Assignment: Automatic Hand Tracking

Goal: Build an automatic pipeline that tracks hand movements in a video.
Tools: cv2, Google MediaPipe, SAM 2

Preliminaries

Data

Download the video test.mp4.

Environment Set Up

- `conda create -n sam2 python=3.12`
- `conda activate sam2`
- `pip3 install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cu118`
    - May take a bit.
    - See PyTorch installation if this doesn't work.
- Follow the instructions for setting up SAM 2 here
- Install **mediapipe** with `pip install -q mediapipe`
- And **cv2** with `pip install opencv-python`

Part 1: Detect Hands in the First Frame

References: Google MediaPipe, Google Colab Sample

Using the references above, write code that can output information on hand location(s) in an image (represented however you wish - e.g., numpy array). This instruction is intentionally vague. Your goal is to return an output that can be used as a SAM 2 prompt (e.g., clicks, bounding boxes). Feel free to look into SAM 2 Video Predictor Example for example prompts.

Part 2: Use Part 1 and SAM 2 to Track Hands

References: SAM 2 Repo, SAM 2 Video Predictor Example

Write a function that uses SAM 2 and the results from Part 1 to generate masks for every frame of the video.

- There can be multiple masks per frame.
- The results from Part 1 are your *input prompts* into SAM 2.

- Ideally, the parameters of this function are an input and output path. The function will write a new video with the masks to the output path.

Deliverables

1) **Link to Github code**: code style and organization are important and will be evaluated! Include a README / pip requirements file for set up.
2) **Output video demo**: an output video with the masked hands