Monica: An amazing virtual anchor

1952214 Ziang Lu

1951976 Linfel Li

1953281 Wenjiong Wang

1 Introduction

1.1 Project Introduction

Monica, named after a character in our favorite show "Friends", is a stunning virtual anchor system. She is another self in the online world, please embrace her! If you are afraid to show your face, you can use it anywhere, such as live streaming, lecturing, various video conferences, etc. I have a nice idea. For online classes, opening the video would be an invasion of privacy, but with Monica, the teacher can indirectly see what each student is listening to, or at least make sure the students are looking at the screen.

1.2 Technical Introduction

MediaPipe

MediaPipe offers cross-platform, customizable ML solutions for live and streaming media. It can provide:

- End-to-End acceleration: Built-in fast ML inference and processing accelerated even on common hardware.
- Build once, deploy anywhere: Unified solution works across Android, iOS, desktop/cloud, web and IoT.
- Ready-to-use solutions: Cutting-edge ML solutions demonstrating full power of the framework.
- Free and open source: Framework and solutions both under Apache 2.0, fully extensible and customizable.

We use MediaPipe to realize motion capture and face detection, namely, holistic recognition. It contains:

- Face detection
- Face Mesh
- Iris Detection
- Hand Detection
- Pose Detection

Electron

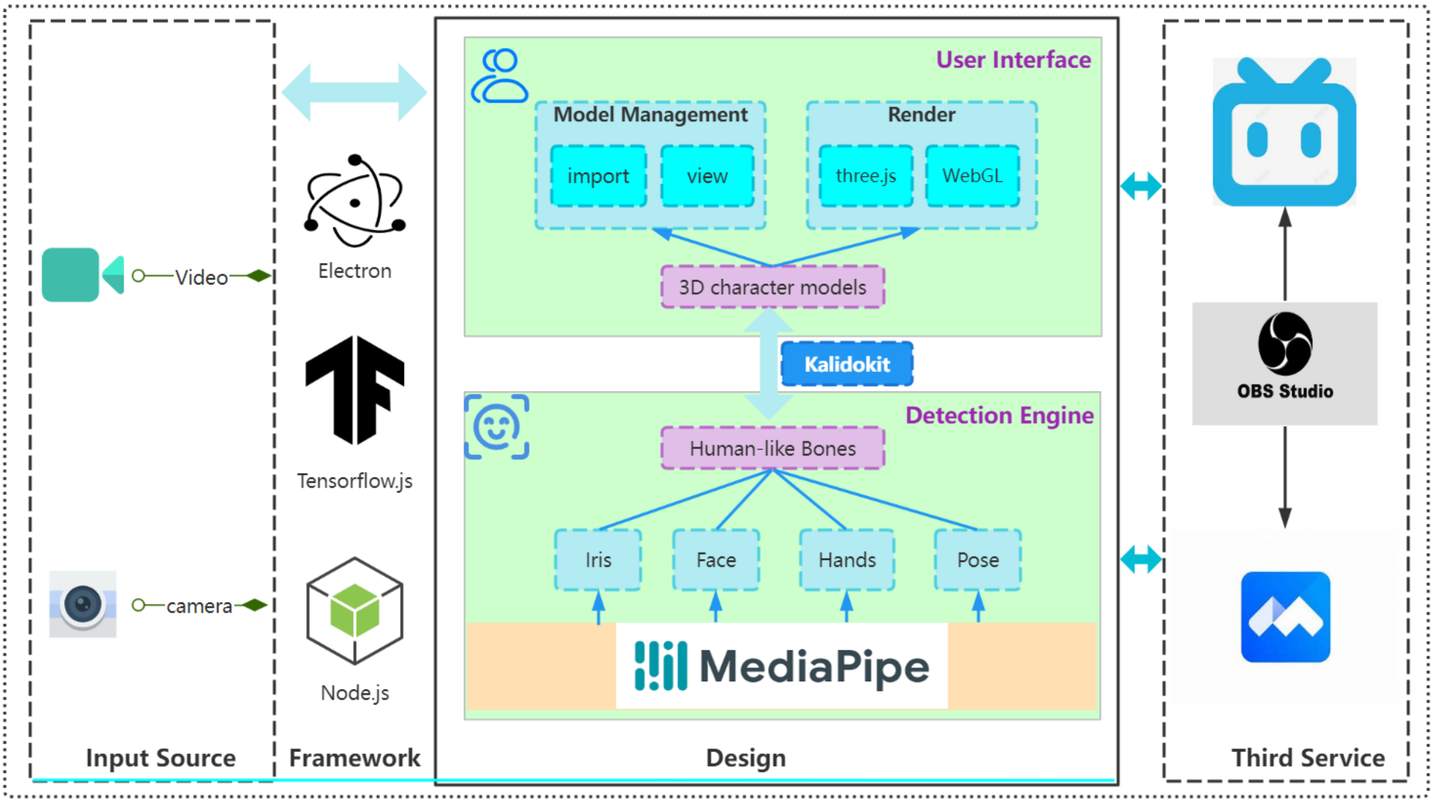
We use Electron to build cross-platform desktop apps with JavaScript, HTML, and CSS.

Kalidokit

Kalidokit is a blendshape and kinematics solver for Mediapipe/Tensorflow.js face, eyes, pose, and hand tracking models, compatible with FaceMesh, Blazepose, Handpose, and Holistic. It takes predicted 3D landmarks and calculates simple euler rotations and blendshape face values.

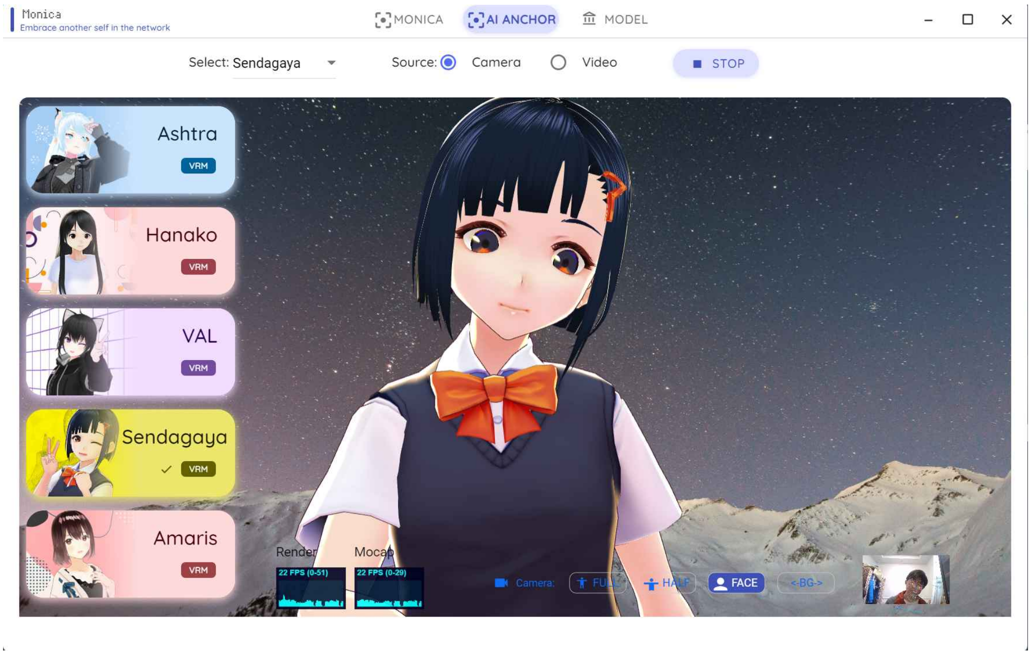
1.3 Structures & Modules

The system structure diagram is shown as follows:



Face Detection Module

Users should open their camera and enter this module (also called face mode). In this module, the system will focus on users' face detection and apply emoticons to virtual models.

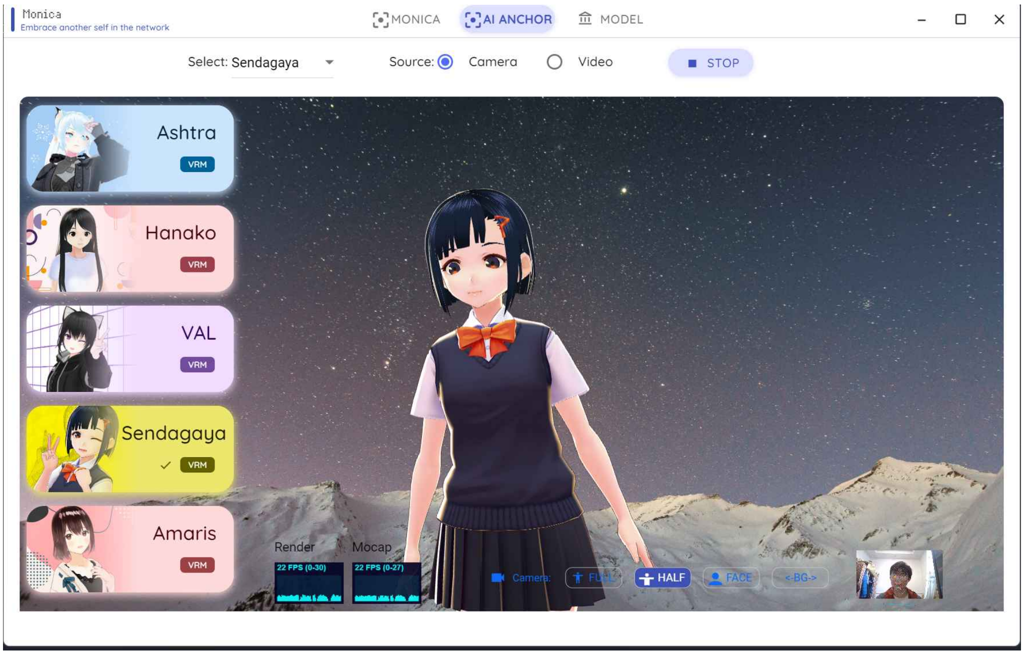


Select: Sendagaya

Source: ● Camera ○ Video

Half-body Capture Module

Users should open their camera and enter this module (also called half-mode). This module actually contains the face detection module because it can show users' half body.

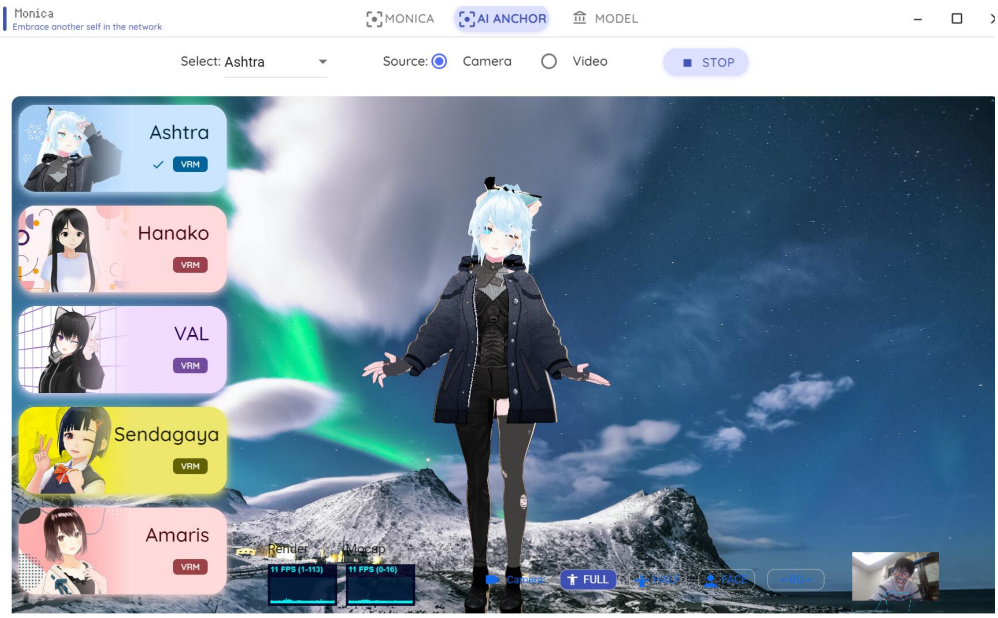


Select: Sendagaya

Source: ● Camera ○ Video

Full-body Capture Module

This module actually contains the half-body capture module and focuses more on posture than facial expression.



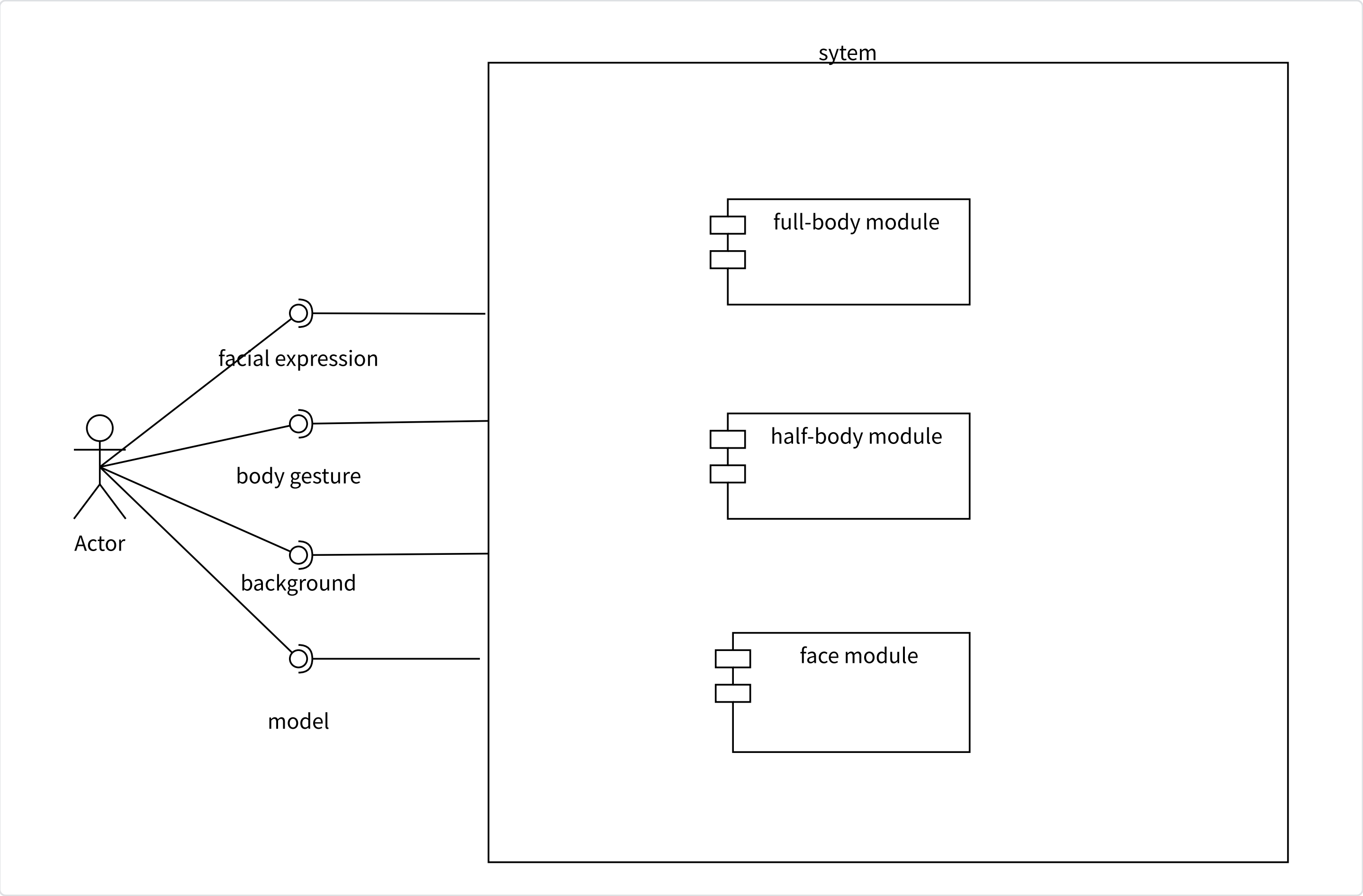
Select: Ashtra

Source: ● Camera ○ Video

■ STOP

1.4 User Interaction

Users can interact with 4 main interfaces provided by the system, only using simple hardware equipment (PC camera).



Furthermore, users can place their virtual characters on meeting applications or social platforms. Take Tencent Meeting as an example, users can use OBS Studio to show their virtual character instead of their real characters.

1.5 Originality

Due to the rapid development of technology as well as the stagnation of the domestic real idol industry and the frequent negative news of netizens and idols, the virtual digital human industry is developing rapidly, and according to the forecast of the "Virtual Digital Human In-depth Industry Report", the overall market size of virtual digital human in China will reach RMB 270 billion by 2030. Certain optimistic analysts even believe that the market size and core market size driven by virtual idols as one of the branches of virtual humans alone is expected to reach 333.47 billion yuan and 20.52 billion yuan respectively in 2023. Virtual digital people are widely used in various fields, including live virtual image, performance, interaction, film, video, game, animation content production, live branding, virtual education, etc. The industry of virtual digital human is in the rapid development period, and it is a veritable blue ocean industry. Therefore, there are many domestic head network high-tech companies and entertainment companies such as Byte Jump, iFLYTEK, and Lehua Entertainment, etc. vying to develop virtual digital human products.

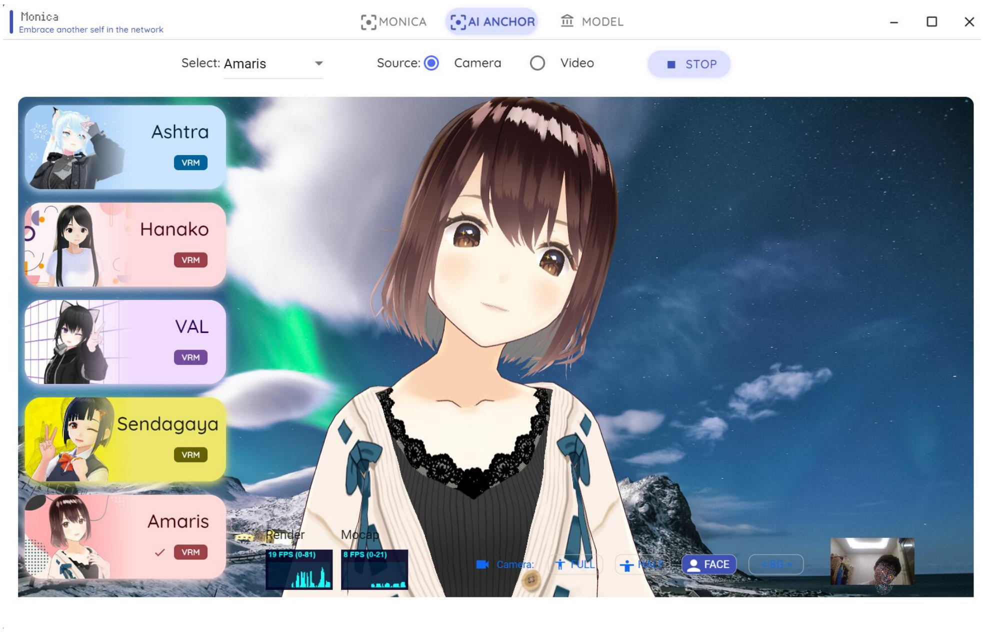
However, at this stage, China's virtual digital human products still have the following defects: high virtual image development cost, no cross platform use ability, and high motion capture cost. In view of the above defects, the project team has conducted targeted thinking and put forward its own solutions.

Compared with the above defects, the innovation of our project is reflected as follows:

1. Cross platform use: the virtual images launched by major manufacturers are only used in their own products, which can not realize platform migration, which brings great inconvenience to users. However, the products of this project can be used by opening the camera, which is very convenient;
2. Low use cost: This product can achieve high-precision motion capture without matching with expensive motion capture clothing, presenting a more vivid and real 3D virtual image, and the user's use cost is low;
3. Multi view: this project realizes multi view switching, such as "whole body", "half body" and "face", which is convenient for users to select their desired view according to their needs and meet the needs of users in various scenarios;

2 The Implemented Requirements

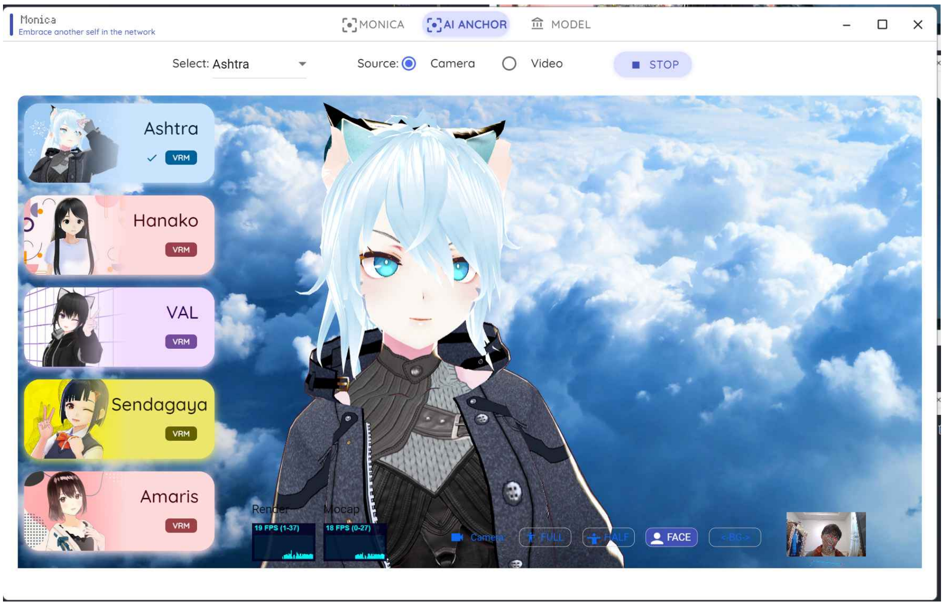
1. Use of virtual character: when the user needs to turn on the camera, he can choose to use the virtual character in this project to replace himself. The virtual character will capture the user's facial expression, gesture, posture and action, and restore the user's behavior in front of the camera with high precision. Users can use this product in online chat, e-commerce delivery, online live broadcast and other scenarios, which can not only meet the needs of users' face-to-face interaction with others, but also protect users' privacy;



Select: Amaris

Source: ● Camera ○ Video

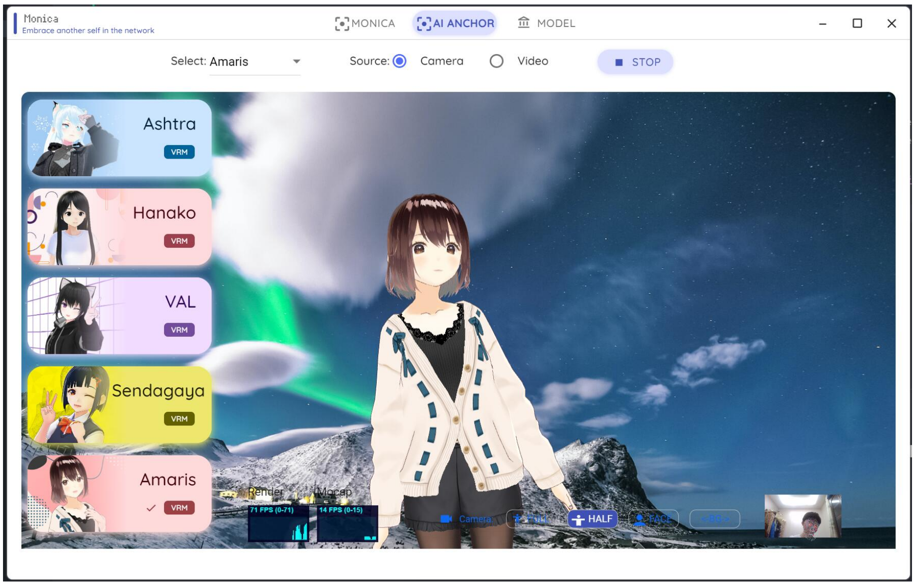
2. Virtual character replacement: when users want to change their virtual character , they can go to the "virtual character library" to select the character they like and replace it. This function keeps users fresh about the product and increases user stickiness;



Select: Ashtra

Source: ● Camera ○ Video

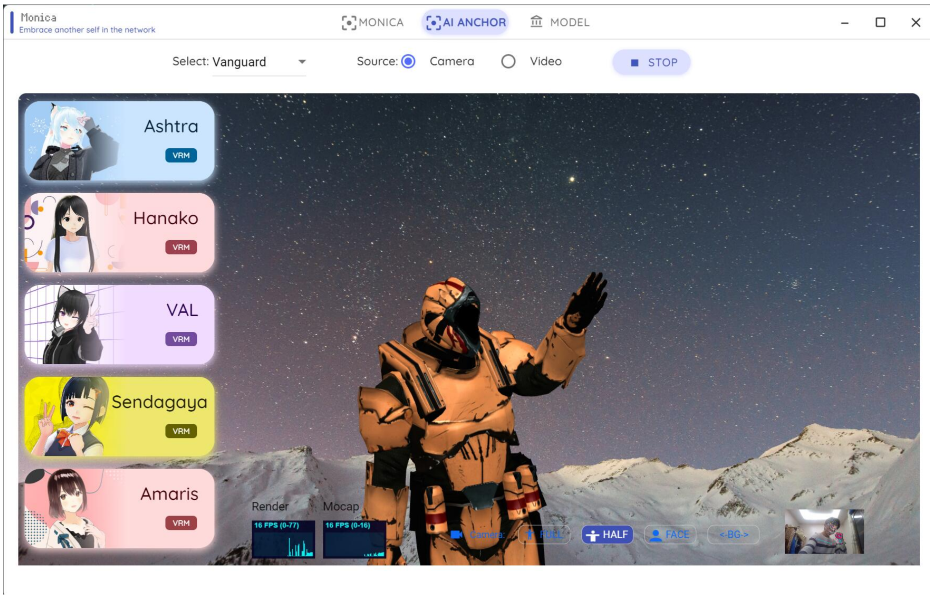
3. Perspective switching: the user can freely switch the three perspectives of "face", "half body" and "full body" to meet the user's needs in different scenarios: if the user only needs to chat face-to-face, he can choose the "face" perspective. When the user needs to interact with the audience, such as teaching fitness movements, dancing and other activities, the user can choose the "full body" or "half body" perspective. Flexible perspective selection can help users use the product more conveniently;



Select: Amaris

Source: ● Camera ○ Video

4. Background switching: users can also freely choose the background of the virtual image, so as to match more different possibilities. This function further increases the usability and friendly interaction of users;



Select: Vanguard

Source: ● Camera ○ Video

3 Advantages & Disadvantages

ID	Advantages	Disadvantages
1	Low requirements for hardware and software equipment leads to low entry threshold for users	Recognition accuracy is not as good as those of wearing sensors on users' body
2	Ability to recognize most expressions and gestures leads to strong usability	Virtualization of sound is not supported
3	The virtual model is exquisite and can match the actual behavior of users	
4	Lightweight and easy to use for new hands	

5	Cater to the recent popular virtual anchor trade and be widely loved	
6	Cross platform	

4 Project Improvement

After referring to the mainstream virtual image generation products in the current market, members of the project team believe that the main enhancement of this project is to improve its real-time recognition capability, which in turn includes the following four aspects：

- Face Detection：At present, the face detection of this product is limited to a single face, and cannot recognize multiple faces appearing in the same camera at the same time in real time, and the speed of detection is relatively slow, while the realization of fast and good performance of face detection function is the basis for the development of other subsequent functions, such as: 3D face key point estimation (e.g., Face Mesh), facial feature or expression classification, and face area segmentation, etc.
- Face Mesh：At present, the number of facial keypoints collected and recognized in this project is relatively small, and the recognition speed is relatively slow and the recognition accuracy is relatively low. If we want to add AR makeup and other functions requiring high accuracy and rendering speed to this product in the future, we need a more lightweight model architecture to accelerate rendering, a more lightweight statistical analysis method to drive robust, high-performance and portable logic, and an increase in the number of facial keypoints recognized.
- Iris Detection：Iris recognition has many applications in the real world, including computational photography (flash reflection) and augmented reality effects (avatars) that rely on accurate tracking of the iris inside the eye. The product's dynamic tracking of the iris and depth measurement capabilities are still relatively low, and enhancements to this aspect will help this project achieve features such as measuring the distance between the person and the camera, and wearing virtual facial decorations (e.g., glasses)
- Pose Detection：The overall recognition of human posture in this project is relatively rigid, and the enhancement for this aspect will help this product to combine with the application of yoga, dance, fitness and other quantitative physical exercise to further develop the market.

5 Contribution

Three team members have the same contribution ratio and share the total score of the project.

Ziang Lu	100%
Linfei Li	100%
Wenjiong Wang	100%