Multimodal Visual 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, wake up the voice module and ask questions about vision or describe the current scene. The program will capture an image of the current environment and upload it to the Alibaba Cloud platform, then the vision model will analyze the image with the question and finally provide an answer.

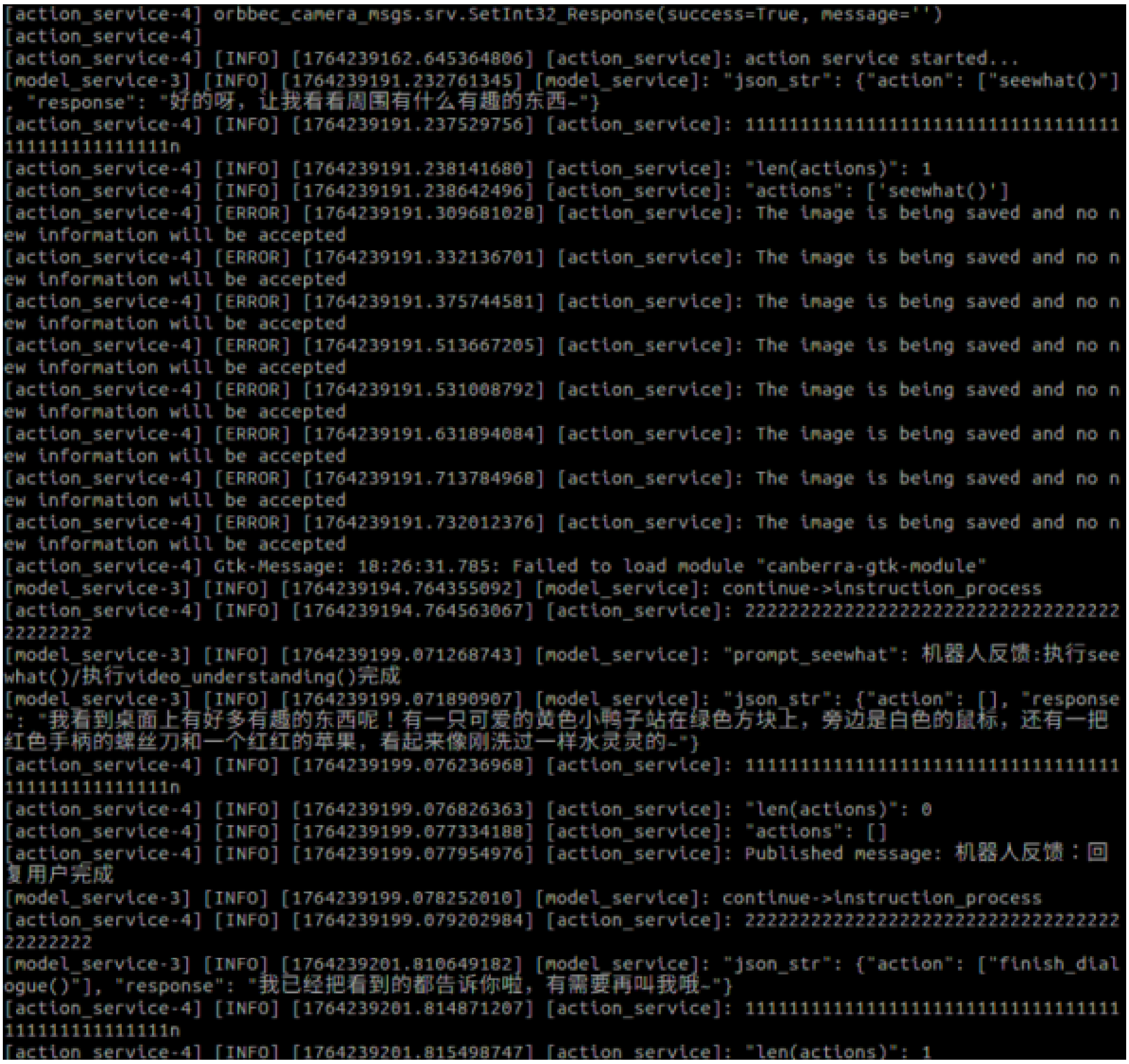
2. Startup

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 vision-related commands. You can refer to the following example,

```
Describe the scene you see.
```



```
[action_service-4] orbbec_camera_msgs.srv.SetInt32_Response(success=True, message='')
[action_service-4]
[action_service-4] [INFO] [1764239162.645364806] [action_service]: action service started...
[model_service-3] [INFO] [1764239191.232761345] [model_service]: "json_str": {"action": ["seewhat()"]
, "response": "好的呀，让我看看周围有什么有趣的东西~"}
[action_service-4] [INFO] [1764239191.237529756] [action_service]: 111111111111111111111111111111111
111111111111111n
[action_service-4] [INFO] [1764239191.238141680] [action_service]: "len(actions)": 1
[action_service-4] [INFO] [1764239191.238642496] [action_service]: "actions": ['seewhat()']
[action_service-4] [ERROR] [1764239191.309681028] [action_service]: The image is being saved and no n
ew information will be accepted
[action_service-4] [ERROR] [1764239191.332136701] [action_service]: The image is being saved and no n
ew information will be accepted
[action_service-4] [ERROR] [1764239191.375744581] [action_service]: The image is being saved and no n
ew information will be accepted
[action_service-4] [ERROR] [1764239191.513667205] [action_service]: The image is being saved and no n
ew information will be accepted
[action_service-4] [ERROR] [1764239191.531008792] [action_service]: The image is being saved and no n
ew information will be accepted
[action_service-4] [ERROR] [1764239191.631894084] [action_service]: The image is being saved and no n
ew information will be accepted
[action_service-4] [ERROR] [1764239191.713784968] [action_service]: The image is being saved and no n
ew information will be accepted
[action_service-4] [ERROR] [1764239191.732012376] [action_service]: The image is being saved and no n
ew information will be accepted
[action_service-4] Gtk-Message: 18:26:31.785: Failed to load module "canberra-gtk-module"
[model_service-3] [INFO] [1764239194.764355092] [model_service]: continue->instruction_process
[action_service-4] [INFO] [1764239194.764563067] [action_service]: 222222222222222222222222222222222
22222222
[model_service-3] [INFO] [1764239199.071268743] [model_service]: "prompt_seewhat": 机器人反馈:执行see
what()/执行video_understanding()完成
[model_service-3] [INFO] [1764239199.071890907] [model_service]: "json_str": {"action": [], "response
": "我看到桌面上有好多有趣的东西呢！有一只可爱的黄色小鸭子站在绿色方块上，旁边是白色的鼠标，还有一把
红色手柄的螺丝刀和一个红红的苹果，看起来像刚洗过一样水灵灵的~"}
[action_service-4] [INFO] [1764239199.076236968] [action_service]: 111111111111111111111111111111111
111111111111111n
[action_service-4] [INFO] [1764239199.076826363] [action_service]: "len(actions)": 0
[action_service-4] [INFO] [1764239199.077334188] [action_service]: "actions": []
[action_service-4] [INFO] [1764239199.077954976] [action_service]: Published message: 机器人反馈：回
复用户完成
[model_service-3] [INFO] [1764239199.078252010] [model_service]: continue->instruction_process
[action_service-4] [INFO] [1764239199.079202984] [action_service]: 222222222222222222222222222222222
22222222
[model_service-3] [INFO] [1764239201.810649182] [model_service]: "json_str": {"action": ["finish_dial
ogue()"], "response": "我已经把看到的都告诉你啦，有需要再叫我哦~"}
[action_service-4] [INFO] [1764239201.814871207] [action_service]: 111111111111111111111111111111111
111111111111111n
[action_service-4] [INFO] [1764239201.815498747] [action_service]: "len(actions)": 1
```

The captured photo is saved at:
`LargeModel_ws/install/largemodel/share/largemodel/resources_file/image.png`

3. Core Code Analysis seehat

You can refer to the content in **3. Core Code Analysis seehat** from tutorial [17. AI Model - Text Version] - [2. Multimodal Visual Understand]. The voice version and text version have the same action functions, only the task command input method is different.