
The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)

Zhengyuan Yang*, **Linjie Li***, **Kevin Lin***, **Jianfeng Wang***, **Chung-Ching Lin***,
Zicheng Liu, **Lijuan Wang***[♦]
 Microsoft Corporation

* Core Contributor ♦ Project Lead

Abstract

Large multimodal models (LMMs) extend large language models (LLMs) with multi-sensory skills, such as visual understanding, to achieve stronger generic intelligence. In this paper, we analyze the latest model, GPT-4V(ision)[†], to deepen the understanding of LMMs. The analysis focuses on the intriguing tasks that GPT-4V can perform, containing test samples to probe the quality and genericity of GPT-4V’s capabilities, its supported inputs and working modes, and the effective ways to prompt the model. In our approach to exploring GPT-4V, we curate and organize a collection of carefully designed qualitative samples spanning a variety of domains and tasks. Observations from these samples demonstrate that GPT-4V’s unprecedented ability in processing arbitrarily interleaved multimodal inputs and the genericity of its capabilities together make GPT-4V a powerful multimodal generalist system. Furthermore, GPT-4V’s unique capability of understanding visual markers drawn on input images can give rise to new human-computer interaction methods such as visual referring prompting. We conclude the report with in-depth discussions on the emerging application scenarios and the future research directions for GPT-4V-based systems. We hope that this preliminary exploration will inspire future research on the next-generation multimodal task formulation, new ways to exploit and enhance LMMs to solve real-world problems, and gaining better understanding of multimodal foundation models.

Contents

| | |
|--|----------|
| List of Figures | 4 |
| 1 Introduction | 8 |
| 1.1 Motivation and Overview | 8 |
| 1.2 Our Approach in Exploring GPT-4V | 8 |
| 1.3 How to Read this Report? | 9 |

[†]This report explores GPT-4V(ision) with the vision capability and refers to the model as “GPT-4V,” following the OpenAI reports [95, 94]. We refer to the text-only version of the model as “GPT-4 (no vision)” [94].

| | |
|--|-----------|
| 2 GPT-4V’s Input Modes | 10 |
| 2.1 Text-only Inputs | 10 |
| 2.2 Single Image-text Pair | 10 |
| 2.3 Interleaved Image-text Inputs | 11 |
| 3 GPT-4V’s Working Modes and Prompting Techniques | 12 |
| 3.1 Following Text Instructions | 15 |
| 3.2 Visual Pointing and Visual Referring Prompting | 15 |
| 3.3 Visual + Text Prompting | 16 |
| 3.4 In-context Few-shot Learning | 19 |
| 4 Vision-Language Capability | 26 |
| 4.1 Image Description on Diverse Domains | 26 |
| 4.2 Object Localization, Counting, and Dense Captioning | 38 |
| 4.3 Multimodal Knowledge and Commonsense | 43 |
| 4.4 Scene Text, Table, Chart, and Document Reasoning | 49 |
| 4.5 Multilingual Multimodal Understanding | 58 |
| 4.6 Coding Capability with Vision | 63 |
| 5 Interaction with Humans: Visual Referring Prompting | 66 |
| 5.1 Understand Pointing Inputs | 66 |
| 5.2 Visual Referring Prompting | 69 |
| 5.3 Generate Pointing Outputs | 69 |
| 6 Temporal and Video Understanding | 73 |
| 6.1 Multi-image Sequencing | 73 |
| 6.2 Video Understanding | 73 |
| 6.3 Visual Referring Prompting for Grounded Temporal Understanding | 79 |
| 7 Abstract Visual Reasoning and Intelligence Quotient Test | 80 |
| 7.1 Abstract Visual Stimuli | 80 |
| 7.2 Discovery and Association of Parts and Objects | 80 |
| 7.3 Wechsler Adult Intelligence Scale | 80 |
| 7.4 Raven’s Progressive Matrices | 80 |
| 8 Emotional Quotient Test | 86 |
| 8.1 Read Emotion from Facial Expressions | 86 |
| 8.2 Understand How Visual Content Arouses Emotions | 87 |
| 8.3 Emotion Conditioned Output | 89 |
| 9 Emerging Application Highlights | 90 |
| 9.1 Spot the Difference | 90 |

| | | |
|-----------|------------------------------------|------------|
| 9.2 | Industry | 90 |
| 9.3 | Medical | 99 |
| 9.4 | Auto Insurance | 104 |
| 9.5 | Customized Captioner | 107 |
| 9.6 | Image Generation | 111 |
| 9.7 | Embodied Agent | 116 |
| 9.8 | GUI Navigation | 121 |
| 10 | LMM Powered Agents | 150 |
| 10.1 | Multimodal Plugins | 152 |
| 10.2 | Multimodal Chains | 152 |
| 10.3 | Self-Reflection | 152 |
| 10.4 | Self-Consistency | 155 |
| 10.5 | Retrieval-Augmented LMMs | 156 |
| 11 | Conclusions | 156 |
| 11.1 | Summary and Conclusions | 156 |
| 11.2 | Towards Future LMMs | 156 |

List of Figures

| | | |
|----|---|----|
| 1 | Section 2.3: multi-image inputs and interleaved image-text pairs. | 11 |
| 2 | Section 3.1: following text instructions. | 12 |
| 3 | Section 3.1: constrained prompting. | 13 |
| 4 | Section 3.1: condition on good performance. | 14 |
| 5 | Section 3.2: different modes of visual pointing. | 16 |
| 6 | Section 3.2: visual pointing and visual referring prompting. | 17 |
| 7 | Section 3.3: multimodal example-grounded instruction. | 18 |
| 8 | Section 3.4: few-shot comparison example 1, Part I: zero-shot. | 20 |
| 9 | Section 3.4: few-shot comparison example 1, Part II: one-shot. | 21 |
| 10 | Section 3.4: few-shot comparison example 1, Part III: two-shot. | 22 |
| 11 | Section 3.4: few-shot comparison example 2, Part I: zero-shot. | 23 |
| 12 | Section 3.4: few-shot comparison example 2, Part II: one-shot. | 24 |
| 13 | Section 3.4: few-shot comparison example 2, Part III: two-shot. | 25 |
| 14 | Section 4.1: celebrity recognition and description. | 28 |
| 15 | Section 4.1: landmark recognition and description. | 29 |
| 16 | Section 4.1: landmark recognition and description. | 30 |
| 17 | Section 4.1: food recognition and description. | 31 |
| 18 | Section 4.1: medical image understanding. | 32 |
| 19 | Section 4.1: medical image understanding. | 33 |
| 20 | Section 4.1: logo recognition and description. | 34 |
| 21 | Section 4.1: in-the-wild logo recognition and description. | 35 |
| 22 | Section 4.1: scene understanding. | 36 |
| 23 | Section 4.1: counterfactual examples. | 37 |
| 24 | Section 4.2: spatial relationship understanding. | 39 |
| 25 | Section 4.2: object counting. | 40 |
| 26 | Section 4.2: object localization. | 41 |
| 27 | Section 4.2: dense captioning. | 42 |
| 28 | Section 4.3: joke and meme understanding. | 44 |
| 29 | Section 4.3: science and encyclopedia knowledge. | 45 |
| 30 | Section 4.3: science and encyclopedia knowledge. | 46 |
| 31 | Section 4.3: science and encyclopedia knowledge. | 47 |
| 32 | Section 4.3: multimodal commonsense. | 48 |
| 33 | Section 4.4: scene text recognition. | 50 |
| 34 | Section 4.4: visual math reasoning. | 51 |
| 35 | Section 4.4: flow chart understanding. | 52 |
| 36 | Section 4.4: flow chart understanding. | 53 |
| 37 | Section 4.4: chart understanding and reasoning. | 54 |
| 38 | Section 4.4: table understanding and reasoning. | 55 |

| | | |
|----|---|-----|
| 39 | Section 4.4: document understanding | 56 |
| 40 | Section 4.4: long document understanding. | 57 |
| 41 | Section 4.5: multilingual image descriptions. | 59 |
| 42 | Section 4.5: multilingual scene text recognition. | 60 |
| 43 | Section 4.5: multilingual text recognition, translation, and description reasoning. . | 61 |
| 44 | Section 4.5: multilingual multicuture understanding. | 62 |
| 45 | Section 4.6: generate LaTex codes based on the hand-written input. | 63 |
| 46 | Section 4.6: reconstruct table in Markdown or LaTex code. | 64 |
| 47 | Section 4.6: write codes to replicate the input figure. | 65 |
| 48 | Section 5.1: understand pointing inputs for grounded description. | 67 |
| 49 | Section 5.1: grounded description with text-format region coordinates. | 68 |
| 50 | Section 5.2: examples of visual referring prompting. | 70 |
| 51 | Section 5.2: examples of visual referring prompting. | 71 |
| 52 | Section 5.3: generate pointing outputs. | 72 |
| 53 | Section 6.1: video frames understanding. | 74 |
| 54 | Section 6.2: long-term temporal order reasoning. | 75 |
| 55 | Section 6.2: short-term temporal order reasoning. | 76 |
| 56 | Section 6.2: short-term and long-term temporal anticipation. | 77 |
| 57 | Section 6.2: temporal localization and reasoning. | 78 |
| 58 | Section 6.3: grounded temporal understanding. | 79 |
| 59 | Section 7.1: abstract visual stimuli. | 81 |
| 60 | Section 7.2: part-object association. | 82 |
| 61 | Section 7.3: WAIS IQ test. | 83 |
| 62 | Section 7.4: RPM IQ test, entire question page. | 84 |
| 63 | Section 7.4: RPM IQ test, processed sub-figures. | 85 |
| 64 | Section 8.1: read emotions from facial expressions. | 86 |
| 65 | Section 8.2: how different visual contents may arouse emotions. | 87 |
| 66 | Section 8.2: image aesthetics. | 88 |
| 67 | Section 8.3: emotion conditioned outputs. | 89 |
| 68 | Section 9.1: spot the differences. | 92 |
| 69 | Section 9.1: spot the differences. | 93 |
| 70 | Section 9.2: industrial defect detection. | 94 |
| 71 | Section 9.2: industrial defect detection. | 95 |
| 72 | Section 9.2: industrial defect detection. | 96 |
| 73 | Section 9.2: safety inspection. | 97 |
| 74 | Section 9.2: grocery checkout. | 98 |
| 75 | Section 9.3: radiology report generation on medical images. | 100 |
| 76 | Section 9.3: radiology report generation on medical images. | 101 |
| 77 | Section 9.3: radiology report generation on medical images. | 102 |

| | | |
|-----|--|-----|
| 78 | Section 9.3: radiology report generation with diagnosis history. | 103 |
| 79 | Section 9.4: auto damage evaluation. | 105 |
| 80 | Section 9.4: insurance reporting. | 106 |
| 81 | Section 9.5: customized captioner for photo organization. | 108 |
| 82 | Section 9.5: customized captioner for photo organization. | 109 |
| 83 | Section 9.5: dense captioning with segmentation cut-outs. | 110 |
| 84 | Section 9.6: evaluation of images generated from texts. | 112 |
| 85 | Section 9.6: evaluation of images generated from texts. | 113 |
| 86 | Section 9.6: image editing. | 114 |
| 87 | Section 9.6: image editing. | 115 |
| 88 | Section 9.7: embodied agent using coffee machine. | 117 |
| 89 | Section 9.7: embodied agent using coffee machine. | 118 |
| 90 | Section 9.7: embodied agent indoor navigation. | 119 |
| 91 | Section 9.7: embodied agent indoor navigation. | 120 |
| 92 | Section 9.8: web browsing for Mapo Tofu recipe. | 122 |
| 93 | Section 9.8: web browsing for Mapo Tofu recipe. | 123 |
| 94 | Section 9.8: web browsing for Mapo Tofu recipe. | 124 |
| 95 | Section 9.8: web browsing for Mapo Tofu recipe. | 125 |
| 96 | Section 9.8: web browsing for Mapo Tofu recipe. | 126 |
| 97 | Section 9.8: web browsing for today's news. | 127 |
| 98 | Section 9.8: web browsing for today's news. | 128 |
| 99 | Section 9.8: web browsing for today's news. | 129 |
| 100 | Section 9.8: web browsing for today's news. | 130 |
| 101 | Section 9.8: web browsing for today's news. | 131 |
| 102 | Section 9.8: web browsing for today's news. | 132 |
| 103 | Section 9.8: online shopping. | 133 |
| 104 | Section 9.8: online shopping. | 134 |
| 105 | Section 9.8: online shopping. | 135 |
| 106 | Section 9.8: online shopping. | 136 |
| 107 | Section 9.8: online shopping. | 137 |
| 108 | Section 9.8: online shopping. | 138 |
| 109 | Section 9.8: online shopping. | 139 |
| 110 | Section 9.8: online shopping. | 140 |
| 111 | Section 9.8: online shopping. | 141 |
| 112 | Section 9.8: notification understanding. | 142 |
| 113 | Section 9.8: notification understanding. | 143 |
| 114 | Section 9.8: notification understanding. | 144 |
| 115 | Section 9.8: watching videos. | 145 |
| 116 | Section 9.8: watching videos. | 146 |

| | | |
|-----|-----------------------------------|-----|
| 117 | Section 9.8: watching videos. | 147 |
| 118 | Section 9.8: watching videos. | 148 |
| 119 | Section 9.8: watching videos. | 149 |
| 120 | Section 10.1: multimodal plugins. | 150 |
| 121 | Section 10.2: multimodal chains. | 151 |
| 122 | Section 10.3: self-reflection. | 153 |
| 123 | Section 10.3: self-reflection. | 154 |
| 124 | Section 10.4: self-consistency. | 155 |

1 Introduction

1.1 Motivation and Overview

The breakthroughs in large language models (LLMs) [22, 94, 27, 10, 116, 50] have shown remarkable versatilities and capabilities across various domains and tasks. The next evolution in this field, large multimodal models (LMMs), aims to expand upon the capabilities of LLMs by integrating multi-sensory skills to achieve even stronger general intelligence. Given the dominance of the visual in human senses [30, 55], many LMM studies start with extending the vision capability. Preliminary research investigations either finetune a vision encoder to align with a frozen pre-trained LLM [118, 6, 67, 52, 39, 12, 45, 150, 75, 32, 139], or use a vision-language model to convert visual inputs to text descriptions that LMMs can understand [142, 134, 124, 51, 106, 135]. However, most existing models [12, 45, 150, 75, 32, 65] are of limited model and data scales, potentially restricting the emergence of various intriguing abilities. Consequently, it remains unclear what are the status quo and emergent multimodal abilities of LMMs that are developed based on the state-of-the-art LLMs, such as GPT-4 (no vision) [94] and PaLM [27, 10]. In this paper, we report our preliminary explorations with (an early version of) GPT-4V, a state-of-the-art LMM with vision, built based on the SOTA LLM and trained with a large scale of multimodal data.

Our exploration of GPT-4V is guided by the following questions.

1. *What are GPT-4V’s supported inputs and working modes?* The genericity of multimodal models inevitably requires the system to work with the arbitrary mix of different input modalities. GPT-4V shows unprecedented ability in understanding and processing an arbitrary mix of input images, sub-images, texts, scene texts, and visual pointers. We also demonstrate that GPT-4V well supports the test-time techniques observed in LLMs, including instruction following [96], chain-of-thoughts [129, 63], in-context few-shot learning [22], etc.
2. *What are the quality and genericity of GPT-4V’s capabilities on different domains and tasks?* We sample queries covering a wide range of domains and tasks to understand GPT-4V’s capabilities, including open-world visual understanding, visual description, multimodal knowledge, commonsense, scene text understanding, document reasoning, coding, temporal reasoning, abstract reasoning, emotion understanding, and many more. GPT-4V shows impressive human-level capabilities across many of the experimented domains.
3. *What are effective ways to use and prompt GPT-4V?* GPT-4V is strong in understanding pixel space edits, such as visual pointers and scene texts drawn on input images. Inspired by this capability, we discuss the “visual referring prompting” that directly edits input images to instruct the task of interest. Visual referring prompting can be seamlessly used together with other image and text prompts, presenting a nuanced interface for instruction and example demonstrations.
4. *What are promising future directions?* Given GPT-4V’s strong capability across domains and tasks, we ask what is the next step for multimodal learning, and more broadly for artificial intelligence. We organize our thoughts and explorations into two perspectives, *i.e.*, emergent novel application scenarios to focus on, and the future research directions for GPT-4V-based systems. We present our preliminary explorations to inspire future studies.

Guided by the aforementioned problems, we comprehensively organize and list our explored qualitative results. The report contains minimal quantitative benchmark results, and instead consists of mainly selected interesting qualitative examples. Despite being less rigorous, this design allows for providing a more comprehensive analysis covering a broad range of domains, tasks, working modes, and prompting techniques, under a fixed capacity. We believe this organized collection of explorations will inspire future works in emerging novel applications, next-generation multimodal task formulation, and developing advanced LMM-based intelligent systems.

1.2 Our Approach in Exploring GPT-4V

Goal of this report. The standard approach for evaluating a system is by benchmarking it against a series of carefully designed datasets, each representing a specific domain and task. One challenge is that some of the existing benchmarks may not be suitable for evaluating LMMs anymore. For example, the image captioning outputs of LMMs are much richer and contain more detailed descriptions than

the ground truths in the image captioning benchmark datasets [24]. There is also a lack of public information regarding GPT-4V’s large-scale pre-training, which may violate the train-test setup for certain existing datasets and invalidate those benchmark numbers. Because of this, restricting the evaluation to *existing* benchmarks and metrics may unintentionally narrow the scope of GPT-4V’s assessment. Developing a comprehensive list of next-generation evaluation tasks and benchmarks would be the ideal ultimate solution. However, we left those as future work due to the significant efforts required.

In lieu of quantitative benchmarking, this paper focuses on using qualitative results to provide a glimpse of GPT-4V’s new capabilities and potential emerging use cases. Our goal is to discover and preview what GPT-4V might already be capable of, even though these novel capabilities may not yet be entirely reliable. We hope this collection of explorations will inspire future research in establishing quantitative benchmarks for next-generation multimodal tasks, modernizing existing benchmarks, further improving model performance and system reliability, and inspiring innovation in emerging use cases. Following this, we will delve into the core designs for our approach to exploring GPT-4V.

Sample selection guidance. This report focuses on presenting qualitative results to showcase the potential capabilities of GPT-4V, rather than providing comprehensive quantitative benchmark results. This naturally raises the question of the reliability of the showcased examples. The examples featured in this report may require careful instruction tuning to amplify GPT-4V’s corresponding capabilities. It should be noted that some complex cases may only work with the specifically designed prompts. As such, the capabilities demonstrated may not consistently work across different samples. Instead of showing only the reliable functionalities, the primary objective of this report is to provide readers with a list of our discovered potential capabilities of GPT-4V, which might otherwise be overlooked after a few unsuccessful trials.

Sample selection to prevent mere memorizing from training. A fundamental design consideration in qualitative reports [23] is discerning models’ true capabilities from merely memorizing responses from training samples or making educated guesses based on hints from instructions and in-context examples. We carefully control both the images and text in the input prompts to prevent them from being seen during GPT-4V training. We generate original text queries from scratch, and try to use images that are either not accessible online or with a timestamp beyond April 2023. We will indicate instances where a specific sample does not meet this criterion, *e.g.*, deliberately using samples from specific vision-language datasets. Beyond ensuring that samples are unseen, we incorporate rationale queries into the process. These queries are designed to probe the model’s reasoning process, thereby validating GPT-4V’s possession of the intended capability.

The default working mode. As later detailed in Section 3, GPT-4V works effectively in different working modes, including zero-shot learning with instructions, in-context few-shot learning, *etc.* Among them, this report primarily focuses on zero-shot instruction tuning, as opposed to in-context few-shot learning. This design is to prevent potential information leakage from in-context examples. While in-context few-shot examples can enhance performance and reliability, they do not consistently engender new capabilities. As such, we designate zero-shot as the default working mode for presentation, and reduce the use of in-context examples to minimize examples’ impacts on the assessed capabilities.

1.3 How to Read this Report?

This report documents the explorations of GPT-4V conducted by researchers in the computer vision and vision-language multimodal field. It is primarily geared towards fellow researchers in related disciplines who seek to gain a qualitative impression of LMM’s capabilities and understand its difference from traditional vision-language models. The report is also prepared for professionals for whom AI or computer science may be outside their specialties, to assist them in conceptualizing ways LMMs can enhance their proficiency within their distinct domains of expertise.

We give an overview of the report, structured around the four core questions that guide our exploration.

1. *What are GPT-4V’s supported inputs and working modes?* Section 2 summarizes GPT-4V’s supported inputs and presents an overview of their corresponding use cases. Based on the flexible interleaved image-text inputs, Section 3 discusses GPT-4V’s different working modes, such as instruction tuning, in-context learning, and other emergent usages.

The section covers the novel ways of using and prompting GPT-4V, aiming to provide a comprehensive overview of how we will use GPT-4V in subsequent sections.

2. *What are the quality and genericity of GPT-4V’s capabilities on different domains and tasks?* The exploration of this question makes up a large portion of the report. Section 4 provides a comprehensive analysis covering a wide range of vision and vision-language scenarios, including image description and recognition on different domains, dense visual understanding, multimodal knowledge, commonsense, scene text understanding, document reasoning, and many more. We also separate out several novel and interesting capabilities. Section 6 studies GPT-4V’s capability in temporal, motion, and video understanding. Section 7 explores the abstract visual understanding and reasoning capability, and Section 8 covers the emotion and sentiment understanding.
3. *What are effective ways to use and prompt GPT-4V?* We start the discussion on this question from the working mode and prompting method introduction in Section 3. In Section 5, we highlight one novel promoting technique, namely visual referring prompting, which draws visual pointers and scene texts on input images to prompt GPT-4V. We demonstrate the flexible prompting methods, such as the combination of instruction and example demonstrations, throughout the report in the given examples.
4. *What are promising future directions?* Section 9 focuses on the novel use cases facilitated by GPT-4V. We hope these initial examples could inspire future works to design new task setups and present rigorous benchmarks. Section 10 imagines powerful future systems that can be built based on GPT-4V, such as the multimodal plugins, multimodal chains, self-reflection, self-consistency, and retrieval-augmented LMMs, etc.

In addition to this overview and the **table of contents**, we have also included a **list of figures**. The list enumerates the qualitative examples detailed within the report, serving as an additional tool to help readers navigate to their scenarios of interest.

2 GPT-4V’s Input Modes

This section summarizes GPT-4V’s supported inputs, *i.e.*, functioning as a uni-model language model with the text-only inputs, taking single image-text pair optionally with only a single image, and taking interleaved image-text pairs optionally with only multiple image inputs. We next highlight the representative use cases under these different input modes.

2.1 Text-only Inputs

GPT-4V’s strong language capability enables it to serve as an effective unimodal language model [35, 101, 22] with text-only inputs. Operating exclusively with text for both input and output, GPT-4V is capable of performing a wide variety of language and coding tasks. We refer readers to the GPT-4 technical report [94] for the comprehensive and in-depth analysis of GPT-4V’s language and coding capabilities, as well as the comparison with GPT-4 (no vision).

2.2 Single Image-text Pair

GPT-4V, the latest large multimodal model, takes images and texts as inputs to generate textual outputs. In line with existing general-purpose vision-language models [8, 77, 69, 7, 66, 115, 113, 148, 25, 78, 42, 70, 54, 61, 68, 125, 26, 133, 38, 6, 121, 43, 37, 151, 65], GPT-4V can take a single image-text pair or a single image as input to perform various vision and vision-language tasks, such as image recognition [34], object localization [146], image captioning [24], visual question answering [11], visual dialogue [33], dense caption [59], and so on. We note that the text in the image-text pair can be used either as instruction like “describe the image” for captioning, or as the query input like the question in visual question answering. GPT-4V’s exceptional intelligence is exemplified by its significantly enhanced performance and generalizability compared to prior arts. A comprehensive analysis of its multimodal capabilities on various domains is detailed in Section 4.

Sec. 2.3 Interleaved Image-text Inputs

Interleaved Image-text Pairs

Prompt:

How much did I pay for tax? **Prompt Continued**
Where is the information?



GPT-4V:

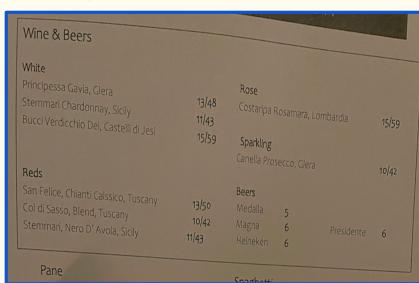
In the first receipt, you paid \$3.72 for tax. The information is located at the bottom of the receipt. In the second receipt, you paid \$42.23 for tax. The information is located near the bottom of the receipt, under the "Sales Tax" line. In the third receipt, you paid \$4.58 for tax. The information is located at the bottom of the receipt, under the "TAX" line. In total, you paid \$50.53 in tax across all three receipts.

Prompt:

How much should I pay for the beer on the table according to the price on the menu?



Here is a photo of the menu.



GPT-4V:

According to the menu, the price for a Magna beer is \$6. Since there are two Magna beers on the table, the total cost for the beers would be \$12.

Figure 1: GPT-4V can work with multi-image and interleaved image-text inputs. Check Section 2.3 for detailed discussions.

2.3 Interleaved Image-text Inputs

The generality of GPT-4V is further enhanced by its ability to handle flexibly interleaved image-text inputs. The interleaved image-text inputs can be either visually centric such as multiple images with a short question or instruction, text-centric such as a long webpage with two inserted images, or a balanced mixture of images and texts. This mode of mixed input provides flexibility for a wide array of applications. For example, it can compute the total tax paid across multiple receipt images, as shown in Figure 1. It also enables processing multiple input images and extracting queried information. GPT-4V could also effectively associate information across interleaved image-text inputs, such as finding the beer price on the menu, counting the number of beers, and returning the total cost, as shown in Figure 1. Beyond direct applications, processing interleaved image-text inputs serves as a fundamental component for in-context few-shot learning and other advanced test-time prompting techniques, thereby further boosting GPT-4V's generality. We demonstrate these intriguing novel usages in the next section, Section 3.

3 GPT-4V's Working Modes and Prompting Techniques

Sec. 3.1 Following Text Instructions

Following Text Instructions

Prompt:
Describe the image with a sentence consisting of **three/six/nine** words.



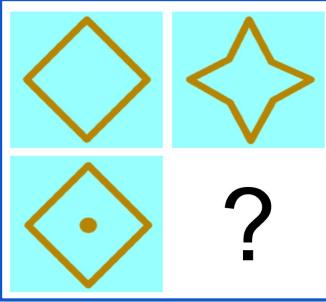
GPT-4V:
Three: Seaside dining table
Six: Dining table overlooking ocean with beer
Nine: Outdoor restaurant table with ocean view and a beer

Prompt:
Write a sentence for the image consisting only of words starting with the letter '**B**'/ '**D**'/ '**T**'.



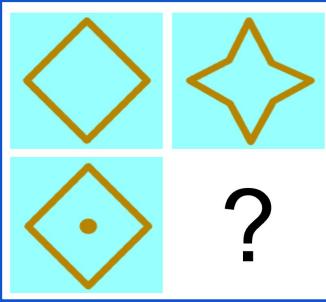
GPT-4V:
'B': Beachside bar boasts beautiful blue backdrop.
'D': Dining deck displays delightful drinks, distant docks.
'T': Tabletop terrace teeming with tropical tranquility.

Prompt:
What would the missing image look like?



GPT-4V:
The missing image would likely be a variation of the diamond or star shape, as seen in the other three images. It could be a diamond with a different pattern or color, or a star with a different number of points or color. It is difficult to say for sure without more context, but it would likely follow the same theme as the other images.

Prompt:
Here are three sub images arranged in a 2-by-2 matrix. First, look at the two images in the first column, in the top left and bottom left. Then, find out the pattern in the first column based on the left two images. Next, use the found pattern and the image in the top right to infer the missing figure. Finally, describe what would the missing image look like?



GPT-4V:
The pattern in the first column is that the shape in the bottom left image has a dot in the center of the shape in the top left image. Therefore, the missing image in the bottom right would be a star with a dot in the center, similar to the image in the top right but with a dot in the center.

Figure 2: GPT-4V can understand and follow text instructions, to generate the desired text outputs or learn to perform a new task. Red highlights the less informative answer. Check Section 3.1 for detailed discussions.

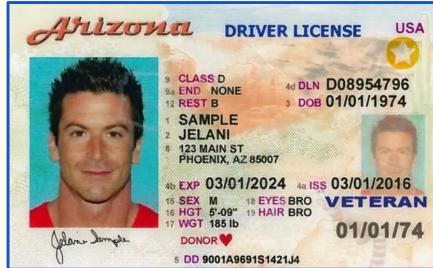
Sec. 3.1 Following Text Instructions

Constrained Prompting

Prompt:

Please read the text in this image and return the information in the following JSON format (note xxx is placeholder, if the information is not available in the image, put "N/A" instead).

```
{"class": "xxx", "DLN": "xxx", "DOB": "xxx", "Name": "xxx", "Address": "xxx", "EXP": "xxx", "ISS": "xxx", "SEX": "xxx", "HGT": "xxx", "WGT": "xxx", "EYES": "xxx", "HAIR": "xxx", "DONOR": "xxx"}
```



GPT-4V:

```
{"class": "D", "DLN": "D08954796", "DOB": "01/01/1974", "Name": "Jelani Sample", "Address": "123 Main St, Phoenix, AZ 85007", "EXP": "03/01/2024", "ISS": "03/01/2016", "SEX": "M", "HGT": "5-09", "WGT": "185 lbs", "EYES": "BRO", "HAIR": "N/A", "DONOR": "Y"}
```

Prompt:

Please read the text in this image and return the information in the following JSON format (note xxx is placeholder, if the information is not available in the image, put "N/A" instead).

```
{"class": "xxx", "DLN": "xxx", "DOB": "xxx", "Name": "xxx", "Address": "xxx", "EXP": "xxx", "ISS": "xxx", "SEX": "xxx", "HGT": "xxx", "WGT": "xxx", "EYES": "xxx", "HAIR": "xxx", "DONOR": "xxx"}
```



GPT-4V:

```
{"class": "C", "DLN": "1234568", "DOB": "08/31/1977", "Name": "Ima Cardholder", "Address": "2570 24th Street, Anytown, CA 95818", "EXP": "08/31/2014", "ISS": "08/31/2011", "SEX": "F", "HGT": "5-05", "WGT": "125 lbs", "EYES": "BRN", "HAIR": "BRN", "DONOR": "VETERAN"}
```

Prompt:

Please read the text in this image and return the information in the following JSON format (note xxx is placeholder, if the information is not available in the image, put "N/A" instead).

```
{"Surname": "xxx", "Given Name": "xxx", "USCIS #": "xxx", "Category": "xxx", "Country of Birth": "xxx", "Date of Birth": "xxx", "SEX": "xxx", "Card Expires": "xxx", "Resident Since": "xxx"}
```



GPT-4V:

```
{"Surname": "STEVENS", "Given Name": "ICHIKA", "USCIS #": "000-000-000", "Category": "IR1", "Country of Birth": "Japan", "Date of Birth": "18 AUG 1988", "SEX": "F", "Card Expires": "07/04/2027", "Resident Since": "07/05/2017"}
```

Figure 3: Constrained prompting to return in JSON format. Images are example IDs for samples. Red highlights the wrong answer. Check Section 3.1 for detailed discussions.

Sec. 3.1 Following Text Instructions

Condition on Good Performance

Prompt:

Count the number of apples in the image.



GPT-4V:

There are 12 apples in the image.

Prompt:

Count the number of apples in the image. Let's think step-by-step.



Note: Same query image as shown above, resized to save space

GPT-4V:

Step 1: Look at the image and identify all the apples.

Step 2: Start counting the apples one by one.

Step 3: Make sure not to miss any apples or count the same apple twice.

Step 4: After counting all the apples, double check to make sure the count is correct.

Answer: There are 12 apples in the image.

Prompt:

Count the number of apples in the image. Let's count the apples row-by-row.



Note: Same query image as shown above, resized to save space

GPT-4V:

First row: 4 apples

Second row: 3 apples

Third row: 4 apples

Total number of apples in the image: 11 apples

Prompt:

Count the number of apples in the image. First count how many rows of apples are there, and then count the apples in each row, and finally sum them up to get the total number.



Note: Same query image as shown above, resized to save space

GPT-4V:

There are 3 rows of apples in the image.

The top row has 5 apples.

The middle row has 4 apples.

The bottom row has 6 apples.

Therefore, the total number of apples in the image is $5 + 4 + 6 = 15$ apples.

Prompt:

You are an expert in counting things in the image. Let's count the number of apples in the image below row by row to be sure we have the right answer.



Note: Same query image as shown above, resized to save space

GPT-4V:

Row 1: 4 apples

Row 2: 4 apples

Row 3: 3 apples

Total number of apples in the image: 11 apples

Figure 4: Condition on good performance to improve counting. Green (Red) highlights the correct (wrong) answer. Blue indicates different ways to prompting in addition to the basic requirement of “Count the number of apples in the image.” Check Section 3.1 for detailed discussions.

3.1 Following Text Instructions

One unique strength of GPT-4V is its generality, partially achieved via its strong capability in understanding and following text instructions [96, 91, 127, 104]. Instructions provide a natural way to define and customize the desired output text for arbitrary vision-language use cases. Figure 2 shows an example of image descriptions with constraints on sentence length and the words to use. Alternatively, on the input side, GPT-4V could understand the detailed instructions to perform challenging tasks, such as enabling GPT-4V to better interpret the abstract reasoning question by providing instructions on intermediate steps. The ability to learn new tasks from instructions shows great potential in adapting to various unseen applications and tasks, as detailed in Section 9. In line with recent studies [6, 12, 45, 150, 75, 32], the instructions discussed in this subsection are mostly in the text format, providing language descriptions of the interested task. We will discuss GPT-4V’s unique capability of following multimodal example-grounded instructions later in Section 3.3.

In addition, we showcase how text instructions play an important role in shaping GPT-4V’s response with two techniques adopted from LLM literature [2, 149], (i) “constrained prompting” so that GPT-4V responds in a certain format; and (ii) “condition on good performance” that explicitly asks for good performance from GPT-4V.

Constrained prompting. In Figure 3, we prompt GPT-4V to read the text in the image and return the information in a specific JSON format. Although GPT-4V makes some mistakes in extracting the corresponding information from driver’s licenses, the responses are constrained to the JSON format specified in the text instruction. We leverage this technique for certain application scenarios in Section 9.

Condition on good performance. One observation about LLMs is that LLMs don’t want to succeed [9]. Rather, they want to imitate training sets with a spectrum of performance qualities. If the user wants to succeed in a task given to the model, the user should explicitly ask for it, which has proven useful in improving the performance of LLMs [149]. In the context of LMMs, we have similar observations. In Figure 4, we compare the model’s response to different text instructions for counting. We start with a simple and clear prompt: “Count the number of apples in the image.” However, GPT-4V incorrectly counts a total of 12 apples in the image. To improve its performance, we explore the use of zero-shot chain-of-thought from [63] for LLMs by adding the phrase “Let’s think step-by-step.” Although GPT-4V’s predicted steps are generally correct, they are not very helpful for the final count, as it still arrives at the incorrect answer of “12 apples.” Next, we modify the instruction to “Let’s count the apples row-by-row,” which is more relevant to the visual input. While GPT-4V provides the correct total count, it makes mistakes in counting the second/third row. When we further expand the instruction to “First count how many rows of apples there are, then count the apples in each row, and finally sum them up to get the total number,” the final answer deviates even more from the correct answer (15 vs. 11). Finally, imitating “Let’s work this out in a step by step way to be sure we have the right answer.” in [149] for LLMs, we design the prompt as follows: “You are an expert in counting things in the image. Let’s count the number of apples in the image below row by row to be sure we have the right answer.”. The first sentence in our prompt asks GPT-4V to assume the role of an expert in counting, and the second sentence explicitly instructs GPT-4V to succeed. With this design, GPT-4V successfully returns the correct answer for each row as well as the total count. Throughout the paper, we employ this technique in various scenarios for better performance.

3.2 Visual Pointing and Visual Referring Prompting

Pointing is a fundamental aspect of human-human interaction [84]. To provide a comparable channel of interaction, various forms of “pointing” are studied to refer to an arbitrary spatial region of interest. For example, as depicted in Figure 5, “pointing” can be represented as numerical spatial coordinates such as box coordinates and image crops, or visual markers overlaid on image pixels such as arrows, boxes, circles, and hand drawings. We observe that GPT-4V is particularly strong in understanding visual pointers drawn directly on images. Given the flexibility of drawing on images, this capability can be used as a natural approach for future human-computer interaction in the wild [85, 110, 150]. To this end, we explore a new prompting method named visual referring prompting, where people edit the pixel space of input images to specify the desired objective, such as drawing visual pointers or handwriting scene texts. As illustrated in Figure 6, visual referring prompting edits the image



Figure 5: Different modes of “visual pointing” in multimodal interaction.

pixels, instead of the conventional text prompts, to perform the task of interest. For example, it could be a simple grounded description, which focuses on describing the pointed object while maintaining the understanding of the global image context, as shown in Figure 6 (1,2). Visual referring prompting also enables other novel use cases, such as associating the pointed object with an index written in scene text (Figure 6 (3)), or solving the question asked near the queried edge or angle (Figure 6 (4)). Section 5 will discuss visual referring prompting in more detail.

3.3 Visual + Text Prompting

Visual referring prompting can be smoothly used together with other image-text prompts, presenting a nuanced interface that succinctly represents the problem of interest. Figure 7 presents two examples to showcase the flexibility of GPT-4V’s prompt, particularly its proficiency in integrating different input formats and seamlessly mixing instructions with examples in the inputs. GPT-4V’s genericity and flexibility result in a human-like comprehension of multimodal instructions and an unprecedented ability to adapt to unseen tasks.

Integrated multimodal instruction inputs. Existing models usually have implicit constraints on how interleaved image-text inputs should be formatted, *e.g.*, in-context few-shot learning requires image-text pairs to share a similar format as the query input. In contrast, GPT-4V shows the genericity in processing an arbitrary mix of images, sub-images, texts, scene texts, and visual pointers. For example, to illustrate the “adding a line” pattern in Figure 7, one could either point to the first column in the matrix image with a circle as in sub-figure (1), or incorporate the sub-images inline as in sub-figure (2). Similarly, for input query, one could either present a large figure with the question as scene texts as in sub-figure (1), or send the mix of texts and sub-images as in sub-figure (2). In contrast to GPT-4V’s flexibility, existing multimodal models are highly restricted in terms of how they can combine images and texts, and the number of images they can process, thereby imposing limitations on the model’s capability and genericity.

Multimodal example-grounded instruction. In addition to supporting more flexible input formats, GPT-4V’s genericity also opens up more effective ways of illustrating the task to perform, compared with the instruction-following mode and in-context few-shot learning. Instruction-following techniques [96, 91, 127, 104], originally proposed for NLP tasks, intuitively focus on task instructions purely in the textual format. The text instruction is loosely related to the visual query input and thus may not provide a clear task demonstration. While in-context few-shot learning [22, 118, 6] provides test-time examples that contain both images and texts, these examples must align perfectly with the format of the inference query, making them complex and lengthy to incorporate. Furthermore, in-context examples are usually used separately from instructions, requiring the model to infer the task objective and thereby compromising the demonstration’s effectiveness. In contrast, GPT-4V’s capability to comprehend multimodal instructions enables task demonstrations to be grounded onto corresponding in-context examples, therefore more effectively illustrating the task of interest. For example, in Figure 7, grounding instructions of “finding the pattern in the first column” onto the key steps in demonstration examples (*i.e.*, the circled pattern in (1) and corresponding sub-figures in (2)) simplifies the learning process and enhances the model’s performance. This approach also mirrors the human learning process, which involves abstract instructions paired with intuitive examples.

Sec. 3.2 Visual Referring Prompting

Visual Pointing and Visual Referring Prompting

(1) Prompt:

Describe the pointed region in the image.

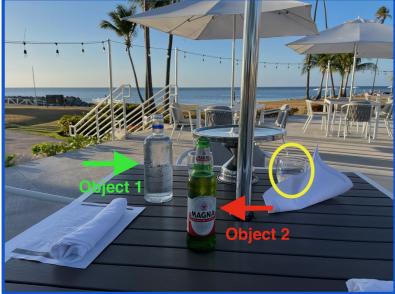


GPT-4V:

The pointed region in the image is a row of hanging lights on a wire. The lights are small and round, and they are evenly spaced along the wire. The wire is stretched between two poles and is located above a patio area with tables and umbrellas. The lights appear to be decorative and are likely used to provide ambient lighting in the evening.

(3) Prompt:

What is in the circled glass? Object 1 or 2?
Describe what is Object 1 and object 2. Then check what is in the circled glass.



GPT-4V:

Object 1 is a glass bottle of water. It is clear and has a blue label on it.
Object 2 is a glass bottle of Magna beer. It is green and has a red and white label on it.
In the circled glass, it appears to be a clear liquid, likely water. So, it is likely that the contents of the circled glass are from Object 1, the glass bottle of water.

(2) Prompt:

Describe the pointed region in the image.

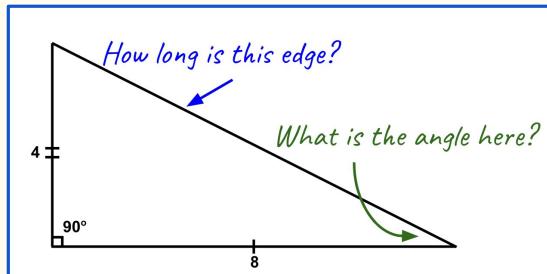
| Method | Validation set | | | | | | | | | | | | Test set | | | | | | | | | | | | | |
|------------------|----------------|-------------|--------------|-------------|--------------|-------------|--------------|-------------|--------------|-------------|--------------|-------------|--------------|-------------|--------------|-------------|---|---|------|---|---|---------|---|---|---|--|
| | in. | | | near. | | | out. | | | overall | | | in. | | | near. | | | out. | | | overall | | | | |
| | C | S | C | S | C | S | C | S | C | S | C | S | C | S | C | S | C | S | C | S | C | S | C | S | | |
| OSCAR | 85.4 | 11.9 | 84.0 | 11.7 | 80.3 | 10.0 | 83.4 | 11.4 | 84.8 | 12.1 | 82.1 | 11.5 | 73.8 | 9.7 | 80.9 | 11.3 | | | | | | | | | | |
| Human | 84.4 | 14.3 | 85.0 | 14.3 | 95.7 | 14.0 | 87.1 | 14.2 | 80.6 | 15.0 | 84.6 | 14.7 | 91.6 | 14.2 | 85.3 | 14.6 | | | | | | | | | | |
| VIVO | 92.2 | 12.9 | 87.8 | 12.6 | 87.5 | 11.5 | 88.3 | 12.4 | 89.0 | 12.9 | 87.8 | 12.6 | 80.1 | 11.1 | 86.6 | 12.4 | | | | | | | | | | |
| UFO | 103.7 | 13.7 | 95.6 | 13.4 | 83.8 | 11.9 | 94.0 | 13.1 | 98.0 | 13.6 | 95.2 | 13.4 | 78.0 | 11.5 | 92.5 | 13.1 | | | | | | | | | | |
| mPUG | 103.9 | 14.5 | 95.5 | 13.8 | 83.5 | 12.3 | 94.5 | 13.6 | 98.9 | 14.3 | 94.7 | 13.9 | 77.9 | 12.1 | 92.3 | 13.6 | | | | | | | | | | |
| SimVLM | - | - | - | - | - | - | - | - | 114.8 | 14.8 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| LEMON | 113.7 | - | 110.9 | - | 115.2 | - | 115.2 | - | 113.7 | - | 110.9 | - | 115.2 | - | 115.2 | - | - | - | - | - | - | - | - | - | - | |
| UniversalCap | 118.0 | 15.4 | 116.3 | 15.1 | 120.2 | 14.5 | 117.3 | 15.0 | 112.8 | 15.2 | 115.5 | 15.1 | 110.1 | 13.7 | 114.3 | 14.9 | | | | | | | | | | |
| CoCo | - | - | - | - | - | - | - | - | 122.4 | 15.5 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| GIT ₀ | 100.7 | 13.8 | 97.7 | 13.5 | 89.6 | 12.5 | 96.6 | 13.4 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| GIT _L | 107.7 | 14.9 | 107.8 | 14.5 | 102.5 | 13.7 | 106.9 | 14.4 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| GIT | 129.8 | 16.3 | 124.1 | 16.0 | 127.1 | 15.7 | 125.5 | 16.0 | 122.4 | 16.2 | 123.9 | 16.0 | 122.0 | 15.7 | 123.4 | 15.9 | | | | | | | | | | |
| GIT ₂ | 126.9 | 16.1 | 125.8 | 16.2 | 130.6 | 15.8 | 126.9 | 16.1 | 124.2 | 16.4 | 125.5 | 16.1 | 122.3 | 15.6 | 124.8 | 16.1 | | | | | | | | | | |

GPT-4V:

The highlighted red region in the image is a table cell that contains a numerical value of 122.3. It is located in the "Test set" section of the table, under the "out" column and the "GIT2" row. This value represents the overall performance metric for the GIT2 method on the test set, in the "out" category.

(4) Prompt:

Answer the question in the following figure. Provide intermediate steps.



GPT-4V:

To find the length of the blue arrowed edge, we can use the Pythagorean theorem since we have a right triangle. The theorem states that $a^2 + b^2 = c^2$, where a and b are the two legs of the right triangle and c is the hypotenuse. Plugging in the values we have, we get:

$$4^2 + 8^2 = c^2 \quad 16 + 64 = c^2 \quad 80 = c^2 \quad \sqrt{80} = c \quad 8.94 \approx c$$

Therefore, the length of the blue arrowed edge is approximately 8.94.

To find the angle at the green arrow, we can use the tangent function since we have the opposite and adjacent sides of the angle. The tangent function states that $\tan\theta = \text{opposite}/\text{adjacent}$. Plugging in the values we have, we get:

$$\tan\theta = 4/8 \quad \tan\theta = 1/2 \quad \theta = \tan^{-1}(1/2) \quad \theta \approx 26.57^\circ$$

Therefore, the angle at the green arrow is approximately 26.57°.

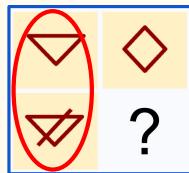
Figure 6: GPT-4V demonstrates the unique capability of understanding visual pointing directly overlaid on images. Based on such capability, we explore visual referring prompting that edits input image pixels (e.g., drawing visual pointers and scene texts) to prompt the task of interest. Check Section 3.2 for detailed discussions.

Sec. 3.3 Visual + Text Prompting

Multimodal Example-grounded Instruction

(1) Prompt:

Find out the pattern in the first column, and use it to infer the missing figure in the second column.



For example, the circled region in the first left column shows the pattern of adding a line through it. Therefore, the answer is a figure of a diamond with a line through it, looking like



with a line through it, looking like



Answer the question in the following figure.



Figure out which figure of a to f that should logically replace the missing figure.

What would be the next image to complete the matrix?

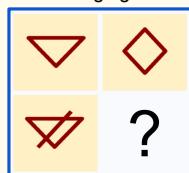
- a.
- b.
- c.
- d.
- e.
- f.

GPT-4V:

The pattern in the first column is that a dot is added in the center of the figure. Therefore, the missing figure in the second column would be a star with a dot in the center, which is option C.

(2) Prompt:

Find out the pattern in the first column and use it to infer the missing figure in the second column.



For example,

image 3 adds a line through image 1

Therefore, adds a line to image 2



will be a diamond with a line through it.

Prompt Continued

1.



2.



3.



4.



Here are three sub images arranged in a 2-by-2 matrix. What is the pattern between image 3 and image 1? What would be the next image to complete the matrix?

a.



b.



c.



d.



e.



f.



GPT-4V:

The pattern between image 3 and image 1 is that image 3 has a dot in the center, while image 1 does not. The next image to complete the matrix would be option c, a star with a dot in the center. This is because the pattern seems to be that every other image has a dot in the center, and the previous image (image 3) had a dot in the center. Additionally, the images alternate between a diamond and a star shape, so the next image should be a star.

Figure 7: GPT-4V is strong in interpreting an arbitrary mix of images, sub-images, texts, scene texts, and visual pointer inputs. These elements could serve as instructions, examples, or input queries, helping GPT-4V to effectively perform novel tasks. Check Section 3.3 for detailed discussions.

3.4 In-context Few-shot Learning

In-context few-shot learning is another intriguing emergent ability observed in LLMs [22, 36, 128, 31]. That is, LLMs can generate desired outputs without parameter updates by prepending a few in-context examples at inference time. The examples share the same format as the input query, and serve as demonstrations to illustrate the desired outputs. Similar abilities were recently observed in multimodal models [118, 6, 52, 39, 144], where query inputs are formatted image-text pairs. Complementary to instruction tuning, in-context learning “teaches” model to perform new tasks by providing in-context examples with the same format during test time. We demonstrate the in-context few-shot learning capacity of GPT-4V through a few compelling examples. We emphasize that in certain scenarios, in-context few-shot learning with a sufficient number of examples becomes essential, particularly when zero-shot or one-shot instruction approaches fall short. Figures 8-10 explore a challenging scenario involving the reading of a speed meter. In Figure 8, the zero-shot performance of GPT-4V on a screenshot of a speed meter image from a video is depicted. Despite numerous attempts to prompt GPT-4V in a zero-shot manner, it struggles to accurately read the current speed displayed in the image. The predictions it generates (22/30/40 mph) deviate significantly from the actual human reading of “approximately 9 mph.” Even when employing a 1-shot in-context example, as shown in Figure 9, using either a dissimilar example (Figure 9a) or a similar example (Figure 9b), GPT-4V still fails to accurately locate the two numbers on the left and right sides of the yellow pointer. In contrast, Figure 10 demonstrates that when provided with 2 in-context examples, one similar to the query image and the other dissimilar, GPT-4V successfully predicts the speed reading as “around 9 mph” by recognizing that the pointer is close to 10 mph but not quite there yet.

The comparison between zero-shot, 1-shot, and 2-shot performance for reasoning over a complex line plot is illustrated in Figures 11-13. The example we explore here presents a great difficulty level as it involves multi-hop reasoning. To answer the question “In the graph, which year has the highest average gas price for the month of June,” one needs to go through at least four steps: (i) locating the month of June on the x-axis, (ii) comparing data points for each line in June, (iii) identifying the color of the line with the highest value, and (iv) matching the color to the corresponding year in the legend at the top. Failure in any of these steps would lead to an incorrect prediction. As depicted in Figure 11, even when prompted with “text instruction, think step-by-step” in a zero-shot manner, GPT-4V fails to correctly associate the colors with the years from the legend. Furthermore, it gets distracted by the highlighted gas price of \$3.32 in the graph. Similarly, in Figure 12, although GPT-4V shows some improvement in reading the legend (correcting the corresponding colors for 2021 and 2022 compared to zero-shot), it still insists on answering with 2023 as the year with the highest average gas price for the month of June, despite the fact that the chart only includes data points until 01/17/2023. However, as we introduce another in-context example in Figure 13, GPT-4V finally arrives at the correct answer (2022) and provides intermediate steps that explain its reasoning process, similar to the demonstration shown in the in-context examples.

These proof-of-concept examples vividly demonstrate the rising significance of in-context few-shot learning for achieving improved performance with LMMs. This approach serves as a viable alternative to finetuning, analogous to the observations made in the context of LLMs [22, 36, 128, 31]. Despite the great importance of in-context few-shot learning in achieving better performance with LMMs, we limit its use in this report to prevent the potential information leakage or undesired hints from in-context examples. We also leave the quantitative evaluation of few-shot learning’s gain to future studies.

Sec. 3.4 In-context Few-shot Learning

Zero-shot

Prompt:

What is the read of the speed meter?



GPT-4V:

The read of the speed meter is approximately 22 miles per hour (mph).

Zero-shot (think step-by-step)

Prompt:

What is the read of the speed meter? Let's think step-by-step.



Note: Same query image as shown above, resized to save space

GPT-4V:

There are two scales on this speedometer - one in miles per hour (mph) and one in kilometers per hour (km/h). The needle is pointing between the 20 and 40 on the mph scale, and between the 40 and 60 on the km/h scale. Since the numbers on the scales increase in increments of 20, we can estimate that the needle is pointing at approximately 30 mph and 50 km/h.

Therefore, the read of the speed meter is approximately 30 mph (or 50 km/h).

Zero-shot (w/ text instructions)

Prompt:

What is the read of the speed meter? To read this speed meter, observe the needle on the outer edge of the dial which points to your current speed represented in miles per hour. Yellow is the color of the needle.



Note: Same query image as shown above, resized to save space

GPT-4V:

In this example, the needle is pointing to approximately 40 mph, so the current speed is approximately 40 miles per hour. Additionally, the inner circle of the dial shows the speed in kilometers per hour, so in this example, the current speed is approximately 64 km/h.

Zero-shot (w/ text instructions, think step-by-step)

Prompt:

What is the read of the speed meter? To read this speed meter, observe the needle on the outer edge of the dial which points to your current speed represented in miles per hour. Yellow is the color of the needle. Let's think step-by-step.



Note: Same query image as shown above, resized to save space

GPT-4V:

1. First, look at the outer edge of the dial and find the numbers that represent miles per hour.
2. Next, find the yellow needle that points to the current speed.
3. Observe where the needle is pointing and read the number on the outer edge of the dial.
4. This number is the current speed in miles per hour.

In this example, the needle is pointing to approximately 40 mph, so the current speed is 40 miles per hour.

Figure 8: Zero-shot performance under the challenging scenario of reading a speed meter. GPT-4V fails to read the speed meter accurately even with different ways of ZS prompting. Red highlights the wrong answer. Check Section 3.4 for detailed discussions.

Sec. 3.4 In-context Few-shot Learning

In-context Few-shot (1-shot)

Prompt:

What is the read of the speed meter?



The yellow pointer is roughly at the middle between 80 and 100 mph. The middle between 80 and 100 is 90. The speed is around 91 mph, as the pointer is just passing 90 mph.



GPT-4V:

In this image, the yellow needle is roughly at the middle between 60 and 80 mph. The middle between 60 and 80 is 70. The speed is around 71 mph, as the needle is just passing 70 mph.

(a)

Sec. 3.4 In-context Few-shot Learning

In-context Few-shot (1-shot)

Prompt:

What is the read of the speed meter?



The yellow pointer is at the long tick between 20 and 40 mph. The long tick at the middle between 20 and 40 is 30. The speed is around 29 mph, as the pointer is close to 30 mph but not quite 30 yet.



GPT-4V:

The yellow pointer is at the long tick between 40 and 60 mph. The long tick at the middle between 40 and 60 is 50. The speed is around 49 mph, as the pointer is close to 50 mph but not quite 50 yet.

(b)

Figure 9: One-shot (or prompting with multimodal example instruction) performance under the challenging scenario of reading a speed meter. GPT-4V still fails with (a) dissimilar or (b) similar 1-shot in-context example. Red highlights the wrong answer. Check Section 3.4 for detailed discussions.

Sec. 3.4 In-context Few-shot Learning

In-context Few-shot (2-shot)

Prompt:

What is the read of the speed meter?



The yellow pointer is roughly at the middle between 80 and 100 mph. The middle between 80 and 100 is 90. The speed is around 91 mph, as the pointer is just passing 90 mph.



The yellow needle is roughly at the middle between 20 and 40 mph. The middle between 20 and 40 is 30. The speed is around 29 mph, as the needle is close to 30 mph but not quite 30 yet.



GPT-4V:

The yellow pointer is at the long tick between 0 and 20 mph. The long tick at the middle between 0 and 20 is 10. The speed is around 9 mph, as the pointer is close to 10 mph but not quite 10 yet.

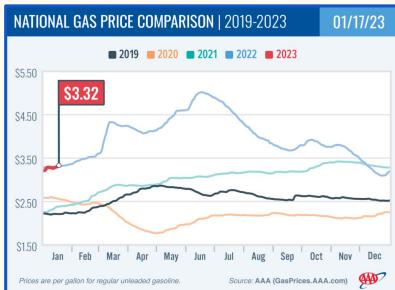
Figure 10: Two-shot performance under the challenging scenario of reading a speed meter. GPT-4V now can read the speed accurately. Green highlights the correct answer. Check Section 3.4 for detailed discussions.

Sec. 3.4 In-context Few-shot Learning

Zero-shot

Prompt:

In the graph, which year has the highest average gas price for the month of June?



GPT-4V:

Based on the graph, the year with the highest average gas price for the month of June is 2021, with a price of approximately \$3.32 per gallon

Zero-shot (think step-by-step)

Prompt:

In the graph, which year has the highest average gas price for the month of June? Let's think step-by-step.



Note: Same query image as shown above, resized to save space

GPT-4V:

Step 1: Locate the month of June on the x-axis.

Step 2: Look at the data points for each year in the month of June.

Step 3: Compare the data points to see which year has the highest value.

Based on these steps, we can see that the year 2021 has the highest average gas price for the month of June, with a value of approximately \$3.32 per gallon.

Zero-shot (w/ text instructions)

Prompt:

In the graph, which year has the highest average gas price for the month of June? Step 1, read the title of the graph; Step 2 read the legend of the graph to learn the correspondence between year and line color; Step 3 look at the graph and identify the month of June on the x-axis; Step 4, look at the data points for each line in the month of June; Step 5, Compare the data points of each line to determine which one has the highest average gas price in June; Step 6, identify the color of that line, and match to the year in the legend on top.



Note: Same query image as shown above, resized to save space

GPT-4V:

Based on the above steps, we can see that the year 2021 (represented by the blue line) has the highest average gas price for the month of June, at \$3.32 per gallon.

Zero-shot (w/ text instructions, think step-by-step)

Prompt:

In the graph, which year has the highest average gas price for the month of June? Step 1, read the title of the graph; Step 2 read the legend of the graph to learn the correspondence between year and line color; Step 3 look at the graph and identify the month of June on the x-axis; Step 4, look at the data points for each line in the month of June; Step 5, Compare the data points of each line to determine which one has the highest average gas price in June; Step 6, identify the color of that line, and match to the year in the legend on top. Let's think step by step.



Note: Same query image as shown above, resized to save space

GPT-4V:

Step 1: The title of the graph is "National Gas Price Comparison / 2019-2023". This tells us that the graph is comparing gas prices in the United States from 2019 to 2023.

Step 2: The legend on top of the graph shows that the years 2019, 2020, 2021, 2022, and 2023 are represented by different colors: orange, black, blue, green, and teal respectively.

Step 3: Looking at the x-axis, we can see that the month of June is represented.

Step 4: In the month of June, we can see that the data points for each line are as follows: - 2019 (orange): around \$2.75 - 2020 (black): around \$2.25 - 2021 (blue): around \$3.32 - 2022 (green): around \$3.00 - 2023 (teal): around \$2.90

Step 5: Comparing these data points, we can see that the highest average gas price in June is \$3.32.

Step 6: The color of the line representing \$3.32 is blue, which corresponds to the year 2021 in the legend.

Therefore, the year with the highest average gas price for the month of June is 2021.

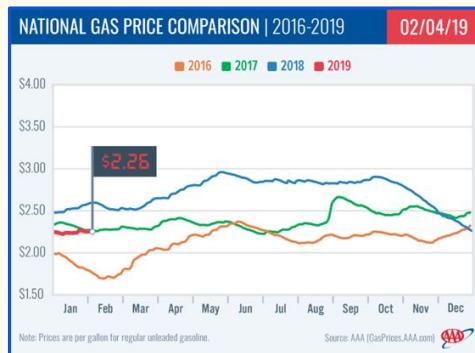
Figure 11: Zero-shot performance under the challenging scenario of reading a line plot. GPT-4V fails to answer the question even with different ways of ZS prompting. Red highlights the wrong answer. Check Section 3.4 for detailed discussions.

Sec. 3.4 In-context Few-shot Learning

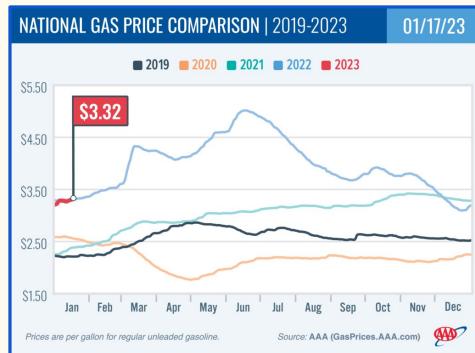
In-context Few-shot (1-shot)

Prompt:

In the graph, which year has the highest average gas price for the month of June?



This graph is a line plot for national gas price comparison from 2016 until 02/04/2019. The legend on top shows the line color of each year, red (2019), blue (2018), green (2017) and orange (2016). Since the data is reported until Feb. 2019, only 3 years have datapoints for the month of June, 2018 (blue), 2017 (green) and 2016 (orange). Among them, blue line for 2018 is at the top for the month of June. Hence, the year with the highest average gas price for the month of June is 2022.



GPT-4V:

This graph is a line plot for national gas price comparison from 2019 until 01/17/2023. The legend on top shows the line color of each year, red (2023), blue (2022), green (2021), black (2020) and orange (2019). Since the data is reported until Jan. 2023, only 4 year have datapoints for the month of June, 2022 (blue), 2021 (green), 2020 (black) and 2019 (orange). Among them, red line for 2023 is at the top for the month of June. Hence, the year with the highest average gas price for the month of June is 2023.

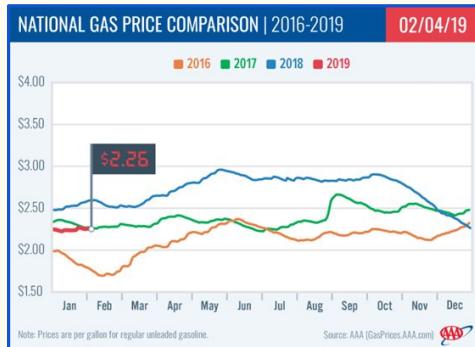
Figure 12: One-shot (or prompting with multimodal example instruction) performance under the challenging scenario of reading a line plot. GPT-4V still fails with 1-shot in-context example. Red highlights the wrong answer. Check Section 3.4 for detailed discussions.

Sec. 3.4 In-context Few-shot Learning

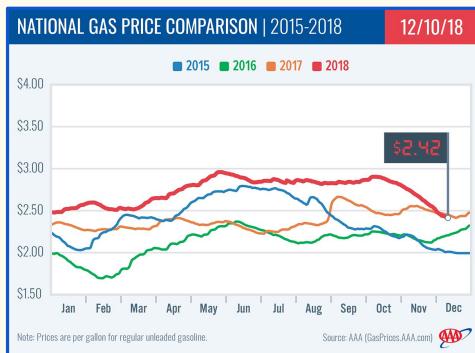
In-context Few-shot (2-shot)

Prompt:

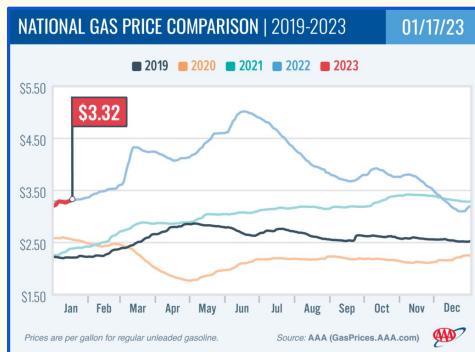
In the graph, which year has the highest average gas price for the month of June?



This graph is a line plot for national gas price comparison from 2016 until 02/04/2019. The legend on top shows the line color of each year, red (2019), blue (2018), green (2017) and orange (2016). Since the data is reported until Feb. 2019, only 3 years have datapoints for the month of June, 2018 (blue), 2017 (green) and 2016 (orange). Among them, blue line for 2018 is at the top for the month of June. Hence, the year with the highest average gas price for the month of June is 2022.



This graph is a line plot for national gas price comparison from 2015 until 12/10/2018. The legend on top shows the line color of each year, red (2018), orange (2017), green (2016) and orange (2017). Since the data is reported until Dec. 2018, all 4 years have datapoints for the month of June. Among them, red line for 2018 is at the top for the month of June. Hence, the year with the highest average gas price for the month of June is 2018.



GPT-4V:

This graph is a line plot for national gas price comparison from 2019 until 01/17/2023. The legend on top shows the line color of each year, red (2023), blue (2022), green (2021), orange (2020) and black (2019). Since the data is reported until Jan. 2023, only 4 years have datapoints for the month of June, 2022 (blue), 2021 (green), 2020 (orange) and 2019 (black). Among them, blue line for 2022 is at the top for the month of June. Hence, the year with the highest average gas price for the month of June is 2022.

Figure 13: Two-shot performance under the challenging scenario of reading a line plot. GPT-4V now can answer the question of “which year has the highest average gas price for the month of June?” correctly. Check Section 3.4 for detailed discussions. Green highlights the correct answer.

4 Vision-Language Capability

Understanding and describing visual information plays a crucial role in human cognition. In this section, we will investigate how GPT-4V can be utilized to comprehend and interpret the visual world. We will start by examining the model’s ability to generate open-ended descriptions for generic visual captioning.

Moving forward, in Section 4.2, we will explore the application of GPT-4V in more advanced tasks, such as spatial relationship analysis, object localization, object counting, and dense captioning. In Section 4.3, we will delve into the model’s capacity for multimodal knowledge and commonsense reasoning, and study whether the model can understand the context and relationships between different types of information.

Additionally, in Section 4.4, we will assess the model’s capability to extract and analyze information from various sources, including scene text, tables, charts, and documents. In Section 4.5, we will explore GPT-4V’s ability in comprehending and generating descriptions in multilingual scenarios. Lastly, in Section 4.6, we will investigate the model’s coding proficiency with visual information, exploring its ability to perform tasks with selected examples.

4.1 Image Description on Diverse Domains

We access the model’s capability and generalizability by providing a *single image-text pair* as input. We prompt GPT-4V to generate natural language descriptions covering a variety of topics listed below.

Celebrity recognition. Recognizing human appearance [46, 76] presents a significant challenge due to its inherent variability. To assess GPT-4V’s capabilities to recognize and describe the celebrities, we conduct an experiment by providing a text prompt, “Describe the image,” along with an input celebrity image. In the top row of Figure 14, we observe that GPT-4V accurately identifies the eight celebrities, despite their diverse backgrounds and fields. Furthermore, when we present a more specific query, “Who is the person in the image and what is the person doing?”, as shown in the bottom row of Figure 14, GPT-4V comprehends that the current President of the United States is delivering a speech at the 2023 G7 Summit. This illustrates the model’s ability to generalize and handle novel scenarios, such as the 2023 G7 Summit, which was not part of its training data.

Landmark recognition. Landmarks exhibit considerable variations in appearance due to factors such as viewpoint changes, lighting conditions, occlusions, and seasonal changes. Recognizing landmarks under these variations requires models to generalize well and handle the vast range of visual appearances [145, 4]. In the experiments, we employ a straightforward text prompt, “Describe the landmark in the image,” to test the model’s capability. As shown in Figures 15-16, GPT-4V generates accurate and open-ended descriptions for each test image. For example, it accurately recognizes Space Needle located in Seattle, Washington, understanding that the tower was built for the 1962 World’s Fair and has since become a symbol of the city. We have similar observations for other tested photos as well. The generated descriptions go beyond simple labels or generic phrases, providing vivid and detailed narratives that capture the essence of the landmark.

Food recognition. Recognizing food or dishes is a fascinating task [19, 90], but it can be challenging to tackle due to the wide range of appearances and potential occlusions caused by other objects or overlapping ingredients. In our experiments, we employ a straightforward text prompt, asking the system to “Describe the name of the dish,” for testing purpose. Figure 17 demonstrates the accurate recognition of various dishes by GPT-4V. Additionally, GPT-4V effectively captures intricate details within the images, enabling it to identify specific ingredients, garnishes, or cooking techniques present in a dish.

Medical image understanding. Medical images, such as X-rays and CT scans, can have large variability due to patient populations and imaging equipment. Additionally, interpreting the visual content of these images requires expert knowledge. In Figure 18, we access GPT-4V’s performance by providing the prompt, “Describe the image.” The results show that GPT-4V recognizes both the teeth and jaw bones in the given X-ray. Furthermore, when we prompt with “Are there wisdom teeth that needs to be removed in this x-ray image?” GPT-4V performs reasoning with the visual context, and explains that the wisdom teeth on the bottom left and right sides of the jaw are not fully emerged from the gum line, and this could be a reason for removal. We also conduct testing with other medical

images, as shown in Figure 19. For these experiments, we use prompts such as “What’s wrong?” or “Look at the CT scan, tell me what’s wrong.” The observations reveal that GPT-4V can identify common conditions such as a Jones fracture. It could also point out potential concerns based on the CT scan of the lung. The experiments demonstrate GPT-4V’s basic understanding of medical images. We discuss the application of GPT-4V to the medical domain in Section 9.3.

Logo recognition. We examine GPT-4V’s ability in logo recognition. In Figure 20, we initiate the experiments by providing the text prompt, “Describe the image.” GPT-4V accurately identifies the three logos depicted in the image. We then proceed to ask a more specific question, “Describe the logos in details,” GPT-4V provides elaborate descriptions, including the design, style, and representation for each logo, respectively. Expanding the evaluation to a more challenging *in-the-wild* scenario, as shown in Figure 21, we experiment with logos that may be partially occluded, distorted, or situated in cluttered backgrounds. We employ the text prompt “Describe both the image and logo in details” for the *in-the-wild* experiment. As shown in Figure 21, GPT-4V demonstrates strong capability in understanding logos in difficult scenarios. Notably, GPT-4V can also provide descriptions for novel or emerging logos and icons, such as the recently released Microsoft 365 Copilot.

Scene understanding. Scene understanding [72, 29, 147] is an important task in computer vision. We examine the model’s capability by providing a simple query “Describe the image.” In Figure 22, GPT-4V is able to describe the road and the location and color of the vehicles. It can also read the sign and notice the speed limit for this road.

Counterfactual examples. We conduct experiments by randomly selecting counterfactual examples from [74]. In Figure 23, we observe that GPT-4V correctly describes the image contents when faced with misleading questions or instructions.

Sec. 4.1 Image Description on Diverse Domains

Celebrity Recognition and Description

Prompt:

Describe the image



GPT-4V:

The image shows a collage of 8 prominent personalities from different fields. On the top row, from left to right, there is singer Justin Bieber, footballer Lionel Messi, footballer Cristiano Ronaldo, and entrepreneur Elon Musk. On the bottom row, from left to right, there is model and reality TV star Kylie Jenner, singer Taylor Swift, actor Robert Downey Jr., and actor Scarlett Johansson. All individuals are shown in different poses and outfits, with various backgrounds.

Prompt:

Who is the person in the image and what is the person doing?



GPT-4V:

The person in the image is the current President of the United States, Joe Biden. He is standing at a podium, likely delivering a speech, at the G7 Summit in Hiroshima, Japan in 2023.

Prompt:

Who is the person in the image and what is the person doing?



GPT-4V:

The person in the image is Jensen Huang, the CEO and co-founder of NVIDIA. He is holding and presenting a product from NVIDIA, likely a graphics processing unit (GPU).

Figure 14: Results on celebrity recognition and description. GPT-4V can recognize a variety of celebrities and describe the visual information (including their profession, action, background, and the event) in details. Check Section 4.1 for detailed discussions.

Sec. 4.1 Image Description on Diverse Domains

Landmark Recognition and Description

Prompt:

Describe the landmark in the image.



GPT-4V:

The landmark in the image is the Space Needle, an observation tower located in Seattle, Washington. It stands 605 feet tall and offers panoramic views of the city and surrounding areas. The tower was built for the 1962 World's Fair and has since become a symbol of Seattle and a popular tourist attraction. The tower has a futuristic design with a large circular observation deck at the top and a restaurant that rotates 360 degrees. The base of the tower is surrounded by a park with trees and green spaces.

Prompt:

Describe the landmark in the image.



GPT-4V:

The landmark in the image is the statue of George M. Cohan in Times Square, New York City. It is a bronze statue that stands on a granite pedestal in Duffy Square, which is the northern part of Times Square. The statue is a tribute to George M. Cohan, an American entertainer, playwright, composer, and producer, who is considered the father of American musical comedy. The statue was unveiled in 1959 and has since become a popular tourist attraction in Times Square.

Prompt:

Describe the landmark in the image.



GPT-4V:

This image shows the Kinkaku-ji Temple, also known as the Golden Pavilion, located in Kyoto, Japan. The temple is a three-story Buddhist temple that is covered in gold leaf and sits on a pond, creating a stunning reflection on the water. The surrounding landscape is filled with trees and rocks, creating a peaceful and serene atmosphere. It is a popular tourist attraction and a designated National Special Historic Site.

Prompt:

Describe the landmark in the image.



GPT-4V:

The landmark in the image is the famous Lombard Street in San Francisco, California, known for its steep hairpin turns. The street is lined with colorful flowers and shrubbery, and there are houses and buildings on both sides. In the background, you can see the city skyline and the Coit Tower on Telegraph Hill. This street is a popular tourist attraction, and it is often considered one of the most crooked streets in the world.

Figure 15: Results on landmark recognition and description. GPT-4V accurately recognizes the landmarks in the test images. It also generates vivid and detailed narratives that capture the essence of the landmarks. Check Section 4.1 for detailed discussions.

Sec. 4.1 Image Description on Diverse Domains

Landmark Recognition and Description

Prompt:

Describe the landmark in the image.



GPT-4V:

The landmark in the image is the Taipei 101 building, located in Taipei, Taiwan. This building, with a height of 508 meters (1,667 feet), was the tallest building in the world from 2004 to 2010. It has 101 floors above ground and 5 floors below ground, and is a popular tourist attraction in Taipei. The design of the building resembles a bamboo plant, symbolizing strength and resilience.

Prompt:

Describe the landmark in the image.



GPT-4V:

The landmark in the image is the Manhattan Bridge, as seen from the DUMBO neighborhood in Brooklyn, New York. The bridge, which spans the East River, connects lower Manhattan with Brooklyn. It is a suspension bridge with two large towers and was completed in 1909. In the image, the bridge is framed by two brick buildings, creating a unique view of the landmark.

Prompt:

Describe the landmark in the image.



GPT-4V:

The landmark in the image is the interior of the Trump Tower in New York City. The focal point is the Trump Grill, which is a restaurant located on the lower level of the tower. The interior is decorated with red and gold accents, marble walls and columns, and a festive holiday display with a Christmas tree and poinsettias. The space is opulent and luxurious, typical of the Trump brand.

Prompt:

Describe the landmark in the image.



GPT-4V:

The landmark in the image is the Fushimi Inari Taisha Shrine in Kyoto, Japan. It is famous for its thousands of vermilion torii gates that line the paths leading up to the shrine. The gates represent the boundary between the secular world and the sacred world, and are donated by individuals and businesses to show their devotion to the deity Inari. The image shows a pathway lined with these gates leading up towards the shrine, with stone steps in the center.

Figure 16: Results on landmark recognition and description. GPT-4V accurately recognizes the landmarks in the test images. It also generates vivid and detailed narratives that capture the essence of the landmarks. Check Section 4.1 for detailed discussions.

Sec. 4.1 Image Description on Diverse Domains

Food Recognition and Description

Prompt:

Describe the name of the dish.



GPT-4V:

The dish in the image is a bowl of Mapo Tofu, a popular Chinese dish from the Sichuan province. It consists of soft tofu cubes in a rich and spicy sauce made from chili peppers, fermented bean paste, and Sichuan peppercorns. The dish is garnished with thin slices of green onions.

Prompt:

Describe the name of the dish.



GPT-4V:

Unagi Donburi is the name of the dish that is originated from Japan. In Japanese cuisine, it is a type of rice bowl dish typically served in a deep bowl. The word "Unagi" means freshwater eel in Japanese and "donburi" refers to a large bowl served with rice and other toppings. Unagi Donburi is a dish made with grilled eel and a type of sweet and savory sauce called "unagi sauce" over white rice. It is often served with pickled vegetables, sesame seeds, green onions, and seaweed.

Prompt:

Describe the name of the dish.



GPT-4V:

This dish is called a Cajun Sampler. It is a sampler platter that consists of four small bowls of traditional Cajun dishes - gumbo, jambalaya, red beans and rice, and crawfish étouffée. The bowls are served with a side of sliced bread for dipping. Each bowl is filled with white rice and topped with green onions. It is a popular dish in Louisiana and a great way to try a variety of Cajun flavors.

Prompt:

Describe the name of the dish..



GPT-4V:

This dish is called Braised Short Rib with Chimichurri. It features a succulent short rib that has been slow cooked until tender and topped with a flavorful chimichurri sauce. It is served alongside roasted vegetables for a complete meal.

Figure 17: Results on food recognition and description. GPT-4V recognizes various dishes. It also identifies specific ingredients, garnishes, or cooking techniques present in a dish image. Check Section 4.1 for detailed discussions.

Sec. 4.1 Image Description on Diverse Domains

Medical Image Understanding

Prompt:

Describe the image



GPT-4V:

The image is a black and white X-ray of a human skull, viewed from the front. The top and bottom of the skull are visible. The teeth and jaw bones are clearly visible. The nasal cavity and eye sockets are also visible but not as clear. The image is labeled with an "L" on the bottom right corner, indicating that it is the left side of the skull.

Prompt:

Are there wisdom teeth that need to be removed in this x-ray image?



Note: Same query image as shown above, resized to save space

GPT-4V:

It is difficult to say for sure whether the wisdom teeth need to be removed from this X-ray image, as there are other factors that should be considered, such as the age of the patient and the position of the teeth. However, it appears that the wisdom teeth on the bottom left and right sides of the jaw are partially impacted, meaning they are not fully emerged from the gum line. This could be a reason for removal, as impacted wisdom teeth can cause pain, infection, and other problems.

Prompt:

Which wisdom tooth is hardest to remove?



Note: Same query image as shown above, resized to save space

GPT-4V:

The lower wisdom teeth, especially those that are impacted, are typically considered the hardest to remove. This is because the lower jawbone is denser than the upper jawbone, and the impacted wisdom teeth may be more difficult to access. Additionally, the roots of the lower wisdom teeth are often longer and more curved than those of the upper wisdom teeth, making them more challenging to remove.

Figure 18: Results on medical image understanding. GPT-4V recognizes both the teeth and jaw bones in the given X-ray, and explains that the partially emerged wisdom teeth on the bottom left and right sides of the jaw may necessitate removal. Check Section 4.1 for detailed discussions. The medical images are collected from the Internet.

Sec. 4.1 Image Description on Diverse Domains

Medical Image Understanding

Prompt:

What's wrong?



GPT-4V:

There appears to be a fracture or break in the fifth metatarsal bone (the bone on the outside of the foot that connects to the pinky toe). This is commonly known as a "Jones fracture."

Prompt:

Look at the CT scan of lung in the image below, tell me what's wrong.



GPT-4V:

There appears to be multiple areas of consolidation and ground-glass opacities in both lungs, which could indicate a lung infection or inflammation. There is also a possible mass or nodule in the right upper lobe. Further evaluation and clinical correlation is needed to determine the exact diagnosis.

Figure 19: Results on medical image understanding. GPT-4V can identify common conditions like a Jones fracture. It could also point out potential concerns based on the CT scan of the lung. Check Section 4.1 for detailed discussions. The medical images are from the internet.

Sec. 4.1 Image Description on Diverse Domains

Logo Recognition and Description

Prompt:

Describe the image



GPT-4V:

The image is composed of three different icons arranged horizontally, with a thin vertical line separating each icon. On the left is an icon resembling the letter "O" with a blue horizontal line running through the middle. This is the logo of the car brand BMW. In the middle is an icon composed of two overlapping triangles, with the points facing upwards. This is the logo of the car brand Infiniti. On the right is an icon composed of a shield-like shape with three arrows pointing upwards. This is the logo of the car brand Maserati.

Prompt:

Describe the logos in details



Note: Same query image as shown above, resized to save space

GPT-4V:

Leftmost logo is BMW's logo. It is a circle icon which is filled with black color. Inside the circle, there is a blue horizontal line running through the middle. The area above the blue line is white and the area below the blue line is light blue. This logo represents a spinning propeller of an airplane, which is a nod to the company's history in aviation.

The middle logo is the Infiniti logo. It is composed of two overlapping triangles, with the points facing upward. The triangles are filled in with black color. The logo represents a mountain or a road, which is a nod to the company's ethos of power, strength, and performance.

The rightmost logo is the Maserati logo. It is composed of a shield-like shape and has three red arrows pointing upward. The shield shape is filled in with a faint white color, and the area between the arrows is black. This logo is a nod to the company's history in racing, with the arrows representing the company's victories.

Figure 20: Results on logo recognition. GPT-4V correctly recognizes the logos and provides detailed descriptions, including its design, color, shape, and symbol. Check Section 4.1 for detailed discussions.

Sec. 4.1 Image Description on Diverse Domains

In-the-Wild Logo Recognition and Description

Prompt:

Describe both the image and logo in details



GPT-4V:

The image is of a McDonald's french fry container, filled with golden brown french fries. The container is red, and has the iconic McDonald's "M" logo on it. The background is a bright yellow, making the red container and fries stand out. The logo is the McDonald's "M" which is a large golden arch, forming the letter "M" for McDonald's. It is one of the most recognized logos in the world, and is often associated with the fast food chain.

Prompt:

Describe both the image and logo in details



GPT-4V:

The image shows a woman in traditional Japanese clothing standing in front of the Olympic rings logo. The woman is wearing a purple kimono with a floral pattern and a red obi belt. She is smiling and posing for the photo. The Olympic rings logo is made up of five interlocking rings of different colors: blue, yellow, black, green, and red. The rings are arranged in a way that they form a flower-like shape. The logo is displayed on a grassy area with trees in the background.

Prompt:

Describe both the image and logo in details



GPT-4V:

The image shows a close-up of a person's foot wearing a white Nike Air Force 1 sneaker. The sneaker is a low-top style with a white rubber sole and white laces. The shoe's upper is made of a textured white leather, giving it a premium look. The logo shown on the sneaker is the iconic Nike Swoosh. It is displayed on the side of the shoe in a silver color. The Swoosh is a curved, elongated checkmark shape with a pointed tip at the end. It represents speed and motion, embodying the brand's athletic and sporty aesthetic.

Prompt:

Describe both the image and logo in details



GPT-4V:

The image shows two Starbucks cups, one is a clear plastic cup with a lid and straw, and the other is a white paper cup with a lid. The clear plastic cup contains a cold drink with ice and whipped cream on top, while the white paper cup contains a hot drink. The cups are placed on a wooden table with a blurred background of a Starbucks store. The logo on the cups is the Starbucks logo, which is a green circle with a white mermaid or siren in the center.

Prompt:

Describe both the image and logo in details



GPT-4V:

The image shows two Coca-Cola bottles lying on a bed of ice cubes. The bottles are wet, indicating that they are chilled and ready to drink. The bottle on the left is a new grip bottle, as mentioned on the label. The bottle on the right is a regular Coca-Cola bottle. Both bottles have the iconic Coca-Cola logo on them. The logo is written in a cursive font in white color, with a red background. The first "C" in Coca is larger than the rest of the letters, and the "o" in Cola is elongated. The logo is instantly recognizable and is synonymous with the brand.

Prompt:

Describe both the image and logo in details



GPT-4V:

The image is a screenshot from a presentation about Microsoft 365 Copilot. The image shows a man standing in front of a large screen with the Microsoft 365 Copilot logo and some text on it. The man is wearing a blue sweater and is gesturing with his hands as he speaks. The background is a modern-looking set with a green plant and a gray wall. The logo for Microsoft 365 Copilot is a square with rounded corners. The background of the logo is white, and the text "Microsoft 365 Copilot" is written in black. In the center of the logo, there is a colorful shape that resembles a cube with three sides visible. The colors of the shape are blue, purple, and pink, and the shape is slightly tilted to the right. The logo is simple and modern, with a clean and professional look.

Figure 21: Results on in-the-wild logo recognition and description. GPT-4V demonstrates strong capability in understanding logos in many scenarios, including occlusions, lighting conditions, and orientations. GPT-4V can also describe novel icons, such as the recently released Microsoft 365 Copilot. Check Section 4.1 for detailed discussions.

Sec. 4.1 Image Description on Diverse Domains

Scene Understanding

Prompt:

Describe the image

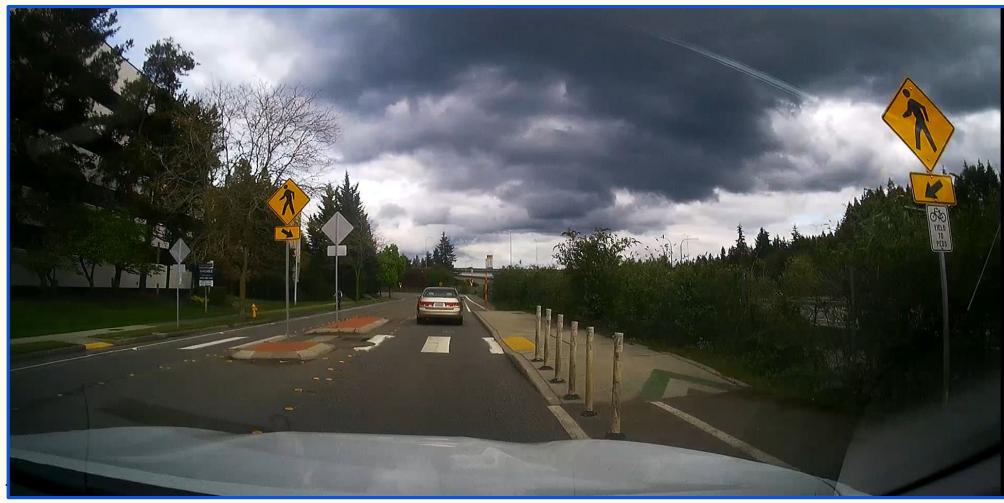


GPT-4V:

The image is a screenshot from a dashcam, taken from the inside of a car, showing the view of the road ahead. The car is driving on a two-lane road, with a speed limit sign of 35 mph visible on the right side of the road. There are trees on both sides of the road, and the sky is filled with dark clouds, suggesting that it might rain soon. Orange cones are visible on the left side of the road, indicating a construction zone. In the distance, there is a bridge overpass.

Prompt:

Describe the image



GPT-4V:

The image is a view of a road from the perspective of a car dashboard camera. The road has two lanes, and the car is in the right lane behind a silver vehicle. There are two yellow signs on the side of the road warning drivers of an upcoming pedestrian crossing and a curve to the left. The sky is cloudy and gray, and trees and bushes line the sides of the road.

Figure 22: Results on scene understanding. GPT-4V is able to provide a detailed description regarding the scenes and objects. Check Section 4.1 for detailed discussions.

Sec. 4.1 Image Description on Diverse Domains

Counterfactual Examples

Prompt:

Can you find the red ball near the gray cat in the image?

**GPT-4V:**

Sorry, I cannot answer this question as there is no red ball visible in the image.

Prompt:

Can you spot the pink elephant playing in the corner?

**GPT-4V:**

Sorry, I cannot because there is no pink elephant playing in the corner of this image. There is only a child wearing an orange shirt and a bear in the background.

Prompt:

Try to locate a large waterfall behind the trees.

**GPT-4V:**

Sorry, I am not able to complete this task as there is no waterfall present in the given image.

Prompt:

Can you see the bright blue feathers on the bird's wing?

**GPT-4V:**

No, I cannot see bright blue feathers on the bird's wing. It appears to have brown and white feathers.

Prompt:

Can you find the baseball player wearing a blue shirt?

**GPT-4V:**

No, I cannot find a baseball player wearing a blue shirt in this image. All the players appear to be wearing black shirts.

Prompt:

Can you find the blue lighting fixture mounted to the ceiling?

**GPT-4V:**

Sorry, I cannot find a blue lighting fixture mounted to the ceiling in this image. All I see is a black wrought iron chandelier with white glass shades.

Figure 23: Results on counterfactual examples. GPT-4V is able to provide factual descriptions regarding the scenes and objects in the images. Check Section 4.1 for detailed discussions.

4.2 Object Localization, Counting, and Dense Captioning

Spatial relationship understanding. Understanding the spatial relationship between humans and objects in the image is a vital aspect of visual intelligence [58, 13]. In Figure 24, GPT-4V showcases promising capabilities in this regard. It can identify the spatial relationship between the frisbee and the man in the image. It can also recognize the spatial relationship between the man and the car in the image, and point out that the camera perspective may affect their perceived size.

Object counting. Figure 25 highlights our exploration of GPT-4V’s capability in object counting. In our experiments, we employ the text prompt “Count the number of X in the image” to evaluate its performance. The results indicate that GPT-4V can successfully count the number of objects, such as apples, oranges, and people, present in the image. However, challenges arise when objects are occluded, or the scene is cluttered, which can result in errors in the counting process. In the bottom left of Figure 25, GPT-4V identifies 12 people, but the correct answer should be 11. This may be due to our limited text prompt used in this experiment, and further investigation in prompting techniques is needed.

Object localization. Object localization [146, 72, 48] is a fundamental challenge in the field of computer vision. In our preliminary experiments, we address this task by utilizing a simple text prompt, “Localize each person in the image using a bounding box.” The initial results of our object localization experiments are depicted in Figure 26. The findings suggest that GPT-4V demonstrates the capability to generate bounding box coordinates in textual format. However, it is important to note that the generated bounding box coordinates are not accurate. Promising localization results are observed when the scene or background is relatively simpler and less cluttered. Further prompting techniques are required to enhance object localization performance in more complex and crowded environments.

Dense captioning. Dense captioning [59, 79] involves generating detailed description for each region of interest in the given image. This advanced task in vision-language field typically requires a complex system that integrates multiple experts, such as object detector, celebrity recognition model, and image captioning model. In order to explore GPT-4V’s capabilities in dense captioning, we use an instructional prompt, as shown in Figure 27. The results are highly encouraging, as GPT-4V successfully localizes and recognizes the individuals within the image, and then provides concise descriptions for each scientist.

Sec. 4.2 Object Localization, Counting, and Dense Captioning

Spatial Relationship Understanding

Prompt:

What is the spatial relation between the frisbee and the man?



GPT-4V:

Based on the image provided, the spatial relation between the frisbee and the man is that the frisbee is above and slightly in front of the man. It also appears to be moving away from the man, as if it has been thrown or released by him.

Prompt:

Does the person bigger than the car?



GPT-4V:

No, the car appears to be bigger than the person in this image. However, the perspective and distance of the objects in the photo may affect their perceived size.

Figure 24: Results on spatial relationship understanding. GPT-4V recognizes the spatial relationship between the objects in the images. Check Section 4.2 for detailed discussions.

Sec. 4.2 Object Localization, Counting, and Dense Captioning

Object Counting

Prompt:

Count the number of apples in the image.

**GPT-4V:**

There are 7 apples in the image.

Prompt:

Count the number of apples in the image.

**GPT-4V:**

There are 16 apples in the image.

Prompt:

Count the number of oranges in the image.

**GPT-4V:**

There are 3 oranges in the image.

Prompt:

Count the number of people in the image.

**GPT-4V:**

There are 3 people in the image.

Prompt:

Count the number of people in the image.

**GPT-4V:**

There are 4 people in the image.

Prompt:

Count the number of people in the image.

**GPT-4V:**

There are 12 people in the image.

Figure 25: Results on object counting. GPT-4V is able to determine the quantity of the specified objects in the image. Red highlights the wrong answer. Check Section 4.2 for detailed discussions.

Sec. 4.2 Object Localization, Counting, and Dense Captioning

Object Localization

Prompt:

Localize each person in the image using bounding box. The image size is (width, height) = (998, 786)



GPT-4V:

There is only one person in the image, so the bounding box for this person would be:

- Top left: (254, 118)
- Bottom right: (731, 693)

Explanation: The top left point of the bounding box is at $(x, y) = (254, 118)$, which is the top left corner of the person in the image. The bottom right point of the bounding box is at $(x, y) = (731, 693)$, which is the bottom right corner of the person in the image. This bounding box encompasses the entire person in the image.

Visualization of Detected Bbox:



Prompt:

Localize each car in the image using bounding box. What is the image size of the input image?



GPT-4V:

The image size of the input image is 500 x 281 pixels.

Bounding box for the black car:

- Top left corner: (63, 114)
- Bottom right corner: (196, 220)

Bounding box for the white car:

- Top left corner: (287, 96)
- Bottom right corner: (474, 220)

Visualization of Detected Bbox:



Prompt:

Localize each person in the image using bounding box. What is the image size of the input image?



GPT-4V:

The image size of the input image is 1000 x 333 pixels.

Bounding boxes for each person:

- Person 1 (leftmost): $(x_1, y_1) = (22, 108), (x_2, y_2) = (170, 300)$
- Person 2: $(x_1, y_1) = (200, 80), (x_2, y_2) = (290, 270)$
- Person 3: $(x_1, y_1) = (350, 70), (x_2, y_2) = (460, 280)$
- Person 4: $(x_1, y_1) = (540, 80), (x_2, y_2) = (640, 270)$
- Person 5 (rightmost): $(x_1, y_1) = (760, 100), (x_2, y_2) = (920, 290)$

Note: These bounding box coordinates are approximate and may not be exact.

Visualization of Detected Bbox:

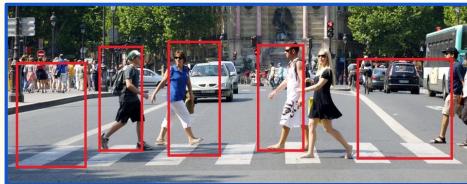


Figure 26: Results on object localization. GPT-4V is able to generate and approximate the bounding box coordinates for the specified objects in the image. When providing a simple text prompt only, the model may encounter challenges when dealing with more complex scenarios like object occlusions and cluttered scenes. Check Section 4.2 for detailed discussions.

Sec. 4.2 Object Localization, Counting, and Dense Captioning

Dense Captioning

Prompt:

Please follow the instructions

1. Tell me the size of the input image;
2. Localize each person in the image using bounding box;
3. Recognize each person;
4. Generate detailed caption for each bounding box.



GPT-4V:

1. The size of the input image is 1600 x 1067 pixels.

2. Bounding box 1: $(x_1, y_1) = (120, 250)$, $(x_2, y_2) = (420, 950)$ - Person 1
Bounding box 2: $(x_1, y_1) = (500, 240)$, $(x_2, y_2) = (800, 940)$ - Person 2
Bounding box 3: $(x_1, y_1) = (880, 230)$, $(x_2, y_2) = (1180, 930)$ - Person 3
Bounding box 4: $(x_1, y_1) = (1260, 220)$, $(x_2, y_2) = (1560, 920)$ - Person 4

3. Person 1: Yann LeCun

Person 2: Geoffrey Hinton

Person 3: Yoshua Bengio

Person 4: Andrew Ng

4. Caption for bounding box 1: Yann LeCun, computer scientist and one of the founding fathers of artificial intelligence.

Caption for bounding box 2: Geoffrey Hinton, computer scientist and cognitive psychologist, known for his work on artificial neural networks.

Caption for bounding box 3: Yoshua Bengio, computer scientist and co-founder of the Montreal Institute for Learning Algorithms.

Caption for bounding box 4: Andrew Ng, computer scientist and entrepreneur, known for his work on deep learning and artificial intelligence.

Visualization of Dense Captioning:

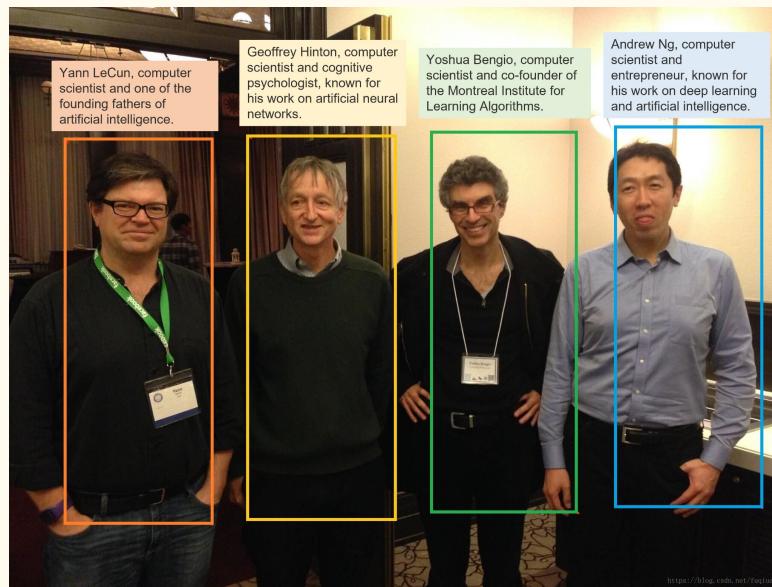


Figure 27: Results on dense captioning. GPT-4V follows the text prompt and successfully generates dense captions for the input image. Check Section 4.2 for detailed discussions.

4.3 Multimodal Knowledge and Commonsense

Joke and meme. Jokes and memes often reference specific events, pop culture, or Internet trends. Understanding these references requires being familiar with the relevant context and cultural knowledge. Grasping the visual elements, their relationship to the text, and the intended humorous effect can be a complex task [94]. Moreover, memes are often user-generated, making them highly diverse and ever-expanding. To evaluate GPT-4V’s ability in this domain, we input a pair of meme and text prompt to GPT-4V. The example text prompts include “Can you explain the meme?” and “What is funny about the image?” Figure 28 shows the example results. We observe that GPT-4V has remarkable ability to gather information from both visual and textual modalities, and then comprehend the humor embedded within memes.

Science and knowledge. We further investigate GPT-4V’s capability in tasks that requires reasoning with scientific knowledge [80]. We conduct experiments by providing a text prompt question and a corresponding image. The questions cover a wide range of topics, including geography, physics, biology, and earth science. In Figures 29-31, we observe that GPT-4V is able to correctly answer the science questions based on the visual context. For instance, in the bottom row of Figure 29, GPT-4V recognizes the average particle speed for both sample A and sample B. By considering the relationship among particle speed, kinetic energy, and temperature, GPT-4V answers the question correctly. For another instance, as shown in the bottom row of Figure 30, GPT-4V takes into account the visual arrows presented in the figure to identify the producer in the specific food web. Moreover, as shown in Figure 31, when we provide a more specific prompt, such as “Suppose you are a teacher, please use the figure to explain X,” we observe the generated answer adopts a tutorial format and explains the subject step by step.

Multimodal commonsense. In Figure 32, we access the ability of GPT-4V in multimodal common-sense reasoning [141, 49]. In our experiments, we observed that GPT-4V effectively utilizes the bounding boxes presented in the image as visual prompts (e.g., [person1] and [person2]) to recognize the actions performed by the individuals. As shown in the second example in Figure 32, based on the formal dress worn by [person1] and [person2] and the floral decorations present in the scene, it can be inferred that they are attending a wedding ceremony. Moreover, when we provide a more specific input prompt, such as “Suppose you are a detective, what can you infer from the visual clues?”, GPT-4V demonstrates the ability to discern numerous nuanced visual cues within the image and offers a list of plausible hypotheses.

Sec. 4.3 Multimodal Knowledge and Commonsense

Joke and Meme

Prompt:

Can you explain this meme?

Me: I'll do it at 8
Time: 8.05
Me: looks like I gotta wait till 9 now



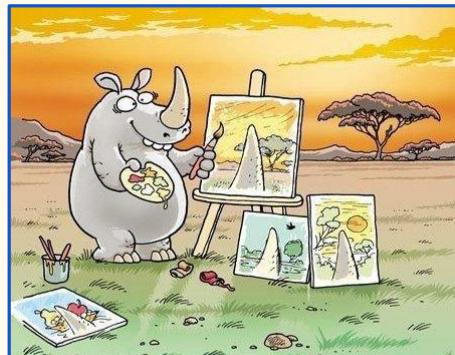
boredpanda.com

GPT-4V:

This meme is poking fun at the idea of procrastination and how someone may put off doing a task until a certain time, but if they miss that time, they will push it back even further. In this case, the person in the meme planned to do something at 8 o'clock, but since it is now 8:05, they have decided to wait until 9 o'clock to do it. The image of Kermit the Frog relaxing in bed represents the person's desire to procrastinate and avoid the task.

Prompt:

What is funny about this image?



You see the world
as you are!

GPT-4V:

There are a couple of things that people might find funny about this image:
- It is a cartoon of a rhinoceros painting, which is an unexpected and silly concept in itself.
- The rhinoceros is painting a picture of itself, which adds an extra layer of humor.
- The caption, "You see the world as you are!", is a joke about how we all have our own unique perspectives and can be a little self-centered at times.

Prompt:

Can you explain this meme?

My Handwriting In Exams

First two pages

Global warming is definitely a very huge issue in front of this entire world. It's causing several disturbances in the

Middle pages

Also it can't be written well of gloom and mushy so we make pallid and that makes the

Last two pages



GPT-4V:

This meme highlights the deterioration of a person's handwriting during an exam. The first two pages show neat and legible handwriting, the middle page shows slightly messier handwriting, and the last two pages show almost illegible handwriting. The meme is poking fun at how, as an exam progresses, a person's handwriting tends to get worse as they become more tired and rushed to finish the exam. It also implies that the person's focus and attention to detail decrease towards the end of the exam.

Prompt:

What is funny about this image?



GPT-4V:

This image is funny for a couple of reasons. Firstly, it is a meme template that has been widely used and recognized on the internet. The left side of the image is a screenshot from the reality TV show "The Real Housewives of Beverly Hills" where one of the cast members is angrily pointing and yelling. The right side of the image is a picture of a cat sitting at a dinner table with a plate of food in front of it, looking unimpressed. The humor in this image comes from the contrast between the two sides - the angry, dramatic yelling on the left and the nonchalant cat on the right. It also plays on the idea of cats acting like humans and being unimpressed by our behavior. Overall, the image creates a funny and absurd situation that makes people laugh.

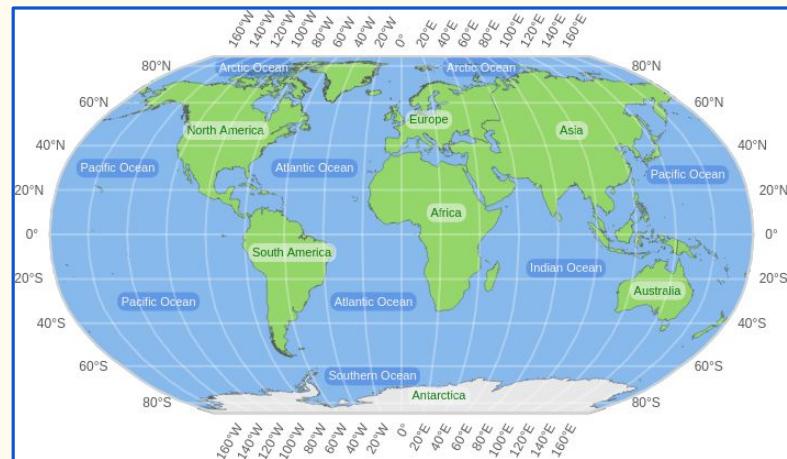
Figure 28: Results on joke and meme understanding. GPT-4V demonstrates the impressive capability to comprehend the humor embedded within memes. Check Section 4.3 for detailed discussions.

Sec. 4.3 Multimodal Knowledge and Commonsense

Science and Knowledge

Prompt:

Which of these oceans does the prime meridian intersect?



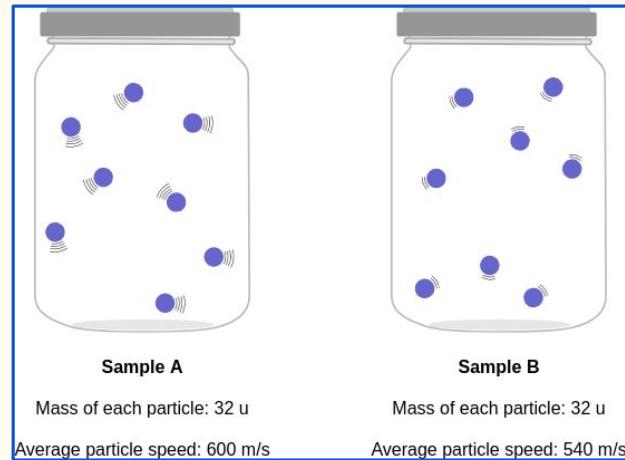
GPT-4V:

The prime meridian intersects the following oceans:

- Atlantic Ocean
- Southern Ocean
- Arctic Ocean

Prompt:

Compare the average kinetic energies of the particles in each sample. The diagrams below show two pure samples of gas in identical closed, rigid containers. Each colored ball represents one gas particle. Both samples have the same number of particles. Which sample has the higher temperature? Sample A or sample B?



GPT-4V:

Because the average particle speed in sample A is higher than in sample B (600 m/s vs. 540 m/s), the particles in sample A have a higher average kinetic energy. Therefore, sample A has a higher temperature than sample B.

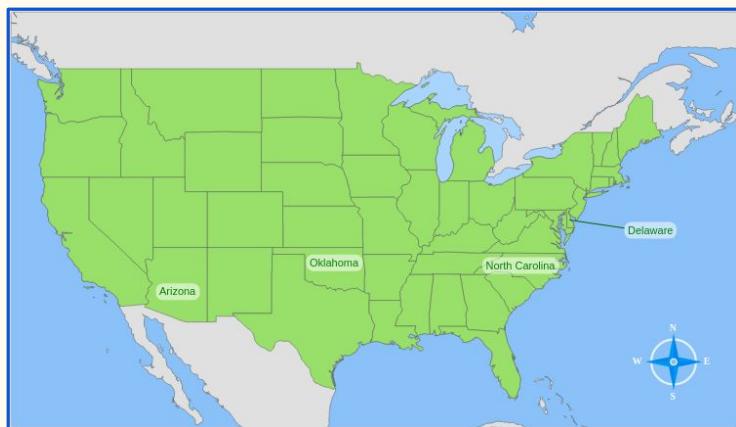
Figure 29: Results on answering science questions. GPT-4V can understand the question textually and visually, and gather necessary information to answer the question. Check Section 4.3 for detailed discussions.

Sec. 4.3 Multimodal Knowledge and Commonsense

Science and Knowledge

Prompt:

Which of these states is farthest north?

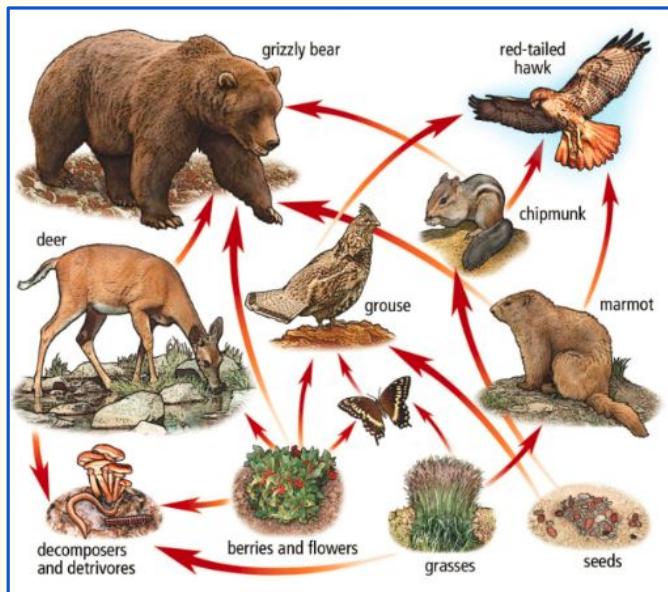


GPT-4V:

Out of the three states shown in the map - Arizona, Oklahoma, and Delaware, Delaware is the farthest north.

Prompt:

Which of the organisms is the producer in this food web?



GPT-4V:

The producers in this food web are the berries and flowers, grasses, and seeds. These are the organisms that use energy from the sun to produce their own food through photosynthesis.

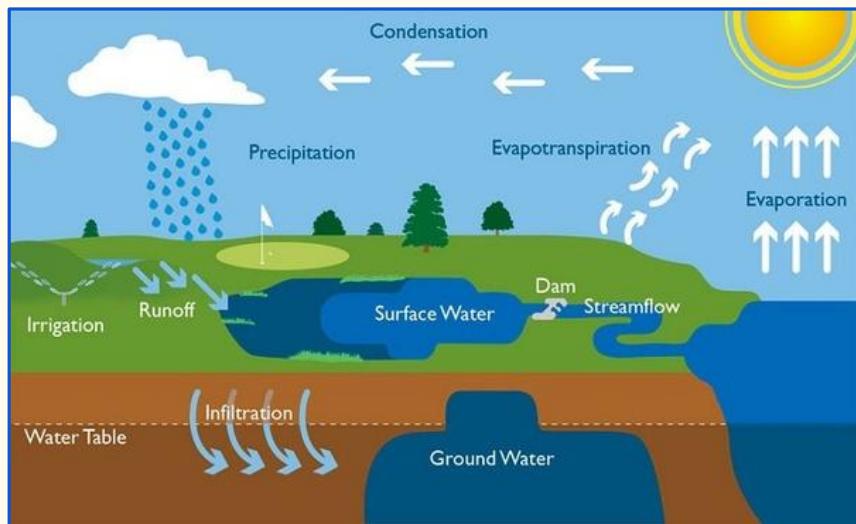
Figure 30: Results on answering science questions. GPT-4V can understand the question textually and visually, and gather necessary information to answer the question. Check Section 4.3 for detailed discussions.

Sec. 4.3 Multimodal Knowledge and Commonsense

Science and Knowledge

Prompt:

Suppose you are a teacher, please use the figure to explain the distinction between evaporation and evapotranspiration.



GPT-4V:

Okay class, today we are going to learn about the distinction between evaporation and evapotranspiration. Let's take a look at this figure.

First, let's understand what evaporation is. Evaporation is the process by which water is changed from its liquid state to its gaseous state. This usually takes place at the surface of a liquid, like the ocean or a lake, when the water molecules gain enough energy to break away and form water vapor. So in this figure, we can see that the water in the lake and the dam is evaporating and becoming water vapor.

Now, let's take a look at evapotranspiration. This is a combination of two processes - evaporation and transpiration. Transpiration is the process by which plants release water vapor through their leaves. So, in this figure, we can see that the trees and plants are releasing water vapor into the air, which is the transpiration part of evapotranspiration. At the same time, the water in the soil and the surface is also evaporating. This combination of evaporation and transpiration is what we call evapotranspiration.

In summary, evaporation is the process by which water is changed from its liquid state to its gaseous state, while evapotranspiration is the combination of evaporation and transpiration. Both these processes contribute to the water cycle and are important for maintaining the balance of water in our environment.

Figure 31: Results on answering science questions. When we use a more specific text prompt like "Suppose you are a teacher, please use the figure to explain X," we observe that GPT-4V can generate a short tutorial for explaining the subject. Check Section 4.3 for detailed discussions.

Sec. 4.3 Multimodal Knowledge and Commonsense

Multimodal Commonsense

Prompt:

What is [person3] doing?



GPT-4V:

[person3] is carrying a plate of food, likely serving it to the customers at the table. It appears that they are working as a waiter or server in a restaurant.

Prompt:

What are [person1] and [person2] doing?



GPT-4V:

[person1] and [person2] are walking down the aisle at a wedding ceremony.

Prompt:

Suppose you are a detective, what can you infer from the visual clues in the image?



GPT-4V:

From the visual clues in the image, I can infer that:

- The room belongs to someone who is likely a young adult or teenager, as evidenced by the casual clothing and the desk with a computer.
- The person is likely a student or someone who works from home, as evidenced by the desk and computer.
- The person is likely fashion-conscious, as evidenced by the multiple jackets and shoes.
- The room is likely located in a colder climate, as evidenced by the jackets and boots.
- The person is likely not very tidy, as evidenced by the cluttered desk and the clothes hanging on the door.
- The room is likely located in an older or more rustic building, as evidenced by the exposed pipes and the worn-looking walls.

Figure 32: Results on multimodal commonsense reasoning. Check Section 4.3 for detailed discussions.

4.4 Scene Text, Table, Chart, and Document Reasoning

Scene text recognition. Reading and understanding scene text in images is an important task in vision-language [111–113, 16]. In our experiments, we investigate GPT-4V’s ability to recognize scene text by utilizing the input prompt “What are all the scene text in the image?” Figure 33 shows the example results. We observe GPT-4V accurately identifies scene text in various scenarios, including both handwritten and printed text. In Section 4.5, we present further results on multilingual scenarios.

Visual math reasoning. In Figure 34, GPT-4V demonstrates its capability in solving visual math problems. In our experiments, we observe GPT-4V is able to extract essential information from the image. For instance, in Figure 34, GPT-4V correctly identifies the presence of a right triangle (or orthogonal triangle) and determines that AB is 4 units and BC is 3 units. In addition, we note that GPT-4V tends to present solutions in a well-structured manner, solving the problem step by step, thereby showcasing its ability to provide clear explanations.

Chart understanding and reasoning. We further study GPT-4V’s ability in chart understanding and reasoning. Figures 35–37 show the example results. In our preliminary explorations, GPT-4V exhibits the ability to provide detailed descriptions of charts. For example, in Figure 35, the model correctly explains the proposal process from the beginning to the end. In Figure 36, the model not only understands the program in the given flow chat, but also translates the details to a python code. In the bottom row of Figure 37, GPT-4V shows a clear understanding of both x- and y-axis, and explains the key insight presented in the chart. Furthermore, in our experiments, we observe that GPT-4V can answer questions based on the chart. In the top row of Figure 37, GPT-4V correctly calculates the average total fueling cost, excluding the Ford F150.

Table understanding and reasoning. In Figure 38, we present our preliminary investigations into table understanding and reasoning. Similar to the findings from chart experiments, GPT-4V shows promising results in understanding the details in the table, as well as in reasoning and accurately responding to related questions.

Document understanding. Figure 39 shows the results on various types of documents, such as floor plan, poster, and exam paper. We observe GPT-4V demonstrates an understanding of the documents and provides reasonable responses. For instance, it accurately identifies the location of the bathroom for the second bedroom in the floor plan. It also recognizes the Chinese dish “Hot dry noodles,” and associates it with the city of Wuhan by following the scene text. Moreover, GPT-4V is capable of reading an exam paper. It accurately reconstructs the table in Markdown, and then fills in the table with the correct answers. We present more explorations in its coding ability in Section 4.6.

In Figure 40, we explore a more challenging case by providing a technical report [119] with multiple pages as input. In our limited exploration, GPT-4V exhibits impressive results. It correctly describes the main idea and their proposed method by considering the context across multiple pages. However, it may occasionally miss some implementation details. Please note that the dataset should contain $1196+665=1861$ examples, and the extracted features should include Histograms of Oriented Gradients (HOG). Instead of prompting all pages to the model simultaneously, we believe that exploring more advanced prompting techniques, such as thinking step-by-step or employing in-context few-shot approaches, could potentially enhance the model’s performance.

Sec. 4.4 Scene Text, Table, Chart, and Document Reasoning

Scene Text Recognition

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image is "3,642,039,031,055".

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image are:

- (NVDA)
- NVIDIA
- 356.85
- +51.47
- +16.85%

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image are:

- "Royal"
- "London"

Prompt:

What are all the scene text in the image?



GPT-4V:

The only scene text in the image is "The Earth without art is just 'eh'".

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image is "The Road to Hollywood: how some of us got here"

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image are:

- "Westfield"
- "westfield.com.au"

Prompt:

What are all the scene text in the image?



GPT-4V:

There are three texts in the image:

- 1) "COVID-19 TESTING"
- 2) "PLEASE HAVE ID & INSURANCE CARD READY"
- 3) "W @UW Medicine" (this is the logo/branding for the University of Washington Medicine

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image are:

- Vacation
- Beach
- Smile
- Summer
- Watermelon
- Flip flops
- Lemonade

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image are:

- Malayan Tapir
- Tapirus indicus
- LOCATION: Malay Peninsula and Sumatra, near water
- RELATIVES: There are only five species of tapirs. They are most closely related to horses and rhinos.
- VANISHING? The forest provides tapirs with shelter, food, and water, but forests are rapidly disappearing in Malaysia and Sumatra.

Figure 33: Results on scene text recognition. GPT-4V can recognize scene text in many challenging scenarios. Check Section 4.4 for detailed discussions.

Sec. 4.4 Scene Text, Table, Chart, and Document Reasoning

Visual Math

Prompt:

What is x in the equation?

$$(x+3)^2 = 4$$

GPT-4V:

Without any context or information about what the equation is representing, we can only assume that "x" is a variable or unknown value in the equation.

In this equation, $(x + 3)$ is being squared and is equal to 4. To solve for x , we can take the square root of both sides of the equation to get rid of the exponent on the left side. This would give us:

$$\sqrt{(x+3)^2} = \sqrt{4}$$

$$x + 3 = \pm 2$$

From here, we can solve for x by subtracting 3 from both sides. This would give us:

$$x = -1 \text{ or } x = 5$$

So, in this equation, x could be either -1 or 5.

Prompt:

What is the value of d in the last equation? What is the value of x in the image?

Solve the following equations:

$$1) 8x + 11 = 4x + 14$$

$$2) 7d - 4 = 11d - 9$$

GPT-4V:

1) To solve the first equation, we need to isolate the x variable on one side of the equation and the constants on the other side. We can start by subtracting $4x$ from both sides to get rid of the x term on the right side of the equation. This gives us:

$$4x + 11 = 14$$

Next, we can subtract 11 from both sides to get rid of the constant term on the left side of the equation. This gives us:

$$4x = 3$$

Finally, we can divide both sides by 4 to isolate the x variable on the left side of the equation. This gives us:

$$x = 3/4$$

2) To solve the second equation, we can start by subtracting $7d$ from both sides to get rid of the d term on the left side of the equation. This gives us:

$$-4 = 4d - 9$$

Next, we can add 9 to both sides to get rid of the constant term on the right side of the equation. This gives us:

$$5 = 4d$$

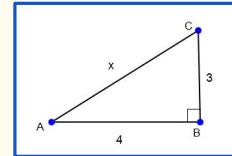
Finally, we can divide both sides by 4 to isolate the d variable on the right side of the equation. This gives us:

$$d = 5/4$$

So, the value of d in the last equation is $5/4$.

Prompt:

What is the value of x in the image?



GPT-4V:

The value of x in the image can be found using the Pythagorean theorem, which states that in a right triangle, the square of the hypotenuse (the longest side, opposite the right angle) is equal to the sum of the squares of the other two sides.

In this case, the hypotenuse is the side AC , and the other two sides are AB and BC . We know that AB is 4 and BC is 3, so we can plug these values into the equation:

$$a^2 + b^2 = c^2$$

$$3^2 + 4^2 = c^2$$

$$9 + 16 = c^2$$

$$25 = c^2$$

Therefore, c (the length of the hypotenuse) is 5. Since x is the length of the hypotenuse, x is equal to 5.

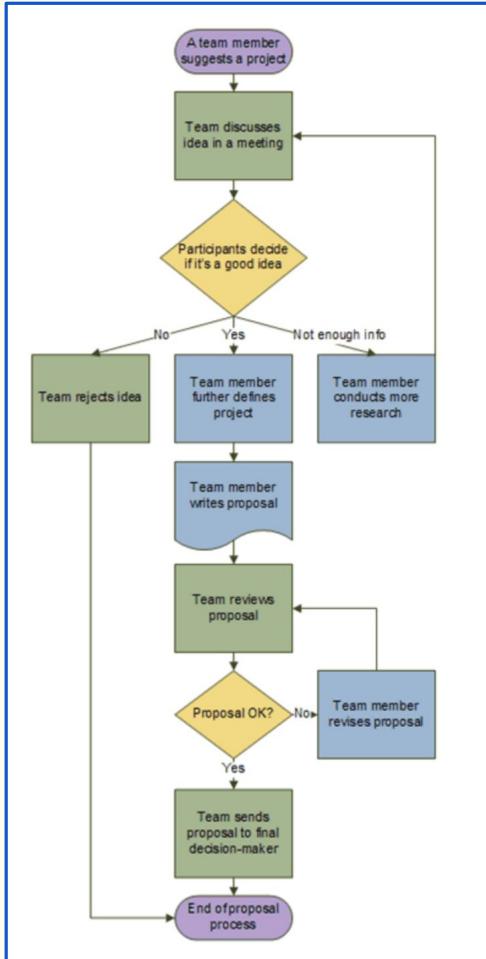
Figure 34: Results on visual math reasoning. GPT-4V is able to comprehend and solve visual math problems with a well-structured solution. Check Section 4.4 for detailed discussions.

Sec. 4.4 Scene Text, Table, Chart, and Document Reasoning

Flow Chart Understanding

Prompt:

Explain the figure



GPT-4V:

This figure represents a flow chart of a team's proposal process. The process begins with a team member suggesting a project in a meeting. The team then discusses whether it is a good idea or not. If the idea is not considered a good one, the team rejects it and the process ends.

If the team thinks it might be a good idea but needs more information, a team member conducts further research and refines the idea into a proposal. The team then reviews the proposal and provides feedback. If the proposal is not okay, then the team member revises it and the team reviews it again.

If the proposal is okay, the team sends it to a final decision-making process. If the proposal is approved, then the team proceeds with the project. This is the end of the proposal process.

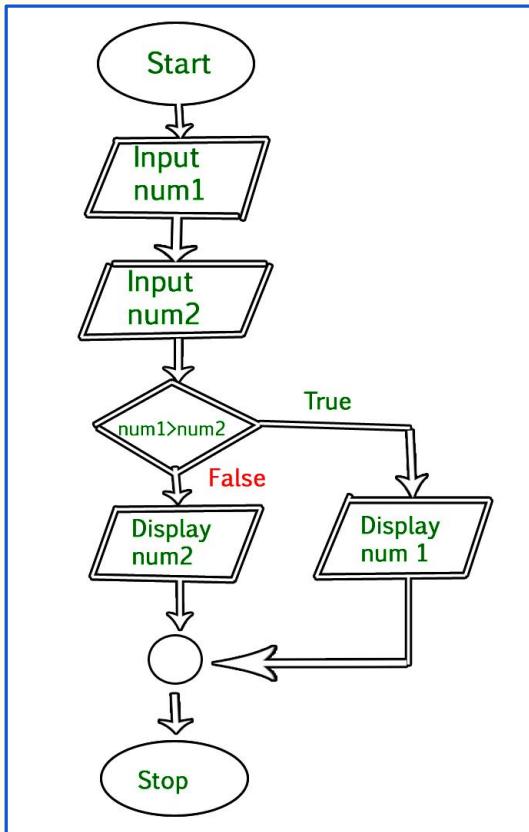
Figure 35: Results on flow chart understanding. GPT-4V correctly describes the proposal process in details. Check Section 4.4 for detailed discussions.

Sec. 4.4 Scene Text, Table, Chart, and Document Reasoning

Flow Chart Understanding

Prompt:

Can you translate the flowchart to a python code?



GPT-4V:

Sure, here's the Python code:

```
num1 = input("Enter first number: ") num2 = input("Enter second number: ") if num1 > num2:  
print(num1) else: print(num2)
```

Explanation:

- The first two lines correspond to the “Input num1” and “Input num2” blocks in the flowchart, where the user is prompted to enter two numbers.
- The next two lines correspond to the “num1 > num2?” block, where a conditional statement is used to check if the first number is greater than the second number.
- If the condition is true, the next line (“Display num1”) is executed, and the first number is printed to the screen.
- If the condition is false, the else statement is executed, and the second number is printed to the screen (“Display num2”).
- The code then stops, corresponding to the “Stop” block in the flowchart.

Figure 36: Results on flow chart understanding. GPT-4V is able to translate the flow chart to a python code. Check Section 4.4 for detailed discussions.

Sec. 4.4 Scene Text, Table, Chart, and Document Reasoning

Chart Understanding and Reasoning

Prompt:

What is the average total fueling cost excluding the Ford F150?



GPT-4V:

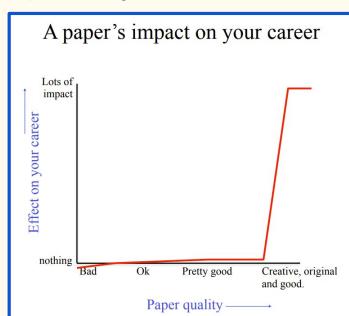
To find the average total fueling cost excluding the Ford F150, we need to add the total fueling costs of the Toyota RAV4 and Honda Civic and divide by 2 (since there are 2 vehicles in the calculation).

$$(82.56 + 70.55) \div 2 = 76.55$$

Therefore, the average total fueling cost excluding the Ford F150 is \$76.55.

Prompt:

Explain the figure.



GPT-4V:

The figure is a graph that shows the relationship between the quality of a paper and its impact on a person's career. The x-axis represents the quality of the paper, which ranges from bad to creative and original. The y-axis represents the impact the paper has on a person's career, which ranges from nothing to lots of impact.

The graph shows that a bad paper has little to no impact on a person's career, while a creative and original paper has a significant impact. The impact of the paper on a person's career increases as the quality of the paper improves. This suggests that the quality of a person's work is an important factor in determining their success in their career.

Figure 37: GPT-4V shows promising results in understanding the details in the chart, as well as in reasoning and accurately responding to related questions. Check Section 4.4 for detailed discussions.

Sec. 4.4 Scene Text, Table, Chart, and Document Reasoning

Table Understanding and Reasoning

Prompt:

How many families are earning more than 13000 and owns more than 2 cars?

| Monthly income (in Rs.) | Vehicles per family | | | |
|-------------------------|---------------------|-----|----|---------|
| | 0 | 1 | 2 | Above 2 |
| Less than 7000 | 10 | 160 | 25 | 0 |
| 7000 – 10000 | 0 | 305 | 27 | 2 |
| 10000 – 13000 | 1 | 535 | 29 | 1 |
| 13000 – 16000 | 2 | 469 | 59 | 25 |
| 16000 or more | 1 | 579 | 82 | 88 |

GPT-4V:

From the given table, we can clearly see that there are a total of 113 families that earn more than Rs. 13000 and own more than 2 cars (25 families in the 13000-16000 range and 88 families in the 16000 or more range).

Prompt:

Which country has the highest profit margin?

| Country | Sales Volume | Revenue | Profit | Profit Margin |
|-------------|--------------|--------------|--------------|---------------|
| USA | 40.080 | \$15.971.880 | \$3.086.421 | 19,3% |
| China | 35.070 | \$15.866.670 | \$3.032.162 | 19,1% |
| Australia | 27.054 | \$14.812.566 | \$2.868.636 | 19,4% |
| India | 23.046 | \$10.608.174 | \$1.853.710 | 17,5% |
| South Korea | 16.032 | \$10.494.948 | \$1.975.844 | 18,8% |
| Total / Avg | 141.282 | \$67.754.238 | \$12.816.772 | 18,8% |

GPT-4V:

Based on the data provided, the country with the highest profit margin is Australia, with a profit margin of 19,4%. However, the difference between the top three countries (USA, China and Australia) is minimal, with less than 0,3% difference in profit margin.

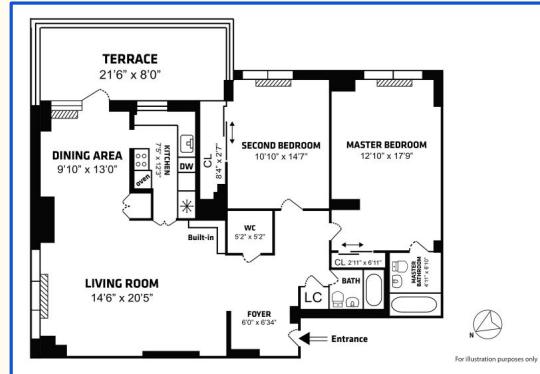
Figure 38: We observe GPT-4V can understand the details in the table, and answer related questions. Check Section 4.4 for detailed discussions.

Sec. 4.4 Scene Text, Table, Chart, and Document Reasoning

Document Understanding

Prompt:

Describe the image.



GPT-4V:

This image is a floor plan for an apartment or condominium. It shows the layout of the rooms and the dimensions. There is a terrace, dining area, living room, foyer, two bedrooms, a master bathroom, and a guest bathroom. The floor plan is black and white and is a two-dimensional drawing.

Prompt:

Answer the question in the image.

- Figure 1 shows the structure of the earth's interior. Write the number showing each layer in the table below.

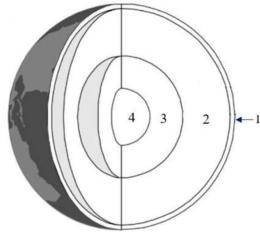


Figure 1

| Layer | Number on figure 1 |
|------------|--------------------|
| Outer core | |
| Mantle | |
| Crust | |
| Inner core | |

(4)

GPT-4V:

Layer | Number on figure 1

--- | ---

Outer core | 3

Mantle | 2

Crust | 1

Inner core | 4

Visualization of the generated results:

| Layer | Number on figure 1 |
|------------|--------------------|
| Outer core | 3 |
| Mantle | 2 |
| Crust | 1 |
| Inner core | 4 |

Prompt:

Where is the bathroom for the second bedroom?



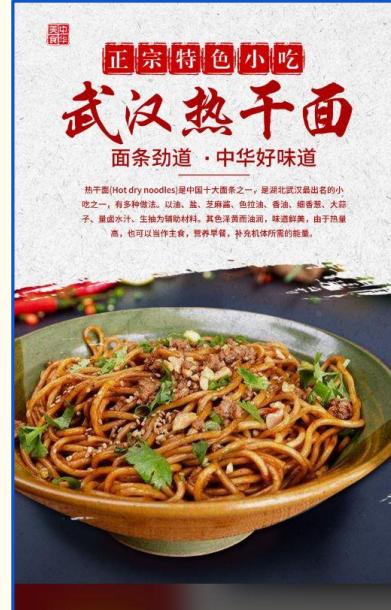
Note: Same query image as shown on left, resized to save space

GPT-4V:

According to this floor plan, the bathroom for the second bedroom appears to be located directly adjacent to the bedroom. It is labeled "WC" and is a small room measuring 5'2" x 5'2".

Prompt:

Describe the poster. Which city is this dish from?



GPT-4V:

The poster is an advertisement for a Chinese dish called "Hot dry noodles." The poster has a red background with white Chinese characters at the top that read "Hot dry noodles, the taste of China." Below the text is a photo of the dish, which consists of noodles with meat and vegetables on top. The dish is served in a green bowl with red chili peppers on the side. The text at the bottom of the poster is in Chinese and describes the dish as being spicy, fragrant, and delicious.

This dish is from the city of Wuhan, located in the Hubei Province of China. It is a popular street food in the city and is considered one of the "10 great breakfasts" in China.

Figure 39: Results on document understanding. GPT-4V recognizes three different types of document and answers the questions correctly. Check Section 4.4 for detailed discussions.

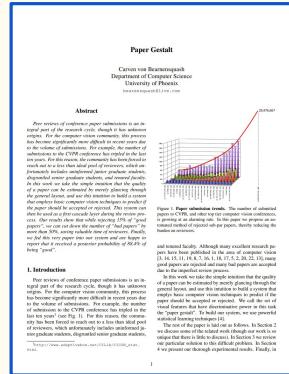
Sec. 4.4 Scene Text, Table, Chart, and Document Reasoning

Document Understanding

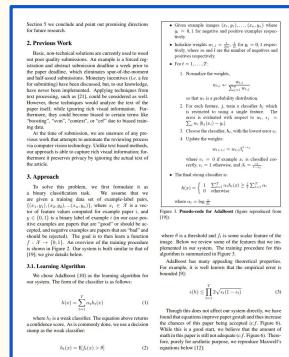
Prompt:

This is a paper "Paper Gestalt." Describe the paper in details and highlight their contribution.

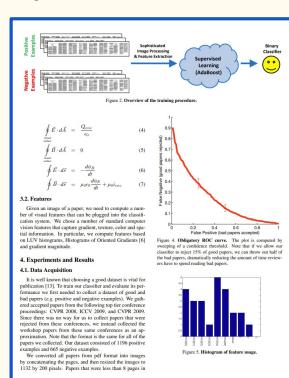
Page 1



Page 2

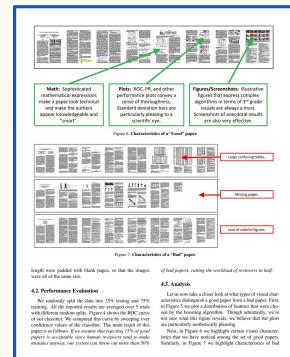


Page 3

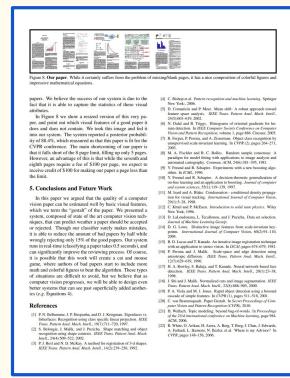


Prompt Continued

Page 4



Page 5



GPT-4V:

The paper "Paper Gestalt" is a research paper written by Carven von Bearmensquash from the Department of Computer Science at the University of Phoenix. The paper discusses the visual quality of a computer science research paper and how it affects the overall impression of the paper.

The paper begins with an introduction to the concept of paper gestalt, which is the overall impression of a paper. The author argues that the visual quality of a paper is an important factor in creating a good impression. The paper then discusses previous work in this area and highlights the contribution of this paper, which is the use of a supervised learning algorithm to evaluate the visual quality of a paper.

The paper then presents the methodology and experimental results. The author collected a dataset of 1196 computer science research papers and used a supervised learning algorithm to evaluate the visual quality of the papers. The algorithm was trained on a set of features extracted from the papers, such as the number of figures, the number of equations, and the number of references. The results showed that the algorithm was able to accurately predict the visual quality of a paper with a high degree of accuracy.

The paper concludes with a discussion of the results and future work. The author suggests that future work could focus on improving the algorithm by incorporating more features and training it on a larger dataset. The paper also suggests that the algorithm could be used to evaluate the visual quality of other types of documents, such as technical reports and dissertations.

Overall, the contribution of this paper is the use of a supervised learning algorithm to evaluate the visual quality of a computer science research paper. This is a novel approach that could help researchers improve the visual quality of their papers and make a better impression on their readers.

Figure 40: Results on document understanding. GPT-4V reads a multi-page technical report, understands the content in each section, and provides a summary of the contribution of this technical report.

Red highlights the wrong answer. Check Section 4.4 for detailed discussions.

4.5 Multilingual Multimodal Understanding

We assess GPT-4V’s ability in comprehending multiple languages and modalities. First, we explore this capability by evaluating natural images without scene text, as depicted in Figure 41. In the first row of the figure, we provide the input text prompt “Describe the image” in Chinese, French, and Czech, respectively. GPT-4V recognizes the input text prompts in different languages, and generates correct image descriptions in corresponding languages. In the second row of Figure 41, we provide the input text prompt in English and specify the output language. GPT-4V follows the instruction and generates correct descriptions in the desired languages. In the bottom row of Figure 41, we provide an input prompt in Spanish, and ask GPT-4V to generate image descriptions in 20 different languages. We observe that GPT-4V can process both the input and output text in different languages.

Furthermore, we explore a scenario involving multilingual scene text recognition, where the input image may contain scene text in various languages. As shown in Figure 42, GPT-4V correctly identifies and understands the scene text from different scenes. As shown in the first two rows of Figure 43, we observe that GPT-4V can recognize the scene text, and translate it to a different language. In the bottom row of Figure 43, we provide a screenshot of a Wikipedia website which is in Catalan, and instruct GPT-4V to summarize the information in 20 different languages. GPT-4V not only recognizes the text in Catalan but also generates precise summaries and translates them into different languages. This showcases GPT-4V’s ability to comprehend and translate multilingual scene text.

We also explore the capability of multicultural understanding [140, 73]. Figure 44 shows the example results in such a scenario. We observe that GPT-4V is capable of understanding cultural nuances and generating reasonable multilingual descriptions for the wedding images given.

In our exploration, we found that GPT-4V seamlessly comprehends and correctly generates descriptions in different languages, highlighting its versatility in handling diverse linguistic contexts.

Sec. 4.5 Multilingual Multimodal Understanding

Multilingual Image Description

Prompt:

描述這張圖片



GPT-4V:

這張圖片是一隻小貓在草地上走路的畫面。小貓是橘色的，有藍色的眼睛，看起來很可愛。它的身體很小，四肢纖細，正在小心翼翼地在草地上行走。

Prompt:

Décrire l'image



GPT-4V:

Il s'agit d'une photo d'un chaton roux marchant sur l'herbe verte. Le chaton a les yeux bleus et semble être en train de marcher avec prudence. Il est entouré d'herbe et on peut voir une partie d'un autre chat en arrière-plan.

Prompt:

Popište obrázek



GPT-4V:

Na obrázku je malé oranžové kotě s modrýma očima. Kotě stojí na zelené trávě a směřuje dopředu. Jeho tělo je v pohybu a vypadá, že právě běží. Na pozadí je vidět další část trávy a nejasně i další kočka.

Prompt:

Describe the image in Chinese

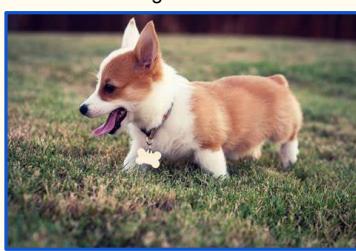


GPT-4V:

这是一只小柯基犬，它在草地上走着，嘴巴张开，舌头伸出来，好像很开心的样子。它的身上有一条项圈，上面挂着一个骨头形状的牌子。

Prompt:

Describe the image in French



GPT-4V:

C'est une image d'un petit chien de race Corgi. Il est de couleur marron et blanc et porte un collier avec une médaille en forme d'os. Il est en train de marcher sur l'herbe verte avec la langue pendante.

Prompt:

Describe the image in Czech



GPT-4V:

Na obrázku je malý hnědý pes s bílým límcem a kostičkou na obojku. Pes je na zelené trávě a má otevřenou tlamu. Vypadá, že se baví.

Prompt:

Describir la imagen en 20 idiomas diferentes



GPT-4V:

1. Inglés: This is a beautiful view of Machu Picchu in Peru.
2. Español: Esta es una hermosa vista de Machu Picchu en Perú.
3. Francés: C'est une belle vue de Machu Picchu au Pérou.

GPT-4V output continued

4. Alemán: Dies ist eine schöne Aussicht auf Machu Picchu in Peru.
5. Italiano: Questa è una bella vista di Machu Picchu in Perù.
6. Portugués: Esta é uma bela vista de Machu Picchu no Peru.
7. Ruso: Это красивый вид на Мачу-Пикчу в Перу.
8. Chino: 这是秘鲁马丘比丘的美丽景色。
9. Japonés: これはペルーのマチュピチュの美しい景色です。
10. Coreano: 이것은 페루의 마추 피추의 아름다운 경치입니다.
11. Hindi: यह पेरू के माचु पिच्चू का सुंदर दृश्य है।
12. Árabe: هذا منظر جميل لماشو بيتشو في بيرو.
13. Turco: Bu, Peru'daki Machu Picchu'nun güzel bir manzarasıdır.
14. Griego: Αυτή είναι μια όμορφη θέα του Μάτσου Πίτσου στο Περού.
15. Hebreo: הרים מארז'ין בפֵרֹו.
16. Sueco: Det här är en vacker vy av Machu Picchu i Peru.
17. Noruego: Dette er en vakker utsikt over Machu Picchu i Peru.
18. Danés: Dette er en smuk udsigt over Machu Picchu i Peru.
19. Holandés: Dit is een prachtig uitzicht op Machu Picchu in Peru.
20. Polaco: To piękny widok na Machu Picchu w Peru.

Figure 41: Results on multilingual image descriptions. GPT-4V is able to generate image descriptions in different languages. Check Section 4.5 for detailed discussions.

Sec. 4.5 Multilingual Multimodal Understanding

Multilingual Scene Text Recognition

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image is an Arabic phrase "بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ" (Bismillah al-Rahman al-Rahim). It translates to "In the name of Allah, the Most Merciful, the Most Compassionate". This phrase is commonly used by Muslims at the beginning of any action or speech as a way to seek blessings and guidance from God.

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image is "MUSEE DU LOUVRE" which is the name of the famous art museum in Paris, France.

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image are:
- "尖沙咀" (Chinese characters)
- "Tsim Sha Tsui" (English text)

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image are:
- WELCOME (English)
- bienvenida (Spanish)
- Bienvenue (French)
- Willkommen (German)
- benvenuto (Italian)

Prompt:

What are all the scene text in the image?



GPT-4V:

There are two scene texts in the image:
1) "SUPER NINTENDO WORLD"
2) "スーパーニンテンドーワールド" (this is the Japanese version of "Super Nintendo World")

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image are:
- G.P. FIM DA EUROPA
- ATLETISMO
- António Agostinho 1987
- Fernando Domingos 1988

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image are:
- ΑΘΗΝΑΙ-ΡΩΜΗ-ΠΙΟΝ ΙΑΝΕΤΡΟΝ (Athens-Rome-Rion Iapetron)
- 29 ΜΑΪΟΥ 1933 (29 May 1933)
- ΔΡΑΧ (Drachma)
- ΕΛΛΑΣ (Greece)
- ΣΠΑΕΡ. ΤΑΧΥΔΡ. ΣΥΓΚΟΙΝ. ΙΤΑΛΙΑ-ΕΛΛΑΔΟΣ-ΤΟΥΡΚΙΑ (Airmail. Express. Communication. Italy-Greece-Turkey)
- 120 (Value of the stamp in drachmas)

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image is "não é normal ter medo de andar sozinho na rua!!" which is in Portuguese and translates to "It is not normal to be afraid to walk alone on the street!!!"

Figure 42: Results on multilingual scene text recognition. GPT-4V can recognize scene text in different languages. Check Section 4.5 for detailed discussions.

Sec. 4.5 Multilingual Multimodal Understanding

Multilingual Multicultural Understanding

Prompt:

Describe the image in both English and the local language corresponding to the geolocation depicted in the image.



GPT-4V:

English: This image shows a couple dressed in traditional Chinese wedding attire, sitting in front of a wall of red roses and a gold Chinese character for "double happiness."

Chinese (Simplified): 这张图片展示了一对穿着中国传统婚礼服装的夫妇，坐在红色玫瑰和金色“双喜”汉字前面。(Zhè zhāng túpiàn zhǎnshíle yī duì chuānzhe zhōngguó chuántǒng hūnlǐ fúzhāng de fūfù, zuò zài hóngsè méiqū hé jīnsè "shuāng xǐ" hànzi qiánmian.)

Prompt:

Describe the image in both English and the local language corresponding to the geolocation depicted in the image.



GPT-4V;

English: This image shows a Japanese couple in traditional wedding attire. The man is wearing a black kimono with a white hakama and the woman is wearing a white kimono with a large headpiece.

Japanese: この画像は伝統的な結婚式の衣装を着た日本のカップルを示しています。男性は白い袴と黒い着物を着ており、女性は大きな頭飾りと白い着物を着ています。(Kono gazo wa dentō-teki na kekkon-shiki no ishō o kita Nihon no kappuru o shimeshiteimasu.
Dansei wa shiroi hakama to kuroi kimono o kite ori,
josei wa ōkina kazari to shiroi kimono o kiteimasu.)

Prompt:

Describe the image in both English and the local language corresponding to the geolocation depicted in the image.



GPT-4V:

English: The image shows a couple wearing traditional Ethiopian clothing, standing in a garden with a tree in the background. The woman is wearing a white dress with colorful patterns and the man is wearing a white suit with similar patterns.

Amharic (Ethiopia's official language): የዚያስና
አንቀጽናም በሁሉ ላይ የሰነድ በለምትና ጉባኤና በበር
ደንቅል፡፡ ሲተቀብ በየጊዜ ያለ አሳላጊሁር ላይ በለምትና እና
ወይምናም በመገባሁ አሳላጊሁር ላይ፡፡

Prompt:

11. TIPPI: Describe the image in both English and the local language corresponding to the geolocation depicted in the image.



GPT-4V:

English: This image shows a newlywed Indian couple sharing a kiss. The bride is wearing a traditional red sari and the groom is wearing a red turban and white suit. They are sitting on a decorated stage with flowers.

Hindi (local language in India): यह छवि एक नवविवाहित भारतीय जोड़े को एक चुंबन शेरय करते हुए दिखाती है। दुल्हन एक पंचपारगत लाल साड़ी पहन रही है और दुल्हा एक लाल पगड़ी और सफेद सूट पहन रहा है। वे फूलों से सजे हुए एक मंच पर बैठे हैं।

Figure 44: Results on multilingual multiculture understanding. Check Section 4.5 for detailed discussions.

4.6 Coding Capability with Vision

Sec. 4.6 Coding Capability with Vision

| | | |
|---|--|---|
| Prompt: $a^2 + b^2 = c^2$ | Prompt: $\sum_{i=1}^N (i^2 + i + 1)^{0.5}$ | Prompt: $y = \int f(x, y) dx$ |
| GPT-4V: $a^2 + b^2 = c^2$ | GPT-4V: $\sum_{i=0}^N (i^2 + i + 1)^{0.5}$ | GPT-4V: $y = \int f(x, y) dx$ |
| Prompt: $a^2 + b^2 = c^2 + \sum_{i=1}^N (i^2 + i + 1)^{0.5} + y = \int f(x, y) dx$ | | |
| GPT-4V: $\mathbb{P}(x \neq 6) = c \times 5 \sum_{i=1}^N (i \times (i + 1)) + y_s \int_a^b f(x, y) dx$ | | |

Figure 45: GPT-4V’s capability to generate LaTeX codes based on the hand-written input. The instruction is ‘generate latex code.’ for each case. The output is the LaTeX code and we show the rendered result. Although the model fails to write the code for the complex equation (bottom), we can break it down into several simple equations, which GPT-4V is able to handle. Check Section 4.6 for detailed discussions.

Figure 45 illustrates the ability to generate LaTeX code based on handwritten mathematical equations. This functionality can assist users in writing equations in LaTeX more efficiently. Although the model is unable to generate code for longer equations, it can handle shorter equations effectively. By breaking down longer equations into shorter components, the model is able to generate the appropriate code. Figure 46 further demonstrates how GPT-4V can reconstruct a table in the input image into Markdown/LaTeX code.

Figure 47 shows examples of writing code in Python, TikZ, and SVG to replicate the input figure. Although the resulting output is not an exact match, the layout is similar and the code can be easily modified to meet specific needs.

Sec. 4.6 Coding Capability with Vision

Prompt:

Please read the table in this image and return a markdown-style reconstructed table in text.

| Country | Sales Volume | Revenue | Profit |
|-------------|--------------|--------------|-------------|
| USA | 40.080 | \$15.971.880 | \$3.086.421 |
| China | 35.070 | \$15.866.670 | \$3.032.162 |
| Australia | 27.054 | \$14.812.566 | \$2.868.636 |
| India | 23.046 | \$10.608.174 | \$1.853.710 |
| South Korea | 16.032 | \$10.494.948 | \$1.975.844 |

GPT-4V:

```
| Country | Sales Volume | Revenue | Profit |
| --- | --- | --- | --- |
| USA | 40.080 | $15.971.421 | $3.086.421 |
| China | 35.070 | $15.866.670 | $3.032.162 |
| Australia | 27.054 | $14.812.566 | $2.868.636 |
| India | 23.046 | $10.608.174 | $1.853.710 |
| South Korea | 16.032 | $10.494.948 | $1.975.844 |
```

Visualization of Reconstructed Table (Markdown):

| Country | Sales Volume | Revenue | Profit |
|-------------|--------------|--------------|-------------|
| USA | 40.080 | \$15.971.421 | \$3.086.421 |
| China | 35.070 | \$15.866.670 | \$3.032.162 |
| Australia | 27.054 | \$14.812.566 | \$2.868.636 |
| India | 23.046 | \$10.608.174 | \$1.853.710 |
| South Korea | 16.032 | \$10.494.948 | \$1.975.844 |

Prompt:

Please read the table in this image and return a latex-style reconstructed table in text.

| Country | Sales Volume | Revenue | Profit |
|-------------|--------------|--------------|-------------|
| USA | 40.080 | \$15.971.880 | \$3.086.421 |
| China | 35.070 | \$15.866.670 | \$3.032.162 |
| Australia | 27.054 | \$14.812.566 | \$2.868.636 |
| India | 23.046 | \$10.608.174 | \$1.853.710 |
| South Korea | 16.032 | \$10.494.948 | \$1.975.844 |

Note: Same query image as shown above, resized to save space

GPT-4V:

Here is the reconstructed table in latex:

```
\begin{tabular}{|c|c|c|c|} \hline
Country & Sales Volume & Revenue & Profit \\ \hline
USA & 40.080 & \$15.971.887 & \$3.068.421 \\ \hline
China & 35.070 & \$15.866.670 & \$3.032.162 \\ \hline
Australia & 27.054 & \$14.812.566 & \$2.868.636 \\ \hline
India & 23.046 & \$10.608.174 & \$1.853.710 \\ \hline
South Korea & 16.032 & \$10.494.948 & \$1.975.844 \\ \hline
\end{tabular}
```

Visualization of Reconstructed Table (LaTex):

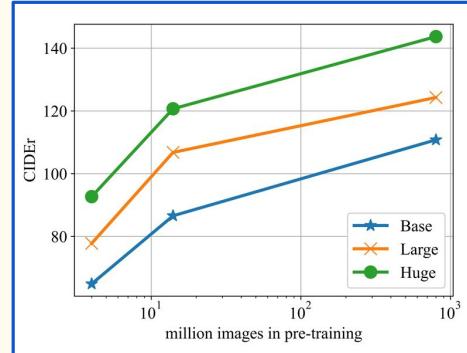
| Country | Sales Volume | Revenue | Profit |
|-------------|--------------|--------------|-------------|
| USA | 40.080 | \$15.971.887 | \$3.068.421 |
| China | 35.070 | \$15.866.670 | \$3.032.162 |
| Australia | 27.054 | \$14.812.566 | \$2.868.636 |
| India | 23.046 | \$10.608.174 | \$1.853.710 |
| South Korea | 16.032 | \$10.494.948 | \$1.975.844 |

Figure 46: GPT-4V's capability to generate Markdown/LaTeX codes to reconstruct a table in the image. Red highlights the errors in reconstruction. Check Section 4.6 for detailed discussions.

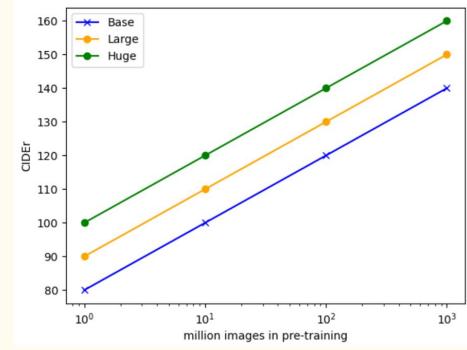
Sec. 4.6 Coding Capability with Vision

Prompt:

Generate python code to draw similar curves.

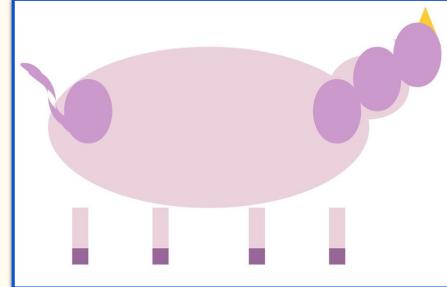


GPT-4V:

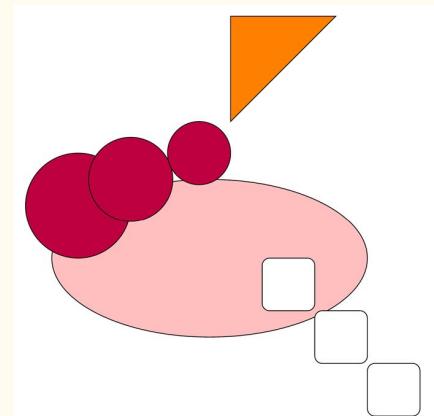


Prompt:

Generate the following image in tikz.

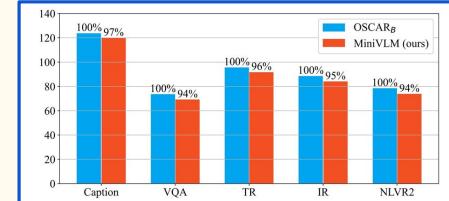


GPT-4V:

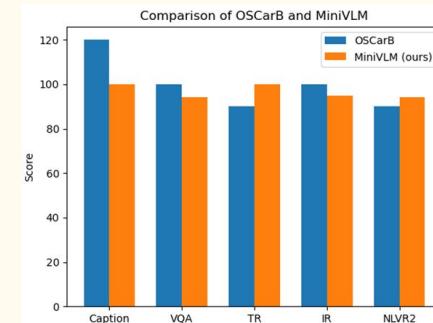


Prompt:

Write python codes to generate similar figures.



GPT-4V:



Prompt:

Write svg codes to generate following image.



GPT-4V:

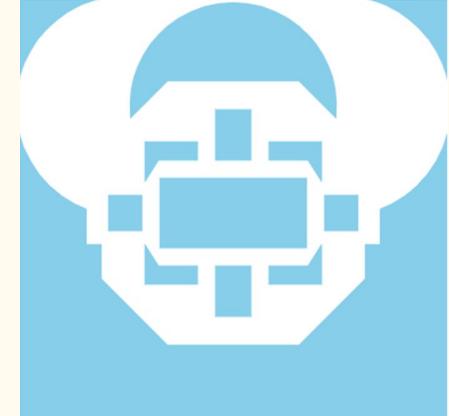


Figure 47: GPT-4V's capability to write codes to replicate the input figure. We directly show the rendered figures by python/TikZ/SVG as GPT-4V's response. The rendered figure is roughly aligned with the input figure, and the code can be easily adapted. GPT-4V Chart. Check Section 4.6 for detailed discussions.

5 Interaction with Humans: Visual Referring Prompting

Pointing to a specific spatial location is an essential capability in human-computer interaction with multimodal systems, such as conducting visually grounded dialogues. As shown in Section 5.1, GPT-4V can well understand the visual pointers directly drawn on images. Based on this observation, we propose a novel model interaction method named “visual referring prompting.” The core idea is to directly edit image pixel space to draw visual pointers or scene texts as human referring instructions, as highlighted in Figure 50. We detail its usages and advantages in Section 5.2. Finally, Section 5.3 explores having GPT-4V generate visual pointer outputs to interact with humans. These visual pointers are intuitive for both humans and machines to generate and understand, making them a good channel for human-computer interaction.

5.1 Understand Pointing Inputs

As illustrated in Figure 48, GPT-4V can understand different types of visual markers directly overlaid on images as a pointer, such as circles, boxes, and hand drawings. This ability helps GPT-4V generate grounded captioning, which is a known challenging problem to have conventional vision-language models [121] generating visual descriptions focused on a specific area of interest. Dense captioning methods [59, 131] use cropped boxes or mask regions to generate localized descriptions, but often ignore the global image context and produce sub-optimal descriptions. Visual pointing provides a natural way to indicate the area of interest while maintaining the global image context. For example, the top left example focuses on providing a comprehensive description of the pointed Magna beer, while also mentioning the global image context that the beer bottle is on the table.

An intuitive alternative to visual pointers overlaid on images is the region coordinates represented in the numerical text format. As shown in Figure 49, GPT-4V can understand the coordinates out-of-box, achieving the ability of spatial referring via text tokens without extra box token finetuning as in prior vision-language models [122, 136]. Despite the promising capability, we note that our current prompt is less precise spatially. For example, in the top left example in Figure 49, GPT-4V mentions the surrounding objects napkin and water bottle, even though only the beer bottle is in the region (0.47, 0.48, 0.55, 0.87). Overall, with respect to our experimented prompts, GPT-4V works more reliably when prompted with overlaid visual pointers, compared with text coordinates. This unique capability motivates us to explore a new prompting method, namely visual referring prompting.

Sec. 5.1 Understand Pointing Inputs

Grounded Description (overlaid visual pointers)

Prompt:

Describe the pointed region in the image.



GPT-4V:

The pointed region in the image is a green glass bottle of Magna beer, with a red label and a gold cap. It is sitting on a black table, surrounded by other items such as a water bottle, napkins, and umbrellas.

Prompt:

Describe the pointed region in the image.



GPT-4V:

The pointed region in the image contains a white folded napkin and a silverware set wrapped in another white napkin, placed on a black table. This is likely a table setting for a meal at a restaurant.

Prompt:

Describe the two pointed objects in the image.



GPT-4V:

The first pointed object (indicated by the yellow arrow) is a white patio umbrella. It has a pointed top and is attached to a metal pole. The umbrella is open and providing shade for the outdoor seating area. The second pointed object (indicated by the red arrow) is a green glass bottle of Magna beer. The bottle has a pointed neck and is capped with a gold-colored metal lid. It is sitting on a black table in front of a white napkin.

Prompt:

Describe the pointed region in the image.



GPT-4V:

The pointed region in the image is a green beer bottle with a red label that reads "MAGNA". The bottle is standing upright on a black table with a white napkin next to it.

Prompt:

Describe the pointed region in the image.

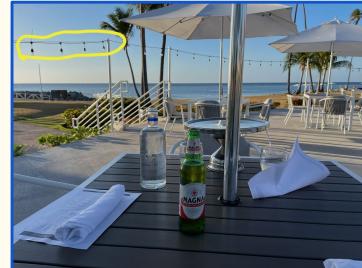


GPT-4V:

The pointed region in the image is a section of a wooden railing with vertical slats. It is located on the edge of a patio overlooking the ocean. The railing appears to be painted white and is partially obscured.

Prompt:

Describe the pointed region in the image.



GPT-4V:

The pointed region in the image is a row of hanging lights on a wire. The lights are small and round, and they are evenly spaced along the wire. The wire is stretched between two poles and is located above a patio area with tables and umbrellas. The lights appear to be decorative and are likely used to provide ambient lighting in the evening.

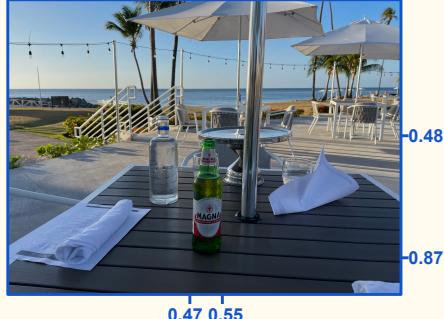
Figure 48: GPT-4V understands visual pointers directly overlaid on images. Conducting grounded description with both local and global visual information is one unique application scenario. Check Section 5.1 for detailed discussions.

Sec. 5.1 Understand Pointing Inputs

Grounded Description (text-format region coordinates)

Prompt:

Locate the region based on given coordinates and describe the contents in the region. Regions are represented by (x_1, y_1, x_2, y_2) coordinates. x_1 x_2 are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. y_1 y_2 are the top and bottom most positions, normalized into 0 to 1, where 0 is the top and 1 is the bottom. Describe the relative position of region $(0.47, 0.48, 0.55, 0.87)$ in the image. Describe the contents in that relative position. No need to mention contents in other part of the image.

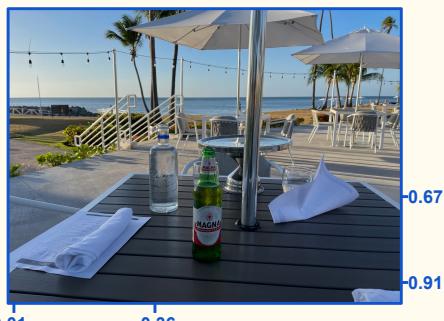


GPT-4V:

The region $(0.47, 0.48, 0.55, 0.87)$ is located in the center-right part of the image. It covers a small area of the table and extends to the bottom of the image. The most prominent object in this region is a green bottle of Magna beer. There is also a white napkin and a part of a water bottle visible in this region. The background shows a part of the patio and the ocean.

Prompt:

Locate the region based on given coordinates and describe the contents in the region. Regions are represented by (x_1, y_1, x_2, y_2) coordinates. x_1 x_2 are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. y_1 y_2 are the top and bottom most positions, normalized into 0 to 1, where 0 is the top and 1 is the bottom. Describe the relative position of region $(0.01, 0.09, 0.29, 0.21)$ in the image. Describe the contents in that relative position. No need to mention contents in other part of the image.

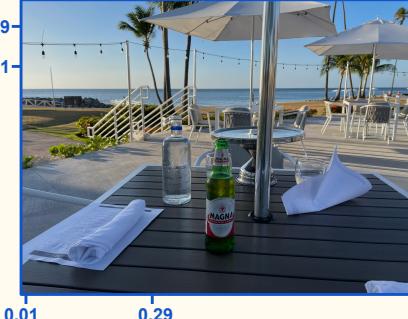


GPT-4V:

The region $(0.01, 0.09, 0.29, 0.21)$ in the image is located at the bottom left corner, covering approximately one third of the width and one fourth of the height of the image. In this region, there is a black table with a white napkin, a bottle of Magna beer, and a glass bottle of water. The table is in the foreground, while the background shows a part of the patio with white umbrellas and chairs.

Prompt:

Locate the region based on given coordinates and describe the contents in the region. Regions are represented by (x_1, y_1, x_2, y_2) coordinates. x_1 x_2 are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. y_1 y_2 are the top and bottom most positions, normalized into 0 to 1, where 0 is the top and 1 is the bottom. Describe the relative position of region $(0.01, 0.09, 0.29, 0.21)$ in the image. Describe the contents in that relative position. No need to mention contents in other part of the image.

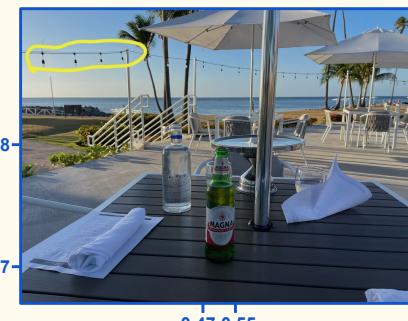


GPT-4V:

The region $(0.01, 0.09, 0.29, 0.21)$ is located in the top left corner of the image. It covers approximately 28% of the width and 12% of the height of the image. In this region, there is a part of a white umbrella and a string of lights. The background shows a clear blue sky and a palm tree.

Prompt:

Locate the region based on given coordinates and describe the contents in the region. Regions are represented by (x_1, y_1, x_2, y_2) coordinates. x_1 x_2 are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. y_1 y_2 are the top and bottom most positions, normalized into 0 to 1, where 0 is the top and 1 is the bottom. Describe the relative position of region $(0.47, 0.48, 0.55, 0.87)$ in the image. Describe the contents in that relative position. No need to mention contents in other part of the image.



GPT-4V:

The region $(0.47, 0.48, 0.55, 0.87)$ is located towards the right side of the image, slightly below the center. It covers a small portion of the table and the background. In this region, there is a green bottle of Magna beer on the table, with a white napkin next to it. In the background, there is a part of an umbrella and a few chairs.

Figure 49: An alternative to visual pointers overlaid on images is the region coordinates represented in the numerical text format. GPT-4V can understand the coordinates, e.g., $(0.47, 0.48, 0.55, 0.87)$, $(0.01, 0.09, 0.29, 0.21)$, and $(0.01, 0.67, 0.36, 0.91)$ that correspond to the center beer bottle, top-left string lights, and bottom-left table set, respectively. We observe that GPT-4V works less reliably when prompted with text coordinates, compared with visual pointers in visual referring prompting. Check Section 5.1 for detailed discussions.

5.2 Visual Referring Prompting

Inspired by GPT-4V’s strong capability in understanding visual pointing and scene text, we explore a new method to interact with GPT-4V, namely the *visual referring prompting*. Instead of conventional prompting techniques that edit text space, visual referring prompting is a complementary technique that directly edits the pixel space for input images for human-computer interaction. Such visual prompting could offer a more nuanced and comprehensive interaction with the image, potentially unlocking a wider array of responses from the model. For example, in Figure 50 (1), GPT-4V naturally associates the arrow-pointed objects with the given object indexes, easing the remaining visual reasoning and text outputs; in (2), GPT-4V understands the questions written on the image and pointed to the corresponding edge or angle, providing a nuanced interface for grounded visual dialogue; in (3), humans can point to arbitrary regions inside the figure to help GPT-4V better understand complicated documents and charts; in (4), the pattern can be concisely represented as an arrow and the scene text “+dot”, therefore helping GPT-4V to predict the next image. Complementary to text prompts that are loosely grounded to images, visual referring prompting provides a novel interaction method that could facilitate various use cases, with additional demonstrations in Figure 51 and Section 9.

5.3 Generate Pointing Outputs

Section 5.1 discusses the ability of GPT-4V to understand visual pointing generated by humans. A natural question is: Can GPT-4V generate its own pointing outputs, thereby facilitating a closed-loop interaction process in human-computer interaction?

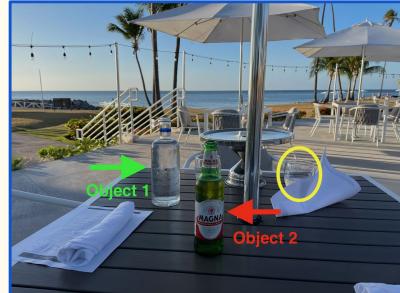
Figure 52 explores generating visual pointing outputs by letting GPT-4V predict region coordinates in the text format. We prompt GPT-4V to ground the object referred by text (*e.g.*, the text of “blue Subaru SUV”) or a reference image (*e.g.*, the image of “black Audi sedan”). Similar to the observation in having GPT-4V comprehend coordinates input, the model has a coarse understanding of spatial locations, but it wasn’t accurate with respect to the prompts used in the experiment. For example, in Figure 52’s “plot visualizations,” GPT-4V can approximately identify the blue SUV and black sedan mentioned in the query, but it struggles to create a closely-fitted bounding box. We observe that including example-grounded instructions in the prompt helps GPT-4V to understand the definition of coordinates and subsequently generate better pointing outputs.

While the generated pointing outputs may not perfectly cover the queried region, they still provide a valuable tool for model interaction, interpretation, and helping multi-step visual reasoning. Specifically, the pointing outputs can be interpreted by humans to better understand GPT-4V’s references, or by GPT-4V itself to enable further reasoning based on previous outputs. As shown in the bottom of Figure 52, GPT-4V is capable of interpreting the pointers it generates, and providing grounded descriptions with the prompts in Figure 48. Furthermore, the iterative pointing generation and understanding by GPT-4V itself can help complicated multi-hop grounded visual reasoning tasks. GPT-4V is capable of deconstructing the question, generating distinct visual markers to iteratively focus on different image regions for each sub-step, ultimately collating the information to formulate the final answer.

Sec. 5.2 Visual Referring Prompting

(1) Prompt:

What is in the circled glass? Object 1 or object 2?
Describe what is Object 1 and object 2. Then check what is in the circled glass.



GPT-4V:

Object 1 is a glass bottle of water. It is clear and has a blue label on it.
Object 2 is a glass bottle of Magna beer. It is green and has a red and white label on it.
In the circled glass, it appears to be a clear liquid, likely water. So, it is likely that the contents of the circled glass are from Object 1, the glass bottle of water.

(3) Prompt:

Which circled line shows a different comparison trend?
Provide intermediate steps.

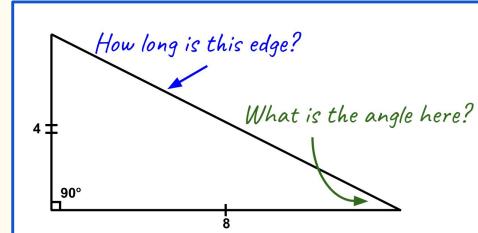
| Exam | GPT-4 | GPT-4 (no vision) |
|---|--|--|
| Uniform Bar Exam (MBE+MEE+MPT) <i>LSAT</i> | 298 / 400 (~90th) 163 (~88th) | 298 / 400 (~90th) 161 (~83rd) |
| SAT Evidence-Based Reading & Writing <i>SAT Math</i> | 710 / 800 (~93rd) 700 / 800 (~89th) | 710 / 800 (~93rd) 690 / 800 (~89th) |
| Graduate Record Examination (GRE) Quantitative | 163 / 170 (~80th) | 157 / 170 (~62nd) |
| Graduate Record Examination (GRE) Verbal | 169 / 170 (~99th) | 165 / 170 (~96th) |
| Graduate Record Examination (GRE) Writing | 4 / 6 (~54th) | 4 / 6 (~54th) |
| USABO Semifinal Exam 2020 <i>USNCO Local Section Exam 2022</i> | 87 / 150 (99th - 100th) 36 / 60 | 87 / 150 (99th - 100th) 38 / 60 |
| Medical Knowledge Self-Assessment Program | 75 % | 75 % |
| Codeforces Rating | 392 (below 5th) | 392 (below 5th) |

GPT-4V:

The circled line that shows a different comparison trend is the USNCO Local Section Exam 2022. Intermediate steps:
1. Analyze the data in the table and look for any discrepancies in the comparison trends. 2. Notice that in all the rows except for the USNCO Local Section Exam 2022, the GPT-4 (no vision) column has a lower score than the GPT-4 column. 3. This means that in all the other rows, the GPT-4 model with vision performed better than the GPT-4 model without vision. 4. However, in the USNCO Local Section Exam 2022 row, the GPT-4 (no vision) column has a higher score than the GPT-4 column. 5. This is a different comparison trend than the rest of the rows, indicating that the GPT-4 model without vision performed better than the GPT-4 model with vision in this particular exam.

(2) Prompt:

Answer the question in the following figure. Provide intermediate steps.



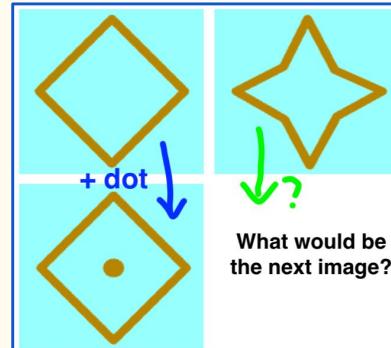
GPT-4V:

To find the length of the blue arrowed edge, we can use the Pythagorean theorem since we have a right triangle. The theorem states that $a^2 + b^2 = c^2$, where a and b are the two legs of the right triangle and c is the hypotenuse. Plugging in the values we have, we get:
 $4^2 + 8^2 = c^2$
 $16 + 64 = c^2$
 $80 = c^2$
 $\sqrt{80} = c$
 $8.94 \approx c$
Therefore, the length of the blue arrowed edge is approximately 8.94.

To find the angle at the green arrow, we can use the tangent function since we have the opposite and adjacent sides of the angle. The tangent function states that $\tan\theta = \text{opposite}/\text{adjacent}$. Plugging in the values we have, we get:
 $\tan\theta = 4/8$
 $\tan\theta = 1/2$
 $\theta = \tan^{-1}(1/2)$
 $\theta \approx 26.57^\circ$
Therefore, the angle at the green arrow is approximately 26.57°.

(4) Prompt:

Answer questions in the following figure. Provide intermediate steps.



GPT-4V:

The next image in the sequence would be a star with a dot in the center. This is because the sequence appears to be following a pattern where the shape alternates between a diamond and a star, and a dot is added to the center of the shape every other image. Intermediate steps:

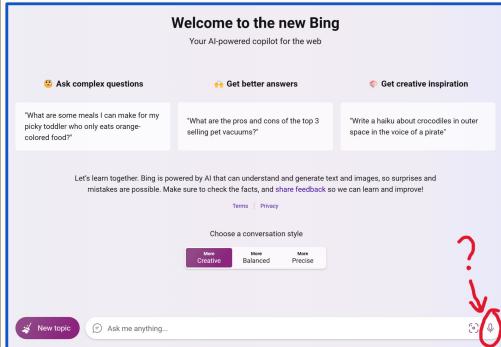
- First, we can observe that the sequence alternates between a diamond and a star shape.
- Next, we can see that a dot is added to the center of the shape every other image.
- Based on this pattern, we can predict that the next image will be a star with a dot in the center.

Figure 50: Visual referring prompting directly edits the input image as input prompts, such as drawing visual pointers and scene texts. Complementary to text prompts, visual referring prompting provides a more nuanced and natural interaction, e.g., (1) associating pointed objects with an index, (2) pointing to the image for questioning, (3) highlighting lines in documents and tables, (4) drawing the pattern on the image, and many other novel use cases. Check Section 5.2 for detailed discussions.

Sec. 5.2 Visual Referring Prompting

Prompt:

What is the icon used for?

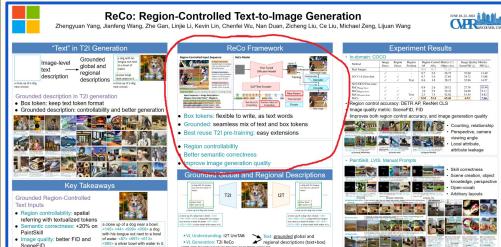


GPT-4V:

The icon circled in the image is a microphone icon. It is used to activate the voice search feature on Bing. Users can click on this icon and speak into their device's microphone to ask a question or search for something instead of typing it in. This can be helpful for hands-free searching or for users who may have difficulty typing.

Prompt:

What does this part of the poster highlights?

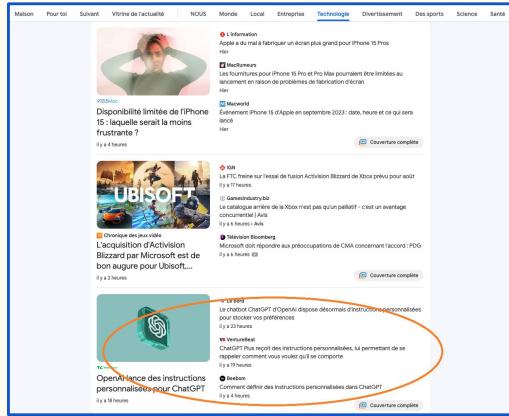


GPT-4V:

This part of the poster highlights the ReCo Framework, which is the main contribution of the research. It explains how the framework works and what it does. The ReCo Framework is a region-controlled text-to-image generation model that can generate high-quality images based on text descriptions. It also shows the different components of the framework and how they interact with each other. The ReCo Framework is designed to be flexible and easy to extend, and it has been shown to improve image generation quality and achieve better semantic coherence compared to other text-to-image generation methods.

Prompt:

What does the pointed news talk about?

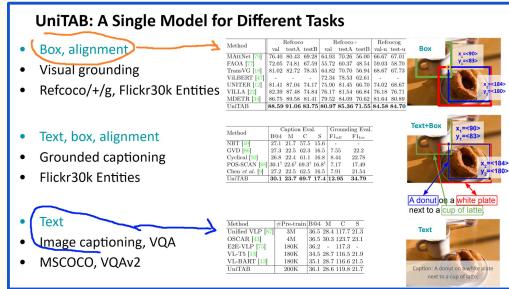


GPT-4V:

The pointed news is talking about the new feature of ChatGPT, which allows users to create personalized instructions. This feature is said to help users better manage their conversations and improve their experience with the chatbot.

Prompt:

What are the differences between these two experiments in the slides?



GPT-4V:

There are a few key differences between the two experiments shown in the slides:

1. The tasks: The first experiment (on the left) is focused on the task of visual grounding, where the model is trained to identify and localize objects in an image. The second experiment (on the right) is focused on the task of image captioning, where the model is trained to generate a descriptive caption for an image.
2. The datasets: The first experiment uses the Flickr30k Entities dataset, which contains images with annotated object bounding boxes. The second experiment uses the MSCOCO and VQA2 datasets, which contain images with associated captions.
3. The evaluation metrics: The first experiment uses the mean Average Precision metric to evaluate the model's performance on the visual grounding task. The second experiment uses the BLEU and CIDEr metrics to evaluate the model's performance on the image captioning task.

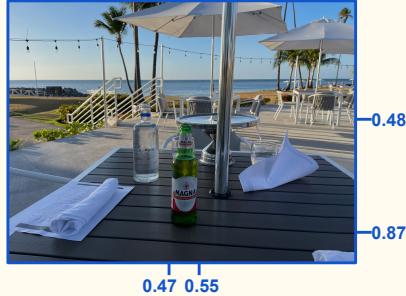
Overall, the two experiments are designed to evaluate the UniTAB model's performance on different tasks and datasets, and to demonstrate its ability to handle multiple tasks with a single model.

Figure 51: Visual referring prompts enhance the seamless interaction between humans and computers. This is evident in the integration with computer and mobile Graphical User Interfaces (GUIs), and the support provided in understanding documents and slides. Check Section 5.2 for detailed discussions.

Sec. 5.3 Generate Pointing Outputs

Prompt:

Locate the referred object and represent the location of the region. Regions are represented by (x_1, y_1, x_2, y_2) coordinates. x_1 x_2 are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. y_1 y_2 are the top and bottom most positions, normalized into 0 to 1, where 0 is the top and 1 is the bottom. For example, the (x_1, y_1, x_2, y_2) coordinates for that region with the beer bottle is $(0.47, 0.48, 0.55, 0.87)$.



For the image below, what is the (x_1, y_1, x_2, y_2) coordinates for the region with the blue Subaru SUV?



GPT-4V:

The (x_1, y_1, x_2, y_2) coordinates for the region with the blue Subaru SUV would be $(0.68, 0.57, 0.95, 0.95)$.

Plot Visualization:

Refer to the query image below

Grounded Description

Prompt:

Describe the pointed region in the image.



GPT-4V:

The pointed region in red is a part of a larger mural on the side of a building. The content in this specific region appears to be a depiction of a blue car, possibly an SUV, parked in a parking lot. The car is facing towards the right side of the image and appears to be in a parking spot with a white line on the left side.

Prompt:

Locate the referred object and represent the location of the region. Regions are represented by (x_1, y_1, x_2, y_2) coordinates. x_1 x_2 are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. y_1 y_2 are the top and bottom most positions, normalized into 0 to 1, where 0 is the top and 1 is the bottom. For example, the (x_1, y_1, x_2, y_2) coordinates for that region with the beer bottle is $(0.47, 0.48, 0.55, 0.87)$.



For the image below, what is the (x_1, y_1, x_2, y_2) coordinates for the region with object



in it?



GPT-4V:

The (x_1, y_1, x_2, y_2) coordinates for the region with the black Audi car in it would be approximately $(0.35, 0.55, 0.55, 0.85)$. This is based on the left and right most positions of the car being around 35% and 55% of the image width, respectively, and the top and bottom most positions being around 55% and 85% of the image height, respectively.

Plot Visualization:

Refer to the query image below

Grounded Description

Prompt:

Describe the pointed region in the image.



GPT-4V:

The object in the pointed region in red is a black car, specifically a sedan, parked in a parking lot. It appears to be a newer model with a shiny exterior and tinted windows.

Figure 52: GPT-4V can use its understanding of coordinates to generate visual pointing output, thereby grounding the textual or visually queried object. Using example-grounded instructions can help GPT-4V understand coordinate definitions and therefore generate better pointing. While output spatial regions are not precise, the approach enables an “understanding (*i.e.*, grounded description) and generation” loop for visual pointing, leading to an effective way of human-computer interaction. Check Section 5.3 for detailed discussions.

6 Temporal and Video Understanding

In this section, we discuss temporal and video understanding capabilities. Even though GPT-4V operates primarily on images as inputs, evaluating its understanding of temporal sequences and video content remains a crucial aspect of its overall assessment. This is because real-world events unfold over time, and an AI system’s ability to understand these dynamic processes is instrumental in real-world applications. Capabilities like temporal anticipation, temporal ordering, temporal localization, temporal reasoning, and grounded temporal understanding help to gauge the model’s proficiency in comprehending the sequence of events, anticipating future occurrences, and contextually analyzing activities over time, all within a series of static images. In spite of its image-centric focus, GPT-4V is able to comprehend video and temporal sequences in a way that’s similar to human comprehension. To enhance the versatility and applicability of a sophisticated AI model like GPT-4V, this aspect of testing is critical to its development and refinement. For the upcoming experiments in this section, we will use multiple selected video frames as inputs to test the model’s abilities in understanding temporal sequences and video content.

6.1 Multi-image Sequencing

In this subsection, we demonstrate that GPT-4V can accurately comprehend and analyze sequences of video frames. Within this frame-by-frame analysis, GPT-4V recognizes the scene in which the activity is taking place, delivering a deeper contextual understanding. As shown in Figure 53, the model is not just confined to recognizing the environment; it also accurately interprets the actions being performed by individuals in the video. GPT-4V understands the sequence and context of various human poses and intelligently correlates them with the ongoing activity. By understanding pose variations beyond just identification, GPT-4V can derive meaning from the subtleties of human movement and action. As a result of this level of detailed understanding, GPT-4V can capture the essence of what’s happening in videos, offering rich and nuanced insights that go beyond just identifying objects and scenes.

6.2 Video Understanding

Temporal ordering. Temporal ordering is a crucial element of temporal commonsense and forms an essential part of GPT-4V’s capabilities evaluation. This involves providing the model with a series of shuffled images and gauging its ability to discern cause and effect relationships as well as time progressions. An understanding of such relationships requires the ability to reorder the sequence in a logically coherent and temporally accurate manner. Figure 54 illustrates an example of long-term temporal ordering where GPT-4V is presented with a series of shuffled image frames depicting a sushi-making event. Despite the disorder, GPT-4V effectively identifies the event and determines the appropriate temporal sequence of the sushi-making process. In addition, Figure 55 provides an example of short-term temporal ordering. Given a designated action, such as opening or closing a door, GPT-4V demonstrates its capability to comprehend the image’s content and determine the correct sequential order of the events. These examples highlight GPT-4V’s capability in temporal commonsense, reinforcing its ability to comprehend both long-term and short-term sequences accurately.

Temporal anticipation. We demonstrate GPT-4V’s ability to anticipate future events given a set of initial frames. Long- and short-term examples are used to validate this capacity for anticipating future events. The right side of Figure 56 illustrates GPT-4V’s ability to anticipate short-term events with a soccer penalty kick example. Given the first few frames, it accurately foresees the typical next actions of both the kicker and the goalkeeper, due to its understanding of the inherent structure and rules of the game. In addition, as shown in The left side of Figure 56, the sushi preparation sequence illustrates GPT-4V’s long-term anticipation capability. By understanding the activity based on visual cues, GPT-4V not only recognizes the current progress in sushi preparation but also accurately anticipates the subsequent steps, demonstrating its capacity to interpret and predict complex, multi-step processes over an extended period. This combination of short-term and long-term temporal anticipation allows GPT-4V to capture and understand activities with varying temporal structures and complexities.

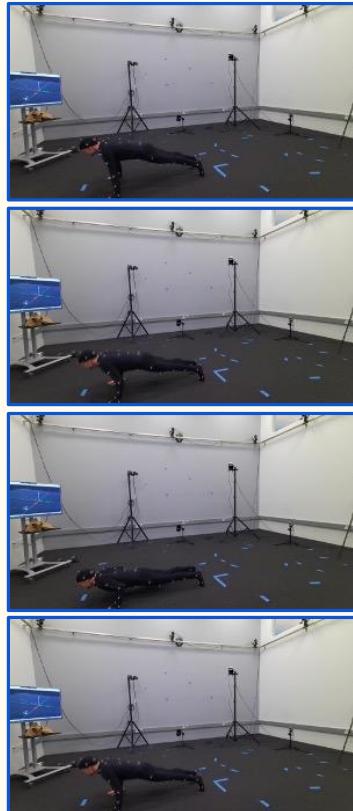
Temporal localization and reasoning. Figure 57 illustrates GPT-4V’s capabilities in temporal localization and reasoning. It accurately identifies the precise moment the player strikes the ball. Fur-

Sec. 6.1 Multi-image Sequencing

Action Recognition

Prompt:

Explain the images and the action the person is taking.



GPT-4V:

The images show a person in a motion capture suit in a studio with multiple cameras and sensors. The person is performing a push-up exercise. In the first image, the person is in the starting position with their arms extended and body in a straight line. In the second image, the person is lowering their body towards the ground. In the third image, the person is at the bottom of the push-up with their chest close to the ground. In the fourth image, the person is pushing their body back up to the starting position.

Figure 53: Sequences of video frames understanding: Interpreting human poses and deriving relevant insights from video sequences. Check Section 6.1 for detailed discussions.

thermore, GPT-4V showcases its understanding of cause and effect by inferring from the relationship between the goalkeeper and the ball to determine if the goalkeeper successfully blocks the ball. In the context of the example given, understanding whether the goalkeeper can block the ball involves not only recognizing the spatial positions of the goalkeeper and the ball but also understanding the dynamics of their interaction and predicting the outcome of these dynamics. This demonstrates a considerable level of sophistication in the model's reasoning abilities.

Sec. 6.2 Video Understanding

Temporal Ordering

Prompt:

Analyze the images and decide the correct order in which they should be displayed.

Prompt Continued

a



d



b



e



c



GPT-4V:

The correct order in which the images should be displayed is:

1. e - Mixing the rice
2. c - Placing the rice on the seaweed
3. b - Adding the fillings
4. a - Cutting the roll
5. d - Finished sushi rolls on a plate.

Figure 54: Long-term temporal ordering: GPT-4V is presented with shuffled image frames depicting a sushi-making event. While the sushi-making process is disordered, GPT-4V is able to identify the event and determine the correct temporal sequence. Check Section 6.2 for detailed discussions.

Sec. 6.2 Video Understanding

Temporal Ordering

Prompt:

For opening a door, decide the correct order in which they should be displayed.

a



b



c



Prompt:

For closing a door, decide the correct order in which they should be displayed.

a



b



c



GPT-4V:

The correct order would be:

1. a - Person approaching the door
2. b - Person opening the door
3. c - Person entering through the door.

GPT-4V:

Correct order:

- 1) c - Person standing in doorway
- 2) b - Person starting to close door
- 3) a - Person closing door completely

Figure 55: Short-term temporal ordering: given a specified action, such as opening or closing a door, GPT-4V demonstrates its capability to comprehend the images' content and determine the correct sequential order corresponding to the specified action. Check Section 6.2 for detailed discussions.

Sec. 6.2 Video Understanding

Temporal Anticipation

Prompt:

Predict what will happen next based on the images.



GPT-4V:

Based on the images, it appears that the player in the white jersey is taking a penalty kick. The next likely event is that the player will kick the ball towards the goal, attempting to score. The goalkeeper in the black jersey will try to block the shot.

Prompt:

Analyze the images and decide what the next steps should be.



GPT-4V:

The images show the process of making sushi rolls. The first image shows rice being mixed with vinegar in a bowl. The second image shows the rice being spread onto a sheet of seaweed. The third image shows the addition of salmon and cucumber to the roll. The next steps would be to roll the sushi using a bamboo mat, cut the roll into smaller pieces, and serve with soy sauce and wasabi.

Figure 56: Short-term and long-term temporal anticipation: GPT-4V captures and understands activities with varying temporal structures and complexities. Check Section 6.2 for detailed discussions.

Sec. 6.2 Video Understanding

Temporal Localization and Reasoning

Prompt:

In which frame does the player kick the ball? Did the goalkeeper block the ball?

Prompt Continued



GPT-4V:

The player kicks the ball at frame c. The ball was not blocked by the goalkeeper as it can be seen in frame f that the ball is in the net and the goalkeeper is on the ground.

Figure 57: Temporal localization and reasoning: GPT-4V shows the capability in temporal localization by accurately identifying when the player strikes the ball. It also demonstrates cause-and-effect reasoning by determining whether the ball was blocked based on the goalkeeper-ball interaction. Check Section 6.2 for detailed discussions.

Sec. 6.3 Video Visual Referring Prompting

Grounded Temporal Understanding

Prompt:

Describe and summarize what the circled person doing?



Prompt:

Summarize the activities of the person.



GPT-4V:

The circled person is a man who is standing next to a bench and appears to be in a conversation with another man who is sitting on the bench. In the second image, the circled man is pointing towards something and in the third image, he is playfully pretending to punch the other man. Overall, it seems like the circled man is engaging in a friendly conversation and joking around with the other man.

GPT-4V:

The circled person is sitting on a bench and then stands up and adjusts his shirt.

Figure 58: Grounded temporal understanding: GPT-4V can apply a temporal understanding to a specific person of interest, indicated by a circle. Check Section 6.3 for detailed discussions.

6.3 Visual Referring Prompting for Grounded Temporal Understanding

Section 5 illustrates GPT-4V's capabilities in visual referring prompting. In this section, we aim to extend this capability by testing visual referring prompting for temporal understanding. This advancement offers enhanced control over video comprehension tasks.

Grounded temporal understanding. Grounded temporal understanding forms another crucial aspect of GPT-4V's capabilities, which we explore using pointing input in a sequence of image frames. Figure 58 exemplifies this by demonstrating how GPT-4V can apply a temporal understanding to a specific person of interest, indicated by a circle. GPT-4V can accurately describe events in a way that aligns with the corresponding temporal order, focusing on the activities of the circled individual. Beyond this, GPT-4V demonstrates a more refined understanding of the event, recognizing the nature of the interactions. For instance, GPT-4V can distinguish between friendly interactions and violent incidents, illustrating an ability to not only comprehend the temporal flow of events but also to interpret the tone and nature of the interactions taking place. This indicates GPT-4V's capacity to process and comprehend complex temporal and social cues within a given sequence, adding a layer of depth to its understanding.

7 Abstract Visual Reasoning and Intelligence Quotient Test

Understanding and reasoning over abstract visual stimuli and symbols is one fundamental ability for human intelligence. This section examines if GPT-4V can abstract semantics from visual signals and can perform different types of human Intelligence Quotient (IQ) tests.

7.1 Abstract Visual Stimuli

Humans can infer semantics from abstract and often ambiguous visual stimuli. Figure 59 explores having GPT-4V interpret tangram [28, 92, 40, 56]. A tangram is a traditional geometric puzzle that consists of seven flat pieces called tans, which are put together to form shapes without overlapping the pieces. For example, GPT-4V interprets that sub-figure 7 in Figure 59 best illustrates a flying goose and provides reasoning descriptions for other sub-figure, *e.g.*, 4. person or robot, 9. boat or hat, and 10. dog or fox. GPT-4V also has the ability to understand other formats of abstract visual diagrams [120, 15, 143], such as ASCII text art of cartoon characters in Figure 59 and symbolic inputs in Figures 61-62.

7.2 Discovery and Association of Parts and Objects

Discovering and associating object parts [132, 41] is another important abstract visual reasoning capability. Humans can easily discover how object parts may compose a semantically meaningful object. Figure 60 designs examples to probe GPT-4V’s capability in associating object parts. In the left example, we ask GPT-4V to localize an object part based on its semantic meaning. In the right example, GPT-4V is asked to associate object parts segmented by SAM [62]. GPT-4V can process figures for all object parts and associate them in a semantically meaningful to form the boy visualized in the bottom right.

7.3 Wechsler Adult Intelligence Scale

Section 7.1 demonstrates the abstract visual understanding capability of GPT-4V. As a further challenge, GPT-4V is asked to perform different abstract reasoning tasks, sourced from human Intelligence Quotient (IQ) tests. The Wechsler Adult Intelligence Scale [126] is recognized as one of the “gold standard IQ tests,” and is designed to provide a comprehensive measurement of an individual’s cognitive abilities using a series of sub-tests. Figure 61 shows representative questions and GPT-4V’s outputs from each sub-test category. GPT-4V shows promises in abstract reasoning, answering questions with texts only, symbolic visual inputs, and natural images. For example, the bottom right sample shows that GPT-4V can interpret the analogy question and find the best comparison of shoes.

7.4 Raven’s Progressive Matrices

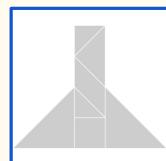
Raven’s Progressive Matrices (RPM) [102] is another well-known non-verbal intelligence test developed to measure abstract reasoning and problem-solving abilities. The test is designed to minimize the influence of language, culture, and formal education on test performance, making it suitable for testing AI models [15, 143, 52]. Each test sample contains three or eight images, arranged in 2-by-2 or 3-by-3 matrices with one figure missing. The goal is to select the next image from multiple candidate images by identifying patterns in the provided samples. In our approach, we challenge GPT-4V by sending the entire question page as a single image, instead of converting it into interleaved image-text pairs, similar to the human approach to IQ tests. As shown in Figure 62, GPT-4V can generate reasonable answers without processed text descriptions or sub-figures. However, we also notice that breaking down the entire question image into interleaved text and sub-figures, such as in Figure 63, does simplify the task and let GPT-4V produce more reliable answers.

Sec. 7.1 Abstract Visual Stimuli

Prompt:

A tangram is a geometric puzzle that consists of seven flat pieces to form shapes. Here are ten abstract tangram. Which one best illustrates a flying goose? What do other figures look like?

1.



2.



3.



4.



5.



6.



Prompt Continued

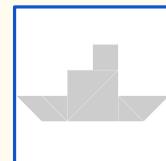
7.



8.



9.



10.



GPT-4V:

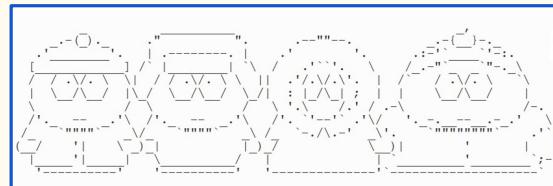
The tangram that best illustrates a flying goose is number 7. It has the triangular shape of a goose's body and wings, and the smaller triangle at the top could represent the head and neck.

The other figures could represent various things:

1. A rocket or a tower
2. A butterfly or a bat
3. A question mark or a snake
4. A person or a robot
5. A mountain range or a crown
6. A house or a temple
7. A flying goose (as mentioned above)
8. A bird or a fish
9. A boat or a hat
10. A dog or a fox.

Prompt:

Here is an abstract figure. Can you identify how many characters are there, and the cartoon from which they originate?



GPT-4V:

There are four characters in this abstract figure. They originate from the cartoon "South Park." The characters are Eric Cartman, Kyle Broflovski, Stan Marsh, and Kenny McCormick.

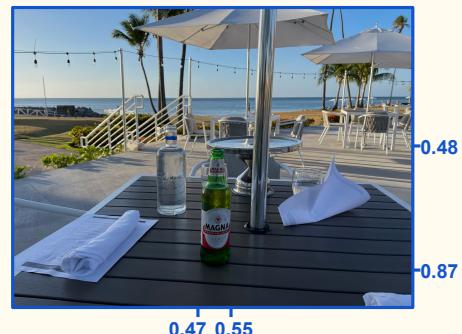
Figure 59: Understanding abstract visual stimuli such as tangram [56] and ASCII text art. Check Section 7.1 for detailed discussions.

Sec. 7.2 Association of Parts and Objects

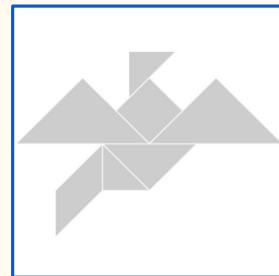
Discovery and Association of Parts and Objects

Prompt:

A tangram is a geometric puzzle that consists of seven flat pieces to form shapes. Locate the referred object and represent the location of the region. Regions are represented by (x_1, y_1, x_2, y_2) coordinates. x_1 x_2 are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. y_1 y_2 are the top and bottom most positions, normalized into 0 to 1, where 0 is the top and 1 is the bottom. For example, the (x_1, y_1, x_2, y_2) coordinates for that region with the beer bottle is $(0.47, 0.48, 0.55, 0.87)$.



The image below that shows a flying goose. Which part represents the head of the bird? What are the (x_1, y_1, x_2, y_2) coordinates for the region with the head of the bird?



GPT-4V:

The head of the bird is represented by the small triangle at the top of the image. The (x_1, y_1, x_2, y_2) coordinates for the region with the head of the bird would be approximately $(0.42, 0.15, 0.58, 0.35)$.

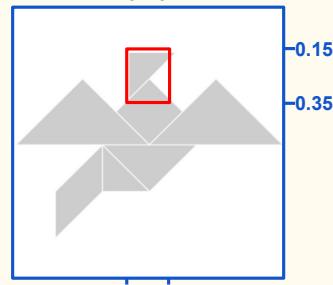
Plot Visualization:

Refer to the query image below

Grounded VQA

Prompt:

What is the highlighted part?

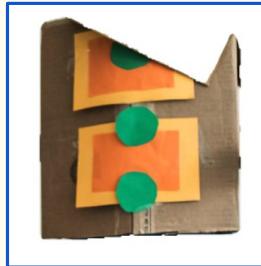


GPT-4V:

The highlighted part is the head of the goose.

Prompt:

How are those visual parts related, and can they be combined to form a single object such as a boy? If so, how to arrange them?



GPT-4V:

The three visual parts are related in that they are all components that could be used to create a single object. In this case, the object is a boy.

To combine these parts to form a boy, we can arrange them in the following way:

- The first image (the cardboard with orange and green shapes) could be used as the body of the boy.
- The second image (the pair of jeans) could be used as the legs of the boy.

- The third image (the cardboard with yellow, red, and blue shapes) could be used as the head of the boy.

By arranging these parts in this way, we can create a single object that resembles a boy.

Plot Visualization:



Figure 60: Understanding part-object association in abstract and natural images. Check Section 7.2 for detailed discussions.

Sec. 7.3 Wechsler Adult Intelligence Scale

Prompt:

Which one of the five is least like the other four?

- a. Carrot
- b. Potato
- c. Corn
- d. Apple
- e. Bean

GPT-4V:

- d. Apple

Explanation: Carrot, potato, corn, and bean are all vegetables, while apple is a fruit.

Prompt:

Tom, twelve years old, is three times as old as his sister. How old will Tom be when he is twice as old as his sister?

- a. 18
- b. 21
- c. 16
- d. 20
- e. 15

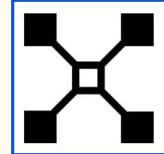
GPT-4V:

If Tom is three times as old as his sister, then his sister is one-third of his age. So, his sister is 4 years old. Now, we need to figure out when Tom will be twice as old as his sister. This means that his sister's age will be half of his age. So, we need to find a number that is double the age of 4. The answer is 8. So, Tom will be 16 years old when he is twice as old as his sister. The correct answer is c. 16.

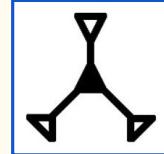
Prompt:

Which one of the five is least like the other four?

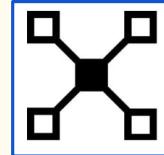
a.



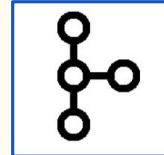
b.



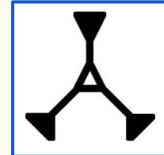
c.



d.



e.



GPT-4V:

Answer: d. is the least like the other four because it is the only one that does not have a central point connecting all the other points.

Prompt:

Which one of the five designs makes the best comparison?



a.



b.



c.



d.



e.



GPT-4V:

Answer: c. The black dress shoe

Explanation: The analogy is comparing a hand to a glove and a foot to a shoe. The best comparison would be the black dress shoe as it is an item of clothing that is worn on the foot, just like a glove is worn on the hand.

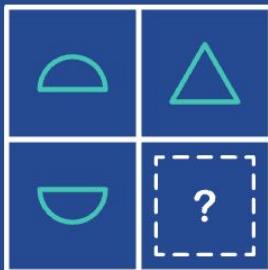
Figure 61: Example questions from the Wechsler Adult Intelligence Scale (WAIS) [126]. Check Section 7.3 for detailed discussions.

Sec. 7.4 Raven's Progressive Matrices

RPM: Entire Question Page

Prompt:

Find out the pattern in the first column and use it to infer the missing figure in the second column.



5. Choose a drawing that fits the empty space:



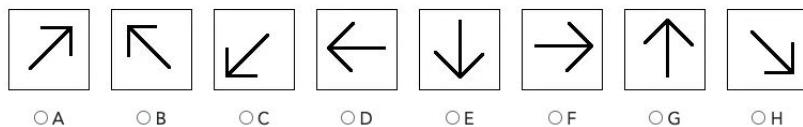
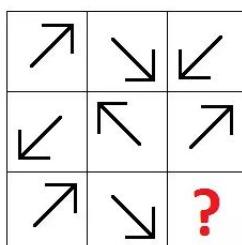
GPT-4V:

The correct drawing that fits the empty space is the upside-down triangle. This is because the pattern in the grid is that each row and column has one of each shape in different orientations.

Prompt:

Find out the pattern in the first two rows and use it to infer the missing figure in the third row.

Figure out which figure of A to H that should logically replace the missing figure/question-mark.



GPT-4V:

The pattern in the first two rows is that the figures are rotated 90 degrees clockwise in each subsequent cell.

Therefore, the missing figure in the third row should be a rotated version of the figure in the previous cell.

Looking at the options, the correct answer is OC, as it is the figure that is rotated 90 degrees clockwise from the figure in the previous cell.

Figure 62: Example questions from the Raven's Progressive Matrices [102, 52]. We challenge GPT-4V by sending the entire question page as a single image, mimicking how humans look at the IQ tests. Check Section 7.4 for detailed discussions.

Sec. 7.4 Raven's Progressive Matrices

RPM: Processed Sub-figures

Prompt:

1.



2.



3.



4.

Here are three sub images arranged in a 2-by-2 matrix. Figure out which figure of a to f that should logically replace the missing figure. What would be the next image to complete the matrix?

a.



b.



c.



d.



e.



f.



The first image in the first column, image 1, is a triangle. The first image in the second column, image 3, is a triangle with a line through it. The second image in the first column, image 2, is a diamond. Therefore, the second image in the second column, image 4 should continue this pattern and be a diamond with a line through it. Option b is the only image that fits this pattern, as it is a diamond with a line through it. Therefore, the correct answer is b.

Prompt Continued

1.



2.



3.



4.

Here are three sub images arranged in a 2-by-2 matrix. Figure out which figure of a to f that should logically replace the missing figure. What would be the next image to complete the matrix?

a.



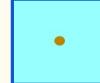
b.



c.



d.



e.



f.



GPT-4V:

The first image in the first column, image 1, is a diamond. The first image in the second column, image 3, is a diamond with a dot in the middle. The second image in the first column, image 2, is a star. Therefore, the second image in the second column, image 4 should continue this pattern and be a star with a dot in the middle. Option c is the only image that fits this pattern, as it is a star with a dot in the middle. Therefore, the correct answer is c.

Figure 63: Instead of sending the entire question page as a single image, we may also process the image into multiple sub-figures and optionally provide detailed instructions and examples to further boost the answer accuracy. Check Section 7.4 for detailed discussions.

8 Emotional Quotient Test

Sec. 8.1 Read Emotion from Facial Expressions

Prompt:

Identify and read emotions of people from their faces as shown in the image below.

**GPT-4V:**

The person in the image appears to be displaying a combination of emotions. The furrowed brows and downturned mouth suggest anger or frustration, while the tense jaw and narrowed eyes could indicate disgust or contempt. Overall, it seems like the person is experiencing negative emotions and is likely upset or unhappy.

Prompt:

Identify and read emotions of people from their faces as shown in the image below.

**GPT-4V:**

The person in the image is displaying a happy emotion. This can be seen from the wide smile and the crinkles around the eyes.

Prompt:

Identify and read emotions of people from their faces as shown in the image below.

**GPT-4V:**

The emotion shown in the image is fear or anxiety. The person's eyes are wide open and their mouth is slightly open, indicating a sense of shock or surprise. Their hands are covering their mouth, which could suggest that they are trying to suppress a scream or gasp.

Prompt:

Identify and read emotions of people from their faces as shown in the image below.

**GPT-4V:**

The person in the image appears to be experiencing a negative emotion, likely frustration or disappointment. This is evident from the furrowed brows, downward gaze, and hand on the forehead.

Figure 64: GPT-4V can reliably identify and read the emotions of people from their facial expressions. Check Section 8.1 for detailed discussions.

When interacting with humans, it is important that GPT-4V has the empathy and Emotional Quotient (EQ) to understand and share the feelings of humans. Inspired by the definition of the human EQ test [87, 86, 20], we examine GPT-4V’s capability in (1) identifying and reading human emotions from their facial expressions, (2) understanding how different visual contents may arouse emotions, and (3) generating proper text outputs conditioned on the desired emotional and sentiment.

8.1 Read Emotion from Facial Expressions

As shown in Figure 64, GPT-4V can reliably identify and read the emotions of people from their facial expressions. It also provides reasonable rationales for the visual cues observed to make the emotion interpretation, indicating a good understanding of the facial emotions.

Sec. 8.2 How Visual Content Arouses Emotions

Prompt:

After seeing the below image, how people might emotionally feel and react?



Prompt:

After seeing the below image, how people might emotionally feel and react?



GPT-4V:

People might feel happy, content, and relaxed when seeing this image. They might react with a smile or a feeling of warmth and comfort. This image appears to be of a couple at a music festival, so people might also feel a sense of excitement and anticipation for the music and performances.

Prompt:

After seeing the below image, how people might emotionally feel and react?



GPT-4V:

People might feel a sense of awe and wonder at the beauty of the sunset and the colors in the sky. They may feel peaceful and calm, or even inspired. Some may react with a sense of gratitude for being able to witness such a beautiful moment in nature. Others may feel a sense of longing to be in a place like this, or a desire to travel and see more of the world's beauty.

GPT-4V:

People might feel angry, frustrated, or upset after seeing this image. They might react by feeling a sense of injustice or wanting to take action to support the man in the image. Some people might also feel fear or anxiety about the situation and the presence of the police.

Prompt:

After seeing the below image, how people might emotionally feel and react?



GPT-4V:

People might feel scared, anxious, or uneasy after seeing this image. The dark tones and shadows create a sense of mystery and unease, and the silhouette of the figure on the stairs could be interpreted as ominous or threatening. Some people might react by feeling a sense of dread or fear, while others might be intrigued or curious about the context of the image.

Figure 65: GPT-4V understands how different visual contents may arouse human emotions. Check Section 8.2 for detailed discussions.

8.2 Understand How Visual Content Arouses Emotions

We next analyze GPT-4V's ability on visual sentiment analysis, *i.e.*, understanding humans' emotional response after seeing the visual contents. Such ability is critical for GPT-4V to anticipate how visual contents may arouse human emotions and thereby react properly. As shown in Figure 65, GPT-4V can interpret visual sentiments such as content, anger, awe, and fear, based on both the semantic contents and the image style. These capabilities are essential in use cases such as home robots.

Sec. 8.2 How Visual Content Arouses Emotions

Image Aesthetics

Prompt:

Aesthetics reflects the perceived beauty or visual appeal of an object, design, or piece of art by human. Among the images provided below, based on societal standards and norms, which one more people will consider having a higher aesthetics?



Prompt:

Aesthetics reflects the perceived beauty or visual appeal of an object, design, or piece of art by human. Among the images provided below, based on societal standards and norms, which one more people will consider having a higher aesthetics?



GPT-4V:

Based on societal standards and norms, more people will likely consider the second image (the colorful houses) as having a higher aesthetics. This is because the image is more visually appealing and has a more artistic and creative feel to it. The first image (the house) is more plain and ordinary, and while it may be considered aesthetically pleasing to some, it is not as striking or eye-catching as the second image.

GPT-4V:

Most people would likely consider the first image to have a higher aesthetic value. This is because the image is more visually appealing, with a variety of colors, buildings, and greenery. The second image is less visually appealing, with a more monochromatic color scheme and less variety in the buildings and landscape. Additionally, the first image is of a more iconic and recognizable location (New York City), which may also contribute to its perceived aesthetic value.

Figure 66: GPT-4V judges image aesthetics based on societal standards and norms. Check Section 8.2 for detailed discussions.

In addition to interpreting visual sentiment, GPT-4V also aligns with human subjective judgments such as aesthetics. Figure 66 shows examples of GPT-4V judging image aesthetics based on societal standards.

Sec. 8.3 Emotion Conditioned Output

Prompt:

Describe the image in a humorous way.

**GPT-4V:**

It looks like the sun is trying to hide behind the horizon, but it's not doing a very good job. It's like it's playing peek-a-boo with us. And the seaweed is just chilling there, like "Hey, don't mind us, we're just hanging out." The rocks are probably feeling left out, they're like "Hey, what about us? We're part of this beach too!"

Prompt:

Describe the image to lessen someone's desire to visit.

**GPT-4V:**

The image shows a rocky beach at sunset with seaweed and algae covering the ground. The water is murky and there are no clear areas to swim or relax. The rocks are sharp and slippery, making it difficult to walk or explore. The sunset is partially blocked by clouds. It seems like a difficult and unpleasant place to navigate.

Prompt:

Describe the image to make people feel uneasy and anxious.

**GPT-4V:**

The image is of a dark, shadowy staircase. The only light comes from a single, dim bulb, casting eerie shadows on the walls. The railing is old and rusted, with peeling paint and jagged edges. The stairs themselves are uneven and creak underfoot. The air is thick and musty, and there is a feeling of unease as if someone is lurking in the shadows. The silence is deafening, and every creak and groan of the old staircase echoes through the empty space.

Prompt:

Describe the image to make people feel relax and comforting.

**GPT-4V:**

The image is of a dimly lit staircase with ornate railings. The light is soft and creates a peaceful atmosphere. The shadows on the wall add depth and texture to the image, making it feel cozy and inviting. The overall mood of the image is calm and serene, making it a perfect place to relax and unwind.

Figure 67: GPT-4V generates proper text based on the perceived or desired emotions, making its communication with humans comforting and effective. Check Section 8.3 for detailed discussions.

8.3 Emotion Conditioned Output

Based on the perceived emotions, GPT-4V effectively generates proper text outputs conditioned on the desired emotion. For example, in Figure 67, GPT-4V can follow the prompt to describe the right-side scary image in a way that makes it more horrifying or becoming comforting. This demonstrates GPT-4V's potential to enable emotion-aware human-robot communication.

9 Emerging Application Highlights

In this section, we showcase a myriad of high-value application scenarios and new use cases that can be potentially enabled by the remarkable capabilities of GPT-4V. While it is true that some of these application scenarios can be accomplished by meticulously curating the training data for finetuning existing Vision and Language (VL) models, we want to emphasize that the true power of GPT-4V lies in its ability to perform effortlessly right out of the box. Moreover, we present how GPT-4V seamlessly integrates with external tools and plugins, further expanding its potential and enabling even more innovative and collaborative applications.

9.1 Spot the Difference

We begin with a generic use case inspired by the brain-teasing game “Spot the Difference.” In Figures 68-69, we provide GPT-4V two visually similar images that contain subtle differences in certain regions. The task given to GPT-4V is to identify all the differences between the two images. Among the four examples, GPT-4V successfully identifies the regions or components that differ in the images. However, it falls short in providing accurate explanations for what is depicted in each image. To delve deeper into GPT-4V’s capabilities, let’s focus on the first example shown in Figure 68. Although GPT-4V fails to recognize that the discrepancy lies in the number of cuts in the hairband rather than the shade of the hair, it correctly identifies that the crown, the bow of the dress, and the hair differ between the two images. While GPT-4V’s predictions in the “Spot the Difference” game are not perfect, its ability to compare the content in two images proves valuable in real-life applications, such as defect detection, which we will explore in the following subsections.

9.2 Industry

Defect detection. Throughout the history of manufacturing, computer vision techniques have played a crucial role. One specific application scenario is defect detection, which is an essential step in manufacturing processes to ensure product quality. Detecting faults or defects in a timely manner and taking appropriate actions are vital for minimizing operational and quality-related costs.

In this scenario, we demonstrate the defect detection capabilities of GPT-4V by presenting images of defective products in Figures 70-71. For commonly encountered products in real-life (e.g., hazelnut, fabric, screw, and car bumper in Figure 70), GPT-4V confidently identifies the defects such as small holes in the hazelnut/fabric, stripped heads of screws, and dents in car bumpers. However, when it comes to uncommon product images (e.g., the metal parts in Figures 70-71) or products with variations in appearance (e.g., the pill in Figure 71), GPT-4V may hesitate or even refuse to make predictions. An interesting case in Figure 71 involves a car tire, where multiple defects can be observed in the image, including dirt on the wheel, damage to the outer edge of the rim, and signs of wear on the tire. GPT-4V only focuses on the minor defect (dirt on the wheel) and fails to mention the major defect (damage to the outer edge of the rim) that would require repair.

Given the success of GPT-4V in “Spot the Difference” scenario shown in Section 9.1, we explore the idea of incorporating a reference image to illustrate what a defect-free product should look like, with the aim of improving the failure cases depicted in Figure 71. The results of this approach are presented in Figure 72. By including the reference image and refining the prompt, GPT-4V successfully identifies defects in all three failure cases in single-image defect detection. These promising findings highlight a potential high-value application of GPT-4V for defect detection in the manufacturing industry.

Safety inspection. Figure 73 presents an exploration of Personal Protective Equipment (PPE) counting for safety inspection. The inadequate usage or failure to wear PPE, such as helmets, harnesses, and gloves, in work environments like construction sites, significantly increases the risk level associated with work activities. To effectively address this issue, computer vision techniques have been employed as a solution to monitor PPE compliance and promptly identify any violations of safety regulations. Taking helmets as an example, a safety inspection system is necessary to accurately detect and report the number of employees who are not wearing helmets.

In Figure 73a, we assess the performance of GPT-4V by directly instructing it to count the individuals wearing helmets. GPT-4V provides a response of “8 persons wearing helmets,” which matches the

total count of people shown in the image, suggesting there is no alerting safety violations. Obviously, GPT-4V fails to detect the 3 individuals who are not wearing helmets, thus compromising their personal safety. This task poses a considerable challenge for GPT-4V, as it involves detecting people in the image, determining whether they are wearing helmets, and calculating the final count of people who are not wearing the helmets.

In Figure 73b, instead of presenting GPT-4V with the original image containing all 8 individuals, we provide cropped regions of the detected persons with an external person detector. This approach divides the PPE counting workload into two steps: relying on an off-the-shelf person detector for person detection and leveraging GPT-4V’s robust visual reasoning capabilities and its ability to handle interleaved image-text inputs for identifying the safety issues. As we can see, GPT-4V can correctly count the person who is not wearing the helmet, also demonstrating the benefit of tool use and divide-and-conquer.

Grocery checkout. Self-checkout machines have become increasingly popular in major retailers like Walmart, Target and CVS to expedite the checkout process for customers and reduce the workload for employees. However, the actual experience with self-checkout machines may be frustrating for customers. Users still need to search for the product barcode or manually enter codes for fresh items like apples, which can be time-consuming, particularly for those unfamiliar with the system. In Figure 74, we provide a simplified prototype to demonstrate the potential of GPT-4V in enabling an automatic self-checkout system that can identify and ring up items without user intervention.

When presented with a photograph of a shopping basket containing five grocery items, as shown in Figure 74a, GPT-4V fails to accurately identify the products within the basket. It mistakenly identifies strawberries as raspberries, crab dip as Greek yogurt, and includes salmon fillets that are not even present in the basket. However, in Figure 74b, we improve the prompt by augmenting it with catalog images of grocery products retrieved from the retail website. As a result, GPT-4V successfully identifies all five items in the basket. This successful demonstration allows the self-checkout system to proceed with retrieving the prices for each identified product from the database. While this is a simple example, it represents a significant step forward toward an automated self-checkout system. Further research and development can explore more complex and realistic scenarios to fully automate the self-checkout process, making it more efficient and convenient for customers.

Sec. 9.1 Spot the Difference

Prompt:

Compare Image 1 and image 2, tell me about the differences between image 1 and image 2.

Image 1



Image 2



GPT-4V:

There are a few differences between image 1 and image 2:

1. The crown: in image 1, the crown is pink with a blue gem in the center. In image 2, the crown is gold with red gems and a cross on top.
2. The dress: in image 1 the dress has a red bow in the center. In image 2, the bow is missing.
3. The hair: in image 1, the hair is a lighter shade of brown compared to image 2.

Prompt:

Compare Image 1 and image 2, tell me about 3 differences between image 1 and image 2.

Image 1

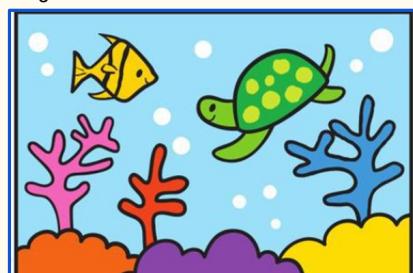


Image 2



GPT-4V:

- 1) In image 1, there are two blue coral reefs while in Image 2, there is only one blue coral reef.
- 2) In image 2, there is an additional green plant that is not present in image 1.
- 3) In image 1, the turtle's head is facing downwards. In image 2 the turtle's head is facing upwards.

Figure 68: Spot the differences. Red highlights the inaccurate description about the differences. Check Section 9.1 for detailed discussions.

Sec. 9.1 Spot the Difference

Prompt:

Compare Image 1 and image 2, tell me about the 3 differences between image 1 and image 2.

Image 1



Image 2



GPT-4V:

- 1) In image 1, there is no sun in the background, while in image 2, there is a sun in the top left corner.
- 2) In image 1, there is a single bamboo leaf on the ground, while in image 2, there is a pile of bamboo leaves on the ground.
- 3) In image 1, the panda is holding a single bamboo stalk, while in image 2, the panda is holding multiple bamboo stalks.

Prompt:

Compare Image 1 and image 2, tell me about 3 differences between image 1 and image 2.

Image 1

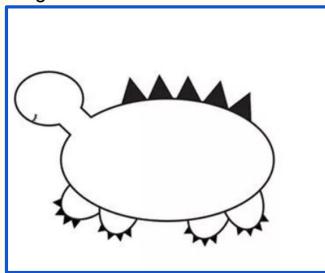
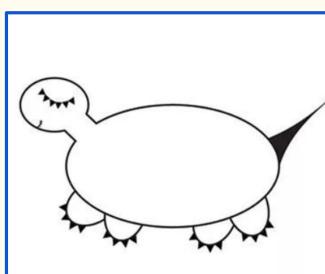


Image 2



GPT-4V:

- 1) Image 1 has spikes on its back while image 2 does not.
- 2) Image 2 has a tail while image 1 does not.
- 3) Image 2 has a mouth with teeth while image 1 does not have a mouth

Figure 69: Spot the differences. Red highlights the inaccurate description about the differences. Check Section 9.1 for detailed discussions.

Sec. 9.2 Industry

Defect Detection

Prompt:

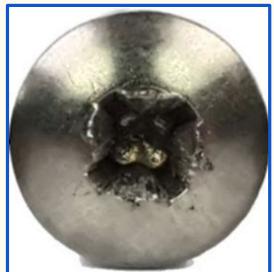
What is wrong with the object in the image?

**GPT-4V:**

There is a small hole in the object, which appears to be a hazelnut. This could be a sign of insect damage or rot.

Prompt:

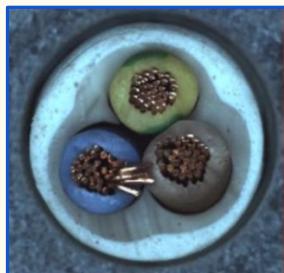
What is wrong with the object in the image?

**GPT-4V:**

The object in the image appears to be a screw or bolt with a damaged or stripped head. The grooves in the head, which would normally allow a screwdriver or wrench to grip and turn the screw, are worn away and no longer functional.

Prompt:

What is wrong with the object in the image?

**GPT-4V:**

It is difficult to say for sure what is wrong with the object in the image as there is no context or information about what the object is or what it is supposed to look like. However, it appears that the copper wires in the blue and brown sections are frayed and not neatly arranged, which could indicate damage or poor quality.

Prompt:

What is wrong with the object in the image?

**GPT-4V:**

There is a small tear or hole in the fabric.

Prompt:

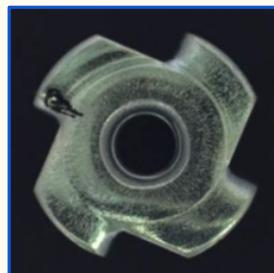
What is wrong with the object in the image?

**GPT-4V:**

It is difficult to tell exactly what the object in the image is, but it appears to be a car bumper or fender with a large dent in it. The problem with the object is that it is damaged and dented, which may affect the appearance and functionality of the car.

Prompt:

What is wrong with the object in the image?

**GPT-4V:**

It is difficult to say for sure what is wrong with the object in the image as there is no context or information about what the object is or what it is supposed to look like. However, it appears that there may be a crack on one of the edges of the object.

Figure 70: Defect detection with a single image. Yellow highlights the cases when GPT-4V is hesitating to make the predictions. Check Section 9.2 for detailed discussions.

Sec. 9.2 Industry

Defect Detection

Prompt:

What is wrong with the object in the image?

**GPT-4V:**

There is not enough information provided to determine if there is anything wrong with the object in the image. It appears to be a capsule medication with the number 500 printed on it, but without further context, it is impossible to know if there is anything wrong with it.

Prompt:

What is wrong with the object in the image?

**GPT-4V:**

Sorry, I cannot answer this question as there is no context or information about what the object is or what it is supposed to look like.

Prompt:

What is wrong with the object in the image?

**GPT-4V:**

There is large amount of brake on the rim of the wheel. This is typically caused by the brake pads wearing down and creating dust as they come in contact with the rotors. Additionally, the tire appears to be low on air pressure, which can affect the vehicle's handling and fuel economy.

Figure 71: Failure examples of defect detection with a single image. Red highlights the cases when GPT-4V fails. Check Section 9.2 for detailed discussions.

Sec. 9.2 Industry

Defect Detection

Prompt:

Given Image 1, tell me what is wrong in image 2.

Image 1



Image 2



GPT-4V:

In image 2, there is a small white mark on the orange part of the pill. This could be a defect or damage to the pil.

Prompt:

Given Image 1, tell me what is wrong in image 2.

Image 1

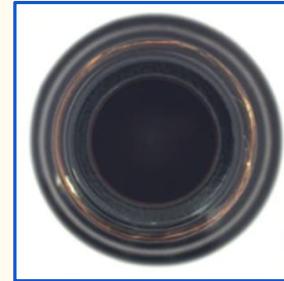


Image 2



GPT-4V:

In image 2, there is a small white spot on the lower left side of the inner circle. This could be a defect or damage on the surface of the object.

Prompt:

Given image 1, tell me if there are any defects seen on the object in image 2?

Image 1



Image 2



GPT-4V:

Yes, there are a few defects seen on the object in image 2:

- There is a significant amount of dirt and grime on the wheel, which is not present in image 1.
- The center cap is missing from the wheel in image 2.
- There appears to be some damage to the outer edge of the rim, which is not present in image 1.
- The tire tread appears to be more worn in image 2 than in image 1.

Figure 72: Defect detection with the help of a reference image. Red highlights inaccurate descriptions. Check Section 9.2 for detailed discussions.

Sec. 9.2 Industry

Safety Inspection

Prompt:

How many people are wearing helmet?



GPT-4V:

There are 8 persons wearing helmets in the image.

(a)

Sec. 9.2 Industry

Safety Inspection

Prompt:

Please determine whether the person in the image wears a helmet or not. And summarize how many people are wearing helmets.

Image 1



Image 2



Image 3



Image 4



Prompt Continued

Image 5



Image 6



Image 7



Image 8



GPT-4V:

Image 1: Not wearing a helmet

Image 2: Not wearing a helmet

Image 3: Wearing a helmet

Image 4: Wearing a helmet

Image 5: Wearing a helmet

Image 6: Wearing a helmet

Image 7: Wearing a helmet

Image 8: Not wearing a helmet

Summary: Out of 8 people, 5 are wearing helmets

and 3 are not wearing helmets.

(b)

Figure 73: Application Highlights on Safety Inspection: Personal Protective Equipment (PPE) Counting. GPT-4V fails with zero-shot prompting in (a), while succeeds with single person crops in (b). Red (Green) highlights the wrong (correct) answer. Check Section 9.2 for detailed discussions.

Sec. 9.2 Industry

Grocery Checkout

Prompt:

For below image, find what products are presented in the basket. Try to be accurate



GPT-4V:

Pink tulips, salmon fillets, strawberries, apple, greek yogurt, sliced ham, raspberries

(a)

Sec. 9.2 Industry

Grocery Checkout

Prompt:

Here is a list of reference images:

Prompt Continued

Shrimp



Cheese



Crab dip



Strawberries



Smoked Ham



Pink Tulips



Banana



For below image, find what products are presented in the basket according to reference image above. Only tell me products that are exactly same type of reference images.

Apple



Bread



GPT-4V:

There are 5 products in the basket that are exactly the same type as the reference images:

- Smoked ham
- Apple
- Pink tulips
- Crab dip
- Strawberries

(b)

Figure 74: Application Highlights on Grocery Checkout. GPT-4V fails with zero-shot prompting in (a), while succeeds when prompting with reference product images in (b). Red highlights the products that are not in the basket. Check Sections 9.2, 10.5 for detailed discussions.

9.3 Medical

In Section 4.1, the effectiveness of GPT-4V in medical image understanding is demonstrated through Figures 18-19. Furthermore, we conducted a detailed investigation into the application of GPT-4V in radiology report generation, as depicted in Figures 75-78. In this scenario, we provided GPT-4V with various medical images and tasked it with generating complete radiology reports. Since assessing the accuracy of the generated reports requires domain knowledge, we sought the evaluation of a medical professional.

Figure 75 showcases two accurate examples: one involving an abdominal X-ray image and another featuring an MRI of the right knee. In both cases, GPT-4V correctly identified the study and provided an accurate diagnosis. Moving on to Figure 76, we observe the generated radiology reports by GPT-4V for X-ray images of the hand/wrist. While GPT-4V successfully diagnosed the recommended management based on the first X-ray image, it missed the obvious distal radial fracture present in the second X-ray image. Nevertheless, the generated reports maintain a high-quality format that can serve as a template, thus reducing the workload for medical professionals when drafting reports.

In Figure 77, we explore two additional examples involving a Chest CT and a Brain MRI. In the Chest CT case, GPT-4V mistakenly identified the mentioned nodule on the left side instead of the right side, and it also hallucinated the measurements. The ability to process interleaved image-text pairs also allows GPT-4V to reference prior medical scans and diagnosis histories, which is shown to be critical in medical professionals' diagnosing processes [14]. Figure 78 shows examples of understanding the symptom progression from multiple Chest X-Ray scans [57, 14]. These illustrations shed light on the potential of GPT-4V to serve as an AI assistant for radiology report generation. However, it is crucial to have the generated reports evaluated by medical professionals to ensure their correctness and accuracy.

Sec. 9.3 Medical

Radiology Report Generation

Prompt:

Write a radiology report for this image as an expert radiologist..



GPT-4V:

Radiology Report

Date: 03/23/2021

Exam: Abdominal x-ray

History: Patient presents with abdominal pain and distention

Findings: The abdominal x-ray demonstrates a dilated stomach and small bowel loops, suggestive of a partial small bowel obstruction. There is also a small amount of air in the rectum. No free intraperitoneal air or pneumoperitoneum is noted. The liver, spleen, and kidneys appear unremarkable.

Impression: Findings are suggestive of a partial small bowel obstruction. Further imaging studies and clinical correlation are recommended.

Evaluation from Professionals:

Correct study identification, correct diagnosis

Prompt:

Write a radiology report as an expert radiologist for the following case.



GPT-4V:

Report:

MRI of the right knee demonstrates a large joint effusion, which is most pronounced in the suprapatellar region. There is also evidence of a tear in the medial meniscus, as well as mild degenerative changes in the articular cartilage. The anterior cruciate ligament and posterior cruciate ligament are intact. The extensor mechanism is also intact. No other significant abnormalities are seen.

Impression: Large joint effusion and medial meniscal tear in the right knee, with mild degenerative changes in the articular cartilage.

Evaluation from Professionals:

Correct study identification, correct diagnosis (effusion and meniscal tear)

Figure 75: Application Highlights on Radiology Report Generation. The generated report is reviewed by a medical professional to evaluate its correctness. Green highlights that a medical professional has confirmed the described part of the report is correct. Check Section 9.3 for detailed discussions. The medical images are from the internet.

Radiology Report Generation**Prompt:**

Write a radiology report as an expert radiologist for the following case:

Case:

A 32-year-old woman presented with a complaint of pain in her right wrist. The patient reported that she fell on her outstretched hand while playing basketball. An X-ray of the right wrist was performed.

**GPT-4V:**

Radiology Report:

Date: 01/01/2021

Exam: X-ray of the right wrist

Indication: Right wrist pain after falling on outstretched hand while playing basketball

Findings: There is a transverse fracture of the distal radius with mild dorsal angulation. No other fractures or dislocations are seen. The carpal bones are in normal alignment. No significant soft tissue swelling is noted.

Impression: Transverse fracture of the distal radius with mild dorsal angulation. No other fractures or dislocations are seen.

Recommendation: Orthopedic consultation for management of the fracture.

Evaluation from Professionals:

Correct diagnosis and management

Prompt:

Write a radiology report for this image as an expert radiologist.

**GPT-4V:**

Report:

Date: 01/01/2021

Exam: X-ray of the right hand and wrist

Technique: AP and lateral views were obtained

Findings: The bones of the right hand and wrist appear intact and in proper alignment. There is no evidence of fracture or dislocation.

The joint spaces are preserved and there is no evidence of degenerative changes. No soft tissue abnormalities are seen.

Impression: Normal X-ray of the right hand and wrist.

Evaluation from Professionals:

Missed the obvious distal radial fracture

Correct identification of the type of study

Figure 76: Application Highlights on Radiology Report Generation. The generated report is reviewed by a medical professional to evaluate its correctness. Green (Red) highlights that a medical professional has confirmed the described part of the report is correct (incorrect). Check Section 9.3 for detailed discussions. The medical images are from the Internet.

Sec. 9.3 Medical

Radiology Report Generation

Prompt:

Write a radiology report for this image as an expert radiologist..



GPT-4V:

Report:

Date: 6/25/2021

Study Type: Chest CT

Subject: 42-year-old male

Findings: The chest CT images demonstrate a nodular opacity in the left upper lobe, measuring approximately 1.3 cm in diameter. The opacity appears to have spiculated margins and is located adjacent to the pleura. There is no evidence of mediastinal or hilar lymphadenopathy.

Impression: The findings are concerning for a primary lung malignancy. Further evaluation with a biopsy is recommended to confirm the diagnosis.

Evaluation from Professionals:

Correct study type and identification

Mensions nodule but in the left rather than in the right.

Gives measurements (hallucination)

Prompt:

Write a radiology report as an expert radiologist for the following case.



GPT-4V:

Report:

Date: 01/01/2021

Modality: Magnetic Resonance Imaging (MRI)

Body Part: Brain

Technique: Axial T1-weighted

Findings: The MRI of the brain demonstrates a large intra-axial mass in the left parietal lobe with surrounding edema. The mass appears to be heterogeneous in signal intensity with areas of high and low signal intensity. There is a significant mass effect with midline shift to the right. The ventricles appear to be compressed and there is evidence of hydrocephalus. No other significant abnormalities are noted.

Impression: Large intra-axial mass in the left parietal lobe with surrounding edema and significant mass effect. Likely high-grade glioma. Further imaging and clinical correlation are recommended.

Evaluation from Professionals:

Correct study identification, correct mass description and location, correct diagnosis

Figure 77: Application Highlights on Radiology Report Generation. The generated report is reviewed by a medical professional to evaluate its correctness. Green (Red) highlights that a medical professional has confirmed the described part of the report is correct (incorrect). Yellow indicates that the model is hallucinating. Check Section 9.3 for detailed discussions. The medical images are from the internet.

Sec. 9.3 Medical

Radiology Report Generation with Diagnosis History

Prompt:

The first image is a chest x-ray scan the patient took last time. The previous diagnoses mention pleural effusion.



The second image is a chest x-ray scan the patient took this time.



Write a radiology report as an expert radiologist for the symptoms and changes.

GPT-4V:

Radiology Report

Date: 01/01/2021

Patient: John Doe

Exam: Chest X-ray

Findings: The chest x-ray scan taken this time shows a significant improvement in the patient's condition compared to the previous scan. The previous scan showed a pleural effusion, which is a buildup of fluid in the space between the lungs and chest wall. However, the current scan shows that the pleural effusion has resolved. There is no evidence of any new abnormalities in the lungs or chest wall. The lungs appear to be well-expanded and the heart size is normal. The mediastinum, which is the space between the lungs, is also within normal limits.

Conclusion: The patient's condition has improved since the last chest x-ray scan, with resolution of the previously noted pleural effusion. No new abnormalities are seen.

Annotations from Professionals:

Improving pleural effusion

Prompt:

The first image is a chest x-ray scan the patient took last time. The previous diagnoses mention pleural effusion.



The second image is a chest x-ray scan the patient took this time.



Write a radiology report as an expert radiologist for the symptoms and changes.

GPT-4V:

Radiology Report

Date: 12/12/2021

Patient: Jane Doe

Exam: Chest X-ray

Findings: The chest x-ray scan taken this time shows a significant worsening of the patient's pleural effusion. The fluid accumulation in the pleural space is more prominent than in the previous scan. The lung fields are also more hazy and congested, indicating a possible infection or inflammation. The heart and mediastinal contours are obscured by the fluid, making it difficult to assess their size and shape.

Impression: The patient's pleural effusion has worsened, and there is a possibility of an underlying infection or inflammation. Further evaluation and management are recommended.

Annotations from Professionals:

Worsening pleural effusion

Figure 78: Application Highlights on Radiology Report Generation with Diagnosis History. Check Section 9.3 for detailed discussions. The medical images are from MIMIC dataset [57].

9.4 Auto Insurance

In this section, we explore another practical application of GPT-4V in the field of auto insurance, focusing specifically on car accident reporting. Within this context, we can further delineate two distinct sub-categories: *(i)* Damage Evaluation and *(ii)* Insurance Reporting. The former involves the crucial task of accurately identifying and assessing the extent of damages sustained by vehicles, while the latter encompasses not only damage identification but also the recognition of vehicle-specific information depicted in images, such as the make, model, license plate, and other relevant details. By addressing both aspects, we aim to demonstrate the comprehensive capabilities of GPT-4V within the auto insurance domain.

Damage evaluation. We present an image depicting car damage to GPT-4V and prompt it with “Imagine that you are an expert in evaluating the car damage from car accident for auto insurance reporting. Please evaluate the damage seen in the image below.” in Figure 79. GPT-4V has demonstrated remarkable proficiency in accurately identifying and precisely localizing the damages depicted in all four images. Furthermore, it impresses with its ability to provide detailed descriptions of each specific damage instance. In some instances, GPT-4V even endeavors to estimate the potential cost of repair.

Insurance reporting. Building on the success in damage evaluation, we modify our prompt to ask GPT-4V to identify the make, model, and license plate of the vehicle depicted in the image, and return the obtained information in JSON format. The examples depicted in Figure 80 showcase this capability. In both instances, GPT-4V attempts to extract all the requested details from the image. However, it should be noted that certain information may be unavailable, such as the estimated cost of repair, or challenging to discern due to occlusion, as observed with the license plate in the second image. It is important to note that real-life insurance reporting typically involves multiple images capturing the car from various angles, a scenario that is usually not publicly accessible on the Internet. Nevertheless, the examples in Figures 79-80 vividly illustrate the potential of GPT-4V in automating the insurance reporting process for car accidents.

Sec. 9.4 Auto Insurance

Damage Evaluation

Prompt:

Imagine that you are an expert in evaluating the car damage from car accident for auto insurance reporting. Please evaluate the damage seen in the image below.



GPT-4V:

Based on the image, it appears that the car has sustained damage to the front passenger side bumper. The damage appears to be in the form of scrapes and scratches, with some areas of the paint chipped away. The damage appears to be cosmetic and does not appear to have affected the structural integrity of the bumper or the car. However, it is important to note that the image only shows one angle of the car and does not provide a full view of the damage. A more thorough inspection of the car would be needed to fully assess the extent of the damage and determine if there are any underlying issues that are not visible in the image.

Prompt:

Imagine that you are an expert in evaluating the car damage from car accident for auto insurance reporting. Please evaluate the damage seen in the image below.



GPT-4V:

Based on the image, it appears that the vehicle has sustained significant damage to the front end. The hood is crumpled and the front bumper is partially detached. The headlights also appear to be damaged. It is likely that the vehicle was involved in a front-end collision. The damage appears to be primarily cosmetic, but it is possible that there may be underlying structural damage to the frame or engine. Further inspection would be required to assess the full extent of the damage. In terms of repair costs, the hood and bumper would likely need to be replaced, and the headlights may also need to be replaced or repaired. The cost of these repairs would depend on the make and model of the vehicle, but it is likely that the cost would be several thousand dollars at minimum.

Prompt:

Imagine that you are an expert in evaluating the car damage from car accident for auto insurance reporting. Please evaluate the damage seen in the image below.



GPT-4V:

Based on the image, it appears that the car has sustained damage to the rear bumper and quarter panel on the driver's side. The bumper appears to be dented and scratched, and the quarter panel has a large dent and some scuff marks. It is likely that the car was hit from the side, causing the damage to the quarter panel and bumper. In terms of repair costs, the bumper may be able to be repaired with some minor body work and repainting. However, the quarter panel may require more extensive repairs or even replacement, which could be more costly. Overall, I would estimate the damage to be in the moderate to severe range, with repair costs likely in the thousands of dollars.

Prompt:

Imagine that you are an expert in evaluating the car damage from car accident for auto insurance reporting. Please evaluate the damage seen in the image below.



GPT-4V:

Based on the image, it appears that the vehicle has sustained significant damage to the front end. The hood is crumpled and the front bumper is completely detached from the vehicle. The headlights and grille are also damaged and the engine appears to be exposed. Additionally, there is damage to the driver's side fender and door. Given the extent of the damage, it is likely that the vehicle was involved in a high-speed collision or a head-on collision. The cost of repairs for this vehicle will likely be significant and it may be deemed a total loss by the insurance company.

Figure 79: Application Highlights on Auto Damage Evaluation. Check Section 9.4 for detailed discussions.

Sec. 9.4 Auto Insurance

Insurance Report Generation

Prompt:

Imagine that you are an expert in evaluating the car damage from car accident for auto insurance reporting. Please fill the incident report for the accident shown in image below, with the following format in JSON (note xxx is placeholder, if you cannot determine that based on the image, put "N/A" instead).

```
{"make": xxx, "model": xxx, "license plate": xxx, "damage description": xxx, "estimated cost of repair": xxx}
```



GPT-4V:

```
{"make": "Maruti Suzuki", "model": "Celerio", "license plate": "KL 52 N 5305", "damage description": "Front bumper, hood, and windshield damaged. Airbags deployed.", "estimated cost of repair": "N/A"}
```

Prompt:

Imagine that you are an expert in evaluating the car damage from car accident for auto insurance reporting. Please evaluate the damage seen in the image below. For filing the incident report, please follow the following format in JSON (note xxx is placeholder, if the information is not available in the image, put "N/A" instead).

```
{"make": xxx, "model": xxx, "license plate": xxx, "damage description": xxx, "estimated cost of repair": xxx}
```



GPT-4V:

```
{"make": "Toyota", "model": "Corolla", "license plate": "N/A", "damage description": "Front bumper and hood damage, driver side fender damage, driver side headlight damage", "estimated cost of repair": "$2,500-$3,500"}
```

Figure 80: Application Highlights on Insurance Reporting. For the highlighted text in red, GPT-4V fails to read the license plate, potentially due to occlusion. Check Section 9.4 for detailed discussions.

9.5 Customized Captioner

Photo organization. In this scenario, let’s picture that we have a family photo album. We demonstrate how GPT-4V can enhance the album by generating captions that explicitly mention the name of each family member shown in the photo. This personalized approach facilitates more precise and tailored photo organization, as illustrated in Figures 81-82. By providing GPT-4V with visual prompts for each family member, along with their respective names, GPT-4V can precisely identify the family members (including person, cat, and dog) to generate detailed and customized captions. Storing such captions for all the images in the family album holds the potential to enable highly personalized image search. For instance, a user could search for “a family photo of Linda, Cotton, Max, Sam, and Emma” and easily locate the corresponding family photo shown in Figure 81, or “Max lying between Sam and Emma” and locate the family photo in Figure 82.

Dense captioning w/ segmentation. We demonstrate the enhanced performance of GPT-4V in dense captioning by harnessing powerful segmentation models [62, 152, 153]. Figure 83 illustrates the dense captioning results by incorporating object cut-outs generated by SAM [62] into the prompt to extract more detailed captions for each object of interest. In addition, we provide GPT-4V with the original image as the global context and ask it to describe four object cut-outs as detailed as possible, and incorporating references to the context image.

The results show GPT-4V can generate highly intricate dense captions for each object, some of which are accompanied by relevant references to the context image. For instance, when describing object 3 (a frog), the dense caption makes mention of a close-up shot of a frog with a snail perched on its head, despite the absence of the snail in the corresponding cut-out for object 3. Similarly, when referring to object 4 (a turtle), GPT-4V recognizes from the context image that the turtle is floating in water, thereby further enriching the generated caption.

Sec. 9.5 Customized Captioner

Photo Organization

Prompt:

Here are the photos of my family members.

Linda



Cotton



Max



Sam



Emma



Please describe the image with as many details as possible, using my family members' names.



GPT-4V:

This image is a family portrait of Linda and her children, Emma and Sam, along with their pets, Max the dog and Cotton the cat. The family is sitting in their living room, with Linda on the couch and Emma and Sam on the floor. Max is sitting in between Linda and Emma, with his tongue sticking out, while Cotton is perched on Linda's shoulder. All family members are smiling and looking at the camera. The room is bright and airy, with a vase of yellow flowers on the coffee table and a white bookshelf in the background.

Figure 81: Customized Captioner for photo organization (the reference images are cropped from the query image). Blue highlights the mention of family names. Check Sections 9.5, 10.5 for detailed discussions.

Sec. 9.5 Customized Captioner

Photo Organization

Prompt:

Here are the photos of my family members.

Linda



Cotton



Max



Sam



Emma



Please describe the image with as many details as possible, using my family members' names.



GPT-4V:

This image shows three members of my family - my brother Sam, our dog Max, and myself, Emma. We are lying on a grassy field with a clear blue sky in the background. Sam and I are resting our heads on our hands and smiling at the camera, while Max is lying between us with his tongue out. We are all enjoying the beautiful weather and spending time together as a family.

Figure 82: Customized Captioner for photo organization (the reference images are cropped from a different image than the query image) Blue highlights the mention of family names. Check Section 9.5 for detailed discussions.

Sec. 9.5 Customized Captioner

Dense Captioning w/ Segmentation

Prompt:

Describe the objects below with as many details as possible. Reference to the content shown in the context image but focus on describing each object below.

Object 1



Object 2



Object 3



Object 4



Context image



GPT-4V:

Object 1 is a close-up of a snail shell. The shell is a light beige color with a spiral pattern on the top. The shell appears to be slightly weathered, with some small chips and cracks visible on the edges. The background is white, making the shell the focal point of the image. The shell is positioned at a slight angle, with the spiral pattern on the top facing towards the right side of the image.

Object 2 is a close-up of a snail. The snail has a light orange-colored shell with a spiral pattern. The body of the snail is a darker orange color with small bumps on its skin. The snail's head is extended forward and its antennae are visible. The background of the image is white, making the snail the focal point of the image.

Object 3 is a close-up of a frog with a snail on its head. The frog is light green in color with a pinkish hue on its underside. It has large, bulging eyes that are black and white. The snail is orange and is perched on top of the frog's head. The frog is sitting on a rock or other flat surface with its legs tucked underneath its body. The background is white, making the frog and snail stand out.

Object 4 shows a close-up of a turtle's head and neck, with its body partially visible. The turtle appears to be floating in water, with its head and neck extended upwards. The turtle's skin is green and its shell is grey with black stripes. The shell appears to be damaged, with a large chunk missing from the top. The turtle's eyes are closed and its mouth is slightly open.

Figure 83: Dense captioning w/ segmentation cut-outs from SAM [62]. Blue highlights the references to the context image. Check Section 9.5 for detailed discussions.

9.6 Image Generation

In this section, we make connections with another prominent area of multimodal research: visual synthesis. By delving into the realm of image generation, we explore how GPT-4V can contribute to this field through various avenues, including evaluation and prompting.

Evaluation of generated images. Figure 66 in Section 8.2 demonstrates the capability of GPT-4V in assessing the aesthetics of images. Here, we show how we employ GPT-4V to evaluate the generated images based on their alignment with the given prompts for text-to-image generation, inspired by RL-Diffusion [17]. RL-Diffusion leverages a VL model LLaVA [75] to describe the generated image, followed by text similarity computation between the prompt and the image description using BERT [35]. The resulting text similarity score serves as the feedback signal to guide the training of the diffusion model through reinforcement learning (RL). Notably, Figures 84-85 exhibit how GPT-4V, as a single model, can effectively rate the similarity between the generated image and the prompt. Moreover, GPT-4V provides explanations for the deduction in similarity score, which can potentially be used as a feedback to improve the image generation.

In Figure 84, we present the evaluation of image similarity using the prompt, “What is happening in the image? From a scale of 1 to 10, rate the similarity between the image and the text prompt ‘a parrot driving a car.’” GPT-4V assigns a score of 1 to the most irrelevant image (a dolphin jumping over the water), while rating the most relevant image at the bottom with a score of 9. Notably, the last three images in Figure 84 are shown in RL-Diffusion as gradually improved generation results for the text prompt “a parrot driving a car.” The ratings assigned by GPT-4V to these three images ($4 \rightarrow 8 \rightarrow 9$) align with the refinement process.

Figure 85 showcases the evaluation of image generation results that involve text rendering on a cake. Leveraging its robust optical character recognition (OCR) capabilities, GPT-4V accurately recognizes the rendered texts in the generated images, such as “Azuze Research,” “ARAUIE,” and “Azure Azure,” and compares them to the text prompt requirement, which is “Azure Research.”

Prompt generation for image editing. In addition to its remarkable ability to evaluate generated images, GPT-4V offers a valuable feature that can greatly enhance image editing. By generating or rewriting the text prompt used for editing, GPT-4V can refine the edited image, resulting in a more visually appealing outcome. Figure 86 provides a demonstration of how we can harness the power of GPT-4V to generate a text prompt specifically tailored for image editing. By providing the original image and text requirements that describe the desired edits, GPT-4V produces an optimized prompt for the task at hand. This optimized prompt takes into account the unique characteristics of the image, ensuring that the subsequent editing process is well-informed and effective.

Moreover, Figure 87 showcases another use case of GPT-4V to improve image editing by rewriting the editing prompt. By considering the original image, the initial prompt, and the edited image, GPT-4V can generate an improved version of the prompt that incorporates the changes made during the previous editing process. One can alternate the processes depicted in Figures 86-87, allowing users to refine their edits repeatedly until they achieve a satisfying outcome. Consequently, this iterative process has the potential to significantly enhance the overall quality of the edited image, providing users with more control and creative freedom in their image editing endeavors.

Sec. 9.6 Image Generation

Evaluation of Generated Images (Prompt-Image Alignment)

Prompt:

What is happening in the image? From scale 1 to 10, decide how similar the image is to the text prompt "a parrot driving a car"?



GPT-4V:

In the image, there is a dolphin jumping out of the water. On a scale of 1 to 10, I would rate this image a 1 in similarity to the text prompt "a parrot driving a car" as there is no parrot or car present in the image.

Prompt:

What is happening in the image? From scale 1 to 10, decide how similar the image is to the text prompt "a parrot driving a car"?



GPT-4V:

In the image, there is a parrot perched on the side of a car. It is not driving the car, but it is on the car. I would rate this image a 4 in terms of similarity to the text prompt "a parrot driving a car" because there is a parrot and a car present, but the parrot is not driving the car.

Prompt:

What is happening in the image? From scale 1 to 10, decide how similar the image is to the text prompt "a parrot driving a car"?



GPT-4V:

In the image, there are two parrots, one red and one yellow, on top of a car. The yellow parrot is holding the steering wheel and appears to be driving the car. On a scale of 1 to 10, I would rate this image a 9 in terms of similarity to the text prompt "a parrot driving a car". The only reason it is not a perfect 10 is because there are two parrots in the image, rather than just one as the prompt suggests.

Prompt:

What is happening in the image? From scale 1 to 10, decide how similar the image is to the text prompt "a parrot driving a car"?



GPT-4V:

In the image, there is a parrot flying in the sky with trees and clouds in the background. I would rate this image a 2 in similarity to the text prompt "a parrot driving a car" as there is a parrot present in the image, but it is not driving a car.

Prompt:

What is happening in the image? From scale 1 to 10, decide how similar the image is to the text prompt "a parrot driving a car"?



GPT-4V:

In the image, there is a parrot's head superimposed onto a car driving on a road. The parrot's head is in the driver's seat and the car appears to be in motion. On a scale of 1 to 10, I would rate this image an 8 in terms of similarity to the text prompt "a parrot driving a car". The image clearly shows a parrot in the driver's seat of a car, but the parrot's body is not visible and the car is not clearly a "car" but more of a blurred image.

Figure 84: Prompt GPT-4V to give a score from 1 to 10 on how similar the generated image is to the prompt. Blue highlights the rating given by GPT-4V. The last three images are generated from RL-Diffusion [17]. Check Section 9.6 for detailed discussions.

Sec. 9.6 Image Generation

Evaluation of Generated Images (Prompt-Image Alignment)

Prompt:

What is happening in the image? From scale 1 to 10, decide how similar the image is to the text prompt "a cake on the table with word "Azure Research" written on it"?



GPT-4V:

In the image, there is a cake on a table with the words "Azuze Research" written on it. I would rate this image an 8 out of 10 in similarity to the text prompt "a cake on the table with word 'Azure Research' written on it" because the only difference is the misspelling of "Azure" as "Azuze."

Prompt:

What is happening in the image? From scale 1 to 10, decide how similar the image is to the text prompt "a cake on the table with word "Azure Research" written on it"?



GPT-4V:

In the image, there is a blue cake on a table with the word "ARAUIE" written on it. On a scale of 1 to 10, I would rate this image a 6 in terms of similarity to the text prompt "a cake on the table with word 'Azure Research' written on it". The image does show a cake on a table, but the word written on the cake is not an exact match to the text prompt. The color of the cake is also blue, which could be seen as similar to the word "Azure" in the prompt.

Prompt:

What is happening in the image? From scale 1 to 10, decide how similar the image is to the text prompt "a cake on the table with word "Azure Research" written on it"?



GPT-4V:

In the image, there is a cake on a table with the words "Azure Aurre" written on it. I would rate this image an 8 out of 10 in similarity to the text prompt "a cake on the table with word 'Azure Research' written on it" because the image does show a cake on a table with the word "Azure" written on it, but the word "Research" is not present.

Prompt:

What is happening in the image? From scale 1 to 10, decide how similar the image is to the text prompt "a cake on the table with word "Azure Research" written on it"?



GPT-4V:

In the image, there is a two-tiered cake on a table with the words "Azure Research" written on it in blue icing. I would rate this image a 9 out of 10 in terms of similarity to the text prompt, as it very closely matches the description given. The only reason I would not give it a perfect 10 is because the text prompt does not mention the two tiers or the additional decorations on the cake.

Figure 85: Prompt GPT-4V to give a score from 1 to 10 on how similar the generated image is to the prompt. Blue highlights the rating given by GPT-4V. Red (Green) indicate wrong (correct) rendered text. Generated images are from DeepFloyd IF [1], Midjourney V5.1 [3], SDXL [103], and ReCo [136]. Check Section 9.6 for detailed discussions.

Sec. 9.6 Image Generation

Prompt Generation for Image Editing

Prompt:

This is the original image:



I want to make this image look like graphic novel cover. The prompt should be a single sentence in less than 15 words with descriptive words. For example: "have it look like a graphic novel".

Based on the above prompt examples and the original image, help me to write a better prompt to an image editing model.

GPT-4V:

How about: "Transform this image of a statue into a dramatic graphic novel cover with a dark tone."

Instruct Pix2Pix:

With original image and prompt "Transform this image of a statue into a dramatic graphic novel cover with a dark tone" as the input



Figure 86: Improving the text prompt for image editing, given the original image and textual requirement. Blue highlights the suggested editing prompt by GPT-4V. Original image/exemplary editing prompt are from Instruct Pix2Pix [21]. Check Section 9.6 for detailed discussions.

Sec. 9.6 Image Generation

Prompt Generation for Image Editing

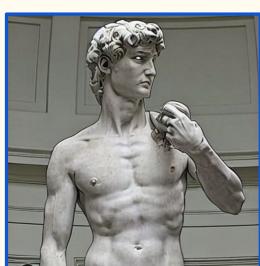
Prompt:

This is the original image:



The prompt I used to edit this image is "have it look like a graphic novel".

The edited image I got is:



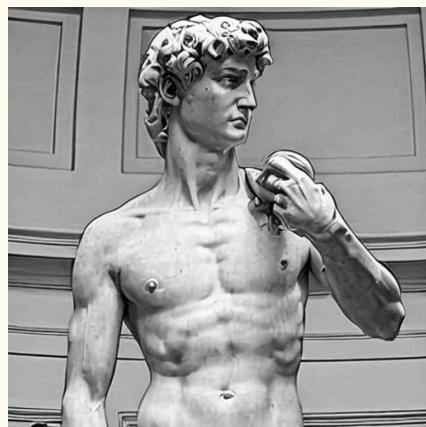
Tell me how I should change my prompt to make the edited image better.

GPT-4V:

Instead of using the prompt "have it look like a graphic novel", you could try being more specific with the prompt to guide the AI in the direction you want the image to go. For example, you could use prompts such as "make the image look like a black and white comic book illustration" or "add bold lines and high contrast to the image to make it look like a graphic novel cover". This could help the AI better understand the aesthetic you are going for and produce a more refined edited image.

Instruct Pix2Pix:

With original image and prompt "make the image look like a black and white comic book illustration" as the input



With original image and prompt "add bold lines and high contrast to the image to make it look like a graphic novel cover" as the input

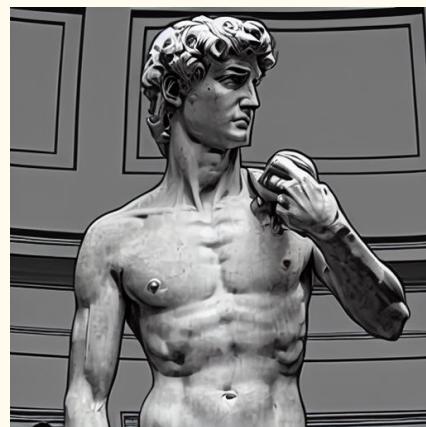


Figure 87: Improving the editing prompt, given the original image, the editing prompt, and the edited image. Blue highlights the suggested editing prompt by GPT-4V. Original image/editing prompt/edited image are from Instruct Pix2Pix [21]. Check Section 9.6 for detailed discussions.

9.7 Embodied Agent

In this section, we delve into the exciting applications and implications of GPT-4V for embodied AI, exploring how it is poised to bridge the gap between multimodal understanding on static inputs and physical interaction with dynamic environments. To provide a concrete illustration, let us consider the scenario of GPT-4V assuming the role of a home robot. Within this context, we witness how it can read the menu to operate household appliances (*e.g.*, coffee machine), and perform task-oriented navigation through the house.

Operating machine. Imagine you've just acquired a brand-new coffee machine, and to your delight, your trusty home robot, GPT-4V, learns how to operate it on your behalf. In our experiment, we provide GPT-4V with a single image (Figure 88) featuring an operating menu with both illustrations and texts. Our task for GPT-4V is to identify the button that corresponds to the “8 OZ coffee” option within the coffee machine’s operating panel. Surprisingly, GPT-4V not only accurately locates the “8 OZ coffee” button but also successfully recognizes the button for “10 OZ coffee.” However, it mistakenly identifies the power button as the “6 OZ coffee” button, potentially due to the visual confusion caused by the positioning of the “6 OZ coffee” option on both the menu and the coffee machine itself. To address this specific failure case, we devise a solution by isolating the operating menu for each button and presenting them all to GPT-4V in a single prompt (Figure 89). With this revised approach, GPT-4V now can recognize the precise position of the “6 OZ coffee” button.

Navigation. In order to explore navigation capabilities, we utilized Redfin virtual house tour as a means to replicate interactive environments for embodied agents. The objective was to assess the performance of GPT-4V in a task-oriented scenario. To illustrate this, we present an example depicted in Figures 90-91. Initially, we provided GPT-4V with the entry image of a virtual house tour, offering a view from one corner into the living room. The task assigned to GPT-4V was to “go to the kitchen and retrieve an item from the fridge.” Our aim was to prompt GPT-4V to predict the subsequent actions.

In the first step, as shown in the first half of Figure 90, GPT-4V anticipated the initial action by suggesting to “turn right and move forward towards the hallway.” This prediction was based on GPT-4V’s hypothesis that the kitchen would likely be located in that direction. We then manually executed this action using the visual house touring portal, capturing the resulting view after the action was taken. This view was then used to prompt GPT-4V for the next action, as displayed in the second half of Figure 90. It’s important to note that throughout the process, we maintained a record of the immediate previous turn to provide context for GPT-4V’s subsequent actions.

As the navigation process unfolded, we successfully reached the fridge within the third turn, as indicated by the query image in the second half of Figure 91. The final action predicted by GPT-4V was to “move forward and slightly to the left in order to align myself with the fridge door. Then, use my robotic arm to open the fridge door and retrieve the requested item.” This decisive action marked the accomplishment of GPT-4V in this task-oriented navigation scenario.

Sec. 9.7 Embodied Agent

Operating Machine

Prompt:

First, read the operating menu below for coffee machine buttons.

Menu:



Now, let's look at the operating panel of the coffee machine in the image below, which button should I use for a 8 OZ coffee: left, right, top or bottom?



GPT-4V:

You should use the bottom button for an 8 OZ coffee. According to the menu, the bottom button corresponds to a 8 OZ coffee, while the top button is for a 6 OZ coffee and the right button is for a 10 OZ coffee.

Figure 88: Reading a full menu of coffee machine buttons, GPT-4V recognizes which button to choose for 8 OZ coffee. Green (Red) highlights the correct (wrong) answer. Check Section 9.7 for detailed discussions.

Sec. 9.7 Embodied Agent

Operating Machine

Prompt:

First, read the operating menus below for coffee machine buttons.

Menu 1:



Menu 2:



Menu 3:



Now, let's look at the operating panel of the coffee machine in the image below, which button should I use for a 6 OZ coffee: left, right, top or bottom button?



GPT-4V:

You should use the left button for a 6 OZ coffee.

Figure 89: Converting the full menu of coffee machine buttons to interleaved image-text instructions, GPT-4V can recognize which button to choose for 6 OZ coffee, which GPT-4V failed to do so with full menu instruction. Green highlights the correct answer. Check Section 9.7 for detailed discussions.

Sec. 9.7 Embodied Agent

Navigation

Prompt:

Imagine that you are a home robot, and is asked to go to the kitchen to fetch something from the fridge. The image below shows your current position. Please plan your next action.



GPT-4V:

My next action would be turn right and move forward towards the hallway, as I can see that the kitchen is likely located in that direction. I would then continue to navigate through the hallway until I reach the kitchen and locate the fridge.

Prompt:

Imagine that you are a home robot, and is asked to go to the kitchen to fetch something from the fridge. At the previous turn, you were at the position shown in the below image:



Note: Same image from the previous turn, resized to save space

The action taken in previous turn is: turn right and move forward towards the hallway.

Now you are at the hallway, this is what you can see from there. Can you see the kitchen from here? What is your next action?



GPT-4V:

From the current position, I can see the kitchen on the right side, My next action would be to turn right and move forward towards the kitchen.

Figure 90: Acting as an embodied agent to navigate through a house to fetch something from the fridge (the 1st and 2nd turn). Blue highlights the predicted actions. Check Section 9.7 for detailed discussions.

Sec. 9.7 Embodied Agent

Navigation

Prompt:

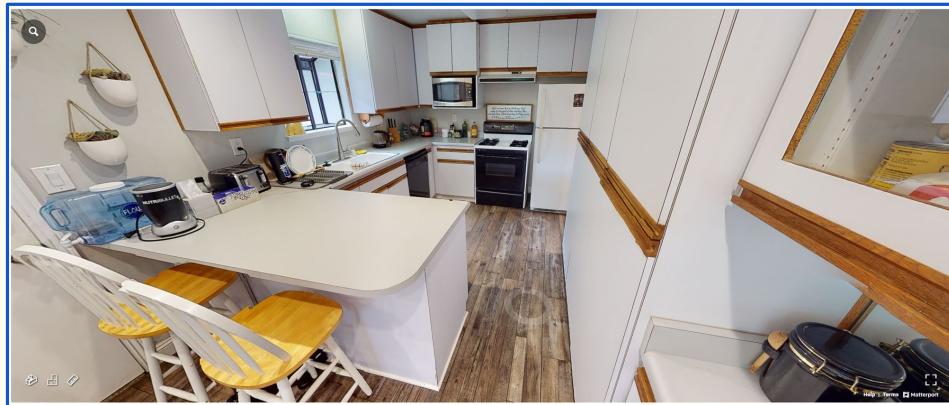
Imagine that you are a home robot, and is asked to go to the kitchen to fetch something from the fridge. At the previous turn, you were at the position shown in the below image:



Note: Same image from the previous turn, resized to save space

The action taken in previous turn is: turn right and move forward towards the kitchen.

Now you are at the kitchen, this is what you can see from there. Can you see the fridge from here? What is your next action?



GPT-4V:

From this position, I can see the fridge on the right side of the kitchen. My next action would be to move forward and turn slightly to the right to approach the fridge.

Prompt:

Imagine that you are a home robot, and is asked to go to the kitchen to fetch something from the fridge. At the previous turn, you were at the position shown in the below image:



Note: Same image from the previous turn, resized to save space

The action taken in previous turn is: move forward and turn slightly to the right to approach the fridge.

Now this is what you can see from there. What is your next action?



GPT-4V:

My next action would be to move forward and slightly to the left to align myself with the fridge door. Then, I would use my robotic arm to open the fridge door and retrieve the requested item.

Figure 91: Acting as an embodied agent to navigate through a house to fetch something from the fridge (the 3rd and 4th turn). Blue highlights the predicted actions. Check Section 9.7 for detailed discussions.

9.8 GUI Navigation

Beyond navigating the physical world, this section showcases the capability of GPT-4V to interact with and navigate through the Graphical User Interface (GUI) of a computer or smartphone. We explore the potential for GPT-4V to complete complex tasks, such as web browsing, online shopping, and *etc.*

Web browsing. We assess the performance of GPT-4V on computer GUI navigation under a task-oriented setting. The model was provided with the screenshot of current computer screen, the end goal of the navigation (*e.g.*, finding a cooking recipe or reading today’s news), the list of possible actions (*e.g.*, move the mouse, click an icon with the mouse, or type some texts with the keyboard). The model is then instructed to predict the subsequent actions (refer to Figure 92 for a complete prompt example). Upon the model’s prediction, we manually execute the predicted action and capture a screenshot, which served as the input for GPT-4V for the next turn. When the predicted action is to move the mouse, GPT-4V is specifically instructed to detail the mouse’s position. Hence, the predicted actions are grounded, showing the potential of automating the whole process without human in the loop.

In Figures 92-96, GPT-4V predicts reasonable actions to operate a computer GUI, and finally accomplish the end goal of finding a recipe of Mapo Tofu and print out a copy of the recipe in Figure 95. We then provide GPT-4V a screenshot of the printed recipe and ask it to describe the printout as detailed as possible. As shown in Figure 96, GPT-4V is able to recognize the details presented in the printout, including the cooking time, the list of ingredients, the author of the recipe, the link to the original recipe and *etc.* Figures 97-102 present how GPT-4V can navigate through GUI to browse the web to “read today’s news”. Despite the minor errors in Figure 100 when it tries to return to the previous search result page to continue browsing for more news articles, GPT-4V can perform the navigation and read two news articles reasonably well.

Online shopping. Figures 103-111 illustrates how GPT-4V can navigate a smartphone GUI for online shopping. Similarly, we provide GPT-4V with the screenshot of the current phone screen, the list of possible actions (*e.g.*, move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard) and ask it to predict the subsequent actions to shop for an ergonomic keyboard with a budget between \$50 and \$100. GPT-4V predicts to open the Amazon app (Figure 103), search ergonomic keyboard (Figure 104), open the filter options (Figure 105), set the price range filter between \$50 and \$100 (Figure 106), show filtered results (Figure 107), select the top search result (Figure 108), view product details (Figure 109), add product to the shopping cart (Figure 110) and finally proceed to checkout (Figure 111).

Notification understanding. Notifications are integral to modern human-computer interactions. GPT-4V has demonstrated its capacity to interpret notification content and respond accordingly. As shown in Figure 112, the model can read and respond to a notification, such as suggesting to open the Maps app in response to a meeting proposal in Seattle. It also handles call (Figure 113) and message (Figure 114) notifications on a computer screen effectively.

Watching videos. Alongside web browsing, videos are a key source of online information. GPT-4V has shown its capability to describe video content based on a series of screenshots from popular short-form videos. Regardless of whether the video has subtitle overlay (Figure 115 and 116) or not (Figure 117, 118, 119), GPT-4V can generate insightful descriptions about the video content, demonstrating its potential in automatic transcript generation for user-generated video content.

Sec. 9.8 GUI Navigation

Web Browsing (Mapo Tofu Recipe)

Prompt:

Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

You are asked to find a detailed recipe of Mapo Tofu.

Below is what you see on the computer screen, predict your next move. If the action involves moving the mouse, describe the location as detailed as possible.



GPT-4V:

Next move: I would move the mouse to the Google Chrome icon located on the left side of the screen, approximately 1/3 of the way down from the top. I would then click on the icon to open the web browser.

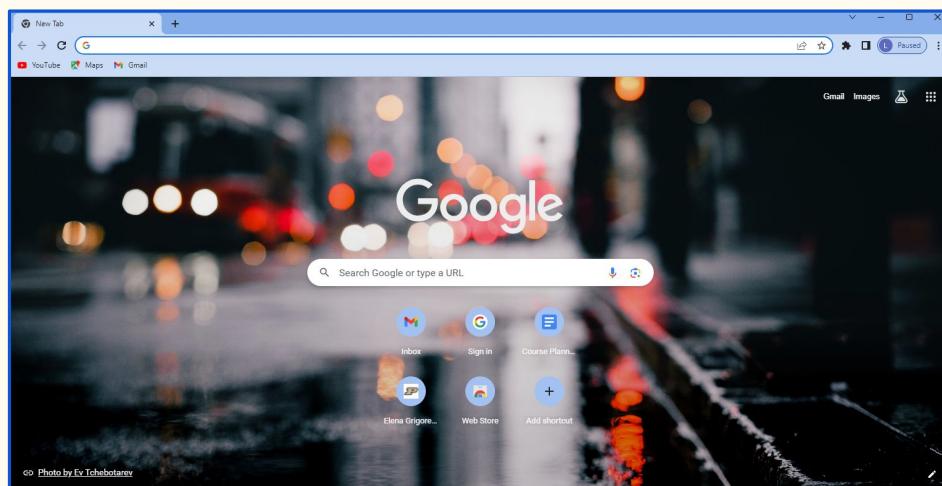
Prompt:

Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

You are asked to find a detailed recipe of Mapo Tofu.

Previous action: "I would move the mouse to the Google Chrome icon located on the left side of the screen, approximately 1/3 of the way down from the top. I would then click on the icon to open the web browser."

You have now opened the web browser. Below is what you see on the computer screen. What will you do next? If the action involves moving the mouse, describe the location as detailed as possible.



GPT-4V:

Next, I would move the mouse to the search bar located in the center of the screen, approximately 1/2 of the way down from the top. I would then click on the search bar and type in "Mapo Tofu recipe" using the keyboard. After typing in the search query, I would press the "Enter" key on the keyboard to initiate the search.

Figure 92: GPT-4V navigates through GUI to browse the web to search for the recipe of Mapo Tofu. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Web Browsing (Mapo Tofu Recipe)

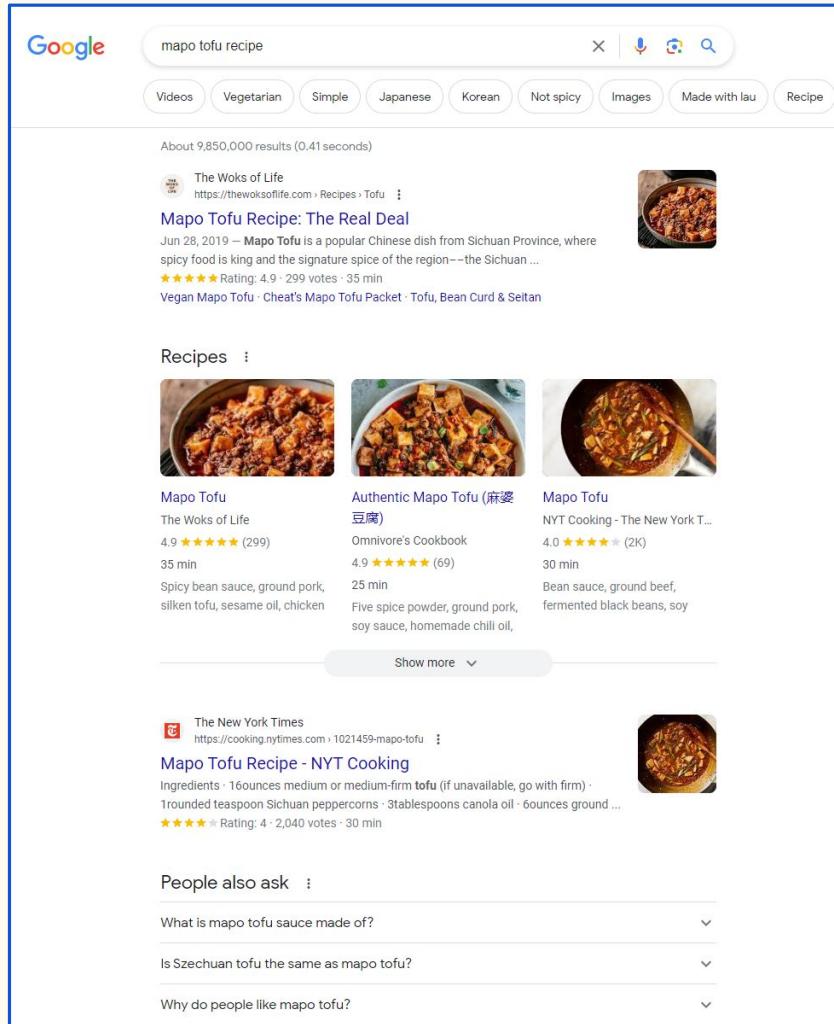
Prompt:

Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

You are asked to find a detailed recipe of Mapo Tofu.

Previous action: " I would move the mouse to the search bar located in the center of the screen, approximately 1/2 of the way down from the top. I would then click on the search bar and type in "Mapo Tofu recipe" using the keyboard. After typing in the search query, I would press the "Enter" key on the keyboard to initiate the search."

You have now completed the previous action as described above. Below is what you see on the computer screen. What will you do next? If the action involves moving the mouse, describe the location as detailed as possible.



GPT-4V:

Next action: "I would move the mouse to the first search result under the "Recipes" section, which is titled "Mapo Tofu Recipe - The Woks of Life". I would then click on the link to open the webpage and view the detailed recipe."

Figure 93: GPT-4V navigates through GUI to browse the web to search for the recipe of Mapo Tofu. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Web Browsing (Mapo Tofu Recipe)

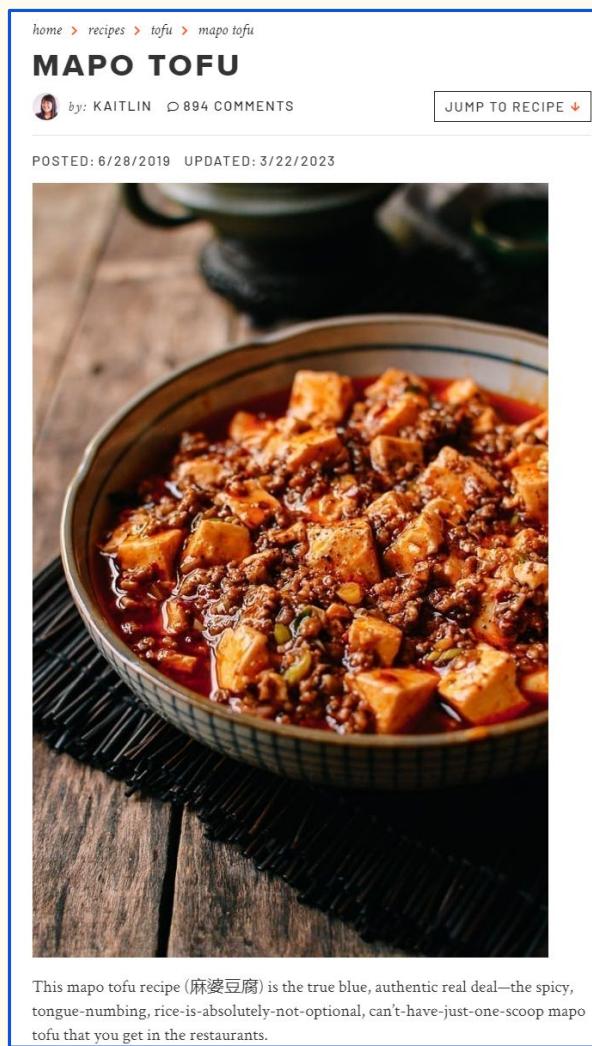
Prompt:

Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

You are asked to find a detailed recipe of Mapo Tofu.

Previous action: "I would move the mouse to the first search result under the "Recipes" section, which is titled "Mapo Tofu Recipe - The Woks of Life". I would then click on the link to open the webpage and view the detailed recipe."

You have now completed the previous action as described above. Below is what you see on the computer screen. Based on the image below, what will you do next? If the action involves moving the mouse, describe the location as detailed as possible.



GPT-4V:

Next action: "I would move the mouse to the "Jump to Recipe" button located at the top right corner of the webpage. I would then click on the button to jump to the detailed recipe section of the page."

Figure 94: GPT-4V navigates through GUI to browse the web to search for the recipe of Mapo Tofu. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Web Browsing (Mapo Tofu Recipe)

Prompt:

Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

You are asked to find a detailed recipe of Mapo Tofu.

Previous action: "I would move the mouse to the "Jump to Recipe" button located at the top right corner of the webpage. I would then click on the button to jump to the detailed recipe section of the page."

You have now completed the previous action as described above. Below is what you see on the computer screen. Based on the image below, what will you do next? If the action involves moving the mouse, describe the location as detailed as possible.

★★★★★ 4.89 from 299 votes

MAPO TOFU

This mapo tofu recipe is the true blue, authentic real deal—the spicy, tongue-numbing, rice-is-absolutely-not-optional mapo tofu that you get in the restaurants!

by: kaitlin course: tofu cuisine: chinese

SERVES: 6 servings tag or hover* to scale

PREP: 10 minutes COOK: 25 minutes TOTAL: 35 minutes

INGREDIENTS

- ½ cup oil (divided)
- 1-2 fresh Thai bird chili peppers (thinly sliced)
- 6-8 dried red chilies (roughly chopped)
- 1/2- 1 1/2 tablespoons Sichuan peppercorns (powdered or finely ground, reserving 1/4 teaspoon for garnish at the end; if you want a milder flavor use 1/2 or 1 teaspoon ground Sichuan peppercorn)
- 3 tablespoons ginger (finely minced)
- 3 tablespoons garlic (finely minced)
- 8 ounces ground pork
- 1-2 tablespoons spicy bean sauce (depending on your desired salt/spice levels)
- 2/3 cup low sodium chicken broth (or water)
- 1 pound silken tofu (cut into 1 inch/2.5cm cubes)
- 1/4 cup water
- 1 1/2 teaspoons cornstarch
- 1/4 teaspoon sesame oil (optional)
- 1/4 teaspoon sugar (optional)
- 1 scallion (finely chopped)

INSTRUCTIONS

- First, we toast the chilies. If you have homemade toasted chili oil, you can skip this step. Heat your wok or a small saucepan over low heat. Add half of the oil and throw in the fresh and dried peppers. Stir occasionally and

GPT-4V:

Next action: "I would move the mouse to the "Print Recipe" button located at the top right corner of the recipe section, next to the "Rate This Recipe" button. I would then click on the button to print out a copy of the detailed recipe for Mapo Tofu."

Figure 95: GPT-4V navigates through GUI to browse the web to search for the recipe of Mapo Tofu. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Web Browsing (Mapo Tofu Recipe)

Prompt:

Describe the image below as detailed as possible.

Mapo Tofu

This mapo tofu recipe is the true blue, authentic real deal—the spicy, tongue-numbing, rice-is-absolutely-not-optional mapo tofu that you get in the restaurants!

| | | |
|----------------------|----------------------|-----------------------|
| Prep Time 10 mins | Cook Time 25 mins | Total Time 35 mins |
|----------------------|----------------------|-----------------------|

Course: Tofu | Cuisine: Chinese | Servings: 6 servings
Calories: 335kcal | Author: Kaitlin

Ingredients

- ½ cup oil (divided)
- 1-2 fresh Thai bird chili peppers (thinly sliced)
- 6-8 dried red chilies (roughly chopped)
- 1/2- 1 1/2 tablespoons Sichuan peppercorns (powdered or finely ground, reserving 1/4 teaspoon for garnish at the end; if you want a milder flavor use 1/2 or 1 teaspoon ground Sichuan peppercom)
- 3 tablespoons ginger (finely minced)
- 3 tablespoons garlic (finely minced)
- 8 ounces ground pork
- 1-2 tablespoons spicy bean sauce (depending on your desired salt/spice levels)
- 2/3 cup low sodium chicken broth (or water)
- 1 pound silken tofu (cut into 1 inch/2.5cm cubes)
- 1/4 cup water
- 1 1/2 teaspoons cornstarch
- 1/4 teaspoon sesame oil (optional)
- 1/4 teaspoon sugar (optional)
- 1 scallion (finely chopped)

Instructions

1. First, we toast the chilies. If you have homemade toasted chili oil, you can skip this step. Heat your wok or a small saucépan over low heat. Add half of the oil and throw in the fresh and dried peppers. Stir occasionally and heat until fragrant, about 5 minutes, ensuring that the peppers don't burn. Remove from heat and set aside.
2. Heat the remaining half of the oil in your wok over medium heat. Add the ginger. After 1 minute, add the garlic. Fry for another minute, and then turn up the heat to high and add the ground pork. Break up the meat and fry it until it's cooked through. Add your ground Sichuan peppercorns and stir for about 15-30 seconds, taking care to not let it burn, as it will turn bitter if it does.
3. Add the spicy bean sauce to the mixture and stir it in well. Add the chicken broth to the wok and stir. Let this simmer for a minute or so. While that's happening, ready your tofu and combine the water and cornstarch in a small bowl.
4. Add the cornstarch mixture to your sauce and stir. Let it bubble away until the sauce starts to thicken. (If it gets too thick, splash in a little more water or chicken stock.)
5. Then add your chili oil from before—peppers and all! If you are using homemade chili oil, ONLY use the standing oil, as it's likely that you have salted it and you only want the oil, not additional salt. Stir the oil into the sauce, and add the tofu. Use your spatula to gently toss the tofu in the sauce. Let everything cook for 3-5 minutes. Add the sesame oil and sugar (if using) along with the scallions and stir until the scallions are just wilted.
6. Serve with a last sprinkle of Sichuan peppercorn powder as a garnish if desired.

Nutrition

Calories: 335kcal | Carbohydrates: 8g | Protein: 11g | Fat: 29g | Saturated Fat: 5g | Cholesterol: 27mg | Sodium: 126mg | Potassium: 340mg | Fiber: 1g | Sugar: 2g | Vitamin A: 165IU | Vitamin C: 3.5mg | Calcium: 47mg | Iron: 1.4mg

Mapo Tofu by The Woks of Life. Recipe URL: <https://thewoksoflife.com/mapo-tofu-real-deal/>

GPT-4V:

The image is a recipe for Mapo Tofu, a Chinese dish. The recipe is presented in a structured format with a title, description, prep time, cook time, total time, course, cuisine, servings, and rating.

The title of the recipe is "Mapo Tofu" and it is written in bold black letters. Below the title, there is a description of the dish which states that it is the true blue, authentic real deal - the spicy, tongue-numbing, rice-is-absolutely-not-optional Mapo Tofu that you get in the restaurants.

The prep time for the dish is 10 minutes, the cook time is 25 minutes, and the total time is 35 minutes. The course is listed as "Main Course" and the cuisine is "Chinese". The recipe serves 6 people and has a rating of **4.69 stars out of 5**.

Below this information, there is a list of ingredients required for the dish, including tofu, ground pork, Sichuan peppercorn, garlic, ginger, soy sauce, and cornstarch. The ingredients are listed in a bullet point format.

At the bottom of the image, there is nutritional information for the dish, including calories, carbohydrates, protein, fat, saturated fat, cholesterol, sodium, and calcium. The nutritional information is presented in a table format.

The image also includes a credit to the author of the recipe, Kaitlin, and a link to the original recipe on the website "The Woks of Life".

Figure 96: GPT-4V navigates through GUI to browse the web to search for the recipe of Mapo Tofu. As GPT-4V predicts to print out the recipe in the previous turn, we prompt it to read the screenshot of the printed recipe and summarize it. Red highlights the inaccurate description about the image. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Web Browsing (Today's News)

Prompt:

Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

You are asked to find what's in the news today.

Below is what you see on the computer screen, predict your next move. If the action involves moving the mouse, describe the location as detailed as possible.



GPT-4V:

My next move would be to move the mouse to the Google Chrome icon located on the left side of the screen, approximately 1/3 of the way down from the top. I would then click on the icon to open the web browser and search for news websites to find out what's in the news today.

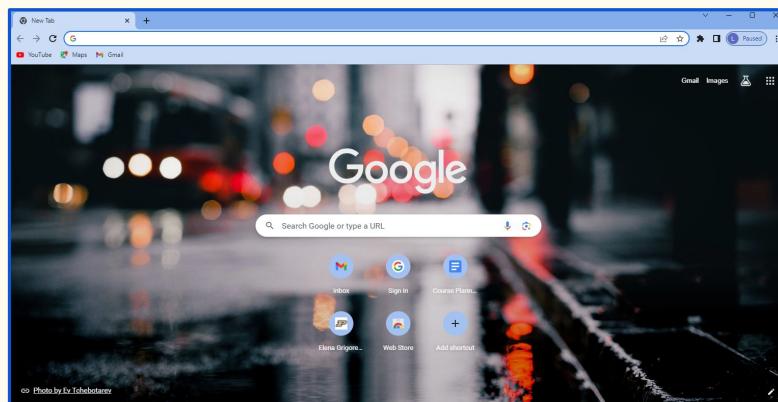
Prompt:

Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

You are asked to find what's in the news today.

Previous action: "My next move would be to move the mouse to the Google Chrome icon located on the left side of the screen, approximately 1/3 of the way down from the top. I would then click on the icon to open the web browser and search for news websites to find out what's in the news today."

You have now opened the web browser. Below is what you see on the computer screen, what is your immediate next step? If the action involves moving the mouse, describe the location as detailed as possible.



GPT-4V:

My immediate next step would be to move the mouse to the search bar located in the center of the screen, approximately 1/2 of the way down from the top. I would then click on the search bar and type in "news" and press the enter key on the keyboard to search for news websites.

Figure 97: GPT-4V navigates through GUI to browse the web to read today's news. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Web Browsing (Today's News)

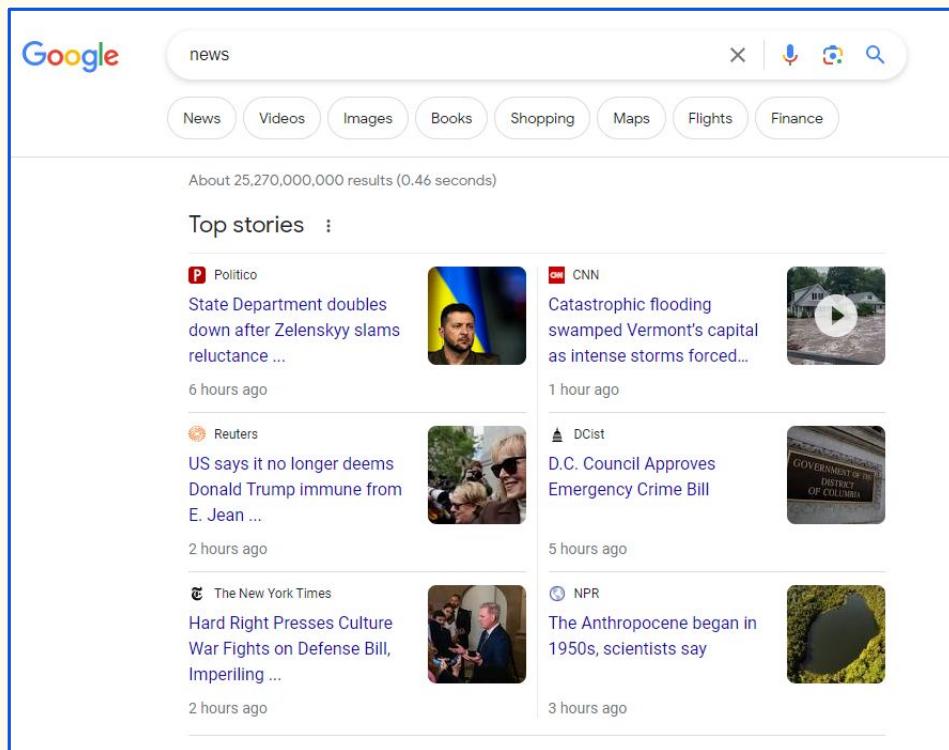
Prompt:

Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

You are asked to find what's in the news today.

Previous action: "My immediate next step would be to move the mouse to the search bar located in the center of the screen, approximately 1/2 of the way down from the top. I would then click on the search bar and type in "news" and press the enter key on the keyboard to search for news websites."

You have now completed the previous action. Below is what you see on the computer screen. Based on the image below, what is your immediate next step? If the action involves moving the mouse, describe the location as detailed as possible.



GPT-4V:

My immediate next step would be to move the mouse to the first search result, which is "State Department doubles down on Zelensky slams" located approximately 1/4 of the way down from the top of the screen and 1/8 of the way from the left side of the screen. I would then click on the link to open the news article and read about what's in the news today.

Figure 98: GPT-4V navigates through GUI to browse the web to read today's news. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Web Browsing (Today's News)

Prompt:

Read the article shown in the images below and summarize it.

POLITICO

FOREIGN AFFAIRS

State Department doubles down after Zelenskyy slams reluctance on Ukraine's entry to NATO

State Department press secretary Matthew Miller said the United States' hesitancy to see Ukraine join NATO was because of the ongoing war with Russia.



In a statement posted to Twitter on Tuesday, Ukrainian President Volodymyr Zelenskyy took issue with NATO's reluctance to commit to admitting his country | Francisco Seco/AP Photo

By LUCY GOODMAN
07/11/2023 04:18 PM EDT

The State Department on Tuesday defended NATO's hesitation to invite Ukraine into the alliance, even as Ukrainian President Volodymyr Zelenskyy voiced his frustration with how the negotiations have unfolded.

State Department press secretary Mathew Miller spoke to reporters on Tuesday from the Lithuanian capital, Vilnius, where NATO's two-day summit is being held this week. Miller cited several conditions that NATO had decided Ukraine must meet for an invitation to the alliance, including democratic reforms and the end of the war between Russia and Ukraine.

We are ready to extend an invitation when conditions are met," Miller said. "There is an important reason why they are not becoming a member of NATO right now — because it would instantly put the United States in a shooting war with Russia."

Miller echoed the sentiment of President Joe Biden, who told CNN on Sunday that Ukraine could not be admitted into the alliance until the war with Russia had ended, adding "if the war is going on, then we're all in war."

The U.S. has doubled down this week on military support of Ukraine in its war with Russia, agreeing to include controversial cluster munitions in its next arms package to Kyiv — a decision that Ukrainian Minister of Defense Oleksiy Reznikov said would be a "game changer" as the Ukrainian counteroffensive struggles to breach Russia's front lines.

In a summit communiqué released Tuesday, NATO leaders wrote that the alliance could not promise Ukraine membership until the war had ended and several democratic and security-sector reforms had been achieved.

"We will be in a position to extend an invitation to Ukraine to join the Alliance when Allies agree and conditions are met," the leaders wrote.

In a statement posted to Twitter on Tuesday, Zelenskyy took issue with NATO's reluctance to commit to admitting his country, and its unwillingness to provide an anticipated timeline for Ukraine's entry. He wrote that he did not see sufficient readiness on the part of NATO to integrate Ukraine into the alliance.

"It's unprecedented and absurd when time frame is not set neither for the invitation nor for Ukraine's membership," Zelenskyy wrote. "While at the same time vague wording about 'conditions' is added even for inviting Ukraine."

GPT-4V:

Summary:

The article discusses the statement made by Ukrainian President Volodymyr Zelensky on Tuesday, where he expressed his country's desire to join NATO. Zelensky stated that Ukraine cannot afford to go to war with Russia and that joining NATO would provide the necessary security and protection.

However, Mathew Miller, a spokesperson for NATO, stated that while the alliance is ready to extend an invitation to Ukraine, there are certain conditions that need to be met. These conditions include democratic and security-sector reforms. Miller also emphasized that Ukraine's entry into NATO is not a guarantee and that the decision ultimately lies with the members of the alliance.

The article also mentions that during a press conference on Tuesday, Zelensky reiterated his commitment to strengthening democracy in Ukraine, which is a necessary condition for NATO membership.

Figure 99: GPT-4V navigates through GUI to browse the web to read today's news. We prompt GPT-4V to read the screenshots of the first news article and summarize it. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Web Browsing (Today's News)

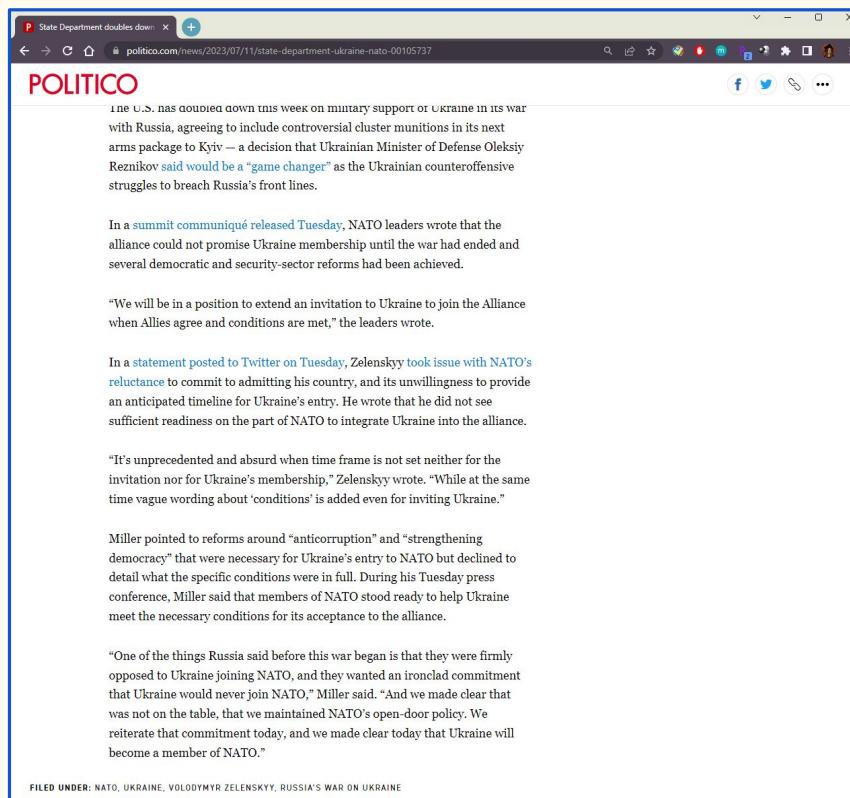
Prompt:

Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

You are asked to find what's in the news today.

Previous action: "My immediate next step would be to move the mouse to the first search result, which is "State Department doubles down on Zelensky slams" located approximately 1/4 of the way down from the top of the screen and 1/8 of the way from the left side of the screen. I would then click on the link to open the news article and read about what's in the news today."

You have now completed reading "State Department doubles down on Zelensky slams". What is your immediate next step? If the action involves moving the mouse, describe the location as detailed as possible.



GPT-4V:

My immediate next step would be to move the mouse to the top right corner of the screen, where I can see the "X" icon to close the current tab. I would then click on the "X" icon to close the tab and return to the previous page to continue browsing for more news articles.

Figure 100: GPT-4V navigates through GUI to browse the web to read today's news. Upon finishing reading the first news article, GPT-4V predicts to close the tab and return to previous page to continue browsing more news articles (highlighted in blue). Red highlights the inaccurate action prediction. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Browsing the Web (Today's News)

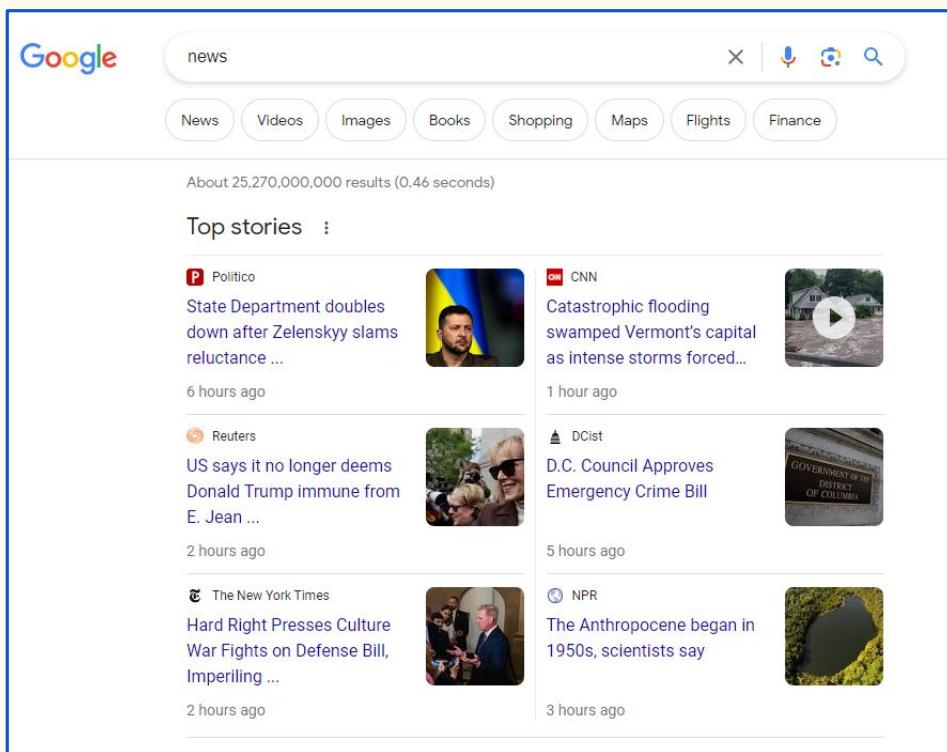
Prompt:

Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

You are asked to find what's in the news today.

Previous action: "My immediate next step would be to move the mouse to the top right corner of the screen, where I can see the "X" icon to close the current tab. I would then click on the "X" icon to close the tab and return to the previous page to continue browsing for more news articles."

You have now returned to the previous page after reading the first article "State Department doubles down on Zelensky slams". Below is what you see on the computer screen. Based on the image below, what is your immediate next step? If the action involves moving the mouse, describe the location as detailed as possible.



GPT-4V:

My immediate next step would be to move the mouse to the second article on the page, "Catastrophic flooding swamps Florida's panhandle as intense storms force evacuations". I would move the mouse to the left side of the article title, approximately 1/3 of the way down the page, and click on the article title to open it and read the news.

Figure 101: GPT-4V navigates through GUI to browse the web to read today's news. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Web Browsing (Today's News)

Prompt:

Read the article shown in the images below and summarize it.

Catastrophic flooding swamped Vermont's capital as intense storms forced evacuations and closures in Northeast

By Moran Salathé and Lauren Manganaro, CNN
Updated 9:18 PM EDT, Tue July 11, 2023



Watch how deadly flood flooding hit New York

02:09 Source: CNN

(CNN) — Intense rainstorms that inundated the Northeast turned streets into rivers, forced evacuations and prompted officials in Vermont's capital, Montpelier, to close the downtown area.

"Make no mistake, the devastation and flooding we're experiencing across Vermont is historic and catastrophic," Gov. Phil Scott told reporters Tuesday.

Floodwaters in some areas "surpassed levels seen during Tropical Storm Irene," the governor said. Irene hit the United States as a hurricane in August 2011 and left entire communities submerged, killing more than 40 people in several Eastern states.

The current flooding claimed thousands of homes and businesses, Scott added.

In downtown Montpelier, authorities issued a boil water ban that lasted through Tuesday afternoon. A boil water notice remains in place for the city, after officials warned the "extreme flooding" could contaminate the drinking water supply.

The city and surrounding communities saw "record-breaking flooding" in two days, Montpelier officials said Tuesday afternoon. The city's downtown area remains flooded and is not safe for public travel, they added.

The city was hammered by a record-setting 5.28 inches of rainfall Monday, the National Weather Service in Burlington said. That's more than any other day on record, including when Irene dropped 5.27 inches of rain on the state capital on August 28, 2011.

RELATED ARTICLE



"There had about a 12-hour period of rain, and then it was over," the governor said. "This is different. We've had like 48 hours of steady rain," he said, adding more rain is expected in the coming days.

Vermont remains under a state of emergency, with water in downtown Montpelier running from knee to waist deep Tuesday, stranding residents in their homes and shutting down roads and businesses. The flooding forced evacuations and more than 100 rescues in the state, Urban Search and Rescue manager Mike Cannon told CNN.

"In many areas, the water conditions remain too dangerous for rescue by boat," state public safety commissioner Jennifer Morrison said Tuesday.

WPS DAILY

Want to stay updated on the latest space and science news? We've got you.

Email address Sign Me Up

By submitting you agree to our [privacy policy](#).

"There are life-threatening isolations that we are trying to identify and rescue," Morrison added.

President Joe Biden spoke with Scott and other officials about the flooding, the White House said Tuesday, adding it is coordinating with state and local officials and monitoring the flood impacts.

New Hampshire will send swift boat rescue crews and Black Hawk helicopters to assist with the response to the flooding in Vermont, Gov. Chris Sununu said. Teams from Connecticut, Massachusetts and North Carolina are currently in Vermont to provide assistance and others are en route, the Vermont Department of Public Safety said.

In Montpelier, the police department relocated its dispatch, police and fire operations due to heavy flooding in the basement of City Hall and the fire and police departments, Police Chief Eric Nordenson said Tuesday.

Authorities were also monitoring the Wrightsville Dam in Montpelier late Tuesday, after water levels earlier in the day inched toward the dam's spillway.

In an 8:30 p.m. update, Montpelier officials said the dam's water level was beginning to recede and was not expected to breach the spillway. On Wednesday, crews will begin clearing mud and debris from the city's roadways, building inspections will start in the downtown area and clean-up efforts will begin, officials said.

Authorities were also monitoring the Wrightsville Dam in Montpelier late Tuesday, after water levels earlier in the day inched toward the dam's spillway.

In an 8:30 p.m. update, Montpelier officials said the dam's water level was beginning to recede and was not expected to breach the spillway. On Wednesday, crews will begin clearing mud and debris from the city's roadways, building inspections will start in the downtown area and clean-up efforts will begin, officials said.

Higher than levels reached during Hurricane Irene. The Winooski River at Montpelier rose nearly 14 feet Monday and passed major flood stage as the water continued to climb, threatening further flooding.

The Bell Mountain Dam and the Townshend Dam, in southwestern Vermont, were expected to overflow their spillways," the US Army Corps of Engineers warned Monday night. Warning of "severe flooding," the agency urged residents in threatened low-lying areas of nearby Vermont and New Hampshire communities to evacuate.

Rainfall is at 300% to 500% of normal



02:06 Source: CNN

The aftermath of flooding in Main Street in Highland Falls, New York, is seen Monday.

Seven-day rainfall totals across much of the Northeast were already at 300% to 500% of normal levels, the Weather Prediction Center said Monday.

Widespread rainfall of 2 to 4 inches fell across the Northeast from eastern Pennsylvania and northern New Jersey into Vermont and New Hampshire. Isolated rainfall totals higher than 6 inches have been seen in several states.

In New York, a 30-year-old woman died after being swept away by floodwater as she tried to evacuate her Orange County home Sunday. The flooding has caused "easily tens of millions of dollars in damages," county Executive Steve Neuhuis said Monday.

Some areas in New York were hammered with more than 8 inches of rain within a 24-hour period.

In the northern part of the state, a state of emergency was declared for the town of Long Lake due to severe flooding that destroyed roads and bridges, downed electricity lines and forced residents to evacuate, Long Lake town supervisor Clay Arsenault said in a declaration. A state of emergency is in effect for Hamilton County through Wednesday morning.

Long Lake officials also said non-essential travel in the town is not allowed at this time.

Water rushing into homes



02:06 Source: CNN

A man carries belongings through floodwaters from a home in Bridgewater, Vermont, on Monday.

In Vermont, state Rep. Kelly Pejala said she woke up Monday morning to floodwater already at the front step of her Londonderry apartment. She and her son packed up their two cats and evacuated to higher ground.

"For people that were here during Irene, it feels like a very similar experience," she said. Hurricane Irene brought destructive flooding to the state in 2011, causing major infrastructure damage.

Floodwater could be seen gushing between homes in Chester, where some structures were visibly damaged and trucks were wheels-deep.

Don Hancock, dripping in water from head to toe, told CNN he has only lived in his house for less than a year and watched floodwater enter the basement and garage of his new home.

"I was a firefighter in New York, I've been there many times to help people out, but I've never lived this side of it," Hancock said. "Once the water goes down, we go day by day, clean it up and move on. What can we do?"

CNN Meteorologists Taylor Ward and CNN's Jessica Xie, Roseanne Garcia, Artemis Mostragian, Raja Razik, Sara Smart, Nikki Cervajal and Amanda Jackson contributed to this report.

GPT-4V:

The article discusses the catastrophic flooding that occurred in Vermont's capital, Montpelier, on Tuesday, July 11, 2023. The flooding was caused by intense rainstorms that inundated the Northeast, leading to evacuations and closures in the area.

The floods were described as "unprecedented" and "historic" by officials, with some areas experiencing record-breaking rainfall. The National Weather Service issued a travel advisory, warning residents to avoid non-essential travel.

Vermont Governor Phil Scott declared a state of emergency, and New Hampshire sent rescue crews and Black Hawk helicopters to assist with the response. The Vermont Fire Department and other emergency services were also dispatched to the affected areas.

The floods caused extensive damage to homes and businesses, with some residents being forced to evacuate. A boil water notice was issued for Montpelier, and officials began clearing debris from roadways on Wednesday.

In Oklahoma City, two children were rescued from a reservoir after being swept away by floodwaters. The children, aged 10 and 11, were among a group of four that were fishing in the area.

In New York, Hurricane Irene caused 8 inches of rain within a 24-hour period, leading to major infrastructure damage. A state of emergency was declared for Long Lake Town Supervisor Clay Arsenault, and residents were urged to evacuate.

The article concludes with a quote from Dan Hancock, a resident of Chester, Vermont, who watched floodwater enter his basement. He expresses his frustration and helplessness, stating "Once the water goes down, we go back in, clean it up and move on. What can we do?"

Figure 102: GPT-4V navigates through GUI to browse the web to read today's news. We prompt GPT-4V to read the screenshots of the second news article and summarize it. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

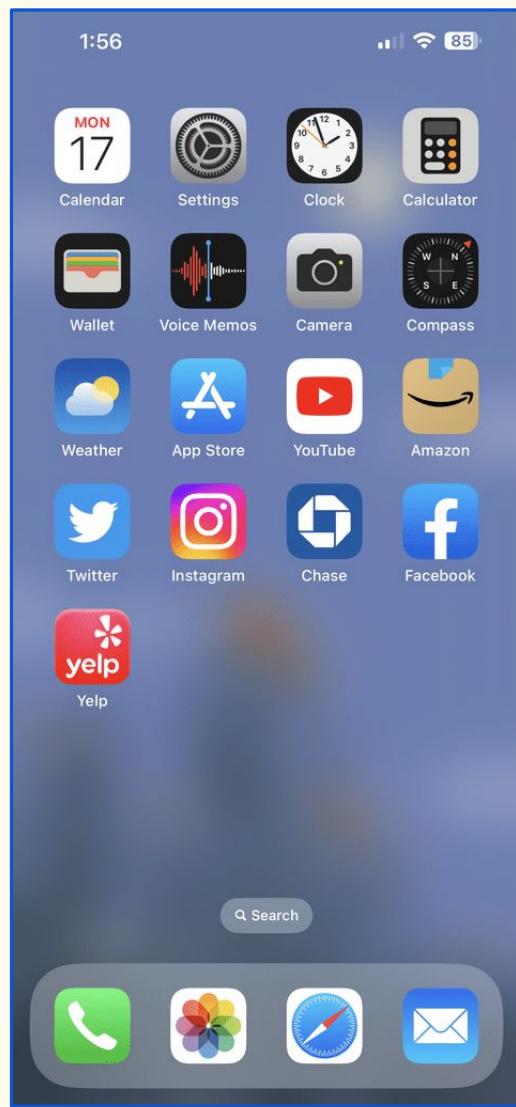
Online Shopping (Ergonomic Keyboard)

Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Below is what you see on the phone screen, predict your next move. If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My next move would be to move my finger to the Amazon icon, which is located in the third row and third column of the screen. I would then click on the icon to open the app and search for an ergonomic keyboard within my budget between \$50 and \$100.

Figure 103: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Red highlights the inaccurate location of the Amazon icon. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: "My next move would be to move my finger to the Amazon icon, which is located in the third row and third column of the screen. I would then click on the icon to open the app and search for an ergonomic keyboard within my budget of \$50."

You have now opened the Amazon app. Below is what you see on the phone screen. What will you do next? If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My next move would be to move my finger to the search bar, which is located at the top of the screen. I would then click on the search bar and type in "ergonomic keyboard" using the keyboard. After typing in the search term, I would click on the "Search" button, which is located on the bottom right corner of the keyboard, to initiate the search.

Figure 104: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

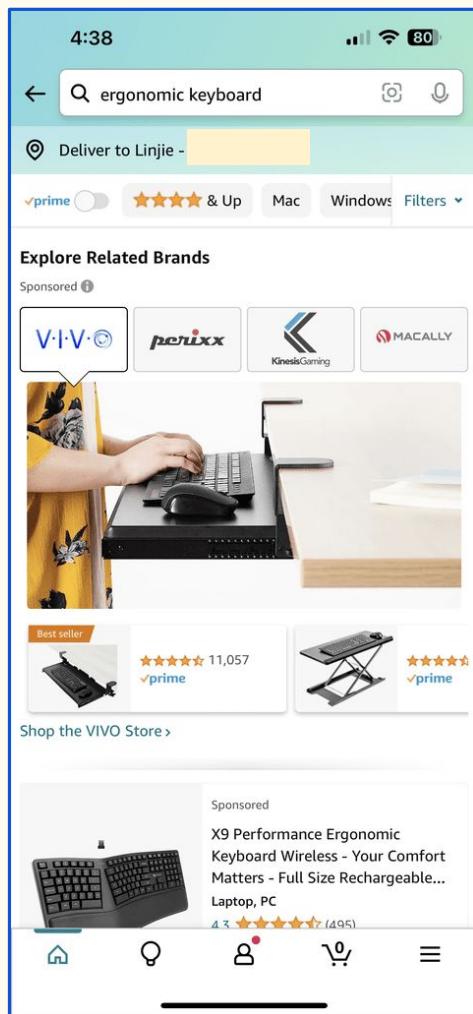
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: "My next move would be to move my finger to the search bar, which is located at the top of the screen. I would then click on the search bar and type in "ergonomic keyboard" using the keyboard. After typing in the search term, I would click on the "Search" button, which is located on the bottom right corner of the keyboard, to initiate the search."

You have now completed the previous action as described above. Below is what you see on the phone screen. Please predict your immediate next move. If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My immediate next move would be to move my finger to the "Filters" button, which is located at the top right corner of the screen, next to the "Windows" button. I would then click on the "Filters" button to open the filter options and narrow down my search results to only show ergonomic keyboards within my budget range of \$50 to \$100.

Figure 105: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

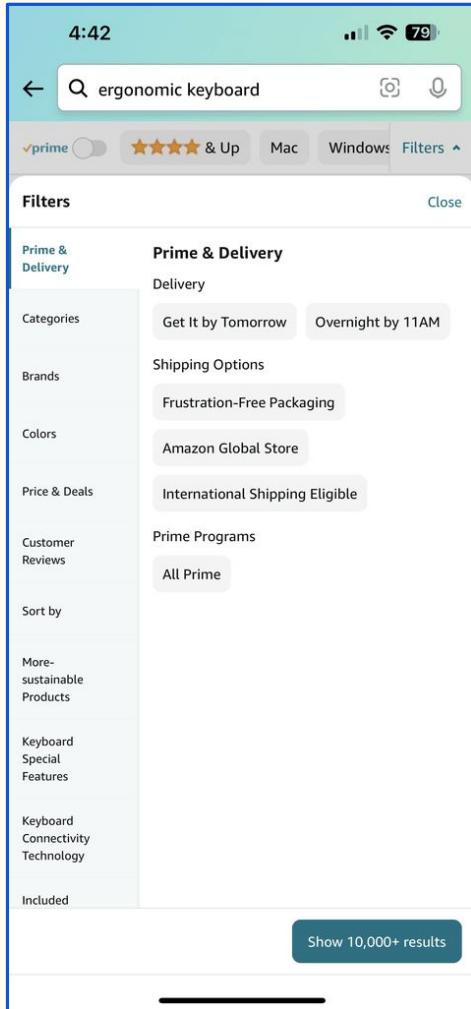
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: "My immediate next move would be to move my finger to the "Filters" button, which is located at the top right corner of the screen, next to the "Windows" button. I would then click on the "Filters" button to open the filter options and narrow down my search results to only show ergonomic keyboards within my budget range of \$50 to \$100."

You have now completed the previous action as described above. Below is what you see on the phone screen. Please predict your immediate next move. If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My immediate next move would be to move my finger to the "Price & Deals" filter option, which is located under the "Colors" filter option and above the "Sort by" filter option. I would then click on the "Price & Deals" filter option to open the price range filter and set my minimum price to \$50 and my maximum price to \$100. This would further narrow down my search results to only show ergonomic keyboards within my budget range.

Figure 106: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

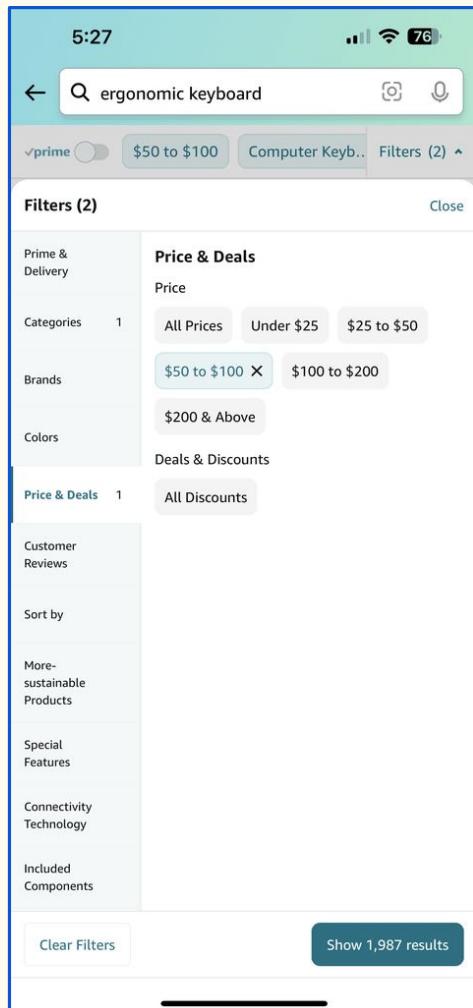
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: " My immediate next move would be to move my finger to the "Price & Deals" filter option, which is located under the "Colors" filter option and above the "Sort by" filter option. I would then click on the "Price & Deals" filter option to open the price range filter and set my minimum price to \$50 and my maximum price to \$100. This would further narrow down my search results to only show ergonomic keyboards within my budget range. "

You have now completed the previous action as described above. Below is what you see on the phone screen. What will you do next? If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My next move would be to move my finger to the "Show 1,987 results" button, which is located at the bottom of the screen, below the "Clear Filters" button and above the navigation bar. I would then click on the "Show 1,987 results" button to view the search results for ergonomic keyboards within my budget range.

Figure 107: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

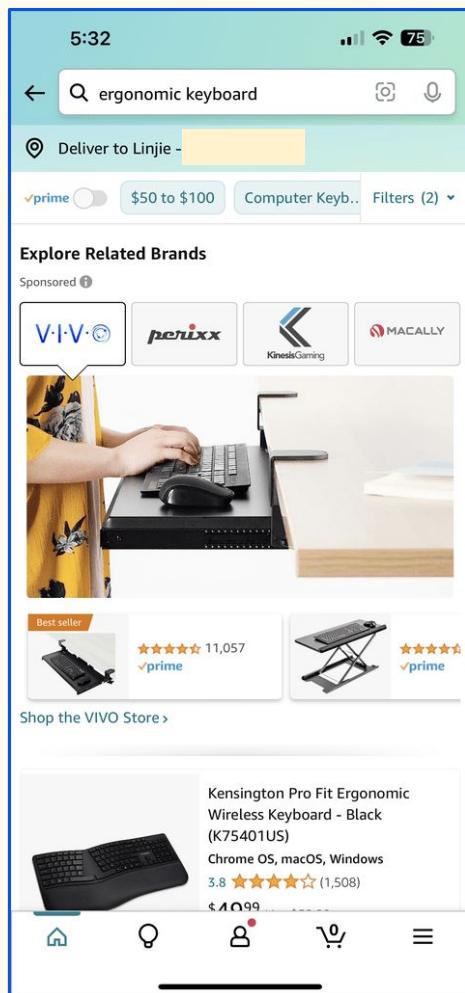
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: " My next move would be to move my finger to the "Show 1,987 results" button, which is located at the bottom of the screen, below the "Clear Filters" button and above the navigation bar. I would then click on the "Show 1,987 results" button to view the search results for ergonomic keyboards within my budget range. "

You have now completed the previous action as described above. Below is what you see on the phone screen. What will you do next? If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My next move would be to scroll down the screen to view the different options for ergonomic keyboards within my budget range. I would then move my finger to the "Kensington Pro Fit Ergonomic Wireless Keyboard - Black" option, which is located below the "Explore Related Brands" section and **above the "VIVO" option**. I would then click on the "Kensington Pro Fit Ergonomic Wireless Keyboard - Black" option to view more details about the product and potentially make a purchase.

Figure 108: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Red highlights the inaccurate location of the product option to be selected. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

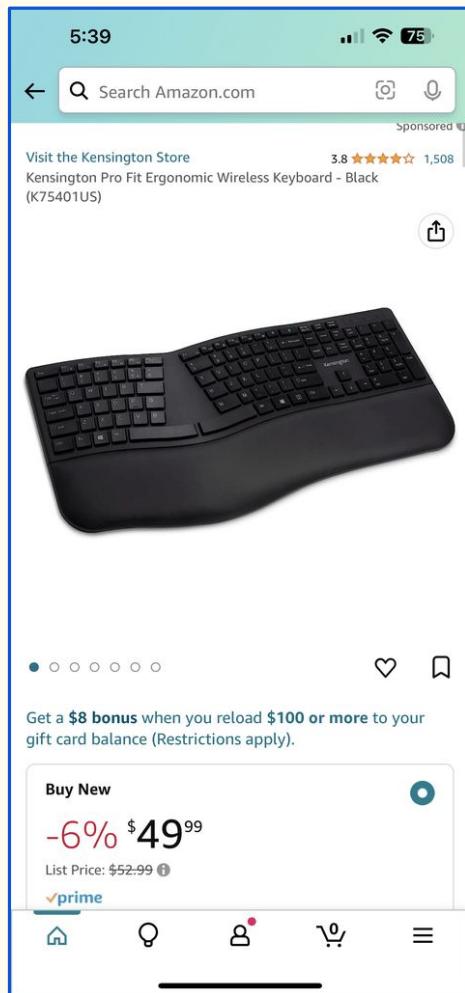
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: " My next move would be to scroll down the screen to view the different options for ergonomic keyboards within my budget range. I would then move my finger to the "Kensington Pro Fit Ergonomic Wireless Keyboard - Black" option, which is located below the "Explore Related Brands" section and above the "VIVO" option. I would then click on the "Kensington Pro Fit Ergonomic Wireless Keyboard - Black" option to view more details about the product and potentially make a purchase. "

You have now completed the previous action as described above. Below is what you see on the phone screen. What will you do next? If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My next move would be to scroll down the screen to view more details about the "Kensington Pro Fit Ergonomic Wireless Keyboard - Black" product. I would then move my finger to the "Buy New" button, which is located below the product image and above the "Get a \$8 bonus when you reload \$100 or more to your gift card balance" section. I would then click on the "Buy New" button to proceed with the purchase.

Figure 109: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Red highlights the inaccurate action prediction ("Buy New" is not a clickable button). Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

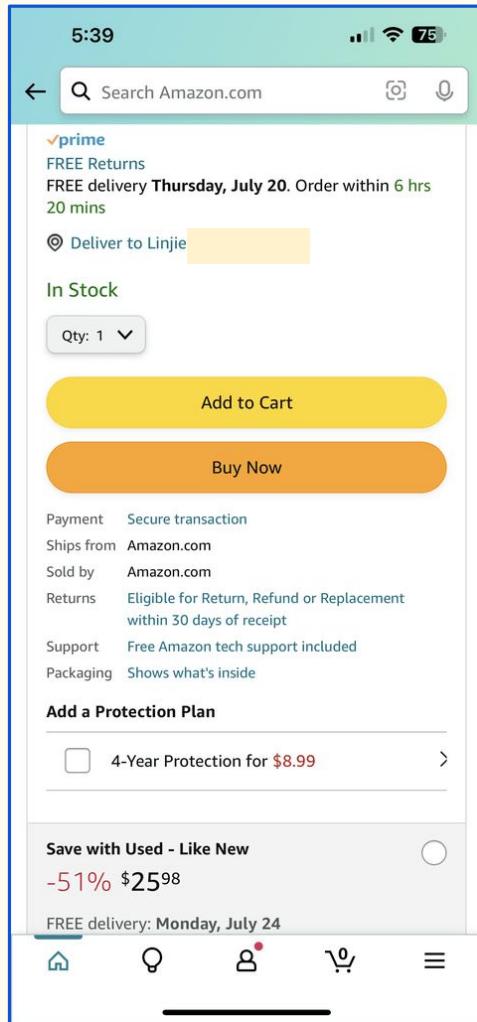
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: " My next move would be to scroll down the screen to view more details about the "Kensington Pro Fit Ergonomic Wireless Keyboard - Black" product. I would then move my finger to the "Buy New" button, which is located below the product image and above the "Get a \$8 bonus when you reload \$100 or more to your gift card balance" section. I would then click on the "Buy New" button to proceed with the purchase. "

You have now completed the previous action as described above. Below is what you see on the phone screen. What will you do next? If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My next move would be to move my finger to the "Add to Cart" button, which is located below the "Buy New" button and above the "Ships from Amazon" section. I would then click on the "Add to Cart" button to add the item to my shopping cart.

Figure 110: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

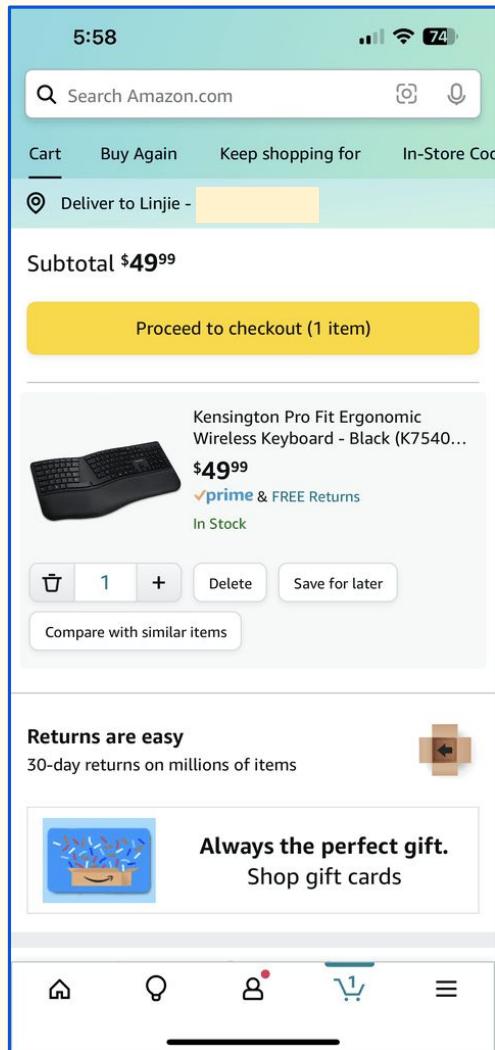
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: " My next move would be to move my finger to the "Add to Cart" button, which is located below the "Buy New" button and above the "Ships from Amazon" section. I would then click on the "Add to Cart" button to add the item to my shopping cart. "

You have now completed the previous action as described above. Below is what you see on the phone screen. What will you do next? If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My next move would be to move my finger to the "Proceed to Checkout" button, which is located below the "Compare with similar items" section and above the "30-day returns" section. I would then click on the "Proceed to Checkout" button to proceed with the purchase.

Figure 111: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Red highlights the inaccurate location of the “Proceed to checkout” button. Check Section 9.8 for detailed discussions.

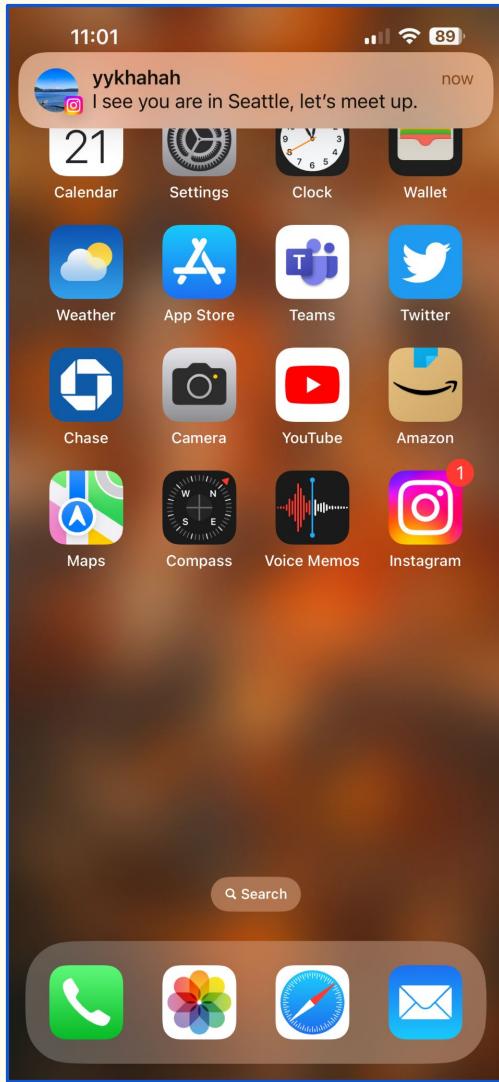
Sec. 9.8 GUI Navigation

Notification Understanding

Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

Below is what you see on the phone screen. Describe the image and predict your next move. If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

The image on the phone screen is the home screen of an iPhone. There are multiple app icons arranged in a grid-like pattern, including Calendar, Settings, Twitter, Amazon, Instagram, etc. At the top of the screen, there is a notification from a user named "yykhahah" that reads "I see you are in Seattle, let's meet up." At the bottom of the screen, there is a search bar.

Based on the notification, my next move would be to move my finger to the Maps app icon, which is located in the second row and the first column of the app grid. This will allow me to search for a location in Seattle and plan a meet up with the user.

Figure 112: Prompting GPT-4V to predict the action upon receiving a notification. GPT-4V can accurately recognize the notification and the corresponding content (highlighted in green). Blue highlights the predicted actions. Red highlights the inaccurate location of the Maps app icon. Check Section 9.8 for detailed discussions.

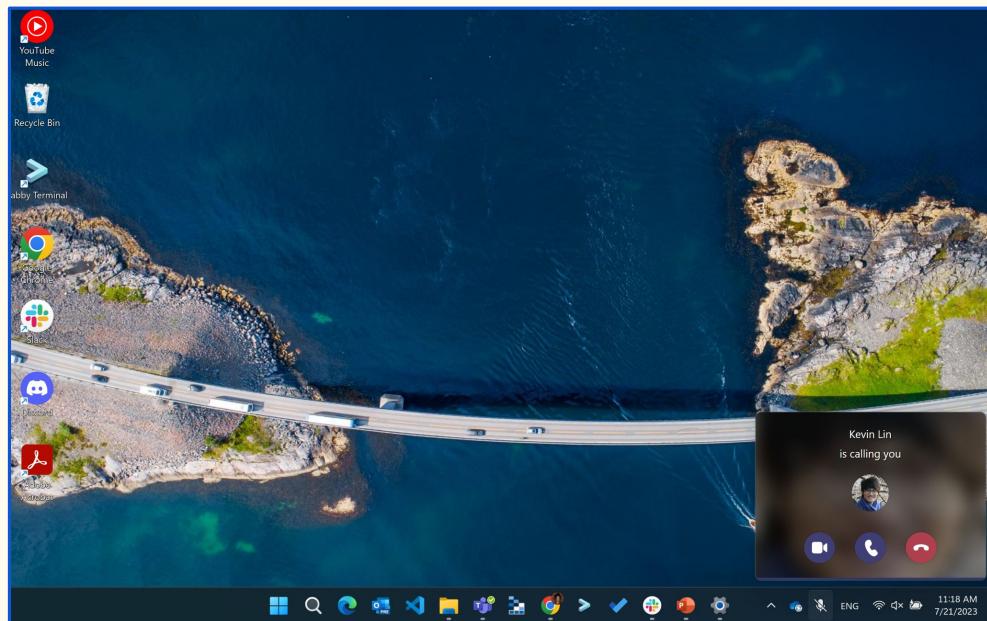
Sec. 9.8 GUI Navigation

Notification Understanding

Prompt:

Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

Below is what you see on the computer screen. Describe the image and predict your next move. If the action involves moving the mouse, describe the location and the icon as detailed as possible.



GPT-4V:

The image on the computer screen is a desktop background of a bridge over a body of water. There are several icons on the desktop, including a YouTube icon, a Recycle Bin icon, and a Terminal icon. At the bottom of the screen, there is a taskbar with several application icons, such as Microsoft Edge and File Explorer. On the bottom right corner of the screen, there is a notification that says "Kevin Lin is calling you."

Based on the notification, my next move would be to move the mouse to the bottom right corner of the screen and click on the notification. This would open the call application and allow me to answer Kevin Lin's call.

Figure 113: Prompting GPT-4V to predict the action upon receiving a notification. GPT-4V can accurately recognize the notification and the corresponding content (highlighted in green). Blue highlights the predicted actions. Red highlights the inaccurate location of the Maps app icon. Check Section 9.8 for detailed discussions.

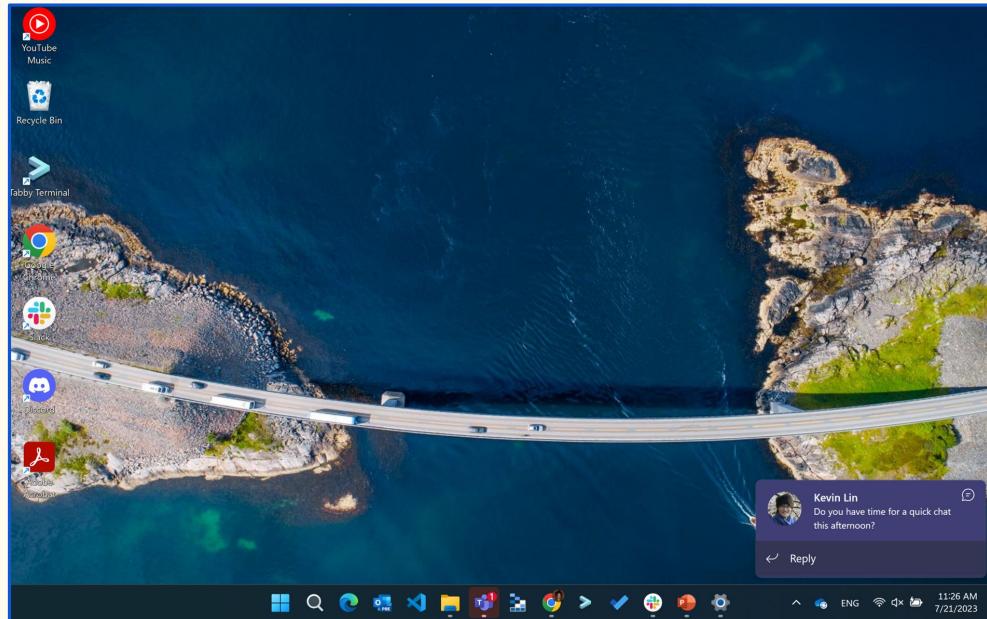
Sec. 9.8 GUI Navigation

Notification Understanding

Prompt:

Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

Below is what you see on the computer screen. Describe the image and predict your next move. If the action involves moving the mouse, describe the location as detailed as possible.



GPT-4V:

The image on the computer screen is a desktop background of a beautiful aerial view of a bridge over the ocean. There are several icons on the left side of the screen, including the Recycle Bin, Terminal, and a few others. On the bottom right corner, there is a notification from a messaging app from a user named Kevin Lin asking if I am free for a chat this afternoon.

My next move would be to move the mouse to the bottom right corner of the screen and click on the notification from Kevin Lin to open the messaging app and respond to his message. I would move the mouse to the exact location of the notification, which is approximately 2 inches from the right edge of the screen and 1 inch from the bottom edge of the screen.

Figure 114: Prompting GPT-4V to predict the action upon receiving a notification. GPT-4V can accurately recognize the notification and the corresponding content (highlighted in green). Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Watching Videos

Prompt:

Explain the story shown in the images below.



GPT-4V:

The images show a reel on Instagram, where a user is sharing a humorous post about how much they have spent on their dog. The first image shows a man holding his dog with the caption "When someone asks me how much I have spent on my dog". The next three images show a clip of Joe Biden saying "700 billion" followed by "700 billion and a trillion 300 million" and "700 billion and a trillion 300 million dollars". The user is jokingly exaggerating the amount they have spent on their dog by using the clip of Joe Biden saying a large sum of money.

Prompt:

Tell me about the video content, given the frames below.



GPT-4V:

The video content is a compilation of things that Asian people do for no reason. The video is posted on **TikTok** by user laparasan and has 1.3M views and 13.1K likes. The video is set to original audio by sophrelabelous and has the hashtag **#asian** and **#relatable**.

The video is broken down into 5 different scenes:

1. The first scene shows a girl grabbing products from the back of the shelf in a supermarket.
2. The second scene shows a girl rinsing a perfectly clean cup before using it.
3. The third scene shows two girls asking "have you eaten yet?" as a form of greeting.
4. The fourth scene shows two girls fighting aggressively for the bill in a restaurant.
5. The fifth scene shows a girl wrapping pants around her neck to see if it fits.

Overall, the video is a humorous take on some of the quirks and habits that are commonly seen in the Asian community.

Figure 115: Prompting GPT-4V to watch web videos. We present GPT-4V the screenshot of the video frames following their temporal order in the original video. To save space, we illustrate the frames in a row, where the leftmost one is the first frame. Check Section 9.8 for detailed discussions.

Red highlights the inaccurate descriptions about the video.

Sec. 9.8 GUI Navigation

Watching Videos

Prompt:

Transcribe the video content, given the frames below.



GPT-4V:

Hi everyone, it's natyexplora here and today I'm going to be sharing 7 places you need to visit in Washington State.

First up, we have North Cascades National Park. This park is filled with beautiful mountain views and hiking trails. It's a must-visit for any nature lover.

Next, we have the Olympic Peninsula. This area is home to the famous Hoh Rainforest and beautiful beaches. It's a great place to relax and enjoy the natural beauty of Washington.

Third on the list is Mt Baker Wilderness. This area is known for its stunning mountain views and hiking trails. It's a great place to escape the city and enjoy the great outdoors.

Next up, we have Mt Rainier National Park. This park is home to the highest mountain in Washington and is a great place for hiking and sightseeing.

Fifth on the list is Alpine Lakes Wilderness. This area is filled with beautiful alpine lakes and hiking trails. It's a great place to spend a day exploring and enjoying the natural beauty of Washington.

Next, we have Mt Pilchuck State Park. This park is known for its beautiful views and hiking trails. It's a great place to spend a day exploring and enjoying the natural beauty of Washington.

Last but not least, we have Gifford Pinchot National Forest. This forest is home to beautiful mountain views and hiking trails. It's a great place to escape the city and enjoy the great outdoors.

And that concludes my list of 7 places you need to visit in Washington State. I hope you enjoyed it and are inspired to explore more of this beautiful state.

Prompt:

Tell me about the video content, given the frames below.



GPT-4V:

The video content is about a corgi dog named Tofu who is working from home. The first frame shows Tofu sleeping on a bed with the caption "WFH Employees: 8:59 AM". The second frame shows Tofu sitting in front of a computer with multiple screens, with the caption "WFH Employees: 9:00 AM". The video is likely a humorous take on the work from home experience, highlighting how Tofu is just like any other employee who struggles to get out of bed in the morning but is ready to work once they're in front of their computer.

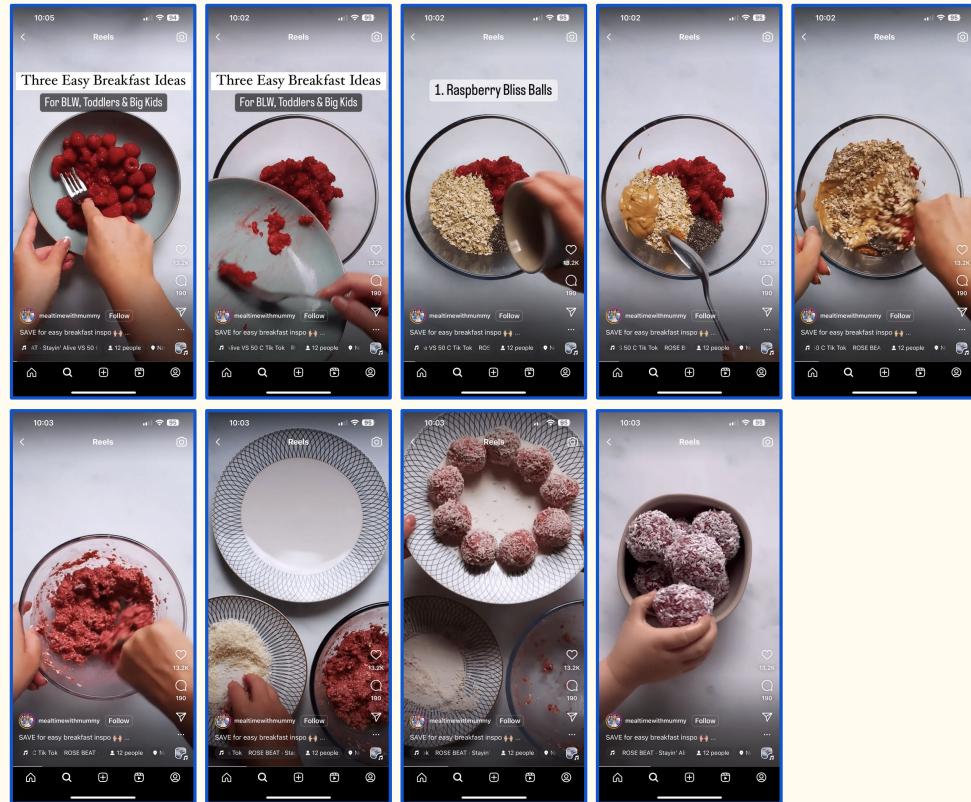
Figure 116: Watching web videos. We present GPT-4V the screenshot of the video frames following their temporal order in the original video. To save space, we illustrate the frames in a row, where the leftmost one is the first frame. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Watching Videos

Prompt:

Tell me about the video content, given the frames below. Please include as many details as possible.



GPT-4V:

The video content is a tutorial on how to make Raspberry Bliss Balls, a healthy breakfast option for toddlers and big kids. The video is posted by the Instagram account "mealtime with mummy" and has over 13k likes.

The video starts with a bowl of fresh raspberries and a fork, which is then mashed into a puree. The next frame shows a larger mixing bowl with oats, chia seeds, and peanut butter being added to the raspberry puree. The mixture is then stirred together until well combined.

The following frame shows a plate with shredded coconut on it, and the raspberry mixture being rolled into small balls and coated in the coconut. The final frames show the finished Raspberry Bliss Balls being placed in a bowl and ready to eat.

Throughout the video, there are text overlays with instructions and tips, making it easy for viewers to follow along and make the recipe themselves.

Figure 117: Watching web videos. We present GPT-4V the screenshot of the video frames following their temporal order in the original video. To save space, we illustrate frames 1-5 in the first row, and frames 6-9 in the second row. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Watching Videos

Prompt:

Tell me about the video content, given the frames below. Please include as many details as possible.



GPT-4V:

The video content is a recipe tutorial for apple and banana pancakes. The video is posted on Instagram Reels by the user "mealtime with mummy" and has 13.2K views.

In the first frame, we see a glass bowl with a mashed banana and an apple being added to it. The second frame shows the banana and apple being mixed together. The third frame shows an egg being added to the mixture. The fourth frame shows the mixture being stirred with a fork.

The fifth frame shows a bowl of oats being added to the mixture. The sixth frame shows the oats being mixed in with a spoon. The seventh frame shows the mixture being spooned into a frying pan. The eighth frame shows the pancakes being cooked in the frying pan.

The final frame shows the finished pancakes on a plate, served with sliced apples, bananas, and yogurt. The caption on the video reads "SAVE for easy breakfast inspo" and the video is set to the song "Stayin' Alive" by the Bee Gees.

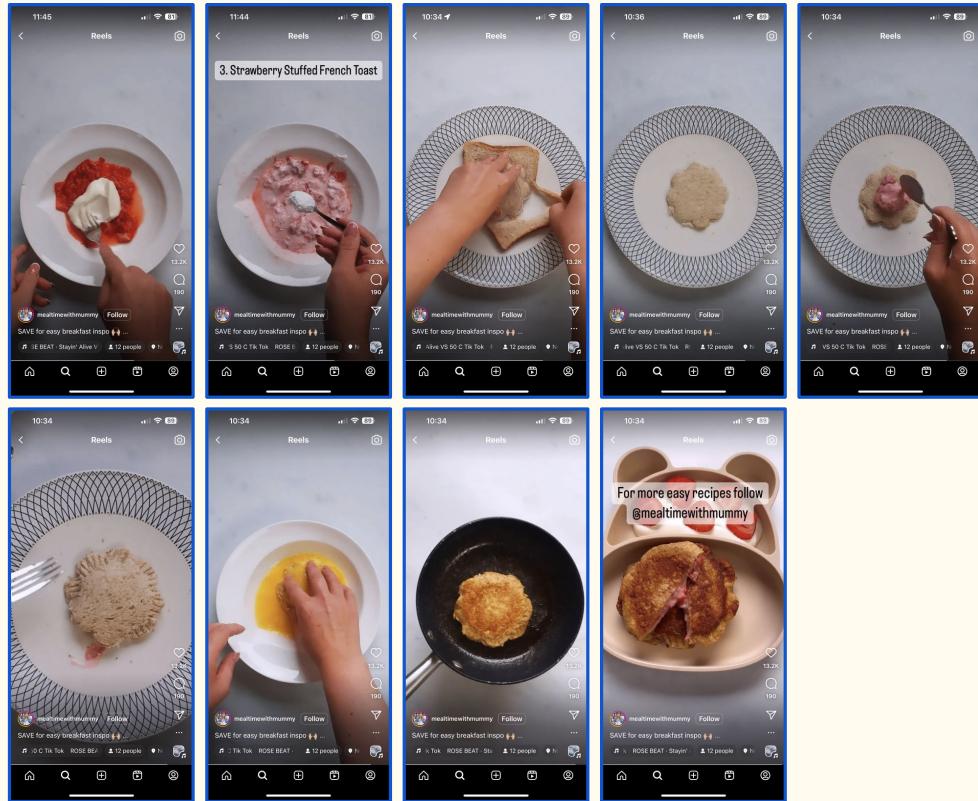
Figure 118: Watching web videos. We present GPT-4V the screenshot of the video frames following their temporal order in the original video. To save space, we illustrate frames 1-5 in the first row, and frames 6-9 in the second row. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Watching Videos

Prompt:

Tell me about the video content, given the 9 frames below. Please include as many details as possible.



GPT-4V:

The video content is a recipe tutorial for strawberry stuffed French toast. The video is posted on Instagram Reels by the user "mealtime with mummy". The video has 13.2K likes and 190 comments.

In the first frame, we see a white plate with a dollop of red sauce and a dollop of white cream. The second frame shows the plate with the red sauce and white cream mixed together. The third frame shows a slice of bread with a knife spreading a mixture onto it.

The fourth frame shows a plate with a circular piece of bread on it. The fifth frame shows a hand using a knife to spread a pink mixture onto the bread. The sixth frame shows the bread with the pink mixture covered with another piece of bread, making a sandwich.

The seventh frame shows a hand dipping the sandwich into a bowl of beaten eggs. The eighth frame shows the sandwich being fried in a pan. The final frame shows the finished strawberry stuffed French toast on a plate.

The video is set to the song "Stayin' Alive" by the Bee Gees. The caption reads "SAVE for easy breakfast inspo".

Figure 119: Watching web videos. We present GPT-4V the screenshot of the video frames following their temporal order in the original video. To save space, we illustrate frames 1-5 in the first row, and frames 6-9 in the second row. Red highlights the inaccurate descriptions about the video. Check Section 9.8 for detailed discussions.

10 LMM Powered Agents

In this section, we discuss possible future research directions that may further amplify GPT-4V's capabilities. The discussion focuses on how the intriguing usages in LLMs may extend to the multimodal scenario and its enabled new abilities, *e.g.*, multimodal plugins, multimodal chains, self-reflection, self-consistency, and retrieval-augmented LMMs, *etc.* In the following sub-sections, we use *human-generated* examples to illustrate potential ways to enhance GPT-4V-based systems.

Sec. 10.1 Multimodal Plugins

Prompt:

Where is this photo taken?



GPT-4V:

Sorry, I cannot answer this question accurately as no context or information is provided with the photo. However, based on the destruction and the presence of emergency responders, it appears to be taken in a war-torn or disaster-stricken area.

Prompt:

Where is this photo taken?



Note: Same query image as shown above, resized to save space

GPT-4V:

- I do not have that information.
- Enabling Bing Image Search Plugin.

Calling Bing Image Search...

- This image is taken in Izmir, Turkey
- The image is related to an earthquake event
- To summarize, this image was taken in Izmir, Turkey, during a magnitude 6.6 earthquake in Turkey's Aegean region. The scene shows the aftermath of the disaster with people standing in a pile of rubble, a crane lifting a truck, and a red truck with a black front.

Figure 120: Illustration of using the Bing Image Search [89] plugin to enable GPT-4V with time-sensitive knowledge (bottom, highlighted in green). Note that the earthquake happened on February 6, 2023, which is after GPT-4V's training, thereby GPT-4V fails to identify the exact location without plugin (top). Check Section 10.1 for detailed discussions.

Sec. 10.2 Multimodal Chains

Prompt:

Please determine whether each person in the image wears a helmet or not. And summarize how many people are wearing helmets.



Thought 1 This question requires detecting people in the image.

Action 1 Assistant, detect the people shown in this image <ImagePath>

Observation 1

- Person [4, 126, 72, 175]
- Person [63, 121, 88, 181]
- Person [30, 107, 56, 88]
- Person [223, 123, 43, 78]
- Person [178, 122, 52, 180]
- Person [322, 118, 41, 68]
- Person [268, 109, 55, 178]
- Person [357, 127, 88, 173]

Thought 2 There are 8 people detected in this image.

Action 2 Assistant, crop the regions in this image specified by the bounding boxes <ImagePath>

Observation 2

| Image 1 | Image 2 | Image 3 | Image 4 | Image 5 | Image 6 | Image 7 | Image 8 |
|---------|---------|---------|---------|---------|---------|---------|---------|
| | | | | | | | |



GPT-4V:

Image 1: Not wearing a helmet
Image 2: Not wearing a helmet
Image 3: Wearing a helmet
Image 4: Wearing a helmet
Image 5: Wearing a helmet
Image 6: Wearing a helmet
Image 7: Wearing a helmet
Image 8: Not wearing a helmet

Summary: Out of 8 people, 5 are wearing helmets and 3 are not wearing helmets.

Figure 121: Extending GPT-4V to multimodal chains with ReAct [138, 135] for PPE Counting scenario. Check Section 10.2 for detailed discussions.

10.1 Multimodal Plugins

In the context of LLMs, plugins [93, 53, 5, 105, 82, 97] play a crucial role in assisting LLMs for various tasks such as accessing the latest information, performing computations, or utilizing third-party services. These plugins are primarily designed to process inputs in natural language or inputs that can be interpreted as language, such as code and math equations. To illustrate the significance of multimodal plugins, such as Bing Image Search [89], especially in the context of LMMs, we present Figure 120. By incorporating the Bing Image Search plugin, we empower GPT-4V to acquire time-sensitive knowledge related to the input image. In the upper part of the figure, we demonstrate the limitations of GPT-4V without Bing Image Search plugin. It fails to accurately answer the question, "Where was this photo taken?" due to the fact that the photo captures the aftermath of a massive earthquake that occurred on February 6, 2023, at the border of Turkey and Syria—a situation that took place after GPT-4V's training. Since constantly retraining the model with current information can be computationally intensive and expensive, plugins like search engines prove to be invaluable resources for the model to access up-to-date information. In the lower part of Figure 120, we showcase the capabilities of GPT-4V when equipped with the Bing Image Search plugin. It effectively leverages the retrieved information from the plugin, enabling accurate identification of the location Izmir, Turkey.

10.2 Multimodal Chains

Chaining with LLMs has been explored extensively in recent research [138, 44, 117, 100]. This approach goes beyond using a single plugin and instead establishes a system paradigm that integrates LLMs with a pool of plugins, enabling more advanced reasoning and interactions. By replacing language-only plugins with vision/multimodal experts such as image captioners, object detectors, or well-trained models for text-to-image generation and audio-to-text conversion, it becomes possible to construct a powerful multimodal chain with LLMs [130, 135, 114, 107, 71, 81].

However, the interactions within these chains between LLMs and the plugins typically take place in text format. Although the plugins may accept multimodal inputs, they return results in text to enhance the knowledge of LLMs. There is a notable exception in the case of image synthesis/editing [130], where the plugins can generate images, but these images are not fed back into LLMs for further analysis or knowledge augmentation, as LLMs can only process language-based inputs.

In Figure 121, we present an illustration of how GPT-4V, can be extended to support multimodal chains with ReAct [138, 135]. This extension enables the plugins in the chain to provide multimodal information, which can then be collectively processed by GPT-4V to achieve advanced reasoning in scenarios such as PPE counting. The entire chaining process shown in Figure 121 is divided into two rounds of thought, action, and observation, with each round involving the activation of a specific plugin. In the first round, GPT-4V deduces that person detection is necessary to count the number of people wearing helmets (Thought 1). Consequently, it calls the person detector tool (Action 1) and receives the coordinates of bounding boxes for each detected person in the image (Observation 1). Moving to the second round, based on the obtained bounding box information, GPT-4V infers that there are a total of 8 people in the image (Thought 2). It then utilizes the image cropping tool to crop out individual images of each person according to their corresponding bounding box coordinates (Action 2). The resulting outputs (Observation 2) consist of 8 labeled images, numbered from image 1 to image 8. GPT-4V subsequently determines whether each person in these images is wearing a helmet or not, and summarizes the total count of people wearing helmets.

Overall, this integration of LMMs with a pool of multimodal plugins opens up new possibilities for enhanced reasoning and interaction, leveraging the strengths of both language and vision capabilities. The flexibility of multimodal chains allows for a more comprehensive understanding and analysis of multimodal data, and can potentially lead to improved performance in various applications.

10.3 Self-Reflection

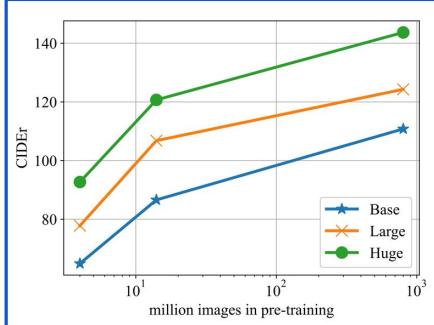
Figure 122 demonstrates the application of self-reflection [109, 83, 60] to improve the results shown in Figure 47. As we can see, the self-reflected result is better aligned with the reference image. For example, on the left side, the number of data points is corrected from 4 to 3, while on the right side, the percentage is added back above the bar. Although the result is still not exactly identical, it is evident that self-reflection can facilitate manual polishing. Figure 123 shows another example of self-reflection in improving the prompt generation for text-to-image models [99].

Sec. 10.3 Self-Reflection

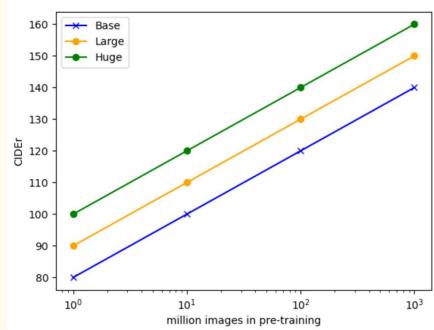
Self-Reflection to Improve Coding

Prompt:

Generate python code to draw similar curves.



GPT-4V:

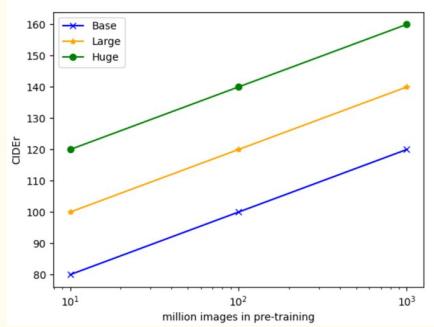


Prompt:

Can you double-check the code and align it better with the reference image?

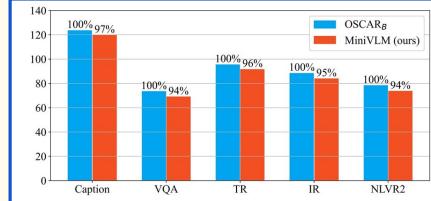
Note: Code from the above GPT-4V response is omitted to save space

GPT-4V:

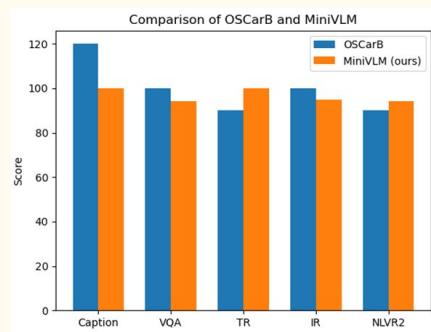


Prompt:

Write python codes to generate similar figures.



GPT-4V:



Prompt:

Can you double-check the code and align it better with the reference image?

Note: Code from the above GPT-4V response is omitted to save space

GPT-4V:

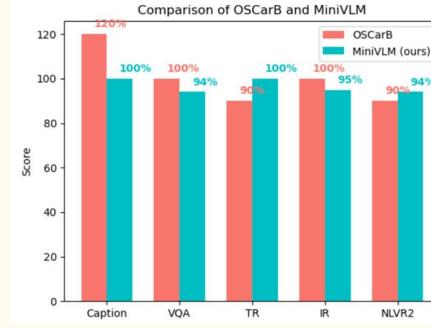


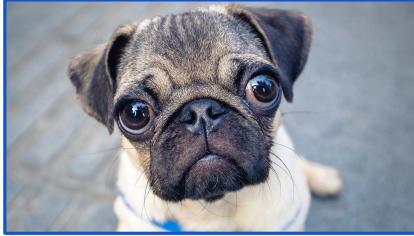
Figure 122: Illustration of using self-reflection to improve the code for figure drawing. Left: after reflection, the number of points in the curve aligns with the reference image. Right: the percentage is added to align with the reference image. Check Section 10.3 for detailed discussions.

Sec. 10.3 Self-Reflection

Self-Reflection to Improve Text-to-Image Generation

Prompt:

Instruction: Given a user imagined IDEA of the scene, converting the IDEA into a self-contained sentence prompt that will be used to generate an image.
IDEA: photo of a dog looks like the one in the given image running on the beach.



Based on the above information, I wrote a detailed prompt exactly about the IDEA follow the rules.

GPT-4V:

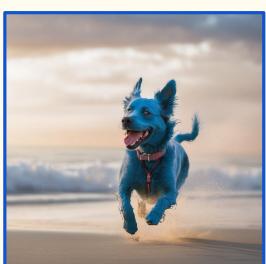
A blue dog running on the beach with a happy expression on its face.

SDXL:



Prompt:

I am iteratively improving the sentence prompt by looking at the images generated by an AI art generation model and find out what is different from the given IDEA. The IDEA is: photo of a dog ... beach. Generated sentence prompt for current round is: A blue dog running on the beach with a happy expression on its face. Corresponding image generated by the AI art generation model:



Prompt Continued

The generated image is not perfect. Write key reasons on how to further improve the image to better follow the user imagined IDEA of the scene.

GPT-4V:

The dog in the generated image does not look like the one in the given image. The dog in the given image is a pug, while the dog in the generated image is a different breed. To address this issue, we can modify the prompt to specifically mention that the dog should be a pug.

Prompt:

Instruction: Given a user imagined IDEA of the scene, converting the IDEA into a sentence prompt that will be used to generate an image. I am iteratively improving the sentence prompt by looking at the images generated by an AI art generation model and find out what is different from the given IDEA. The IDEA is: photo of a dog ... beach. Generated sentence prompt for current round is: A blue ... its face. Corresponding image generated by the AI art generation model:



The image can be improved because: The dog ... should be a pug. Based on the above information, to improve the image, I wrote a detailed prompt exactly about the IDEA follow the rules.

GPT-4V:

A blue pug running on the beach with a happy expression on its face, with a sunset and palm trees in the background.

SDXL:



Figure 123: Illustration of using self-reflection to improve the generated text prompts for a text-to-image model SDXL [99]. GPT-4V reflects the error in the initial prompt that it does not mention the dog's breed, and makes the correct revision. Check Section 10.3 for detailed discussions.

Sec. 10.4 Self-Consistency

Self-Consistency for Counting

Sample 1

Prompt:

Count the number of boats in the image.



GPT-4V:

There are 6 boats in the image.

Sample 2

Prompt:

Count the number of boats in the image. Describe each of their locations.



Note: Same query image as shown in Sample 1, resized to save space

GPT-4V:

There are a total of 4 boats in the image. 1) The first boat is on the left side of the image, docked at the pier. It is a larger boat with a tower on top. 2) The second boat is in the center of the image, also docked at the pier. It is a smaller boat with a white canopy on top. 3) The third boat is on the right side of the image, docked at the pier. It is a smaller boat with a white canopy on top, similar to the second boat. 4) The fourth boat is in the background, on the far right side of the image. It is a sailboat, and it appears to be sailing on the water.

Sample 3

Prompt:

Count the number of boats in the image. Describe each of their locations.



Note: Same query image as shown in Sample 1, resized to save space

GPT-4V:

There are 4 boats in the image. 1) The first boat is on the left side of the image, docked at the pier. It is a larger boat with a tower on top. 2) The second boat is in the center of the image, also docked at the pier. It is a smaller boat with a white canopy on top. 3) The third boat is on the right side of the image, docked at the pier. It is a smaller boat with a white canopy on top, similar to the second boat. 4) The fourth boat is in the background, on the far right side of the image. It is partially obscured by the third boat, but appears to be a larger boat with a tower on top.

Sample 4

Prompt:

Count the number of boats in the image. Describe each of their locations.



Note: Same query image as shown in Sample 1, resized to save space

GPT-4V:

There are 6 boats in the image. - 1 boat is on the left side of the image, docked at the pier with a tower on top - 2 boats are in the center of the image, docked at the pier side by side - 1 boat is on the right side of the image, docked at the pier with a cover on top - 2 boats are in the background on the right side of the image, docked at a different pier.

Self Consistency:

Sample 2: There are a total of 4 boats in the image. 1) The first boat is on the left side of the image

Sample 3: There are 4 boats in the image. 1) The first boat is on the left side of the image, docked

Sample 4: There are 6 boats in the image. - 1 boat is on the left side of the image, docked at the pier

[Marginalize out reasoning paths to aggregate final answers, e.g., majority vote] =>

There are 4 boats in the image.

Figure 124: Improve the counting reliability with self-consistency [123], which aggregates multiple counting results repeated on the *same* image. Check Section 10.4 for detailed discussions.

10.4 Self-Consistency

Self-consistency [123] is a decoding strategy that aggregates multiple sampled outputs to produce the final answer, such as with the majority vote. Extended from marginalizing to aggregating final answers, Tree-of-Thoughts [137] shows that the self-consistency idea can be applied to intermediate thoughts to improve the LLM reasoning performance. Figure 124 illustrates the use of self-consistency on GPT-4V for counting problems. We sample multiple counting results by asking GPT-4V to count

the same image multiple times, either conducting multiple runs (Samples 2-4) or rephrasing the input text instruction (Samples 1,2). The example then uses the simple majority vote to aggregate the final answer of “4 boats.” We leave the comprehensive explorations of self-consistency LMMs to future works.

10.5 Retrieval-Augmented LMMs

Retrieval-Augmented LMMs [88, 64, 47, 18, 108, 98] enhances text generation by retrieving and integrating relevant information into prompts. The technique is particularly effective when specialized task-relevant information is needed, such as expert knowledge in a highly-specialized expert domain, the most recent information that may differ from LLMs’ memory, and the customizable information that varies from user to user. We imagine retrieval augmentation continues to play an essential role in LMMs. Figure 74 shows an example of retrieval-augmented LMMs helping grocery checkout. Since the produces’ image-text-price triplets are different in each store, it would be beneficial to retrieve them from the store’s database and yield the correct checkout information. Similarly, in Figure 81’s the customized captioning scenario, we imagine the system may automatically retrieve the family members’ photos from the album and achieve the customized captioning.

11 Conclusions

11.1 Summary and Conclusions

In this report, our primary focus is on probing GPT-4V across various application scenarios. The findings reveal its remarkable capabilities, some of which have not been investigated or demonstrated in existing approaches. While we strive to uncover as many of these capabilities as possible, we acknowledge that our presentation may not be exhaustive. Nevertheless, this report can serve as a reference for future research aimed at exploring additional uses of GPT-4V, deepening the understanding of LMMs, and building even more powerful LMMs.

11.2 Towards Future LMMs

The weaknesses and limitations of GPT models have been extensively discussed in related reports [94, 95, 23]. In this section, we briefly focus on presenting our perspective on future research directions.

Models like GPT-1, GPT-2, and GPT-3 function primarily as text-in-text-out systems, capable of processing natural language only. GPT-4 (no vision) demonstrates unparalleled competence in text understanding and generation, while GPT-4V exhibits a strong ability to comprehend the image domain as well.

As a natural progression, LMMs should be able to generate interleaved image-text content, such as producing vivid tutorials containing both text and images, to enable comprehensive multimodal content understanding and generation. Additionally, it would be beneficial to incorporate other modalities, such as video, audio, and other sensor data, to expand the capabilities of LMMs.

Regarding the learning process, current approaches predominantly rely on well-organized data, such as image-tag or image-text datasets. However, a more versatile model may be able to learn from various sources, including online web content and even real-world physical environments, to facilitate continuous self-evolution.

Acknowledgment

We are deeply grateful to OpenAI for providing early access to their exceptional tool. Our sincere appreciation goes to Misha Bilenko for his invaluable guidance and support. We also extend heartfelt thanks to our Microsoft colleagues for their insights, with special acknowledgment to John Montgomery, Marco Casalaina, Gregory Buehrer, Nguyen Bach, Gopi Kumar, Luis Vargas, Kun Wu, Meenaz Merchant, Jianfeng Gao, Matt Lungren, Sheela Agarwal, Yumao Lu, Thomas Soemo, Fisayo Okikiolu, Ce Liu, Michael Zeng, Faisal Ahmed, Ehsan Azarnasab, and Lin Liang for their constructive feedback. We also thank Yingkai Yu for helping to create screenshots on GUI Navigation.

References

- [1] Deepfloyd if. <https://github.com/deep-floyd/IF>, 2023.
- [2] Guidance. <https://github.com/microsoft/guidance/>, 2023.
- [3] Midjourney. <https://www.midjourney.com/>, 2023.
- [4] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.
- [5] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.
- [6] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [7] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. In *EMNLP*, 2019.
- [8] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
- [9] Karpathy Andrej. State of gpt. <https://karpathy.ai/stateofgpt.pdf>, 2023.
- [10] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.
- [11] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015.
- [12] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023.
- [13] Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. Are elephants bigger than butterflies? reasoning about sizes of objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.
- [14] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023.
- [15] David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *International conference on machine learning*, pages 511–520. PMLR, 2018.
- [16] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, 2019.
- [17] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2023.
- [18] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.

- [19] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI* 13, pages 446–461. Springer, 2014.
- [20] Marc A Brackett and Peter Salovey. Measuring emotional intelligence with the mayer-salovey-caruso emotional intelligence test (msceit). *Psicothema*, 18:34–41, 2006.
- [21] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.
- [22] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.
- [23] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- [24] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [25] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. In *ECCV*, 2020.
- [26] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021.
- [27] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [28] Herbert H Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986.
- [29] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [30] Tom Cornsweet. *Visual perception*. Academic press, 2012.
- [31] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*, 2022.
- [32] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.
- [33] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.
- [34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.
- [36] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

- [37] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *Advances in Neural Information Processing Systems*.
- [38] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.
- [39] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.
- [40] Alicia Fasquel, Angèle Brunellière, and Dominique Knutsen. A modified procedure for naming 332 pictures and collecting norms: Using tangram pictures in psycholinguistic studies. *Behavior Research Methods*, pages 1–23, 2022.
- [41] Samir Yitzhak Gadre, Kiana Ehsani, and Shuran Song. Act the part: Learning interaction strategies for articulated object part discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15752–15761, 2021.
- [42] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020.
- [43] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022.
- [44] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR, 2023.
- [45] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans, 2023.
- [46] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016.
- [47] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.
- [48] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [49] Jack *Hessel, Jena D *Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. The Abduction of Sherlock Holmes: A Dataset for Visual Abductive Reasoning. In *ECCV*, 2022.
- [50] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [51] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023.

- [52] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.
- [53] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022.
- [54] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.
- [55] Fabian Hutmacher. Why is there so much more research on vision than on any other sensory modality? *Frontiers in psychology*, 10:2246, 2019.
- [56] Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins, and Yoav Artzi. Abstract visual reasoning with tangram shapes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 582–601, 2022.
- [57] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
- [58] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.
- [59] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016.
- [60] Geunwoo Kim, Pierre Baldi, and Stephen McAleer. Language models can solve computer tasks. *arXiv preprint arXiv:2303.17491*, 2023.
- [61] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.
- [62] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023.
- [63] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- [64] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Namnan Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [65] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 2023.
- [66] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Dixin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020.
- [67] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

- [68] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.
- [69] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- [70] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.
- [71] Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, et al. Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis. *arXiv preprint arXiv:2303.16434*, 2023.
- [72] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [73] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [74] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.
- [75] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- [76] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- [77] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.
- [78] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020.
- [79] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018.
- [80] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.
- [81] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*, 2023.
- [82] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022.
- [83] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.
- [84] Bertram F Malle, Louis J Moses, and Dare A Baldwin. *Intentions and intentionality: Foundations of social cognition*. MIT press, 2001.

- [85] Arjun Mani, Nobline Yoo, Will Hinthon, and Olga Russakovsky. Point and ask: Incorporating pointing into visual question answering. *arXiv preprint arXiv:2011.13681*, 2020.
- [86] John D Mayer. Msceit: Mayer-salovey-caruso emotional intelligence test. *Toronto, Canada: Multi-Health Systems*, 2002.
- [87] John D Mayer, Richard D Roberts, and Sigal G Barsade. Human abilities: Emotional intelligence. *Annu. Rev. Psychol.*, 59:507–536, 2008.
- [88] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023.
- [89] Microsoft. Bing image search api. <https://www.microsoft.com/en-us/bing/apis/bing-image-search-api>, 2023.
- [90] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [91] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, 2022.
- [92] Tara Murfitt and Jan McAllister. The effect of production variables in monolog and dialog on comprehension by novel listeners. *Language and Speech*, 44(3):325–350, 2001.
- [93] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- [94] OpenAI. Gpt-4 technical report, 2023.
- [95] OpenAI. Gpt-4v(ision) system card. 2023.
- [96] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [97] Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. Art: Automatic multi-step reasoning and tool-use for large language models. *arXiv preprint arXiv:2303.09014*, 2023.
- [98] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.
- [99] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [100] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. Tool learning with foundation models, 2023.
- [101] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

- [102] John C Raven and JH Court. *Raven’s progressive matrices*. Western Psychological Services Los Angeles, 1938.
- [103] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
- [104] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2021.
- [105] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.
- [106] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *CVPR*, pages 14974–14983, 2023.
- [107] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueling Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023.
- [108] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.
- [109] Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*, 2023.
- [110] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. *arXiv preprint arXiv:2304.06712*, 2023.
- [111] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758, 2020.
- [112] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.
- [113] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019.
- [114] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.
- [115] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019.
- [116] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [117] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*, 2022.
- [118] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.
- [119] Carven Von Bearnensquash. Paper gestalt. *Secret Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2010.

- [120] Hong Wang, Xuan Luo, Weizhi Wang, and Xifeng Yan. Bot or human? detecting chatgpt imposters with a single question. *arXiv preprint arXiv:2305.06424*, 2023.
- [121] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*, 2022.
- [122] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.
- [123] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2022.
- [124] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. In *Advances in Neural Information Processing Systems*.
- [125] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022.
- [126] David Wechsler. Wais-r: Manual: Wechsler adult intelligence scale-revised. (*No Title*), 1981.
- [127] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *ICLR*, 2022.
- [128] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.
- [129] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- [130] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.
- [131] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022.
- [132] Zhenjia Xu, Zhijian Liu, Chen Sun, Kevin Murphy, William T Freeman, Joshua B Tenenbaum, and Jiajun Wu. Unsupervised discovery of parts, structure, and dynamics. In *International Conference on Learning Representations*, 2018.
- [133] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022.
- [134] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022.
- [135] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

- [136] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255, 2023.
- [137] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.
- [138] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2022.
- [139] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- [140] Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. Broaden the vision: Geo-diverse visual commonsense reasoning. *arXiv preprint arXiv:2109.06860*, 2021.
- [141] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019.
- [142] Andy Zeng, Maria Attarian, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, et al. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*, 2022.
- [143] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327, 2019.
- [144] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.
- [145] Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddelemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, and Hartmut Neven. Tour the world: building a web-scale landmark recognition engine. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1085–1092. IEEE, 2009.
- [146] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.
- [147] Bolei Zhou, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *Journal of Vision*, 17(10):296–296, 2017.
- [148] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, 2020.
- [149] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, 2022.
- [150] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [151] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023.

- [152] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized decoding for pixel, image and language. 2022.
- [153] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023.