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

Supplementary Material

## 1. Supplemental material

The supplementary material is organized as follows: Sec. 1.1 provides supplemental videos to show additional qualitative results; Sec. 1.2 presents additional details of the network architectures; Sec. 1.3 introduces evaluation metrics; Sec. 1.4 provides statistical characteristics of RapVerse dataset; Sec. 1.5 presents additional ablation studies; Sec. 1.6 shows additional qualitative results; Sec. 1.7 discusses limitations of the work; Sec. 1.8 discusses broader societal impacts of the work.

### 1.1. Website Demo

In order to provide more vivid and clear qualitative results, we make a supplemental website demo to demonstrate the generation quality of our proposed system. We encourage readers to view the results at https://jiabenchen.github.io/RapVerse/.

### 1.2. Implementation Details

**Motion Tokenizer.** We train three separate Vector Quantized Variational Autoencoders (VQ-VAE) for face, body and hand, respectively. We adopt the same VQ-VAE architecture based on [2, 5, 18]. For loss functions optimizing the motion tokenizers, we use the L1 smooth reconstruction loss, the embedding loss, and the commitment loss. The commitment loss weight is set to 0.02. In line with [2, 5], strategies such as the exponential moving average and the codebook reset technique [14] are implemented to optimize codebook efficacy throughout the training process. We take 512 for the codebook size and set the dimension of each code to 512. We set the temporal down-sampling rate to 4. We train the VQ-VAEs with a batch size of 256 and a sequence window length of 72. We adopt Adam with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and a learning rate of 0.0002 as the optimizer.

**Vocal Tokenizer.** For the semantic encoder, we adopt a BASE 12 transformer of HuBERT [4] pre-trained on the 969-hour LibriSpeech corpus [11]. Following [6, 12], we derive the feature activations from its sixth layer. This process allows the HuBERT model to transform input audio into a 768-dimensional vector space. Subsequently, we employ the k-means algorithm with 500 centroids to get quantized discrete content codes. For the F0 encoder, a VQ-VAE framework is utilized to discretize the F0 signal into quantized F0 tokens. We adopt the Exponential Moving Average updates during training the VQ-VAE following [1, 12]. We set the codebook size of the VQ-VAE to 20 entries. Moreover, as the original work directly normalizes the extracted F0 values for each singer respectively, we don't explicitly use the singer's statistics but adopt a windowed convolutional layer with both the audio input (sliced into the window size) and singer's embedding as input. Finally, we adopt a similar architecture as [3] for the singer encoder.

**General Auto-regressive Model.** The auto-regressive model consists of a T5 Embedder and a Foundation Model. We use a T5-Large Encoder as our Embedder, with 24 layers and 16 heads. The Embedder is freezed during foundation model training. The foundation model is based on the Decoder-only transformer architecture, which has 12 layers and 8 heads. We use Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and a learning rate of 0.0002. We do not use dropout in our training. Our training batch size is 384 for 100 epochs.

### 1.3. Evaluation Metrics

To evaluate the motion generation quality, we utilize the following metrics:

1. **Frechet Inception Distance (FID)**: This metric measures the distribution discrepancy between the ground truth and generated motions of body and hand gestures. Specifically, we train an autoencoder based on [2] as the motion feature extractor.
2. **Diversity (DIV)**: DIV evaluates the diversity of generated motions, where we calculate the variance from the extracted motion features.
3. **Vertex MSE**: Following [16], we compute the mean L2 error of lip vertices between generated face motions and ground truth.
4. **Landmark Velocity Difference (LVD)**: Introduced by [17], LVD calculates the velocity difference of generated facial landmarks and ground truth.
5. **Beat Constancy (BC)** [7]: BC evaluates the synchrony of generated motions and singing vocals by calculating the similarity between the rhythm of gestures and audio beat. Specifically, we extract audio beats using librosa [9], and we compute the kinematic beats as the local minima of the velocity of joints. Then the alignment score is derived from the mean proximity of motion beats to the nearest audio beat.

For the evaluation of singing vocal generation quality, the Mean Opinion Score (MOS) is employed. It reflects the perceived naturalness of the synthesized vocal tones, with human evaluators rating each sample on a scale from 1 to 5, thereby offering a subjective measure of vocal synthesis fidelity.

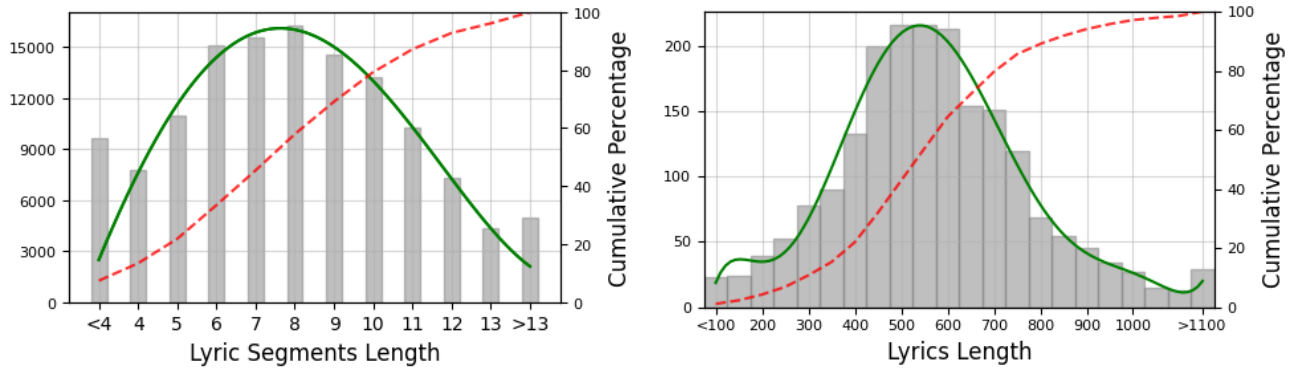


Figure 1. **Lyric length distribution** in RapVerse, from segmented clips (left), and original videos (right).

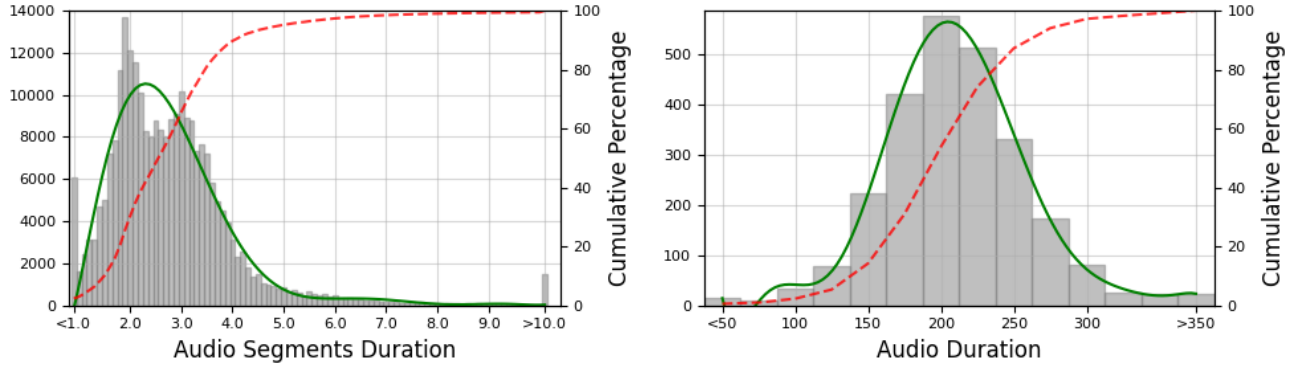


Figure 2. **Audio duration distribution** in RapVerse, including segmented audio clips (left), and original audios (right).

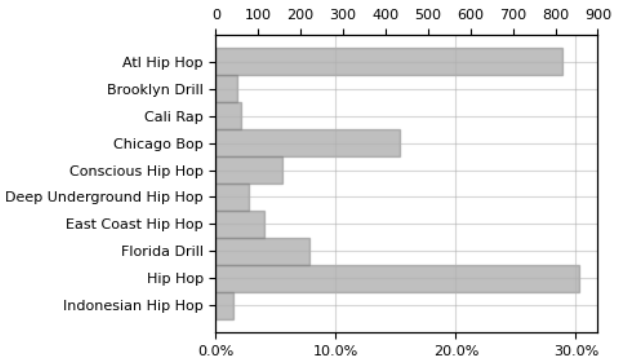


Figure 3. **Genre distribution** of data in RapVerse. We show the distribution of different genres under the rap genre.

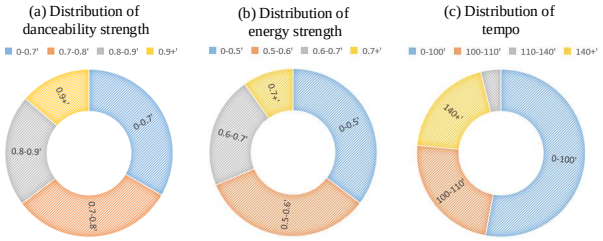


Figure 4. **Additional distribution** of data in RapVerse. We show the distribution of danceability, energy and tempo in our dataset.

## 1.4. Dataset Characteristics

We provide additional statistical characteristics of the overall RapVerse dataset. Fig. 1 and Fig. 2 show the distribution of lyric lengths and track durations, respectively. For lyric lengths and durations, we provide statistics for both the original soundtracks and the segmented clips in our dataset.

Since all audio in our dataset falls under the general rap genre, we calculated the distribution of different subgenres within rap in Fig. 3. Following Music4All dataset [15], we additionally calculate distributions for audio attributes including tempo (speed of the song measured in beats per minute), danceability (a value from 0.0 to 1.0 representing how suitable the song is for dancing) and energy (a value from 0.0 to 1.0 representing the song's intensity and activity) in Fig. 4. All these metadata for each song will be released for further research use.

| Design | MPJPE↓ | PAMPJPE↓ | ACCL↓ |
|---|---|---|---|
| K = 256 | 63.4 | 39.1 | 9.14 |
| K = 512 | 60.9 | 35.4 | 8.90 |
| K = 1024 | 62.2 | 36.8 | 9.08 |
| Single | 65.2 | 38.8 | 9.32 |
| Ours | **60.9** | **35.4** | **8.90** |

Table 1. **Evaluation of our motion tokenizer.** We follow [8] to evaluate the motion reconstruction errors of our motion VQVAE model $V_m$. MPJPE and PAMPJPE are measured in millimeters. ACCL indicates acceleration error.

| Method | CER↓ | GPE↓ | VDE↓ |
|---|---|---|---|
| GT | 41.88 | - | - |
| K = 100 | 76.39 | 2.64 | 11.98 |
| K = 500 (**Ours**) | 69.21 | **2.29** | **8.84** |
| K = 2000 | **67.46** | 2.54 | 9.08 |
| Original F0 post-process | 68.93 | 2.71 | 9.19 |

Table 2. **Evaluation of our unit2wav model.** We follow [12] to evaluate the speech resynthesis errors of our unit2wav model. WER, GPE, and VDE, expressed as percentages, indicate the character error rate, the grand pitch error and the voicing decision error.

## 1.5. Additional Ablation Studies

**Ablation on Motion Tokenizer.** We study different designs of our motion tokenizer by comparing the reconstruction results. Specifically, we explore VQ-VAEs with different codebook sizes, and study the effect of using a single VQ-VAE to model full-body motions instead of multiple VQ-VAEs for different body parts. As is demonstrated in Tab. 1, we find that using separate VQ-VAEs for face, body and hands has lower reconstruction error. And we select a codebook size of 512 for our final model.

**Ablation on Vocal Tokenizer.** We also study different designs for our audio tokenizer by comparing the reconstruction results. Specifically, we explore different codebook sizes for the semantic encoder by changing the K-Means number. We also compare the effect with our singer embedding in F0 value postprocessing. We use the following metrics to measure the reconstruction quality of the vocal tokenizer:

1. **Character Error Rate (CER)**: We use Whisper [13] to transcribe the ground truth and synthesized audios, and then take the corresponding ground truth lyrics as the reference to calculate the CER of the synthesized audios.
2. **Gross Pitch Error (GPE)**: The percentage of pitch estimates that have a deviation of more than 20% around the true pitch. Only frames considered pitched by both the ground truth and the synthesized audio are considered.
3. **Voicing Decision Error (VDE)** [10]: The portion of frames with voicing decision error, that is, the results using ground truth and synthesized audio to determine whether the frame is voiced are different.

The analysis of the results indicates that even the original audio exhibits a high CER, which could be attributed to the rapid speech rate associated with rapping. In some instances, the lyrics may not be distinctly recognizable even by human listeners. Upon comparing different codebook sizes, it is observed that they achieve comparable GPE values. This similarity in GPE is expected since the same F0 model is employed across all codebook sizes. The CER, which serves as a direct measure of the semantic information preserved in the code, suggests that larger codebooks tend to retain more semantic information. However, the difference in CER between codebook sizes of $K = 500$ and $K = 2000$ is minimal. Given that $K = 500$ demonstrates better GPE and VDE, we select $K = 500$.

Additionally, we ablate a design without the singer embedding in the F0 preprocessing, instead normalizing the F0 values for each singer. It shows that this approach resulted in significantly inferior performance, particularly in pitch prediction, compared to the modified version that includes the singer embedding.

| Method | MOS↑ | FID↓ | DIV↑ | BC↑ | MSE↓ | LVD↓ |
|---|---|---|---|---|---|---|
| *Text-to-Speech* | | | | | | |
| FastSpeech2 | 2.37 ± 0.14 | - | - | - | - | - |
| *Speech-to-Motion* | | | | | | |
| Talkshow | - | 25.21 | 11.85 | 0.485 | 2.06 | 9.31 |
| *Text-to-Vocal+Motion* | | | | | | |
| **Ours** | 3.64 ± 0.15 | 17.58 | 14.08 | 0.485 | 2.03 | 7.23 |

Table 3. **Comparison with models trained on speech or motion data.** We compare inference results with our model against those from FastSpeech2, trained on a standard speech corpus, calculating the Mean Opinion Score (MOS) with 95% confidence intervals for song samples. Additionally, we compare results with the Talkshow model, trained on speech-to-motion data, to evaluate our model's performance.

**Ablation on benefits of RapVerse dataset.** We conduct additional experiments, as shown in Table. 3, to demonstrate the benefits of using our RapVerse dataset. The results show that models trained on normal speech audio cannot produce convincing rap motion sequences and vocals. Specifically, text-to-speech models trained only on speech corpora are unable to generate plausible singing vocals, as measured by the MOS metric, compared to our model. Additionally, audio-to-motion models trained on speech-to-motion dataset fail to produce natural body singing gestures, as measured by the FID and DIV metrics. This is due to the inherent differences in motion, rhythm, and expressiveness

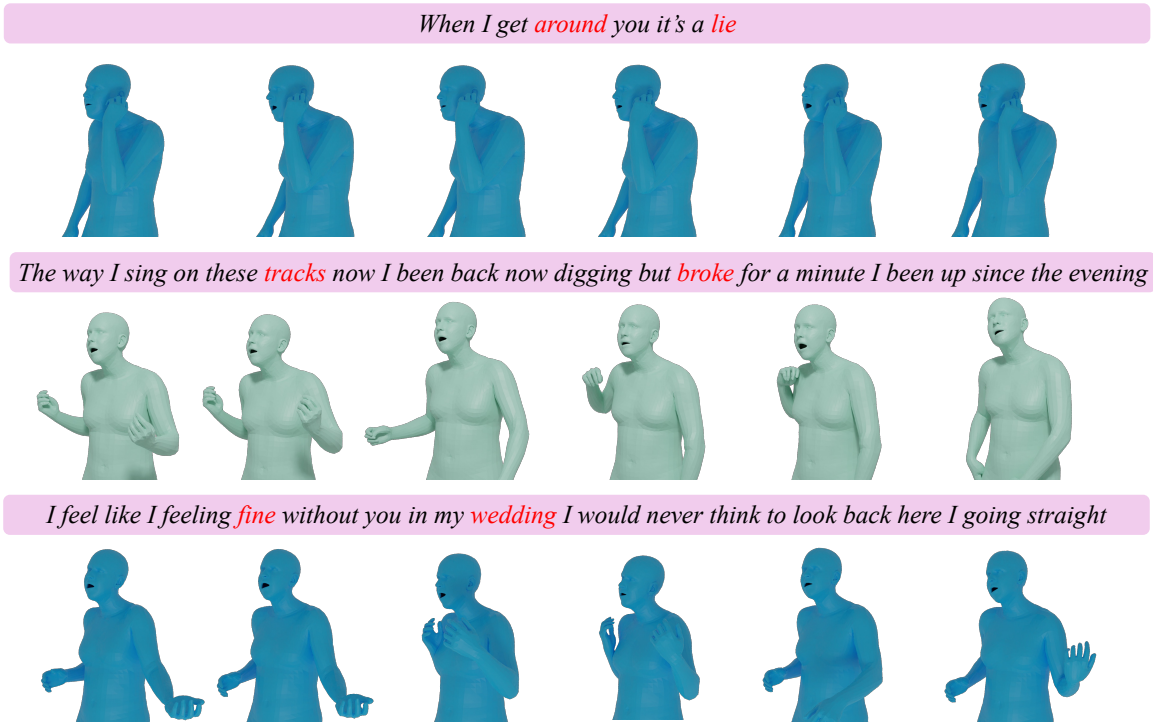


Figure 5. **Additional qualitative results.** Our method can generate diverse whole-body motions from input lyrics.

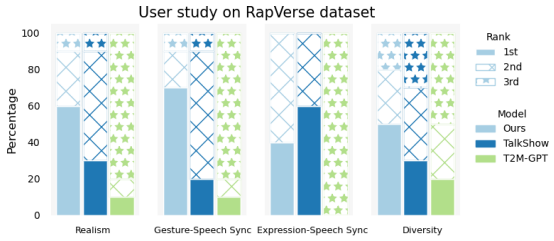


Figure 6. **Additional user study** on RapVerse dataset. The figure demonstrates user preference percentage in terms of four aspects: holistic realism of human motion, gesture-vocal synchronism, expression-vocal synchronism, and holistic motion diversity.

between singing and speaking. These findings highlight the necessity of our dataset, which opens up new opportunities for research in generating coordinated lyrics-driven motions and vocals. Furthermore, we plan to make our dataset publicly available to facilitate further advancements in this promising field.

## 1.6. Additional Qualitative Results

We show additional qualitative results in Fig. 5. Our model adeptly generates comprehensive whole-body motions that embody the essence of the input lyrics. These include authentic gesture movements that resonate with the song's rhythm and synchronized lip motions that articulate the lyrics.

## 1.7. Limitations

Currently, our method does not generate root motions; we focus solely on local body motions, including body gestures and facial expressions. In the future, we plan to explore generating root motions to provide a more comprehensive representation of full-body dynamics. Regarding voice selection, our framework allows for choosing from over a hundred different singer embeddings, offering a wide range of vocal timbres. We aim to expand this selection further in future work, allowing for even greater diversity and customization in voice generation. At this stage, our model does not support selecting specific rapping styles or tempos based on input lyrics. We recognize the importance of this feature and plan to investigate methods to incorporate style and tempo selection to enhance the flexibility and expressiveness of our system.

## 1.8. Broader Impacts

This research contributes to advancements in generating synchronized vocals and human motion from textual lyrics, aiming to enhance virtual agents' ability to provide immersive and interactive experiences in digital media. The potential positive impact of this work lies in its ability to create more lifelike and engaging virtual performances, such as in virtual concerts and gaming, where characters can per-

form and react in ways that are deeply resonant with human expressions. This could significantly enhance user engagement in virtual reality settings, and provide innovative solutions in entertainment industries.

However, this capability carries inherent risks of misuse. The technology's ability to generate realistic human-like actions and singing vocals from mere text raises concerns about its potential to create misleading or deceptive content. For example, this could be exploited to produce fake videos or deepfakes where individuals appear to sing and perform that never actually occurred, which could be used to spread misinformation or harm reputations. Recognizing these risks, it is crucial to advocate for ethical guidelines and robust frameworks to ensure the responsible use of such technologies.

## References


[1] 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. 1

[2] 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. 1

[3] 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. 1

[4] 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. 1

[5] 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

[6] Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021. 1

[7] 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. 1

[8] 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. 3

[9] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *SciPy*, pages 18–24, 2015. 1

[10] Tomohiro Nakatani, Shigeaki Amano, Toshio Irino, Kentaro Ishizuka, and Tadahisa Kondo. A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments. *Speech Communication*, 50(3):203–214, 2008. 3

[11] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015. 1

[12] 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. 1, 3

[13] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. 3

[14] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 1

[15] Igor André Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Catharin, Rafael Biazus Mangolin, Valéria Delisandra Feltrim, Marcos Aurélio Domingues, et al. Music4all: A new music database and its applications. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 399–404. IEEE, 2020. 2

[16] 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. 1

[17] 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. 1

[18] 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. 1