# **Deploying LLaVA with Gradio web server**

## ****1. Clone and install LLaVA****

git clone https://github.com/haotian-liu/LLaVA.git

cd LLaVA

pip install -e .

pip install gradio transformers torch

## ****2. Start the controller****

## python -m llava.serve.controller --host 0.0.0.0 --port 21001

## This will manage communication between web server and model worker.

## ****3. Start the model worker****

python -m llava.serve.model\_worker \

--host 0.0.0.0 --controller http://localhost:21001 \

--port 40000 \

--model-path liuhaotian/llava-v1.5-13b \

--load-4bit

## ****4.**** Launch the Gradio Web Server

Serves your UI and provides a REST API endpoint:

python gradio\_web\_server.py \

--host 0.0.0.0 --port 7860 \

--controller-url http://localhost:21001 \

--model-list-mode reload

## ****5.**** Grab the API Endpoint

Check network logs or hover over the “Send” button to find the URL:

POST http://localhost:7860/api/chat

## ****6.**** Integrate with Streamlit

## API\_URL = "http://localhost:7860/api/chat"

## response = requests.post(API\_URL, json={"data":[final\_prompt]})

## if response.ok:

## answer = response.json()["data"][0]

## # Use `answer` just like before: display and store.