RapVerse: Coherent Vocals and Whole-Body Motion Generation from Text

Jiaben Chen[1] Xin Yan[2] Yihang Chen[3] Siyuan Cen[1] Zixin Wang[1] Qinwei Ma[4]
Haoyu Zhen[1] Kaizhi Qian[5] Lie Lu[6] Chuang Gan[1]

[1]UMass Amherst [2]Wuhan University [3]UC San Diego [4]Tsinghua University
[5]MIT-IBM Watson AI Lab [6]Dolby Laboratories

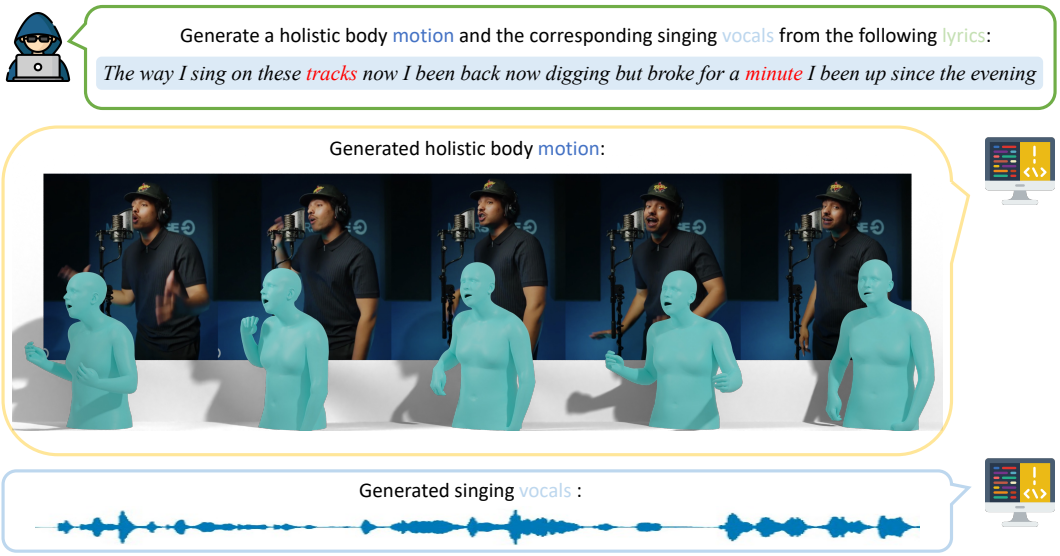

Figure 1. **RapVerse.** We present a unified text-conditioned multi-modality generation framework, for jointly generating holistic body motions and singing vocals from textual lyrics inputs only. Note that the corresponding video frames are just shown for reference.

Abstract

In this work, we introduce a challenging task for simultaneously generating 3D holistic body motions and singing vocals directly from textual lyrics inputs, advancing beyond existing works that typically address these two modalities in isolation. To facilitate this, we first collect the RapVerse dataset, a large dataset containing synchronous rapping vocals, lyrics, and high-quality 3D holistic body meshes. With the RapVerse dataset, we investigate the extent to which scaling autoregressive multimodal transformers across language, audio, and motion can enhance the coherent and realistic generation of vocals and whole-body human motions. For modality unification, a vector-quantized variational autoencoder is employed to encode whole-body motion sequences into discrete motion tokens, while a vocal-to-unit model is leveraged to obtain quantized audio tokens preserving content, prosodic information and singer identity. By jointly performing transformer modeling on these three modalities in a unified way, our framework ensures a seamless and realistic blend of vocals and human motions. Extensive experiments demonstrate that our unified generation framework not only produces coherent and realistic singing vocals alongside human motions directly from textual inputs, but also rivals the performance of specialized single-modality generation systems, establishing new benchmarks for joint vocal-motion generation. Video demonstration and more information can be found on the project page[1].

1. Introduction

In the evolving landscape of multi-modal content generation in terms of sound and motion, significant strides have been made in individual modalities, including text-to-music [1, 24, 51], text-to-vocal [34], text-to-motion [7, 16, 26, 35,

[1]https://jiabenchen.github.io/RapVerse/.

66], and audio-to-motion [18, 33, 65] generation. These developments have paved the way for creating more dynamic and interactive digital content. Despite these advancements, existing works predominantly operate in silos, addressing each modality in isolation. However, there's strong psychological evidence that for human beings, the generation of sound and motion are highly related and coupled [30]. A unified system for joint generation allows for a more expressive and nuanced communication of emotions, intentions, and context, where the generation of one modality could guide and assist the other in a coherent and efficient way.

In this paper, we tackle a crucial problem: can a machine not only sing with emotional depth but also perform with human-like expressions and motions? We propose a novel task for generating coherent singing vocals and whole-body human motions (including body motions, hand gestures, and facial expressions) simultaneously, see Fig. 1. This endeavor holds practical significance in fostering more immersive and naturalistic digital interactions, thereby elevating virtual performances, interactive gaming, and the realism of virtual avatars.

An important question naturally arises: what constitutes a good model for unified generation of sound and motion? Firstly, we consider textual lyrics as the proper form of inputs for the unified system, since text provides a highly expressive, interpretable flexible means of conveying information by human beings, and could serve as a bridge between various modalities. Previous efforts explore scores [34], action commands [7, 26, 66], or audio signals [65] as inputs, which are inferior to textual inputs in terms of semantic richness, expressiveness and flexible integration of different modalities.

Secondly, we reckon that a joint generation system that could produce multi-modal outputs simultaneously is better than a cascaded system that executes the single-modal generation sequentially. A cascaded system, combining a text-to-vocal module with a vocal-to-motion module, risks accumulating errors across each stage of generation. For instance, a misinterpretation in the text-to-vocal phase can lead to inaccurate motion generation, thereby diluting the intended coherence of the output. Furthermore, cascaded architectures necessitate multiple training and inference phases across different models, substantially increasing computational demands.

To build such a joint generation system, the primary challenges include: 1) the scarcity of datasets that provide lyrics, vocals, and 3D whole-body motion annotations simultaneously; and 2) the need for a unified architecture capable of coherently synthesizing vocals and motions from text. In response to these challenges, we have curated **RapVerse**, a large-scale dataset featuring a comprehensive collection of lyrics, singing vocals, and 3D whole-body motions. Despite the existence of datasets available for text-to-vocal [11, 25, 34, 52], text-to-motion [16, 32, 37, 43], and audio-to-motion [5, 8, 12, 15, 18, 62], the landscape lacks a unified dataset that encapsulates singing vocals, whole-body motion, and lyrics simultaneously. Most notably, large text-to-vocal datasets [25, 67] are predominantly in Chinese, limiting their applicability for English language research and lacking any motion data. And text-to-motion datasets [16, 32, 43] typically focus on text descriptions of specific actions paired with corresponding motions without audio data, often not covering whole body movements. Moreover, audio-to-motion datasets [34, 36] focus primarily on speech rather than singing. A comparison of existing related datasets is demonstrated in Table. 1. The RapVerse dataset is divided into two distinctive parts to cater to a broad range of research needs: 1) a Rap-Vocal subset containing a large number of pairs of vocals and lyrics, and 2) a Rap-Motion subset encompassing vocals, lyrics, and human motions. The Rap-Vocal subset contains 108.44 hours of high-quality English singing voice in the rap genre without background music. Paired lyrics and vocals are crawled from the Internet from 32 singers, with careful cleaning and post-processing. On the other hand, the Rap-Motion subset contains 26.8 hours of rap performance videos with 3D holistic body mesh annotations in SMPL-X parameters [41] using the annotation pipeline of Motion-X [32], synchronous singing vocals and corresponding lyrics.

With the RapVerse dataset, we explore how far we can push by simply scaling autoregressive multimodal transformers with language, audio, and motion for a coherent and realistic generation of vocals and whole-body human motions. To this end, we unify different modalities as token representations. Specifically, three VQVAEs [60] are utilized to compress whole-body motion sequences into three-level discrete tokens for head, body, and hand, respectively. For vocal generation, previous works [10, 34, 38, 38] share a common paradigm, producing mel-spectrograms of audio signals from input textual features and additional music score information, following with a vocoder [40, 59, 64] to reconstruct the phase. We draw inspiration from the speech resynthesis domain [44], and learn a self-supervised discrete representation to quantize raw audio signal into discrete tokens while preserving the vocal content and prosodic information. Then, with all the inputs in discrete representations, we leverage a transformer to predict the discrete codes of audio and motion in an autoregressive fashion. Extensive experiments demonstrate that this straightforward unified generation framework not only produces realistic singing vocals alongside human motions directly from textual inputs but also rivals the performance of specialized single-modality generation systems.

To sum up, this paper makes the following contributions:

- We collect RapVerse, a large dataset featuring synchronous singing vocals, lyrics, and high-quality 3D

holistic SMPL-X parameters.

- We design a simple but effective unified framework for the joint generation of singing vocals and human motions from text with a multi-modal transformer in an autoregressive fashion.
- To unify representations of different modalities, we employ a vocal-to-unit model to obtain quantized audio tokens and utilize compositional VQVAEs to get discrete motion tokens.
- Experimental results show that our framework rivals the performance of specialized single-modality generation systems, setting new benchmarks for joint generation of vocals and motion.

2. Related Work

2.1. Text to Vocal Generation

Text-to-audio Dataset. There exists several singing vocal datasets, yet they each has constraints. For instance, PopCS [34] and OpenSinger [25] are limited to Chinese, while NUS-48E [11] and NHSS [52] have only a few hours of songs. Nevertheless, JVS-MuSiC [55] and NUS-48E [11] offer a few hours of songs from dozens of singers, whereas OpenSinger [25] provides a more extensive collection with tens of hours from a single singer. Notably, our dataset represents the first to be specifically curated for rap songs of multiple singers with 108 hours.

Text-to-vocal Models. Recent advancements in text-to-speech (TTS) models, including WavNet [40], FastSpeech 1 and 2 [48, 49], and EATS [10], have significantly improved the quality of synthesized speech. However, singing voice synthesis (SVS) presents a greater challenge due to its reliance on additional musical scores and lyrics. Recent generation models [19, 19, 29, 34, 71] perform excellently in generating singing voices. However, their performance tends to deteriorate when encountering out-of-distribution data. To handle this, StyleSinger [70] introduces a Residual Style Adaptor and an Uncertainty Modeling Layer Normalization to handle this.

2.2. Text to Motion Generation

Text-to-motion Dataset. Current text-motion datasets, such as KIT [43], AMASS [37], and HumanML3D [16], are constrained by limited data coverage, typically spanning only tens of hours and lacking whole-body motion representation. To overcome these shortcomings, Motion-X [32] was developed, providing an extensive dataset that encompasses whole-body motions.

Text-to-motion Models. Text-to-motion models [7, 16, 17, 35, 42, 68] have gained popularity for their user convenience. Recent advancements have focused on diffusion models [28, 58, 68], which, unlike deterministic generation models [2, 14], enable finer-grained and diverse output generation. Chen et al. [7] proposed a motion latent-based diffusion model to enhance generative quality and reduce computational costs. However, the misalignment between natural language and human motions presents challenges. Lu et al. [35] introduced the Text-aligned Whole-body Motion generation framework to address these issues by generating high-quality, diverse, and coherent facial expressions, hand gestures, and body motions simultaneously.

2.3. Audio to Motion Generation

Audio-to-motion Dataset. Co-Speech Datasets can be classified into 1: pseudo-labeled (PGT) and 2: motion-captured (mocap). PGT datasets [5, 15, 18, 65] utilize disconnected 2D or 3D key points to represent the body, allowing extracting at a lower cost but with limited accuracy. On the other hand, mocap datasets provide annotations for limited body parts, with some focusing solely on the head [8, 12, 62] or the body [13, 54]. Differing from them, BEATX [33] contains mesh data of both head and body. However, these datasets typically focus on speech to motion generation. Our RapVerse, in contrast, contains paired text-vocal-motion data, enabling simultaneous motion and vocal generation.

Audio-to-motion Models. Audio-to-motion models [3, 4, 6, 33, 53, 56, 65, 69] aim to produce human motion from audio innputs. Recognizing the intricate relationship between audio and human face, TalkSHOW [65] separately generates face parts and other body parts. However, this approach exhibits limitations, such as the absence of generating lower body motion. Motioncraft [4] proposes a DiT structure for motion generation. EMAGE [33] utilizes masked gestures together with vocals for a generation. Our model, however, goes a step further by being the first to generate paired audio-motion data directly from text.

3. RapVerse Dataset

In this section, we introduce RapVerse, a large rap music motion dataset containing synchronized singing vocals, textual lyrics and whole-body human motions. A comparison of our dataset with existing datasets is shown at Table. 1. The RapVerse dataset is divided into two subsets to cater to a broad range of research needs: a Rap-Vocal subset and a Rap-Motion subset. The overall collection pipeline of RapVerse is shown at Fig. 2.

3.1. Rap-Vocal Subset

The Rap-Vocal subset contains 108.44 hours of high-quality English singing voice in the rap genre with paired lyrics. We will introduce each step in detail.

Data Crawling. In a bid to obtain a large number of rap songs and corresponding lyrics from the Internet, we utilize Spotdl and Spotipy to collect songs, lyrics, and metadata of different rap singers. To ensure the quality of the dataset, we

Dataset	Language	#Singers	Hours
NUS-48E [11]	English	12	1.91
NHSS [52]	English	10	7
JVS-MuSiC [55]	Japanese	100	2.28
Tohoku Kiritan [39]	Japanese	1	1
PopCS [34]	Chinese	1	5.89
OpenSinger [25]	Chinese	66	50
Opencpop [61]	Chinese	1	5.25
M4Singer [67]	Chinese	20	29.77
Ours	English	32	108.44

(a) Audio Datasets

Dataset	Modality	Whole†	Hours
KIT-ML [43]	Text	55	11.2
AMASS [37]	Text	55	40.0
HumanML3D [16]	Text	55	28.6
Motion-X [32]	Text	51	127.1
HA2G [34]	Audio	55	33
Yoon et.al [36]	Audio	55	30
Speech2Gesture [15]	Audio	55	144
Talkshow [65]	Audio	51	27.0
Ours	Text-Audio	51	26.8

(b) Motion Datasets

Table 1. **Comparison of Audio and Motion Datasets.** (a) compares our Rap-Vocal Subset with existing singing vocal datasets. (b) compares our Rap-Motion Subset with existing motion datasets. †: "Whole" means whole-body motion including face, body, and hand.

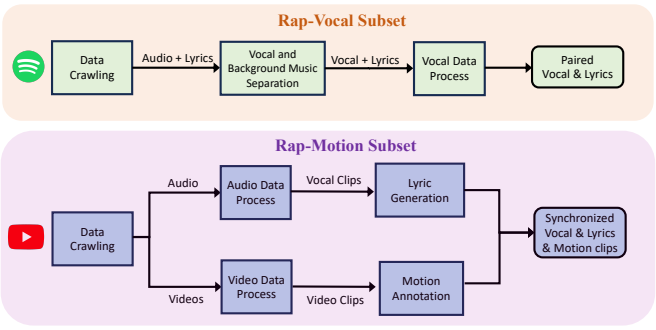

Figure 2. **RapVerse** dataset collection pipeline. There are two pathways for rap-vocal and rap-motion subsets, respectively.

perform cleaning on the crawled songs by removing songs with misaligned lyrics and filtering out songs that are too long or too short.

Vocal and Background Music Separation. Since the crawled songs are mixed with rapping vocals and background music, and we aim to synthesize singing vocals from separated clean data, we utilize Spleeter [21], the state-of-the-art open-source vocal-music source separation tool to separate and extract rap vocal voices and accompanying background music from the collected songs. Following [50], we normalize the loudness of the vocal voices to a fixed loudness level.

Vocal Data Processing. The raw crawled lyrics from the Internet are in inconsistent formats, we conduct data cleaning on the lyrics by removing meta information (singer, composer, song name, bridging words, and special symbols). To ensure that the lyrics are aligned with the singing vocals, we collect lyrics only with the correct timestamps of each sentence, and we separate each song into around 10-second to 20-second segments for model training.

3.2. Rap-Motion Subset

The Rap-Motion subset contains 26.8 hours of rap performance videos with 3D holistic body mesh annotations in SMPL-X parameters [41], synchronous singing vocals, and corresponding lyrics. We introduce the collection pipeline of this subset as follows.

Data Crawling. We crawled over 1000 studio performance videos from YouTube under the Common Creative License. We filter out low-quality videos manually to ensure the videos meet the following criteria: stable camera work, performers centered in the frame, clear visibility of the performer's entire body to capture detailed motion data, and high-quality audio for accurate vocal analysis.

Audio Data Processing. Similar to the Rap-Vocal subset, we leverage Spleeter [22] to isolate singing vocals from accompanying music. Given that YouTube videos typically lack paired lyrics, we employ an ASR model, Whisper [45], to accurately transcribe vocals into corresponding text.

Video Data Processing. To ensure the collection of high-quality video clips for motion annotation, we implemented a semi-automatic process to filter out undesirable content, such as advertisements, transition frames, changes in shots, and flashing lights. Initially, we applied YOLO [47] for human detection to discard frames where no humans were detected. Subsequently, we utilized RAFT [57] to assess the motion magnitude, employing a threshold to eliminate frames affected by camera instability. We then perform meticulous manual curation on the extracted clips, retaining only those of the highest quality. Finally, we follow the pipeline of the optimized-based method Motion-X [32] to extract 3D whole-body meshes from monocular videos. Specifically, we adopt the SMPL-X [41] for motion representations, given a T-frame video clip, the corresponding pose states $\mathcal{M}$ are represented as:

$$\mathcal{M} = \{\mathcal{M}_f, \mathcal{M}_b, \mathcal{M}_h, \zeta, \epsilon\} \quad (1)$$

where $\mathcal{M}_f \in \mathbb{R}^{T\times 3}, \mathcal{M}_b \in \mathbb{R}^{T\times 63}$ and $\mathcal{M}_h \in \mathbb{R}^{T\times 90}$ denotes jaw poses, body poses and hand poses, respectively. $\zeta \in \mathbb{R}^{T\times 100}$ and $\epsilon \in \mathbb{R}^{T\times 3}$ are the facial expression and global translation.

4. Method

Given a piece of lyrics text, our goal is to generate rap-style vocals and whole-body motions, including body movements, hand gestures, and facial expressions that resonate with the lyrics. With the help of our RapVerse dataset, we propose a novel framework that not only represents texts, vocals, and motions as unified token forms but also integrates token modeling in a unified model. As illustrated in Fig. 3, our model consists of multiple tokenizers for motion (Sec. 4.2) and vocal (Sec. 4.3) token conversions, as well as a general Large Text-Motion-Audio Foundation Model (Sec. 4.4) that targets for audio token synthesize and motion token creation, based on rap lyrics.

4.1. Problem Formulation

Let $\mathcal{L}$ be the rap lyric, such as "*I going to make it a flashback set a picture from way back.*", our model will compose the text-related vocal $\mathcal{V} \in \mathbb{R}^{t\times d_v}$ and whole-body motion $\mathcal{M} \in \mathbb{R}^{t\times d_m}$, where t, d_v, d_m denote the temporal length and the feature dimensions for vocal and motion in every time unit, respectively. Modality tokenizer encoders $\phi^{\mathcal{X}}(\mathcal{T} \mid \mathcal{X})$ and decoders $\varphi^{\mathcal{X}}(\mathcal{X} \mid \mathcal{T})$ will bridge each modality with the token format $\mathcal{T}$ utilizing different tokenizers, where $\mathcal{X} \in \{\mathcal{L}, \mathcal{V}, \mathcal{M}\}$. The large foundation model $\psi^{\mathcal{X}}$ will incorporate tokens in unified modeling. The whole problem of coherent vocal and model generation via lyric text can be formulated as follows:

$$\Theta^* = \arg\max_{\Theta} \mathbb{P}_{\Theta}(\mathcal{V}, \mathcal{M} \mid \mathcal{L}), \tag{2}$$

where Θ denotes the whole model parameters and $\mathbb{P}_{\Theta}(\cdot)$ denotes the model distributions.

4.2. Motion VQ-VAE Tokenizer

Similar to prior arts, a motion VQ-VAE is adopted to convert 3D human motions into discrete tokens, where the tokenizer encoder ϕ generates motion tokens enriched with high-frequency information, while the decoder φ seamlessly reconstructs discrete tokens into continuous motion sequences. Specifically, the tokenizer will maintain a learnable codebook $\mathcal{C}$, which firstly encodes the input motion features $\mathcal{M}$ to vectors $z^{\mathcal{M}}$, and then the motion quantizer $\mathcal{Q}^{\mathcal{M}}$ will look up the nearest neighbor in codebook $\mathcal{C}^{\mathcal{M}} = \{c_k^{\mathcal{M}}\}_{k=1}^{d_c}$ for the motion representation. The quantized code can be calculated as

$$Q(z^{\mathcal{M}}; \mathcal{C}^{\mathcal{M}}) = \arg\min_{k} \|z^{\mathcal{M}} - c_k^{\mathcal{M}}\|_2. \tag{3}$$

The tokenizer decoder φ will project the quantized code back into the human motion after quantization. The reconstruction objective in the tokenizer training stage can be represented as

$$\mathbb{E}_{\mathcal{T}\sim\phi_{\mathcal{M}}(\mathcal{T}|\mathcal{M})} \left[\log \varphi_{\mathcal{M}}(\mathcal{M} \mid \mathcal{T})\right]. \tag{4}$$

As the whole-body motion $\mathcal{M}$ can be decomposed into face $\mathcal{M}_f$, body $\mathcal{M}_b$, and hand $\mathcal{M}_h$, where $\mathcal{M} \in \mathbb{R}^{t\times d}$, $d_f + d_b + d_h = d_m$, we build three tokenizers to obtain tokens for these three parts.

4.3. Vocal2unit Audio Tokenizer

Overall, we leverage the self-supervised framework [44] in speech resynthesis domain to learn vocal representations from the audio sequences. Specifically, we train a Vocal2unit audio tokenizer to build a discrete tokenized representation for the human singing voice. The vocal tokenizer consists of three encoders and a vocoder. The encoders include three different parts: (1) the semantic encoder; (2) the F0 encoder; and (3) the singer encoder. We will introduce each component of the model separately.

Semantic Encoder. In order to extract the semantic information of the audio into discrete tokens, we use a pretrained Hubert encoder [23] as the main component of our semantic encoder network E^h. Given a raw audio sequence v, the output of the Hubert encoder is a continuous representation. Then we apply a K-means algorithm on the representations to finally generate discrete units $z^{\mathcal{S}} = \{z_i^{\mathcal{S}}\}_{i=1}^{L_s}$ where $z_i^{\mathcal{S}} \in \{1, 2, \cdots, K\}$. L_s is the length of hubert token sequence, and K represents the number of clusters in K-means algorithm.

F0 Encoder. We use the YAAPT algorithm [27] to extract the F0 from the input audio signal v, which is then fed into an encoder E_p to generated low-frequency F0 pitch representation. The encoder E^p is trained using a VQ-VAE framework with Exponential Moving Average updates similar to [9], maintaining a learnable codebook $\mathcal{C}^p = \{c_k^p\}_{k=1}^{d_c}$, where d_c is the number of codes in the codebook. The raw F0 sequence is first preprocessed according to the singer, converted into a sequence of latent vectors, then mapped to its nearest neighbors in the codebook c^p, which could be represented using integer number $k \in \{1, 2, \cdots, d_c\}$ as its index in the codebook. The index sequence will be further formatted as the token *¡pitch_k¿* to produce pitch token sequence $z^{\mathcal{P}} = \{z_i^{\mathcal{P}}\}_{i=1}^{L_p}$, where L_p is the length of the pitch token sequence. Finally, the decoder will reconstruct the F0 representation leveraging the pitch token sequence.

Singer Encoder. As the last component of our encoders, the singer encoder E_{si} from [20, 44] is used to extract the Mel-spectrogram from the raw audio sequence and to output a singer representation $z^{\mathcal{I}} \in \mathbb{R}^{256}$. The singer embedding only depends on the singer and is trained globally.

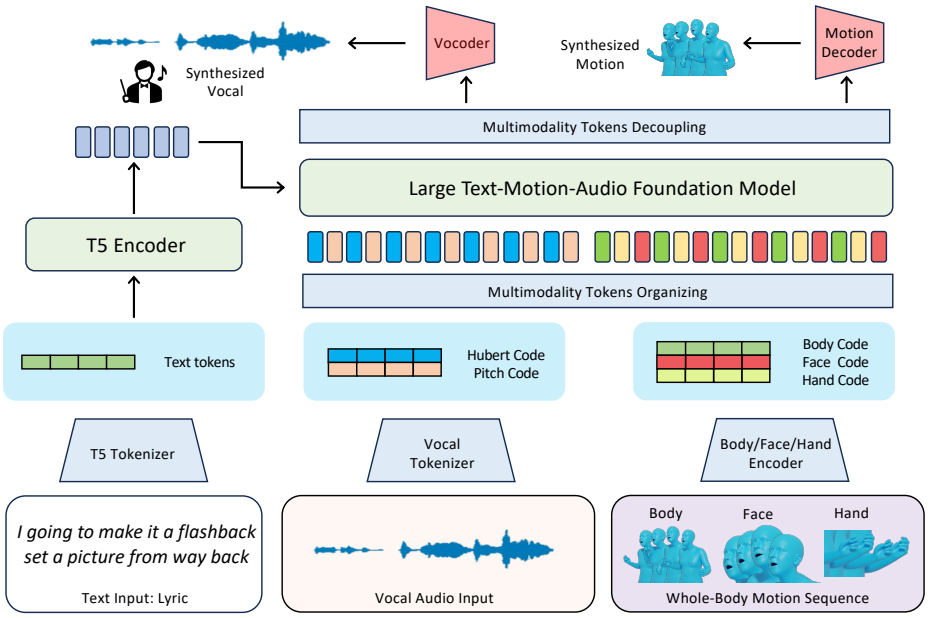

Figure 3. **Pipeline overview.** We first pre-train all tokenizers on vocal-only and motion-only data. After we have pretrained the modality tokenizers, we can unify text, vocal, and motion in the same token space. We adopt a mixing organizing algorithm for input tokens to align via the temporal axis. These mixed input tokens are fed into the large Text-Motion-Audio foundation model to train on token prediction tasks, guided by the encoded features from textual input.

Vocoder. In order to decode the vocal signal from the discrete tokens, we adopt a modified version of the HiFi-GAN neural vocoder similar to the one in [44], consisting of a generator and multiple discriminators. The generator takes the encoded discrete representation $z^{\mathcal{S}}$ and $z^{\mathcal{P}}$, and the singer embedding vector $z^{\mathcal{I}}$ as inputs. After a lookup table and a series of blocks composed of transposed convolution and a residual block with dilated layers, it finally outputs waveform audio. The discriminators include multiple sub-discriminators, analyzing on different scales and different periodic structures.

4.4. General Auto-regressive Modeling

Model Architectures. The whole model of our general modeling pipeline consists of several pre-trained vocal and motion tokenizers and a large Text-Motion-Audio foundation model. The vocal and motion tokenizers are based on VQ-VAE architectures, enabling us to represent these two modalities as discrete tokens. We also adopt a T5-Tokenizer [46] to convert lyric text into tokens, so that all three modalities $\mathcal{X} \in \{\mathcal{L}, \mathcal{V}, \mathcal{M}\}$ are united in the token space $\mathcal{T}^{\mathcal{X}} = \{t_k^{\mathcal{X}}\}_{k=1}^{\mathcal{N}^{\mathcal{X}}}$, where $\mathcal{N}^{\mathcal{X}}$ is the token sequence length for each modality. The large foundation model, a temporal transformer based on the decoder-only architecture, performs the next-token-prediction task based on textual features encoded by the T5-Encoder.

Specifically, each token is generated after the start token by the probability distribution from the foundation model $p_\delta(t^{\mathcal{X}}; h^{\mathcal{L}}) = \prod_i p_\delta(t_i^{\mathcal{X}} \mid t_{<i}^{\mathcal{X}}; h_{\mathcal{L}})$. During the training period, we calculate the cross-entropy loss to measure the likelihood between ground truth tokens and predicted token probabilities as

$$-\sum_{i=1}^{\mathcal{N}^{\mathcal{X}}} \log p_\delta(t_i^{\mathcal{X}}; h^{\mathcal{L}}). \tag{5}$$

After optimizing via this training objective, our model learns to predict the next token, which can be decoded into different modality features. This process is similar to text word generation in language models, while the "word" in our method such as *¡face_0123¿*, does not have explicit semantic information, but can be decoded into continuous modality features.

Multimodality Tokens Organization. Since we have designed a coherent vocal and motion generation pipeline with multiple token types in our method, including *hubert* and *pitch* tokens in rap vocals, and *face*, *body*, *hand* in whole-body motions, the organization method of all these tokens in one sentence is important for training. Similar to previous works [26, 35], we adopt an interleaved style within each modality. Specifically, with the input whole-body motion tokens $\mathcal{T}^{\mathcal{M}_s} = \{t_k^{\mathcal{M}_s}\}_{k=1}^{\mathcal{N}^{\mathcal{M}_s}}$, where $s = f, b, h$ which denotes face, body and hand respectively, and $\mathcal{N}^{\mathcal{M}_s}$ denotes the token length for each part. Therefore the motion token sequence can be formulated as $\mathcal{T}^{\mathcal{M}} = \{t_1^{\mathcal{M}_f}, t_1^{\mathcal{M}_b}, t_1^{\mathcal{M}_h}, t_2^{\mathcal{M}_f}, ...\}$. Similarly, the vocal to-

kens consist of the hubert tokens $t^{\mathcal{V}_h}$ and pitch tokens $t^{\mathcal{V}_p}$, which can be formulated as $\mathcal{T}^{\mathcal{V}} = \{t_1^{\mathcal{V}_h}, t_1^{\mathcal{V}_p}, t_2^{\mathcal{V}_h}, t_2^{\mathcal{V}_p}, ...\}$. This interleaved way of organization allows us to align the timesteps of each part. When performing coherent vocal and motion generation, we mix these two modality sequences sequentially with modality-wise start token as $\mathcal{T} = \{t_{start}^{V}, \mathcal{T}^{V}, t_{start}^{M}, \mathcal{T}^{M}\}$. Note that we put vocal sequences before motion sequences because 1) the vocals are more related to the lyrics, and 2) motion tokens have a relationship with previously generated vocals, *e.g.* face motions with lips movement are directly related to vocals. Joint training of mixed modalities allows the transformer to capture the temporal dependencies and dynamics inherent in vocal and motion data.

Inference and Decoupling. In the inference stage, we use different start tokens to specify which modality to generate. The textual input is encoded as features to guide token inference. We adopt a top-k algorithm to control the diversity of the generated content by adjusting the temperature, as generating vocals and motions based on lyrics is a creation process with multiple possible answers. After token prediction, a decoupling algorithm is used to process output tokens to make sure tokens from different modalities are separated and temporally aligned. These discrete tokens will be further decoded into text-aligned vocals and motions.

5. Experiments

In this section, we evaluate our proposed model on our proposed benchmark designed for joint vocal and whole-body motion generation from textual inputs.

5.1. Experimental Setup

Metrics. To evaluate the generation quality of singing vocals, we utilize the Mean Opinion Score (MOS) to gauge the naturalness of the synthesized vocal. For motion synthesis, we evaluate the generation quality of the body hand gestures and the realism of the face, respectively. Specifically, for gesture generation, we use Frechet Inception Distance (FID) based on a feature extractor from [16] to evaluate the distance of feature distributions between the generated and real motions, and Diversity (DIV) metric to assess the motions diversity. For face generation, we compare the vertex MSE [63] and the vertex L1 difference LVD [65]. Finally, we adopt Beat Constancy (BC) [31] to measure the synchrony of generated motion and singing vocals.

Baselines. We compare the vocal generation quality with the state-of-the-art vocal generation method DiffSinger [34]. And we also adapt the text-to-speech model FastSpeech2 [49] for vocal generation. For motion generation, we compare our method with both text-to-motion methods and audio-to-motion methods. For text-to-motion methods, since there is no existing open-sourced work for text to whole-body motion generation, we compare with transformer-based T2M-GPT [66] and MLD [7] for body generation. For the audio-to-motion generation, we compare with Habibie et al. [18], Talkshow [65], Motioncraft [4] and EMAGE [33]. We report all the results on RapVerse with an 85%/7.5%/7.5% train/val/test split.

5.2. Main Results Analysis

Evaluations on joint vocal and whole-body motion generations. We compared both text-driven and audio-driven motion generation baselines in Table. 2 (a). To be noted, our setting is different from all existing methods in the following ways. First, we use rap lyrics as our textual input instead of motion textual descriptions, which contain direct action prompt words, such as *walk* and *jump*; Second we use text to jointly generate both audio and motion, instead of using audio to generate motion as audio-driven methods did. As is demonstrated, our model rivals with both text-to-motion and audio-to-motion methods in terms of metrics measuring body motion quality and face motion accuracy.

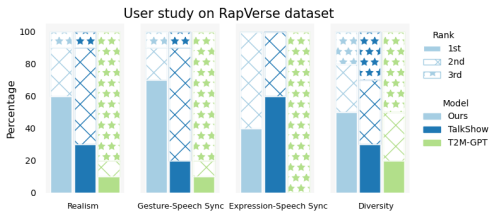

Figure 4. **User study** on generated motion.

Furthermore, the cornerstone of our approach lies in the simultaneous generation of vocals and motion, aiming to achieve temporal alignment between them. This objective is substantiated by our competitive results on the BC metric, which assesses the synchrony between singing vocals and corresponding motions, underscoring our success in closely synchronizing the generation of these two modalities. For the cascaded system, we integrate the text-to-vocal model DiffSinger with the audio-to-motion model Talkshow. Compared with the cascaded system, our joint-generation pipeline demonstrates superior outcomes while also reducing computational demands during both training and inference phases. In the cascaded architectures, errors tend to accumulate through each stage. Specifically, if the text-to-vocal module produces unclear vocals, it subsequently hampers the audio-to-motion model's ability to generate accurate facial expressions that align with the vocal content.

To better evaluate the generation quality, we conducted an additional user study (see Fig. 4) using ten 10-second-long videos from the RapVerse dataset, sampled from the test set. We recruited 20 participants to evaluate four aspects: the holistic realism of human motion, expression-vocal synchronism, gesture-vocal synchronism, and holistic diversity, using real data as a reference. Participants

Method	FID↓	DIV↑	BC↑	MSE↓	LVD↓
Text-to-Motion					
T2M-GPT [66]	23.45	11.75	-	-	-
MLD [7]	26.34	12.15	-	-	-
Audio-to-Motion					
Habibie et al. [18]	32.14	10.08	0.476	2.13	9.54
Talkshow [65]	18.23	13.14	0.482	2.05	9.20
EMAGE [33]	21.18	12.65	**0.488**	**1.96**	8.45
MotionCraft [4]	17.75	13.76	0.482	2.06	9.23
Cascaded Result	23.42	12.87	0.479	2.09	9.38
Text-to-Audio+Motion					
Ours	**17.58**	**14.08**	0.485	2.03	**7.23**

(a) Motion Generation

Method	MOS↑ †
GT	4.45 ± 0.06
Reconstruction	4.02 ± 0.08
FastSpeech2 [49]	3.41 ± 0.18
DiffSinger [34]	3.72 ± 0.12
Ours	3.64 ± 0.15

(b) Vocal Generation

Table 2. **Quantitative results of generated motion and vocal.** (a) compares with different motion generation baselines, best result is shown in **bold** and the second-best result is shown in underline. (b) compares with text-to-vocal generation baselines. †: The MOS is calculated with 95% confidence intervals of song samples.

Method	FID↓	DIV↑	BC↑	MSE↓	LVD↓
GT	0	14.53	0.499	0	0
Pretrained LLM	48.25	12.65	0.462	3.28	12.15
Single motion token	19.15	12.75	0.477	2.18	10.12
Ours	**17.58**	**14.08**	**0.485**	**2.03**	**7.23**

Table 3. **Ablation study.** We compare with common designs in motion generation frameworks.

were asked to sort shuffled videos generated by different methods based on these criteria. As shown in the results, our model achieved the highest scores in human motion realism, gesture-vocal synchronism, and holistic diversity when compared with T2MGPT (text-to-motion) and Talkshow (audio-to-motion).

Evaluations on vocal generations. We compare our method against other state-of-the-art text-to-vocal generation baselines in Table. 2 (b). While our unified model is trained to simultaneously generate vocals and motion, a task considerably more complex than generating vocals alone, its vocal generation component still manages to achieve results comparable to systems designed solely for vocal generations.

5.3. Ablation Study

We present the outcomes of our ablation study in Table. 3. Initially, we explored the integration of a pre-trained large language model [46] for multi-modality generation, akin to the approach in [26]. However, the efficacy of utilizing pre-trained language models significantly lags behind our tailored design, underscoring that pre-training primarily on linguistic tokens does not facilitate effective prediction across multiple modalities like vocal and motion. Additionally, we study the impact of our compositional VQ-VAEs on motion generation. In contrast, a baseline employing a single VQVAE for the joint quantization of facial, body, and hand movements was implemented. This approach led to a noticeable degradation in performance, particularly marked by a **-2.89** decrease in LVD. This decline can be attributed to the preponderance of facial movements in a singer's performance. Using a single VQ-VAE model for full-body dynamics compromises the detailed representation of facial expressions, which are crucial for realistic and coherent motion synthesis.

6. Conclusion

In this work, we present a new framework for the simultaneous generation of 3D whole-body motions and singing vocals, directly from textual lyrics. To address this challenging task, we first collect RapVerse, a large dataset containing synchronous rap vocals, alongside lyrics and 3D whole-body motions. Utilizing RapVerse, we demonstrate that simply scaling autoregressive transformers across language, audio, and motion yields a coherent generation of singing vocals and 3D holistic human motions. We anticipate that this work will inspire novel avenues in the joint modeling of text, audio, and motion.

References

[1] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023. 1

[2] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, pages 719–728. IEEE, 2019. 3

[3] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–20, 2023. 3

[4] Yuxuan Bian, Ailing Zeng, Xuan Ju, Xian Liu, Zhaoyang Zhang, Wei Liu, and Qiang Xu. Adding multi-modal controls to whole-body human motion generation. *arXiv preprint arXiv:2407.21136*, 2024. 3, 7, 8

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2, 3

[6] Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7361, 2024. 3

[7] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 1, 2, 3, 7, 8

[8] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. 2, 3

[9] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 5

[10] Jeff Donahue, Sander Dieleman, Mikołaj Bińkowski, Erich Elsen, and Karen Simonyan. End-to-end adversarial text-to-speech. *arXiv preprint arXiv:2006.03575*, 2020. 2, 3

[11] Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye Wang. The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–9. IEEE, 2013. 2, 3, 4

[12] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591–598, 2010. 2, 3

[13] Ylva Ferstl and Rachel McDonnell. Investigating the use of recurrent motion modelling for speech gesture generation. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 93–98, 2018. 3

[14] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1396–1406, 2021. 3

[15] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 2, 3, 4

[16] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 1, 2, 3, 4, 7

[17] Chuan Guo, Xinxin Zuo, Sen Wang, and Li 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. 3

[18] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 101–108, 2021. 2, 3, 7, 8

[19] Jinzheng He, Jinglin Liu, Zhenhui Ye, Rongjie Huang, Chenye Cui, Huadai Liu, and Zhou Zhao. Rmssinger: Realistic-music-score based singing voice synthesis. *arXiv preprint arXiv:2305.10686*, 2023. 3

[20] Georg Heigold, Ignacio Moreno, Samy Bengio, and Noam Shazeer. End-to-end text-dependent speaker verification. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5115–5119. IEEE, 2016. 5

[21] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154, 2020. 4

[22] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154, 2020. Deezer Research. 4

[23] Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: How much can a bad teacher benefit asr pre-training? In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6533–6537. IEEE, 2021. 5

[24] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al. Noise2music: Text-conditioned music generation with diffusion models. *arXiv preprint arXiv:2302.03917*, 2023. 1

[25] Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. Multi-singer: Fast multi-singer singing

voice vocoder with a large-scale corpus. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3945–3954, 2021. 2, 3, 4

[26] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 6, 8

[27] Kavita Kasi and Stephen A Zahorian. Yet another algorithm for pitch tracking. In *2002 ieee international conference on acoustics, speech, and signal processing*, pages I–361. IEEE, 2002. 5

[28] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8255–8263, 2023. 3

[29] Sungjae Kim, Yewon Kim, Jewoo Jun, and Injung Kim. Muse-svs: Multi-singer emotional singing voice synthesizer that controls emotional intensity. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023. 3

[30] George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 2003 edition, 1980. First published January 1, 1980. 2

[31] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 7

[32] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 4

[33] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via masked audio gesture modeling. *arXiv preprint arXiv:2401.00374*, 2023. 2, 3, 7, 8

[34] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11020–11028, 2022. 1, 2, 3, 4, 7, 8

[35] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978*, 2023. 1, 3, 6

[36] Shuhong Lu, Youngwoo Yoon, and Andrew Feng. Co-speech gesture synthesis using discrete gesture token learning. *arXiv preprint arXiv:2303.12822*, 2023. 2, 4

[37] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 2, 3, 4

[38] Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. Meta-stylespeech: Multi-speaker adaptive text-to-speech generation. In *International Conference on Machine Learning*, pages 7748–7759. PMLR, 2021. 2

[39] Itsuki Ogawa and Masanori Morise. Tohoku kiritan singing database: A singing database for statistical parametric singing synthesis using japanese pop songs. *Acoustical Science and Technology*, 42(3):140–145, 2021. 4

[40] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016. 2, 3

[41] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2, 4

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

[43] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2, 3, 4

[44] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. Speech resynthesis from discrete disentangled self-supervised representations, 2021. 2, 5, 6

[45] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 4

[46] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 6, 8

[47] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 4

[48] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32, 2019. 3

[49] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020. 3, 7, 8

[50] Yi Ren, Xu Tan, Tao Qin, Jian Luan, Zhou Zhao, and Tie-Yan Liu. Deepsinger: Singing voice synthesis with data mined from the web. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1979–1989, 2020. 4

[51] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. Mo\^ usai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*, 2023. 1

[52] Bidisha Sharma, Xiaoxue Gao, Karthika Vijayan, Xiaohai Tian, and Haizhou Li. Nhss: A speech and singing parallel database. *Speech Communication*, 133:9–22, 2021. 2, 3, 4

[53] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized talking head synthesis. *arXiv preprint arXiv:2301.03786*, 2023. 3

[54] Kenta Takeuchi, Souichirou Kubota, Keisuke Suzuki, Dai Hasegawa, and Hiroshi Sakuta. Creating a gesture-speech dataset for speech-based automatic gesture generation. In *HCI International 2017–Posters' Extended Abstracts: 19th International Conference, HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part I 19*, pages 198–202. Springer, 2017. 3

[55] Hiroki Tamaru, Shinnosuke Takamichi, Naoko Tanji, and Hiroshi Saruwatari. Jvs-music: Japanese multispeaker singing-voice corpus. *arXiv preprint arXiv:2001.07044*, 2020. 3, 4

[56] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. A deep learning approach for generalized speech animation. *ACM Transactions On Graphics (TOG)*, 36(4):1–11, 2017. 3

[57] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 4

[58] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3

[59] Jean-Marc Valin and Jan Skoglund. Lpcnet: Improving neural speech synthesis through linear prediction. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5891–5895. IEEE, 2019. 2

[60] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[61] Yu Wang, Xinsheng Wang, Pengcheng Zhu, Jie Wu, Hanzhao Li, Heyang Xue, Yongmao Zhang, Lei Xie, and Mengxiao Bi. Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis. *arXiv preprint arXiv:2201.07429*, 2022. 4

[62] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, et al. Multiface: A dataset for neural face rendering. *arXiv preprint arXiv:2207.11243*, 2022. 2, 3

[63] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023. 7

[64] Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 492–498. IEEE, 2021. 2

[65] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023. 2, 3, 4, 7, 8

[66] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052*, 2023. 2, 7, 8

[67] Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, et al. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. *Advances in Neural Information Processing Systems*, 35:6914–6926, 2022. 2, 4

[68] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

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

[70] Yu Zhang, Rongjie Huang, Ruiqi Li, JinZheng He, Yan Xia, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. Stylesinger: Style transfer for out-of-domain singing voice synthesis. *arXiv preprint arXiv:2312.10741*, 2023. 3

[71] Zewang Zhang, Yibin Zheng, Xinhui Li, and Li Lu. Wesinger: Data-augmented singing voice synthesis with auxiliary losses. *arXiv preprint arXiv:2203.10750*, 2022. 3