Visual Question Answering (VQA) Mini Project

Duration: 1 Day (Mini Project)

Level: Beginner to Intermediate

Core Focus: Image Understanding + Vision-Language Modeling

Objective

To build a web-based application that allows users to upload an image and ask a question related to that image. The application will generate an answer using a pre-trained Visual Question Answering (VQA) model. This combines computer vision and natural language processing to deliver interactive image-based insights.

Tech Stack & Tools

Component	Tools / Technologies
Image Handling	PIL or OpenCV
Vision-Language Model	Pre-trained VQA model (BLIP-2)
Frontend	Streamlit or simple web UI
Voice Output (Optional)	gTTS or any TTS module
Voice Input (Optional)	SpeechRecognition library

Features to Implement

1. Image Upload: Allow users to upload an image for analysis.

2. Question Input: Enable users to ask a question related to the uploaded image.

3. VQA Model Inference: Use a pre-trained model to answer the question based on visual content.

4. Answer Display: Present the answer clearly in the user interface.

5. Optional: Add voice input for asking the question and voice output for the answer.

Step-by-Step Instructions

Phase 1: Environment Setup

- Install necessary libraries for computer vision and NLP.

- Set up the basic user interface for image and question input.

Phase 2: Core Logic Development

- Preprocess uploaded image (if needed).

- Pass image and question to the VQA model.

- Extract and display the model's answer.

Phase 3: Interface Integration

- Integrate components for file upload, question entry, and response display.

- Optionally include voice features for input/output.

Deliverables

- Source code hosted on a version control platform (e.g., GitHub).

- Sample images and questions used for testing.

- Short demo video or screenshots of working app.

- A user guide documenting usage, expected inputs/outputs, and known limitations.