HoloSinger: Semantics and Music Driven Motion Generation with Octahedral Holographic Projection

Zeyu Jin
Zixuan Wang
Qixin Wang
Tsinghua University
Beijing, China
13269992899@163.com

Jia Jia*
Tsinghua University
Beijing National Research Center for Information Science and Technology
Beijing, China
jjia@tsinghua.edu.cn

Ye Bai
Yi Zhao
Hao Li
Xiaorui Wang
Kuaishou Technology
Beijing, China
zhaoyi07@kuaishou.com

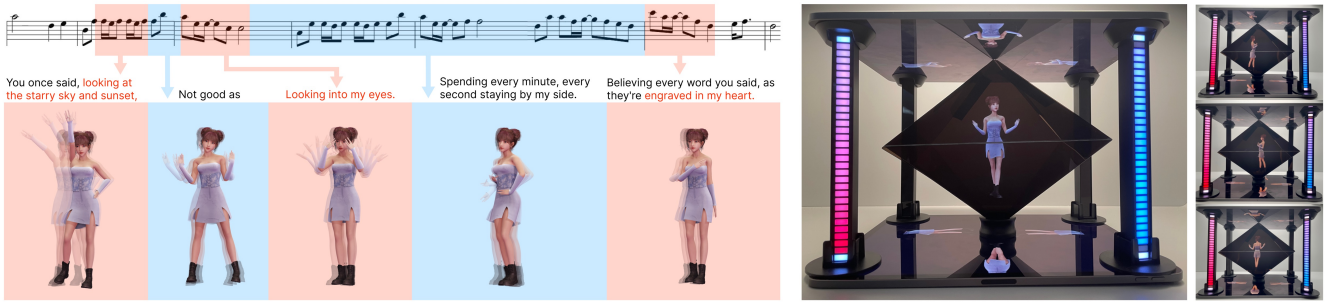

Figure 1: Left: Our model automatically generates motion based on the semantics or music modality of a song. Right: A holographic projection system displaying the singing avatar, simultaneously providing four views with an octahedral structure.

ABSTRACT

Lyrics and music are both significant for a singer to perform a song. Therefore, it is important in singer's motion generation to model both semantic and acoustic correlation with motions at the same time. In this paper, we propose HoloSinger, a novel comprehensive system that synthesizes singing motions according to the given song. Additionally, we present singing avatar with octahedral holographic projection. For singing motion generation, we introduce a Transformer-VAE generative model to decompose lyrics and music, then fuse their impacts to synthesize singer's motions. Extensive experiments and user studies show that our method automatically generates realistic motions that adhere to musical choreography and reflect the lyric semantics appropriately. Furthermore, we design a desktop-level holographic projection device with an octahedral structure. It achieves high-definition holographic projection effects with smaller volume, larger imaging area ratio, and the ability of real-time AI interaction.

*The corresponding author

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612674

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *Interaction paradigms*;

KEYWORDS

dance synthesis, text2motion, holographic projection system

ACM Reference Format:
Zeyu Jin, Zixuan Wang, Qixin Wang, Jia Jia*, and Ye Bai, Yi Zhao, Hao Li, Xiaorui Wang. 2023. HoloSinger: Semantics and Music Driven Motion Generation with Octahedral Holographic Projection. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3581783.3612674

1 INTRODUCTION

Music express the soul, while lyrics articulate its tale. However, most existing methods consider barely music or text-conditioned motion generation for a song, which is insufficient in creating a fully developed singing avatar. To enhance a singer's performance, we divide body movements into two categories: 1) *Music Driven Choreography*, which involves movements synchronized with music beats, or conveying emotional feelings of the song. 2) *Semantic gestures related to specific lyrics*. Hence, this work aims to address and balance these two types of actions within a single multi-modality system. To achieve this goal, we employ a Transformer-VAE generative model and incorporates the embeddings from two modalities before a joint decoder.

In order to demonstrate the singing avatar performance with a panoramic view, we introduce the holography technique. Compared to the widely used VR/AR system, holography provides 3D visual experience in physical space without additional wearable devices. Although ideal autostereoscopy is currently not feasible, many devices on the market provide general holographic capability. Pyramid-shaped holographic display cabinets, commonly seen in museums, are a notable example. Such displays are lively but typically bulky, playing pre-recorded videos with relatively confined central viewing areas. Following the same basic principle, we devised an octahedral holographic projection device to simultaneously showcase four different angles of the avatar's movements. Besides, our device offers high-definition visual effects with smaller size, larger imaging area, and the ability to engage in real-time AI interaction. To summarize, our contributions are three-fold:

- We propose HoloSinger, a novel comprehensive system to generate and display avatar motions with input songs.
- We devise a Transformer-VAE based model to automatically synthesize singing motion by decomposing acoustic and semantic features from song and fusing their effects to motions.
- We introduce an octahedral holographic projection technique to demonstrate the real-time holographic avatar singing performance with an improved imaging structure.

2 ALGORITHM FRAMEWORK

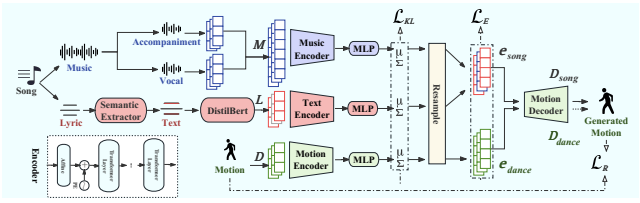

Figure 2: System Framework of Motion Generation System

As shown in Figure2, our model consists of three modality-specific encoders to map source codes to a joint latent space, and a motion decoder to reconstruct motion sequence from embeddings.

2.1 Music Modality

We used the paired data $M \in \mathbb{R}^{T_m \times Dim_m}$ and $D \in \mathbb{R}^{T_d \times Dim_d}$ to train $Encoder_{music}$, $Encoder_{dance}$ and $Decoder_{dance}$ simultaneously. The parameters of the networks are learned via reconstruction loss (L_R), Kullback-Leibler divergences loss (L_{KL}) and cross-modal embedding similarity loss (L_E).

The music modality model was trained on a 3D singing motion dataset with 3.7 hours data paired with lyrics, music, and motions.

2.2 Text Modality

We implemented a pipeline of semantic words selection from the prescribed lyrics. The singing motion dataset was labeled with 43 motion categories. For each line of input lyrics, we extract keywords, and relate them to Top3 similar motion label categories as the target semantic motions. With the help of ASR tool MFA [2], we get the timeline of each word and conduct frame by frame replacement on the music driven results at the point of target keywords. Cubic spline interpolation is utilized to make the adjustment smooth.

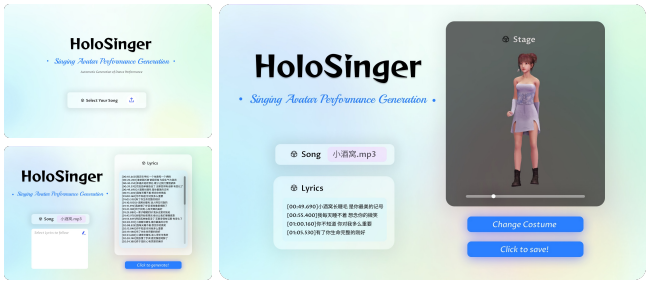

Figure 3: User Interface

Text modality model shares basically the same structure with music modality. With limited labeled semantic motions, we firstly trained the model's text branch on BABEL [3] (a large annotated action dataset), then fine-tuned it with finely labeled fragments.

3 EXPERIMENTS AND USER STUDY

We presented 27 users with 7 pieces of song clips. Participants were asked to use the Likert 5-point scale ('5' equals satisfied) to evaluate our synthesized singing motions from three aspects:

- **Motion quality.** Our method got 3.84 (standard deviation= 0.82) , which indicates that it outputs realistic motions.
- **Motion style differences between verse and chorus.** Over 68% users preferred us, showing that our method can better interpret emotional changes along with the song.
- **Motion consistency with lyrics.** We got 3.80 (standard deviation=0.90), which implies that the generated motions are semantically appropriate.

Besides, our model has equivalent effects comparing with Bailando [4] (SOTA of Music2Motion) on objective evaluation indicators like FID↓ (**4.94** v.s. 5.21) and Beat Align Score [1]↑ (**0.2313** v.s. 0.1956), showing its competence for music driven choreography.

4 DEMOSTRATION

As revealed in Figure3, the input user interface of our system helps conduct efficient interactive tasks, including avatar costume exchange, song and lyrics generation assignment.

As shown in Figure1 on the right, we developed a holographic projection device for singing avatar performance. With an octahedral structure, our holo system features two electronic screens acting as light sources from top and bottom. Having solved the issue of display blending on the octahedron edges, the device can achieve high-definition holographic projection effects at desktop levels with a smaller volume and a larger imaging area ratio, distinguishing it from the common pyramid-shape device. Its interactive capability breaks away from traditional practices where holographic projection products merely act as prefabricated content players.

5 CONCLUSION

This article proposes a comprehensive system for 3D avatar motion generation and practical demonstration. It holographically presents the singing avatar driven by random song and lyrics. Motions generated by our method are visually realistic and semantically appropriate. Attempts to conduct entity interaction with holography technique open up new possibilities for this technology.

REFERENCES

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

[2] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi.. In *Interspeech*, Vol. 2017. 498–502.

[3] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. 2021. BABEL: bodies, action and behavior with English labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 722–731.

[4] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. 2022. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11050–11059.