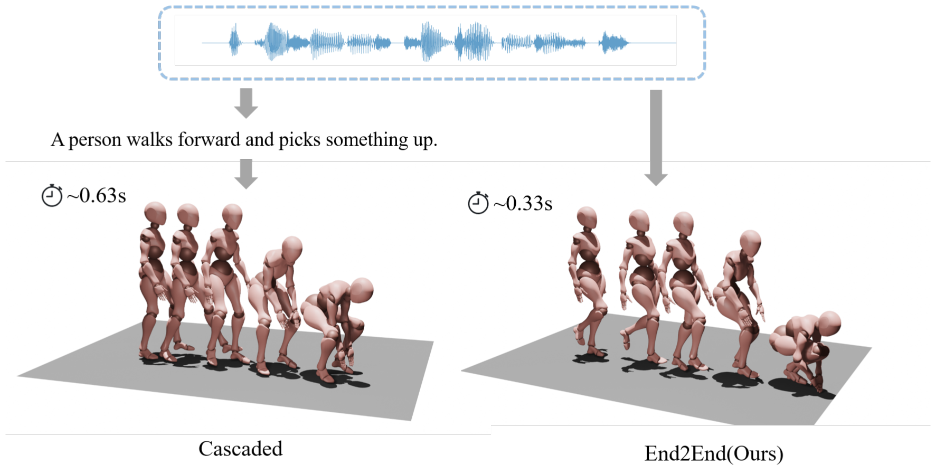


Figure 7. **Comparison with Cascaded Method.** Our end-to-end method can generate more than 50% faster than cascaded approach, while maintaining high generation quality.

Framework	Pipeline	Speed (sample/sec)
End2End (Ours)	audio→motion	2.6951
Cascaded	audio→text→motion	1.5093

Table 4. **Comparison to Cascaded Method.** Our end-to-end framework achieves over a 50% improvement in generation efficiency compared to the two-stage cascaded approach. The speed means the number of samples generated per second.

Models	FID↓	R-top1↑	R-top2↑	R-top3↑	MM Dist↓
Encodec [5]	2.920	0.233	0.370	0.464	5.548
ImageBind [15]	2.712	0.143	0.239	0.309	7.295
WavLM [9]	**0.113**	**0.426**	**0.645**	**0.771**	**2.817**

Table 5. **Ablation on Pre-trained Audio Feature Extractors.** WavLM's strong generalization across multiple downstream tasks leads to better performance in our audio-instructed motion generation, emphasizing the importance of effective pre-trained models in enhancing motion synthesis quality.

Whisper [36] for speech recognition. The output texts were then passed our text-based framework for motion generation, where texts are encoded with CLIP as input conditions. Performance testing was conducted on the same NVIDIA A6000 hardware. As shown in Tab. 4, our end-to-end framework achieves over a 50% improvement in generation efficiency compared to the two-stage cascaded approach, while maintaining high generation quality, as shown in Fig. 7. This highlights the practical advantages of bypassing text conversion in favor of direct audio input for motion generation.

Pre-trained Audio Feature Extractors. In audio research, numerous pre-trained models are available. To evaluate the effectiveness of our chosen WavLM model, we replaced it with alternative pre-trained models and conducted training and testing on the Original KIT-ML dataset. The results, shown in Tab. 5, demonstrate that WavLM significantly outperforms the alternatives. This improvement is

Audio Encoding	FID↓	R-top1↑	R-top2↑	R-top3↑	MM Dist↓
AvgPool-8	0.733	0.376	0.539	0.627	4.244
AvgPool-32	1.258	0.406	0.581	0.677	3.706
AvgPool-64	0.571	0.408	0.591	0.686	3.640
Conv1d	2.000	0.342	0.503	0.602	4.471
Transformer	0.523	0.496	0.680	0.778	3.259
Mem-Retr(Ours)	**0.121**	**0.519**	**0.713**	**0.808**	**1.221**

Table 6. **Ablation on Audio Feature Compression for Conditioning.** Our proposed memory-retrieval based module demonstrates a significant improvement over other architectural designs. This improvement highlights its effectiveness in compressing audio features as conditions, leading to better performance in motion generation tasks.

likely due to WavLM's strong generalization across multiple downstream tasks, whereas other models are often optimized for specific tasks.

Audio Feature Compression for Conditioning. As discussed in Sec. 3.2, the extracted audio features exhibit significant variation in sequence length across different samples, making an effective compression strategy essential for subsequent audio-conditioned motion generation. We explored several methods for compressing sequence lengths, including Average Pooling (AvgPool) to fixed lengths (e.g., 8, 32, 64), Conv1d, and Transformer Encoder, for encoding audio conditions. The results, presented in Tab. 6, are based on models trained and tested on the Original HumanML3D dataset. The results indicate that AvgPool delivers unsatisfactory performance, likely due to its lack of learnable parameters, which limits its adaptability to varying audio features. Similarly, Conv1d performs poorly, as its fixed receptive field struggles to capture long-range dependencies. The Transformer Encoder, although more capable, still appears to incorporate irrelevant audio content that is not directly related to motion description. In contrast, our proposed feature compression module effectively handles the significant variation in audio sequence lengths, producing more refined and semantics-aware audio conditions. This improvement leads to enhanced performance in generating accurate, audio-conditioned motions.

6. Conclusion

In conclusion, this paper introduces a novel task: motion generation based on audio instructions. Compared to text-based interactions, audio offers a more convenient and natural mode of user interaction in real-world scenarios. We propose an end-to-end framework built on a masked generative transformer, enhanced with a memory-retrieval based attention module to effectively address the challenges of sparse and extended audio signals. Additionally, we augment existing text-based datasets by rephrasing textual descriptions into a more conversational, spoken language style and synthesizing corresponding audio with diverse speaker identities. Our experimental results demonstrate the effectiveness and efficiency of this framework in generating human motion directly from audio instructions. For future work, we plan to explore the unique characteristics of instructive audio as conditioning signals, focusing on aspects that text cannot capture, such as pronunciation, intonation, and their influence on motion generation.

Acknowledgments

The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions. The work was partially supported by grants from Shenzhen Science and Technology Program (Grant No. JSGG20220831105002004) and the Shenzhen Key Laboratory of Computer Vision and Pattern Recognition.

References

[1] C. Ahuja and L.-P. Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision*, pages 719–728. IEEE, 2019. 3

[2] T. Ao, Q. Gao, Y. Lou, B. Chen, and L. Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. *ACM Transactions on Graphics*, 41(6):1–19, 2022. 1, 3

[3] T. Ao, Z. Zhang, and L. Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics*, 42(4):1–18, 2023. 1, 3

[4] J. Betker. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*, 2023. 2, 7, 8

[5] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE transactions on audio, speech, and language processing*, 31:2523–2533, 2023. 3, 5, 6, 11

[6] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. Murphy, W. T. Freeman, M. Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[7] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 2, 4, 6

[8] K. Chen, Z. Tan, J. Lei, S.-H. Zhang, Y.-C. Guo, W. Zhang, and S.-M. Hu. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics*, 40(4):1–13, 2021. 1, 3

[9] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. 2, 3, 4, 5, 11

[10] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 10

[11] R. Dabral, M. H. Mughal, V. Golyanik, and C. Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9760–9770, 2023. 1

[12] J. Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[13] A. Ghosh, N. Cheema, C. Oguz, C. Theobalt, and P. Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE international conference on computer vision*, pages 1396–1406, 2021. 3

[14] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 3

[15] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 4, 5, 11

[16] K. Gong, D. Lian, H. Chang, C. Guo, Z. Jiang, X. Zuo, M. B. Mi, and X. Wang. Tm2d: Bimodality driven 3d dance generation via music-text integration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9942–9952, 2023. 1

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. Publisher: ACM New York, NY, USA. 4

[18] A. Graves and A. Graves. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45, 2012. 3

[19] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng. Momask: Generative masked modeling of 3d human motions.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 1, 2, 3, 6, 9, 10

[20] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 1, 2, 3, 4, 5, 7, 8, 10

[21] C. Guo, X. Zuo, S. Wang, and L. Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 10

[22] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4, 8

[23] J. Jiang, C. Liang, J. Yang, G. Lin, T. Zhong, and Y. Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024. 5

[24] J. Kim, J. Kim, and S. Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8255–8263, 2023. 3

[25] D. P. Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[26] T. Kucherenko, D. Hasegawa, G. E. Henter, N. Kaneko, and H. Kjellström. Analyzing input and output representations for speech-driven gesture generation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 97–104, 2019. 3

[27] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, and J. Kautz. Dancing to music. *Advances in neural information processing systems*, 32, 2019. 3

[28] J. Li, D. Kang, W. Pei, X. Zhe, Y. Zhang, Z. He, and L. Bao. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 11293–11302, 2021. 1, 3

[29] R. Li, S. Yang, D. A. Ross, and A. Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 13401–13412, 2021. 1

[30] A. S. Lin, L. Wu, R. Corona, K. Tai, Q. Huang, and R. J. Mooney. Generating animated videos of human activities from natural language descriptions. *Learning*, 1(2018):1, 2018. 3

[31] A. J. OpenAI, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. 2023. *URL: https://arxiv. org/abs/2303.08774*, 2024. 1, 2, 7, 8

[32] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. 4

[33] M. Petrovich, M. J. Black, and G. Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 3

[34] M. Plappert, C. Mandery, and T. Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 1, 2, 4, 5, 7, 8

[35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3, 5

[36] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 1, 4, 5, 11

[37] T. Tang, J. Jia, and H. Mao. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1598–1606, 2018. 3

[38] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 3

[39] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3, 10

[40] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[41] A. Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 5

[42] Y. Yuan, J. Song, U. Iqbal, A. Vahdat, and J. Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE international conference on computer vision*, pages 16010–16021, 2023. 10

[43] J. Zhang, Y. Zhang, X. Cun, Y. Zhang, H. Zhao, H. Lu, X. Shen, and Y. Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 1, 3, 10

[44] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 8, 10

[45] M. Zhang, X. Guo, L. Pan, Z. Cai, F. Hong, H. Li, L. Yang, and Z. Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 364–373, 2023. 10

[46] Z. Zhang, T. Ao, Y. Zhang, Q. Gao, C. Lin, B. Chen, and L. Liu. Semantic gesticulator: Semantics-aware co-speech gesture synthesis. *ACM Transactions on Graphics*, 43(4):1–17, 2024. 1, 3

[47] W. Zhu, X. Ma, D. Ro, H. Ci, J. Zhang, J. Shi, F. Gao, Q. Tian, and Y. Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 3