

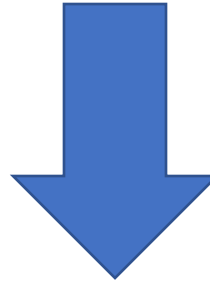
Paper Review : Visual ChatGPT

Talking, Drawing and Editing with Visual Foundation Models

Visual ChatGPT : Motivation

Background :

- Current ChatGPT model is limited within language domain.
- Training large foundation model that includes many modalities is very expensive.
- Is it possible to create a ChatGPT-like system that can understand and generate vision modality without doing training?



Key Idea : Combine ChatGPT with existing vision foundation models !

Note :

- Vision foundation models are multimodal models that also includes vision, but with much smaller parameters than ChatGPT.



why you cannot support image understanding and generation



As a language model, my primary focus is on processing and generating natural language text. While I am capable of understanding and processing some types of visual information, such as textual descriptions of images, I do not have the capability to directly process or generate images.



Image understanding and generation typically require specialized models and techniques, such as computer vision algorithms and deep learning models that are specifically designed to work with visual data. These models typically operate on pixel-level information and use advanced techniques such as convolutional neural networks (CNNs) and generative adversarial networks (GANs) to process and generate images.

While there are some research efforts focused on creating models that can integrate natural language and visual information, these are still relatively new and experimental areas of research. As such, it is currently not within my capabilities to support image understanding and generation to the same degree as specialized image processing models.

Figure 1. Example of ChatGPT output

Some Examples of Vision Foundation Models

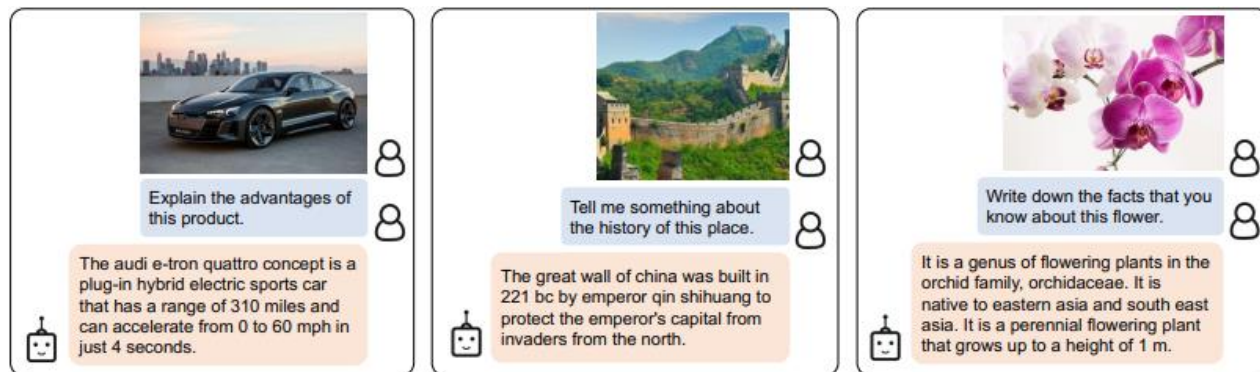


Figure 2. Illustration of BLIP [2] model to understand and provide the description of an image.



Figure 3. Examples of Stable Diffusion [3] model to synthesize an image based on text prompts.

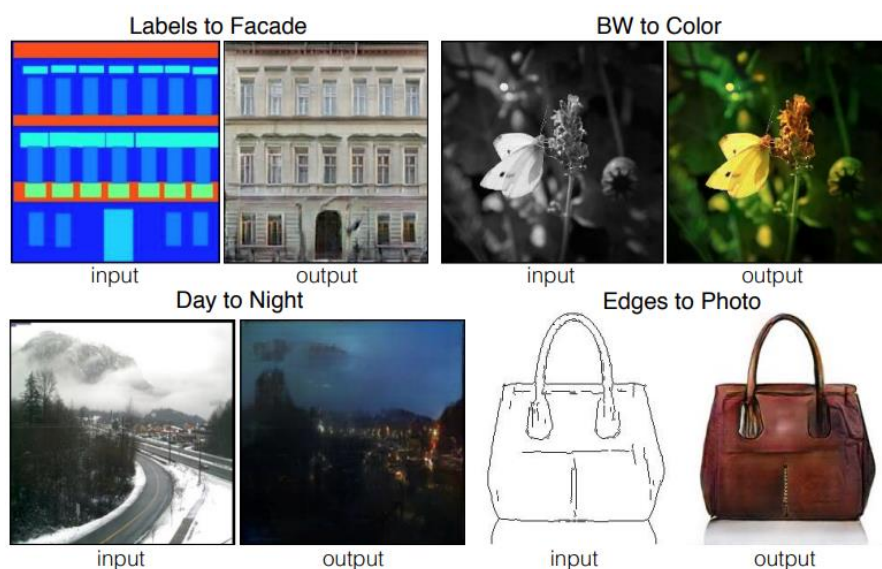


Figure 4. Illustration of Pix2Pix [4] model to generate image-to-image translation.

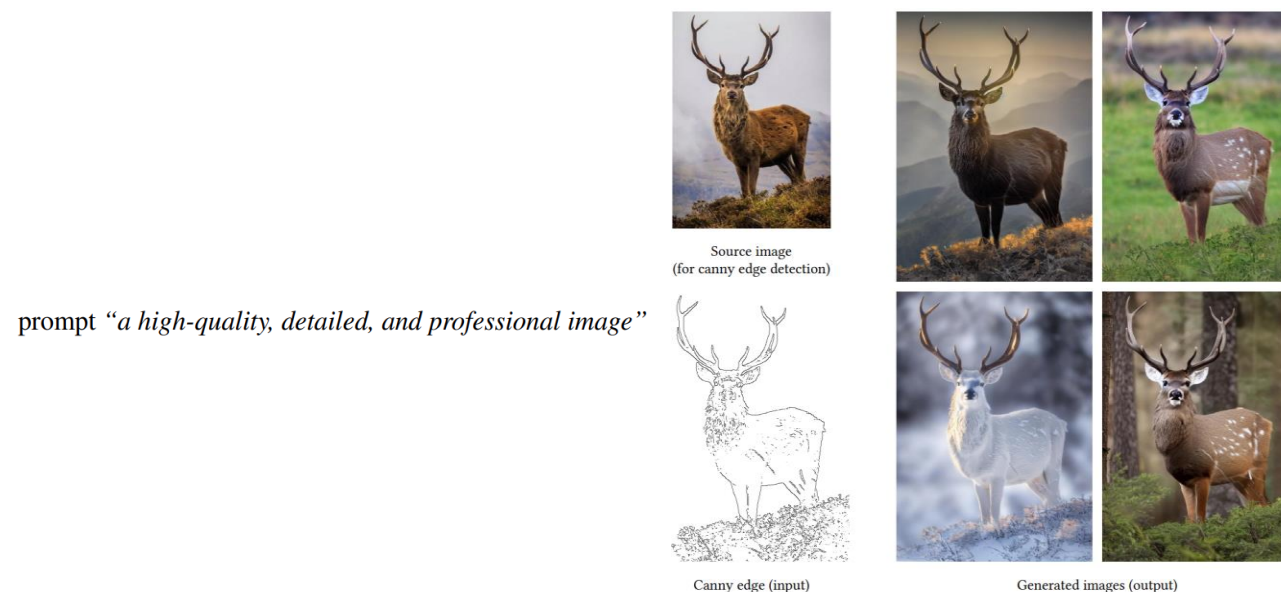


Figure 5. Illustration of ControlNet [5] model to control generated images₃ from large diffusion models conditioned on canny images.

Visual ChatGPT : Major Contributions

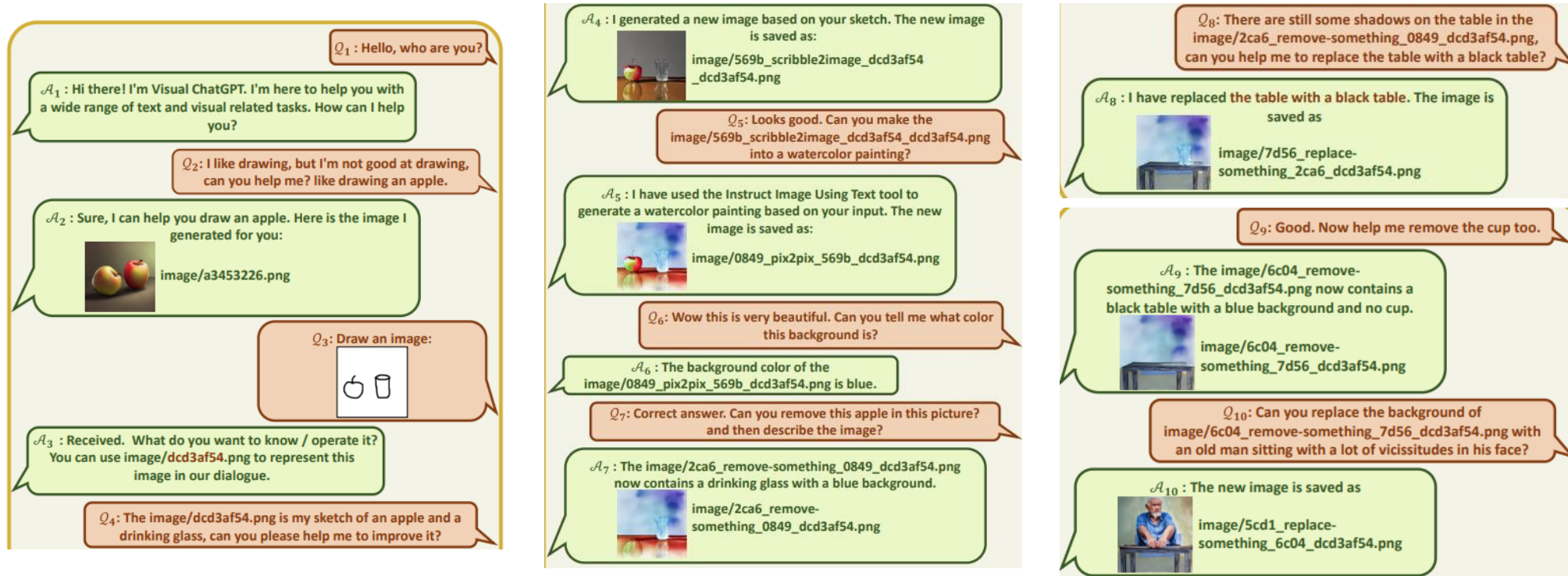


Figure 6. Example of Visual ChatGPT [1] multiple rounds of conversation.

Major contributions :

- First work to **combine** the power of **ChatGPT** and existing Vision Foundation Models (**VFM**s).
- Design a novel module called **prompt manager** that act as an operator to interact with ChatGPT and 22 different VFMs.

Visual ChatGPT : Notation Preliminaries

History of Dialogue $\mathcal{H}_{<i}$

User query \mathcal{Q}_i

History of Reasoning $\mathcal{R}_i^{(<j)}$

Intermediate Answer $\mathcal{A}^{(j)}$

Prompt Manager \mathcal{M}

Notation meaning :

- History of Dialogue ($\mathcal{H}_{<i}$) : Concatenation of previous QA pairs.
- User query (\mathcal{Q}_i) : linguistic and/or vision input.
- History of Reasoning ($\mathcal{R}_i^{(<j)}$) : All previous reasoning histories from j invoked VFMs for i-th round of conversation.
- Intermediate Answer ($\mathcal{A}^{(j)}$) : The answer generated step-by-step to achieve the final answer for user.
- Prompt Manager (\mathcal{M}) : Convert all the visual signals into language so that ChatGPT can understand.

Visual ChatGPT Brain : ReAct Framework

Published as a conference paper at ICLR 2023

REACT: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS

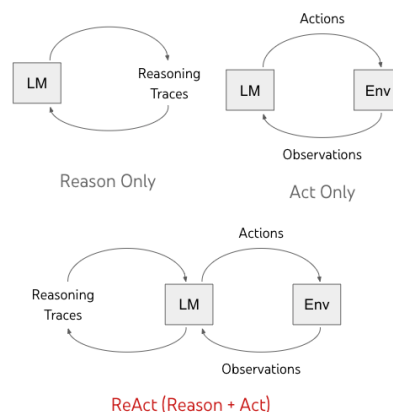
Shunyu Yao^{*1}, Jeffrey Zhao², Dian Yu², Nan Du², Izhak Shafran², Karthik Narasimhan¹, Yuan Cao²

¹Department of Computer Science, Princeton University

²Google Research, Brain team

¹{shunyuy, karthikn}@princeton.edu

²{jeffreyzhao, dianyu, dunan, izhak, yuancao}@google.com



Key Lessons :

- Novel technique that combine reasoning and acting to solve various tasks in Language Model (LM).
- Environment can be external APIs (e.g., Wikipedia API, etc) or take an action from custom tools.

Note:

- Thought : Reasoning trace
- Action : Using LM to take an action based on its reasoning
- Observation : Results from taking the action

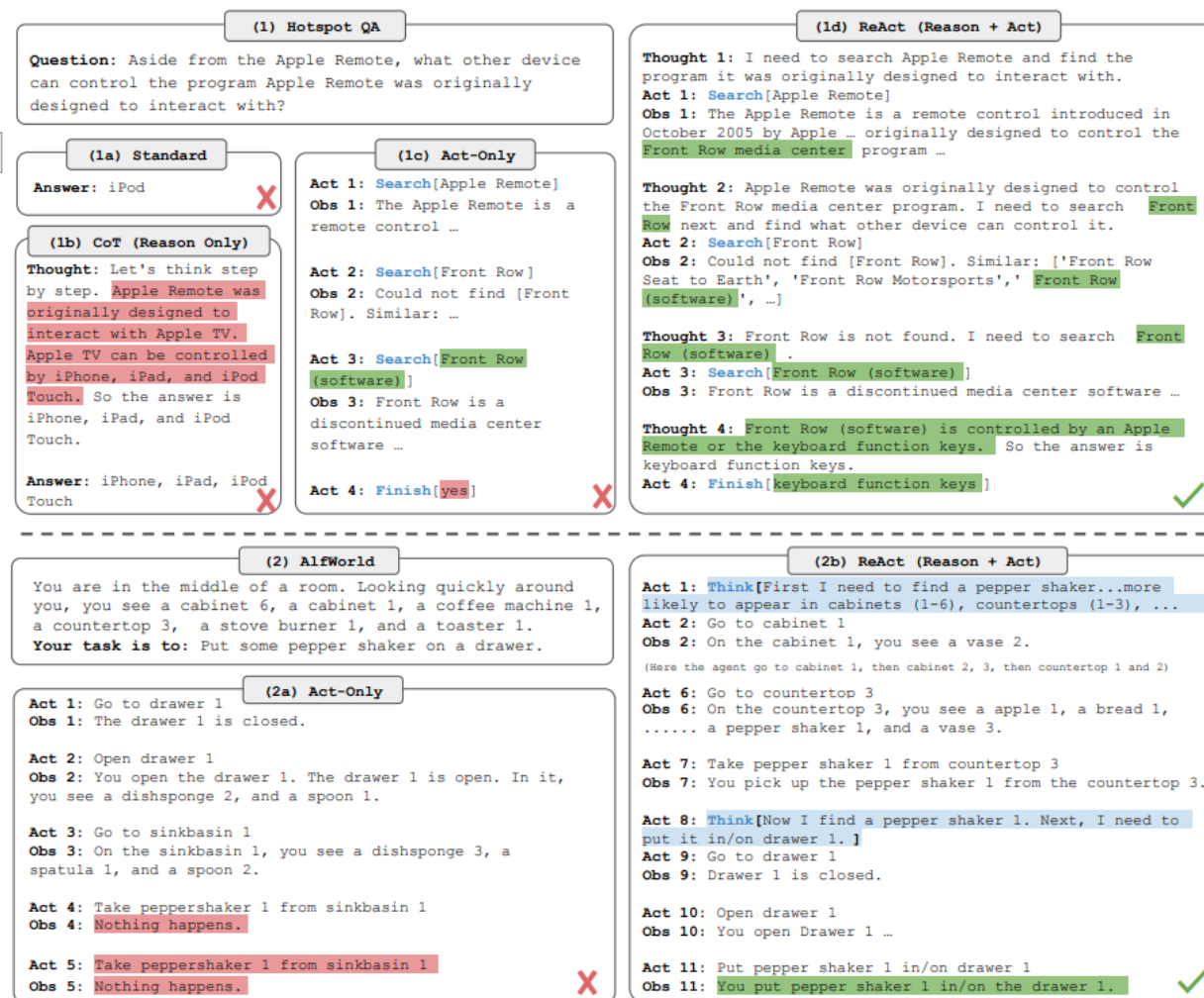
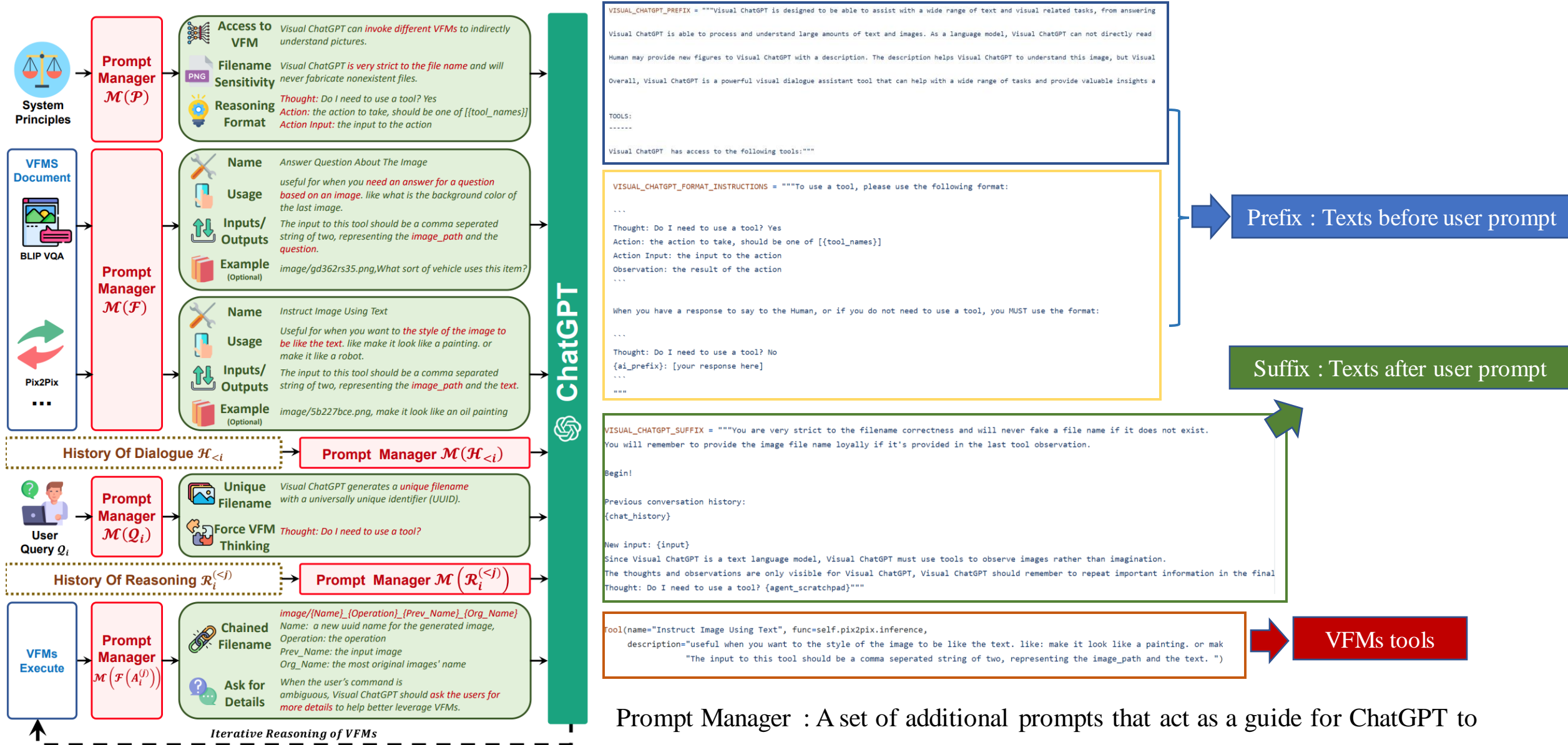


Figure 6. Example of outputs generated by ReAct framework [6].

Visual ChatGPT : Prompt Manager



Prompt Manager : A set of additional prompts that act as a guide for ChatGPT to understand and create visual outputs with the help from other VFMs.

Visual ChatGPT : Prompt Manager

```
VISUAL_CHATGPT_PREFIX = """Visual ChatGPT is designed to be able to assist with a wide range of text and visual related tasks, from answering  
  
Visual ChatGPT is able to process and understand large amounts of text and images. As a language model, Visual ChatGPT can not directly read  
Human may provide new figures to Visual ChatGPT with a description. The description helps Visual ChatGPT to understand this image, but Visual  
Overall, Visual ChatGPT is a powerful visual dialogue assistant tool that can help with a wide range of tasks and provide valuable insights a  
  
TOOLS:  
-----  
  
Visual ChatGPT has access to the following tools: """
```

Explaining the system message (the first part of conversation) for Visual ChatGPT.

```
You are ChatGPT, a large language model trained by OpenAI. Answer as  
concisely as possible. Knowledge cutoff: {knowledge_cutoff} Current date:  
{current_date}
```

Figure 7. System message used by OpenAI for ChatGPT [7].

```
VISUAL_CHATGPT_FORMAT_INSTRUCTIONS = """To use a tool, please use the following format:  
  
...  
Thought: Do I need to use a tool? Yes  
Action: the action to take, should be one of [{tool_names}]  
Action Input: the input to the action  
Observation: the result of the action  
...  
  
When you have a response to say to the Human, or if you do not need to use a tool, you MUST use the format:  
  
...  
Thought: Do I need to use a tool? No  
{ai_prefix}: [your response here]  
...  
"""
```

Set of instructions to make sure that Visual ChatGPT works as expected (based on ReAct framework).

```
VISUAL_CHATGPT_SUFFIX = """You are very strict to the filename correctness and will never fake a file name if it does not exist.  
You will remember to provide the image file name loyally if it's provided in the last tool observation.  
  
Begin!  
  
Previous conversation history:  
{chat_history}  
  
New input: {input}  
Since Visual ChatGPT is a text language model, Visual ChatGPT must use tools to observe images rather than imagination.  
The thoughts and observations are only visible for Visual ChatGPT, Visual ChatGPT should remember to repeat important information in the final  
Thought: Do I need to use a tool? {agent_scratchpad}"""
```

Set of prompts to make sure that ChatGPT doesn't hallucinate and only depend on previous outputs.

```
Tool(name="Instruct Image Using Text", func=self.pix2pix.inference,  
      description="useful when you want to the style of the image to be like the text. like: make it look like a painting. or mak  
      "The input to this tool should be a comma seperated string of two, representing the image_path and the text. ")
```

List of VFMs tools used by Visual ChatGPT to understand and generate images.

Prompt Manager : A set of additional prompts that act as a guide for ChatGPT to understand and create visual outputs with the help from other VFMs.

Visual ChatGPT : Overall Architecture

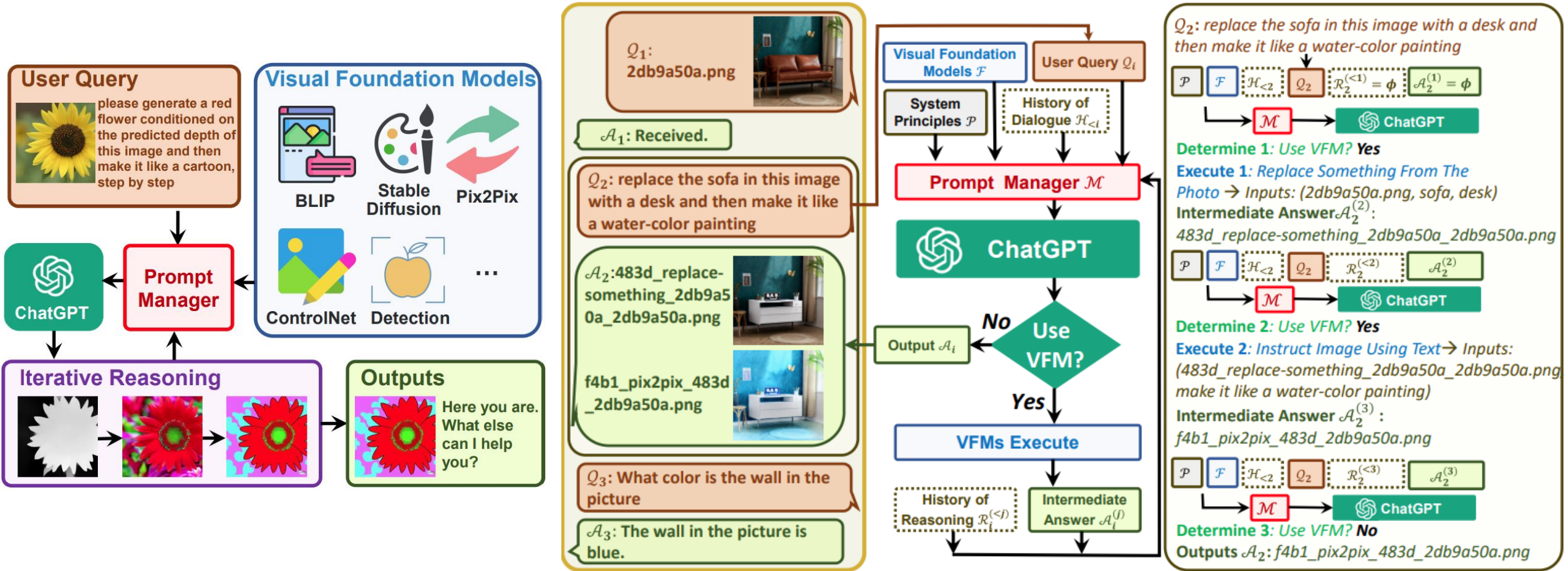


Figure 8. Architecture of Visual ChatGPT [1].

Visual ChatGPT : Limitations

Some limitations from the proposed Visual ChatGPT model :

- Heavily rely on existing ChatGPT and VFMs.
- Requires many prompt engineering (e.g., set of instructions, guidelines, confirmations, etc).
- Cannot achieve real-time results if it triggers many VFMs to answer the user query.
- Current ChatGPT API has maximum number of tokens which are 4096 (already solved in GPT-4).
- Need quite large storage (~ 70 GB GPU RAM) to include all 22 different VFMs mentioned in the paper.
- Need internet connection since ChatGPT API can only be accessed by those who have the internet.

Visual ChatGPT : Possible Research Directions

Some possible research directions in vision-language tasks based on the recent work, Visual ChatGPT :

- Research about efficient prompt that doesn't require expertise from specific field.
- Investigating techniques to speed up the interaction between ChatGPT and VFMs for enabling real-time results.
- Research the possibility to be run in CPU or resource-constrained devices.

Additional Details in Visual ChatGPT

Remove Objects from Image [10, 35]	
Replace Objects from Image [10, 35]	
Change Image by the Text [35]	
Image Question Answering [23]	
Image-to-Text [23]	Text-to-Image [35]
Image-to-Edge [45]	Edge-to-Image [53]
Image-to-Line [16]	Line-to-Image [53]
Image-to-Hed [44]	Hed-to-Image [53]
Image-to-Seg [24]	Seg-to-Image [53]
Image-to-Depth [34, 33]	Depth-to-Image [53]
Image-to-NormalMap [34, 33]	NormalMap-to-Image [53]
Image-to-Sketch [44]	Sketch-to-Image [53]
Image-to-Pose [6]	Pose-to-Image [53]

Figure 9. 22 different VFM models used in Visual ChatGPT [1].

GPU memory usage

Here we list the GPU memory usage of each visual foundation model, you can specify which one you like:

Foundation Model	GPU Memory (MB)
ImageEditing	3981
InstructPix2Pix	2827
Text2Image	3385
ImageCaptioning	1209
Image2Canny	0
CannyText2Image	3531
Image2Line	0
LineText2Image	3529
Image2Hed	0
HedText2Image	3529
Image2Scribble	0
ScribbleText2Image	3531
Image2Pose	0
PoseText2Image	3529
Image2Seg	919
SegText2Image	3529
Image2Depth	0
DepthText2Image	3531
Image2Normal	0
NormalText2Image	3529
VisualQuestionAnswering	1495

Figure 10. GPU memory needs for including VFMs in Visual ChatGPT[1].