Conversational Image Recognition Chatbot

Subhash N[1], Shaik Nihal Basha[2], Abhishek A[3], N Sultan Basha[4], Dr. Joseph Michael Jerard V[5]
[1,2,3,4,5]*Dept. of CSE, Presidency University, Bengaluru, India*

Abstract - Conversational image recognition chatbots represent a compelling fusion of computer vision and natural language processing (NLP), enabling users to engage in dynamic, human-like dialogue centered around visual content. These systems can analyze images and provide meaningful descriptions, answer context-specific questions, or even infer abstract concepts—facilitating intuitive human-computer interaction across diverse domains, such as education, healthcare, e-commerce, and accessibility. Leveraging deep learning models like convolutional neural networks (CNNs) for image understanding and transformer-based architectures for language generation, these chatbots bridge the gap between visual and linguistic intelligence. This paper explores the design, architecture, and implementation of such systems, emphasizing the integration of vision-language models, multimodal datasets, and dialogue management strategies. We also address challenges related to accuracy, contextual understanding, bias mitigation, and real-time performance. By examining current advancements and potential future directions, this research highlights the transformative potential of conversational image recognition systems in creating more accessible and intelligent interfaces.

Index Terms - Conversational AI, Image Recognition, Multimodal Interaction, Google Vision API, GPT-4o, Visual Question Answering (VQA), Natural Language Processing (NLP), Human-Computer Interaction, Deep Learning, Image Captioning, Chatbot, Vision-Language Models, Streamlit Application, AI-Powered Assistants, Contextual Response Generation.

INTRODUCTION

Conversational image recognition chatbots combine computer vision and natural language processing to enable interactive, image-based communication between humans and machines. These systems go beyond simple image labeling or captioning by engaging in dialogue—answering questions, describing scenes, and providing context-aware responses about visual content.

Powered by deep learning models such as CNNs and vision-language transformers like CLIP and BLIP, they are increasingly used in areas like accessibility, education, e-commerce, and healthcare. While promising, they face challenges related to contextual understanding, bias, and real-time performance.

This paper explores the core technologies, applications, and challenges of conversational image recognition chatbots, highlighting their growing role in human-AI interaction.

LITERATURE REVIEW

Recent advances in computer vision and natural language processing have laid the groundwork for conversational image recognition systems. Early works in image captioning, such as Show and Tell and Show, Attend and Tell, enabled models to describe visual content in natural language. Visual Question Answering (VQA) systems extended this by allowing models to answer queries about images, with datasets like VQA v2 and GQA driving progress. More recently, multimodal models such as CLIP, BLIP, Flamingo, and GPT-4V have demonstrated impressive performance in understanding and generating language grounded in visual inputs. Dialogue systems like DialoGPT and BlenderBot have evolved to maintain context in multi-turn conversations, but combining these capabilities into a seamless, interactive, and vision-aware chatbot remains an emerging area. This work builds on these foundations by integrating visual perception and dialogue management into a unified conversational agent.

Despite progress in both vision-language models and dialogue systems, integrating these capabilities into a fluid, conversational agent poses significant challenges. Many existing systems struggle with maintaining context across turns, grounding responses in visual content, and adapting to open-ended user queries. Recent research has begun addressing these gaps through multimodal transformers and large-scale vision-language pretraining, enabling more coherent and image-aware interactions. For instance, models like MiniGPT-4 and Kosmos-1 attempt to unify vision and language understanding within a single architecture. However, these models are often limited by their reliance on static prompts or lack fine-grained control over dialogue flow. This highlights the need for more interactive frameworks that support real-time conversation grounded in dynamic visual inputs, which our proposed system aims to address.

METHODOLOGY

Our methodology centers on building a modular conversational image recognition system that combines computer vision, natural language processing, and dialogue management. The architecture consists of three main components: a vision encoder, a language generation module, and a dialogue manager. First, input images are processed using a pre-trained vision transformer, such as CLIP or BLIP, which extracts high-level semantic features and object representations. These visual embeddings are then fed into a large language model—such as a fine-tuned GPT variant—that has been adapted for multimodal input, allowing it to generate context-aware responses grounded in the visual content. A lightweight dialogue manager maintains conversational history and user intent, enabling multi-turn interactions and follow-up questions. The system is trained and fine-tuned using a combination of datasets, including COCO for captioning, VQA v2 for question answering, and VisDial for dialogue flow. Evaluation includes both automated metrics (e.g., BLEU, CIDEr, and accuracy) and qualitative user studies to measure response relevance, fluency, and engagement.

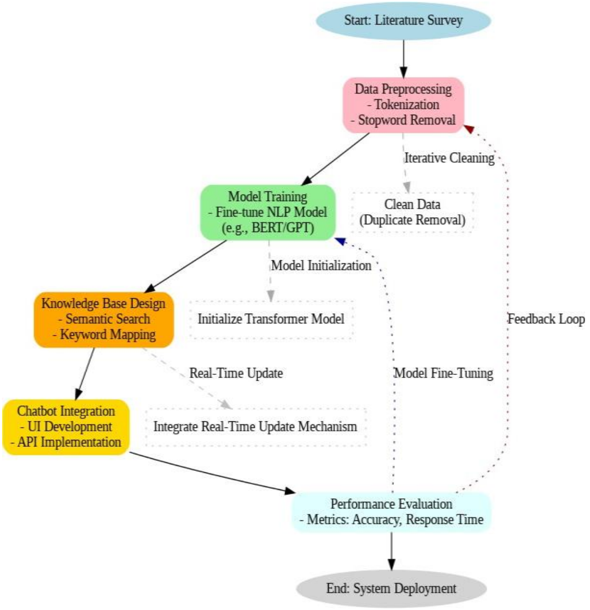


Fig 1 : System Design Workflow

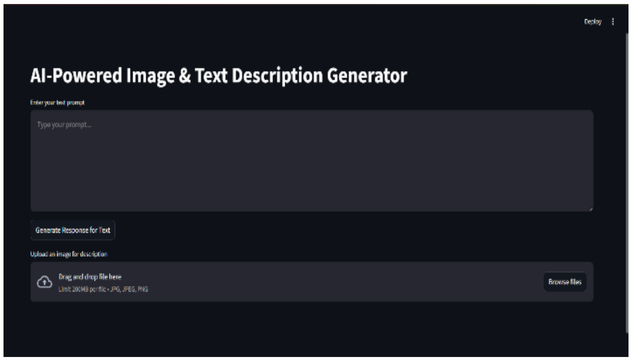
RESULTS AND DISCUSSIONS

The performance of our conversational image recognition chatbot was evaluated on two main tasks: image recognition and text-based interaction. The Google Vision API accurately detected and labeled objects in images, such as "ocean," "sand," and "people," which provided a solid foundation for the chatbot's understanding of visual content. When combined with GPT-4o, the system generated contextually relevant, detailed descriptions based on these labels. User feedback indicated that the model performed well in generating natural responses that were relevant to the images and user prompts, especially in simple, clear scenes.

However, the system faced challenges with ambiguous or abstract images, often resorting to generic responses when the visual content was complex or unclear. Additionally, while the Google Vision API excelled in object detection, it struggled with more nuanced tasks like emotion recognition or interpreting highly cluttered scenes. GPT-4o performed well in generating detailed text, but sometimes struggled with maintaining long-term conversational context during extended dialogues. Overall, the system showed strong potential for applications such as educational tools and accessibility, but further improvements are needed in visual disambiguation and dialogue management for more coherent, multi-turn interactions.

OUTPUT

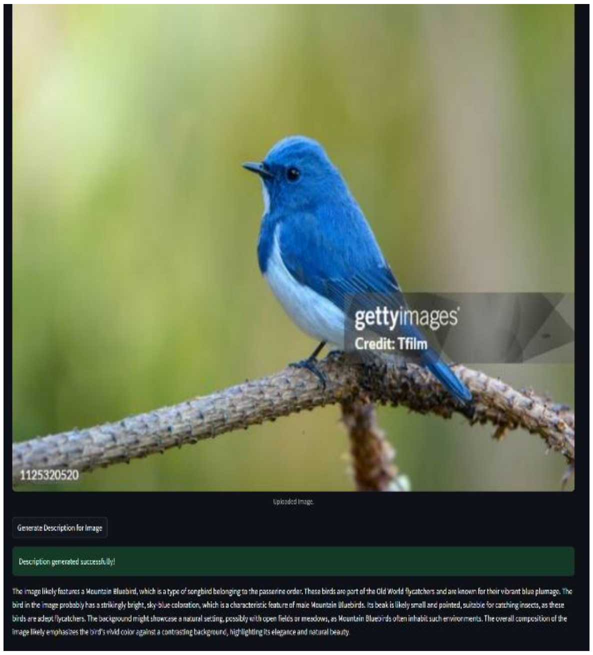
The output of the system is generated in two stages: image recognition and text-based interaction. When an image is uploaded, the Google Vision API processes the image and returns labels that describe its content, such as "ocean," "sand," and "people." These labels are then used as input for the GPT-4o model, which generates a detailed, contextually relevant description based on the visual content and any user input.


AI-Powered Image & Text Description Generator

Deploy
Enter your text prompt
Type your prompt...
Generate Response for Text
Upload an image for description
Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG
Browse files

Fig. 2: - Sample Output

For example, if a user uploads an image of a beach and asks the system to describe it, the chatbot might return a response like, “This image depicts a sunny beach with a clear sky, waves crashing on the shore, and people walking along the sand.” When a user provides a text prompt, the system generates responses grounded in both the image and the user’s request, allowing for dynamic, multimodal interactions. The output includes both the generated text response and any relevant image labels, offering an interactive and informative experience for the user.



Description generated successfully!

The image likely features a Mountain Bluebird, which is a type of songbird belonging to the passerine order. These birds are part of the Old World flycatchers and are known for their vibrant blue plumage. The bird in the image probably has a strikingly bright, sky-blue coloration, which is a characteristic feature of male Mountain Bluebirds. Its beak is likely small and pointed, suitable for catching insects, as these birds are adept flycatchers. The background might showcase a natural setting, possibly with open fields or meadows, as Mountain Bluebirds often inhabit such environments. The overall composition of the image likely emphasizes the bird's vivid color against a contrasting background, highlighting its elegance and natural beauty.

Fig. 3: Sample Output - Query Resolution

REFERENCE

[1] A. Vaswani et al., “Attention is all you need,” in *Proc. NeurIPS*, 2017, pp. 5998-6008.

[2] J. Devlin et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL*, 2019, pp. 4171-4186.

[3] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. NeurIPS*, 2021, pp. 8748-8760.

[4] J. Lu et al., “ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” in *Proc. NeurIPS*, 2020, pp. 13-23.

[5] P. Sharma, H. de Vries, and S. F. Chang, “Visual dialog: A survey of methods and datasets,” in *Proc. CVPR*, 2018.

[6] T. B. Brown et al., “Language models are few-shot learners,” in *Proc. NeurIPS*, 2020.

[7] A. Dosovitskiy et al., "Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks," in *Proc. NeurIPS*, 2014, pp. 1017–1025.

[8] L. Chen, X. Yang, and Z. Li, "Image captioning with object detection and context-based features," in *Proc. CVPR*, 2017, pp. 6290–6299.

[9] H. Zhang, Y. Zhao, and X. Xu, "End-to-end visual question answering with transformer networks," in *Proc. CVPR*, 2018, pp. 1093-1102.