Dean&Francis

ISSN 2959-6157

Character Pose Generation and Editing Method Based on Large Model and Multi-modal Prompt Words

Boxiao Xu

The University of Waikako Joint Institute at Zhejiang University City College, Zhejiang University City College, Zhejiang, China
32211052@stu.hzcu.edu.cn

Abstract:

Character pose generation and editing is an important technology for character shaping in the fields of film, animation, and game character development. The current mainstream posture editing methods usually focus on text and image input, but overlook the equally critical audio modality; Meanwhile, existing technologies generally cannot run on consumer grade graphics cards. Therefore, this paper proposes a new technical approach: using a large model as a multimodal processor, a stable diffusion model as a transition device to generate four views after attitude changes, and finally using 3D Gaussian multi view generation models to generate the final model. Then, regarding the fusion of audio modalities, this paper introduces the "persona mask" mechanism to set a unified audio analysis method for the large model in advance, achieving accurate recognition of character emotions from speech semantics, and synchronously mapping emotional features to character actions. During this process, all models are called using interfaces. Through experiments, it can generate poses that match the character design very well on the 4070 graphics card. Hope this technology can be integrated with skeletal animation in the future, so as to complete the editing of character action animations.

Keywords: Pose editing; Large model; Multi-modal.

1. Introduction

The core role of character pose generation and editing technology in film, animation, and game character development is to create static poses that conform to character settings for skeleton animation keyframes. It can transform abstract character personalities and plot contexts into intuitive action forms, and this transformation effect directly determines the narrative appeal and user immersion of virtual characters.

Previous studies have conducted a series of explorations around "multimodal pose generation":

In the direction of multimodal generation and editing frameworks [1], UniPose proposed a framework that can use images, text, and 3D SMPL poses as input

signals to achieve pose understanding, generation, and editing tasks [2]; ChatPose has the ability to infer poses from text; On this basis [3], CoT Pose deepens its reasoning ability for abstract information [4]; PointT2I introduces a feedback mechanism to ensure consistency between the generated pose and the textual description [5].

In the pose prior direction, MOPED introduces SMPL parameters to make the prior conditional on images/text to enhance generative control [6]; DPoser-X proposed a mechanism that can capture the correlation between various parts of the body [7]; PoseEmbroider proposed a modal fusion processing method to combine visual+-pose+semantic representation.

In the direction of detailed manipulation of posture [8], new research such as AutoComPose automatically generates prompt words required for posture transformation [9]; Text Driven Human Image Generation with Texture and Pose Control adds joint control over texture and pose [10]; PoseLaVA treats posture as an independent modality and collaborates with text/image modalities for control.

Although significant progress has been made in posture editing by predecessors, there are still two unresolved issues:

First, there is a key blind spot in mode coverage - all mainstream methods are not included in audio mode. In film and game creation, character dubbing (such as passionate lines and low-pitched monologues) is the core carrier for conveying emotions. However, existing technology cannot convert the emotional semantics in audio into posture features, resulting in generated postures that are disconnected from the character's audio context;

Secondly, the hardware threshold is too high and the adaptability is insufficient. The overall model operation usually requires high-performance computing resources, which are generally unable to be implemented on consumer grade graphics cards

So the paper designed a new lightweight technology framework using interfaces - "character model → 2D slicing → 2D pose editing → 3D reconstruction". Secondly, the paper designed a multi-mode information coupling module driven by a large model, innovatively introduced a three mode input system consisting of audio (dubbing, background music), images and text, and established the priority rules of "image action framework → text details modification → audio emotion assistance". This module can transform personality roles and emotions into quantifiable action parameters, so as to solve the problem of disconnection between posture and dubbing emotion, plot setting.

2. Method implementation

2.1 Character model → 2D slicing → 2D pose editing → 3D reconstruction

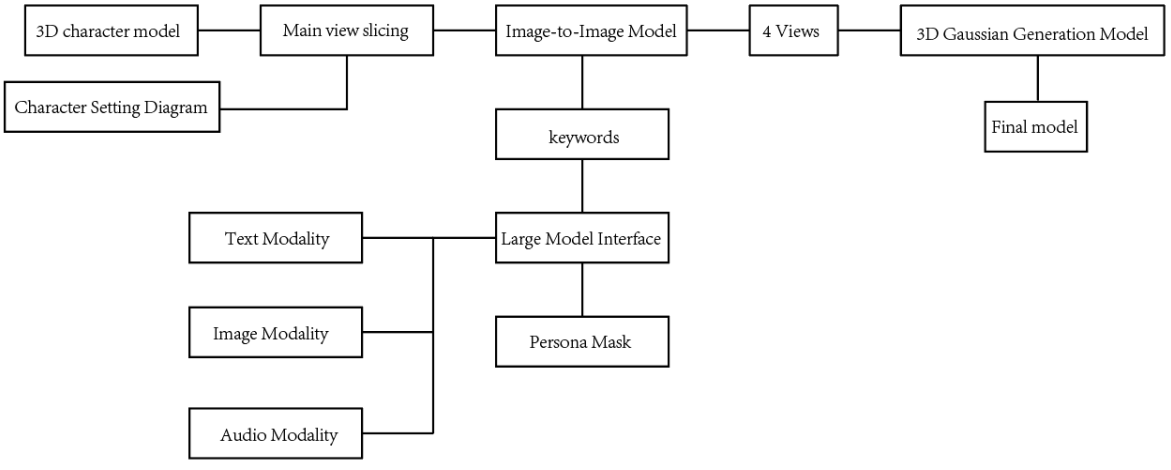


Fig. 1 Principle explanation diagram

As shown in the figure 1, the technical implementation of the method mainly relies on three core modules, namely the multimodal input processing module, the 2D action redraw module, and the 3D model reproduction module. The multimodal input processing module and the 2D action redraw module are both implemented through OpenAI interfaces, while the 3D model reproduction module is built based on the Hunyuan3D model.

The multimodal input processing module uses a large model and several personality masks to integrate the information input from text, images, and audio modalities, then input the keywords required for the image generation

ISSN 2959-6157

module.

The 2D action redrawing module takes the 3D character model input by the user and the keywords generated by the previous module as inputs, and generates four different perspective images of the person with posture adjustment through the stable diffusion model. In the input stage, users have flexible selection space - they can directly input the character setting diagram, or use the 3D model main view obtained (through slicing algorithm processing) as input material.

The 3D pose generation module takes 4 multi view 2D images as inputs to the 3D model reproduction module (Hunyuan 3D model), and finally outputs a 3D character model with modified actions that conform to physical laws and personality characteristics.

2.2 Multimodal input processing module

2.2.1 Contradiction Handling

This module is connected to a large model interface with multimodal understanding capabilities and outputs logic through persona constraints. Ensure that it follows the contradiction handling rule of "text mode>image mode>audio mode", and finally generate accurate 3D character action keywords. All action keywords must be accompanied by a unified suffix "remove background, change the actions of characters inside" to clarify the editing objectives for subsequent image generation modules.

The contradiction handling between multiple modalities mainly follows the following rules - using image modalities as the basic framework for actions; The audio modality is used to assist in action details, analyze emotional features and convert them into adjusting action amplitude and adding some action details, without changing the overall action framework; The text modality is used for final action calibration, where when the text description action does not match the image extraction action, the text action takes precedence.

2.2.2 Audio Modal

The audio modal persona combines semantic content and intonation features (such as word repetition frequency, speech rate, volume fluctuations, etc.) in the audio. Analyze the speaker's emotional tendencies through this information. Then, using the "Emotion Type and Action Comparison Table" (which can be omitted as the table mainly quantifies the amplitude of actions), corresponding character posture keywords are generated. These keywords are only used to adjust the details of actions and do not change the overall action framework determined by the image modality. Emotion judgment needs to be based on intonation frequency and semantic keywords, where low intonation corresponds to negative emotions, and rising intonation corresponds to positive emotions. Words such as "sad" and "lost" in semantics point to negative emotions, while words such as "happy" and "excited" point to positive emotions. If there is a conflict between semantics and intonation, for example, if the intonation is low but the semantics contain "happy", the core emotion needs to be judged based on intonation characteristics, such as "depressed".

2.2.3 Image Modal

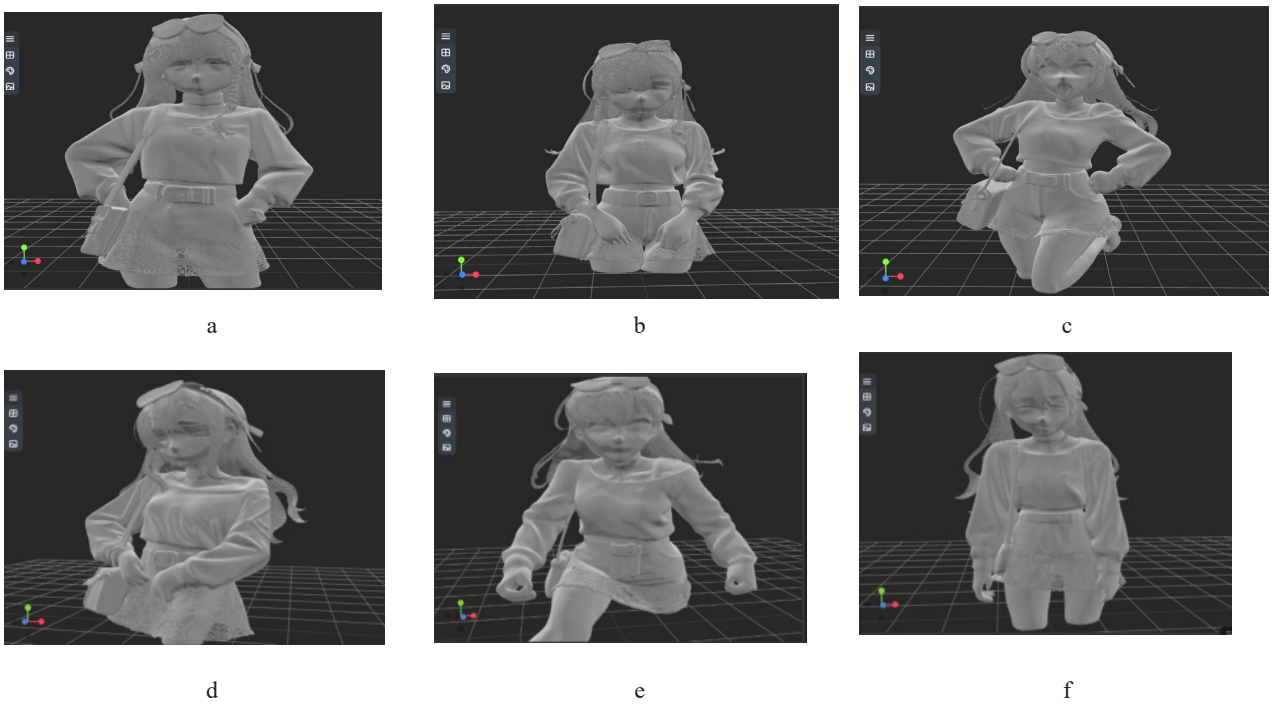
The image modal personality mask focuses on the normal physiological organs of the human body, excluding non-human physiological structures such as wings and tails, extracts action states, and generates high-precision action keywords. At the same time, it completes the "left-right exchange" calibration to ensure that the action description is consistent with actual needs. The calibration rule for "left-right swapping" is that, since the input image is from a positive perspective, the "left hand/foot" of the character in the image corresponds to the "right hand/right foot" of the actual character.

2.2.4 Text Modal

The text modal persona is responsible for analyzing the input text content and generating action keywords required for the graph generation model. In terms of personality/scene text processing, if the text contains personality descriptions such as "gloomy" and "bold", or scene descriptions such as "combat scene" and "leisure scene", it is necessary to first examine and combine audio modalities to analyze the emotional characteristics of the characters, and convert them into action parameters.

3. Results

3.1 Audio Modal



a b c

d e f

Fig. 2 Model for audio modality generation (a: Delight, b: Nervous, c: Angry, d: Swinging along with the song e: Battle BGM, f: Despair)

Audio modal input can be character dubbing, songs, scene music, etc.

For example, figures 2a, 2b, 2c, and f are al l character voiceovers. The semantic meaning of Figure 2a is happiness, the tone is high, and the emotion detected is "delight". At this time, action keywords such as "increased amplitude of action (such as raising the arm by 15 °), clenched fists with both hands, slightly raised shoulders, and smiling eyes" are generated; The semantics of Figure 2b are intermittent words with low intonation, and the detected emotion is "tension". At this point, keywords such as "reduced amplitude of action, tense expression, messy hair, unconscious curled fingers, and slight body tremors" are generated; Figure 2c, with a semantic meaning of blame, a high tone, and an emotion of "anger", generates keywords such as "clenching fists with both hands, raising shoulders, furrowing brows, and pressing the corners of the mouth"; Among them, Figure 2f shows the situation where the semantic tone does not match, with the semantic meaning being happy and the tone being low, recognized as "despair", generating the keywords "haggard face, weak hands, messy hair".

And Figure 2d uses songs as audio input, generating the keyword 'swing with music'; Figure 2e uses passionate background music input to generate keywords such as "significantly increased movement amplitude (such as large arm swings, increased leg strides), upright body posture, shoulder extension, and slight head elevation".

3.2 Image Modal

The large model will accurately describe the overall posture of the character in the image, as well as the movement status of the torso, limbs (arms, palms, legs, feet), and head. It includes quantifiable parameters such as angle and attitude type.

When describing a character's posture, the overall posture is often presented as "standing posture" or "half squatting posture (with knees bent at 45 °)"; In terms of limb movements, usually the right

arm is extended backwards to shoulder height, the palm of the right hand is raised 30 degrees, the left hand naturally hangs to the side of the body, the left leg is slightly bent, and the right leg is upright; The head movement is often described as "head turning to the right, face facing forward".

ISSN 2959-6157

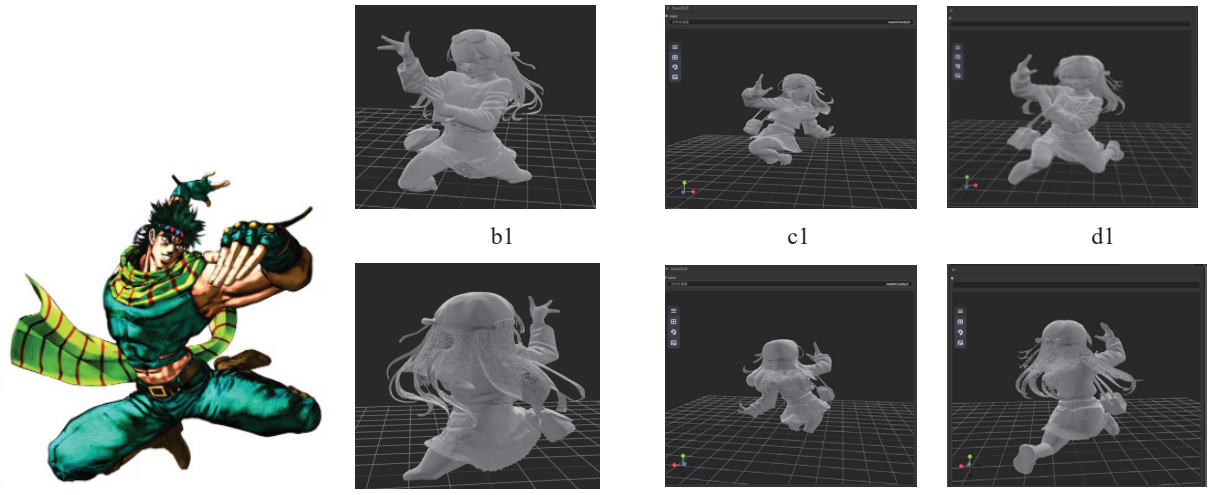


Figure 3a serves as the original action input, while Figures 3b, 3c, and 3d are edited with text to create a set of keyframes for a usable skeletal animation.

Fig. 3 Model for image modality generation (a: Input image, b1 & b2: Generate motion, c1 & c2: Action adjustment 1, d1 & d2: Action adjustment 2)

3.3 Text Modal

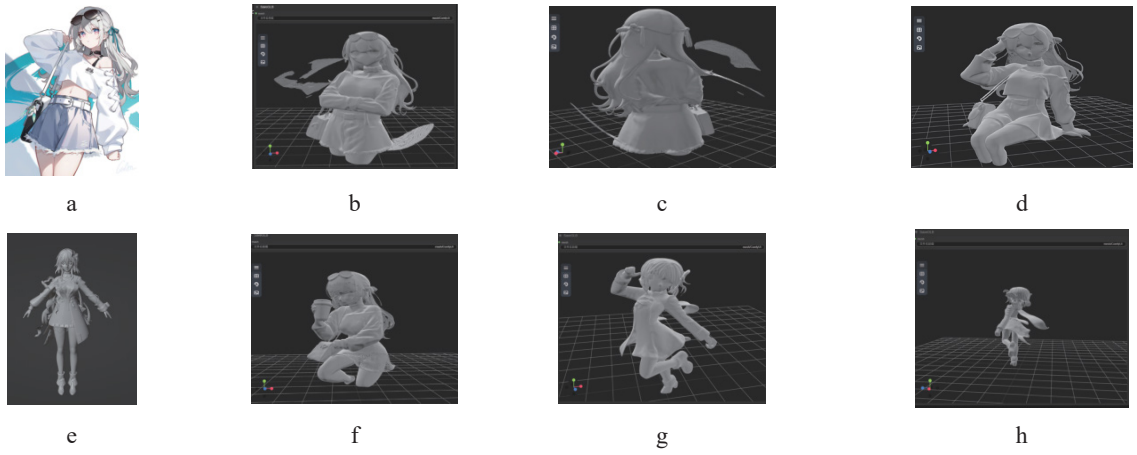


Fig. 4 Model for text modality generation (a: Input image, b: Cross one's arms, c: Back of Figure b, d: sitting posture e: Input model, f: Scene of finding something while drinking coffee, g: Mischievous personality, taking photo scenes, h: Back of Figure g)

The information input in text mode will be recognized into two types: one is the text that directly edits the action posture, and the other is the text about the character design. The text about the character design will be combined with the emotions extracted from the audio mode to adjust the micro actions.

Figures 4a-4d tested a model generated by inputting a set diagram and describing actions in text; Figure 4f is a model generated by describing a scene of drinking coffee in text; Figures 4e and 4g-4h show the model generated by inputting the main view of the model and describing the actions in text

4. Conclusion

To enable better application of character pose generation and editing techniques in fields such as movies, anime, and games. This article proposes an interface based lightweight technology framework and successfully runs it on the RTX 4070 graphics card. In addition, this article also introduces the persona mechanism in the big model, clarifies the importance of audio modality, proposes an

emotion recognition mechanism for audio modality, and a quantifiable mapping mechanism for emotion action during modality fusion, allowing posture to fit character attributes and plot context. Finally, hope that character pose editing technology can truly enter production and daily life scenarios.

References

[1] Li Y H, Hou R B, Chang H, Shan S G, Chen X L, UniPose: A Unified Multimodal Framework for Human Pose Comprehension, Generation and Editing, *arXiv preprint arXiv:2411.16781*, 2024.

[2] Feng Y, Lin J, Dwivedi S K, Sun Y, Patel P, Black M J, ChatPose: Chatting about 3D Human Pose, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[3] Cha J, Kim J, CoT-Pose: Feng Y, Lin J, Dwivedi S K, Sun Y, Patel P, Black M J, ChatPose: Chatting about 3D Human Pose, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[4] Feng Y, Lin J, Dwivedi S K, Sun Y, Patel P, Black M J, ChatPose: Chatting about 3D Human Pose, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[5] Ta C K, et al., Multi-modal Pose Diffuser: A Multimodal Generative Conditional Pose Prior (MOPED), *arXiv preprint arXiv:2410.14540*, 2024.

[6] Lu J Z, Lin J, Dou H K, Zeng A L, Deng Y, Liu X, Cai Z G, Yang L, Zhang Y L, Wang H Q, Liu Z W, DPoser-X: Diffusion Model as Robust 3D Whole-body Human Pose Prior, *arXiv preprint arXiv:2508.00599*, 2025.

[7] Delmas G, Weinzaepfel P, Moreno-Noguer F, Rogez G, PoseEmbroider: Towards a 3D, Visual, Semantic-aware Human Pose Representation, *arXiv preprint arXiv:2409.06535*, 2024.

[8] Shen Y T, Eum S, Lee D, Shete R, Wang C Y, Kwon H, Bhattacharyya S S, AutoComPose: Automatic Generation of Pose Transition Descriptions for Composed Pose Retrieval Using Multimodal LLMs, *arXiv preprint arXiv:2503.22884*, 2025.

[9] Jin Z D, Xia G Y, Yang P K, Wang M X, Sun Y B, Liu Q S, Text-driven Human Image Generation with Texture and Pose Control, *Neurocomputing*, 2025.

[10] Feng D, Guo P, Peng E, Zhu M, Yu W, Wang P, PoseLLaVA: Pose Centric Multimodal LLM for Fine-Grained 3D Pose Manipulation, *AAAI Conference on Artificial Intelligence*, vol. 39, no. 3, 2025.