Language Is Not All You Need: Aligning Perception with Language Models

Shaohan Huang∗ , Li Dong∗ , Wenhui Wang∗ , Yaru Hao∗ , Saksham Singhal∗ , Shuming Ma∗ , Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, Furu Wei†

Microsoft

Introduction

- LLMs have successfully served as a **general-purpose interface** across various natural language tasks. The LLM-based interface can be adapted to a task as long as we are able to transform the input and output into texts.

- It is still struggling to natively use LLMs for multimodal data, such as image, and audio.

- Multimodal perception is a necessity to achieve artificial general intelligence, in terms of knowledge acquisition and grounding to the real world. Unlocking multimodal input greatly widens the applications of language models to more high-value areas, such as multimodal machine learning, document intelligence, and robotics.

key takeaways

1. From LLMs to MLLMs:

1. multimodal perception enables LLMs to acquire commonsense knowledge beyond text descriptions.
2. aligning perception with LLMs opens the door to new tasks.
3. the capability of perception unifies various APIs, as graphical user interfaces are the most natural and unified way to interact with.

- KOSMOS-1 models are trained on **web-scale multimodal corpora**, which ensures that the model robustly learns from diverse sources. We not only use a large-scale text corpus but also mine high-quality image-caption pairs and arbitrarily interleaved image and text documents from the web.

key takeaways

2. Language models as general-purpose interfaces:

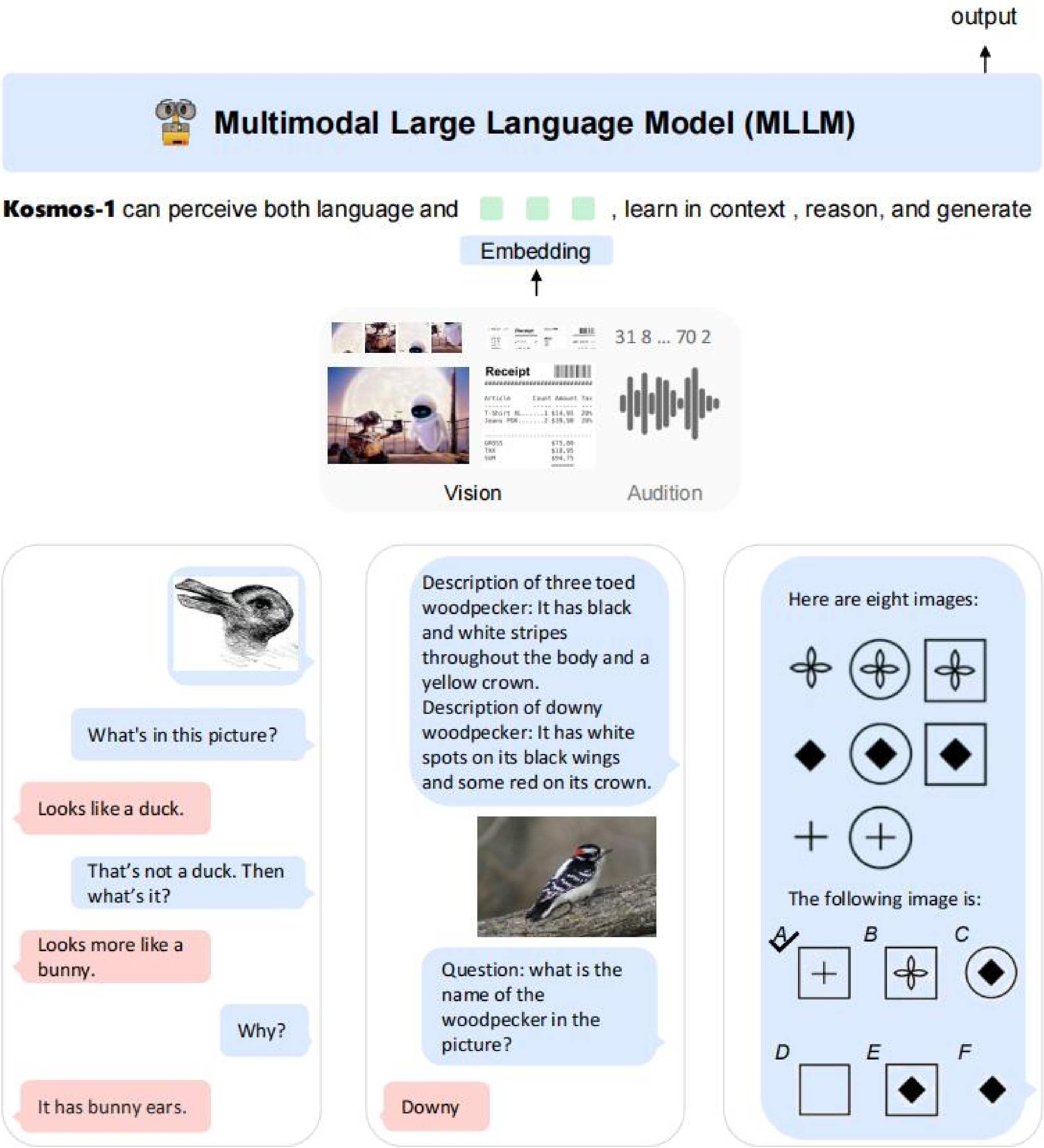
1. because of the open-ended output space, we are able to unify various task predictions as texts.
2. natural-language instructions and action sequences can be well handled by language models.
3. LLMs also serve as basic reasoners, which is complementary to perception modules on complex tasks.

- So it is natural to align world, action, and multimodal perception with the general-purpose interface.

key takeaways

3. New capabilities of MLLMs:

1. conduct zero- and few-shot multimodal learning by using natural language instructions and demonstration examples.
2. observe promising signals of nonverbal reasoning by evaluating the Raven IQ test, which measures the fluid reasoning ability of humans.
3. MLLMs naturally support multi-turn interactions for general modalities, such as multimodal dialogue.


output
Multimodal Large Language Model (MLLM)
Kosmos-1 can perceive both language and , learn in context , reason, and generate
Embedding
Receipt
31 8 ... 70 2
Vision
Audition
What's in this picture?
Looks like a duck.
That's not a duck. Then what's it?
Looks more like a bunny.
Why?
It has bunny ears.
Description of three toed woodpecker: It has black and white stripes throughout the body and a yellow crown.
Description of downy woodpecker: It has white spots on its black wings and some red on its crown.
Question: what is the name of the woodpecker in the picture?
Downy
Here are eight images:
The following image is:
A
B
C
D
E
F

KOSMOS-1

- **Input Representation**
  - use <s> and </s> to denote start and end-of sequence.
  - special tokens <image> and </image> indicate the beginning and end of encoded image embeddings.
  - An embedding module is used to encode both text tokens and other input modalities into vectors:
    - use a lookup table to map tokens into embeddings.
    - represent modalities of continuous signals inputs as discrete code and then regard them as “foreign languages”

Datasets	Format Examples
Text	<s> KOSMOS-1 can perceive multimodal input, learn in context, and generate output. </s>
Image-Caption	<s> <image> Image Embedding </image> WALL-E giving potted plant to EVE. </s>
Multimodal	<s> <image> Image Embedding </image> This is WALL-E. <image> Image Embedding </image> This is EVE. </s>

KOSMOS-1

- **Multimodal Large Language Models (MLLMs)**
  - MAGNETO:

    a Transformer variant, as the backbone architecture.

    (It introduces an extra LayerNorm to each sublayer (i.e., multi-head self-attention, and feed-forward network))

  - XPOS:

    employ XPOS relative position encoding for better long-context modeling.

    The method can better generalize to different lengths.

KOSMOS-1

- **Training Objective**
  - training is conducted on web-scale multimodal corpora, including monomodal data, cross-modal paired data, and interleaved multimodal data
  - The models are trained with the next-token prediction task. The training objective is to maximize the log-likelihood of tokens in examples. Notice that only discrete tokens, such as text tokens, are accounted for in the training loss.

Model Training

- Datasets
  - Text Corpora:
    1. The Pile : 800 GB English text corpus combining 22 diverse sources.
    2. Common Crawl

- **Academic**: NIH Exporter
- **Internet**: Pile-CC, OpenWebText2, Wikipedia (English), CC-2020-50, CC-2021-04, Realnews
- **Prose**: BookCorpus2, Books3, Gutenberg [RPJ+20], CC-Stories

Datasets	Tokens (billion)	Weight (%)	Epochs
OpenWebText2	14.8	21.8%	1.47
CC-2021-04	82.6	17.7%	0.21
Books3	25.7	16.2%	0.63
CC-2020-50	68.7	14.7%	0.21
Pile-CC	49.8	10.6%	0.21
Realnews	21.9	10.2%	0.46
Wikipedia	4.2	5.4%	1.29
BookCorpus2	1.5	1.1%	0.75
Gutenberg (PG-19)	2.7	1.0%	0.38
CC-Stories	5.3	1.0%	0.19
NIH ExPorter	0.3	0.2%	0.75

Table 20: Language datasets used to train the KOSMOS-1 model.

Model Training

- Datasets
  - Image-Caption Pairs:
    1. LAION-2B
    2. LAION-400M
    3. COYO-700M
    4. Conceptual Captions.

  - Interleaved Data:

    collect a large corpus of 2 billion web pages from the snapshots of common crawls. After applying filters, end up with about 71 million documents for training.

Model Training

- Language-Only Instruction Tuning

Continue-train the model with the instruction data in the format of (instructions, inputs, and outputs). The instruction data is language-only, which is mixed with training corpora. The tuning process is conducted as language modeling. Notice that instructions and inputs are not accounted for in the loss.

Datasets:

1. Unnatural Instructions: generate instructions for various natural language processing tasks
2. FLANv2: cover diverse types of language understanding tasks, such as reading comprehension, commonsense reasoning, and closed-book question answering.

Evaluation

- Language tasks
  - Language understanding
  - Language generation
  - OCR-free text classification
- Cross-modal transfer
  - Commonsense reasoning
- Nonverbal reasoning
  - IQ Test (Raven's Progressive Matrices)
- Perception-language tasks
  - Image captioning
  - Visual question answering
  - Web page question answering
- Vision tasks
  - Zero-shot image classification
  - Zero-shot image classification with descriptions

Evaluation

1. Perception-Language Tasks
   - Image captioning

     datasets：MS COCO Caption， Flickr30k

     prompt：“An image of”

   - Visual question answering

     datasets：VQAv2， VizWiz

     prompt：“Question: {question} Answer: {answer}”

Model	COCO CIDEr	COCO SPICE	Flickr30k CIDEr	Flickr30k SPICE
ZeroCap	14.6	5.5	-	-
VLKD	58.3	13.4	-	-
FewVLM	-	-	31.0	10.0
METALM	82.2	15.7	43.4	11.7
Flamingo-3B*	73.0	-	60.6	-
Flamingo-9B*	79.4	-	61.5	-
KOSMOS-1 (1.6B)	**84.7**	**16.8**	**67.1**	**14.5**

Model	COCO			Flickr30k		
	$k=2$	$k=4$	$k=8$	$k=2$	$k=4$	$k=8$
Flamingo-3B	-	85.0	90.6	-	72.0	71.7
Flamingo-9B	-	93.1	**99.0**	-	72.6	**73.4**
KOSMOS-1 (1.6B)	**99.6**	**101.7**	96.7	**70.0**	**75.3**	68.0

Model	VQAv2	VizWiz
Frozen	29.5	-
VLKDViT-B/16	38.6	-
METALM	41.1	-
Flamingo-3B*	49.2	28.9
Flamingo-9B*	**51.8**	28.8
KOSMOS-1 (1.6B)	51.0	**29.2**

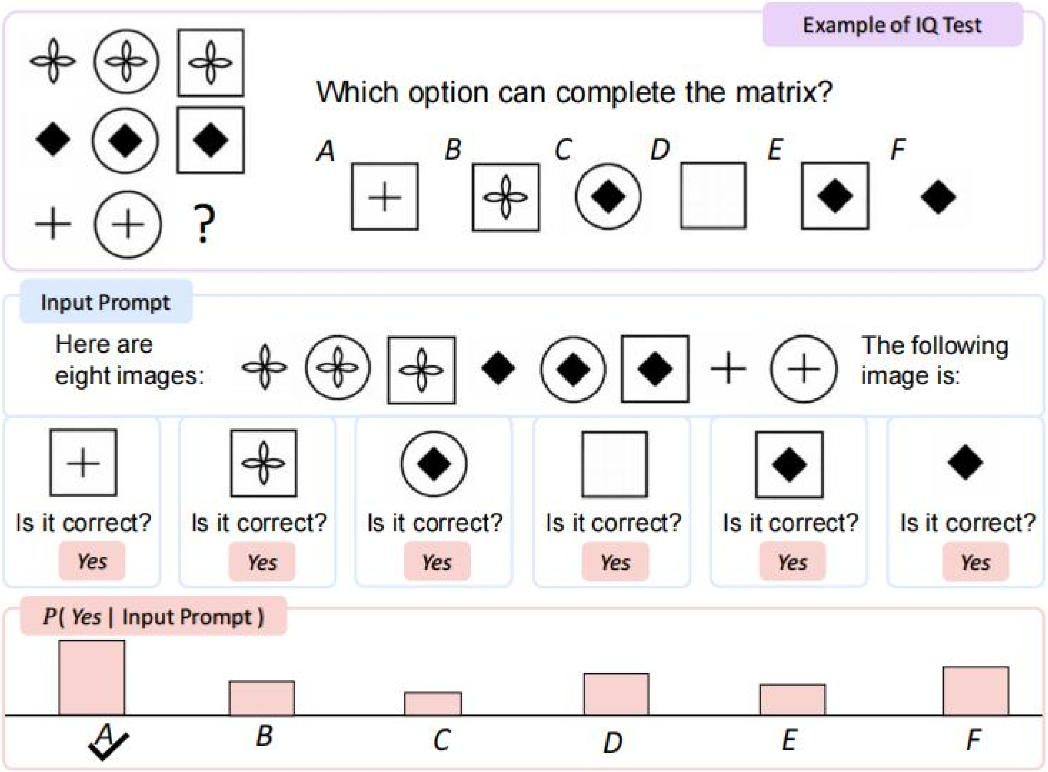
Model	VQAv2			VizWiz		
	$k=2$	$k=4$	$k=8$	$k=2$	$k=4$	$k=8$
Frozen	-	38.2	-	-	-	-
METALM	-	45.3	-	-	-	-
Flamingo-3B	-	53.2	55.4	-	34.4	38.4
Flamingo-9B	-	**56.3**	**58.0**	-	34.9	**39.4**
KOSMOS-1 (1.6B)	**51.4**	51.8	51.4	**31.4**	**35.3**	39.0

Evaluation

2. IQ Test: Nonverbal Reasoning

- **Raven's Progressive Matrices**: the models have to recognize abstract concepts and identify the underlying patterns of given images

  prompt: "Here are three/four/eight images:", "The following image is:", "Is it correct?"



Method	Accuracy
Random Choice	17%
KOSMOS-1	22%
w/o language-only instruction tuning	**26%**

Table 6: Zero-shot generalization on Raven IQ test.

Evaluation

3. OCR-Free Language Understanding

- The task evaluates a model's ability to read and comprehend the meaning of words and sentences directly from the images
- datasets： Rendered SST-2， HatefulMemes
- prompt: "Question: what is the sentiment of the opinion? Answer: {answer}" "Question: does this picture contain real hate speech? Answer: {answer}"

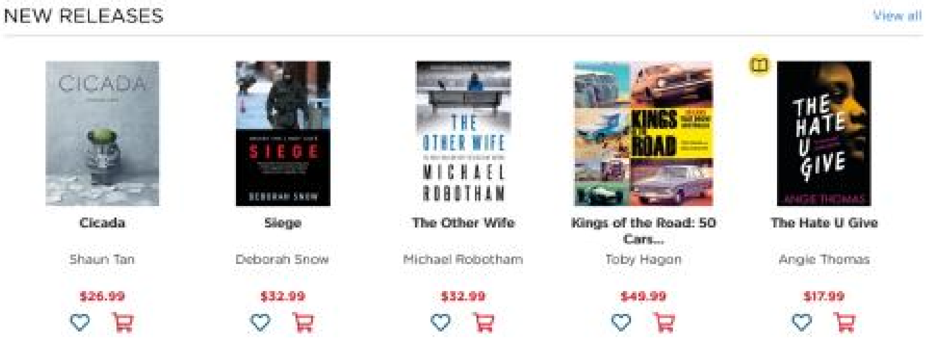
Model	HatefulMemes	Rendered SST-2
CLIP ViT-B/32	57.6	59.6
CLIP ViT-B/16	61.7	59.8
CLIP ViT-L/14	63.3	64.0
Flamingo-3B	53.7	-
Flamingo-9B	57.0	-
KOSMOS-1 (1.6B)	**63.9**	**67.1**

Evaluation

4. Web Page Question Answering

- The task evaluate model's ability to understand the semantics and the structure of web pages(table, html...)
- datasets: Web-based Structural Reading Comprehension (WebSRC)
- prompt: "Given the context below from web page, extract the answer from the given text like this: Qusestion: Who is the publisher of this book? Answer: Penguin Books Ltd. Context: {WebText} Q: {question} A: {answer}"

(a) Question is "*What is the type of this drive?*"



(b) Question is "*Who is the author of "Cicada"?*"

Models	EM	F1
Using extracted text		
LLM	7.6	17.9
KOSMOS-1	**15.8**	**31.3**
Without using extracted text		
KOSMOS-1	3.8	10.6

Evaluation

5. Multimodal Chain-of-Thought Prompting

- rationale first, then answer.
- datasets: Rendered SST-2
- prompt: “Introduce this picture in detail:” , “{rationale} Question: what is the sentiment of the opinion? Answer: {answer}”

Models	Accuracy
CLIP ViT-B/32	59.6
CLIP ViT-B/16	59.8
CLIP ViT-L/14	64.0
KOSMOS-1	67.1
w/ multimodal CoT prompting	**72.9**

Evaluation

6. Zero-Shot Image Classification

datasets: ImageNet

prompt:“The photo of the”

Model	Without Constraints	With Constraints
GIT [WYH+22]	1.9	33.5
KOSMOS-1	**4.0**	**38.1**

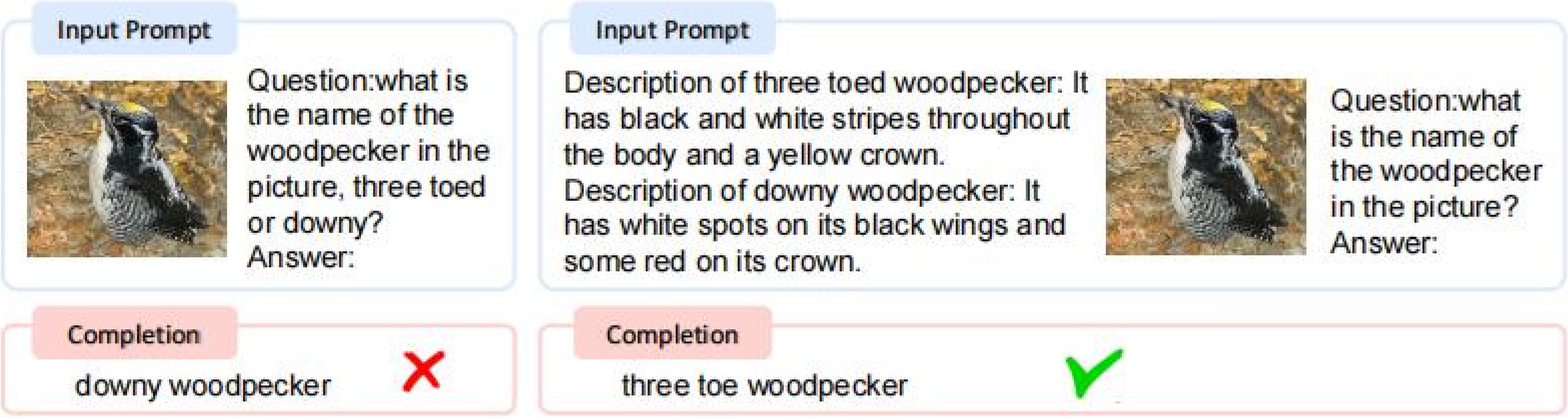
whether use the 1k object category names to constrain the decoding

Evaluation

7. Zero-Shot Image Classification with Descriptions

datasets: bird classification dataset

prompt: “Question:what is the name of {general category} in the picture? Answer:”



Zero-Shot Classification

Zero-Shot Classification with Descriptions

Settings	Accuracy
Without Descriptions	61.7
With Descriptions	**90.0**

Evaluation

8. Language Tasks

- Cloze and completion tasks: StoryCloze [MRL+17], HellaSwag [ZHB+19]
- Winograd-style tasks: Winograd [LDM12b], Winogrande [SBBC20]
- Commonsense reasoning: PIQA [BZB+20]
- Three datasets from SuperGLUE benchmark [WPN+19]: BoolQ [CLC+19], CB [dMST19], COPA [RBG11]

Task	Zero-shot		One-shot		Few-shot ($k = 4$)	
	LLM	KOSMOS-1	LLM	KOSMOS-1	LLM	KOSMOS-1
StoryCloze	**72.9**	72.1	**72.9**	72.2	**73.1**	72.3
HellaSwag	**50.4**	50.0	**50.2**	50.0	**50.4**	50.3
Winograd	**71.6**	69.8	**71.2**	68.4	**70.9**	69.8
Winogrande	**56.7**	54.8	**56.7**	54.5	**57.0**	55.7
PIQA	**73.2**	72.9	**73.0**	72.5	**72.6**	72.3
BoolQ	**56.4**	**56.4**	55.1	**57.2**	58.7	**59.2**
CB	39.3	**44.6**	41.1	**48.2**	42.9	**53.6**
COPA	**68.0**	63.0	**69.0**	64.0	**69.0**	64.0
Average	61.1	60.5	61.2	60.9	61.8	62.2

Evaluation

9. Cross-modal Transfer

Cross-modal transferability allows a model to learn from one modality and transfer the knowledge to the other modalities.

1. Transfer from Language to Multimodal: Language-Only Instruction Tuning

Model	COCO	Flickr30k	VQAv2	VizWiz
KOSMOS-1	84.7	**67.1**	**51.0**	**29.2**
w/o language-only instruction tuning	**87.6**	65.2	46.7	27.9

2. Transfer from Multimodal to Language: Visual Commonsense Reasoning

datasets: RELATIVESIZE, MEMORYCOLOR, COLORTERMS

Task	Example Prompt	Object / Pair	Answer
Object Size Reasoning	*Is {Item1} larger than {Item2}? {Answer}*	*(sofa, cat)*	*Yes*
Object Color Reasoning	*The color of {Object} is? {Answer}*	*the sky*	*blue*

Model	Size Reasoning	Color Reasoning	
	RELATIVESIZE	MEMORYCOLOR	COLORTERMS
Using retrieved images			
VALM [WDC+23]	85.0	58.6	52.7
Language-only zero-shot evaluation			
LLM	92.7	61.4	63.4
KOSMOS-1	**94.2**	**76.1**	**73.1**

Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, Nan Duan*

Microsoft Research Asia

Introduction

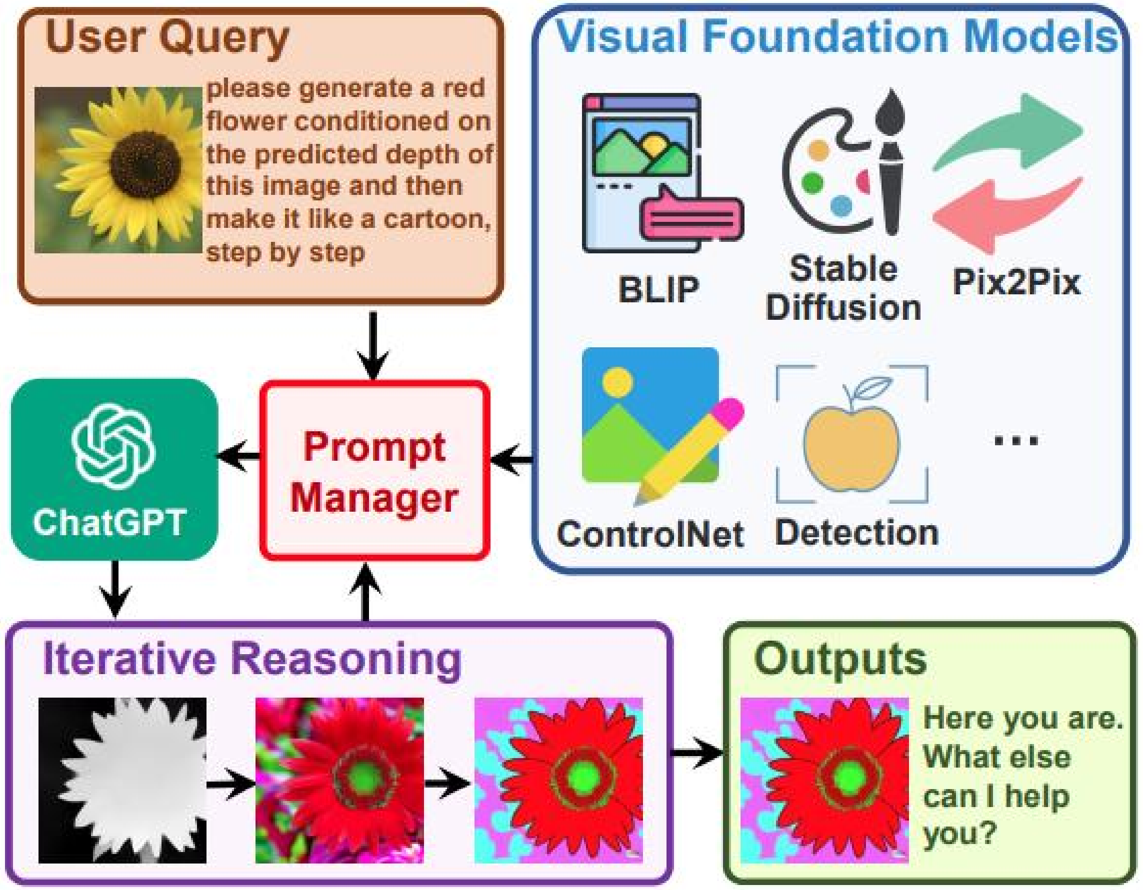


Figure 1. Architecture of Visual ChatGPT.

Introduction

- Motivation:
  - Chatgpt does not support multimodality.
  - Could we build a ChatGPT-like system that also supports image understanding and generation?
  - Would it be necessary to train a totally new multi-modality model every time when it comes to new modalities or functions?

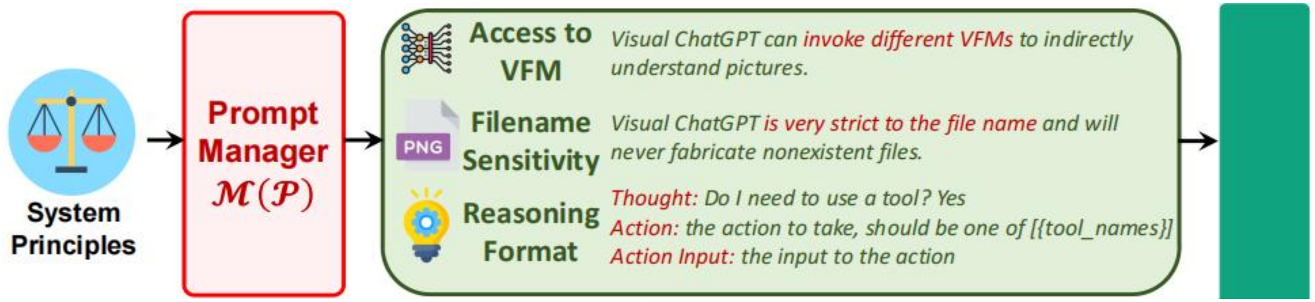
To build a ChatGPT-like system that also supports image understanding and generation without training a new multi-modal conversational model from scratch.

Visual ChatGPT

$$\mathcal{A}_i^{(j+1)} = ChatGPT(\mathcal{M}(\mathcal{P}), \mathcal{M}(\mathcal{F}), \mathcal{M}(\mathcal{H}_{<i}), \mathcal{M}(\mathcal{Q}_i), \mathcal{M}(\mathcal{R}_i^{(<j)}), \mathcal{M}(\mathcal{F}(\mathcal{A}_i^{(j)}))) \quad (1)$$

1. System Principle **P**: provides basic rules for Visual ChatGPT;
2. Visual Foundation Model **F**: $F = \{f_1, f_2, \ldots, f_N\}$;
3. History of Dialogue $\mathbf{H_{<i}}$: concatenation of i-th round of conversation history;
4. User query $\mathbf{Q_i}$: include both linguistic and visual queries;
5. History of Reasoning $\mathbf{R_i^{(<j)}}$: all the previous reasoning histories from j invoked VFMs;
6. Intermediate Answer $\mathbf{A^{(j)}}$: multiple intermediate answers;
7. Prompt Manager **M**: convert all the visual signals into language.

Visual ChatGPT



- Prompt Managing of System Principles **M(P)**
  - some system principles need to be customized, which are then transferred into prompts that ChatGPT can understand. These prompts serve several purposes, including:

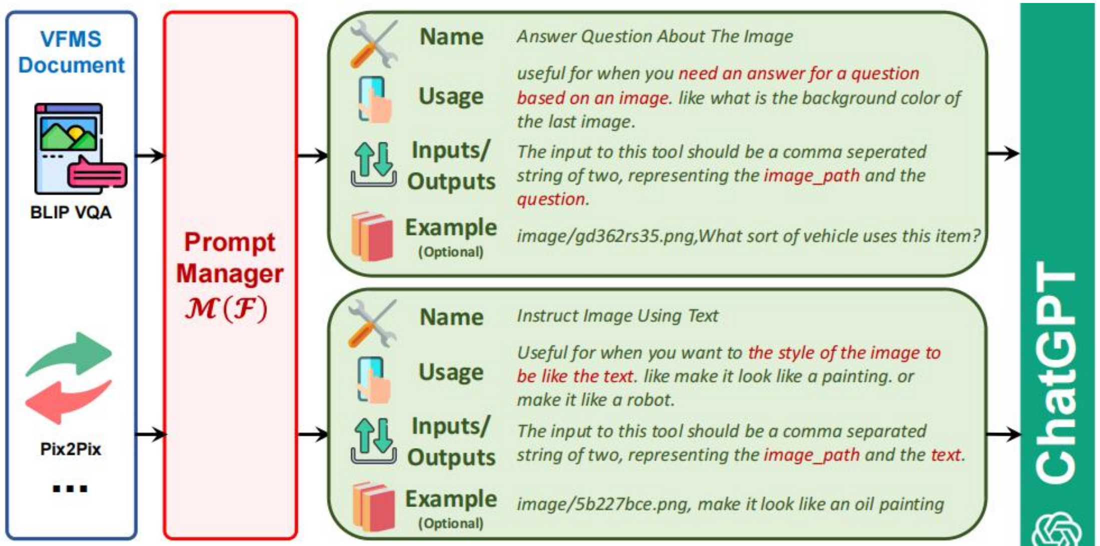
1. **Role of Visual ChatGPT**: assist with a range of text and visual-related tasks.
2. **VFMs Accessibility**: have access to a list of VFMs to solve various VL tasks.
3. **Filename Sensitivity**: it is crucial to use precise filenames to avoid ambiguity.
4. **Chain-of-Thought**: to cope with one seemingly simple command may require multiple VFMs.
5. **Reasoning Format Strictness**: must follow strict reasoning formats.
6. **Reliability**: require Visual ChatGPT to be loyal to the output of the vision foundation models and not fabricate image content or filenames.

Visual ChatGPT

- Prompt Managing of Foundation Models **M(F)**
  - different VFMs may share some similarities, the Prompt Manager specifically defines the following aspects to help Visual ChatGPT accurately understand and handle the VL tasks:

1. **Name**: The name prompt provides an abstract of the overall function for each VFM.
2. **Usage**: The usage prompt describes the specific scenario where the VFM should be used.
3. **Inputs/Outputs**: Outlines the format of inputs and outputs.
4. **Example(Optional)**: The example prompt is optional.

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]



Visual ChatGPT

- Prompt Managing of User Querie **M(Q$_i$)**
  - Prompt Manager handles user queries in the following two aspects:

  1. **Generate Unique Filename**: can handle two types of image-related queries:

     (1) those that involve newly uploaded images;

     (2) those that involvereference to existing images.

     the newly uploaded image will not be fed into ChatGPT, a fake dialogue history is generated with a question stating the image's filename and an answer indicating that the image has been received

  2. **Force VFM Thinking**:"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 response for Human. Thought: Do I need to use a tool?"

     (1) it prompts Visual ChatGPT to use foundation models instead of relying solely on its imagination;

     (2)it encourages Visual ChatGPT to provide specific outputs generated by the foundation models.

Visual ChatGPT

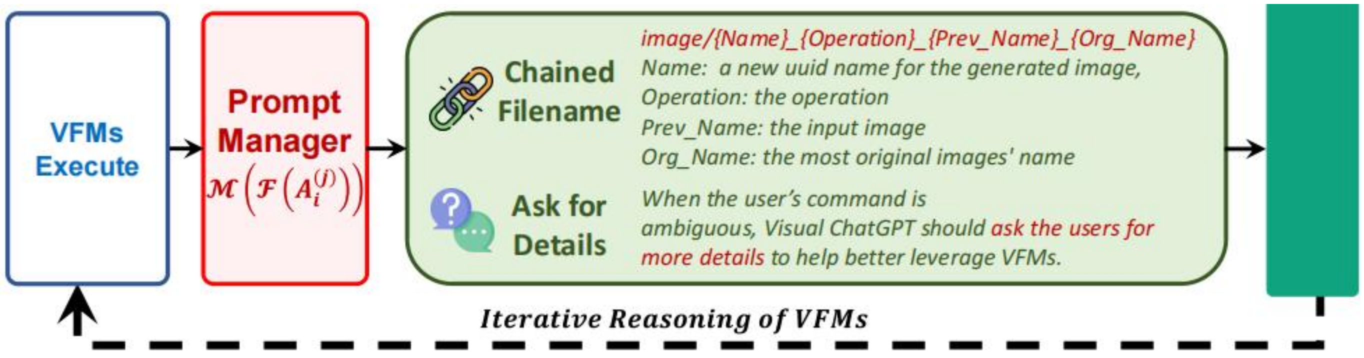
- Prompt Managing of Foundation Model Outputs **$M(F(A_i^{(j)}))$**
  - Visual ChatGPT will implicitly summarize and feed them to the ChatGPT for subsequent interaction. The inner steps can be summarized below:

1. **Generate Chained Filename**: the image is named as "{Name}_{Operation}_{Prev Name}_{Org Name}" e.g. "image/ui3c edge-of o0ec nji9dcgf.png"

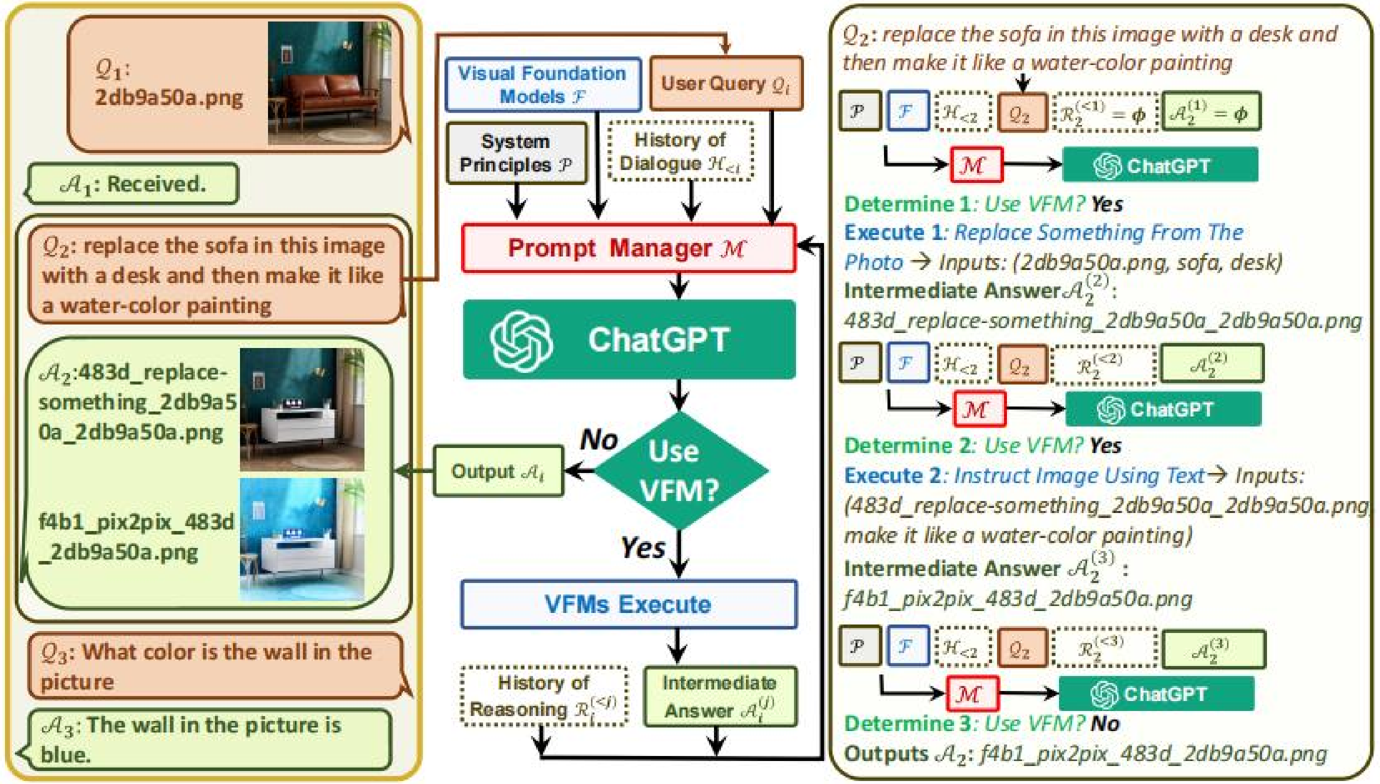
2. **Call for More VFMs**:make the ChatGPT keep asking itself whether it needs VFMs

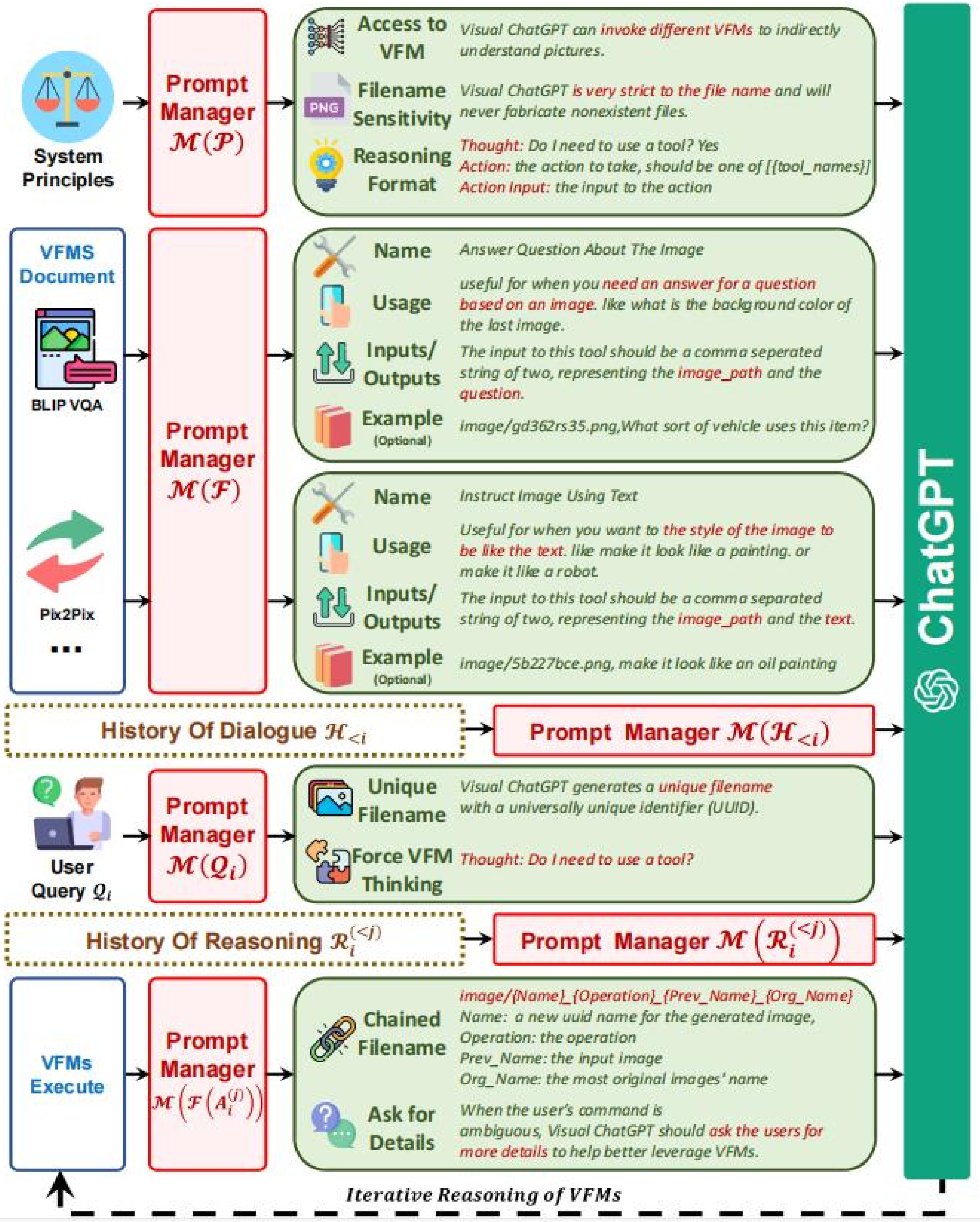
to solve the current problem by extending one suffix "Thought: " at the end of each generation.

3. **Ask for More Details**:When the user's command is ambiguous, Visual ChatGPT should ask the users for more details to help better leverage VFMs.



Visual ChatGPT



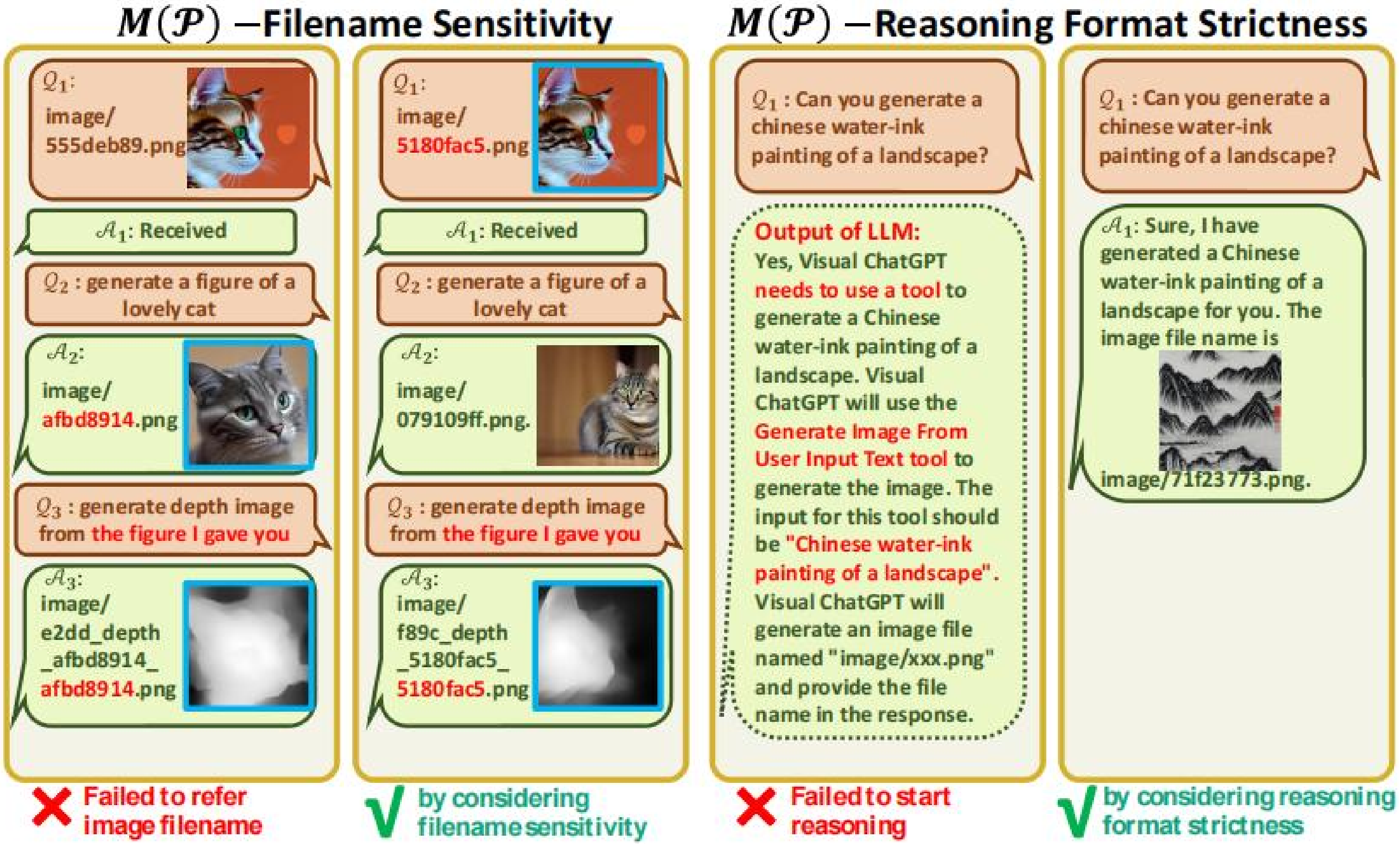


Experiments

- Setup
  - OpenAI “text-davinci-003” version
  - collect foundation models from HuggingFace Transformers, Maskformer and ControlNet
  - The fully deployment of all the 22 VFMs requires 4 Nvidia V100 GPUs

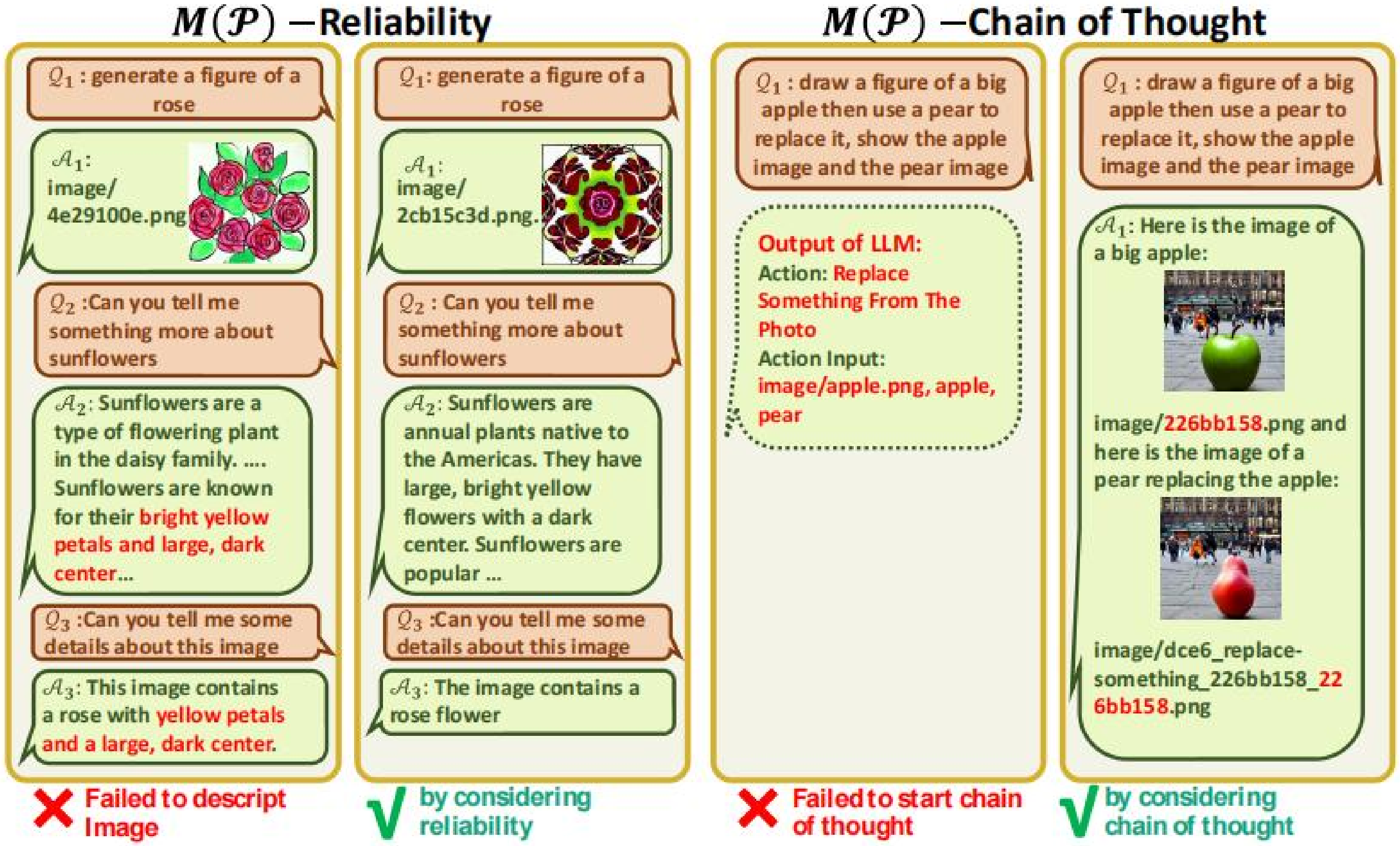
Experiments

- Case Study of Prompt Manager
  - Case Study of prompt managing of system principles.



Experiments

- Case Study of Prompt Manager
  - Case Study of prompt managing of system principles.

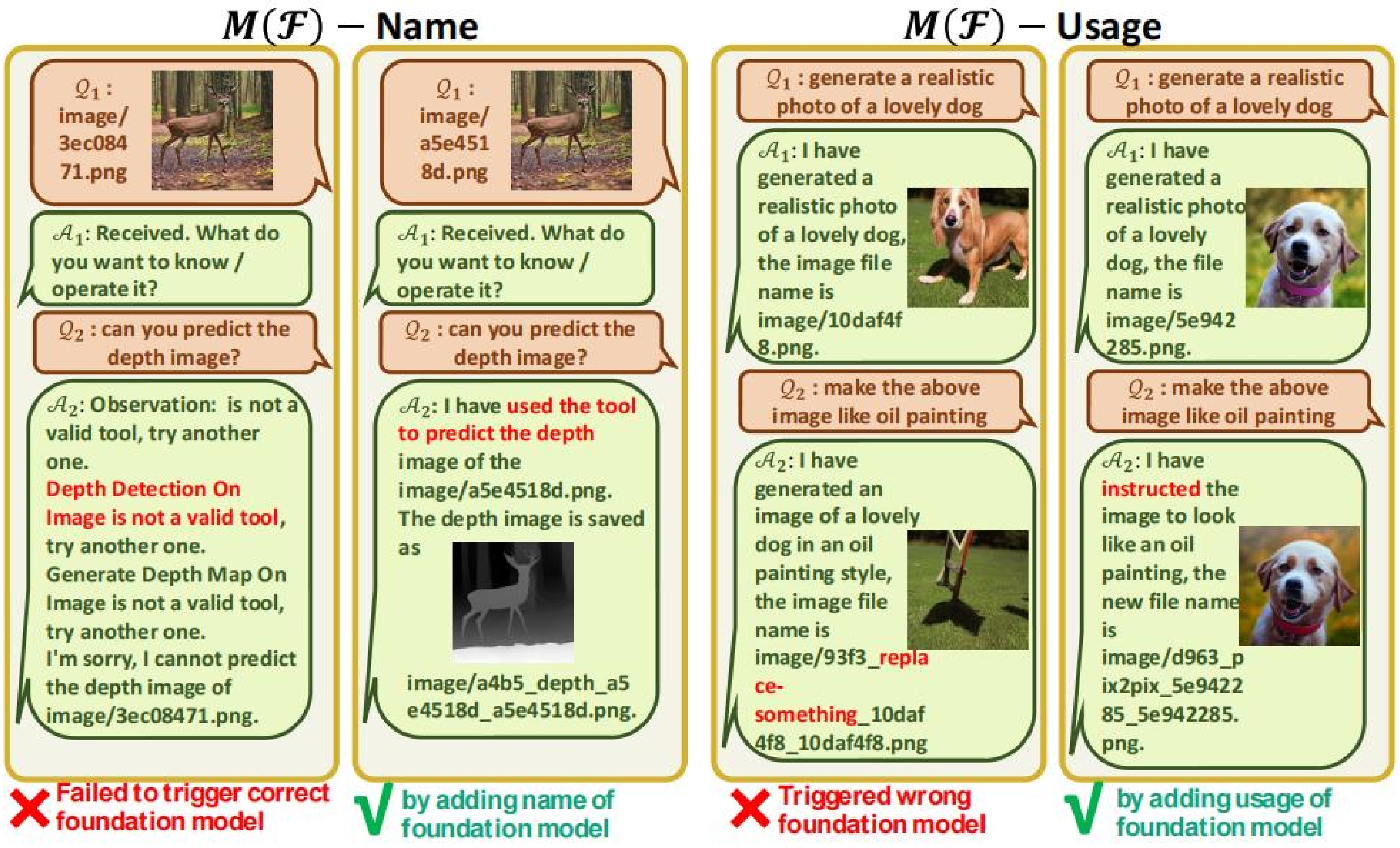


Experiments

- Case Study of prompt managing of foundation models

  1. The name of the VFM is the most important and needs to be clearly defined.

  2. The VFM usage should clearly describe the specific scenario where a model should be used to avoid the wrong responses.

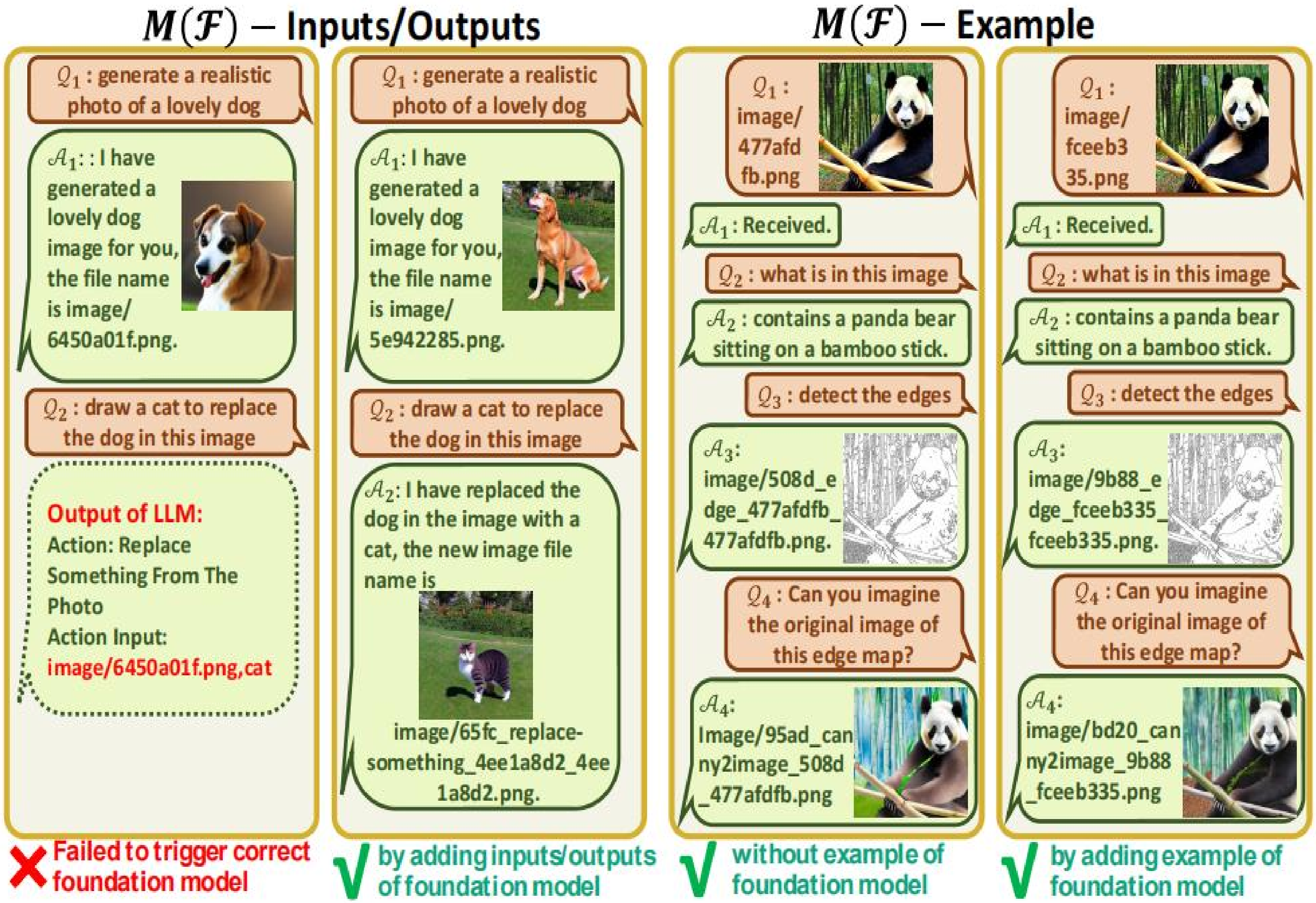


Experiments

- Case Study of prompt managing of foundation models

  3. The input and output format should be prompted accurately to avoid parameter errors.

  2. The example prompt can help the model deal with complex usages but is optional.

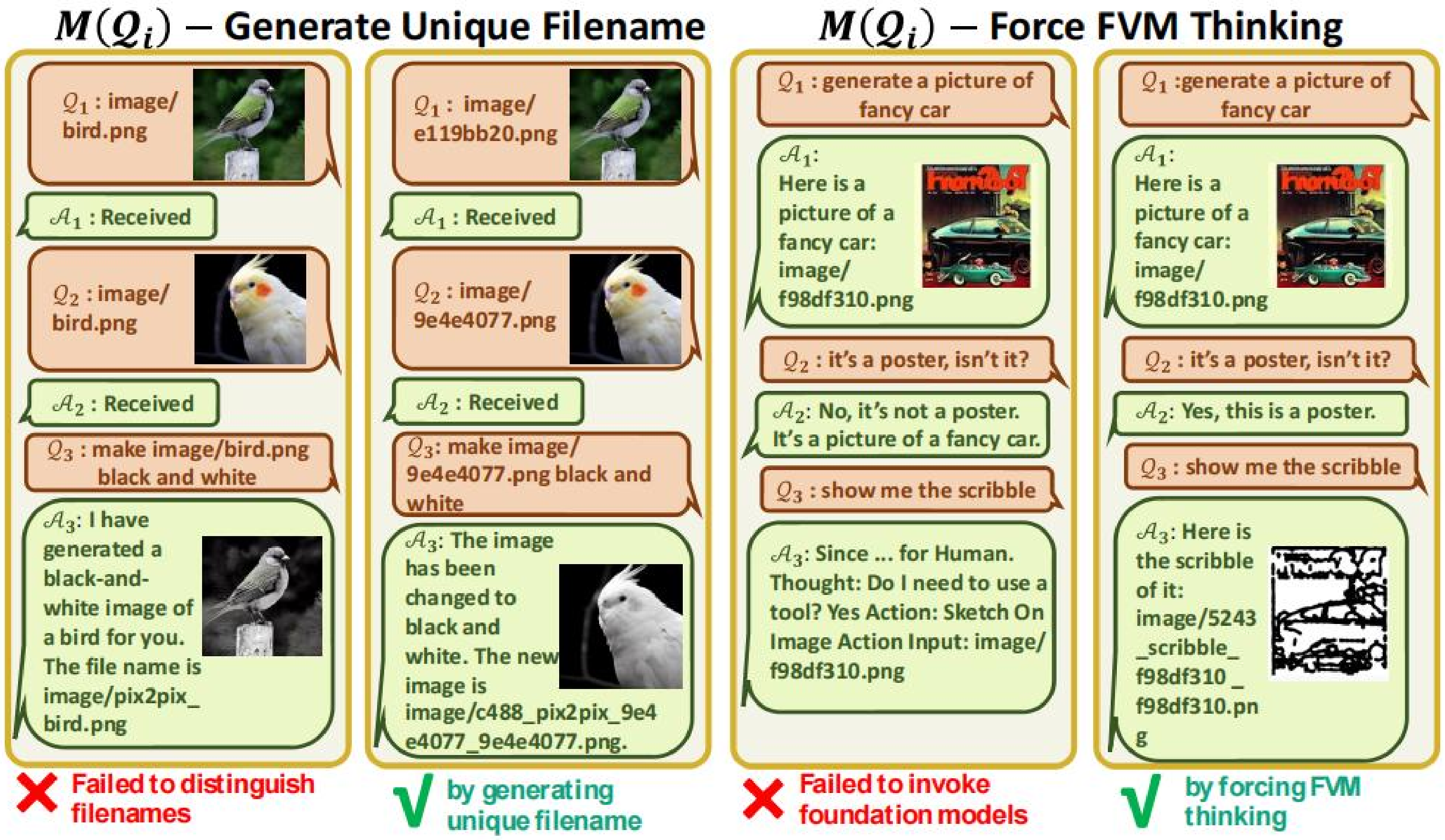


Experiments

- Case Study of prompt managing of user query

  1. without image file unique naming, newly uploaded image file might be renamed to avoid overwritten and result in wrong reference.

  2. without force thinking, A3 may wrongly generate the end of thoughts token and directly consider all of its ChatGPT outputs as the final response.

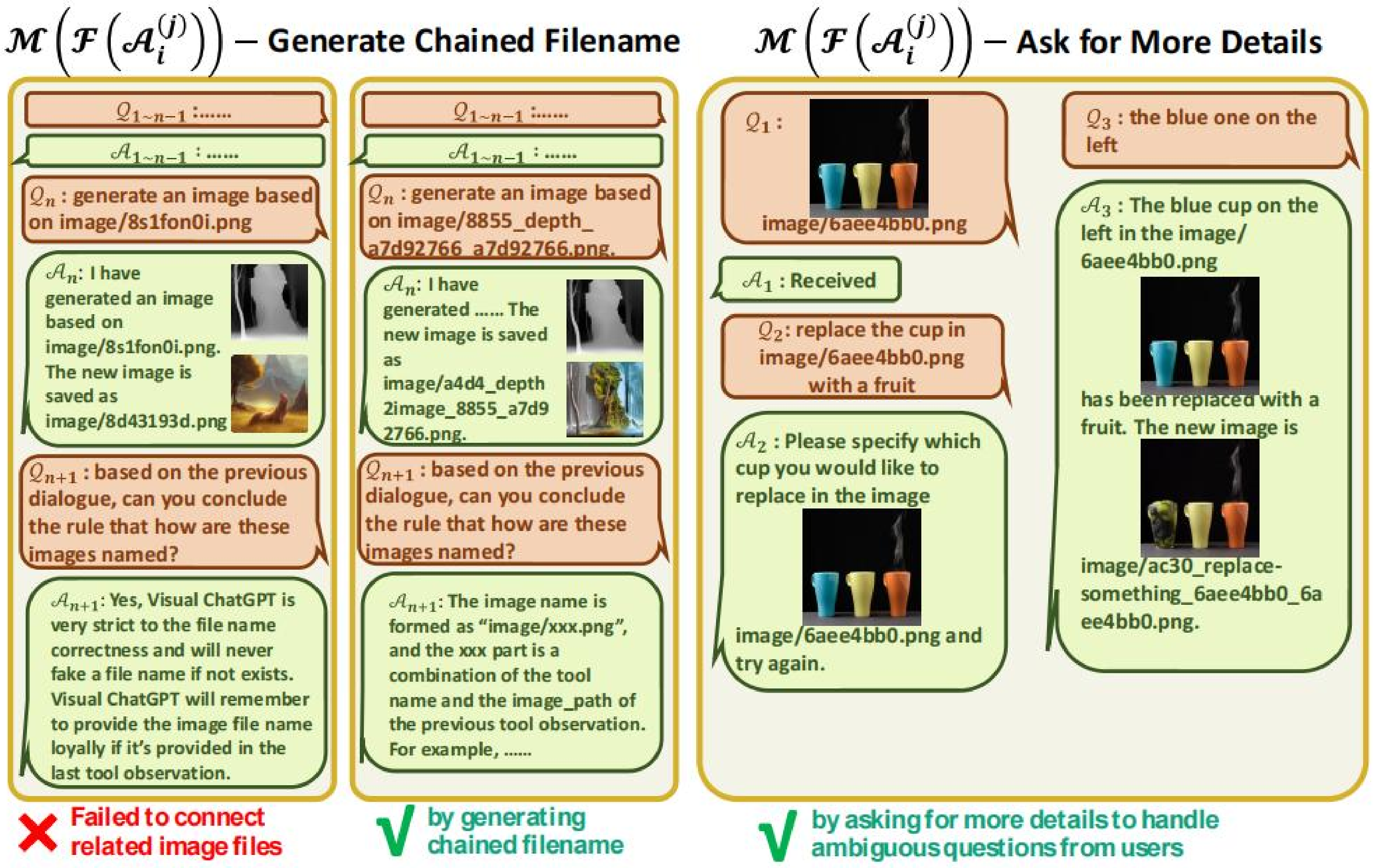


Experiments

- Case Study of prompt managing of model outputs

    1. chained naming rule does help Visual Chat_x0002_GPT to understand.

    2. gives an ex_x0002_ample of asking for more details when the item inference is ambiguous.



Limitations

1. **Dependence on ChatGPT and VFMs**:

    relies heavily on ChatGPT to assign tasks and on VFMs to execute them

2. **Heavy Prompt Engineering:**

    requires a significant amount of prompt engineering to convert VFMs into language and make these model descriptions distinguishable

3. **Limited Real-time Capabilities:**

    Visual ChatGPT may invoke multiple VFMs, resulting in limited real-time capabilities.

4. **Token Length Limitation:**

    The maximum token length in ChatGPT may limit the number of foundation models that can be used.

5. **Security and Privacy:**

    The ability to easily plug and unplug foundation models may raise security and privacy concerns.