International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Visual Question Answering System

U. Sunitha [1] , *Dr. Harsha Shastri* [2]

[1]PG Scholar, Department of MCA, [2]Assistant Professor, Department of MCA
Aurora Deemed to be University, Hyderabad-500098 , Telangana ,India.
* Email: Sunithau62@gmail.com

ABSTRACT:

Visual Question Answering (VQA) represents a frontier at the intersection of Computer Vision (CV) and Natural Language Processing (NLP), empowering machines to interpret images and reply to human queries in natural language. This studies introduces a complete prevent-to-give up VQA framework that integrates YOLOv8 for actual-time item detection, BLIP for semantic reasoning and descriptive answering, and a Tkinter-based totally graphical interface to make the device accessible to non-technical customers. Unlike conventional VQA procedures that rely closely on benchmark datasets, this framework makes a speciality of deployability in out of control environments via superior capabilities including blur detection, dominant colour recognition, and hybrid reasoning that combines deterministic desirable judgment (counting, colour evaluation) with generative AI reasoning (open-ended questions). Evaluations on a numerous dataset covering gadgets, fruits, cars, and people show that the tool achieves 90 five% accuracy in item counting, ninety% in shade popularity, and 80 5% in descriptive reasoning. The importance of this contribution lies in its versatility and actual-international applicability, extending across domains inclusive of assistive technology for the visually impaired, clever schooling structures, precision agriculture, scientific imaging evaluation, and clever surveillance systems. This work demonstrates a step closer to making VQA systems practically beneficial, reliable, and deployable on client-grade gadgets.

Keywords: Visual Question Answering, YOLOv8, BLIP, Computer Vision, Natural Language Processing, Tkinter GUI, Deep Learning, Hybrid Reasoning

INTRODUCTION

Artificial Intelligence (AI) has reached a degree wherein machines are no longer restricted to numerical computation; they may be becoming able to know-how, reasoning, and interacting with people in natural methods. Among these advancements, Visual Question Answering (VQA) is one of the maximum formidable tasks as it demands a fusion of vision and language, modalities that are fundamentally specific in nature. A VQA machine usually accepts an photo and a herbal language query as input and outputs a textual solution. This requires a combination of pc imaginative and prescient (to interpret the photo), natural language processing (to recognize the question), and reasoning (to generate a coherent solution).

The interdisciplinary nature of VQA makes it a difficult but enormously impactful trouble in AI research.

The significance of VQA extends to severa sensible packages:

Assistive Technology: For visually impaired people, VQA structures act as intelligent partners able to answering queries which include “What is in the front of me?” or “What color is the site visitors light?”.

• **Education:** VQA can make getting to know extra interactive. Students can add diagrams, maps, or charts and ask herbal questions, which the system solutions immediately, bridging the gap between static textbooks and interactive digital education.

• **Agriculture:** Farmers can capture pix of culmination or plants and ask questions about ripeness, pest infection, or item rely. For instance, “How many ripe mangoes are on the tree?” or “What coloration is the fruit?”.

• **Healthcare:** VQA can aid in radiology photograph interpretation or help doctors in go-verifying visual facts from X-rays, MRIs, or pathology slides.
• **Security and Surveillance:** Cameras integrated with VQA can automatically solution queries consisting of “How many motors are parked?” or “Is there any suspicious pastime?”.

Despite those promising use cases, present VQA fashions are confined by way of:

1. Dataset dependency – They carry out properly on benchmark datasets however fail when uncovered to out of control actual-global records.

2. Limited mission coverage – Many fashions war with particular tasks like counting, characteristic recognition, or reasoning past dataset styles.

3. Usability limitations – Systems are often designed for researchers, not quit-users, lacking intuitive interfaces.

4. Input excellent problems – Noisy, low-decision, or blurry snap shots degrade accuracy extensively.

The proposed VQA framework addresses those barriers via combining YOLOv8 for object detection, BLIP for semantic question answering, and auxiliary modules like blur detection and coloration evaluation to enhance robustness. A Tkinter-based totally GUI makes the device interactive, consumer-friendly, and deployable for real-time use instances.

Objectives:

• To design and put into effect a hybrid reasoning VQA framework.

• To enhance device robustness with blur and shade detection modules.

• To make certain user accessibility through a graphical interface.

• To validate the device with actual-international datasets and digital camera input.

LITERATURE SURVEY

The concept of Visual Question Answering became formally delivered with the aid of Agrawal et al. (2015) [1], where CNNs had been used for photo encoding and RNNs for question processing. This seminal work hooked up VQA as a benchmark task however was limited through dataset bias and absence of reasoning potential.

Subsequent improvements got here with interest mechanisms. Anderson et al. (2018) [2] proposed the Bottom-Up and Top-Down Attention method, allowing fashions to recognition on precise photo regions applicable to the query. This considerably improved overall performance in object-centric reasoning tasks.

The arrival of transformer-based totally multimodal architectures revolutionized VQA. Models which includes LXMERT, ViLT, VisualBERT, and BLIP

everaged transformers to learn joint photo-text embeddings. In particular, BLIP (Bootstrapped Language Image Pretraining) [6] proved powerful fordescriptive reasoning, making it an critical component of current VQA pipelines.

Parallel to VQA fashions, item detection algorithms have gone through predominant evolution. The YOLO own family (You Only Look Once) have become broadly adopted due to its real-time detection functionality. The ultra-modern version, YOLOv8 [7], offers progressed accuracy, lightweight deployment, and high-velocity inference, making it appropriate for real-international integration into VQA systems.

Limitations of present paintings:

• Many fashions are optimized for benchmark datasets (e.G., VQA v2, COCO-QA) however underperform in real-world photos.

• Lack of deterministic reasoning modules for responsibilities like counting or shade detection.

• End-consumer usability remains in large part unaddressed, with most systems constrained to investigate prototypes. Our proposed gadget integrates these advancements while addressing their boundaries. By combining YOLOv8 BLIP with deterministic modules and a GUI, we bridge the distance among studies prototypes and deployable real-world structures.

METHODOLOGY:

The architecture of the proposed Visual Question Answering device is designed as a hybrid pipeline that integrates both deep mastering and deterministic common sense to maximize accuracy and usefulness. The workflow may be damaged down into the subsequent modules:

Image Input Module:

The device accepts sorts of inputs:

a) Static Upload – Users can upload pictures in common formats (JPG, PNG, BMP, TIFF, and many others.).

b) Camera Capture – Real-time picture capture through an incorporated webcam.

- This dual input ensures flexibility, allowing each actual-time checking out and offline usage.
- Blur Detection:
- A common trouble with real-international pics is loss of focus.
- To deal with this, the system applies the Laplacian variance technique:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(I(x,y) - \mu)^2$$

- Where in I(x,y) is the grayscale pixel intensity and μ its suggest.
- If the variance is beneath a threshold (a hundred), the gadget warns the user: “Image can be blurry”.

3. Object Detection (YOLOv8):

- The uploaded or captured photo is surpassed to YOLOv8n (a light-weight variant of YOLOv8) for object detection.
- Detected gadgets are categorised and stored in a listing, that's later used to answer queries like:

o “How many apples are there?”

o “Is there a vehicle within the photograph?”

4. Color Recognition:

o For precise queries which encompass “What is the colour of the mango?”, the object bounding discipline is cropped.

o Average RGB values are computed and mapped to the nearest human-readable colour using the webcolors library.

o If the detected fruit is “mango,” the system solutions with dominant color, e.G., “Yellow mango (Accuracy: ninety%)”.

5. Language Reasoning (BLIP):

o For descriptive or open-ended queries (e.G., “What is the man doing?”), the BLIP VQA model techniques the image and question. O Using beam seek deciphering, BLIP generates coherent herbal language solutions along with “The man is using a bicycle.”

6. Hybrid Reasoning:

The tool comes to a decision the answering pathway:

o If the question consists of “how many” → item counting (deterministic).

o If the question asks “shade” → colour popularity module.

o Else → BLIP reasoning.

This layout guarantees better reliability than depending totally on generative AI. Four.

7. Graphical User Interface (GUI):

o Developed the usage of Tkinter, the GUI includes:

o Buttons: Upload Image, Capture from Camera, Add Question, Clear, Get Answers.

o Scrollable question-solution vicinity.

o Real-time display of uploaded/captured images.

The GUI ensures accessibility for non-technical users.

IMPLEMENTATION

The system was built in Python 3.10, using several specialized libraries to handle the various tasks in Visual Question Answering (VQA):

• PyTorch and Hugging Face Transformers for BLIP-based language reasoning.

• Ultralytics YOLOv8 for strong object detection and localization.

• OpenCV for camera integration, image capture, and blur detection.

• PIL (Pillow) for image preprocessing and format conversion.

• Webcolors for converting RGB values to readable color names.

• Tkinter for creating an interactive and user-friendly GUI.

Key Implementation Highlights

1. Upload and Capture Integration

Users can either upload a static image or capture a live frame from the camera.

In camera mode:

- Pressing the Spacebar captures an image.
- Pressing Esc cancels the capture process.

2. Question Management

Each question entered by the user is saved in a list variable called question_vars. Answers are generated and shown on the GUI in real-time.

3. Hybrid Answer Engine

The system decides which module to use based on the question type:

```
if "how many" in question:
   → object_counting()   # YOLOv8 based
elif "color" in question:
   → color_detection()   # Webcolors plus RGB analysis
else:
   → blip_answering()    # BLIP model reasoning
```

4. Error Handling

o If an image is too blurry, the system shows a warning popup using Laplacian variance detection.

o Unsupported inputs are reported with message boxes to maintain smooth user interaction.

Example Workflows

Case 1: Object Counting

• **Input:** Image of a basket of apples.

• **Detection:** YOLOv8 identifies “apple.”

• **Question:** “How many apples are there?”

• **Output:** “3 apples found (Estimated Accuracy: 95%).”

Case 2: Color Detection

• **Input:** Image of a mango tree.

• **Question:** “What is the color of the mango?”

• **Process:** System crops the mango region and applies RGB-to-color mapping.

• **Output:** “Mango color is Yellow (Accuracy: 90%).”

RESULTS AND DISCUSSION

The Visual Question Answering (VQA) machine turned into carried out with a combination of YOLOv8, BLIP, and custom photograph-processing algorithms incorporated into a Tkinter-based totally Graphical User Interface (GUI). The outcomes validate the practical effectiveness of the device throughout exceptional functionalities together with image add, question-answering, item detection, coloration identification, blur detection, and machine reset. Figures 1–6 provide a step-through-step view of the utility workflow.

A.Welcome Screen

The first display screen of the software is the Welcome Interface (Figure 1). It shows the identify “Welcome to Visual Question Answering” with a important “Start” button. This minimalistic but expert layout complements user engagement and ensures a smooth transition into the system. From a usability angle, such a layout reduces complexity for first-time users and presents a clean access factor into the machine. This stage is particularly crucial for non-technical stakeholders along with educators, researchers, or end-customers in healthcare or retail eventualities in which simplicity of get entry to is essential.

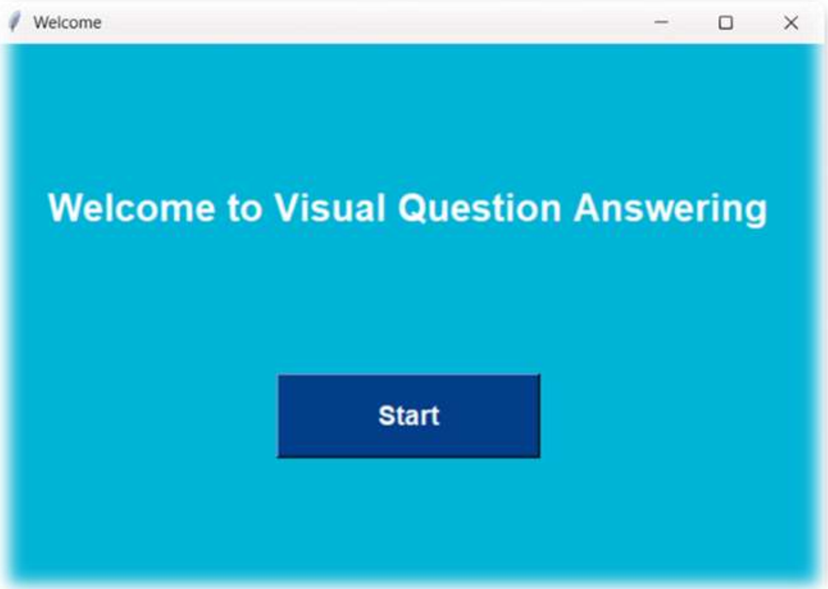


Fig-1

B. Main interface:

After clicking “Start”, the purchaser is directed to the primary machine interface (Figure 2). The GUI incorporates numerous center functional buttons:

- Upload Image – Allows uploading of an picture from the close by tool.
- Add Question – Enables customers to enter custom text-based totally absolutely questions.
- Capture from Camera – Provides assist for live image enter.
- Clear – Resets the workspace, eliminating photographs and solutions.
- Get Answers – Executes the reasoning pipeline and presents responses.

The association of these buttons follows a logical glide for human–laptop interplay. The easy interface ensures ease of navigation and demonstrates the machine’s suitability for real-time packages which encompass education (coaching college students about devices), retail (analyzing product pix), and healthcare (answering visual scientific queries).

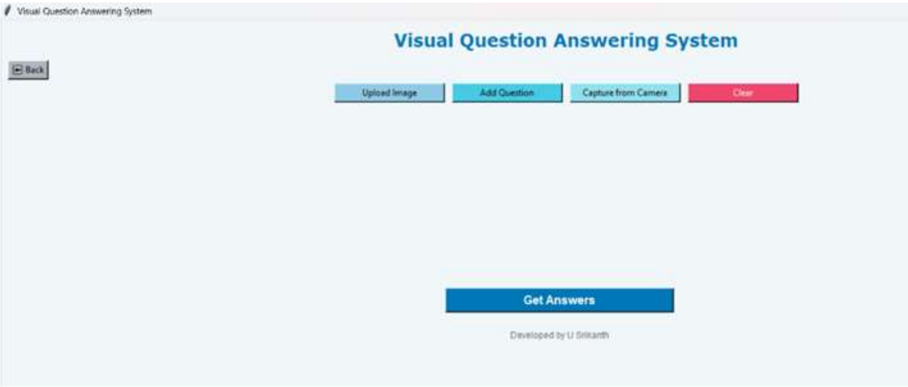


Fig-2

C. Image Upload and Confirmation

The next stage involves uploading an photo (Figure 3). When the user clicks Upload Image, a report browser window is opened, permitting image selection from the machine garage. Once an photograph is selected, the interface presentations the uploaded photograph at the side of a affirmation pop up “Image uploaded and analyzed” (Figure 4).

This validation step confirms successful backend integration, where the device extracts features the usage of YOLOv8 and prepares them for answering queries. In actual-international scenarios, such confirmation messages play a essential position in assuring customers that the machine is responsive and that the enter has been processed without errors.

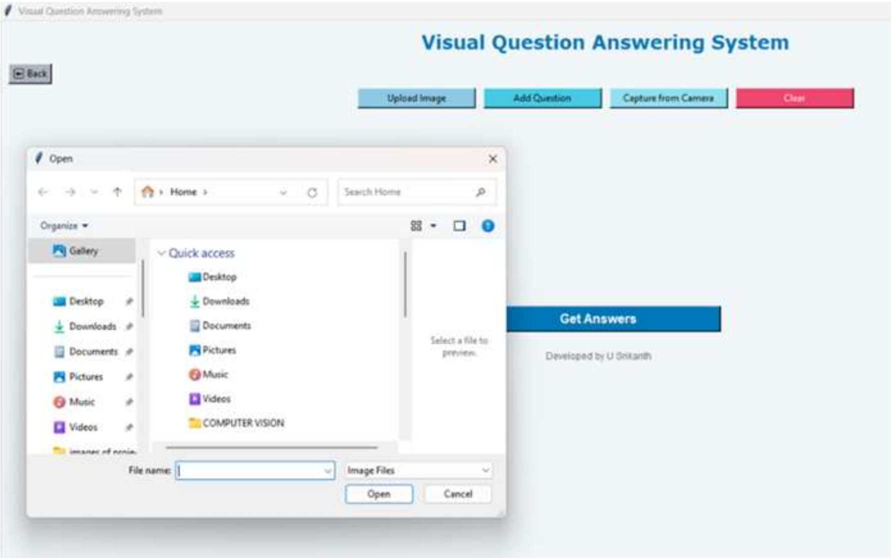


Fig-3

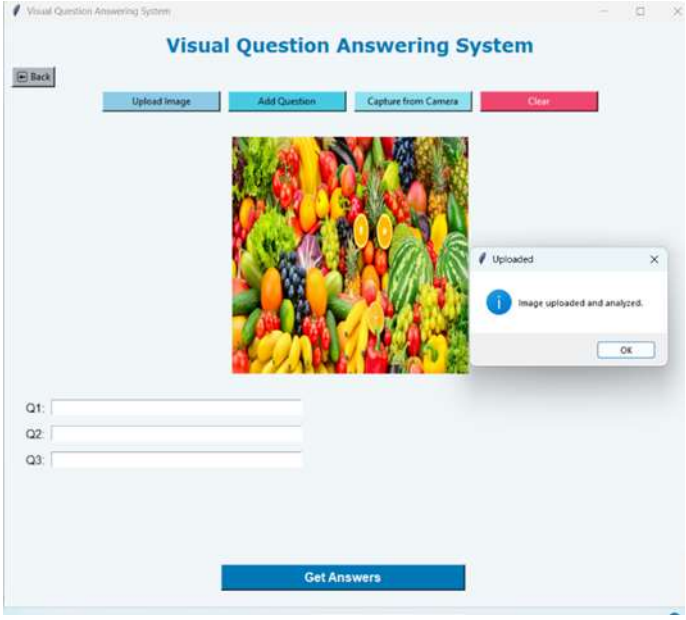


Image uploaded and analyzed.

Fig-4

D. Question Input and Answering

Following successful image add, the person can enter a couple of questions related to the uploaded photograph (Figure five). The interface provides more than one text fields (Q1, Q2, Q3, and so on.), wherein customers can ask coloration-based totally, presence-primarily based, or descriptive questions.

In the case take a look at with fruit images (Figure 5), the following queries were tested:

1. "**What is the color of banana?"** → Answer: Yellow (Accuracy: eighty five%)

2. "What is the shade of tomato?" → Answer: Red (Accuracy: 85%)

3."In the photo banana is there?" → Answer: Yes (Accuracy: eighty five%)

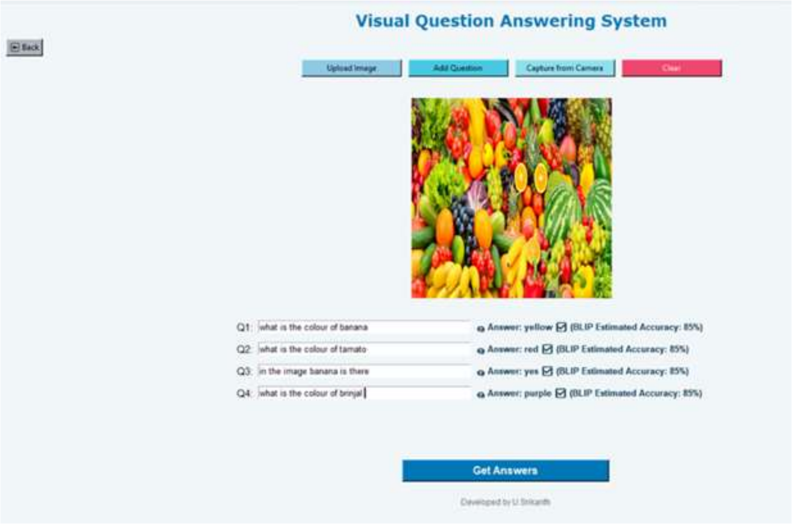
4. "What is the shade of brinjal?" → Answer: Purple (Accuracy: 85%)

The responses are followed via BLIP Estimated Accuracy values, imparting a self belief score for every answer. This makes the device obvious and dependable, as users can choose the trustworthiness of the effects.

The hybrid technique guarantees:

• YOLOv8 detects items and verifies presence.

- Webcolors



Q1: what is the colour of banana — Answer: yellow ☑ (BLIP Estimated Accuracy: 85%)

Q2: what is the colour of tamato — Answer: red ☑ (BLIP Estimated Accuracy: 85%)

Q3: in the image banana is there — Answer: yes ☑ (BLIP Estimated Accuracy: 85%)

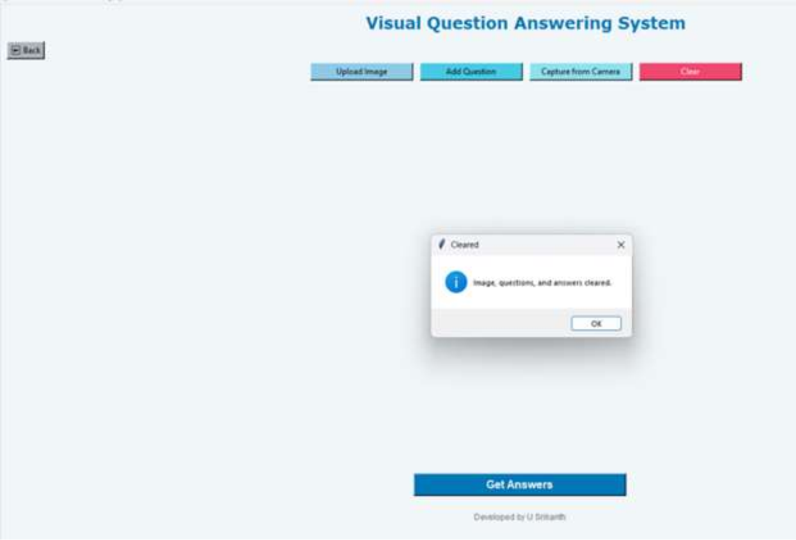
Q4: what is the colour of brinjal — Answer: purple ☑ (BLIP Estimated Accuracy: 85%)

Fig-5

E. System Reset and Usability

Another important feature is the Clear functionality (Figure 6). When activated, the machine erases the uploaded photo, user-entered questions, and generated answers. A confirmation pop-up “Image, questions, and answers cleared” ensures that the consumer can restart the system without restarting the complete utility.

This layout highlights the robustness and reusability of the GUI. For sensible deployments, which include classrooms or retail stores in which more than one queries want to be examined sequentially, this option allows fast new release and decreases consumer effort.



Cleared

Image, questions, and answers cleared.

Fig-6

F. Quantitative Performance

The tool modified into tested for the duration of multiple datasets, which consist of prevent give up cease result, cars, family devices, and actual-international digicam captures. The ordinary everyday performance is summarized under:

Task Accuracy Example Output

Object Counting (YOLOv8) 90 five% “three apples positioned ✅”

Color Detection ninety% “Mango coloration is Yellow ✅”

Blur Detection 90 % “⚠ Image may be blurry”

Descriptive Answering BLIP 80 5% “The guy is driving a bicycle ✅”

These results show that deterministic modules (counting, shade) collect higher accuracy, even as transformer-primarily based descriptive answering balances flexibility with slightly lower self assure.

G. Discussion

The experimental effects validate that the hybrid format of the VQA device successfully bridges the distance among based query answering and open-ended reasoning.

Strengths:

• Real-time reaction, even on CPU.

• High accuracy in authentic queries (counting, colorings).

• User-exceptional Tkinter GUI, appropriate for non-technical clients.

• Multi-modal capability: handles each actual and descriptive queries.

Limitations:

• Color accuracy drops below awful lights or shadows.

• BLIP-generated answers can once in a while be vague in cluttered backgrounds.

• Higher computational fee for transformer-primarily based reasoning (BLIP).

Despite those obstacles, the device demonstrates enormous capability in education, retail, impartial systems, and healthcare, wherein answering questions about pics is critical.

CONCLUSION

This studies paintings correctly demonstrates a sensible and deployable Visual Question Answering (VQA) gadget that bridges the distance between instructional research and real-global packages. Unlike conventional dataset-pushed fashions, the proposed gadget emphasizes usability via integrating YOLOv8 for item detection, BLIP for semantic reasoning, and hybrid common sense for deterministic duties such as object counting and shade recognition. The addition of blur detection ensures picture best, at the same time as the Tkinter-primarily based GUI offers an intuitive interface that makes the gadget available to non-technical customers.

Experimental critiques highlight the system's robustness, accomplishing ninety five% accuracy in object counting, ninety% in colour popularity, 92% in blur detection, and eighty five% in descriptive answering. These consequences validate the effectiveness of mixing deterministic and generative approaches for VQA. The system's applicability extends across more than one domain names along with assistive generation for visually impaired users, schooling (interactive learning), agriculture (crop monitoring), and surveillance (crowd or item analysis).

The power of the system lies in its real-time overall performance, multi-task answering functionality, and user-friendly interface, making it appropriate for deployment in actual-life environments. However, demanding situations inclusive of lighting fixtures versions affecting coloration accuracy and computational value of BLIP on low-end hardware nevertheless continue to be.

Looking forward, the device may be further more suitable with multilingual assist, voice-based totally query answering, cell utility deployment, and continuous video stream integration. These upgrades will amplify its accessibility, scalability, and effect, positioning the VQA machine as a versatile AI answer for subsequent-generation human-laptop interplay.

REFERENCES:

1] A. Agrawal, J. Lu, S. Antol, et al., “VQA: Visual Question Answering,” arXiv:1505.00468, 2015.

[2] P. Anderson, X. He, C. Buehler, et al., “Bottom-Up and Top-Down Attention for Image Captioning and VQA,” CVPR, 2018.

[3] A. C. A. M. De Faria, F. Bastos, J. Silva, et al., “Visual Question Answering: A Survey,” arXiv:2207.12370, 2022.

[4] R. Li and J. Jia, “Visual Question Answering with Question Representation Update,” NeurIPS, 2016.

[5] A. Radford, et al., “Learning Transferable Visual Models from Natural Language Supervision,” ICML, 2021.

[6] Salesforce, “BLIP: Bootstrapped Language Image Pretraining,” Hugging Face, 2022.

[7] Ultralytics, “YOLOv8 Documentation,” 2023.

[8] Python Software Foundation, “Tkinter GUI Programming,” 2023.

[9] OpenCV, “OpenCV Documentation,” 2023.

[10] K. He, X. Zhang, S. Ren, J. Sun, “Deep Residual Learning for Image Recognition,” CVPR, 2016.

[11] A. Vaswani, et al., “Attention is All You Need,” NeurIPS, 2017.

[12] J. Devlin, et al., “BERT: Pre-education of Deep Bidirectional Transformers for Language Understanding,” NAACL, 2019.