Detailed Project Plan

--Xu, Shuyang (UID: 3035947740)

Project Topic: Human Motion Planning in Spatial Audio Field.

1. Background

For a long time, methods of creating human motions have been largely based on keyframe animation, motion capture, stop motion animation, and so forth. However, these methods are often very labor intensive and time-consuming. This inefficiency becomes more serious as the number of required motions increases. Even for characters that require only simple reactions rather than intricate movements to some given circumstances, animators have to invest significant time in creating and refining the animations.

There do exist a few ways to lower the cost of producing a large number of animations to some extent. For instance, a rule-based animation system, which involves establishing rules for characters and guiding their actions based on these rules, can effectively animate a great number of characters, particularly non-player characters (NPCs). However, there are still several drawbacks to this approach. The animations produced by this method are often monotonous, as they all adhere to the same rules. Furthermore, as more and more diverse motions are required, the rules may become overly complicated, making the process of establishing, adjusting, and unifying such a set of rules costly and time-consuming.

In such a context, with the explosive development of machine learning and artificial intelligence (AI), the state-of-the-art research now revolves around the fusion of machine learning with human motion. Instead of producing animation frame by frame, animators can now simply input specific scenarios in text, audio, or other formats into a pre-trained model, and the model will subsequently generate appropriate animations according to the given conditions. Animators thus only need to select the most suitable one from a range of motions and can apply it with minor modifications. The utilization of machine learning, which not only elevates the diversity of the motions but also significantly saves time and cost, has become the prevailing trend in the animation and game industries.

There are currently a lot of studies on using machine learning to generate human motion based on given conditions. Tevet et al. (2022) developed the Montion Diffusion Model (MDM), which is a model capable of synthesizing human motion with high fidelity from given textual descriptions or action classes using the diffusion process—a method that has been proved to be effective in generating pictures in the field of

computer vision. Based on the Motion Diffusion Model (MDM), Zhou at al. (2023) put forward the Efficient Motion Diffusion Model (EMDM), which significantly accelerates the generation of the motions by allowing much fewer diffusion steps. Zhang et al. (2023) innovatively combined Vector Quantized-Variational AutoEncoder (VQ-VAE) with Generative Pretrained Transformer (GPT) to synthesize human motion based on textual descriptions of the requirements and proved that this approach is very competitive against diffusion.

In addition to various studies on text-to-motion human motion synthesis, there is also a wealth of research focusing on music- or audio-to-motion human motion synthesis. Siyao et al. (2022) proposed the music-to-dance framework Bailando—a combination of a choreographic memory used for summarizing 3D pose sequences to a quantized codebook and a Generative Pretrained Transformer (GPT)—to synthesize dance motions that have high temporal coherency with the given music as well as high fluency. Zhuang et al. (2022) also applied an autoregressive generative model, DanceNet, to synthesize dance motions that have high realism and diversity, as well as excellent consistency with the style, rhythm, and melody of the music.

Despite a number of studies on text-to-motion and music-to-motion frameworks, as I have mentioned in previous paragraphs, research in human motion synthesis in the spatial audio field still remains a blank paper. For the present models, characters know how to dance given music signals, but their motions cannot jump out of this box. Music is only a small proportion of the definition of audio, not to mention that the direction of arrival of the audio and the distance between the audio source and the character should all have significant influence on human motions. For example, a character should react differently to a bomb from far away and nearby. This research gap is gaining more and more attention these years. Wang et al. (2024) proposed a geometry-based model which can effectively predict the room impulse response (RIR) at novel points in a room, and also established a dataset for it. Nevertheless, the spatial acoustic field prediction models have not been integrated to human motion synthesis yet. This project will put forward a new baseline to complement this gap in human motion synthesis, from creating a new dataset to fitting a model to resolve the difficulty of generating human motion in the spatial audio field.

2. Objective

The objective of this project mainly focuses on the following points:

1) In the area of machine learning and human motion synthesis, none of the present datasets includes both spatial information of the audio and corresponding human motion. Currently, the most satisfying dataset is AIST++ proposed by Li et al. (2021), which is a dataset containing 5.2 hours of 3D dance motions covering 10

genres. However, this dataset still includes no spatial information for the audio, and the dance motions are not related to the spatial position of the audio as well. Therefore, one of the main purposes of this project is to build such a dataset to make it feasible to train a model that can synthesize human motion with high fidelity given a spatial audio field.

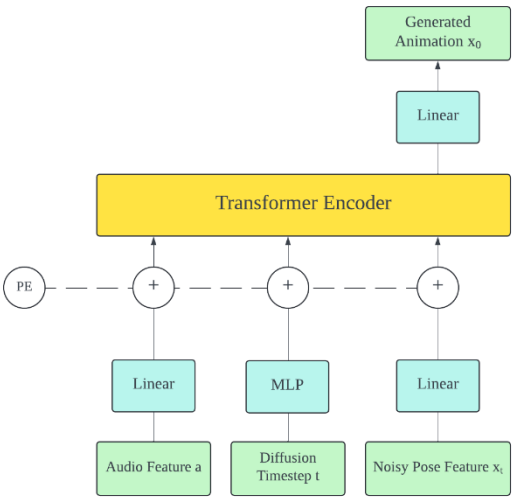
2) As this is the first attempt to use machine learning methods to generate human motion in a spatial audio field, this project also aims to establish a research baseline for this area. This project thus has another objective of building a model that can output robust human motions given audio in a spatial field as input. The development of this model will be largely based on past models that have been proved to be highly effective, especially the Motion Diffusion Model (MDM) (Tevet et al., 2022), Bailando (Siyao et al., 2022), and CAMDM (Chen et al., 2024).

3) As the baseline for a new sub-area in machine learning and human motion synthesis, a nonnegligible objective of this project is to propose metrics that are effective for evaluating the fidelity, fluency, and diversity of the generated human motion in a spatial audio field. Due to various similarities within the area of human motion synthesis, the metrics will still be proposed upon the original ones (e.g., Fréchet Inception Distance (FID)), and new metrics will be added to include the spatial information of the audio.

3. Methodology

It is unpractical to first make a new dataset and then develop a fresh model to train upon this dataset when encountering a field for the first time, because there is not a clear understanding of exactly what kind of data is needed. Rushing into data collection may result in a dataset found to be incomplete in the later process, and it would be rather time-consuming to recollect or calibrate the dataset. Therefore, this project starts from exploring a model that is efficient at generating high-quality human motions for the AIST++ dataset (Li et al., 2021), which is currently the most used dataset in the music-to-motion human motion synthesis task.

Building a new model for human motion synthesis may take a lot of time and has a risk of failure. Therefore, the model of this project will be based on previous successful models. The structure of the model will largely inherit the Motion Diffusion Model (MDM) (Tevet et al., 2022), the Efficient Motion Diffusion Model (EMDM), and the CAMDM (Chen et al., 2024), which all



apply the diffusion process that is originally used in machine learning for image generation. In terms of data processing and loading, the methods applied by Bailando (Siyao et al., 2022) can be used for reference, as it also performs its training and evaluation using the AIST++ dataset (Li et al., 2021).

The audio and pose features, as inputs of the model, are of great importance because they are one of the most direct factors affecting the power and integrity of the model. The audio features used in the model will be very similar to those used in Bailando (Siyao et al., 2022), but there will still be some improvements to satisfy the objective of this project. For example, constant-Q chromagram, which is used as one of the audio features for training in Bailando (Siyao et al., 2022), will be removed because it is a feature for music only, while music is not the only kind of audio considered in this project. On the other hand, some other features will be added to meet the need for the spatial information of the audio so that the character will perform differently according to different locations of the audio source. The root mean square (rms) energy, for instance, can be added as an indicator for the distance of the audio source from the character. The model proposed by Wang et al. (2024) can be used for reference as well, in terms of simulation of the acoustic field. For the pose features, the model will inherit directly from those suggested by Guo et al. (2022), using both the global positions and the local rotations of the joints.

In the training process, the losses of both the training data and the validation data will be calculated and outputted at every epoch so that the trend of the losses can be monitored using Tensorboard, and modifications can thus be made immediately when the trend is found abnormal. Specifically, the loss used by the model will be mainly the mean squared error (mse) loss of the predicted pose vectors and the ground truth, with assistance of some other losses in order to make the predicted human motion more delicate. For instance, foot contact loss, which is a loss that calculates how well the feet of the character are in contact with the ground, will also be applied to prevent the foot sliding problem. The predicted human motions will also be visualized and outputted every 50 epochs to check manually whether the model is training in the right way.

Building a new dataset that is suitable for training a model that can synthesize human motions given an audio in space will be the next step of the project. Since the motions are expected to be conducted in a relatively large space, an inertial motion capture system will be the best choice because an inertial system is less affected by distance compared with an optical one, and the data obtained is also accurate and easy to process. For audio acquisition, it is expected that headphones (or AirPods) will be used to collect the binaural sound at the position of the character. The binaural sound will be useful in predicting the direction of arrival of the audio and the distance of the audio source from the character when the data is later inputted into the model. In order

to ensure the integrity and validity of the dataset, the same audio will be played at a set of different positions related to the character, and corresponding reactions to the audio will be recorded. All the data collected will then be integrated together and split into training, validation, and test sub-datasets.

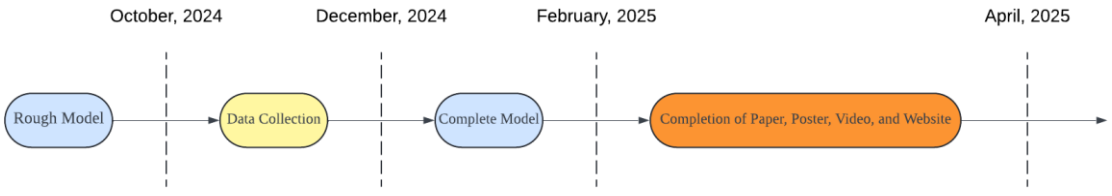
After the establishment of the new dataset, the model will be improved to fit the dataset. Calculation of metrics such as Fréchet Inception Distance (FID) will show the performance of the model with the new dataset numerically. In addition, new metrics will also be designed and calculated to demonstrate the model's ability of generating human motions that have high consistency with the spatial information of the audio. The same metrics will be calculated for similar models such as DanceNet (Zhuang et al., 2022) and Bailando (Siyao et al., 2022) and compared with those obtained from the model of this project to further demonstrate its strengths.

After completing all of the above studies, a paper will be written and published to review the achievements.

4. Schedule and Milestones

This project is scheduled on a monthly basis.

- **September, 2024**—Complete the detailed project plan and a rough project webpage.
- **October, 2024**—Build a model for the "old" dataset AIST++ (Li et al. 2021).
- **November, 2024**—Collect data for the new dataset using a mocap system.
- **December, 2024**—Build the new dataset.
- **January, 2025**—Complete the interim report.
- **February, 2025**—Improve the model based on the new dataset.
- **March, 2025**—Write the paper for the research achievements.
- **April, 2025**—Complete the final report, the webpage, the poster, and the video for the project. End the project with the final presentation.



5. References

Chen, R., Shi, M., Huang, S., Tan, P., Komura, T., & Chen, X. (2024, July). Taming Diffusion Probabilistic Models for Character Control. In ACM SIGGRAPH 2024 Conference Papers (pp. 1-10).

Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., & Cheng, L. (2022). Generating diverse and natural 3d human motions from text. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5152-5161).

Li, R., Yang, S., Ross, D. A., & Kanazawa, A. (2021). Ai choreographer: Music conditioned 3d dance generation with aist++. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13401-13412).

Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., ... & Liu, Z. (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 (pp. 11050-11059).

Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., & Bermano, A. H. (2022). Human motion diffusion model. arXiv preprint arXiv:2209.14916

Wang, M. L., Sawata, R., Clarke, S., Gao, R., Wu, S., & Wu, J. (2024). Hearing Anything Anywhere. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11790-11799).

Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., ... & Shan, Y. (2023). Generating human motion from textual descriptions with discrete representations. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 14730-14740).

Zhou, W., Dou, Z., Cao, Z., Liao, Z., Wang, J., Wang, W., ... & Liu, L. (2023). Emdm: Efficient motion diffusion model for fast, high-quality motion generation. arXiv preprint arXiv:2312.02256, 2.

Zhuang, W., Wang, C., Chai, J., Wang, Y., Shao, M., & Xia, S. (2022). Music2dance: Dancenet for music-driven dance generation. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 18(2), 1-21.