

Conversational Image Recognition Chatbot Using Computer Vision and Natural Language Processing

Mr. Premkumar Pawar¹, Prof. Shubham Lotwala²

Research Scholar, Department of Computer Applications, SSBT COET, Jalgaon, Maharashtra, India¹

Associate Professor, Department of Computer Engineering, SSBT COET, Jalgaon, Maharashtra, India²

Abstract: Conversational image recognition combines image understanding with natural language interaction so that a user can upload an image and ask questions about its content. This research proposes a practical framework for a conversational image recognition chatbot using Computer Vision and Natural Language Processing (NLP). The proposed approach accepts an image and a natural-language query, processes visual and textual information, combines the resulting representations, and generates a conversational response. Visual recognition can use Convolutional Neural Networks or Vision Transformers for extracting image features, while language processing handles the user question and conversation context. Vision-language models such as CLIP, BLIP, BLIP-2, and LLaVA provide useful directions for connecting visual and textual information. The chatbot can support image description, object-related questions, simple visual question answering, and follow-up questions. Evaluation will consider recognition quality, answer relevance, response consistency, error cases, and processing time. The study is designed as a practical research framework in which the final performance values will be obtained after implementation and testing.

Keywords: Conversational AI, Image Recognition, Computer Vision, Natural Language Processing (NLP), Visual Question Answering, Image Captioning, Vision-Language Models, Multimodal Chatbot.

I. INTRODUCTION

Conversational image recognition is becoming an important application of Artificial Intelligence because users often need to understand an image through a natural conversation rather than through a single classification label. Traditional image recognition systems usually return predefined object or class labels. A conversational system can instead allow a user to upload an image, ask a question, receive an answer, and continue with a related question.

An image may contain several objects, people, text regions, colors, and spatial relationships. A useful chatbot therefore needs to connect visual information with the meaning of the user's question. For example, after uploading an image, a user may ask what objects are present, what color an object has, or what activity is shown. The system must process both the image and the language used in the query.

Computer Vision provides methods for image feature extraction and recognition, while NLP provides methods for understanding questions and generating responses. Vision-language models provide a common direction for combining these two information sources. A conversational interface can use these capabilities to present the result in a simple question-and-answer form.

The motivation of this research is to study a practical conversational framework that combines image recognition, question understanding, multimodal processing, and response generation. The proposed system is designed as an information tool: it provides responses based on the uploaded image and the user's question, while uncertain or ambiguous cases can be identified for further review.

II. LITERATURE REVIEW

The reviewed studies show a progression from image recognition and image captioning toward unified vision-language and conversational systems. Image encoders provide visual representations, while VQA and captioning methods connect those representations with natural language. Recent vision-language models further combine visual and language information for instruction-based interaction. The proposed research brings these directions together in a simple conversational workflow consisting of image input, visual analysis, question processing, multimodal understanding,

response generation, and conversational follow-up. This combination is proposed as a research framework; its effectiveness must be established experimentally on the selected images and questions.

Research Gap and Proposed Direction

The literature indicates that visual recognition alone is not sufficient for conversational image understanding. Image encoders such as ResNet and ViT provide visual representations [4], [5], while attention-based captioning and VQA demonstrate the connection between images and natural-language responses [1]–[3]. CLIP, BLIP, and BLIP-2 provide stronger image-text alignment and multimodal generation capabilities [6]–[8]. LLaVA and Context-VQA further demonstrate the importance of visual instruction and conversational context [9], [10]. These studies support a framework in which image features, question meaning, and conversation context are combined before generating the final response.

Synthesis of the Literature

10. Naik et al. (2023) – Context-Aware Visual Question Answering [10]

Naik et al. proposed Context-VQA for context-aware and purposeful visual question answering. Their work shows that the surrounding context of a question can influence the interpretation of a visual query. This is relevant to a conversational chatbot because follow-up questions may depend on information from earlier turns.

9. Liu et al. (2023) – Visual Instruction Tuning [9]

Liu et al. introduced LLaVA, a vision-language conversational approach that connects visual features with a large language model through visual instruction tuning. The work demonstrates a practical direction for answering natural-language questions about images and supports the conversational aspect of the proposed framework.

8. Li et al. (2023) – BLIP-2 for Vision-Language Understanding and Generation [8]

Li et al. presented BLIP-2, which connects frozen image encoders with large language models through a lightweight querying transformer. The approach shows how visual information can be transferred to a language model for generation and understanding.

7. Li et al. (2022) – BLIP for Unified Vision-Language Understanding and Generation [7]

Li et al. proposed BLIP as a unified vision-language pre-training framework for understanding and generation tasks. Its combination of image-text learning and generation is relevant to image captioning and visual question answering in a chatbot.

6. Radford et al. (2021) – CLIP for Image-Text Alignment [6]

Radford et al. introduced CLIP, which learns transferable visual representations from natural language supervision. The study is important for the proposed system because image-text alignment can help connect a user's words with visual content.

5. Dosovitskiy et al. (2021) – Vision Transformer [5]

Dosovitskiy et al. introduced Vision Transformer (ViT), showing that transformer architectures can be applied directly to image patches for image recognition. ViT provides an alternative visual encoder for extracting image representations.

4. He et al. (2016) – Deep Residual Learning for Image Recognition [4]

He et al. introduced residual networks that enable very deep neural networks to learn effective visual representations. ResNet provides a widely used foundation for image feature extraction and can be used as a visual encoder in a multimodal system.

3. Xu et al. (2015) – Show, Attend and Tell [3]

Xu et al. proposed an attention-based image captioning model that learns to focus on relevant image regions while generating words. Visual attention is useful for connecting parts of an image with the words in a generated response.

2. Vinyals et al. (2015) – Show and Tell [2]

Vinyals et al. demonstrated neural image caption generation by connecting visual features with a language generation model. The work provides an early foundation for converting image content into natural-language descriptions.

1. Antol et al. (2015) – Visual Question Answering [1]

Antol et al. introduced the VQA task, in which a system answers natural-language questions about images. This work directly motivates the question-answering component of the proposed conversational image recognition chatbot.

1. Antol et al. (2015) – Visual Question Answering [1]

Study	Main method	Main contribution	Relevance to chatbot
Antol et al. (2015) [1]	VQA	Image + question → answer	Foundation for visual question answering
Vinyals et al. (2015) [2]	Image captioning	Image → natural-language description	Supports image description
Xu et al. (2015) [3]	Visual attention	Focuses on relevant image regions	Supports visual grounding

He et al. (2016) [4]	ResNet	Deep visual representation	Visual encoder foundation
Dosovitskiy et al. (2021) [5]	Vision Transformer	Transformer image encoding	Alternative visual encoder
Radford et al. (2021) [6]	CLIP	Image-text alignment	Connects visual and language concepts
Li et al. (2022) [7]	BLIP	Unified vision-language learning	Supports understanding and generation
Li et al. (2023) [8]	BLIP-2	Frozen encoder + language model	Efficient multimodal generation
Liu et al. (2023) [9]	LLaVA	Visual instruction tuning	Supports visual conversation
Naik et al. (2023) [10]	Context-VQA	Context-aware VQA	Supports conversational follow-up

III. METHODOLOGY

The research follows an experimental and applied design to develop and evaluate a prototype conversational image recognition chatbot. The methodology consists of five main stages.

1. Image Collection and Preprocessing: Images will be collected from suitable datasets or user-provided samples. Images will be resized and normalized before visual processing. Basic preprocessing will be applied to maintain a consistent input format.

2. Visual Feature Extraction and Image Understanding: A suitable visual encoder such as ResNet or Vision Transformer will be used to extract image features. Object recognition, image description, and visual region information may be obtained according to the selected model.

3. Question and Conversation Processing: The user's question will be tokenized and represented for language understanding. The system will retain the relevant previous question and answer so that follow-up questions can be interpreted in the correct conversational context.

4. Multimodal Fusion and Response Generation: Visual features and language features will be combined using a vision-language model or multimodal fusion mechanism. The resulting representation will be passed to a response-generation component to produce an answer related to the image and the user's question.

5. Evaluation and Tools: Answer relevance, recognition quality, response consistency, error cases, and average processing time will be measured. Python, OpenCV, NumPy, PyTorch or TensorFlow, Transformers, and a simple Flask/FastAPI interface can be used for the prototype.

Fig. 1. Overall architecture of the proposed conversational image recognition chatbot.

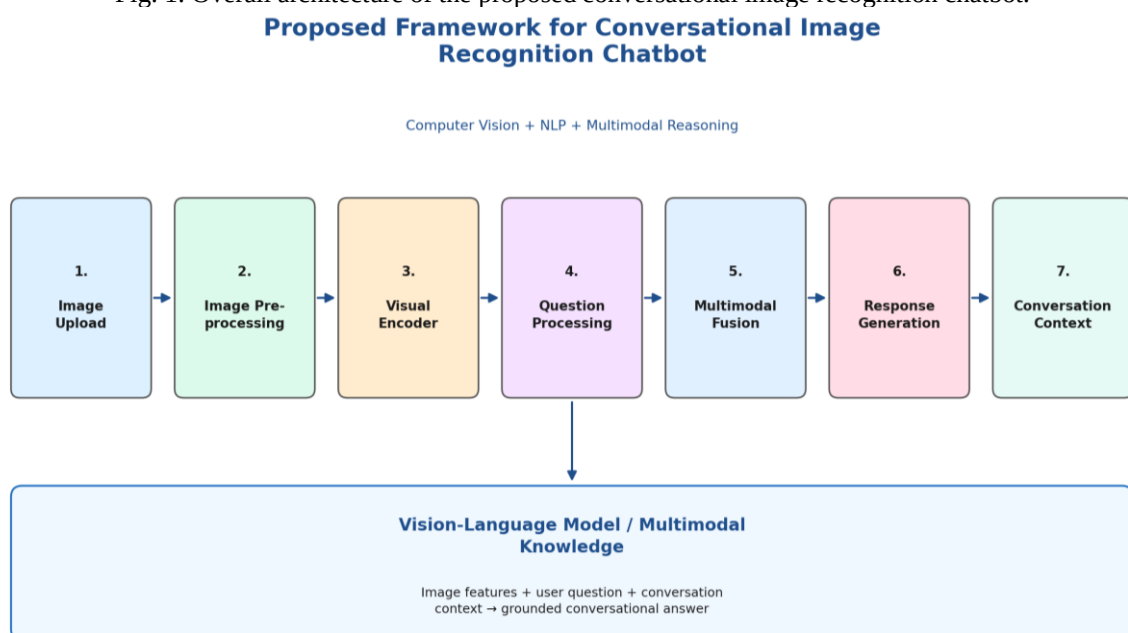
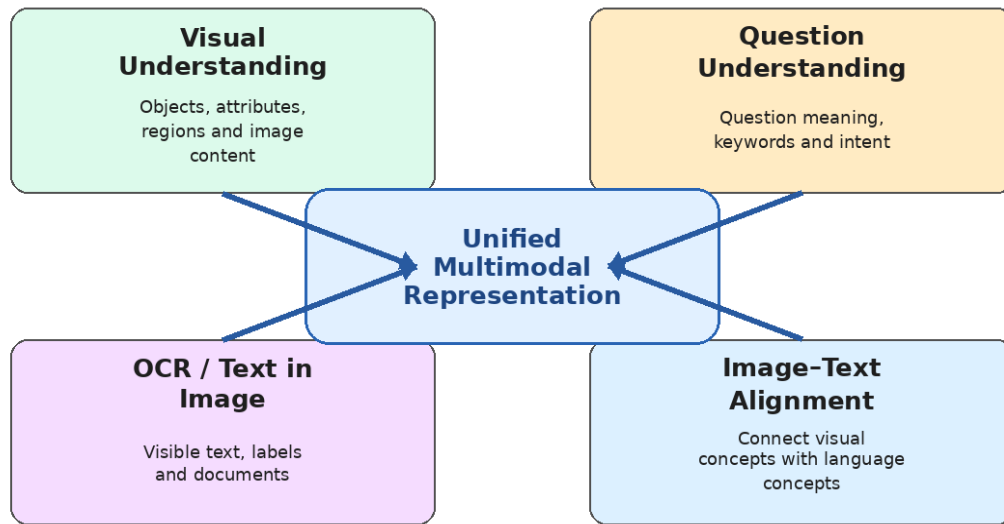


Fig. 2. Multimodal image understanding framework.

Multimodal Image Understanding Framework



Conversational Image Question-Answer Process

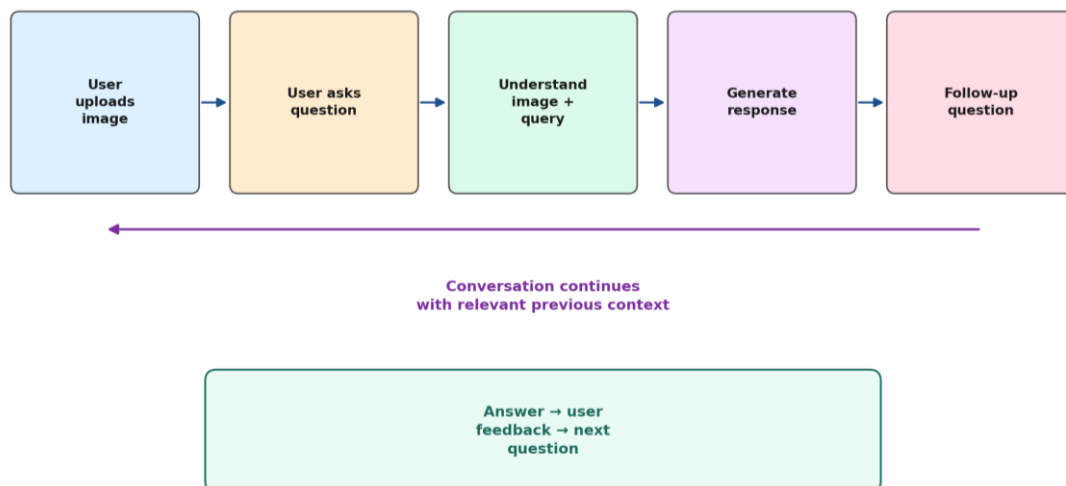


Fig. 3. Conversational image question-answer and follow-up process.

The proposed framework is illustrated through five complementary diagrams. The architecture combines image preprocessing, visual feature extraction, question processing, multimodal fusion, response generation, and conversational context. The evaluation framework further considers recognition quality, answer relevance, dialogue consistency, error cases, and processing time. These figures represent the proposed system design and evaluation plan; they do not claim experimental results before testing.

IV. RESULTS

The research is designed to produce results that can be interpreted directly from the chatbot interaction. Since the final experiments depend on the selected image dataset and question set, numerical performance values are not assumed in

advance. The result analysis will compare the proposed multimodal chatbot with simpler image-only or single-turn baselines.

Evaluation Aspect	Measure / Output	Purpose
Recognition	Accuracy, Precision, Recall, F1-score	Measures how correctly the system identifies or answers visual content.
Question Answering	Answer accuracy and relevance	Evaluates whether the response addresses the user's visual question.
Captioning	BLEU, ROUGE or CIDEr	Measures similarity between generated and reference image descriptions.
Dialogue Quality	Context consistency and follow-up accuracy	Checks whether the chatbot maintains relevant information across turns.
Error Analysis	Incorrect answers and ambiguous cases	Identifies cases where visual or conversational understanding fails.
Efficiency	Average response time	Checks whether the chatbot can provide responses efficiently.

The primary result will be a conversational response for every submitted image and user question. The system should provide an answer that is relevant to the visual content and the question. Follow-up questions will be examined to determine whether the chatbot maintains the required context. Model performance will be summarized using the selected recognition, language, and dialogue measures. Processing time will also be reported because an interactive chatbot requires practical response time. Incorrect and ambiguous cases will be examined separately because they can arise from visual recognition, language interpretation, or multimodal reasoning. The figures summarize the proposed workflow and evaluation plan; they are not claims of experimental accuracy before testing.

Workflow for Conversational Image Recognition

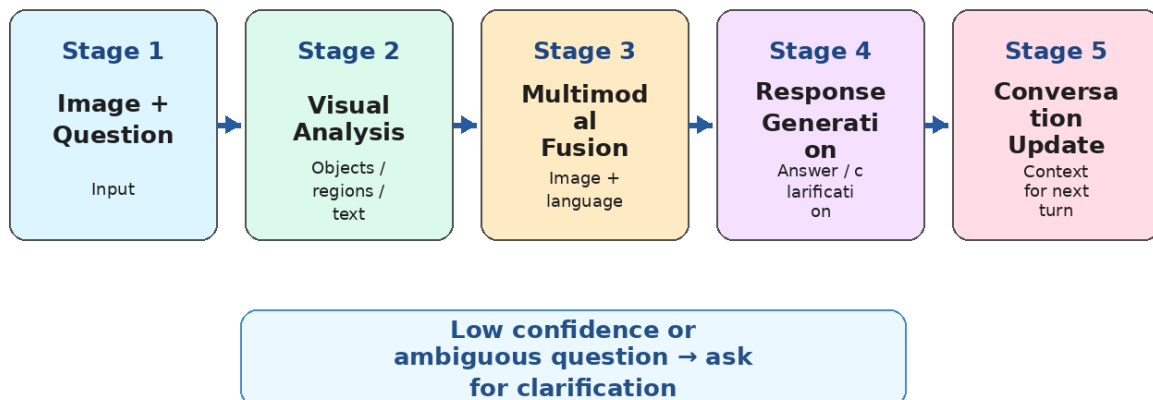


Fig. 4. Workflow for conversational image recognition and response generation.

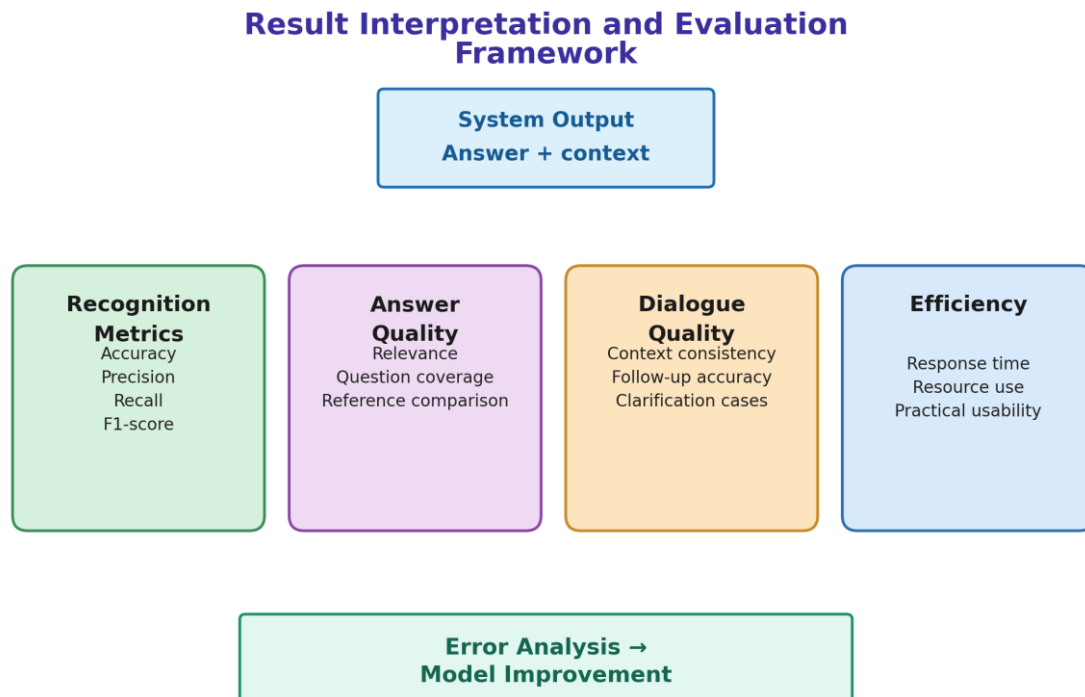


Fig. 5. Evaluation framework for result analysis and model improvement.

V. DISCUSSION

The proposed framework combines methods that operate at different levels of image and language understanding. Visual encoders provide image representations, while NLP methods process the user's question and conversational context. Multimodal models can connect the two information sources and generate a response that is related to the image.

The most important practical issue is the consistency of the generated answer with the actual image and the user's question. Incorrect visual recognition may lead to an incorrect response, while an ambiguous question may require clarification. Recognition measures, answer relevance, dialogue consistency, and error-case inspection are therefore necessary.

The quality and diversity of the image and question dataset will directly affect reliability. Images containing small objects, complex scenes, poor lighting, or text may require additional visual processing. Follow-up questions also require the system to maintain useful conversational context. Therefore, the system is best viewed as an interactive image-understanding tool that provides information through natural language conversation.

VI. CONCLUSION

This research proposes a practical Computer Vision and NLP framework for a conversational image recognition chatbot. The approach begins with image preprocessing and visual feature extraction, processes the user's question, combines visual and language information, and generates a conversational response.

The literature indicates that image recognition, visual question answering, image captioning, and vision-language models provide complementary capabilities for multimodal interaction. The staged methodology therefore provides a practical structure for implementing a chatbot that can answer questions about uploaded images and continue the conversation.

The final system will present image-related answers and follow-up responses so that users can interact with visual information through natural language. Evaluation using recognition metrics, answer relevance, dialogue consistency, error analysis, and processing time will determine the suitability of the proposed approach for the selected dataset.

REFERENCES

- [1]. S. Antol et al., “VQA: Visual Question Answering,” in Proc. IEEE International Conference on Computer Vision (ICCV), pp. 2425–2433, 2015.
- [2]. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and Tell: A Neural Image Caption Generator,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
- [3]. K. Xu et al., “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” in Proc. International Conference on Machine Learning (ICML), 2015.
- [4]. K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.
- [5]. A. Dosovitskiy et al., “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in Proc. International Conference on Learning Representations (ICLR), 2021.
- [6]. A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision,” in Proc. International Conference on Machine Learning (ICML), vol. 139, pp. 8748–8763, 2021.
- [7]. J. Li et al., “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation,” in Proc. International Conference on Machine Learning (ICML), vol. 162, pp. 12888–12900, 2022.
- [8]. J. Li et al., “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models,” in Proc. International Conference on Machine Learning (ICML), vol. 202, pp. 19730–19742, 2023.
- [9]. H. Liu et al., “Visual Instruction Tuning,” arXiv:2304.08485, 2023.