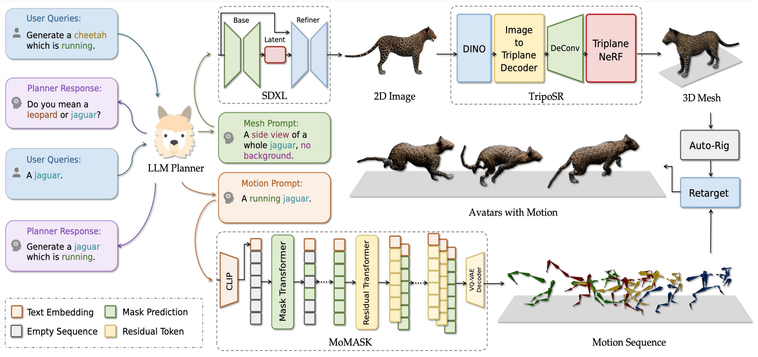
Motion Avatar: Generate Human and Animal Avatars with Arbitrary Motion

Zeyu Zhang[12]*, Yiran Wang[2]*, Biao Wu[3]*, Shuo Chen[4], Zhiyuan Zhang[5], Shiya Huang[5], Wenbo Zhang[5], Meng Fang[6], Ling Chen[3], Yang Zhao[1]
[1]La Trobe University, [2]The Australian National University, [3]University of Technology Sydney,
[4]Monash University, [5]The University of Adelaide, [6]University of Liverpool
*Equal contribution.

Abstract

In recent years, there has been significant interest in creating 3D avatars and motions, driven by their diverse applications in areas like film-making, video games, AR/VR, and human-robot interaction. However, current efforts primarily concentrate on either generating the 3D avatar mesh alone or producing motion sequences, with integrating these two aspects proving to be a persistent challenge. Additionally, while avatar and motion generation predominantly target humans, extending these techniques to animals remains a significant challenge due to inadequate training data and methods. To bridge these gaps, our paper presents three key contributions. Firstly, we proposed a novel agent-based approach named Motion Avatar, which allows for the automatic generation of high-quality customizable human and animal avatars with motions through text queries. The method significantly advanced the progress in dynamic 3D character generation. Secondly, we introduced a LLM planner that coordinates both motion and avatar generation, which transforms a discriminative planning into a customizable Q &A fashion. Lastly, we presented an animal motion dataset named Zoo-300K as shown in the figure below, comprising approximately 300,000 text-motion pairs across 65 animal categories and its building pipeline ZooGen, which serves as a valuable resource for the community.



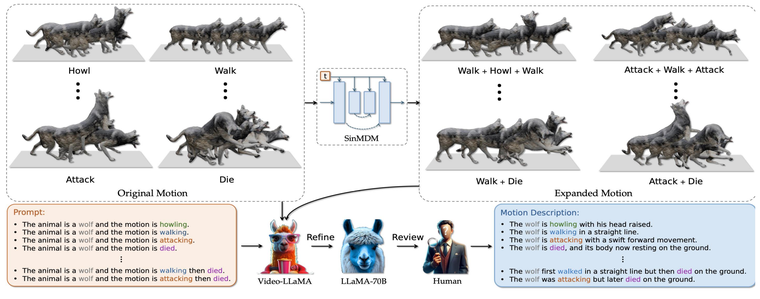
Introduction

The field of computational modeling for dynamic 3D or 4D avatars holds significant importance across various domains such as robotics, virtual, and augmented reality, computer gaming, and multimedia. Creating high-quality and user-friendly animated avatars is a desired goal of the 3D computer vision community. This involves not only ensuring the visual appeal of the avatars but also prioritizing their functionality and ease of use. Traditional methods involve extracting information directly from video recordings, then using the data to model and reconstruct dynamic avatars in both spatial and temporal dimensions. Other approaches involve integrating 3D reconstruction with video diffusion to bring 3D meshes to life through animation. However, these methods often suffer from imprecise and limited motion control, or exhibit inconsistencies between multiple views of the 3D mesh. These limitations hinder the seamless manipulation and other real-world applications of dynamic avatars in interactive settings.

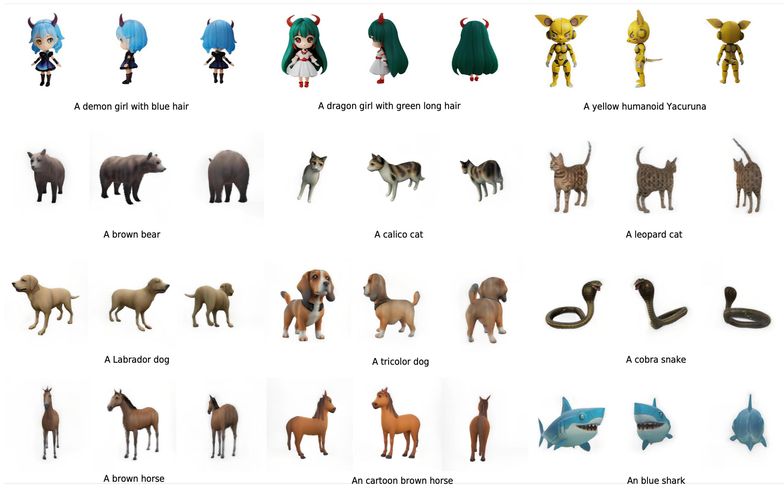
Methods

The generation pipeline consists of three parts as shown in Figure 2:

1. The objective of LLM planner design is to efficiently and automatically extract useful information from the user prompt, then use external components to create customizable animated avatars effectively. To efficiently deploy the planner on a local device, we developed the LLM planner leveraging the LLaMA-7B framework. This planner was finely tuned using the avatar Q&A dataset to better suit the task of generating motion avatars.
2. Motion generation with MoMask involves a two-stage training process. Firstly, a residual VQ-VAE compresses a motion sequence into a latent vector, which is then replaced with its nearest code entry in a codebook. Then the decoder projects the quantized code sequence back to motion space for motion reconstruction.
3. For generation of avatar meshes, the Motion Avatar initially employed Stable Diffusion XL to utilize the mesh prompt generated by LLM planner, producing a 2D image of the avatar's front or side view tailored to requirements. Given a 2D input, TripoSR initially obtains the image representation using a DINO encoder, followed by the image-to-triplane decoder that transforms the image embedding into the triplane-NeRF representation.



Presently, a major obstacle in generating animal motion is the lack of sufficient data pairs containing both animal motion and text descriptions. To address this gap, we introduced Zoo-300K, a dataset containing around 300,000 pairs of text descriptions and corresponding animal motions spanning 65}different animal categories. Based on the Truebones Zoo free dataset, which comprises human-cultivated animal motion data annotated with textual labels denoting animal and motion categories, we have introduced a dataset construction pipeline named ZooGen, which facilitates the creation of a text-driven animal motion dataset, shown in the figure on the left-hand side.



Experiments

The Table below demonstrated that our method achieves high-quality animal motion generation. The results indicate that our approach produces highly realistic and promising animations, highlighting the effectiveness and potential of our technique in generating detailed and lifelike animal motions.

The figure above illustrates our model's capability to seamlessly generate animal motion based on text conditions, showcasing potential applications in fields like computer gaming and film making. Besides, the Figure below illustrates randomly selected examples of 3D avatars generated by our model.

Method	R-Prec Top 1 ↑	R-Prec Top 2 ↑	R-Prec Top 3 ↑	FID ↓	MultiModal-Dist ↓	Diversity →
Average (GT)	0.437	0.650	0.767	0.024	2.559	11.790
Average (Ours)	**0.069**	**0.144**	**0.173**	**62.785**	**9.371**	**4.158**