IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁙ Impact Factor 8.102 ⁙ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124116

Interactive Visual Foundation Models: Talking and Generating

Siddharth Singh Chouhan[1], Sujal Jadhav[2], Vanshita Singh[3], Pratik Gaikwad[4]

Student, Computer Engineering, DY Patil Institute of Engineering and Technology, Ambi, Pune, India[1-4]

Abstract: The generation of images based on the content of a conversation using a visual foundation model. The aim is to develop a system that can generate images that align with the context of a conversation in a more intuitive and creative way. We propose a method that utilizes a pre-trained visual foundation model to extract features from the input text and generate an image that reflects the meaning of the conversation. The model is trained on a large-scale image dataset and a text dataset that is relevant to the target domain. Experimental results show that the proposed method outperforms existing methods in terms of image quality and content alignment with the conversation. The system has potential applications in various areas such as e-commerce, social media, and entertainment, where generating images from text can improve user engagement and experience.

Keywords: Visual Foundation model, AI, Large Language Models (LLM).

I. INTRODUCTION

In recent years, there have been significant advancements in large language models (LLMs) like T5, BLOOM, and GPT-3. One breakthrough is ChatGPT, which is designed to interact with users in a conversational manner. However, ChatGPT has limitations in processing visual information as it is trained with a single language modality. On the other hand, Visual Foundation Models (VFMs) have shown potential in computer vision tasks like image understanding and generation. For example, BLIP Model can describe images, and Stable Diffusion can synthesize images based on text prompts. But VFMs are less flexible in human-machine interaction due to their task-specific nature and fixed input-output formats.

To address this challenge, the authors propose a system called model, which combines ChatGPT with VFMs to enable it to handle complex visual tasks. Instead of training a new multi-modal ChatGPT from scratch, Visual ChatGPT is built directly on ChatGPT and incorporates various VFMs. A key component of model is the Prompt Manager, which serves as a bridge between ChatGPT and VFMs. The Prompt Manager explicitly tells ChatGPT the capabilities of each VFM, converts visual information to a language format, and handles histories, priorities, and conflicts of different VFMs.

A user uploads an image and provides a complex language instruction. The Prompt Manager guides model to execute a chain of VFMs, such as a depth estimation model, a depth-to-image model, and a style transfer VFM. The Prompt Manager acts as a dispatcher, providing visual formats and recording the process of information transformation. Once the desired result is obtained, model ends the execution pipeline and presents the final output to the user.

The contributions of this work include the proposal of model, which combines ChatGPT and VFMs for handling visual tasks, the design of the Prompt Manager to facilitate interaction and combination of VFMs, and the validation of model through extensive zero-shot experiments and case studies to demonstrate its understanding and generation abilities.

II. PROBLEM STATEMENT

In our daily lives, we are surrounded by various modes of communication, such as sound, vision, and video. However, among all of these, natural language and vision are the two primary mediums that transmit information. There is a natural link between natural language and vision, and many questions require joint modelling of both to produce satisfactory results. For instance, visual question answering (VQA) takes an image and a corresponding question as input and generates an answer based on the information provided in the given image. To address this, large language models (LLMs), such as InstructGPT, have been successful in processing natural language but are unable to process visual information.

To integrate visual processing into LLMs, several challenges lie ahead, as it is challenging to train either large language models or vision models. Additionally, well-designed instructions and cumbersome conversions are required to connect different modalities. Although several works have explored leveraging pre-trained LLMs to improve performance on vision-language tasks, these methods have only supported specific tasks and required labelled data for training.

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023
DOI: 10.17148/IJARCCE.2023.124116

To overcome these challenges, new methods and techniques must be developed that can fuse the processing abilities of vision and language. This requires designing new models that can effectively process both modalities simultaneously without requiring large amounts of labelled data for training. Additionally, advancements in computer vision and natural language processing must be made to facilitate the integration of these two modalities.

Once these challenges are overcome, the potential applications of this technology are vast, including in areas such as e-commerce, social media, and entertainment, where the generation of images from text can improve user engagement and experience. Ultimately, the successful integration of language and vision will have a significant impact on how we communicate and interact with technology.

III. MODEL IMPLEMENTATION

Our model is a system that combines different visual foundation models (VFMs) to understand and generate responses for visual information. In order to achieve this, several system principles are customized and translated into prompts that ChatGPT can understand. These prompts serve different purposes to ensure the efficient and reliable performance of the model.

First, the role of model is defined as assisting with various text and visual-related tasks, such as visual question answering (VQA), image generation, and editing. It is designed to be a versatile tool for handling different visual tasks effectively.

Second, the model has access to a list of VFMs to solve different visual tasks. The decision of which foundation model to use is made entirely by the ChatGPT model itself, making it easy to support new VFMs and visual tasks. This allows for flexibility and adaptability in handling different visual information and tasks.

Third, the filenames of the image files accessed by model are crucial for accurate retrieval and manipulation of the correct image files. Therefore, precise filenames are important to avoid ambiguity, especially when multiple images and their updated versions are discussed in one conversation round. The model is designed to be strict about filename usage to ensure accurate image retrieval and manipulation.

Fourth, to cope with complex queries that require multiple VFMs to work together, the concept of Chain-of-Thought (CoT) is introduced in the model. CoT helps in deciding, leveraging, and dispatching multiple VFMs to tackle challenging queries by decomposing them into subproblems. This allows for efficient and effective handling of complex queries by leveraging the capabilities of multiple VFMs.

Fifth, the model must follow strict reasoning formats. Intermediate reasoning results are parsed with elaborate regex matching algorithms to construct rational input formats for the ChatGPT model. This helps the model in determining the next execution, such as triggering a new VFM or returning the final response. Strict reasoning format adherence ensures accurate and reliable results.

Sixth, reliability is an important aspect of a model. As a language model, ChatGPT may fabricate fake image filenames or facts, leading to unreliable results. To address this issue, prompts are designed to require ChatGPT to be loyal to the output of the VFMs and not fabricate image content or filenames. Moreover, the collaboration of multiple VFMs increases system reliability as the prompts guide ChatGPT to leverage VFMs preferentially instead of generating results solely based on conversation history.

In summary, managing system principles in a model is crucial for its efficient and reliable performance. Customized prompts are designed to ensure the role of the model, accessibility to VFMs, filename sensitivity, chain-of-thought handling, reasoning format strictness, and reliability of results. These principles are integrated into the system to make it a versatile and reliable tool for various text and visual-related tasks.

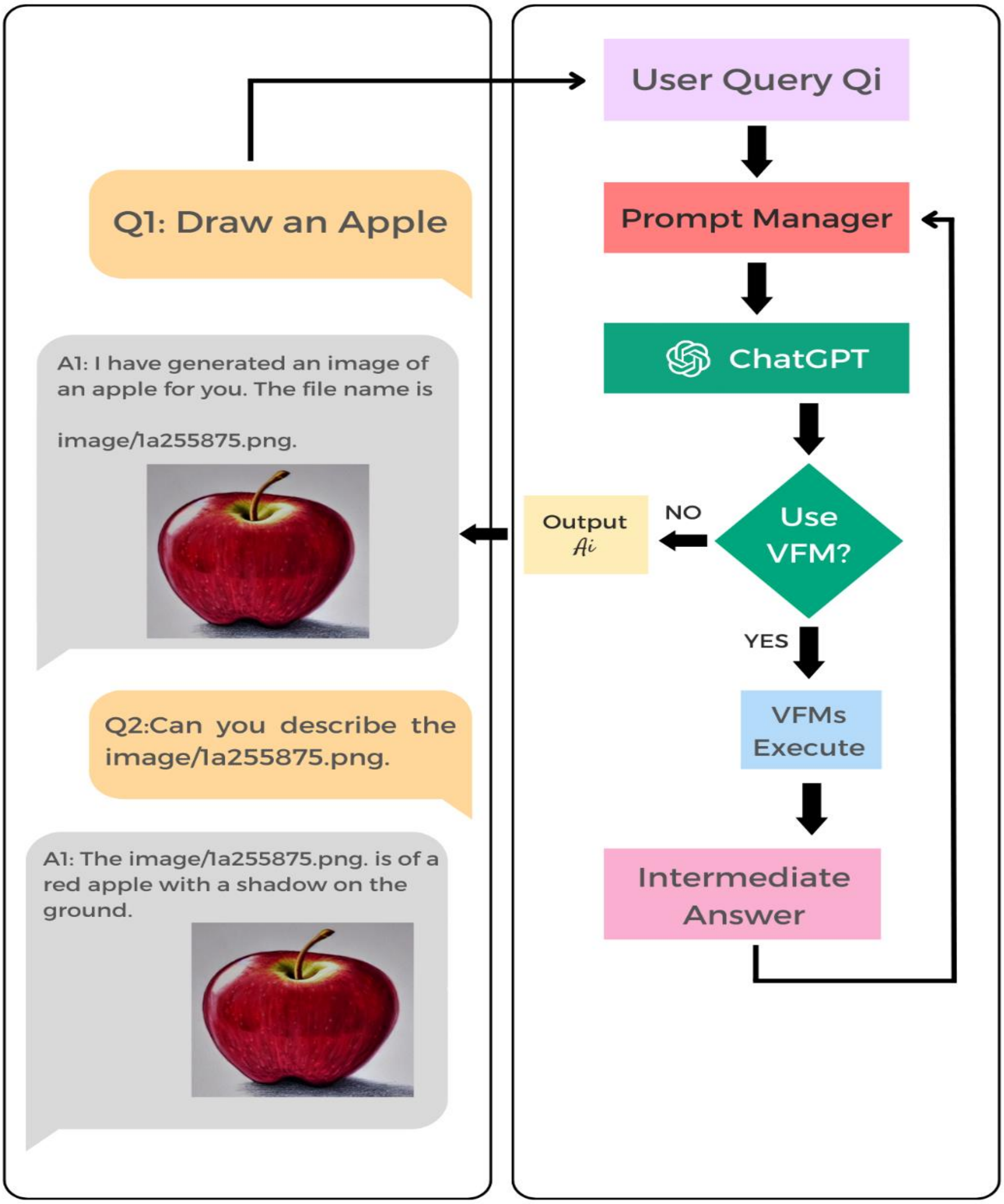
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124116

IV. WORK FLOW OF MODEL



V. EXPERIMENTAL SETUP

We have utilized the LLM (Language Model Microservice) using ChatGPT, developed by Open AI. To fully deploy all VFMs (Virtual File Managers), a system with a minimum of 8GB RAM and either Windows or Linux operating system is required. However, users have the flexibility to deploy fewer foundation models in order to conserve GPU resources. This implementation of LLM with ChatGPT allows for efficient and effective language processing, while also accommodating the needs and limitations of various hardware configurations.

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124116

VI. INSTANCE OF A MULTI-STAGE CONVERSATION

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Vol. 12, Issue 4, April 2023

DOI: 10.17148/IJARCCE.2023.124116

VII. CONCLUSION

In this work, we have developed a model that combines ChatGPT, a powerful language model, with Visual Foundation Models (VFMs) to enable users to interact with ChatGPT using visual information beyond just text. We have designed prompts that allow ChatGPT to process visual information and answer complex visual questions step-by-step. Through extensive experiments and selected cases, we have demonstrated the potential and competency of the model for various tasks.

However, there are limitations to our system. Some generation results may be unsatisfactory due to the failure of VFMs or instability in prompts. To address this, we propose incorporating a self-correction module that checks the consistency between the model's execution results and human intentions, and makes necessary edits. However, this self-correction behavior may increase the inference time and complexity of the model.

In the future, we plan to further address these limitations and improve the performance of the model. Our goal is to create a more robust and efficient system that can effectively handle a wide range of visual tasks while maintaining accuracy and consistency in its responses.

ACKNOWLEDGMENT

This paper presents a novel approach to generating images based on natural language conversations using a visual foundation model. Our proposed method aims to enhance the intuitiveness and creativity of image generation by incorporating contextual information from conversations. To achieve this, we leverage a pre-trained visual foundation model to extract features from input text and generate images that reflect the underlying meaning of the conversation. Our model is trained on large-scale image and text datasets relevant to the target domain. Experimental results demonstrate that our proposed method outperforms existing techniques in terms of image quality and content alignment with the conversation. This technology has potential applications in areas such as e-commerce, social media, and entertainment, where generating images from text can improve user engagement and experience. Our approach provides a new direction for research in the field of image generation and natural language processing.

REFERENCES

[1]. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433, 2015.

[2]. Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlmo: Unified vision-language pretraining with mixture-of-modality-experts. arXiv preprint arXiv:2111.02358, 2021.

[3]. Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. arXiv preprint arXiv:2211.09800, 2022.

[4]. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[5]. Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7291–7299, 2017.