A Digital Human Interaction System: A Three-Step Construction from a Single Image to Real-time Conversation Avatar

Jing Wang
The Hong Kong University of Science and Technology
jwangke@connect.ust.hk

Abstract

Simply upload a portrait or personified image of characters from famous games or animations (e.g., a personified child toy), and it can come to life. In just three steps, you can build a customizable digital avatar from scratch and have real-time intelligent and emotional interactions. The first step is to upload the image and use Vision Language Models (VLM) to analyze the descriptive text. Then, you can optimize the prompts in the Stable Diffusion web UI[1] or ComfyUI[2] and their plugins to customize your preferred expressions, poses, backgrounds, and styles using image-to-image or Low-Rank Adaptation (LoRA) fine-tuning[3]. The second step involves using the official LAM (Large Avatar Model for One-shot Animatable Gaussian Head) platform[4] to generate a chatting avatar corresponding to the uploaded image. Finally, import the avatar into the Open Avatar Chat platform[5] and you can interact with your digital avatar in real time. This project uses Streamlit to build a user-friendly, aesthetically pleasing GUI that links the entire process while ensuring adequate legality checks (available at Github[6]).

1 Introduction

In recent years, the development of digital avatars has revolutionized the way we interact with virtual characters. The ability to create highly customizable, lifelike avatars has opened up new possibilities for entertainment, education, and digital communication. This paper presents a novel approach to building such digital avatars, allowing users to bring their own creative characters to life using a combination of cutting-edge machine learning techniques and interactive platforms.

The process begins with uploading an image, which is analyzed using a VLM to generate descriptive text, helping optimize and customize the avatar's appearance. Users can further refine the image with tools like Stable Diffusion or ComfyUI, adjusting expressions, poses, and styles, while LoRA enables efficient fine-tuning of pre-trained models.

Next, the LAM is used to create a dynamic chatting avatar with realistic facial features, allowing it to interact based on user input. Finally, the avatar is imported into the Open Avatar Chat platform, enabling real-time, intelligent, and emotionally responsive conversations, resulting in a highly customizable and interactive digital persona.

To ensure the usability and accessibility of this process, I have built a user-friendly GUI using

---

[1] https://github.com/AUTOMATIC1111/stable-diffusion-webui

[2] https://github.com/comfyanonymous/ComfyUI

[3] https://github.com/Akegarasu/lora-scripts

[4] https://www.modelscope.cn/studios/Damo_XR_Lab/LAM_Large_Avatar_Model

[5] https://github.com/HumanAIGC-Engineering/OpenAvatarChat

[6] https://github.com/WillongWang/Single-Image-to-Real-time-Conversation-Avatar-powered-by-LAM-and-Cosyvoice-v2

Streamlit, which seamlessly connects all the steps involved in creating and interacting with a digital avatar. This interface simplifies the complex process, making it accessible to users with varying technical backgrounds. Furthermore, legal and ethical considerations are incorporated into the platform to ensure responsible use of digital avatars in interactive environments.

In summary, this approach combines state-of-the-art technologies in AI, offering a powerful tool for creating personalized, interactive digital avatars. The proposed workflow is not only efficient but also adaptable, opening the door to a wide range of applications in digital media, entertainment, and virtual communication.

2 Related Works

2.1 Stable Diffusion and LoRA

Stable Diffusion is a latent diffusion model that generates high-quality images from text prompts by compressing images into a latent space and applying diffusion processes. Its loss function is mathematically formulated as:

$L_{LDM} = \mathbb{E}_{\epsilon(x),y,\epsilon\sim N(0,1),t}[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2]$

which incorporates conditioning information y, such as text prompts or semantic maps, to control the generation process. The conditioning is processed by a function $\tau_\theta(y)$, typically another neural network, which projects y into a suitable representation. This processed conditioning is then used as an additional input to the noise prediction network ϵ_θ, alongside z_t and t (the noisy latent representation at time step t).

The attention mechanism is used to integrate conditioning information into the UNet backbone of the LDM. In this standard scaled dot-product attention,

$Q = W_Q^{(i)} \cdot \varphi_i(z_t),\ K = W_K^{(i)} \cdot \tau_\theta(y),\ V = W_V^{(i)} \cdot \tau_\theta(y)$

Here, $\varphi_i(z_t)$ denotes a flattened intermediate representation of the UNet implementing ϵ_θ and W are learnable projection matrices.

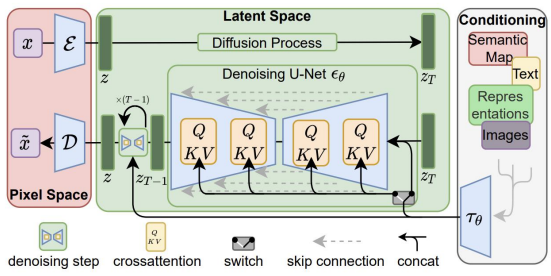

Figure 1: A figure of latent diffusion model (Rombach, 2022)

LoRA, introduced by (Hu, 2021), is a parameter-efficient fine-tuning technique that adds low-rank matrices to the model's weights, particularly in cross-attention layers. The weight update is decomposed as $\Delta W = A \bullet B$, where A and B are low-rank matrices, reducing trainable parameters and computational demands, making it ideal for customizing Stable Diffusion for specific styles or characters.

2.2 Digital Avatars

Recent advancements in digital human research focus on enhancing realism and interactivity. VLMs, which integrate vision and language processing, are pivotal, as seen in examples like Marianne, a digital receptionist for Tour de France, using LLMs for real-time, human-like responses (NTT Data, 2024). Platforms like PARSONII leverage Generative AI for customizable digital humans with micro-expressions and real-time dialogues, improving customer engagement.

Text-to-Speech (TTS) technologies are crucial, with recent developments including deep learning-based pipelines for virtual digital humans with emotional talking faces and lip-syncing, driven by TTS for natural body and lip movements (ScienceDirect, 2023). These advancements, detailed in studies like "Generation of virtual digital human for customer service industry", highlight the integration of VLM and TTS for natural, empathetic interactions.

2.3 LAM

LAM, presented at SIGGRAPH 2025 by (He, 2025), is a state-of-the-art model for creating animatable 3D Gaussian heads from a single image. Unlike traditional methods requiring video sequences, LAM generates immediately animatable and renderable heads in a single forward pass, enabling real-time animation on platforms including mobile devices. Its canonical Gaussian attributes generator uses FLAME canonical points and a Transformer to predict attributes, ensuring efficiency and cross-platform compatibility.

3 Methodology

3.1 Process Overview

The methodology employed in this study can be broken down into three major steps, as shown in the algorithm flowchart below.

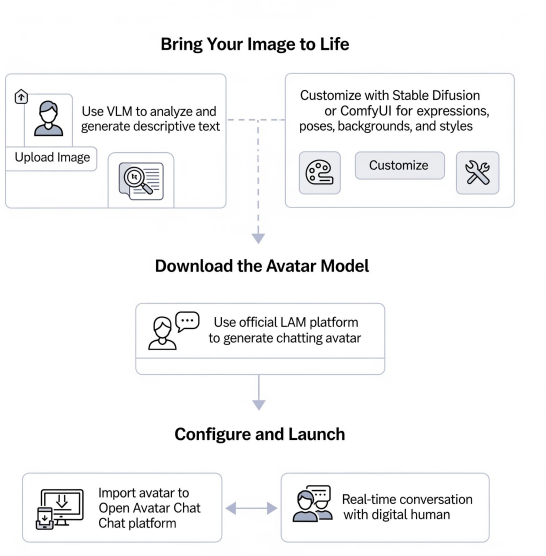

Figure 2: Workflow

3.2 Bring Your Image to Life

To bring an image to life, we utilized the official code for the Image-Text-to-Text task from Hugging Face[7]. This process begins by uploading an image for analysis, allowing for the customization of text prompts. For instance, we uploaded an image of the character Mi Yue from the popular mobile game "Honor of Kings". In the Stable Diffusion web UI, a v1.5 anime-style model from LibLibAI[8] is selected, and four additional Mi Yue images are used as a dataset for LoRA fine-tuning. This process takes only a few minutes, producing a more detailed and refined Mi Yue image.

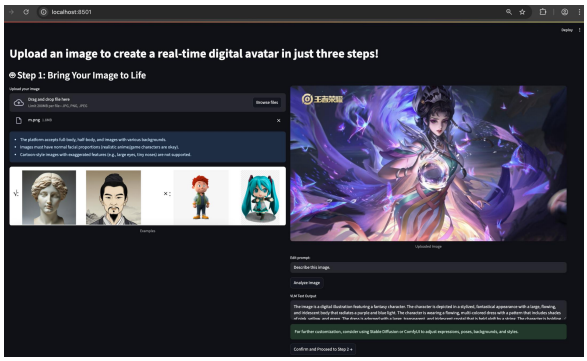

Figure 3: GUI sample

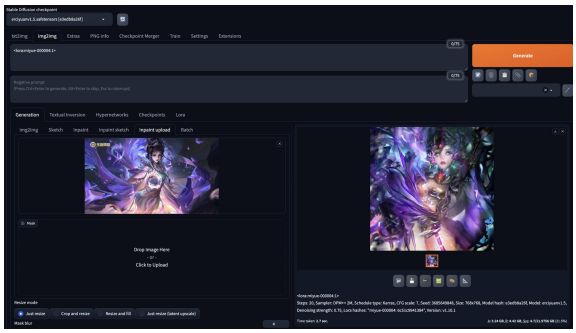

Figure 4: LoRA in Stable Diffusion web UI

3.3 Download the Avatar Model

The next step leverages the LAM platform, where a personified image and a short video are input to generate a digital avatar file. The platform accepts full-body, half-body, or varied-background images (not limited to frontal portraits) and automatically detects and processes facial features. However, images must have normal facial proportions (e.g., coordinated eye, nose, and mouth ratios, akin to realistic anime/game characters). Cartoon-style images with exaggerated features (e.g., large round eyes, tiny noses, and mouths) are not supported.

7 https://huggingface.co/tasks/image-text-to-text

8 https://www.liblib.art

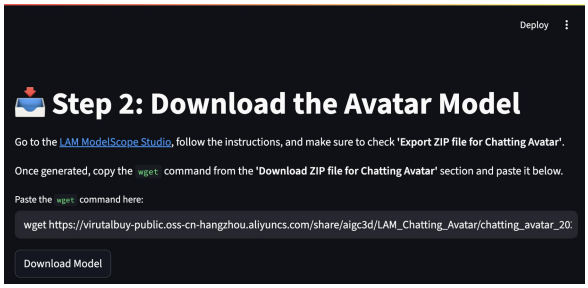

Figure 5: GUI sample

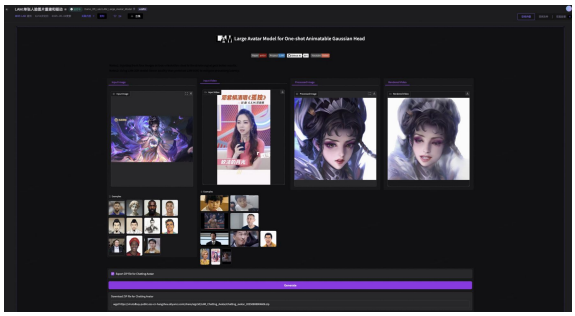

Figure 6: LAM platform

3.4 Configure and Launch

After the avatar has been generated, the next step is to import it into the Open Avatar Chat platform for deployment. The avatar can be deployed on the Compshare GPU server platform, which allows for real-time interaction (image by @十字鱼[9]. Open Avatar Chat supports multiple large language models (LLMs) such as qwen-plus, qwen-turbo, and qwen-max, each offering varying levels of response speed and inference accuracy. Users can also input system prompts to define the avatar's persona. For text-to-speech (TTS), the model uses cosyvoice-v2[10], which provides different voice styles, including "浪漫风情女声" (Romantic female voice), "温暖春风女声" (Warm spring breeze female voice), "甜美娇气女声" (Sweet delicate female voice), "知性粤语女声" (Intellectual Cantonese female voice), "知性英文女声" (Intellectual English female voice), "豪放可爱童女声" (Bold cute child female voice), and "元气甜美童女声" (Energetic sweet child female voice). Users are free to choose from these LLMs and voices based on their preferences, covering child/adult voices with styles like sunny, mature, gentle, and delicate. Additionally, SileroVad, a pre-trained enterprise-grade Voice Activity Detector, is integrated for improved audio recognition and interaction. This setup ensures an engaging, responsive experience for users.

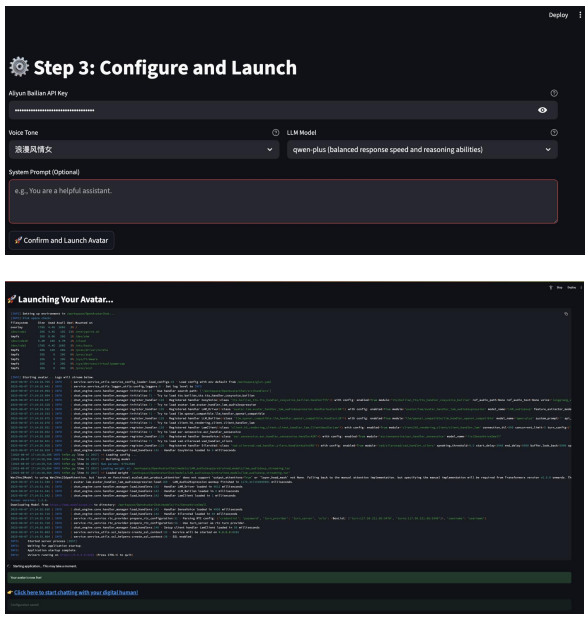

Figure 7: GUI sample

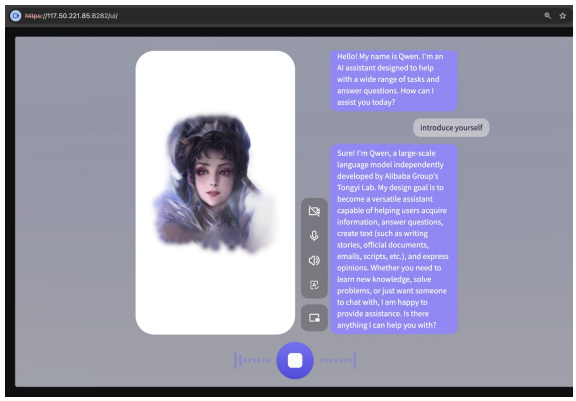

Figure 8: Open Avatar Chat platform

4 Results

The integration of LAM and Open Avatar Chat has led to several impressive outcomes. The LAM platform enables ultra-realistic 3D avatar creation from a single image in just seconds, followed by super-fast cross-platform animation and rendering, ensuring compatibility with a variety of devices. Moreover, the platform supports a low-latency SDK for real-time interactive chatting avatars, providing an average response delay of

9 https://www.compshare.cn/images/63f27744-54ee-4ea2-a536-a72075f4b28e

10 https://help.aliyun.com/zh/model-studio/cosyvoice-python-sdk

approximately 2.2 seconds. Open Avatar Chat further enhances this experience by supporting multi-modal language models, including text, audio, and video input, which allow the avatars to generate intelligent and emotionally engaging responses.

5 Future Directions

In the future, there are several potential improvements that could enhance the user experience. For example, using more advanced plugins within the Stable Diffusion Web UI or ComfyUI could result in more detailed and beautiful images. Additionally, further exploration of the components supported by Open Avatar Chat may help overcome some of the limitations of the LAM platform, such as the restriction on image uploads. Furthermore, reducing latency times or improving the level of intelligent and emotional response from the avatars would significantly enhance the interactivity and realism of the digital humans.

References

Rombach, R., et al. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684-10695).

Hu, E., Shen, Y., Allen, P., & Fang, C. (2021). LoRA Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.

NTT Data. (2024). Digital humans: the future of human-like technology. Retrieved from https://nttdata-solutions.com/us/blog/digital-humans-humanizing-technology-and-improving-user-experiences/

ScienceDirect. (2023). Generation of virtual digital human for customer service industry. Computers & Graphics, 115, 359-370. DOI: https://doi.org/10.1016/j.cag.2023.07.018

He, Y., Gu, X., Ye, X., Xu, C., Zhao, Z., Dong, Y., Yuan, W., Dong, Z., & Bo, L. (2025). LAM: Large Avatar Model for One-shot Animatable Gaussian Head. In Proceedings of SIGGRAPH 2025.

A Appendices

Additional examples:

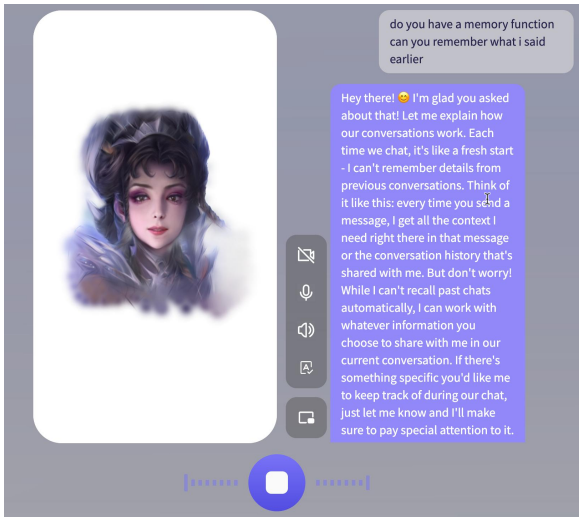

Figure 9: English Version

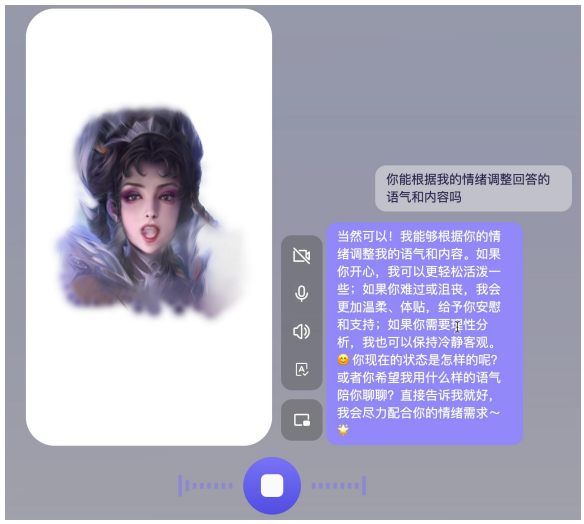

Figure 10: Mandarin Version

Figure 11: Cantonese Version

B Supplementary Material

Minutes of the 1st Project Meeting

Date	June 16, 2025
Place	Zoom
Present	Jing Wang Prof Nevin Lianwen ZHANG AOVA co-supervisors
Note-taker	Jing Wang

Discussion items

The meeting discussed the goals of this project, which include enabling it to come alive by uploading an image, allowing real-time interaction with users, featuring low latency, emotional intelligence, and reflecting the user's characteristics through unique human-computer interaction methods. The potential technical approaches could include VLM, TTS, text-to-image-to-video generation, digital humans, and others.

Meeting adjournment and next meeting

The next meeting will be held in late June.

Minutes of the 2nd Project Meeting

Date	June 27, 2025
Time	15:30
Place	Room 2541
Present	Jing Wang Prof Nevin Lianwen ZHANG
Note-taker	Jing Wang

Discussion items

I reported on the literature review and established the basic plan: uploading an image to generate text using VLM, then using various plugins in Stable Diffusion to enhance the image and optimize it with LoRA fine-tuning. After that, the avatar is generated using LAM and imported into Open Avatar Chat. The professor's feedback was that LoRA fine-tuning may not be realistic due to the large data requirements and time consumption, and a preliminary implementation should be developed as soon as possible.

Meeting adjournment and next meeting

The next meeting will be held in July. The meeting adjourned at 16:00.

Minutes of the 3rd Project Meeting

Date	July 19, 2025
Time	12:20
Place	Zoom
Present	Jing Wang Prof Nevin Lianwen ZHANG
Note-taker	Jing Wang

Discussion items

I completed the task of uploading a photo to generate an avatar using LAM and importing it into Open Avatar Chat. Using a GPU server and the Bailian API, I successfully deployed and tested the system. I also identified limitations and outlined the remaining tasks (GUI and image optimization). The results were praised by the professor.

Meeting adjournment and next meeting

The next meeting will be held in August. The meeting adjourned at 12:40.

Minutes of the 4th Project Meeting

Date	August 8, 2025
Time	21:30
Place	Zoom
Present	Jing Wang Prof Nevin Lianwen ZHANG
Note-taker	Jing Wang

Discussion items

I delivered the final comprehensive report, including the complete system architecture, operational processes, tech stack, as well as the achieved results and future improvement directions. Using Streamlit GUI, I developed an interactive digital human with real-time capabilities, emotional intelligence, low latency, cross-platform support, and customizable voice tones. The outcome was well-received by the professor, who praised my work as the best.

Meeting adjournment and next meeting

The meeting adjourned at 22:00.