For chatbot demo:

```
pip install -r requirements_o2.6.txt

python web_demos/minicpm-o_2.6/chatbot_web_demo_o2.6.py
```

Open `http://localhost:8000/` in browser and enjoy the vision mode chatbot.

Inference

Model Zoo

Model	Device	Memory	Description
MiniCPM-o 2.6	GPU	18 GB	The latest version, achieving GPT-4o level performance for vision, speech and multimodal live streaming on end-side devices.
MiniCPM-o 2.6 gguf	CPU	8 GB	The gguf version, lower memory usage and faster inference.
MiniCPM-o 2.6 int4	GPU	9 GB	The int4 quantized version, lower GPU memory usage.
MiniCPM-V 2.6	GPU	17 GB	Strong end-side multimodal performance for single image, multi-image and video understanding.
MiniCPM-V 2.6 gguf	CPU	6 GB	The gguf version, lower memory usage and faster inference.
MiniCPM-V 2.6 int4	GPU	7 GB	The int4 quantized version, lower GPU memory usage.