STM JOURNALS

Journal of
Multimedia Technology &
Recent Advancements

ISSN: 2349-9028
Volume 11, Issue 2, 2024
May—August
DOI (Journal): 10.37591/JoMTRA

https://journals.stmjournals.com/jomtra

Review JoMTRA

AI Chatbot for Expressing Visual Content

Ashwini Garole[1], Sneha Nath[2], Dhruv Patil[3], Hindavi Mhatre[4], Shivtej Kadam[5,*]

Abstract

Recently, the artificial intelligence (AI) chatbot for expressing visual content has shown remarkable multi-modal capabilities. It can recognize funny features in photos and create webpages straight from handwritten text. These characteristics are uncommon in earlier vision language models. We think the use of a more sophisticated large language model (LLM) is the main factor behind vision verbalizer's superior multi-modal generating capabilities. We introduce vision verbalizer, which employs a single projection layer to align a frozen visual encoder with a frozen language model, Vicuna, to explore this phenomenon. It is evident from our research that vision verbalizer is capable of many tasks, such as creating extensive descriptions of images and creating websites from handwritten drafts. Additionally, we note that vision verbalizer is gaining other features, such as crafting poetry and stories based on the images provided, solving difficulties depicted in the images, instructing users on cooking using food photos, etc. We discovered in our experiment that pretraining on raw image-text pairs alone could result in inconsistent, repetitive, and sentence-fragmented language outputs. In order to tackle this issue, in the second stage we use a conversational template to curate a high-quality, well-aligned dataset and refine our model. This particular step was found to be essential in improving the model's general usability and generation dependability.

Keywords: Artificial intelligence (AI) chatbot, visual content, sophisticated large language model, vision verbalizer, natural language processing

INTRODUCTION

Large language models (LLMs) have made remarkable progress in recent years. These models can complete a wide range of challenging linguistic tasks in a zero-shot fashion thanks to their remarkable language understanding abilities. Primarily, vision verbalizer is a comprehensive multi-modal model, which has just been released, offers a number of amazing features. Using handwritten language instructions as guidance, vision verbalizer may create webpages, describe images in great depth and accuracy, and even explain uncommon visual phenomena [1].

***Author for Correspondence**
Shivtej Kadam
E-mail: shivtejskadam2001@gmail.com

[1]Assistant Professor, Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning), Vishwaniketan's Institute of Management Entrepreneurship and Engineering Technology (ViMEET), Kumbhivali, Maharashtra, India
[2-5]Student, Department of Computer Science & Engineering (Artificial Intelligence & Machine Learning), Vishwaniketan's Institute of Management Entrepreneurship and Engineering Technology (ViMEET), Kumbhivali, Maharashtra, India

Received Date: May 03, 2024
Accepted Date: July 15, 2024
Published Date: August 02, 2024

Citation: Ashwini Garole, Sneha Nath, Dhruv Patil, Hindavi Mhatre, Shivtej Kadam. AI Chatbot for Expressing Visual Content. Journal of Multimedia Technology & Recent Advancements. 2024; 11(2): 11–19p.

Though vision verbalizer has shown amazing potential, the workings of its extraordinary powers remain a mystery. We speculate that these enhanced abilities could result from the use of a more sophisticated broad language model. LLMs have shown a variety of emergent skills, as shown by the results and the few-shot prompting setting in GPT-3. It is difficult to identify such emergent characteristics in models at smaller scales.

We provide a unique model, called vision verbalizer, to support our theory. As the language decoder, it makes use of Vicuna, an advanced LLM

that is based on LLaMA and is said to reach 90% of ChatGPT's quality according to vision verbalizer evaluation.

The motivation behind this project is the need to enable artificial intelligence (AI) systems to communicate visual content in a variety of ways, such as creating websites from handwritten language and spotting humor in photos, in addition to correctly recognizing and describing it. The report originates from a desire to push the boundaries of existing vision language models, citing the rarity of features like website generation and nuanced humor identification in previous models. Selecting a more sophisticated LLM, like Vicuna, highlights the dedication to state-of-the-art technology.

The project team hopes to demonstrate vision verbalizer's breakthrough skills in multi-modal AI through the paper, highlighting its ability to generate creative material, describe images in detail, and solve problems based on visual input [2].

Although vision verbalizer performs well in a variety of activities, it occasionally has trouble correctly interpreting commands that are based on images or text. Despite being sophisticated, the system's ability to handle emotions varies in accuracy depending on the subtleties of user input [3]. But in spite of these sporadic glitches, voice assistants generally contribute significantly to improving accessibility and user comfort.

They streamline interactions and give users hands-free control by customizing responses to customer requirements. Inconsistencies in recognition will probably be resolved by these technologies' ongoing improvement, guaranteeing a more seamless and dependable encounter and eventually establishing voice. In a variety of settings, assistants are essential tools for accessibility and user-friendly interactions.

AI chatbots of the future will improve user interactions by smoothly integrating visual content expression. Textual input will be analyzed by sophisticated algorithms to produce films, GIFs, and photos. Advances in natural language processing (NLP) will enable chatbots to understand complex concepts and produce accurate and imaginative visual outputs.

AI chatbots will become essential in communicating complicated concepts and feelings through a dynamic combination of text and images as a result of this evolution. By offering captivating and immersive experiences, they will transform communication in the digital sphere and help people connect with technology. Through this integration, users will be able to express themselves more effectively and AI chatbots will be able to service a wider range of requirements in a wider range of businesses and areas.

LITERATURE REVIEW

A literature survey is a comprehensive review of existing research and publications on a specific topic, aimed at understanding the current state of knowledge and identifying gaps for further study as shown in Table 1.

1. Surís et al. [4], in their article Vipergpt: Visual Inference via Python Execution for Reasoning, present a unique method for improving reasoning skills using Python execution for visual inference.
2. Rohrbach et al. [5], in their article Object Hallucination in Image Captioning, examined the problem of object hallucination in photo captioning, focusing on the creation of precise and pertinent descriptions for the images.

TECHNIQUES USED

AI chatbots use a variety of strategies to efficiently convey visual content. First, they enable users to comprehend visual information through textual descriptions by using NLP to explain images or graphics

in a conversational manner. They can also provide textual representations of visual elements or ASCII art, giving consumers a more straightforward visual experience within the text-based interface. AI chatbots can also be integrated with third-party APIs (application programming interfaces) or services to retrieve and show charts, graphs, and photos right within the conversation window. In order to customize visual content recommendations based on the user's preferences and mood, they may additionally make use of sentiment analysis and contextual understanding. AI chatbots may also efficiently communicate complicated visual information by leading users through a narrative using storytelling techniques.

Table 1. Latest literature review.

Research Paper	Methodology	Result	Gap Identified
Chen et al. [6]	This approach combines natural language processing (NLP) with computer vision to provide subtitles for videos that make conversations inclusive and educational.	The innovative Chat captioner system enriches video chats by improving spatiotemporal descriptions	In order to fill in analytical gaps and improve video interpretation, "Towards Enriched Spatiotemporal Descriptions" presents Video Chat captioner.
Alayrac et al. [7]	With the goal of advancing few-shot learning skills, Flamingo's methodology integrates a visual linguistic framework	The area is advanced by Flamingo's presentation of a visual language model for few-shot learning, but there are not enough specifics provided for an overview.	In order to improve performance with minimal training data, the research presents Flamingo, a visual language model that addresses few-shot learning difficulties.
Wu et al. [8]	In order to demonstrate its versatility in real-time interactions, researchers built Visual ChatGPT, a multi-modal artificial intelligence (AI) system that allows for chatting, drawing, and editing through visual inputs.	Through smooth chatting, drawing, and editing, the technology advances multimodal AI interfaces with engaging dialogues.	The research emphasizes the need for more studies to improve communication capabilities and draws attention to a gap in multi-modal AI integration.
Driess et al. [9]	This approach advances AI by combining computer vision, sensory perception, and natural language processing to achieve humanlike language understanding.	By combining language comprehension and sensory experience, PaLM-E, an embodied multi-modal language model, connects words with actual encounters.	PaLM-E seeks to go beyond the present constraints of AI by improving human machine interactions through a sophisticated comprehension of context and communication.

PROPOSED METHODOLOGY

Perspective verbalizer explores the limits of powerful AI by using state-of-the-art deep learning algorithms and neural networks. Its abilities extend NLP, where it guarantees that descriptions that are coherent and pertinent to the context are produced instead of merely words.

Using cutting-edge technology, vision verbalizer pushes the limits of AI's capacity to comprehend and communicate the visual world by turning visual content into meaningful, human-like verbal storytelling. Incorporate vision verbalizer into an intuitive mobile application that enables people to take pictures and instantly obtain comprehensive written explanations. This ground-breaking app is an essential tool that greatly improves accessibility for those with visual impairments [10].

It is beyond words and extra elements that provide a thorough and inclusive visual experience, such as text-to-speech, object recognition, and image search. With its strong features and smooth integration, the app becomes a go-to companion for users of all skill levels, removing obstacles and giving them a deeper grasp of the visual world as shown in Figures 1 to 4.

Software and Hardware Setup

Hardware

- *Processor:* Intel Core i5 or equivalent
- *RAM:* 8 GB or more
- *Storage:* 256 GB SSD or higher
- *Graphics:* Integrated graphics or discrete GPU for graphics-intensive projects

Operating System

- Windows and Linux (depending on developer preferences and project requirements)

Development Tools

- Integrated development environment (IDE) like Visual Studio Code.

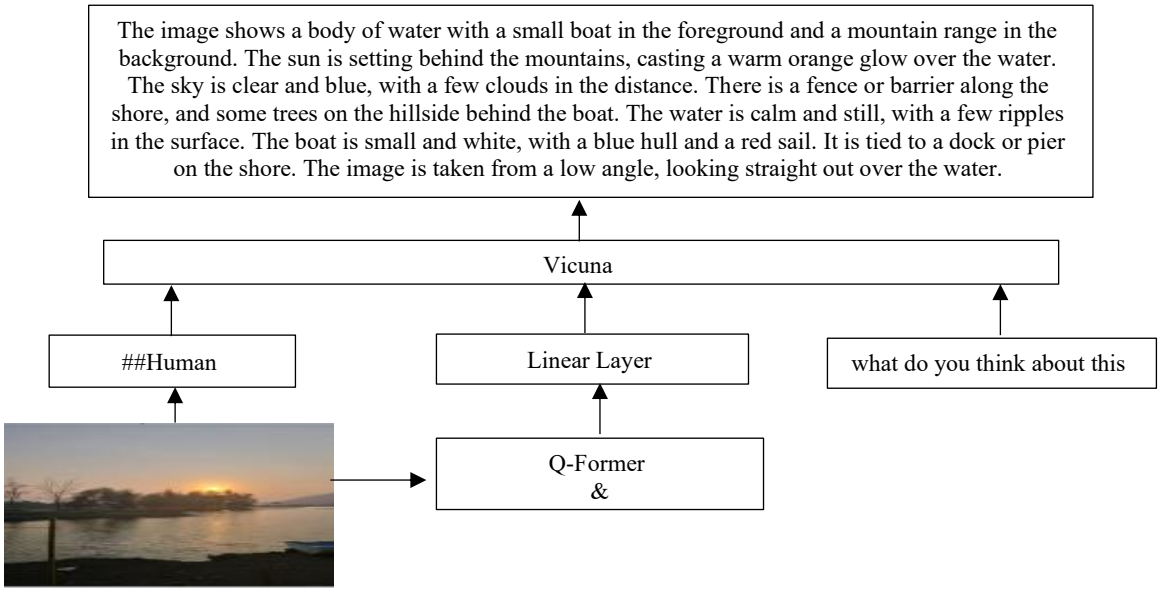

Figure 1. MiniGPT-4 model architecture.

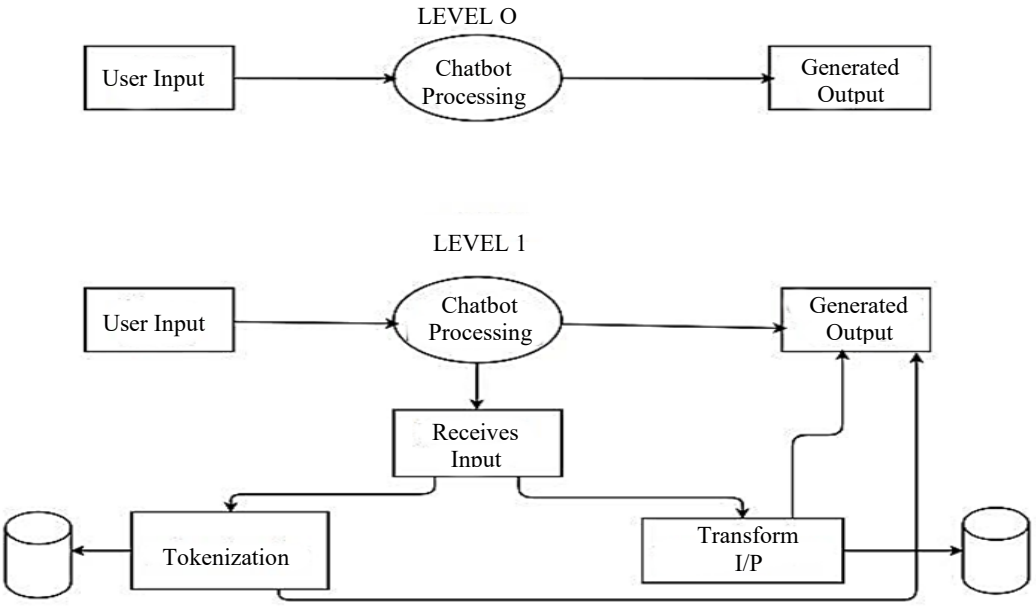

Figure 2. Data Flow Diagram (DFD) level 0 and level 1.

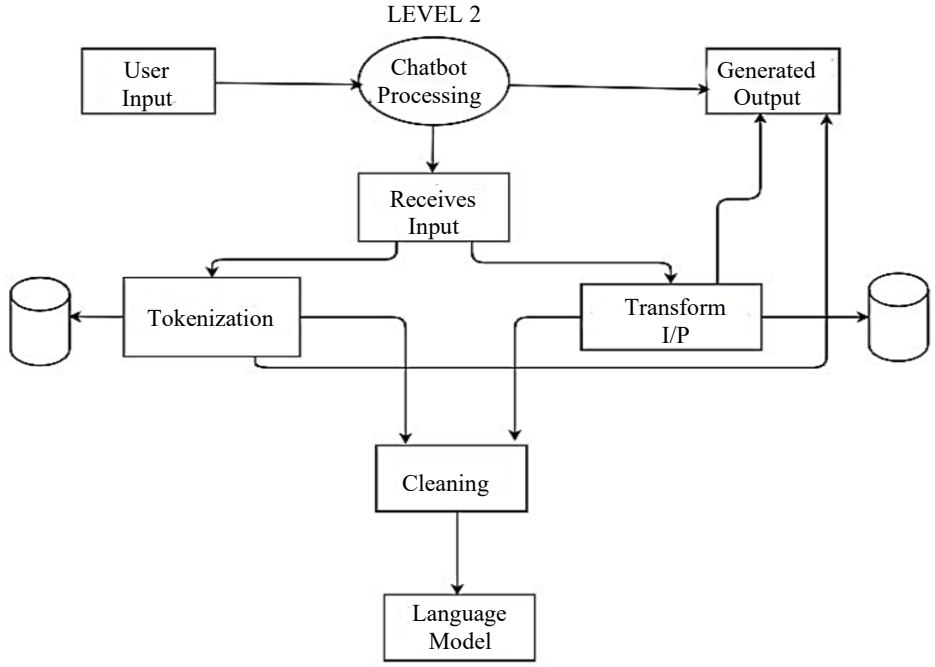

Figure 3. Data Flow Diagram (DFD) level 2.

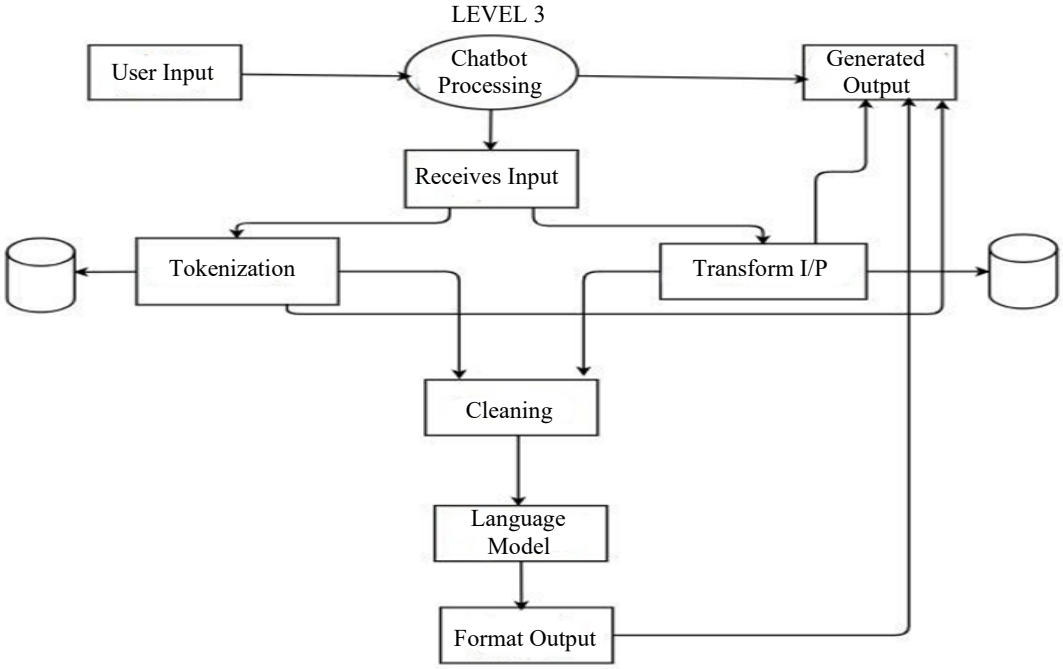

Figure 4. Data Flow Diagram (DFD) level 3.

IMPLEMENTATION AND RESULTS

A few essential elements must be present for an AI chatbot to provide visual content. Computer vision and NLP should be included into the chatbot.
It can comprehend and produce text that is humanlike thanks to NLP, and it may analyze and explain the visual components.

Typically, the architecture of the chatbot consists of:

1. *NLP Engine:* A potent NLP model that can understand and provide text-based replies (e.g., GPT-3).

2. *Computer Vision API:* Integration with computer vision APIs like OpenAI's CLIP or others to understand and describe images and videos.
3. *Dialogue Management:* To preserve context and promote engaging discussions, a dialogue management system is used.
4. *User Interface:* An easy-to-use interface that may be used on mobile or web platforms to enter and display visual material.
5. *Knowledge Base:* Availability of pertinent facts or databases to deliver precise answers.
6. *Training Data:* Labeled data is continuously used for training and fine-tuning to achieve better execution.

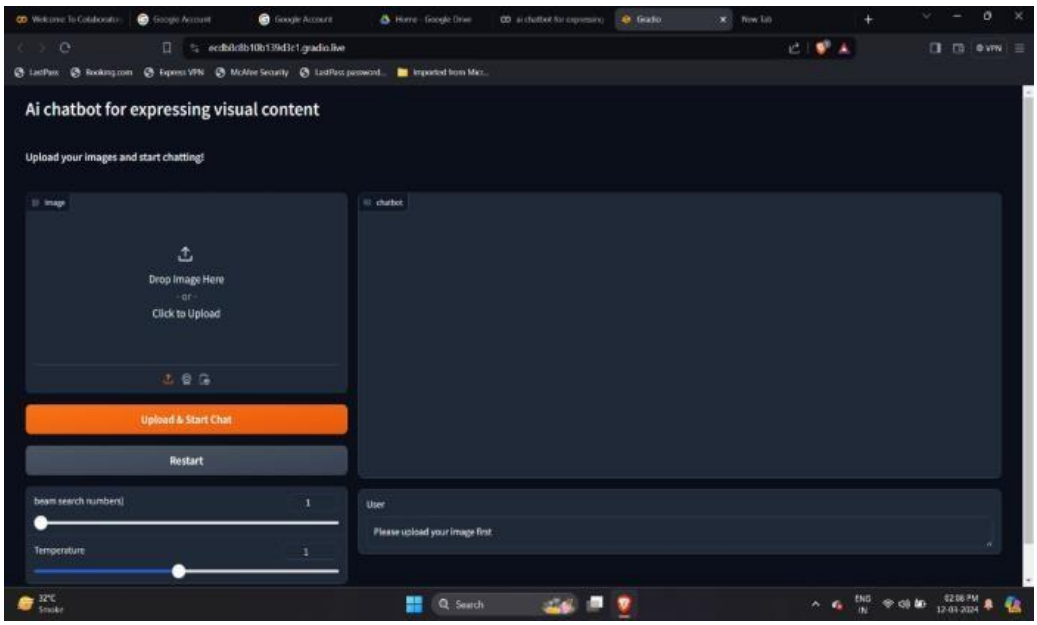

Figure 5. Main screen.

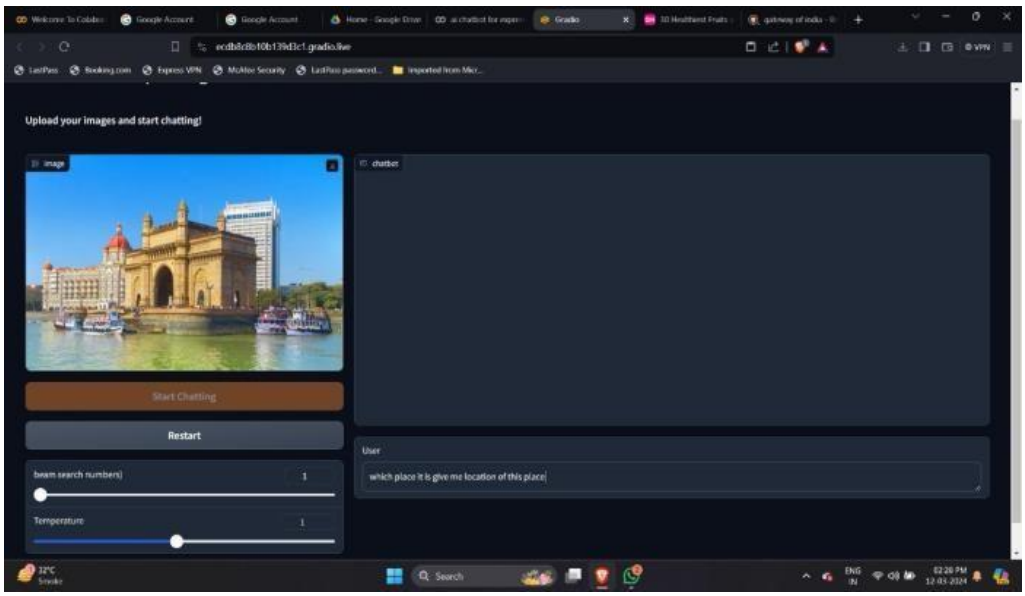

Figure 6. Image upload screen.

In the end, an AI chatbot is a useful tool for conveying visual content since it can take in photos or videos, analyze them using computer vision, and then produce interesting, contextually aware, and detailed verbal responses as shown in Figures 5 to 7.

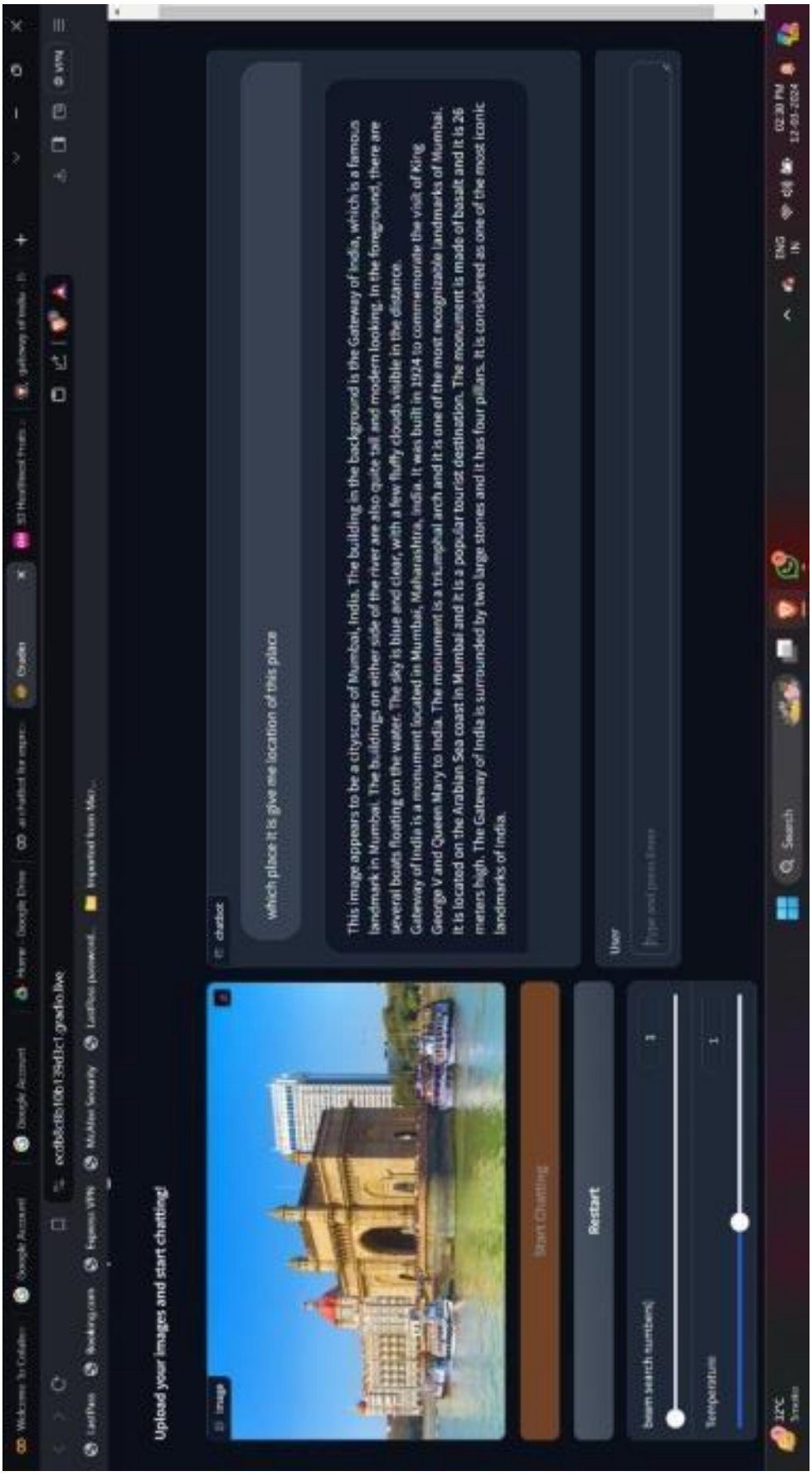

Figure 7. Fetching screen.

CONCLUSION

As a result, vision verbalizer becomes more than just a description generator; it becomes a portal to a more accepting and intelligent digital world. Beyond that of a simple chatbot, its ability to interpret photographs, provide descriptions that are humanlike, and retrieve more data sets it apart from average AI helper. In contrast, it turns into a reliable ally in our investigation of the visual realm, providing a deep comprehension and enhancing our digital encounters.

In addition to its technical features, vision verbalizer represents a change in the way we engage with and interpret visual information, promoting a more meaningful and approachable relationship between people and the enormous visual space of the digital world. With the rapid advancement of technology, vision verbalizer is leading the way as a tool as well as a catalyst for a digital future that is more inclusive and enlightened.

Future Scope

AI chatbots of the future will improve user interactions by smoothly integrating visual content expression. Textual input will be analyzed by sophisticated algorithms to produce films, GIFs, and photos. Advances in NLP will enable chatbots to understand complex concepts and produce accurate and imaginative visual outputs.

AI chatbots will become essential in communicating complicated concepts and feelings through a dynamic combination of text and images as a result of this evolution. By offering captivating and immersive experiences, they will transform communication in the digital sphere and help people connect with technology. Through this integration, users will be able to express themselves more effectively and AI chatbots will be able to service a wider range of requirements in a wider range of businesses and areas.

REFERENCES

1. BigScience Workshop, Scao TL, Fan A, Akiki C, Pavlick E, Ilić S, Hesslow D, Castagné R, Luccioni AS, Yvon F, Gallé M, et al. Bloom: a 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100. November 9, 2022. Available at https://arxiv.org/abs/2211.05100
2. Schwenk D, Khandelwal A, Clark C, Marino K, Mottaghi R. A-OKVQA: a benchmark for visual question answering using world knowledge. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, editors. European Conference on Computer Vision 2022. Cham, Switzerland: Springer Nature; 2022. pp. 149–162.
3. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog. 2019; 1 (8): 9.
4. Surís D, Menon S, Vondrick C. ViperGPT: visual inference via Python execution for reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, October 1–6, 2023. pp. 11888–11898.
5. Rohrbach A, Hendricks LA, Burns K, Darrell T, Saenko K. Object hallucination in image captioning. arXiv preprint arXiv:1809.02156. September 6, 2018. Available at https://arxiv.org/abs/1809.02156
6. Chen J, Zhu D, Haydarov K, Li X, Elhoseiny M. Video ChatCaptioner: towards enriched spatiotemporal descriptions. arXiv preprint arXiv:2304.04227. April 9, 2023. Available at https://arxiv.org/abs/2304.04227
7. Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M, Ring R. Flamingo: a visual language model for few-shot learning. Adv Neural Inform Process Syst. 2022; 35: 23716–23736.
8. Wu C, Yin S, Qi W, Wang X, Tang Z, Duan N. Visual ChatGPT: talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671. March 8, 2023. Available at https://arxiv.org/abs/2303.04671

9. Driess D, Xia F, Sajjadi MS, Lynch C, Chowdhery A, Ichter B, Wahid A, Tompson J, Vuong Q, Yu T, Huang W, Chebotar Y, Sermanet P, Duckworth D, Levine S, Vanhoucke V, Hausman K, Toussaint M, Greff K, Zeng A, Mordatch I, Florence P. PaLM-E: an embodied multimodal language model. arXiv preprint arXiv:2303.03378. March 6, 2023. Available at https://arxiv.org/abs/2303.03378
10. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020; 21 (140): 1–67.

