Multimodal Large Model+video understanding（Voice Version）

Before running the function, you need to close the App and large programs. For the closing method, refer to [4. Preparation] - [1. Manage APP control services].

1. Function Description

After the program runs, input questions about video through the terminal. The large model will first record a video, then analyze the video content with the question, and finally reply with the answer to the question. You can also first record a video of specified duration, then ask the large model questions related to the video based on the video content.

2. Startup

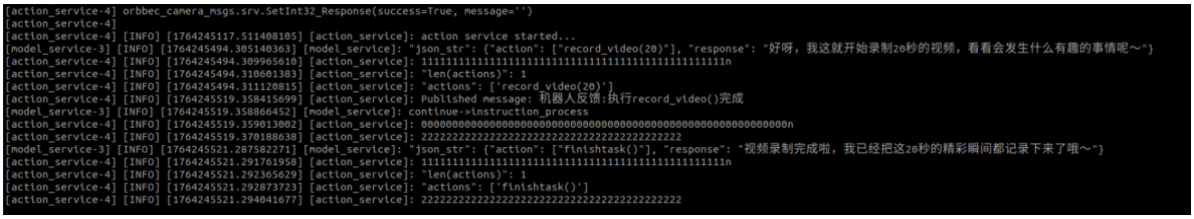
Users with Jetson-Nano mainboard version need to enter the docker container first and then input the following command. Users with Orin mainboard can directly open the terminal and input the following command:

```
ros2 launch largemodel largemodel_control.launch.py
```

After waking up the module, give video recording commands. You can refer to the following example:

```
Record a 20-second video
```

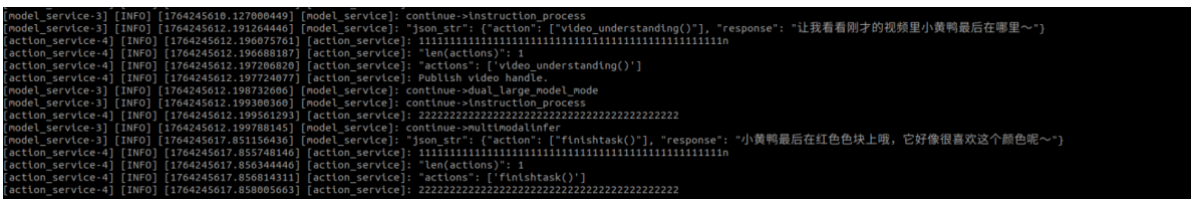
After the buzzer beeps once,



After the buzzer beeps once, recording ends. Then based on this video content, wake up the voice module and say:

```
In the video just recorded, which color block was the small yellow duck on at the end?
```

The question here should be asked based on the actual recorded video content,



The location where recorded video files are saved: For Orin mainboard, videos are saved in /home/jetson; for Jetson-nano mainboard, videos are saved in /root inside the docker container.

3. Core Code Analysis

You can refer to the content in **3. Core Code Analysis** from tutorial [17. AI Model - Text Version] - [1. Multimodal Large Model+video understanding]. The voice version and text version have the same action functions, only the task command input method is different.

4. Combination with Robotic Arm

We can also combine robotic arm + video understanding to develop new gameplay. For example, prepare several opaque paper cups. After the video starts, we place an object under one of the opaque paper cups, then slowly change the positions of the paper cups. After the video recording ends, we ask the large model which cup the object is finally under, and finally have the robotic arm point to that cup. To test this function, you can wake up the voice module and input by voice:

```
Let's play a game. First record a video, then guess which cup has the small
yellow duck at the end of the video?
```

The program will record a video, then analyze which cup contains the small yellow duck, and finally control the robotic arm to point to the cup hiding the little duck.

The task planning should be:

```
1. Call record_video(20) function to record a 20-second video;
2. Call video_understanding() function to analyze the recorded 20-second video to
determine which cup hides the object at the end of the video.
If the analysis result shows that an object is indeed hidden under a certain cup,
continue with the following steps:
3. Call point_to(x1, y1, x2, y2) function to point to the cup hiding the object,
where x1, y1, x2, y2 are the outer bounding box coordinates of the cup hiding the
object.
```