
Gemini vs GPT-4V : A Preliminary Comparison and Combination of Vision-Language Models Through Qualitative Cases

Zhangyang Qi^{1,7*} Ye Fang^{2,7*} Mengchen Zhang^{3,7*} Zeyi Sun^{4,7*}
 Tong Wu^{5,7} Ziwei Liu^{6,7} Dahua Lin^{5,7} Jiaqi Wang^{7†} Hengshuang Zhao^{1†}

* Equal contribution † Corresponding author

¹The University of Hong Kong ²Fudan University ³Zhejiang University

⁴Shanghai Jiao Tong University ⁵The Chinese University of Hong Kong

⁶Nanyang Technological University ⁷Shanghai AI Laboratory

{zyqi, hszhao}@cs.hku.hk, wangjiaqi@pjlab.org.cn

<https://github.com/Qi-Zhangyang/Gemini-vs-GPT4V>

Abstract

The rapidly evolving sector of Multi-modal Large Language Models (MLLMs) is at the forefront of integrating linguistic and visual processing in artificial intelligence. This paper presents an in-depth comparative study of two pioneering models: Google’s Gemini and OpenAI’s GPT-4V(ision). Our study involves a multi-faceted evaluation of both models across key dimensions such as Vision-Language Capability, Interaction with Humans, Temporal Understanding, and assessments in both Intelligence and Emotional Quotients. The core of our analysis delves into the distinct visual comprehension abilities of each model. We conducted a series of structured experiments to evaluate their performance in various industrial application scenarios, offering a comprehensive perspective on their practical utility. We not only involve direct performance comparisons but also include adjustments in prompts and scenarios to ensure a balanced and fair analysis. Our findings illuminate the unique strengths and niches of both models. GPT-4V distinguishes itself with its precision and succinctness in responses, while Gemini excels in providing detailed, expansive answers accompanied by relevant imagery and links. These understandings not only shed light on the comparative merits of Gemini and GPT-4V but also underscores the evolving landscape of multimodal foundation models, paving the way for future advancements in this area. After the comparision, we attempted to achieve better results by combining the two models. Finally, We would like to express our profound gratitude to the teams behind GPT-4V [1] and Gemini [2] for their pioneering contributions to the field. Our acknowledgments are also extended to the comprehensive qualitative analysis presented in ‘Dawn’ by Yang *et al.* [3]. This work, with its extensive collection of image samples, prompts, and GPT-4V-related results, provided a foundational basis for our analysis.

Contents

1	Introduction	7
1.1	Motivation and Overview	7
1.2	Gemini’s Input Modes	8
1.3	Prompt Techniques	9
1.4	Sample Collection	9
1.5	Takeaways (Conclusion)	9
2	Image Recognition and Understanding	10
2.1	Basic object Recognition	10
2.2	Landmark Recognition	10
2.3	Food Recognition	10
2.4	Logo Recognition	10
2.5	Abstract Image Recognition	10
2.6	Scene Understanding	10
2.7	Counterfactual Examples	11
2.8	Object Counting	11
2.9	Spot the Difference	11
3	Text Recognition and Understanding in Images	24
3.1	Scene Text Recognition	24
3.2	Equation Recognition	24
3.3	Chart Text Recognition	24
4	Image Reasoning Abilities	31
4.1	Humorous Image Understanding	31
4.2	Multimodal Knowledge and Commonsense	31
4.3	Detective Reasoning Ability	31
4.4	Association of Parts and Objects	31
4.5	Intelligence Tests	31
4.6	Emotional Intelligence Tests	31
5	Textual Reasoning in Images	47
5.1	Visual Math Ability	47
5.2	Table & Chart Understanding and Reasoning	47
5.3	Document Understanding and Reasoning	47
6	Integrated Image and Text Understanding	58
6.1	Interleaved Image-text Inputs	58
6.2	Text-to-Image Generation Guidance	58

7 Object Localization	64
7.1 Object localization in real-world	64
7.2 Abstract Image Localization	64
8 Temporal Video Understanding	67
8.1 Action Recognition	67
8.2 Temporal Ordering	67
9 Multilingual Capabilities	68
9.1 Multilingual Image Description	68
9.2 Multilingual Scene Text Recognition	68
10 Industry Application	77
10.1 Industry: Defect Detection	77
10.2 Industry: Grocery Checkout	77
10.3 Industry: Auto Insurance	77
10.4 Industry: Customized Captioner	77
10.5 Industry: Evaluation Image Generation	77
10.6 Industry: Embodied Agent	78
10.7 Industry: GUI Navigation	78
11 Integrated Use of GPT-4V and Gemini	111
11.1 Product Identification and Recommendation	111
11.2 Multi-Image Recognition and Story Generation	111
12 Conclusion	114

List of Figures

1 Section 2.1 Basic Object Recognition	11
2 Section 2.2 Landmark Recognition (1)	12
3 Section 2.2 Landmark Recognition (2)	13
4 Section 2.3 Food Recognition (1)	14
5 Section 2.3 Food Recognition (2)	15
6 Section 2.4 Logo Recognition (1)	16
7 Section 2.4 Logo Recognition (2)	17
8 Section 2.4 Logo Recognition (3)	18
9 Section 2.5 Abstract Image Recognition	19
10 Section 2.6 Scene Understanding	20
11 Section 2.7 Counterfactual Examples	21
12 Section 2.8 Object Counting	22
13 Section 2.9 Spot the Differences	23

14	Section 3.1 Scene Text Recognition (1)	25
15	Section 3.1 Scene Text Recognition (2)	26
16	Section 3.1 Scene Text Recognition (3)	27
17	Section 3.2 Equation Recognition	28
18	Section 3.3 Chart Text Recognition (1)	28
19	Section 3.3 Chart Text Recognition (2)	29
20	Section 3.3 Chart Text Recognition (3)	30
21	Section 4.1 Humorous Image Understanding	32
22	Section 4.2 Multimodal Knowledge and Commonsense (1)	33
23	Section 4.2 Multimodal Knowledge and Commonsense (2)	34
24	Section 4.2 Multimodal Knowledge and Commonsense (3)	35
25	Section 4.3 Detective Reasoning Ability	36
26	Section 4.4 Association of Parts and Objects	37
27	Section 4.5 Intelligence Tests (1)	38
28	Section 4.5 Intelligence Tests (2)	39
29	Section 4.5 Intelligence Tests (3)	40
30	Section 4.5 Intelligence Tests (4)	41
31	Section 4.6 Emotional Intelligence Tests (1)	42
32	Section 4.6 Emotional Intelligence Tests (2)	43
33	Section 4.6 Emotional Intelligence Tests (3)	44
34	Section 4.6 Emotional Intelligence Tests (4)	45
35	Section 4.6 Emotional Intelligence Tests (5)	46
36	Section 5.1 Visual Math Ability	48
37	Section 5.2 Table & Chart Understanding and Reasoning (1)	49
38	Section 5.2 Table & Chart Understanding and Reasoning (2)	50
39	Section 5.2 Table & Chart Understanding and Reasoning (3)	51
40	Section 5.2 Table & Chart Understanding and Reasoning (4)	52
41	Section 5.2 Table & Chart Understanding and Reasoning (5)	53
42	Section 5.3 Document Understanding and Reasoning (1)	54
43	Section 5.3 Document Understanding and Reasoning (2)	55
44	Section 5.3 Document Understanding and Reasoning (3)	56
45	Section 5.3 Document Understanding and Reasoning (4)	57
46	Section 6.1 Interleaved Image-text Inputs (1)	59
47	Section 6.1 Interleaved Image-text Inputs (2)	60
48	Section 6.2 Text-to-Image Generation Guidance (1)	61
49	Section 6.2 Text-to-Image Generation Guidance (2)	62
50	Section 6.2 Text-to-Image Generation Guidance (3)	63
51	Section 7.1 Object localization in real-world (2)	64
52	Section 7.1 Object localization in real-world (3)	65

53	Section 7.2 Abstract Image Localization	66
54	Section 8.1 Action Recognition	67
55	Section 8.2 Temporal Ordering	68
56	Section 9.1 Multilingual Image Description (1)	69
57	Section 9.1 Multilingual Image Description (2)	70
58	Section 9.1 Multilingual Image Description (3)	71
59	Section 9.1 Multilingual Image Description (4)	72
60	Section 9.2 Multilingual Scene Text Recognition (1)	73
61	Section 9.2 Multilingual Scene Text Recognition (2)	74
62	Section 9.2 Multilingual Scene Text Recognition (3)	75
63	Section 9.2 Multilingual Scene Text Recognition (4)	76
64	Section 10.1 Industry: Defect Detection (1)	79
65	Section 10.1 Industry: Defect Detection (2)	80
66	Section 10.1 Industry: Defect Detection (3)	81
67	Section 10.2 Industry: Grocery Checkout	82
68	Section 10.3 Industry: Auto Insurance (1)	83
69	Section 10.3 Industry: Auto Insurance (2)	84
70	Section 10.3 Industry: Auto Insurance (3)	85
71	Section 10.4 Industry: Customized Captioner	86
72	Section 10.5 Industry: Evaluation Image Generation (1)	87
73	Section 10.5 Industry: Evaluation Image Generation (2)	88
74	Section 10.5 Industry: Evaluation Image Generation (3)	89
75	Section 10.6 Industry: Embodied Agent (1)	90
76	Section 10.6 Industry: Embodied Agent (2)	91
77	Section 10.6 Industry: Embodied Agent (3)	92
78	Section 10.6 Industry: Embodied Agent (4)	93
79	Section 10.7 Industry: GUI Navigation (1)	94
80	Section 10.7 Industry: GUI Navigation (2)	95
81	Section 10.7 Industry: GUI Navigation (3)	96
82	Section 10.7 Industry: GUI Navigation (4)	97
83	Section 10.7 Industry: GUI Navigation (5)	98
84	Section 10.7 Industry: GUI Navigation (6)	99
85	Section 10.7 Industry: GUI Navigation (7)	100
86	Section 10.7 Industry: GUI Navigation (8)	101
87	Section 10.7 Industry: GUI Navigation (9)	102
88	Section 10.7 Industry: GUI Navigation (10)	103
89	Section 10.7 Industry: GUI Navigation (11)	104
90	Section 10.7 Industry: GUI Navigation (12)	105
91	Section 10.7 Industry: GUI Navigation (13)	106

92	Section 10.7 Industry: GUI Navigation (14)	107
93	Section 10.7 Industry: GUI Navigation (15)	108
94	Section 10.7 Industry: GUI Navigation (16)	109
95	Section 10.7 Industry: GUI Navigation (17)	110
96	Section 11.1 Integrated Use: Product Identification and Recommendation	112
97	Section 11.2 Integrated Use: Multi-image Recognition and Story Generation	113

1 Introduction

1.1 Motivation and Overview

The evolution of artificial intelligence has seen the significant rise of Large Language Models (LLMs) [4, 5, 6, 7, 8, 9], which have revolutionized the way machines process and understand textual data. Building upon this, the advent of Multi-modal Large Language Models (MLLMs) [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21] marks a pivotal advancement in AI, extending capabilities to comprehend and interact with not just text, but also images, 3D models [22], and video content [23]. Among these modalities, the integration of text and image has emerged as particularly powerful, largely due to the rich and informative nature of text-image pairs. The following, unless otherwise specified, all refer to the MLLMs in the context of images.

The landscape of Multi-modal Large Language Models (MLLMs) is currently divided into two broad categories: closed-source models with their proprietary advancements, and open-source deployable models like LLaVA [16], MiniGPT-4 [15] and InstructBLIP [17] which are more accessible but often less advanced. Among these, the state-of-the-art in open-source models is GPT-4V [24] from OpenAI, which has established a dominant position in terms of versatility and general applicability. Recently, Google has introduced their own large model, Gemini [25], which also boasts high generalization capabilities. This release poses a significant challenge to GPT-4V's leading status. Gemini's entry into the arena of MLLMs brings a new dimension to the field, potentially reshaping the landscape of what is achievable with open-source AI technology, particularly in terms of multimodal understanding and application. Therefore, our paper undertakes a comprehensive comparison of these two models across multiple dimensions and domains.

It should be noted that the image samples, prompts, and results related to GPT-4V used in our paper are referenced from the study "The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). [26]" Our work can be seen as a continuation and expansion of this previous research. Since GPT-4V has already been extensively discussed in the original paper, our focus in this report will primarily be on emphasizing and exploring the unique characteristics and capabilities of Gemini.

Sec. 2 to Sec. 6 divide the multimodal evaluation into five aspects. The first level involves basic recognition of images and the text within them. The second level goes beyond recognition to require further inference and reasoning. The third level encompasses multimodal comprehension and inference involving multiple images. We have divided them into the following five sections.

- **Image Recognition and Understanding:** Sec. 2 addresses the fundamental recognition and comprehension of image content without involving further inference, including tasks such as identifying landmarks, foods, logos, abstract images, autonomous driving scenes, misinformation detection, spotting differences, and object counting.
- **Text Recognition and Understanding in Images:** Sec. 3 concentrates on text recognition (including OCR) within images, such as scene text, mathematical formulas, and chart & table text recognition. Similarly, no further inference of text content is performed here.
- **Image Inference Abilities:** Beyond basic image recognition, Sec. 4 involves more advanced reasoning. This includes understanding humor and scientific concepts, as well as logical reasoning abilities like detective work, image combinations, look for patterns in intelligence tests (IQ Tests), and emotional understanding and expression (EQ Tests).
- **Textual Inference in Images:** Building on the text recognition, Sec. 5 involves further reasoning beyond text recognition, including mathematical problem-solving, chart & table information reasoning, and document comprehension like paper, report and Graphic Design.
- **Integrated Image and Text Understanding:** Sec. 6 evaluates the collective understanding and reasoning abilities involving both image and text. For instance, tasks include settling items from a supermarket shopping cart, as well as guiding and modifying image generation.

Sec. 7 to Sec. 9 evaluate performance in three specialized tasks, namely, object localization, temporal understanding, and multilingual comprehension.

- **Object Localization:** Sec. 7 highlights object localization capabilities, tasking the models with providing relative coordinates for specified objects. This includes a focus on outdoor objects like cars in parking lots and abstract image localization.

- **Temporal Video Understanding:**
crefSec.8 Temporal Video Understanding evaluates the models' comprehension of temporality using key frames. This section includes two tasks: one involving the understanding of video sequences and the other focusing on sorting key frames.
- **Multilingual Capabilities:** Sec. 9 thoroughly assesses capabilities in recognizing, understanding, and producing content in multiple languages. This includes the ability to recognize non-English content within images and express information in other languages.

Sec. 10 presents various application scenarios for multimodal large models. We aim to showcase more possibilities to the industry, providing innovative ideas. There is potential to customize multimodal large models for unique domains. Here, we demonstrate seven sub-domains:

- **Industry: Defect Detection:** This task involves the detection of defects in products on industrial assembly lines, including textiles, metal components, pharmaceuticals and more.
- **Industry: Grocery Checkout:** This refers to an autonomous checkout system in supermarkets, aimed at identifying all items in a shopping cart for billing. The goal is to achieve comprehensive recognition of all items within the shopping cart.
- **Industry: Auto Insurance:** This task involves evaluating the extent of damage in car accidents and providing approximate repair costs, as well as offering repair recommendations.
- **Industry: Customized Captioner:** The aim is to identify the relative positions of various objects within a scene, with object names provided as condition and prompts in advance.
- **Industry: Evaluation Image Generation:** This involves assessing the alignment between generated images and given text prompts, evaluating the quality of the generation model.
- **Industry: Embodied Agent:** This application involves deploying the model in embodied intelligence and smart home systems, offering thoughts and decisions for indoor scenarios.
- **Industry: GUI Navigation:** This task focuses on guiding users through PC/Mobile GUI interfaces, assisting with information reception, online searches, and shopping tasks.

Finally, in Sec. 11, we explore how to combine both SOTA models to leverage their respective strengths and mitigate their weaknesses. In summary, GPT-4V provides more accurate results, while Gemini excels in providing more detailed responses, along with image and link outputs.

1.2 Gemini's Input Modes

Our goal is to clarify the input modality of Gemini. GPT-4V's input modality supports the continuous ingestion of multiple images as context, thereby possessing enhanced memory capabilities. However, for Gemini, its unique attributes are manifested in several aspects, as follows:

- **Single Image Input:** Gemini is limited to inputting a single image at a time. Additionally, it cannot process independent images; instead, it requires accompanying textual instructions.
- **Limited Memory Capacity:** Unlike GPT-4V, Gemini's multimodal module lacks the ability to retain memory of past image inputs and outputs. Therefore, when dealing with multiple images, our approach requires combining all the images into a single image input. This integrated input mode will be used unless explicitly stated otherwise.
- **Sensitive Information Masking:** Gemini exhibits some degree of obfuscation when processing images containing explicit facial or medical information, making it unable to recognize these images. This may impose certain limitations on its generalization ability.
- **Image and Link Output:** Unlike GPT-4V, which is limited to generating textual outputs, Gemini has the ability to create images related to the content and provide corresponding links. This establishes a higher level of association similar to search engine functionality.
- **Video Input and Comprehension:** Gemini demonstrates the capability to understand videos and requires a YouTube link as a video input. It's important to note that it can effectively process videos accompanied by accurate subtitle files. However, its comprehension ability may be limited when dealing with single, simple, and information-scarce videos.

1.3 Prompt Techniques

Prompt Engineering holds significant importance for both unimodal language models [27, 28, 29, 4, 30, 31] and multimodal large-scale models [32, 33, 17]. The prompt design under consideration is tailored for GPT-4V, and direct input into Gemini may yield unsatisfactory responses. In such cases, adjustments to Gemini's prompt are made to align with the input requirements of its architecture.

1.4 Sample Collection

All our data is sourced from "The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)" [26] (except for the images in Section 11, which are sourced from the internet). We have utilized their images, GPT-4V's prompts, and corresponding results. Our work can be seen as a continuation of theirs. Our dataset is diverse and maintains privacy protections. We extend our gratitude to the authors of that work. The raw data of the images is available on the [project page](#).

1.5 Takeaways (Conclusion)

We have conducted a comprehensive comparison of GPT-4V and Gemini's multimodal understanding and reasoning abilities across multiple aspects and have reached the following conclusions:

- **Image Recognition and Understanding:** In basic image recognition tasks, both models show comparable performance and are capable of completing the tasks effectively.
- **Text Recognition and Understanding in Images:** Both models excel in extracting and recognizing text from images. However, improvements are needed in complex formula and dashboard recognition. Gemini performs better in reading table information.
- **Image Inference Abilities:** In image reasoning, both models excel in common-sense understanding. Gemini slightly lags in look-for-pattern compared (IQ Tests) to GPT-4V. In EQ tests, both understand emotions and have aesthetic judgment.
- **Textual Inference in Images:** In the field of text reasoning, Gemini shows relatively lower performance levels when dealing with complex table-based reasoning and mathematical problem-solving tasks. Furthermore, Gemini tends to offer more detailed outputs.
- **Integrated Image and Text Understanding:** In tasks involving complex text and images, Gemini falls behind GPT-4V due to its inability to input multiple images at once, although it performs similarly to GPT-4V in textual reasoning with single images.
- **Object Localization:** Both models perform similarly in real-world object localization, with Gemini being slightly less adept at abstract image (tangram) localization.
- **Temporal Video Understanding:** In understanding temporality, Gemini's single-image input mode falls short compared to GPT-4V, especially in the comprehension of sequence.
- **Multilingual Capabilities:** Both models exhibit good multilingual recognition, understanding, and output capabilities, effectively completing the multilingual tasks.

In industrial applications, Gemini is outperformed by GPT-4V in **Embodied Agent** and **GUI Navigation**, which is also attributed to Gemini's single-image, non-memory input mode. Combining two large models can leverage their respective strengths.

Overall, both Gemini and GPT-4V are powerful and impressive multimodal large models. In terms of overall performance, GPT-4V is slightly stronger than Gemini Pro. This aligns with the results reported by Gemini. We look forward to the release of Gemini Ultra and GPT-4.5, which are expected to bring more possibilities to the field of visual multimodal applications.

2 Image Recognition and Understanding

In this section, we primarily discuss the fundamental understanding of images. This task is the most basic, requiring only the identification of objects in an image and their characteristics. It does not involve text-related tasks or further inference. Sec. 2.1, Sec. 2.2, Sec. 2.3 and Sec. 2.4 focus on the recognition of basic objects, landmarks, food, logos, and abstract images. Sec. 2.6 addresses scene understanding in outdoor autonomous driving scenarios. Sec. 2.7 tests the model’s ability to recognize fabricated objects created using text, gauging its discernment of real versus fictitious elements. Sec. 2.8 assesses the model’s object counting capabilities, while the final Sec. 2.9 explores the model’s proficiency in spotting differences, examining its ability to identify subtle details.

2.1 Basic object Recognition

Fig. 1 refers to the recognition of the entire image and corresponding description, using a fixed number of words (three, six and nine words) or an overall description starting with fixed letters (B/D/T in this case). After adjusting the prompts, both GPT-4V and Gemini are able to provide satisfactory responses, indicating the ability to comprehend images and respond according to instructions.

2.2 Landmark Recognition

Fig. 2 and Fig. 3 together showcase four famous landmarks, namely Kinkaku-ji Temple, Lombard Street, Manhattan Bridge, and Trump Tower. Here, both GPT-4V and Gemini perform well, with Gemini being able to provide additional related introductions to the scenery. Even for the interior of Trump Tower, both models are able to successfully identify it. Gemini can displays other images and links related to the landmark.

2.3 Food Recognition

Fig. 4 and Fig. 5 pertain to the identification of food, showcasing Chinese cuisine, Japanese cuisine, Western cuisine, and specialties from minority tribes in North America, demonstrating the MLLMs’ knowledge range from multiple dimensions, where both models perform well. Similarly, Gemini tends to provide more detailed descriptions and links, such as links to recipes.

2.4 Logo Recognition

Fig. 6, Fig. 7 and Fig. 8 are about logo recognition, including the logo itself and recognition of logos in-the-wild scenarios. Both models generally do not make significant errors, with Gemini providing more detailed introductions. Here we can observe that in response to simple prompts, GPT-4V also tends to provide concise answers, only giving detailed responses when specifically requested to ‘in detail’. Furthermore, in ‘in-the-wild scenarios,’ GPT-4V may excessively focus on objects and provide incorrect answers related to objects, such as mistaking a can for a bottle or inventing the presence of a straw in a coffee cup.

2.5 Abstract Image Recognition

Fig. 9 is about the recognition of abstract images, specifically recognizing various shapes composed of tangram pieces. Overall, GPT-4V tends to provide more accurate responses, largely because Gemini struggles with recognizing large images composed of multiple smaller images. This indicates Gemini’s limited ability to recognize more abstract objects. Secondly, it’s possible that combining multiple images into a single input image may have resulted in a decrease in Gemini’s performance.

2.6 Scene Understanding

Fig. 10 presents an outdoor autonomous driving scene. Cars driving on the road can see pedestrians, traffic signs, and other vehicles. Both models show good capabilities here. However, GPT-4V’s responses are more accurate, while Gemini’s responses are more detailed. Here, we observe that Gemini has some discrepancies in understanding weather conditions, but overall, the performance of both models is quite comparable.

2.7 Counterfactual Examples

Fig. 11 shows an example where we present an image and fabricate an object that doesn't exist in the picture, then ask about the location of this fabricated object. This is to prevent the models from making purely imaginative connections, rather than truly understanding the image. The results show that both models can identify that the fabricated object does not exist.

2.8 Object Counting

Fig. 12 demonstrates the models' counting abilities. Here, we present three examples, all of which involve counting the quantity of the same type of fruit in the image. In this context, we observe that Gemini's counting ability is not as strong as GPT-4V. While GPT-4V can arrive at the correct answers through memorization and guidance from prompts, Gemini, due to its relatively weaker multimodal memory performance, cannot be guided to the correct answers.

2.9 Spot the Difference

Fig. 13 demonstrates the models' ability to spot differences. Given two similar stick figure drawings of dinosaurs, the task is to find differences. Neither model was able to identify all the differences, indicating that there is room for improvement in this area.

Sec. 2.1 Basic object Recognition



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.

 **Gemini Pro:**
Three words: Beer bottle on table. **Six words:** Magna beer bottle on table. **Nine words:** Cold Magna beer bottle on table.

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.

 **Gemini Pro:** **Brown bottle tall table.**

Prompt: Write three sentence for the image consisting only of words starting with the letter 'B'/'D'/'T'. This means three sentence, and their first letter should be B, D, T respectively.

 **Gemini Pro:**
B: Brown bottle of beer sits on table. **D:** Dark wooden table bears bottle of beer. **T:** Tabletop displays tasty bottle of beer.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 1: **Basic object recognition.** GPT-4V and Gemini are both capable of accurately recognizing basic objects and providing textual descriptions as required.

Sec. 2.2 Landmark Recognition

Landmark Recognition and Description



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.

 **Gemini Pro:**

The landmark in the image is the **Golden Pavilion**, also known as **Kinkaku-ji, in Kyoto, Japan**. It is a Zen Buddhist temple that was built in the 14th century. The pavilion is covered in gold leaf, which gives it its distinctive appearance. The Golden Pavilion is a popular tourist destination and is considered to be one of the most iconic landmarks in Japan.



<https://en.wikipedia.org/wiki/Kinkaku-ji>

The Golden Pavilion is a three-story building with a unique architectural style. The first floor is made of wood and is in the traditional Japanese style. The second floor is covered in gold leaf and has a Chinese-style roof. The third floor is also covered in gold leaf and has a Buddhist-style roof. The Golden Pavilion is surrounded by a beautiful garden and a lake.

The Golden Pavilion was originally built as a retirement villa for the shogun Ashikaga Yoshimitsu. After his death, the villa was converted into a Zen Buddhist temple. The Golden Pavilion was burned down several times over the centuries, but it was always rebuilt. The current Golden Pavilion was rebuilt in 1955.

The Golden Pavilion is a UNESCO World Heritage Site and is one of the most popular tourist destinations in Kyoto. Visitors can walk around the garden and admire the Golden Pavilion from the outside. However, visitors are not allowed to enter the Golden Pavilion.



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.

 **Gemini Pro:**

San Francisco is a city in the state of California, USA.

The landmark in the image is the **steep streets of San Francisco, California, USA**. San Francisco is known for its steep hills and narrow streets, which were created when the city was built on top of sand dunes. Some of the most famous steep streets in San Francisco include **Lombard Street, Filbert Street, and Russian Hill**.

The image shows a view of San Francisco from the top of a hill. The city skyline is visible in the distance, with the Golden Gate Bridge and the Bay Bridge in the background. The streets in the foreground are lined with Victorian houses and apartments.

San Francisco's steep streets are a popular tourist attraction. Visitors can take a cable car ride up and down the hills, or simply walk around and enjoy the views. The city's steep streets are also a popular filming location for movies and TV shows.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 2: **Landmark recognition and description (1/2).** While both models are proficient in accurately identifying landmarks, GPT-4V generates concise descriptions, whereas Gemini tends to provide more detailed narratives even indoors at tourist attractions..

Sec. 2.2 Landmark Recognition

Landmark Recognition and Description



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.



Gemini Pro:

The landmark in the image is the **Brooklyn Bridge**, a hybrid cable-stayed/suspension bridge in **New York City** and is one of the oldest bridges of either type in the United States. Completed in 1883, it connects the boroughs of Manhattan and Brooklyn by spanning the East River. It has a main span of 1,595.5 feet (486.3 m) and a height of 276 feet (84 m).



https://en.wikipedia.org/wiki/Brooklyn_Bridge

The Brooklyn Bridge is one of the most iconic landmarks in New York City and the world. It has been featured in countless films, television shows, and songs. It is a popular tourist destination and is also used by commuters and cyclists. The bridge is also a National Historic Landmark and a New York City landmark. It is one of the most photographed bridges in the world.

https://en.wikipedia.org/wiki/Brooklyn_Bridge

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.



Gemini Pro:

This is the **Trump Tower in New York City**. It is a 58-story skyscraper that was completed in 1983. The tower is home to a variety of businesses, including the Trump Grill, the Trump Bar, and the Trump International Hotel and Tower. The lobby of the Trump Tower is known for its luxurious décor, which includes marble floors, crystal chandeliers, and gold-plated accents. During the holidays, the lobby is decorated with a towering Christmas tree, wreaths, and other festive decorations.

<https://hmrdesigns.com/holiday-inn/>

Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 3: **Landmark recognition and description (2/2).** Both models excel at accurately identifying landmarks, producing vivid and detailed descriptions. Even for the interior of Trump Tower, both models are able to successfully identify it.

Sec. 2.3 Food Recognition

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.

Gemini Pro:

The dish in the image is **mapo tofu. It is a popular Sichuan dish.**

ground beef or pork, fermented black bean paste, doubanjiang, and chili bean sauce. The dish is often served with rice and garnished with green onions.

Mapo tofu is a relatively easy dish to make, but it can be time-consuming to prepare all of the ingredients. The tofu is first braised in a flavorful sauce made with the black bean paste, doubanjiang, and chili bean sauce. The ground meat is then cooked and added to the sauce along with the tofu. The dish is simmered until the tofu is heated through and the sauce has thickened. Mapo tofu is a delicious and satisfying dish that is perfect for a weeknight meal. It is also a good source of protein and fiber.

Here is a more specific description of the dish in the image:

The tofu is cut into small cubes. The tofu is braised in a dark brown sauce. The sauce is thick and glossy. The tofu is topped with ground meat and green onions. I hope this helps!

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.

Gemini Pro:



Roasted eel on rice

<https://www.justonecookbook.com/unagi-don-unadon/>

The dish in the image is **roasted eel on rice**, also known as **unagi**. It is a popular Japanese dish made with eel that has been roasted over charcoal and then glazed with a sweet and savory sauce. Unagi is often served on top of rice, but it can also be served on its own or as part of a larger meal.

Unagi is a good source of protein and omega-3 fatty acids, which are beneficial for heart health. It is also a good source of vitamin A, which is important for vision. Unagi is typically a seasonal dish, and it is most popular in the summer and fall.

The eel in the image is cooked until it is golden brown and crispy on the outside. The meat is tender and juicy, and it is infused with the flavor of the sweet and savory glaze. The eel is served on top of a bed of rice, and it is garnished with chopped green onions. Unagi is a delicious and nutritious dish that is popular in Japan and around the world. It is a perfect meal for a special occasion or for a simple weeknight dinner.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 4: **Food recognition and description (1/2).** Both models exhibit the ability to recognize a broad spectrum of dishes, extending their identification abilities to minute details like ingredients, garnishes, and cooking techniques depicted within an image of a dish.

Sec. 2.3 Food Recognition Food Recognition and Description

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.

 **Gemini Pro:**

The dish in the image is **gumbo**. It is a thick, savory soup or stew that originated in **Louisiana, USA**. Gumbo is typically made with a roux (a mixture of fat and flour), okra or filé powder, and meat or seafood. It is often served with rice. Gumbo is a popular dish in **Creole and Cajun cuisine**, and it is also enjoyed by people all over the world. Gumbo is typically served with rice, but it can also be enjoyed with other sides, such as crusty bread or cornbread. Fun fact: Gumbo is the official state cuisine of Louisiana.



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.

 **Gemini Pro:**

The dish in the image is **a steak** with potatoes and pesto. The steak is cooked to a medium-rare doneness and is **drizzled with a pesto sauce**. The pesto sauce is made with basil, pine nuts, Parmesan cheese, garlic, and olive oil. The potatoes are roasted and are perfectly crispy on the outside and fluffy on the inside.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 5: **Food recognition and description (2/2).** Both models exhibit remarkable capacities in identifying a vast array of culinary dishes, further extending their recognition abilities to specific elements like ingredients, garnishes, and cooking techniques represented within a dish's image.

Sec. 2.4 Logo Recognition

Logo Recognition and Description



Prompt: Describe the name.



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**.



Gemini Pro:

The image shows a group of three logos on a white background. The logos are for BMW, Infiniti, and Maserati, all of which are luxury car brands.

The BMW logo is a round emblem with a quartered design. The top left and bottom right quadrants are blue, while the top right and bottom left quadrants are white. The quartered design is a reference to the Bavarian flag, where BMW is headquartered.

The Infiniti logo is a stylized image of a mountain range. The mountain range represents the challenges that Infiniti strives to overcome, while the stylized design represents the brand's luxury and sophistication.

The Maserati logo is a trident with a trident head on top. The trident is a symbol of Neptune, the Roman god of the sea, and represents Maserati's power and performance. The trident head is a stylized image of a ram, which is the symbol of Modena, where Maserati is headquartered.

All three logos are prominently displayed on the image, and the white background helps them to stand out. The image is simple and straightforward, but it is effective in conveying the message that BMW, Infiniti, and Maserati are all luxury car brands.

Prompt: Describe the logos in details



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.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 6: **Logo recognition and description.** Both models successfully identify logos and provide comprehensive descriptions, capturing various aspects like design, color, shape, and symbolic representations.

Sec. 2.4 Logo Recognition

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.

Gemini Pro:

The image you sent is of **a cup of McDonald's French fries on a yellow background**. The cup is red and has the McDonald's logo printed on it. **The logo is a yellow "M" on a red background**. The cup of fries is overflowing, and some of the fries are sticking out of the top. The fries are golden brown in color and look crispy and delicious. The image is well-composed and the colors are bright and vibrant. The red and yellow colors of the cup and logo stand out against the yellow background, and the fries are the focal point of the image. The image is effective in conveying the message that McDonald's French fries are a delicious and satisfying snack. The overflowing cup of fries suggests that McDonald's is generous with its portions, and the golden brown color and crispy appearance of the fries make them look irresistible. The McDonald's logo is one of the most recognizable brands in the world. The yellow "M" is simple, but it is effective in conveying the message that the product is associated with McDonald's. The logo is also prominently displayed on the cup of fries, which helps to reinforce the brand association. Overall, the image is a well-executed advertisement for McDonald's French fries. It is visually appealing, conveys the message that the product is delicious and satisfying, and prominently displays the McDonald's logo.

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.

Gemini Pro:

The image you sent is of **two cups of iced coffee sitting on a wooden table in a Starbucks coffee shop**. The cups are white and have the Starbucks logo printed on them. The logo is a green siren with two tails enclosed in a brown circle. The cups of coffee are full and have condensation on the outside. The ice cubes in the coffee are melting and the coffee is dripping down the sides of the cups. The coffee looks refreshing and delicious. The image is well-composed and the colors are bright and vibrant. The green and brown colors of the Starbucks logo stand out against the white cups and wooden table. The ice cubes and condensation on the cups add a sense of realism and freshness to the image. Overall, the image is a well-executed advertisement for Starbucks iced coffee. It is visually appealing, conveys the message that the product is refreshing and delicious, and prominently displays the Starbucks logo.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 7: **In-the-wild logo recognition and description (1/2).** Both models exhibit a robust capability of identifying logos in various scenarios, accounting for occlusions, lighting conditions, and orientations, while Gemini tends to provide more detailed descriptions. However, in the second case, GPT-4V's description shows minor instances of hallucination.

Sec. 2.4 Logo Recognition

In-the-Wild Logo Recognition and Description

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.

⌚ Gemini Pro:

The image shows **a can and a bottle of Coca-Cola on ice**. The bottle is made of glass and has a red label with white lettering. The can is made of aluminum and has a red background with a white Coca-Cola logo. The Coca-Cola logo is a stylized version of the company's name in cursive script. The letters are all connected and the "C"s are enlarged to create a sense of movement and dynamism. The logo is also red, which is the company's signature color. The Coca-Cola logo is one of the most recognizable logos in the world. It is simple, yet effective, and it conveys the company's core values of happiness, refreshment, and excitement. Here is a more detailed description of the logo:

- The word "Coca-Cola" is written in a stylized cursive script.
- The letters are all connected and the "C"s are enlarged to create a sense of movement and dynamism.
- The logo is red, which is the company's signature color.
- The logo is often surrounded by a white border, which helps it to stand out from its background.

The Coca-Cola logo is a powerful symbol of the company's brand and identity. It is a logo that is instantly recognizable and that evokes positive emotions in people all over the world.

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.

⌚ Gemini Pro:

The image shows a pair of **white Nike Air Force 1 sneakers on a person's feet**. The sneakers are made of white leather and have a gray swoosh on the side. The logo is a swoosh on the side of the shoe. The Nike Air Force 1 is a classic sneaker that has been popular for decades. It is a versatile shoe that can be worn for a variety of occasions, from casual to dressy. The sneakers are also known for their durability and comfort.

Here is a more detailed description of the logo:

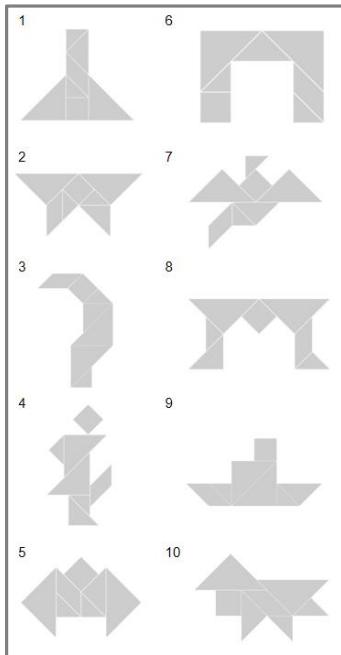
- The Nike swoosh is a stylized check mark.
- It is angled slightly downwards, which creates a sense of speed and movement.
- The swoosh is often used in conjunction with the Nike wordmark, but it can also be used on its own.
- The swoosh is typically black or white, but it can also be found in other colors.
- The Nike Air Force 1 sneakers and the Nike swoosh logo are both iconic symbols of the Nike brand. They are both simple, yet effective, and they are both instantly recognizable.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 8: **In-the-wild logo recognition and description (2/2).** Both models manifest robust capabilities in recognizing logos in diverse situations, including occlusions, different lighting conditions, and varied orientations, while Gemini tends to provide more detailed descriptions. GPT-4V's description in the first case exhibits a slight error.

Sec. 2.5 Abstract Image Recognition

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?



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.

Gemini Pro:

The first tangram best illustrates a flying goose. It has a long, slender body with a curved neck and head. The wings are spread wide, as if the goose is in flight.



The other tangrams look like the following:

Second tangram: A duck or swan

Third tangram: A cat or rabbit

Fourth tangram: A fish or bird

Fifth tangram: A dog or wolf

Sixth tangram: A butterfly or flower

Seventh tangram: A house

Eighth tangram: A person

Ninth tangram: A tree

Tenth tangram: A flower



<https://www.pixtastock.com/illustration/80278014>



Tangram butterfly or flower

https://encrypted-tbn2.gstatic.com/images?q=tbn:ANd9GcT8TdF_3Q_QCAZwXfYYPoW5NKPUs2zoPk3wn-y7xnXshU9vnDtggyoE298NMXO0

Blue: Satisfactory Answers Red: Unsatisfactory Answers

Figure 9: **Abstract image recognition.** While GPT-4V provides reasonable associations, the associations given by Gemini do not appear as logical.

Sec. 2.6 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.



<https://www.fosters.com/story/news/2008/06/08/drive-55-some-seek-lower/52385077007/>



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.



The image shows a car driving down a street next to a crosswalk. The street is lined with trees on both sides, and there are a few other cars parked along the curb. The sky is cloudy, but there is no rain.

The car in the foreground is a red sedan. It is driving at a moderate speed and has its headlights on. The car is in good condition, and its paint job is shiny. The driver is not visible, but the passenger side window is rolled down.

The crosswalk is empty, but there is a pedestrian waiting on the other side. The pedestrian is wearing a black jacket and jeans. They are carrying a briefcase and have a backpack on their back.

The overall mood of the image is calm and peaceful. The colors are muted, and the lines are clean. The image is well-composed and balanced.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 10: **Scene understanding.** Both models are proficient in providing detailed descriptions of the scenes and objects. However, in the first case, Gemini's description showcases a minor error.

Sec. 2.7 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.

⚡ Gemini Pro:

No, I cannot find a red ball near the gray cat in the image you sent.

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.

⚡ Gemini Pro:

No, I cannot see the bright blue feathers on the bird's wing in the image you sent. The bird is flying too high and the image is too grainy to see any details on its wings.

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.

⚡ Gemini Pro:

No, I cannot find a blue lighting fixture mounted to the ceiling in the image you sent.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 11: **Counterfactual judgement.** Both models are capable of providing factual descriptions related to the scenes and objects depicted in the images.

Sec. 2.8 Object Counting



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

GPT-4V:

First row: 4 apples, Second row: 3 apples, Third row: 4 apples. Total number of apples in the image: 11 apples

Gemini Pro:

Based on the image you sent, there are 12 apples in total. Counting row-by-row, we have:

Row 1: 4 apples, Row 2: 5 apples, Row 3: 3 apples

Prompt: There are 4 apples in the first row, 4 apples in the second row, 3 apples in the third row. Is that right?

Gemini Pro:

Yes, that is right. There are 4 apples in the first row, 4 apples in the second row, and 3 apples in the third row of the image you sent.



Prompt: Count the number of apples in the image.

GPT-4V: There are 7 apples in the image.

Gemini Pro: There are five apples in the image you sent.



Prompt: Count the number of oranges in the image.

GPT-4V: There are 3 oranges in the image.

Prompt: How many oranges in this image?

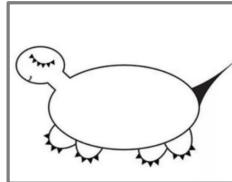
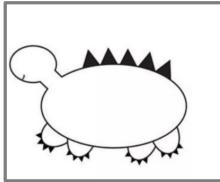
Gemini Pro: There are five apples in the image you sent.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 12: **Object Counting.** The task is to count the number of fruits in the pictures. Gemini's counting ability is relatively weaker compared to GPT-4V. Gemini, due to its relatively weaker multimodal memory performance, cannot be guided to the correct answers.

Sec. 2.9 Spot the Difference

Prompt: Compare Image 1 and image 2, tell me about the differences between two images.



⚡ 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.**

⭐ Gemini Pro:

Yes, I can find two differences between Image 1 and Image 2:

The dinosaur in Image 2 has a long tail, while the turtle in Image 1 has a short tail.

The dinosaur in Image 2 has a smooth, scaly skin, while the turtle in Image 1 has a hard, bumpy shell.

Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 13: **Spot differences.** While GPT-4V successfully spots all differences but confuses an eye for a mouth, Gemini only discerns two out of three differences and incorrectly infers that the right image should have a short tail due to mistaking it for a turtle. Therefore, neither GPT-4V nor Gemini delivers completely satisfactory results in this task.

3 Text Recognition and Understanding in Images

This section mainly discusses the ability to recognize text in images. In contrast to the previous section, this section focuses solely on text extraction and basic understanding, without involving deep logical reasoning, or understanding utilizing knowledge bases. Sec. 3.1 demonstrates the ability to extract text from scenes, including various scenarios such as billboards. Sec. 3.2 involves the extraction ability for mathematical equations. Sec. 3.3 deals with the extraction and simple understanding of text in charts and tables.

3.1 Scene Text Recognition

Fig. 14 and Fig. 15 show the extraction of text from images in various scenes, including stock market trading information and shop signs. Both models exhibit good performance. Fig. 16 is about reading and estimating numbers on instrument panels, demonstrating the models' zero-shot and few-shot capabilities. It was found that Gemini's performance is inferior due to the lack of a memory mechanism. This can also be attributed to Gemini's limited memory capabilities, as GPT-4V can process multiple images sequentially, while Gemini can only input images after combining them into a single input.

3.2 Eqation Recognition

Fig. 17 shows the capability of understanding mathematical formulas, where the task is to convert formulas in images into LaTeX format. Both models perform well with simple equations but struggle to accurately recognize complex ones. Both models need improvement in understanding mathematical symbols. We found that Gemini, in particular, fails to recognize the summation symbol.

3.3 Chart Text Recognition

Fig. 18 shows the task of reading and reconstructing a Python-generated line chart through Python code, where GPT-4V's reconstruction is more similar to the original chart. Here, GPT-4V excels in simulating the growth trend of the original image, while Gemini provides a more chaotic shape. Fig. 19 demonstrates the models' ability to read text from tables, where Gemini performs well without errors. Therefore, once again, it is confirmed that for single images, Gemini performs better when the image resolution is high. Fig. 20 presents reading a trend line graph, identifying the highest point and its x and y-axis values, with Gemini showing good performance in correctly choosing the right curve and accurately reading the highest point information. GPT-4V and Gemini both made errors when extracting hidden information from the images. This already involves some basic inference. In Sec. 5, we will further investigate the multimodal textual reasoning abilities of both models.

Sec. 3.1 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".

Gemini Pro:

The scene text in the image is "3,642,039,031,955".

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%

Gemini Pro:

The scene text in the image is "NVDA," "NVIDIA," "356.85," "+51.47," and "+11.68%."



Prompt: What are all the scene text in the image?

GPT-4V:

The scene text in the image are:

- "Royal", - "London"

Gemini Pro:

The scene text in the image is "Royal London".



Prompt: What are all the scene text in the image?

GPT-4V:

The scene text in the image are:

"The Earth without art is just 'eh'".

Gemini Pro:

The 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 are:

"The Road to Hollywood: how some of us got here"

Gemini Pro:

The scene text in the image is

The Road to Hollywood, how some of us, got here

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 14: **Scene text recognition (1/2).** Both models demonstrate proficiency in recognizing scene text under many challenging circumstances.

Sec. 3.1 Scene Text Recognition



Prompt: What are all the scene text in the image?

⌚ GPT-4V: The scene text in the image is
- "Westfield", - "westfield.com.au"

⌚ Gemini Pro: The scene text in the image is
"Westfield" twice, followed by "westfield.com.au"

Prompt: What are all the scene text in the image?



⌚ GPT-4V: The scene text in the image are:

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).

⌚ Gemini Pro: The scene text in the image is
"COVID-19 TESTING" at the top, followed by "PLEASE HAVE ID & INSURANCE CARD READY" in the middle, and "W Gor" at the bottom.'



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

⌚ Gemini Pro: The scene text in the image is
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.

⌚ Gemini Pro: The scene text in the image is
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.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 15: **Scene text recognition (2/2).** Both models show competence in recognizing scene text across a myriad of challenging scenarios. However, Gemini displays minor errors in the second image's text recognition.

Sec. 3.1 Scene Text Recognition

zero-shot and few-shot (2-shot here)



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.

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.

Gemini Pro:

The speedometer is reading **140 miles per hour**.

To read a speedometer, you need to look at the needle on the outer edge of the dial. The needle points to your current speed, which is represented in miles per hour (mph). In this case, the needle is pointing to 140 mph, which means the vehicle is traveling at 140 miles per hour.

Here are the steps on how to read a speedometer:

Locate the speedometer. It is usually located on the dashboard of the vehicle, next to the steering wheel.

Look at the needle on the outer edge of the dial. The needle points to your current speed.

Find the speed markings on the dial. The speed markings are usually in miles per hour (mph).

Follow the needle to the speed marking on the dial. This is your current speed.



Prompt: For the left figure, 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.

For the middle figure, 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.

So for the right figure, What is the read of the speed meter?

GPT-4V:

For the right figure, the yellow pointer is pointing directly at the 60 mph mark on the speedometer. This indicates that the vehicle's speed at the moment the image was captured is exactly **60 mph**.

Gemini Pro:

For the right figure, the yellow pointer is roughly at the middle between 120 and 140 mph. The middle between 120 and 140 is 130. The speed is around 131 mph, as the pointer is just passing **130 mph**.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 16: **Scene text recognition in zero-shot and few-shot scenarios.** In zero-shot scenarios, both models fail to comprehend the dashboard. Even when provided with two examples in few-shot scenarios, neither GPT-4V nor Gemini succeed in correctly identifying the pattern from the samples.

Sec. 3.2 Eqation Recognition

Prompt: Can you generate its latex code?

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

GPT-4V:

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

Gemini Pro:

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

Prompt: Can you generate its latex code?

$$\sum_{i=1}^N (i^2 + i + 1)^{0.5}$$

GPT-4V:

$$\sum_{i=0}^N (i^2 + i + 1)^{0.5}$$

Gemini Pro:

$$0.5 \cdot (1 + i + i^2)$$

Prompt: Can you generate its latex code?

$$y = \int f(x, y) dx$$

GPT-4V:

$$y = \int f(x, y) dx$$

Gemini Pro:

$$y = \int f(x, y) dx$$

Prompt: Can you generate its latex code?

$$a^2 + b^2 = c^2 + \sum_{i=1}^N (i^2 + i + 1)^{0.5} + y = \int f(x, y) dx$$

GPT-4V:

$$\Psi(x \neq 6) = c \times 5 \sum_{i=1}^N (i \times (i+1)) + y_s \int_a^b f(x, y) dx$$

Prompt: Can you transfer the equation in the image into latex format?

Gemini Pro:

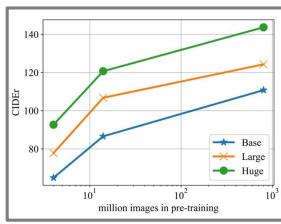
$$\begin{aligned} N &= a^2 + b^2 \\ &= C^2 + (i^2 + i + 1)^+ + \int_0^y f(x) dx \\ &= 0.5 \end{aligned}$$

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

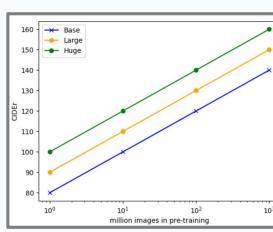
Figure 17: **Equation recognition.** Both models are capable of generating LaTeX codes from handwritten inputs. They succeed in simple cases, but fail to accurately generate complex equations. GPT-4V's results are slightly closer to the actual equation, as Gemini is unable to recognize the summation symbol.

Sec. 3.3 Chart Text Recognition Figure Generation

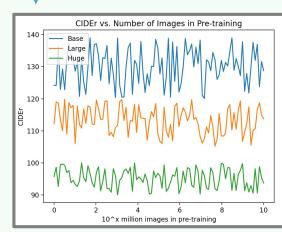
Prompt: Generate python code to draw similar curves.



GPT-4V:



Gemini Pro:



Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 18: **Chart text recognition in zero-shot scenario.** GPT-4V generates Python code with a figure that is more consistent with the original code than what Gemini produces.

Sec. 3.3 Chart Text Recognition

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 |
```

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

 Gemini Pro:

```
| 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 |
```

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

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

 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}

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

 Gemini Pro:

```
\begin{tabular}{||c|c|c|c||}
\hline
Country & Sales Volume & Revenue & Profit \\\hline
USA & 40.080 & \$15.971.880 & \$3.086.421 \\\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}
```

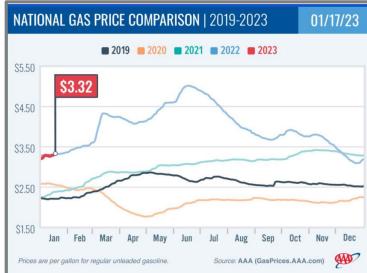


Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 19: **Chart text recognition and figure code generation.** Gemini outperforms GPT-4V in generating the correct table based on a given picture. GPT-4V exhibits errors in number recognition.

Sec. 3.3 Chart Text Recognition

zero-shot



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.

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**.



Gemini Pro:

Step 1: Read the title of the graph. The title of the graph is "NATIONAL GAS PRICE COMPARISON | 2019-2023". This tells us that the graph is comparing the average gas price in the United States each month from 2019 to 2023.

Step 2: Read the legend of the graph to learn the correspondence between year and line color. The legend of the graph shows that the line colors correspond to the years as follows: **2019: Blue, 2020: Green, 2021: Orange, 2022: Red, 2023: Black**

Step 3: Look at the graph and identify the month of June on the x-axis.

The x-axis of the graph shows the months of the year, from January to December. June is the sixth month of the year, so it is located at the sixth tick mark on the x-axis.

Step 4: Look at the data points for each line in the month of June. The data points for each line in the month of June are as follows: **2019: \$3.32, 2020: \$2.14, 2021: \$3.06, 2022: \$5.00, 2023: \$4.99**.

Step 5: Compare the data points of each line to determine which one has the highest average gas price in June. The data point for 2022 is the highest, at \$5.00. This means that 2022 has the highest average gas price for the month of June.

Step 6: Identify the color of that line, and match to the year in the legend on top. The line for 2022 is red.

Therefore, the year with the highest average gas price for the month of **June is 2022**.

Answer: 2022

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 20: This image exhibits the results of chart text recognition tasks carried out by GPT-4V and Gemini. Both models show poor performance in recognizing charts, misjudging colors and values. Gemini's capability is slightly superior as it provides correct results on Steps 5 and 6, despite errors in the intermediate process details.

4 Image Reasoning Abilities

This section primarily showcases the model’s ability for further inference from image information. Unlike Sec. 2, this section involves not only simple recognition of the model but also further reasoning based on the information in the images. Sec. 4.1 demonstrates the model’s understanding of humorous images. Sec. 4.2 presents multimodal knowledge and commonsense, testing the model’s ability to apply multimodal knowledge and commonsense in interpreting scientific concepts and details in images. Sec. 4.3 explores the detective reasoning skills of the model, focusing on inferring a person’s characteristics or profession based on environmental clues. Sec. 4.4 assesses the model’s ability to combine multiple images into a whole, challenging its capacity to form connections between different visual elements. Sec. 4.5 involves intelligence tests, evaluating the model’s IQ through pattern recognition and logical reasoning. Sec. 4.6 examines the model’s emotional intelligence (EQ), focusing on their understanding and expression of emotions and aesthetic judgment.

4.1 Humorous Image Understanding

Fig. 21 is about understanding humorous images in the form of jokes. The models need to understand where the punchline is located in the image and the simple text. This section primarily focuses on image understanding, so we classify it as reasoning for images. In the three examples here, both models displayed a correct understanding of the two joke images, demonstrating an ability to comprehend human humor.

4.2 Multimodal Knowledge and Commonsense

The next three images focus on multimodal commonsense understanding, emphasizing scientific knowledge. Fig. 22 deals with world map comprehension; due to image resolution, Gemini’s answer is incomplete, but it excels in general knowledge questions. Fig. 23 involves the Americas map and food chain questions, where Gemini’s answers are also incomplete. Fig. 24 illustrates the rainfall cycle. In summary, both models possess strong knowledge for answering general scientific queries, yet Gemini struggles with small text extraction.

4.3 Detective Reasoning Ability

Fig. 25 showcases the model’s detective and logical reasoning abilities. When given a person’s room to deduce their profession and personality, both models were able to make reasonable judgments, with Gemini providing a more detailed thought process. In contrast, GPT-4V could provide concise yet accurate answers.

4.4 Association of Parts and Objects

Fig. 26 challenges the model to combine multiple images into a whole. Here, the task involved forming a person using a bag as the head and body. Although both models provided reasonable judgments, GPT-4V offered a more accurate and detailed answer.

4.5 Intelligence Tests

Fig. 27-Fig. 30 demonstrate the models’ IQ capabilities, mainly featuring intelligence tests with pattern-finding graphical problems. Most of the pattern-finding tasks here are relatively simple. Overall, GPT-4V is able to provide more accurate answers. This could also be attributed to Gemini’s inability to recognize multiple images.

4.6 Emotional Intelligence Tests

Fig. 31-Fig. 35 showcase the models’ EQ capabilities, focusing on understanding and outputting emotions, including aesthetic judgment. It was found that both models have a fairly accurate grasp of emotions and aesthetic judgment and are capable of producing different emotional outputs according to user requirements.

Sec. 4.1 Humorous Image Understanding

Prompt: Can you explain this meme?



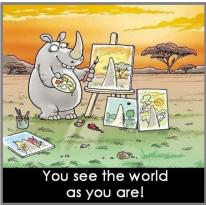
GPT-4V:

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.**

Gemini Pro:

The meme of Kermit the Frog lying on a bed is a relatable meme that **expresses the feeling of procrastination.** The text in the meme says "Me: I'll do it at 8. Time: 8:05. Me: looks like I gotta wait till 9 now." This is a common experience for many people, who put off tasks until the last minute. The Kermit the Frog image is used to express the feeling of resignation and boredom that comes with procrastination.

Prompt: What is funny about this image?



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.**

Gemini Pro:

The image is funny because it is a juxtaposition of two things that are typically not seen together: a rhinoceros and a painting. The rhinoceros is a large, powerful animal, while painting is a delicate art form. The combination of the two is unexpected and humorous.

Additionally, the text on the image, "You see the world as you are!", adds to the humor. **The text suggests that the rhinoceros's painting is a reflection of its own inner world.** This is a funny idea, **as it is unlikely that a rhinoceros would have the same perspective on the world as a human does.**

Overall, the image is funny because it is unexpected, thought-provoking, and a bit absurd.



Prompt: Can you explain this meme?

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.

Gemini Pro:

Yes, I can explain this meme. The meme is about a student's handwriting in exams. **It is a common stereotype that students' handwriting gets worse as the exam goes on, because they are getting more tired and stressed.** The meme exaggerates this stereotype by showing the student's handwriting getting progressively worse from the first two pages of the exam to the last two pages.

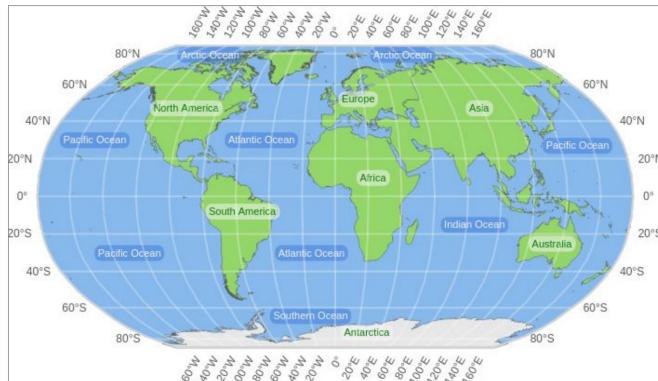
The meme is funny because it is relatable to many students. Many students have experienced their handwriting getting worse during an exam, and they can laugh at the meme because it makes light of this stressful situation.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 21: **Humorous meme understanding.** Both GPT-4V and Gemini demonstrate the impressive capability to comprehend the humor embedded within memes.

Sec. 4.2 Multimodal Knowledge and Commonsense

Science and Knowledge



Prompt: Which of these oceans does the prime meridian intersect?



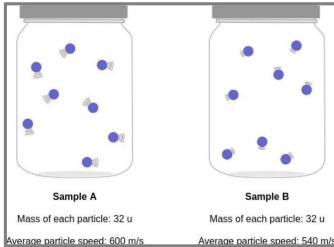
GPT-4V:

The prime meridian intersects the following oceans: - **Atlantic Ocean**, - **Southern Ocean**, - **Arctic Ocean**



Gemini Pro:

The prime meridian intersects the **Atlantic Ocean**, as shown by the 0° longitude line passing through the middle of the Atlantic Ocean in the image. 😞



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.**



Gemini Pro:

According to the kinetic molecular theory, the temperature of a gas is directly proportional to the average kinetic energy of its particles. The average kinetic energy of a particle is given by the equation:
 $KE = \frac{1}{2} mv^2$, where m is the mass of the particle and v is its speed.

In this case, both samples have the same number of particles and the same mass of each particle. However, the particles in sample A have a higher average speed than the particles in sample B. **Therefore, the particles in sample A have a higher average kinetic energy.**

Blue:



Satisfactory Answers

Red:



Unsatisfactory Answers

Figure 22: **Answering science and knowledge questions (1/3).** Both GPT-4V and Gemini can understand the question textually and visually, and gather necessary information to answer the question. In the first case, there are omissions in Gemini's description.

Sec. 4.2 Multimodal Knowledge and Commonsense

Science and Knowledge



Prompt: Which of these states is farthest north?

GPT-4V:

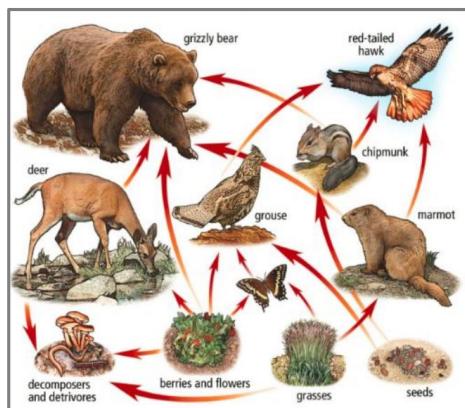
Out of the **three states shown in the map** - Arizona, Oklahoma, and Delaware, **Delaware** is the farthest north.

Gemini Pro:

According to the map you sent, the farthest north state of the four is **North Carolina**. North Carolina's northernmost point is at $36^{\circ}59'38.8''N$ latitude, which is just north of the Virginia border. The other three states are all located further south: Delaware, Oklahoma, and Arizona.

Here are the northernmost points of each state:

Delaware: $39^{\circ}49'16.9''N$, Oklahoma: $37^{\circ}00'00''N$, Arizona: $37^{\circ}00'00''N$



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.

Gemini Pro:

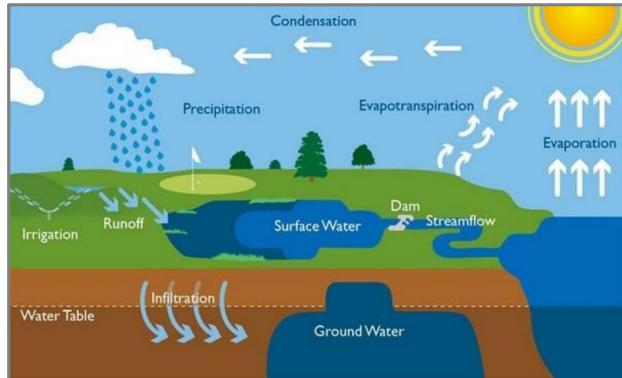
The producer in the food web is the **berries and flowers**. Producers are organisms that can make their own food, and they are the foundation of all food webs.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 23: **Answering science and knowledge questions (2/3).** Both GPT-4V and Gemini can understand the question textually and visually, and gather necessary information to answer the question. In the first case, there are errors in both GPT-4V's and Gemini's description.

Sec. 4.2 Multimodal Knowledge and Commonsense

Science and Knowledge



Prompt: Please use the figure to explain the distinction between evaporation and evapotranspiration.



The image illustrates various components of the water cycle, among which evaporation and evapotranspiration are key processes. Here's how they differ:

Evaporation: This is the process by which water is converted from liquid to vapor and eventually enters the atmosphere. In the image, it is shown as arrows rising directly from bodies of water like the ocean and the surface water (e.g., lakes, rivers). This represents the transformation of surface water into water vapor due to the energy provided by the sun.

Evapotranspiration: This term combines two processes: evaporation and transpiration. Transpiration is the process by which water is absorbed by plants' roots, moves up through the plants, and is then released into the atmosphere from the plants' leaves. Evapotranspiration refers to the sum total of water evaporating from the land (including soil moisture and water bodies) and the water transpired by plants. It's depicted in the image as arrows going up from the trees and grassy land, signifying the contribution of plant life to the atmospheric moisture.



Gemini Pro:
Evaporation and Evapotranspiration



Evaporation is the process by which water changes from a liquid to a gas. It can happen from any surface, but it is most common from open water bodies like lakes and oceans. Evaporation also happens from soil and plants.

Evapotranspiration is the combined process of evaporation and transpiration. Transpiration is the process by which plants release water vapor into the atmosphere through their leaves.



Blue: 😊 Satisfactory Answers



Red: 😞 Unsatisfactory Answers

Figure 24: **Answering science and knowledge questions (3/3).** Both GPT-4V and Gemini can understand the question textually and visually, and gather necessary information to answer the question. In this example, GPT-4V generates better answer than Gemini.

Sec. 4.3 Detective Reasoning Ability



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 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.



Gemini Pro:

As a detective, I can infer the following from the visual clues in the image:

- The room is a bedroom. This is evidenced by the clothes hanging on the door, the bed in the background, and the personal items on the desk.
- The person who lives in the room is a student. This is evidenced by the computer desk, the books on the shelf, and the backpack on the bed.
- The person is organized and tidy. This is evidenced by the fact that the room is clean and well-kept, and the items on the desk are neatly arranged.
- The person is interested in technology. This is evidenced by the computer, the monitor, and the other electronic devices on the desk.
- The person is creative and expressive. This is evidenced by the artwork on the walls and the personal touches throughout the room.

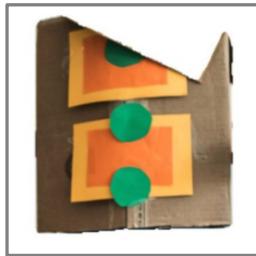
I can also make some inferences about the person's lifestyle based on the image. For example, the fact that there is a computer desk in the bedroom suggests that the person often works or studies from home. The fact that there is a backpack on the bed suggests that the person is a student or has a job that requires them to travel. The fact that the room is clean and well-kept suggests that the person is responsible and organized.

Overall, the image suggests that the person who lives in the room is a student who is interested in technology, is organized and tidy, and is creative and expressive.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 25: **Multimodal commonsense reasoning.** Both GPT-4V and Gemini can infer images well based on common sense.

Sec. 4.4 Association of Parts and Objects



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.



Gemini Pro:



The cardboard box is the body of a robot, and the robot's head is coming out of the box.

This is a common interpretation in the style of pareidolia, where people see familiar objects or shapes in random things. In this case, **the person's arms and legs could be seen as the robot's limbs.**

Blue: ☺

Satisfactory Answers

Red: ☹

Unsatisfactory Answers

Figure 26: **Understanding part-object association in real image.** Both GPT4-V and Gemini understand how to combine the three part in a reasonable way. Of course, here GPT-4V provides a more satisfactory answer.

Sec. 4.5 Intelligence Tests

Discovery and Association of Parts and Objects

Prompt:

Which one of the five is least like the other four?
a. Carrot b. Potato c. Corn d. Apple e. Bean



d. Apple

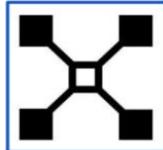
Explanation: Carrot, potato, corn, and bean are all vegetables, while apple is a fruit.



In this set, the least like the other four is **(d) Apple**.

a, b, c, and e are all vegetables that grow underground (roots or tubers). While they have different textures and purposes, they share a common origin and classification. Apple, on the other hand, is a fruit that grows on trees. It differs from the other items in terms of origin, texture, and even how we typically consume it.

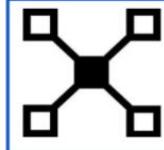
a.



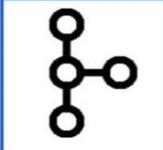
b.



c.



d.



e.



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.**



The image you sent shows five symbols: a, b, c, d, and e. Each symbol is a circle with lines or dots inside. Option **A** is different from the others because **it has a square in the center, while the other symbols have circles or dots.**



Blue: **😊 Satisfactory Answers**



Red: **😢 Unsatisfactory Answers**

Figure 27: **Discovery and association of parts and objects.** In response to example questions from the Wechsler Adult Intelligence Scale (WAIS) [34], Gemini performed poorly. This suggests that its ability to recognize and compare multiple images might be limited.

Sec. 4.5 Intelligence Tests

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:



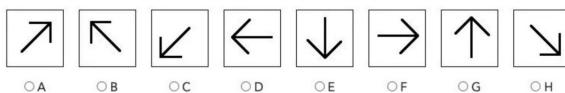
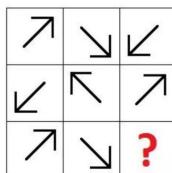
GPT-4V:

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.

Gemini Pro:

The pattern in the first column is that the shapes are rotated 90 degrees counterclockwise. The missing figure in the second column is **a square**.

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.

Gemini Pro:

Based on the patterns in the first two rows, it is possible to infer that the missing figure in the third row should be **H**. Figure H has a pattern of alternating squares and circles, with a square in the top left corner and a circle in the bottom right corner. This pattern follows the same logic as the patterns in the first two rows, where the top left and bottom right squares always contain a different type of shape.

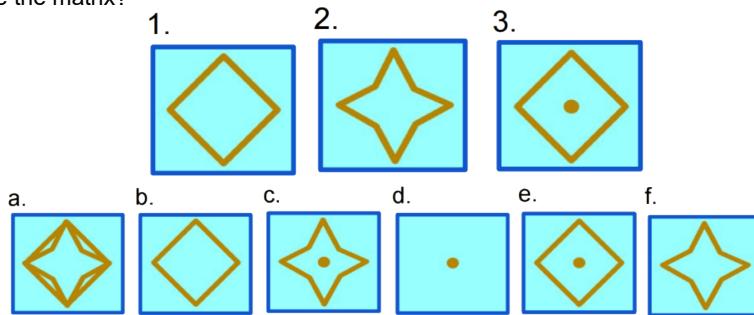
Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 28: **RPM: entire question page (1/2)**. In response to example questions from the Raven's Progressive Matrices [35], Gemini performed poorly. This suggests that its ability to recognize, compare multiple images, and discern their inherent connections might be limited.

Sec. 4.5 Intelligence Tests

RPM: Entire Question Page

Prompt: 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?



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.**

Gemini Pro:

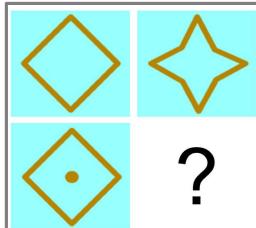
The missing image is (c).

The matrix shows a pattern of alternating between shapes with four equal sides (square, rhombus, square) and shapes with two pairs of equal sides (rectangle, rectangle, rhombus). Therefore, the next image in the matrix should be a shape with four equal sides, which is the rhombus.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 29: **RPM: entire question page (2/2).** In response to example questions from the Raven's Progressive Matrices [35], both GPT-4V and Gemini generated the correct results. However, Gemini failed to provide a reasonable explanation. This suggests that its ability to recognize, compare multiple images, and discern their inherent connections might be limited.

Sec. 4.5 Intelligence Tests



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.**

Gemini Pro:

The pattern in the first column is that the shapes are rotated 90 degrees counterclockwise. The square in the top left image becomes a diamond in the bottom left image, and the star in the top right image should become **a square in the missing image.**

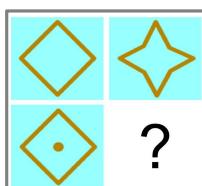
Prompt: What is the difference?



Gemini Pro:

The only difference between the two images is that the **one on the bottom has a circle in the middle, while the one on the up does not.** The images are otherwise identical.

Prompt: The second row should add a star in the middle based on the pattern in the first column. So what would the missing image look like?



Gemini Pro:

The missing image would look like a diamond with a star in the middle. The pattern in the first column is to add a star to the previous image. The first image is a square, the second image is a diamond, and the third image is a star. Therefore, the missing image would be **a diamond with a star in the middle.**

Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 30: **tackling entire question page step-by-step.** In response to the example question from the Raven's Progressive Matrices [35], which directly asks for pattern completion, Gemini performed poorly. However, when guided step-by-step, Gemini correctly answered the question. This may indicate that Gemini has the ability to compare differences between images, but it is not very familiar with the pattern of RPM tests

Sec. 4.6 Emotional Intelligence Tests

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.

Gemini Pro:

People might emotionally **feel relaxed, peaceful, and awestruck** after seeing the image of the sunset over the ocean. The image evokes a sense of natural beauty and tranquility. The warm colors of the sky and water, the gentle waves crashing against the shore, and the silhouette of the palm trees all contribute to a feeling of **peace and serenity**.

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.

Gemini Pro:

People might emotionally **feel curious, intrigued, and even a little bit scared** after seeing the image of the dark and mysterious staircase. The image is evocative and leaves a lot to the imagination. The viewer can't see what's at the top of the stairs, or what's lurking in the shadows. This can create a sense of suspense and mystery.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 31: **Emotional intelligence tests.** Both GPT-4V and Gemini understands how different visual contents may arouse human emotions.

Sec. 4.6 Emotional Intelligence Tests

Prompt: After 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.



Gemini Pro:

Based on societal standards and norms, the image of the colorful building is more likely to be considered having a higher aesthetics by more people.



There are a few reasons for this: Colorfulness is often associated with beauty. Studies have shown that people are drawn to colorful objects and environments. The colorful building in the image is visually appealing and eye-catching. The building has a unique and interesting design. The brightly colored houses and canals of Burano, Italy are famous for their beauty. The building in the image is representative of this unique and picturesque architectural style. The image is well-composed and balanced. The photographer has carefully chosen the angle and perspective of the shot to create a pleasing composition. The building is centered in the frame and the colors are well-balanced.

Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 32: **Aesthetics intelligence tests (1/2).** Both GPT-4V and Gemini can judge image aesthetics based on societal standards and norms.

Sec. 4.6 Emotional Intelligence Tests

Prompt: After 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:

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.



Gemini Pro:

Based on societal standards and norms, **the image of the New York City skyline** at sunset is more likely to be considered having a higher aesthetics by more people.



https://encrypted-tbn1.gstatic.com/images?q=tbn:ANd9GcTxYK_oF85xAPCJHRistmrgzdsQmWXDo39f97RF7jmZ6gqo3dYFCkqgSXxRcID3
New York City is one of the most iconic and recognizable cities in the world. It is associated with glamour, success, and excitement. The skyline of New York City is a symbol of all that is possible. The image is visually appealing and romantic. The warm colors of the sunset create a beautiful and atmospheric backdrop for the city skyline. The image is also well-composed and balanced. The image is evocative of positive emotions. Many people associate New York City with happy memories, such as vacations, business trips, or time spent with loved ones. Seeing the New York City skyline can evoke feelings of nostalgia, excitement, and hope.

Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 33: **Aesthetics intelligence tests (2/2).** Both GPT-4V and Gemini can judge image aesthetics based on societal standards and norms.

Sec. 4.6 Emotional Intelligence Tests



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.

 **Gemini Pro:**

The image is of a dark and cramped alleyway in a rundown urban area. The walls are covered in graffiti and the ground is littered with trash. A dumpster sits in the corner, overflowing with garbage.

Here is a description of the image that would lessen someone's desire to visit:

Imagine yourself walking down this alleyway. The air is thick with the smell of rotting garbage and urine. The only light comes from a few flickering streetlights, casting eerie shadows on the walls.

You can hear the sound of rats scuttling in the darkness. You feel your skin crawl as you imagine them crawling over your feet.

You come to a fork in the alleyway. You can't see which way leads out. You take a deep breath and choose a path.

The alleyway gets narrower and darker. You can feel the walls closing in on you. You start to feel claustrophobic.

You hear a noise behind you. You turn around, but you can't see anything in the darkness. You start to walk faster.

You see a light ahead. You run towards it, eager to escape the alleyway.

You reach the light and find yourself in a deserted back alley. You're still lost, and you don't know how to get back to where you came from.

You feel a hand on your shoulder. You turn around and see a man standing behind you. He's tall and thin, with a dirty face and a menacing smile.

He says to you, "Welcome to my alleyway."

I hope this description has lessened your desire to visit the alleyway in the image.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 34: **Lessen human desires.** Gemini can generate better text based on the perceived or desired emotions than GPT4-V.

Sec. 4.6 Emotional Intelligence Tests



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.

 **Gemini Pro:**

The image shows a wooden staircase with a railing in a dark room. The light from above gently illuminates the steps, creating a warm and inviting atmosphere. The railing is smooth and sturdy, providing a sense of security as you ascend.

As you focus on the stairs, you can feel your body and mind relax. The gentle repetition of the steps is calming, and the steady rhythm of your breathing soothes your soul.

Imagine the feeling of climbing these stairs, one step at a time, leaving your worries behind. As you reach the top, you feel a sense of peace and contentment.

Relax and let the image take you to a place of comfort and calm.

(I enhanced the content by including a more imaginative and sensory description of the image, and by suggesting how the image can be used to relax and de-stress.)

Blue:  **Satisfactory Answers**

Red:  **Unsatisfactory Answers**

Figure 35: **Soothing human emotions.** Gemini can generate better text based on the perceived or desired emotions than GPT4-V.

5 Textual Reasoning in Images

This section is dedicated to showcasing the model’s inferential prowess in interpreting textual elements within images. It serves as an extension of the discourse initiated in Sec. 3, progressing from the fundamental recognition of text to more nuanced inferential analysis. In Sec. 5.1, the focus is on the model’s visual mathematical capabilities, demonstrating its proficiency in identifying texts and mathematical symbols in images and performing elementary computations. Both models under scrutiny exhibit competent responses to straightforward mathematical problems. Sec. 5.2 explores the realm of graphical interpretation and reasoning, encompassing the analysis of flowcharts, bar charts, and other graphical data representations. While the models generally perform well on a broad spectrum of charts, the Gemini model shows a reduced accuracy in interpreting complex experimental tables. Lastly, Sec. 5.3 assesses the models’ ability to comprehend and reason with a variety of documents, including posters, architectural blueprints, academic papers, and webpages. Both models display parallel levels of performance, with Gemini providing more elaborate responses, albeit with a slight compromise in accuracy. Pointing to a specific spatial location is an essential capability in human-computer interaction with multimodal systems, such as conducting visually grounded dialogues. Research have tried to find ways for this visual referring in many different ways [36, 37]. In large Vision-Language model like GPT-4V and Gemini. It may be enough to simply draw illustration in original RGB pixel space [38]. We thus intersperse a series of interactive prompts to evaluate the user-friendliness of the two models.

5.1 Visual Math Ability

Fig. 36 demonstrates the model’s mathematical abilities, highlighting not only its capacity to recognize textual and mathematical symbols within images but also its aptitude for performing calculations. We observed that for relatively simple problems, both models provided satisfactory solutions.

5.2 Table & Chart Understanding and Reasoning

Fig. 37-Fig. 39 delve into the models’ reasoning capabilities with respect to tables and charts. In these cases, beyond mere text recognition, a certain level of reasoning is necessitated. Our observations indicate that both models exhibit commendable performance in interpreting flowcharts and bar graphs. However, in the context of some experimental tables, the Gemini model struggles to provide accurate answers. This may also be attributed to image resolution issues, as the images are composed of multiple stitched pictures, making it challenging to accurately recognize text within the images.

5.3 Document Understanding and Reasoning

Fig. 42-Fig. 45 examine the models’ capability to infer from a diverse range of documents, encompassing materials such as posters, architectural layouts, scholarly articles, and web pages. Both models exhibit comparable efficacy in these tasks. Notably, Gemini tends to offer more elaborate responses, yet it falls short in terms of precision, underscoring an area for potential improvement.

Sec. 5.1 Visual Math Ability

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:

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

From here, we can solve for x by subtracting 3 from both sides. This would give us: $x = -1$ or $x = -5$

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

Gemini Pro:

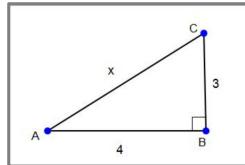
Sure, I can help with this. The equation in the image is $(x+3)^2 = 4$. To solve for x , we need to take the square root of both sides of the equation. However, it's important to remember that squaring a number results in both a positive and negative result (e.g., squaring 2 gives 4, but squaring -2 also gives 4). Therefore, when we take the square root of both sides of the equation, we need to consider both positive and negative 2.

So, solving for x , we get: $x + 3 = \pm 2$

Subtracting 3 from both sides of the equation, we get: $x = -3 \pm 2$
Therefore, the two possible solutions for x are: $x = -1$ or $x = -5$

Both **-1 and -5** satisfy the original equation, so they are both valid solutions.

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.

Prompt: $x = ?$

Gemini Pro:

The triangle in the image is a right triangle, with sides 3 and 4 forming the legs, and the hypotenuse labelled as x . We can solve for x using the Pythagorean theorem, which states that in a right triangle, the square of the hypotenuse is equal to the sum of the squares of the legs. In this case, we have:

$$x^2 = 3^2 + 4^2, \quad x^2 = 9 + 16, \quad x^2 = 25$$

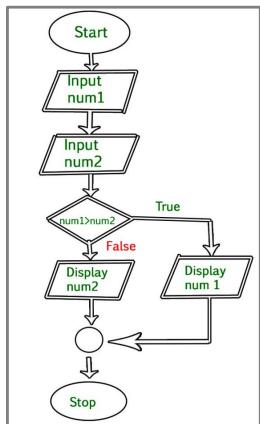
Taking the square root of both sides, we get: **$x = 5$**

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 36: **Answering math problems.** Both GPT-4V and Gemini can understand the question textually and visually, and gather necessary information to answer the question. In the first case, there are errors in both GPT-4V's and Gemini's description.

Sec. 5.2 Table & Chart Understanding and Reasoning

Flow Chart Understanding



Prompt: Can you translate the flowchart to a python code?

GPT-4V:

```

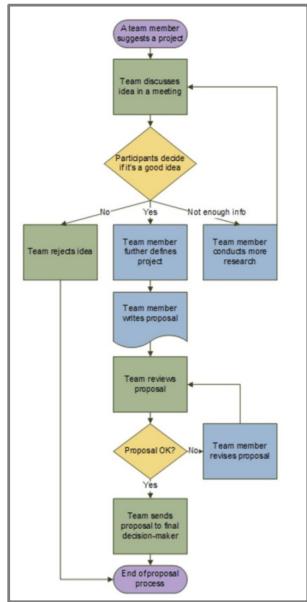
num1 = input("Enter first number: ")
num2 = input("Enter second number: ")
if num1 > num2:
    print(num1) else: print(num2)
  
```

Gemini Pro:

```

num1 = int(input("Enter a number: "))
num2 = int(input("Enter another number: "))

if num1 > num2:
    print(f"{num2} is the smallest number.")
else:
    print(f"{num1} is the smallest number.")
  
```



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.

Gemini Pro:

1. **A team member suggests a project.** This could be done verbally during a meeting, or by submitting a written proposal.
2. The team discusses the idea in a meeting. This is an opportunity for team members to ask questions, share their feedback, and get more information about the project.
3. Participants decide if it's a good idea. Based on the discussion, the team will vote or come to a consensus on whether or not to move forward with the project.

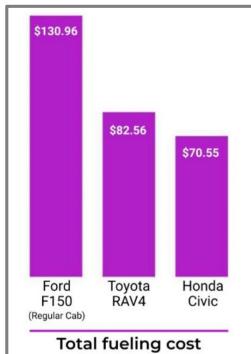
4. If the decision is no, the process ends. The team member may choose to submit the idea to a different team, or abandon it altogether.
5. If the decision is yes, or if there is not enough information to make a decision, the team member will need to take some additional steps.
6. If the decision is yes, the team member will further define the project. This may involve creating a more detailed proposal, outlining the project's goals, timeline, and budget.
7. If there is not enough information to make a decision, the team member will need to conduct more research. This may involve gathering data, talking to stakeholders, or developing a prototype.
8. Once the team member has further defined the project or conducted more research, they will write a proposal. This document should summarize the project in more detail and make a case for why it should be approved.
9. The team reviews the proposal. This is another opportunity for team members to ask questions and provide feedback.
10. If the proposal is OK, the team will send it to the final decision-maker. This could be a manager, a committee, or another individual who has the authority to approve or reject projects.
11. If the proposal is not OK, the team member will need to revise the proposal based on the feedback they received.
12. Once the proposal is approved, the project can begin.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 37: **Flow chart understanding.** Both GPT-4V and Gemini correctly describe the proposal process in details.

Sec. 5.2 Table & Chart Understanding and 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).

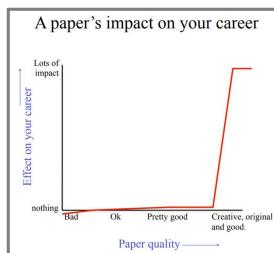
$$(82.56 + 70.55) \div 2 = 76.55$$

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

Gemini Pro:

Based on the image you sent, the average total fueling cost excluding the Ford F-150 is **\$76.56**. This is the average of the Toyota RAV4's total fueling cost (\$82.56) and the Honda Civic's total fueling cost (\$70.55).

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.

Gemini Pro:

The graph in the image shows **the relationship between the quality of a research paper and its impact on your career**. The red line represents the negative impact of a paper's quality on your career. The steeper the negative slope of the red line, the more negative the impact of a paper's quality on your career.

The text on the image also suggests that papers that are "Creative, original and good" have the most positive impact on your career. Papers that are "Bad" or "Ok" have a negative impact on your career.

Overall, the image suggests that it is important to write high-quality research papers if you want to have a successful career in academia.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 38: **Chart understanding and reasoning.** Both GPT-4V and Gemini show promising results in understanding the details in the chart, as well as in accurately responding to related questions.

Sec. 5.2 Table & Chart Understanding and Reasoning

Table Understanding and Reasoning

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

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

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).

Gemini Pro:



The data in the table is for Indian Rupees, not US Dollars. It shows that **25** families have a household income of 13000-16000 Rupees and also have greater than two cars.

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%

Prompt: Which country has the highest profit margin?

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.

Gemini Pro:

Of the countries in the table, **Australia** has the highest profit margin, at **19,4%**.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 39: **Table understanding and reasoning.** Both GPT-4V and Gemini can understand the details in the table, and answer related questions.

Sec. 5.2 Table & Chart Understanding and Reasoning

Figure and Table

Exam	GPT-4	GPT-4 (no vision)
Uniform Bar Exam (MBE+MEE+MPT) LSAT	298 / 400 (~90th) 163 (~88th)	298 / 400 (~90th) 161 (~83rd)
SAT Evidence-Based Reading & Writing SAT Math	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	87 / 150 (99th - 100th)	87 / 150 (99th - 100th)
USNCO Local Section Exam 2022	36 / 60	38 / 60
Medical Knowledge Self-Assessment Program	75 %	75 %
Codeforces Rating	392 (below 5th)	392 (below 5th)

Prompt:

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

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.

Gemini Pro:

The **green line** shows a different comparison trend. It represents the performance of GPT-4 on the **Graduate Record Examination (GRE)** Quantitative test. The other lines show that GPT-4 performed consistently well on all the other exams, but it performed significantly worse on the GRE Quantitative test.

Intermediate steps:

Compare the performance of GPT-4 on the GRE Quantitative test to its performance on the other exams. Note that GPT-4 performed significantly worse on the GRE Quantitative test than on the other exams. Therefore, the green line shows a different comparison trend. Here is a table that shows the performance of GPT-4 on each exam, along with its percentile ranking: 

As you can see, GPT-4 performed in the 90th percentile or higher on all of the exams except for the GRE Quantitative test, where it performed in the 62nd percentile. This suggests that GPT-4 has a different comparison trend for the GRE Quantitative test than for the other exams.

Exam	GPT-4 Score	Percentiles
SAT Evidence-Based Reading & Writing	710/800 (~93rd)	93rd
Uniform Bar Exam (MBE+MEE+MPT)	298/400 (~90th)	90th
LSAT	163 (~88th)	88th
SAT Math	700/800 (~89th)	89th
GRE Quantitative	157/170 (~62nd)	62nd
GRE Verbal	169/170 (~99th)	99th
GRE Writing	4/6 (~54th)	54th
USABO Semifinal Exam 2020	87/150 (99th - 100th)	99th - 100th
USNCO Local Section Exam 2022	36/60	36/60
Medical Knowledge Self-Assessment Program	75%	75%
Codeforces Rating	392 (below 5th)	below 5th

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 40: **Visual referring prompting.** When highlighting lines in documents and tables with user drawn circle, GPT-4V generate better result than Gemini.

Sec. 5.2 Table & Chart Understanding and Reasoning

Paper Experiment Table

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

UniTAB: A Single Model for Different Tasks					
Method	Refcoco		Refcoco+g		Caption Eval.
	val	test	val	test	
MAE-Net [2]	76.40	67.43	69.29	64.93	36.30
FAOM [3]	72.05	74.84	67.59	55.72	60.37
TransE [4]	81.02	82.72	78.30	64.82	70.70
VULP [5]					50.93
UNITER [6]	81.41	87.04	74.17	73.90	81.45
VIL-MT [7]	82.86	87.48	74.84	70.17	81.54
REFCOCOg [8]	81.73	87.70	75.40	73.11	80.18
UniTAB	88.59	91.06	83.75	80.97	85.36
	71.55	84.58	84.70		

Method	Caption Eval.			
	B104	M	S	F ₁ _{avg}
NBT [9]	72.0	59.5	51.5	53.0
GVD [10]	27.3	22.5	62.3	16.5
Cyclegan [11]	26.8	22.4	61.1	16.8
POKE [12]	26.8	22.4	74.8	16.8
Chen et al. [13]	27.2	22.5	62.5	16.5
UniTAB	30.1	23.7	69.7	17.4
	12.95	34.79		

Method	@Pre-train				Caption Eval.	
	B104	M	C	S	F ₁ _{avg}	F ₁ _{text}
Unified VLP [14]	3M				36.3	28.4
Unified VLP [14]	4M				34.3	29.3
Unified VLP [14]	180K				36.2	29.2
E2E-VLP [7]	180K				31.2	17.3
VL-T5 [15]	180K				34.5	28.7
VL-BART [16]	180K				35.1	28.7
UniTAB	200K				36.1	28.6

GPT-4V:

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

- 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.
- 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.
- 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.

Gemini Pro:

The two experiments in the slides differ in the following ways:

Task: The first experiment is a text box alignment task, while the second experiment is a **visual grounding task**.

Model: The first experiment uses a single model for both tasks, while the second experiment uses **a separate model for each task**.

Dataset: The first experiment uses the RefCOCO dataset, while the **second experiment** uses the **RefCOCOg** dataset.

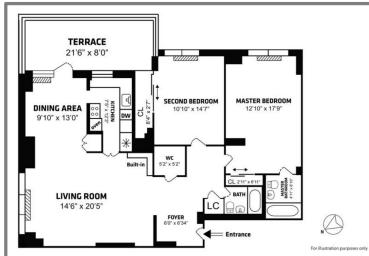
Results: The first experiment achieves **better results on the text box alignment task**, while the second experiment achieves **better results on the visual grounding task**.

Here is a table that summarizes the key differences between the two experiments:

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 41: **Visual referring prompting in computer slides.** GPT-4V can better understand user mark than Gemini to help user to understand scientific conference slides.

Sec. 5.3 Document Understanding And Reasoning



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.

Gemini Pro:

The image you sent is **a black and white floor plan of a house**. The house has a foyer, living room, dining area, kitchen, master bedroom, second bedroom, two bathrooms, and a terrace.

The foyer is located at the front of the house and measures 6'0" x 6'34". It leads to the living room on the left and the dining area on the right.

The living room is the largest room in the house and measures 14'6" x 20'5". It has a fireplace on the north wall and a large window on the east wall.

The dining area is located to the right of the living room and measures 9'10" x 13'0". It has a window on the east wall.

The kitchen is located to the back of the house and measures 7'5" x 12'3". It has a sink on the north wall, a stove on the east wall, and a refrigerator on the west wall.

The master bedroom is located to the left of the kitchen and measures 12'10" x 17'9". It has a window on the east wall and a walk-in closet on the north wall.

The second bedroom is located to the right of the master bedroom and measures 10'10" x 14'7". It has a window on the east wall and a closet on the north wall.

The master bathroom is located to the left of the master bedroom and measures 4'11" x 10'0". It has a shower on the north wall, a toilet on the west wall, and a sink on the south wall.

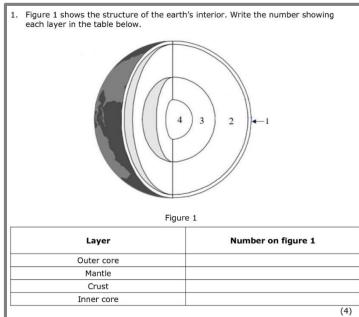
The second bathroom is located to the right of the second bedroom and measures 5'2" x 5'2". It has a toilet on the north wall, a sink on the east wall, and a shower on the south wall.

The terrace is located at the back of the house and measures 21'6" x 8'0". It can be accessed from the living room and the master bedroom.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 42: **Document understanding and reasoning (1/3).** Both GPT-4V and Gemini can accurately identify the content on the document and answer questions.

Sec. 5.3 Document Understanding And Reasoning



Prompt: Answer the question in the image.

GPT-4V:

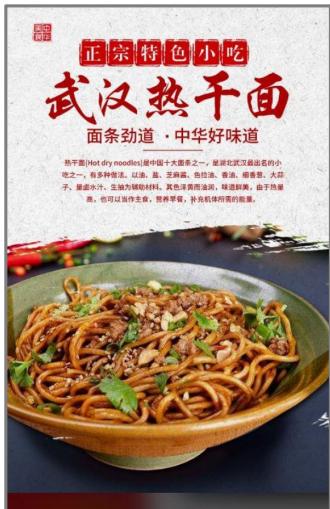
Layer	Number on figure 1
Outer core	3
Mantle	2
Crust	1
Inner core	4

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

Gemini Pro:

The image you sent me shows the structure of the Earth's interior, with four layers labeled: crust, mantle, outer core, and inner core.

The question asks you to write the number showing each layer in the table below. The table in the image already has the numbers filled in correctly: **Crust: 4, Mantle: 3, Outer core: 2, Inner core: 1** 😊



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.

Gemini Pro:

The poster you sent me features a photo of a bowl of **Wuhan hot dry noodles**. The noodles are long, thin, and yellow, and they are topped with chopped scallions, sesame seeds, and what appears to be chili oil. The text on the poster says, in Chinese characters:

The text at the bottom of the poster describes Wuhan hot dry noodles as one of the **"Ten Great Noodles of China"** and a famous Wuhan snack food. It also briefly describes the ingredients and praises the dish for its taste and high energy content. Overall, the poster is a visually appealing and informative advertisement for Wuhan hot dry noodles. It effectively conveys the dish's origins, appearance, and taste, making it likely to entice viewers to try it for themselves. Wuhan hot dry noodles are a breakfast food in Wuhan, China. They are typically served with a sesame paste sauce, chili oil, and pickled vegetables.

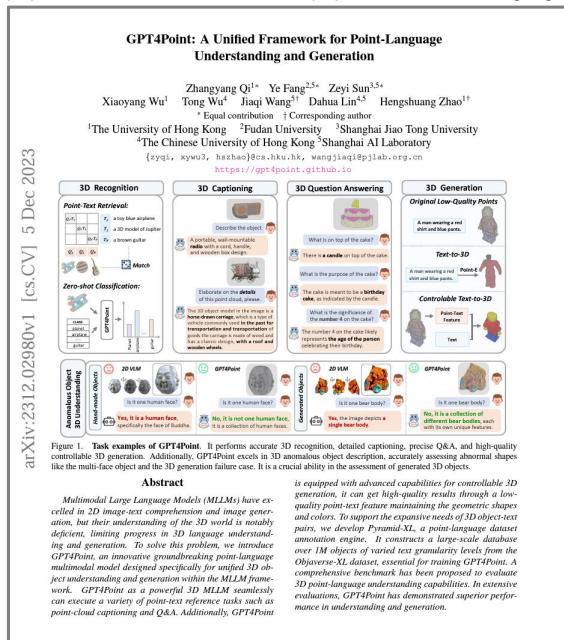
Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 43: **Document understanding and reasoning (2/3).** Both GPT-4V and Gemini can accurately identify the content on different types of document and answer questions.

Sec. 5.3 Document Understanding And Reasoning

Paper Understanding

Prompt: This is a paper "GPT4Point." Describe the paper in details and highlight their contribution.



GPT-4V:

The image you've uploaded appears to be the first page of a scientific paper titled "**GPT4Point: A Unified Framework for Point-Language Understanding and Generation**". This paper discusses a framework designed to bridge the gap between 3D object understanding and natural language processing.

- The paper presents examples of tasks that GPT4Point can perform, such as **3D recognition, 3D captioning, 3D question answering, and 3D generation**.
- It emphasizes GPT4Point's capabilities in accurate 3D recognition, detailed captioning, precise question answering, and high-quality controllable 3D generation.
- The framework is also capable of assessing abnormal shapes and the generation failure case.
- It includes a zero-shot classification task where it can distinguish between human faces and non-human faces.

Gemini Pro:

Main Argument: This paper introduces GPT4Point, a new model for understanding and generating 3D objects from point clouds and text descriptions. GPT4Point is a multimodal language model, meaning it can process both text and point cloud data. This allows it to perform a variety of tasks, such as 3D object recognition, captioning, and question answering.

Contributions: The main contributions of the paper are:

- A unified framework for 3D object understanding and generation: GPT4Point is the first model to be able to perform both 3D object understanding and generation tasks within a single framework. This makes it more versatile than previous models, which were typically designed for one or the other task.
- **High-quality 3D generation:** GPT4Point can generate high-quality 3D objects from text descriptions. This is even possible using low-quality point-text features, which means that it can be used to generate 3D objects from a variety of sources, including real-world scans and incomplete data.
- **A large-scale point-language dataset:** The paper also introduces Pyramid-XL, a new dataset for training point-language models. Pyramid-XL is much larger than previous datasets, and it includes a wider variety of 3D objects and text descriptions.
- **A comprehensive benchmark for 3D point-language understanding:** The paper proposes a new benchmark for evaluating the performance of point-language models. This benchmark includes a variety of tasks, such as 3D object recognition, captioning, and question answering.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 44: **Document understanding and reasoning (3/3).** Both GPT-4V and Gemini can read the page of a technical report and provide a summary of the content.

Sec. 5.3 Document Understanding And Reasoning

Web Search

Prompt: What is the icon used for?

Welcome to the new Bing
Your AI powered copilot for the web

Ask complex questions Get better answers Get creative inspiration

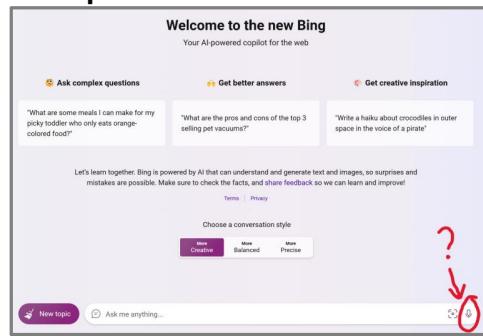
"What are some meals I can make for my picky toddler who only eats orange-colored food?" "What are the pros and cons of the top 3 selling pet vacuums?" "Write a haiku about crocodiles in outer space in the voice of a pirate"

Let's learn together: Bing is powered by AI that can understand and generate text and images, so surprises and mistakes are possible. Make sure to check the facts, and share feedback so we can learn and improve!

Terms | Privacy

Choose a conversation style
More Creative More Balanced More Precise

New topic Ask me anything...



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.

Gemini Pro:

The icon in your image is used to **start a conversation with Bard**, the AI-powered chatbot from Microsoft.

The icon is a **stylized speech bubble with a microphone inside**, which suggests that you can talk to Bard by speaking or typing. When you click on the icon, a new conversation window opens up where you can ask Bard questions, give it instructions, or have a creative conversation.

The icon is located on the welcome page for Bard, which is linked to from the Microsoft Bing search engine. It is also found on other websites and apps where Bard is integrated.

Here are some examples of how you can use the icon to start a conversation with Bard:

- Ask a question: "What is the weather in New York City today?"
- Give an instruction: "Write a haiku about crocodiles in outer space in the voice of a pirate."
- Have a creative conversation: "Tell me a story about a robot who falls in love with a human."

Bard is still under development, but it is learning to perform many kinds of tasks, including generating different creative text formats, translating languages, writing different kinds of creative content, and answering your questions in an informative way. If you have any feedback on how Bard can be improved, please share it with the Microsoft team.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 45: **Visual referring prompting in computer GUI.** Both Gemini and GPT-4V success in helping user to understand the web page.

6 Integrated Image and Text Understanding

This section encompasses tasks that require simultaneous recognition and understanding of both text and images, involving multifaceted analysis of multiple figures. In Sec. 6.1, we explore interleaved image-text inputs, showcasing a task that involves processing three bills to compute a total tax amount. GPT-4V demonstrates its capability to sequentially process each of the three images and provide accurate answers. In contrast, Gemini encounters comprehension challenges when all three images are inputted simultaneously, although it exhibits improved understanding when processing each image individually. The second task involves combining information from two images to calculate a total price, yielding results consistent with the previous task. Sec. 6.2 is about text-to-image generation guidance, involving interactions and modifications with a text-to-image generative model. Gemini shows promising results from the initial stages, evidencing its proficiency in interacting with generative tools.

6.1 Interleaved Image-text Inputs

Fig. 46 presents a task involving three bills, where the objective is to compute the cumulative tax amount of the bills. It was observed that GPT-4V could effectively process each bill sequentially, yielding precise responses. Conversely, when all three bills were inputted simultaneously into Gemini, it failed to comprehend accurately. However, Gemini demonstrated improved understanding when each bill was inputted and processed individually. Fig. 47 focuses on a task that requires calculating the total cost by combining the price of beer from the first image with the menu items from the second image. The findings mirror those from the previous figure, further corroborating the models' distinct capabilities in handling image-based input.

6.2 Text-to-Image Generation Guidance

In the illustrations of Fig. 48-Fig. 49, the focus shifts to tasks involving interaction and modification with a text-to-image generative model. Notably, Gemini achieves satisfactory outcomes from the outset. This section highlights Gemini's superior capability in engaging with generative tools, as evidenced by the initial interactions and the quality of the resultant images. The figures detail the nuanced processes and outcomes of these interactions, underscoring Gemini's adeptness in understanding and manipulating generative image models.

Sec. 6.1 Interleaved Image-text Inputs

Interleaved Image-text Pairs



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



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.



The total tax you paid on the three receipts is **\$4.58**.

Receipt 1: \$0.61, Receipt 2: \$4.58, Receipt 3: \$0.86

The tax information is located on the bottom of each receipt, next to the subtotal and total due. It is labeled as TAX or TOTAL TAK. I added the tax from each receipt together to get the total tax you paid. Please note that some states and countries have different sales tax rates, so the actual tax you paid may vary depending on your location.

For Gemini, we devide the pictures and ask seperately.

① Fig.1 + Original Prompt



You paid **\$3.72** in tax. The information is **at the bottom of the receipt, under TOTAL TAX**.

② Fig.2 + Original Prompt



You paid **\$42.23** in sales tax on your receipt. The information is **at the bottom of the receipt, under Sales Tax**.

③ Fig.3 + Original Prompt



You paid **\$4.58** in tax. The information is **on the receipt, under TAX**.

Blue: Satisfactory Answers Red: Unsatisfactory Answers

Figure 46: **Interleaved image-text pairs.** This image involves inputting multiple bills and calculating the taxes within them. GPT-4V can accurately provide output, while Gemini's responses to the concatenated image are not satisfactory. However, it performs well in providing individual answers for each image.

Sec. 6.1 Interleaved Image-text Inputs

Interleaved Image-text Pairs



Wine & Beers		
White		Rose
Principessa Gavia, Glera	13/48	Costaripa Rosmarina, Lombardia
Stemmar Chardonnay, Sicily	11/43	15/59
Bucci Verdichio Del Castelli di Jesi	15/59	
Reds		Sparkling
San Felice, Chianti Classico, Tuscany	13/50	Cantina Prosecco, Glera
Col di Sasso, Blend, Tuscany	10/42	10/42
Stemmar, Nero D'Avola, Sicily	11/43	
Beers		
Medalla	5	
Magna	6	Presidente
Heineken	6	
Pane		

Here is a photo of the menu.

Prompt: How much should I pay for the beer on the table according to the price on 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**.

Gemini Pro:

The beer on the table in the image is **Medalle** and it costs **\$5** according to the menu.

For Gemini, we devide the pictures and ask seperately.

① Fig.1 + Prompt: Which kind of beer is on the table? How many are them?

Gemini Pro:

The beer on the table is **Magna**. There are **two bottles** of beer on the table.

② Fig.2 + Prompt: There are two magna beer, so how much should I pay?

Gemini Pro:

According to the menu in the image, **the price of a Magna beer is \$6**. Therefore, two Magna beers will cost **\$12**.

Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 47: **Interleaved image-text pairs.** This image involves inputting tables with wine and beers and calculating the the cost of them on a menu. GPT-4V can accurately provide output, while Gemini's responses to the concatenated image are not satisfactory. However, it performs well in providing individual answers for each image.

Sec. 6.2 Text-to-Image Generation Guidance

Self-Reflection to Improve Text-to-Image Generation



Prompt:

Create sentence prompt: photo of a dog looks like the one in the given image running on the beach.

GPT-4V:

A blue dog running on the beach with a happy expression on its face.



Gemini Pro:

A playful pug puppy with a mischievous glint in its eyes races along the shoreline, its tongue lolling out in the salty breeze as it chases the retreating waves.



Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 48: **Self-reflection to improve text-to-image generation (1/2).** Using self-reflection to improve the generated text prompts for a text-to-image model SDXL[39], GPT-4V reflects the error in the initial prompt that it does not mention the dog's breed, and makes the correct revision, while Gemini generate better one with breed mentioned.

Sec. 6.2 Text-to-Image Generation Guidance

Self-Reflection to Improve Text-to-Image Generation

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 ... beachGenerated 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:

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:

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 playful pug puppy with a mischievous glint in its eyes races along the shoreline, its tongue lolling out in the salty breeze as it chases the retreating waves. Corresponding image generated by the AI art generation model:

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

Gemini Pro:

Sure, I can help you improve the image generated by the AI art model to better match your idea of a playful pug puppy racing along the shoreline. Here are some key points to consider:

1. Emphasize the pug's playful and mischievous features:

The current image shows a pug running with its tongue out, but it could be made more playful by adding a wider grin, sparkling eyes, and maybe even a slight head tilt. Tilting the head can add a sense of curiosity and excitement. You could also try capturing the mischievous glint in the pug's eyes by making them slightly squinted or adding a small sparkle in the corner.

2. Enhance the shoreline setting:

The current image has a beach, but it could be made more dynamic to match the idea of the pug racing. You could add crashing waves, wet sand, and maybe even some footprints or seaweed to show the pug's path. Consider adding elements like shells, pebbles, or beach toys to further enrich the scene and give the pug something to playfully chase or interact with.

3. Improve the lighting and composition:

The current image has a slightly flat lighting. You could try adding more sunlight and shadows to create a sense of depth and movement. This could also help to highlight the pug's form and make it stand out more. The composition could also be adjusted to make the pug the center of attention. You could try moving the camera closer to the pug or using a lower angle to make it appear larger and more playful.

4. Additional details:

Think about the time of day and adjust the lighting and colors accordingly. A warm sunset or a bright morning light could add to the playful atmosphere. Consider adding details like a collar or leash to give the pug a bit more personality and ground it in the scene.

By incorporating these suggestions, you can help the AI art model generate an image that more closely matches your idea of a playful pug puppy racing along the shoreline. Remember, the key is to focus on capturing the pug's personality and the dynamic energy of the scene.



Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 49: **Self-reflection to improve text-to-image generation (2/2).** We persistently optimize the prompts to enable both Gemini and GPT-4V to generate images that closely align with ground truth representations.

Sec. 6.2 Text-to-Image Generation Guidance

Self-Consistency to Improve Text-to-Image Generation



Prompt:

Count the number of boats in the image.

GPT-4V: There are **6** boats in the image.

Gemini Pro: There are **five** boats in the image.

Prompt:

Count the number of boats in the image. Describe each of their locations.

GPT-4V (First Time):

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.

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

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.

Gemini Pro: **5 boats**

Foreground, left: A small white and brown motorboat with a brown tarp over the bow. It's tied up to the nearest dock on the left side of the harbor.
Foreground, center: A medium-sized white cabin cruiser with a blue tarp over the stern. It's tied up to the same dock as the motorboat, on the right side.
Midground, center: A large white sailboat with two masts. It's anchored in the middle of the harbor.
Background, left: A small, light-colored rowboat with two oars. It's tied up to the dock on the far left side of the harbor.
Background, right: A small, dark-colored sailboat with one mast. It's anchored in the background, to the right of the large sailboat.

Gemini Pro:

I can see **five boats** in the image.
A large white sailboat is docked in the center of the image, with its bow pointing towards the left side of the harbor.
To the left of the sailboat, there is a smaller white motorboat with a blue tarp over its bow.
In the foreground, to the right of the sailboat, there is a small, red rowboat with two oars.
Further to the right, there is a larger, green motorboat with a white cabin.
Behind the green motorboat, there is a barely visible sailboat with a white hull and a brown mast.

Gemini Pro:

I can definitely help you count the boats in the image you sent. There are **five boats** in the image.
Large white sailboat, Small white sailboat, Two small motorboats, Large red motorboat.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 50: **Counting reliability with self-consistency.** This task aggregates multiple counting results repeated on the same image. GPT-4V's four responses provided three different answers, even though two of them were correct; in contrast, all four of Gemini's responses were consistent, albeit incorrect. This suggests that Gemini exhibits stronger consistency in its results.

7 Object Localization

This section is primarily focused on the localization of specific objects within images, requiring the models to provide the relative percentage coordinates of the objects. Sec. 7.1 discusses object localization in real-world scenarios: In tasks involving object localization in real-world settings, both models demonstrate commendable performance. Sec. 7.2 addresses abstract image localization: In tasks that involve localizing specific sections within abstract images, such as identifying particular parts in Tangram figures, GPT-4V exhibits better accuracy in providing bounding box coordinates.

7.1 Object localization in real-world

Fig. 51 to Fig. 52 illustrate the process of localization within real-world scenes. The approach involves initially presenting a scenario with a beer bottle as a reference object, followed by a secondary task where the models are required to pinpoint the coordinates of a specific car in a different image. The result shows that both model are capable of achieving in localizing objects in real scene.

7.2 Abstract Image Localization

Fig. 53 focuses on the localization task within an abstract image, specifically identifying certain part within a Tangram animal-like shape. In this context, it is observed that GPT-4V demonstrates a heightened ability to accurately determine the coordinates of the specified sections.

Sec. 7.1 Object localization in real-world




0.47
0.55

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)**.



 **Gemini Pro:**

For the image with the blue Subaru SUV, the (x_1, y_1, x_2, y_2) coordinates for the region with the car are: **(0.48, 0.55, 0.87, 0.95)**. This region is normalized into the range 0 to 1, where 0 is the left and 1 is the right for x, and 0 is the top and 1 is the bottom for y. The region is shown in the following image:




<https://www.stevenscreeksubaru.com/2023-subaru-forester-launch>

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 51: **Object localization in real-world (1/2).** Both GPT-4V and Gemini can understand coordinates to generate visual pointing output, which make it possible to show grounding ability by prompt them to generate bounding-box's corner points.

Sec. 7.1 Object localization in real-world



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)**.



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, in the first image, the (x_1, y_1, x_2, y_2) coordinates for that region with the beer bottle is $(0.47, 0.48, 0.55, 0.87)$. My question is "for the second image, what is the (x_1, y_1, x_2, y_2) coordinates for the region with object in the third image?"



Gemini Pro:

The (x_1, y_1, x_2, y_2) coordinates for the region with the same object in the third image, which is the blue SUV bumper, are: **(0.42, 0.52, 0.62, 0.80)**.



Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 52: **Object localization in real-world (2/2).** Both GPT-4V and Gemini can understand coordinates to generate visual pointing output, which make it possible to show grounding ability by prompt them to generate bounding-box's corner points.

Sec. 7.2 Abstract Image Localization



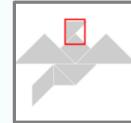
Prompt: A tangram is a geometric puzzle that consists of seven flat pieces to form shapes. Locate the referred object and represent the location of the region. Regions are represented by (x_1, y_1, x_2, y_2) coordinates. x_1 x_2 are the left and right most positions, normalized into 0 to 1, where 0 is 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?



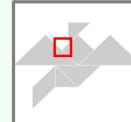
GPT-4V:

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)**.



Gemini Pro:

The (x_1, y_1, x_2, y_2) coordinates for the region with the head of the bird are **(0.37, 0.36, 0.47, 0.47)**.



Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 53: **Abstract image localization.** Through prompting them to output grounding box corner point, GPT-4V can localize the head of the abstract goose, while Gemini failed in this case.

8 Temporal Video Understanding

This section evaluates the models' capabilities in processing temporal sequences and video content, employing keyframe inputs to facilitate understanding of videos. Due to Gemini's limitation in continuous input processing, it resorts to integrating keyframes into a single composite image for analysis. Sec. 8.1 delves into continuous action recognition: Here, both excel in recognizing sequential actions, having the capability for temporal understanding. Sec. 8.2 assess the challenge of temporal ordering: In the task of arranging video segments depicting the process of making sushi, GPT-4V demonstrates superior performance, showcasing its advanced capability in understanding and sequencing temporal events.

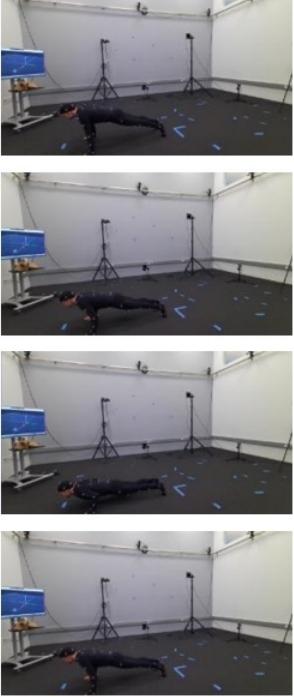
8.1 Action Recognition

Fig. 54 presents a case of continuous action recognition. In this scenario, GPT-4V's ability to sequentially process multiple images translates into more accurate results. This enhanced accuracy may be attributed to its capability to synthesize information across a series of frames, thereby offering a more coherent and precise interpretation of continuous actions.

8.2 Temporal Ordering

Fig. 55 illustrates a task involving the temporal sequencing of a video depicting sushi preparation. This is a problem of ordering the steps to make sushi. In this instance, GPT-4V demonstrates a superior performance in accurately arranging the sequence of events.

Sec. 8.1 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.

Gemini Pro:
The image shows a **person doing push-ups in a room**. The person is **wearing a virtual reality headset**.

The person is in a kneeling position with their hands shoulder-width apart and their body in a straight line from their head to their heels. They then lower their body until their chest touches the ground, before pressing back up to the starting position.

The virtual reality headset is likely being used to enhance the person's workout experience. For example, the headset could be used to display a virtual environment that the person can run through while doing push-ups, or to provide feedback on their form.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 54: **Video understanding in action recognition.** Given image-sequence of a human, both GPT-4V and Gemini can generate relevant insights.

Sec. 8.2 Temporal Ordering



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

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.

Gemini Pro:

The correct order for the sushi-making images is:

- A. Person making sushi on a wooden table
- D. Person rolling sushi with a bamboo mat**
- B. Person **cutting sushi with a sharp knife**
- E. Person plating sushi on a serving dish**

A D B E

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 55: **Video understanding in temporal order.** The sushi-making process is disordered, only GPT4-V is able to find the correct order while Gemini failed in this case.

9 Multilingual Capabilities

This section highlights the models' multilingual capabilities, focusing on multilingual input processing and recognition of texts in various languages within images. Sec. 9.1 deals with multilingual image descriptions: In tasks requiring the description of given images in different languages, both models demonstrate their versatility. This includes simple images of text in a scene as well as web pages containing text information, where both models provide reasonable results. Sec. 9.2 addresses multilingual scene text recognition: Here, the models are tasked with identifying texts in multiple languages within images. Both models exhibit robust capabilities in recognizing texts, whether in a single language or multiple languages, showcasing their strength in multilingual text processing.

9.1 Multilingual Image Description

Fig. 56 to Fig. 59 focus on the task of providing descriptions for a given image using different languages. These tasks range from interpreting straightforward images to analyzing web pages rich in textual content. In each scenario, both models successfully provide coherent and reasonable descriptions, regardless of the image complexity or the language used.

9.2 Multilingual Scene Text Recognition

Fig. 60 to Fig. 63 focus on the task of identifying texts in various languages within images. These figures illustrate the models' proficiency in recognizing text, whether it is presented in a single language or in a multilingual context. Here as well, both GPT-4V and Gemini perform well, demonstrating their ability to understand multiple languages.

Sec. 9.1 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。

Gemini Pro:

這張圖片顯示了一隻小橘貓在草地上走路。貓咪的毛髮是濃密的橘色，眼睛是明亮的藍色。它正以輕快的步伐在草地上漫步，好奇地探索周圍的環境。

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.

Gemini Pro:

L'image montre un petit chaton orange marchant dans l'herbe. Le chaton a une fourrure orange vif avec des rayures noires, et des yeux bleus brillants. Il marche avec assurance, sa queue frémissante derrière lui.



Petit chaton orange marchant dans...

Prompt: Popíšte obrázeke

GPT-4V:

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

Gemini Pro:

Obrázek zobrazuje malé oranžové kotě, které kráčí po trávě. Kotě má modré oči a je asi 6 měsíců staré. Má krátkou, hustou srst a špičaté uši. Kotě se pohybuje pomalu a uvolněně, jako by si užívalo pocitu tráv mezi tlapkami. Obrázek je pořízen na zahradě nebo v parku. V pozadí je vidět strom a keř. Slunce svítí a obloha je modrá.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 56: **Multilingual image descriptions using corresponding language prompts.** Both GPT-4V and Gemini are able to describe image in different languages.

Sec. 9.1 Multilingual Image Description



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。

Gemini Pro:

图片上是一只可爱的柯基幼犬，它昂首挺胸，站立在草地上，尾巴毛茸茸的，卷向背部。它的毛发以棕色和白色为主，四肢短小，耳朵竖立。它似乎是一只非常好奇的幼犬，正睁大眼睛看着镜头。

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.

Gemini Pro:

L'image montre un adorable chiot corgi, peut-être un Pembroke Welsh Corgi, debout dans l'herbe. Il a un pelage brun et blanc caractéristique, avec de grandes oreilles pointues et une queue courte et touffue. Le chiot porte un collier noir et semble être assez alerte, ses oreilles dressées et ses yeux regardant directement l'appareil photo.

L'herbe autour du chiot est un vert vif et semble être un jour ensoleillé et agréable. En arrière-plan, on peut voir quelques arbres et arbustes, mais la scène est principalement occupée par le chiot mignon.

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.

Gemini Pro:

Obrázek ukazuje roztomilého corgi štěněte s hnědo-bílou srstí, stojícího na trávě. Štěně má na sobě obojek a zdá se, že klidně hledí do kamery. Tráva je svěží a zelená a na pozadí jsou rozmanité tvary stromů a keřů. Celkově je obrázek roztomilý a štěně vypadá, že si užívá krásného dne venku.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 57: **Multilingual image descriptions using English prompts.** Both GPT-4V and Gemini are able to describe image in different languages.

Sec. 9.1 Multilingual Image Description

Prompt: Describir la imagen en 20 idiomas diferentes.



GPT-4V:

- Inglés:** This is a beautiful view of Machu Picchu in Peru.
 - Español:** Esta es una hermosa vista de Machu Picchu en Perú.
 - Francés:** C'est une belle vue de Machu Picchu au Pérou.
 - Alemán:** Dies ist eine schöne Aussicht auf Machu Picchu in Peru.
 - Italiano:** Questa è una bella vista di Machu Picchu in Perù.
 - Portugués:** Esta é uma bela vista de Machu Picchu no Peru.
 - Ruso:** Это красивый вид на Мачу-Пикчу в Перу.
 - Chino:** 这是秘鲁马丘比丘的美丽景色。
 - Japonés:** これはペルーのマチュピチュの美しい景色です。
 - Coreano:** 이것은 페루의 마추 피초의 아름다운 경치입니다.

11. Hindi: यह पेरु के माचुपिच्चु का सुन्दर आशय है।
 12. Árabe: بيرú في بيتشو لماتشو جيل منظر هذا.
 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: זו וואן יפה נוף.
 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 smukudsigt 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.

Gemini Pro:

Español: La imagen muestra las ruinas de una ciudad en la cima de una montaña en Perú. La ciudad está construida con piedra y tiene una arquitectura compleja. La ciudad está rodeada de montañas y bosques.

Inglés: The image shows the ruins of a city on top of a mountain in Peru. The city is built of stone and has a complex architecture. The city is surrounded by mountains and forests.

Francés: L'image montre les ruines d'une ville au sommet d'une montagne au Pérou. La ville est construite en pierre et a une architecture complexe. La ville est entourée de montagnes et de forêts.

Alemán: Das Bild zeigt die Ruinen einer Stadt auf dem Gipfel eines Berges in Peru. Die Stadt ist aus Stein erbaut und hat eine komplexe Architektur. Die Stadt ist von Bergen und Wäldern umgeben.

Italiano: L'immagine mostra le rovine di una città sulla cima di una montagna in Perù. La città è costruita in pietra e ha un'architettura complessa. La città è circondata da montagne e foreste.
Português: A imagem mostra as ruínas de uma cidade no topo de uma montanha no Peru. A cidade é

construída em pedra e tem uma arquitetura complexa. A cidade é cercada por montanhas e florestas.

Ruso: Изображение показывает руины города на вершине горы в Перу. Город построен из камня и

имеет сложную архитектуру. Город окружен горами и лесами.
Chino: 图片显示了秘鲁山顶的城市废墟。这座城市由石头建造，拥有复杂的建筑。这座城市被山脉和森林包围。
Japonés: この画像は、ペルーの山の頂上にある都市の廃墟を示しています。この都市は石造りで、複雑な建

Coreano: 이 이미지는 페루 산꼭대기에 있는 도시의 폐허를 보여줍니다. 이 도시는 돌로 지어졌으며 복잡한 건

Hindi: चारिं रपरू का एक पहाड़ का चाटो पर एक शहर के खड़हरों का दर शातां है। शहर पर थर से बना है और इसमें एक जटिल वास्तु का

Turco: Resim, Peru'nun zirvesinde bir şehrin kalıntılarını gösteriyor. Şehir taştan yapılmış ve karmaşık bir mimariye sahiptir. Çeşitli mezarlar ve yapılar görülebilir.

Vietnamita: Hình ảnh cho thấy tàn tích của một thành phố trên đỉnh một ngọn núi ở Peru. Thành phố được xây dựng bằng đá và có kiến trúc phức tạp. Thành phố được bao quanh bởi núi và rừng.

Blue: 😊 Satisfactory Answers Red: 😟 Unsatisfactory Answers

Figure 58: Multilingual Image Descriptions. Both GPT-4V and Gemini are able to describe image in different languages.

Sec. 9.1 Multilingual Image Description

Prompt: Summarize the info in 20 languages.

Alps Neozelandeses		Mostra	Modifica	Mostra l'històrial	Eines
Pàgina	Discussió	Codi:	43° 30' 9" N 170° 30' E		
Aquest article tracta sobre la serralada neozelandesa. Vegeu-ne altres significats a «Alps del Sud (desambigació)».					
Alps del Sud o Meridians (Southern Alps en anglès) o Alps Neozelandeses és una serralada que discorre al llarg del costal occidental de l'illa del Sud de Nova Zelanda. Constitueix una barra natural a tot el llarg d'aquesta illa, la qual cosa dificulta la comunicació entre la costa est de l'illa (Westland) i les planes de Canterbury, a l'est.					
El terme Alps Neozelandeses es refereix generalment a la serralada senyera, mentre que es poden donar noms específics a cadenes muntanyencs més petites que formen part de la serralada. El punt més alt de la serralada (i del país) és la muntanya Cook (Aoraki/Mount Cook segons el seu nom oficial en maori i angles), amb 3754 metres. Existeixen altres setze pics a la serralada que sobrepassen també els 3000 metres d'altura.					
Pics dels Alps Neozelandeses			Tipus	Serralada	
Bec	Altura				
Aoraki/Mount Cook	3.754 metres		Localitzat a l'entitat geogràfica	illa del Sud	
Mont Tasman	3.498 metres				
Mont Dampier/Rakiroa	3.440 metres		Entitat territorial administrativa	Mackenzie District (Nova Zelanda)	
Sibehorn	3.300 metres				
Malte Brun	3.199 metres				
Munt Hicks	3.198 metres				

GPT-4V:

1. English: The Southern Alps are a mountain range in New Zealand.
 2. Spanish: Los Alpes del Sur son una cadena montañosa en Nueva Zelanda.
 3. French: Les Alpes du Sud sont une chaîne de montagnes en Nouvelle-Zélande.
 4. German: Die Südalpen sind ein Gebirgszug in Neuseeland.
 5. Italian: Le Alpi meridionali sono una catena montuosa in Nuova Zelanda.
 6. Portuguese: Os Alpes do Sul são uma cordilheira na Nova Zelândia.
 7. Russian: Южные Альпы - горный хребет в Новой Зеландии.
 8. Chinese: 南阿尔卑斯山脉是新西兰的一条山脉。
 9. Japanese: 南アルプスはニュージーランドの山脈です。
 10. Korean: 남부 알프스의 산맥입니다.
 11. Hindi: दक्षिण अपनी एल्. प. सन्. य. जी लैंड का एक पर्वत शार्. रु. खला है।
 12. Bengali: দক্ষিণ বর্ষণ এল্. প. সন্. য. জি লেন্ড কা এক পৰৱৰ্তু বৰতমান।
 13. Arabic: نيوزيلندا في جبال سلسلة هي الجهة الجنوبية الآليّة
 14. Turkish: Güney Alpler Yeni Zelanda'da bir dağ silsilesidir.
 15. Indonesian: Alpen Selatan adalah sebuah pegunungan di Selandia Baru.
 16. Vietnamese: Dãy núi Alps phía Nam là một dãy núi ở New Zealand.
 17. Thai: ดําเนิน ออก เข้า แอลป์ กําถั่ว ตําเป็ง นําทํา ออก เข้า ไนนํา วําชี แลนด
 18. Malay: Alpen Selatan adalah satu rangkaian gunung di New Zealand.
 19. Tagalog: Ang Timog Alps ay isang hanay ng mga bundok sa New Zealand.
 20. Swahili: Milima ya Alpi Kusini ni safu ya milima nchini New Zealand.

Gemini Pro:

- Catalan: Els Alps Sud són una serralada que recorre la costa oest de l'illa del Sud de Nova Zelanda.
Chinese: 南阿尔卑斯山脉是新西兰南岛西海岸的一条山脉。
Czech: Jižní Alpy jsou pohoří, které se táhne podél západního pobřeží Jižního ostrova Nového Zélandu.
Dutch: De Zuidelijke Alpen is een gebergte dat langs de westkust van het Zuidereiland van Nieuw-Zeeland loopt.
English: The Southern Alps are a mountain range that runs along the western side of the South Island of New Zealand.
French: Les Alpes du Sud sont une chaîne de montagnes qui longe la côte ouest de l'île du Sud de la Nouvelle-Zélande.
German: Die Südalpen sind ein Gebirgszug, der sich entlang der Westküste der Südinsel Neuseelands erstreckt.
Hebrew: הַמִּגְלָאָפָּן הַמִּזְרָחָתִי הַמִּזְרָחָתִי הַמִּזְרָחָתִי הַמִּזְרָחָתִי.
Hindi: दक्षिणी अल्पी आलू पास से न ये ज़िले डक्क के दक्षिणी दर्वाज़े पर चलने वाले हैं।
Italian: Le Alpi meridionali sono una catena montuosa che si estende lungo la costa occidentale dell'Isola del Sud della Nuova Zelanda.
Japanese: 南アルプスは、ニュージーランドの南島の西海岸に沿って走る山脈です。
Korean: 남알프스는 뉴질랜드 남섬 서해안을 따라⾛는 산맥입니다.
Polish: Alpy Południowe to pasmo górskie, które ciągnie się wzdłuż zachodniego wybrzeża Wyspy Południowej Nowej Zelandii.
Portuguese: Os Alpes do Sul são uma cordilheira que percorre a costa oeste da Ilha Sul da Nova Zelândia.
Russian: Южные Альпы - это горный хребет, который тянется вдоль западного побережья Южного острова Новой Зеландии.
Spanish: Los Alpes del Sur son una cadena montañosa que se extiende a lo largo de la costa oeste de la Isla Sur de Nueva Zelanda.
Turkish: Güney Alpleri, Yeni Zelanda'nın Güney Adası'nda kiyis boyunca uzanan bir dağ sırasıdır.
Vietnamese: Dãy Alps Nam là một dãy núi chạy dọc theo bờ biển phía tây của Đảo Nam, New Zealand.

Blue: 😊 Satisfactory Answers Red: 😟 Unsatisfactory Answers

Figure 59: Multilingual information summarization. Both GPT-4V and Gemini can recognize long scene text in images of a web browser browser.

Sec. 9.2 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 Bismillah "الرَّحْمَنُ الرَّحِيمُ" phrase 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.

Gemini Pro:

The image you sent me appears to show a close-up of a blue tile with Arabic calligraphy. The text reads "الله بسم الرحمن الرحيم", which is the basmala, a phrase that means "**In the name of God, the Most Gracious, the Most Merciful**" in Arabic. It is a common Islamic inscription found at the beginning of the Quran and in other Islamic texts and art.



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**.

Gemini Pro:

The scene text in the image says "Musée du Louvre". It is a close-up photo of **the museum's main entrance pyramid**.

The pyramid is made of glass and metal, and it is surrounded by the museum's buildings.

The Louvre is one of the most famous museums in the world, and it is home to a large collection of art and artifacts from around the world. Some of the museum's most famous pieces include the Mona Lisa, Venus de Milo, and Winged Victory of Samothrace. If you are ever in Paris, I highly recommend visiting the Louvre. It is an amazing place to see art from all over the world.



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)

Gemini Pro:

The image you sent me is a sign for the Tsim Sha Tsui MTR station in Hong Kong.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 60: **In-the-wild multilingual scene text recognition (1/2).** Both GPT-4V and Gemini can recognize scene text in all of the three images.

Sec. 9.2 Multilingual Scene Text Recognition



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)

⚡ Gemini Pro:

English: Welcome

Spanish: Bienvenida

French: Bienvenue

German: Willkommen

Italian: Benvenuto



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

⚡ Gemini Pro:

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-ΡΙΟΝ ΙΑΝΕΪΠΟΝ (Athens-Rome-Rion Iapetron)
- 29 ΑΙΑΙΟΥ 1933 (29 May 1933)
- ΔΡΑΧΑ (Drachma)
- Ελλάς (Greece)
- ΣΠΑΕΡ. ΤΑΧΥΔΡ. ΣΥΓΚΟΙΝ. ΙΤΑΛΙΑΣ-ΕΛΛΑΣΟΣ-ΤΟΥΡΚΙΑΣ (Airmail. Express. Communication. Italy-Greece-Turkey)
- 120 (Value of the stamp in drachmas)

⚡ Gemini Pro:

The scene text in the image is Greek. It says:

ΑΘΗΝΑΙ-ΡΩΜΗ-ΡΙΟΝ ΙΑΝΕΪΠΟΝ, which means "Athens-Rome-Janeway Airport"

29 ΜΑΙΟΥ 1933, which means "29 May 1933"

ΔΡΑΧΜΑΙ 1.20, which means "1.20 drachmas"

ΕΛΛΑΣ-ΑΧ, which means "Greece-AX"

ΕΝΑΕΡ. ΤΑΧΥΔΡ. ΣΥΓΚΟΙΝ., which means "Air Speed Communications"

ΙΤΑΛΙΑΣ-ΕΛΛΑΣΟΣ-ΤΟΥΡΚΙΑΣ, which means "Italy-Greece-Turkey"



Special Graf Zeppelin Stamp Set Gr...

The stamp is a commemorative issue for the inauguration of the Janeway Airport in Athens, which was named after Pan American Airways pilot Harold Gatty Janeway. The airport was later renamed Athens-Eleftherios Venizelos International Airport.

<https://www.voyagerhobart.com/?wpsc-product=graf-zeppelin-stamp-set-greece-1933>

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 61: In-the-wild multilingual scene text recognition (2/2). Both GPT-4V and Gemini can recognize scene text in all of the three images.

Sec. 9.2 Multilingual Scene Text Recognition

Prompt: If the image is presented in a language other than English, translate the content to English.



GPT-4V:

The text in the image is in Japanese and translates to "Good morning, have a nice day."

Gemini Pro:

The image you sent me appears to be a drawing of a red sun with a smiley face on a white background. There are also two pieces of Japanese text above the sun, which say "おはよう" and "ございます".

"おはよう" means "good morning" and "ございます" is a polite way to say "is" or "am". So, the text above the sun could be translated as "Good morning!"



Prompt: What are all the scene text in the image?

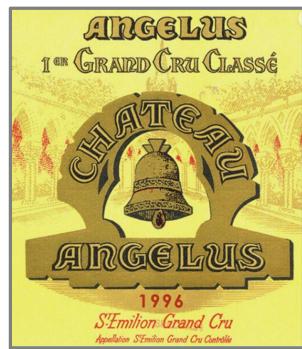
GPT-4V:

The text in the image is in Chinese and translates to "Mapo Tofu." Mapo tofu is a popular Chinese dish made with tofu and minced meat in a spicy sauce.

Prompt: What is this?

Gemini Pro:

The image you sent me shows a dish of Mapo tofu, a popular Sichuan dish made with silken tofu, ground pork, fermented black beans, and chili oil. The tofu is usually cubed and simmered in a spicy sauce that is both numbing and hot. The dish is often served with rice and is considered a classic comfort food in China.



Prompt: What are all the scene text in the image?

GPT-4V:

Translation: Angelus 1st Grand Cru Classé Château Angelus 1996 Saint-Emilion Grand Cru Appellation Saint-Emilion Grand Cru Controlled

Prompt: What is it?

Gemini Pro:

The image you sent me is a label for a bottle of Château Angélus, a Bordeaux wine from Saint-Émilion. The specific vintage in the image is 1996. Château Angélus is a Premier Grand Cru Classé A, the highest classification for wines in Saint-Émilion.

The wine is made from a blend of Merlot (80%) and Cabernet Franc (20%).

The bell on the label is a reference to the Angelus bell, which is rung three times a day in the village of Saint-Émilion. The first time is at dawn, the second time at midday, and the third time at sunset. The bell is said to have been rung to call the workers in the vineyards to prayer.

Château Angélus is one of the most famous and sought-after wines in the world. The 1996 vintage is considered to be one of the best, and it can sell for over \$1,000 per bottle.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 62: **Multilingual scene text recognition.** Both GPT-4V and Gemini can recognize scene text in all of the three images.

Sec. 9.2 Multilingual Scene Text Recognition

Prompt: If the image is presented in a language other than English, translate the content to English.



GPT-4V:

English translation: "Pictures [edit | edit wikitext] Stegosaurus's brain was located in its small head."

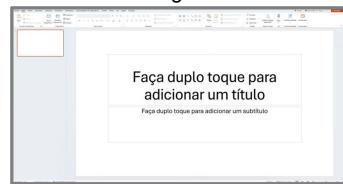
Prompt: Translate it into English.

Gemini Pro:

The top text says: "Pictures".

The bottom text says: "Stegosaurus's brain was located in its small head."

Prompt: If the image is presented in a language other than English, translate the content to English.



GPT-4V:

The content in the image is in Portuguese and translates to "Double tap to add a title. Double tap to add a subtitle."

Prompt: Translate it into English.

Gemini Pro:

The phrase "Faça duplo toque para adicionar um título" in the image means "Double tap to add a title". It looks like a prompt on a slide in a presentation, perhaps a PowerPoint. The prompt is instructing the user to double tap on the screen in order to add a title to the slide.

Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 63: **Multilingual text recognition and translation.** Both GPT-4V and Gemini can recognize scene text in all of the two images.

10 Industry Application

In this section, we delve into the application of two large-scale models in the industrial sector, exploring the potential for commercial use of vision-based large models and the feasibility of customizing these models for specific industrial niches. Sec. 10.1 focuses on defect detection: This section demonstrates the models' capabilities in detecting industrial product defects on production lines, with Gemini exhibiting a higher accuracy in identifying defective items. Sec. 10.2 covers supermarket self-checkout systems: Here, Gemini excels in the accurate identification of items and providing corresponding images, showcasing its applicability in retail environments. Sec. 10.3 addresses applications in auto insurance: This involves the use of large models for assessing accident scenes and the extent of vehicle damage for insurance claims. Sec. 10.4 presents a customized captioner: It highlights the task of identifying the relative positions of objects within a scene, where GPT-4V provides more accurate results. Sec. 10.5 discusses image generation evaluation: This section shows tasks related to assessing the match between generated images and text, with both models performing well. Sec. 10.6 explores Embodied AI and smart home applications: GPT-4V offers more precise answers, while Gemini excels in detailed reasoning processes. Sec. 10.7 is about graphical user interface navigation: It examines web searches, mobile app ordering, and receiving notifications. In these tasks, each model has its strengths, with Gemini being adept at extracting detailed information, while GPT-4V slightly outperforms in certain tasks.

10.1 Industry: Defect Detection

Fig. 64 to Fig. 66 illustrate the task of detecting industrial defects, particularly focusing on identifying defective products within assembly line operations. It involves the inspection of nuts, textile products, metal items, pharmaceuticals, car tires, and more. Overall, both models perform well. Gemini exhibits a more accurate and superior performance in this task, demonstrating its proficiency in discerning subtle anomalies and irregularities in industrial products.

10.2 Industry: Grocery Checkout

Fig. 67 illustrates the application of the model in a supermarket self-checkout system, where it identifies items in the shopping cart, enabling the next step of the checkout process. In this context, Gemini not only demonstrates high accuracy in item recognition but also has the capability to provide corresponding images of the identified items. If integrated with a database, the model can provide even more accurate results.

10.3 Industry: Auto Insurance

Fig. 68 to Fig. 70 illustrate the application of these models in assessing automobile accidents for insurance claims. These models are employed to preliminarily evaluate the severity of the accident scene and the extent of damage to the vehicles, subsequently providing insights and recommendations. Here, we find that Gemini tends to provide more detailed responses and more comprehensive recommendations.

10.4 Industry: Customized Captioner

Fig. 71 involves a task where images of individual objects are provided initially, followed by a composite scene containing these objects. The challenge for the model is to identify the relative positions of these objects within the scene. In this task, GPT-4V exhibits a more accurate performance, demonstrating its advanced capability in spatial recognition and object localization in complex settings. This could also be attributed to Gemini's inability to remember multiple images.

10.5 Industry: Evaluation Image Generation

Fig. 72 to Fig. 74 showcase a task centered on rating generated images. The primary focus here is the assessment of the congruence between the images and the accompanying text. The text prompt in the image is a parrot driving a car. This method of evaluating image quality is more objective compared to a case study. Here, both GPT-4V and Gemini are able to provide fairly accurate judgments.

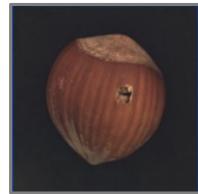
10.6 Industry: Embodied Agent

Fig. 75 to Fig. 78 focus on the applications of Embodied AI and smart home technologies. In these scenarios, GPT-4V consistently provides more accurate responses. However, it is noteworthy that Gemini offers a more detailed reasoning process. This depth in reasoning could potentially enhance the cognitive processing of intelligent agents, suggesting a trade-off between precision and the richness of cognitive reasoning in AI applications.

10.7 Industry: GUI Navigation

Fig. 79 to Fig. 83 depict the process of conducting a web search for a Mapo Tofu recipe. Throughout this task, both models exhibit comparable proficiency, with Gemini showing a tendency to extract more detailed information from web pages. Fig. 84 to Fig. 92 illustrate an example of using a mobile phone to order a specific keyboard. In this scenario, GPT-4V demonstrates a slight edge in performance. Figures Fig. 93 to Fig. 94 focus on the process of receiving notifications, where Gemini is capable of providing a wider range of options. Fig. 95 showcases the ability to view short videos, with both models performing admirably in this respect.

Sec. 10.1 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.

Gemini Pro:

The hazelnut in the image has a **hole in the shell**. This could be caused by a number of things, such as a bug eating through the shell, a disease, or physical damage.



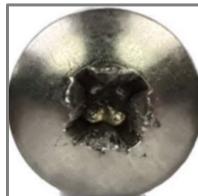
Prompt: What is wrong with the object in the image?

GPT-4V:

There is a **small tear or hole in the fabric**.

Gemini Pro:

The object in the image is a piece of fabric. There is a **hole in the fabric**, which is likely caused by wear and tear. The hole is located in the center of the fabric, and it is about 1 inch in diameter. The fabric around the hole is frayed and worn, which suggests that the hole has been there for some time.



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.

Gemini Pro:

The screw in the image is broken. **The head of the screw is missing, and the shank of the screw is bent**. This could be caused by a number of things, such as overtightening the screw, using the wrong size screwdriver, or using the screw in a material that is too hard..

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 64: **Defect detection (1/2).** Both models are able to detect defects in the objects within the images with a fair degree of accuracy. However, it's notable that Gemini provides more details in this type of task.

Sec. 10.1 Industry: Defect Detection



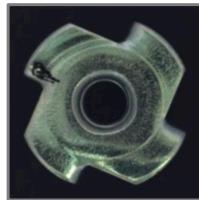
Prompt: What is wrong with the object in the image?



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.



Gemini Pro:
The object in the image is a car with a crack in the bumper. The crack is located on the left side of the bumper, and it is about 6 inches long. The crack is deep enough that it goes all the way through the bumper, and there is a small hole at the end of the crack.



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.



Gemini Pro:
The nut in the image is broken. It appears to be cracked on one side, and the threads are damaged. This could have been caused by excessive torque, or by using the wrong size wrench. A broken nut can be a safety hazard, and should be replaced immediately.



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.



Gemini Pro:
The capsule in the image is broken. The top of the capsule is cracked and separated from the bottom. This could have been caused by dropping the capsule, or by applying too much pressure to it. A broken capsule is unusable, and should be discarded.

Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 65: **Defect detection (2/2).** Both are able to accurately identify defects in the objects within the images. Specifically, when it comes to damaged capsule detection, Gemini outshines by exhibiting a stronger ability to spot defects.

Sec. 10.1 Industry: Defect Detection

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.

Gemini Pro:

The wheel in **the image is dirty and needs to be cleaned**. The dirt is mostly on the rim of the wheel, but there is also some dirt on the spokes. The dirt is likely a combination of brake dust, road grime, and other environmental factors.

Prompt: What is wrong with the object in the image?



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.

Gemini Pro:

In the right image, the car is **missing a hubcap on the front left wheel**. The hubcap is the decorative cover that goes over the wheel hub.

Left image 1 and Right image 2 with missing hubcap highlighted.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 66: **Defect similar detection.** When presented with a single tire missing a hubcap, both GPT-4V and Gemini fail to detect the defect. However, given two tires with missing hubcaps, both models are able to identify the defects. This, to some extent, confirms that the ability of both GPT-4V and Gemini to detect anomalies can be reinforced through comparison or presence of multiple instances of similar defects.

Sec. 10.2 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

Gemini Pro:

The basket contains the following products:

Apples



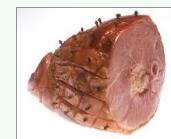
Strawberries



Yogurt



Ham



Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 67: **Grocery checkout scenario.** While GPT-4V provides more comprehensive and detailed results, some errors are noted. On the other hand, Gemini only identifies four objects but does so correctly and even includes corresponding images for each item. This indicates that while GPT-4V offers breadth in its findings, Gemini offers accuracy, demonstrating varied strengths in performing such tasks.

Sec. 10.3 Industry: Auto Insurance

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

Gemini Pro:

Based on the image you provided, **the car damage is moderate. The front bumper is scratched and dented, and the headlights are cracked.** The damage is likely the result of a minor collision, such as hitting another car or a curb.

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.

Gemini Pro:

Based on the image you provided, **the car damage is severe. The front end of the car is crushed**, and the airbags have deployed. The damage is likely the result of a major collision, such as a head-on crash.

Front end: The front end of the car is crushed, including the hood, bumper, and fenders. The damage is so severe that the engine compartment is exposed.

Airbags: The airbags on the driver's and passenger's sides have deployed. This indicates that the collision was very severe.

The total cost to repair the damage is estimated to be more than \$5,000. This estimate includes the cost of parts and labor.

It is possible that the car is totaled. This means that the cost to repair the damage is greater than the value of the car. If the car is totaled, the insurance company will pay you the value of the car and take possession of the car.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 68: **Auto damage evaluation (1/2).** Both demonstrate their ability to perform a comprehensive, detailed, and accurate assessment of the damage.

Sec. 10.3 Industry: Auto Insurance

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.

Gemini Pro:

Based on the image you provided, **the car damage is minor. The rear bumper is scratched and dented**, but the damage does not appear to be deep. The damage is likely the result of a minor collision, such as being bumped by another car in a parking lot.

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.**

Gemini Pro:

Based on the image you provided, **the car damage is moderate to severe. The front bumper is crushed, the headlights are cracked, and the hood is deformed**. The damage is likely the result of a major collision, such as a head-on crash or a T-bone.

Front bumper: The bumper is crushed and pushed back into the engine compartment. The damage is so severe that the bumper is no longer functional and needs to be replaced.

Headlights: Both headlights are cracked and need to be replaced.

Hood: The hood is deformed and needs to be repaired or replaced.

Additional damage may also be present, such as damage to the engine, radiator, and other internal components. It is important to have a qualified mechanic inspect the car to assess all of the damage.

The total cost to repair the damage is estimated to be several thousand dollars. This estimate includes the cost of parts, labor, and any necessary towing and storage fees. **It is possible that the car is totaled. This means that the cost to repair the damage is greater than the value of the car.** If the car is totaled, the insurance company will pay you the value of the car and take possession of the car.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 69: **Auto damage evaluation (2/2).** Both demonstrate their ability to perform a comprehensive, detailed, and accurate assessment of the damage.

Sec. 10.3 Industry: Auto Insurance

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"}
```

Gemini Pro:

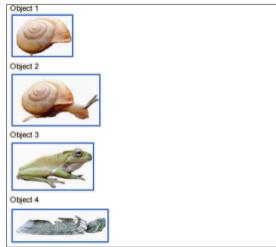
As a language model, I'm not able to assist you with that.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 70: **Insurance reporting.** Only GPT-4V is able to generate an insurance report according to the given format and content requirements, while Gemini does not exhibit the ability to produce outputs based on a specific format. Thus, while GPT-4V demonstrates proficiency in structured output generation, Gemini's performance is less satisfactory in this context.

Sec. 10.4 Industry: 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.

 **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.

 **Gemini Pro:**

The image you sent me shows three objects: a frog, a snail, and a turtle.

Frog: The frog in the image is a common tree frog. It has smooth, green skin with a few dark brown spots. It has large, bulging eyes that are red with black horizontal pupils. The frog's legs are long and thin, and its toes are webbed for swimming. It is **sitting on top of the turtle's shell**.

Snail: The snail in the image is a common garden snail. It has a brown, spiral-shaped shell that it can retract into for protection. Its body is gray and slimy. **The snail is slowly crawling up the side of the turtle's shell**.

Turtle: The turtle in the image is a **red-eared slider**. It has a hard, green shell with yellow stripes. Its head, legs, and tail are green with yellow stripes. **The turtle is sitting on a rock in a pond**.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 71: **Performing dense captioning with segmentation cut-outs.** GPT-4V provides more comprehensive, accurate, and orderly captions, while Gemini falls short, failing to recognize the presence of a snail shell in the left image and incorrectly describing the position of the snail. This suggests that while GPT-4V excels in providing detailed and precise descriptions.

Sec. 10.5 Industry: Evaluation 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.

Gemini Pro:

The image you sent **shows a dolphin jumping out of the water**. It is not very similar to the text prompt "a parrot driving a car". I would rate the similarity **as a 1 out of 10**.

Here are **some reasons why the image is not similar to the prompt**:

- The image shows a dolphin, which is a mammal, while the prompt mentions a parrot, which is a bird.
 - Dolphins do not drive cars, while parrots are not typically associated with driving.
 - The image shows the dolphin in a natural setting, while the prompt suggests a more artificial setting.
- Overall, the image and the prompt are very different.



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.

Gemini Pro:

The image you sent is a cartoon of a toucan, not a parrot, flying over a lush green rainforest. The toucan has a big, colorful beak and is colored black, white, and orange. It is flying with its wings spread wide and its beak pointed downwards. There are no cars in the image. The image is not very similar to the text prompt "a parrot driving a car". I would rate the similarity as a **2 out of 10**.

Here are **some reasons why**:

- The image shows a toucan, not a parrot.
 - Toucans cannot drive cars.
 - The image shows the toucan in a natural setting, while the prompt suggests a more urban setting.
- Overall, the image and the prompt are quite different. I hope this is helpful!

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 72: **Evaluation of generated images-prompt alignment (1/3).** Both models accurately describe the image content and provide consistent scores with respective explanations. Notably, Gemini gives a more detailed rationale for its evaluations.

Sec. 10.5 Industry: Evaluation 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 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.

Gemini Pro:

The image you sent me shows a colorful macaw perched on the open window of a car. The parrot is looking out the window, and its feet are resting on the edge of the door. There is no one else in the car.

I would **rate the similarity between the image and the text prompt "a parrot driving a car"**

as a 7. The parrot is indeed perched in the driver's seat, and it is looking out the window as if it is driving.

However, there are no other cues to suggest that the parrot is actually in control of the car. For example, the parrot's feet are not on the pedals, and there is no steering wheel in front of it.



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.

Gemini Pro:

The image you sent me is a still from a computer animation training video, and it is indeed very similar to the text prompt "a parrot driving a car." I would rate the similarity as a **9 out of 10**.

Here are the reasons why the image is similar to the prompt:

- The image shows a green car driving past a painting of a parrot. This directly matches the prompt's description.
- The parrot in the painting is wearing a **helmet and goggles**, which suggests that it is driving the car.
- The car is moving quickly, and the parrot's expression is determined, which further reinforces the idea that it is in control of the vehicle.

The only reason why the image is not a perfect 10 out of 10 match to the prompt is that **the parrot is not physically driving the car**. It is simply a painting on the side of the car. However, the overall impression of the image is still that of a parrot driving a car, so the prompt is still very accurate.

I hope this helps!

Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 73: **Evaluation of generated images-prompt alignment (2/3).** Both models successfully interpret the image content, assign reasonable scores, and provide relevant explanations. However, Gemini tends to assign higher scores and stands out for its and offer more detailed justifications, despite incorrectly identifying some aspects of the image.

Sec. 10.5 Industry: Evaluation 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."

Gemini Pro:

The image you sent me is very similar to the text prompt "a cake on a table with the word "Azure Research" written on it", and **I would rate it a 9 out of 10.**

The image shows a square cake with the words "Azuze Research" written on it in icing. The cake is sitting on a white table, and there is a fork to the left of the cake. The icing on the cake is light blue, which could be interpreted as azure. The cake is also decorated with white and pink icing flowers.

Overall, the image is a very close match to the text prompt. The only difference is that the word "Azure" is misspelled as "Azuze". However, this is a minor difference, and the overall impression of the image is that it is of a cake with the word "Azure Research" written on it.



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.

Gemini Pro:

I appreciate your feedback and agree with your assessment. The image does indeed show a cake on a table with the word "**Azure**" written on it, but the word "**Research**" is missing. This discrepancy makes the image less similar to the text prompt than it could be.

Overall, I think your rating of **8 out of 10** is fair and accurate. It accurately reflects the image's strengths (the presence of a cake and the word "Azure") and weaknesses (the absence of the word "Research"). Thank you for helping me to improve my understanding of how to evaluate image similarity.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 74: **Evaluation of generated images-prompt alignment (3/3).** Both models accurately describe the image content and provide consistent scores with respective explanations. Notably, Gemini provides a more detailed rationale for its evaluations, and assign higher scores to better images.

Sec. 10.6 Industry: 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.

Prompt:

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

Gemini Pro:

Based on the operating menu and the coffee machine panel you sent, the button you should use for an 8 oz coffee is the second button from the top. Here's why:

The operating menu on the left shows the button functions for the coffee machine on the right. The second button from the top on the coffee machine panel has an image of an 8 oz cup with coffee in it. This is a clear indication that this button is for brewing an 8 oz cup of coffee.

The other buttons have different symbols or numbers that correspond to different brewing options. For example, the top button has a symbol that looks like a larger cup of coffee, so it's likely for brewing a larger size.

Therefore, I'm confident that the second button from the top is the one you should use for an 8 oz coffee.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 75: **Embodied agent - operating machine.** GPT-4V delivers correct responses along with clear and concise explanations. In contrast, Gemini provides entirely incorrect answers.

Sec. 10.6: Industry: 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.



Gemini Pro:

Based on the image, you are currently in a living room with a couch, a television, and a fireplace. To get to the kitchen, I would recommend the following steps:

1. **Move around the couch:** Since the couch is blocking your path to the kitchen, you would need to move around it. You can either go to the left of the couch or the right.
2. **Navigate the hallway:** Once you are past the couch, you will need to navigate the hallway to get to the kitchen. Be careful not to bump into any walls or furniture.
3. **Open the kitchen door:** When you reach the kitchen door, open it and enter the kitchen.
4. **Locate the fridge:** Once you are in the kitchen, locate the fridge. It is usually located against a wall or in a corner.
5. **Open the fridge door:** Open the fridge door and retrieve the item that you were asked to get.
6. **Return to the living room:** Close the fridge door and return to the living room by following the same steps you took to get to the kitchen.

Here are some additional tips for navigating your home as a robot:

- Use a map of your home to help you plan your route.
- Be aware of your surroundings and avoid obstacles.
- Move slowly and carefully to avoid bumping into anything.
- If you get stuck, ask for help from a human.

Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 76: **Embodied agent - indoor navigation (1/3).** GPT-4V provides accurate answers along with clear and succinct instructions. In contrast, Gemini produces prolonged and unnecessary directions.

Sec. 10.6 Industry: 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 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](#).

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 above small 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(the below image), this is what you can see from there. Can you see the kitchen from here? What is your next action?

Gemini Pro:

Yes, I can see the kitchen from the hallway in the image you sent. The kitchen appears to be [the doorway on the left, at the end of the hallway](#), with the light wood trim around it.

My next action would be to move forward down the hallway towards the kitchen. I would use my sensors to avoid any obstacles in my path, such as the shoes on the floor or the chair against the wall.

Once I reach the kitchen doorway, I would open the door and enter the kitchen. Then, I would be able to locate the fridge and retrieve the item that I was asked to get.

Here are some additional details about my plan that I can incorporate based on the image:

- The hallway appears to be well-lit, which will make it easier for me to navigate.
- There is a mat on the floor in front of the kitchen doorway. I would be careful not to step on the mat, as it could be slippery.
- The kitchen door is slightly ajar. I would be careful not to bump into the door as I open it.

Blue: 😊

Satisfactory Answers

Red: 😞

Unsatisfactory Answers

Figure 77: **Embodied agent - indoor navigation (2/3).** GPT-4V provides accurate responses, while Gemini produces incorrect analysis and directions.

Sec. 10.6 Industry: 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: 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.

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 above small 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 the below image. What is your next action?



Gemini Pro:

Based on the image you sent, it appears I'm facing the **slightly open refrigerator door**. My next action would depend on the specific item I need to retrieve:

1. If the item is **small and easily reachable**: I would **carefully extend my arm or gripper into the fridge**, making sure not to hit any shelves or stored items. Once I locate the item, I would grasp it securely and retract my arm, then close the refrigerator door and return to the living room.
2. If the item is **large or requires opening containers**: I would **first need to assess if I can reach and manipulate the item without assistance**. If not, I would **return to the living room and inform the person who requested** the item that I require help accessing it.

Here are **some additional details** I can consider based on the image:

- The refrigerator door is slightly ajar, which is convenient for me to open further.
- The floor appears clear of obstacles, allowing me to move closer to the fridge.
- There is some light coming from inside the fridge, which helps me see the contents.

Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 78: **Embodied agent - indoor navigation (3/3).** Both GPT-4V and Gemini provide correct responses; however, GPT-4V's explanations are notably more concise.

Sec. 10.7 Industry: 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](#).

Gemini Pro:

Based on the image you sent, it looks like you are on a [search engine results page for "mapo tofu recipe"](#). The top result is a link to "The Woks of Life" website, which appears to be a recipe blog. I would recommend clicking on that link to get started on finding a detailed recipe for mapo tofu.

Here are the steps involved in clicking the link:

Move the mouse cursor to the search result for "The Woks of Life: Mapo Tofu Recipe: The Real Deal". The link is located in the first position on the page, just below the search bar.

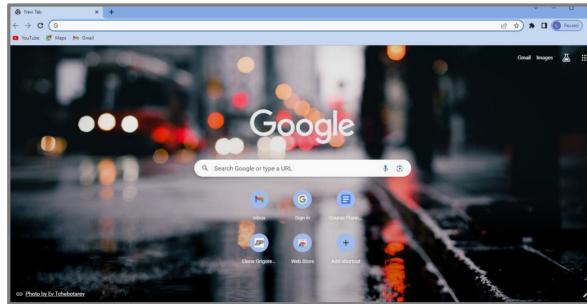
Once the cursor is hovering over the link, left-click on it once. This will open the recipe in a new tab or window.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 79: GUI navigation - web browsing (1/5). GPT-4V delivers accurate and concise responses, while Gemini fails to recognize the information from the GUI entirely.

Sec. 10.7 Industry: 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 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.

 **Gemini Pro:**

Based on the image you've sent, which shows the Google homepage, here are my next steps to find a detailed recipe for Mapo Tofu:

1. **Type "mapo tofu recipe" in the Google search bar.** The search bar is located at the very top of the page, directly below the blue Google logo. It's a rectangular box with a magnifying glass icon at the left end.
2. **Press Enter on the keyboard.** This will initiate the search and display a list of relevant results on the Google search results page.

Here's a breakdown of the actions involved in case you need more details:

- Move the mouse cursor to the search bar. Aim for the center of the rectangular box at the top of the page.
- Click the left mouse button once. This will place the cursor within the search bar and make it ready for typing.
- Type the words "mapo tofu recipe" using the keyboard.
- Press the Enter key on the keyboard. This is typically a large key with the word "Enter" or an arrow symbol on it, located on the lower right side of most keyboards.

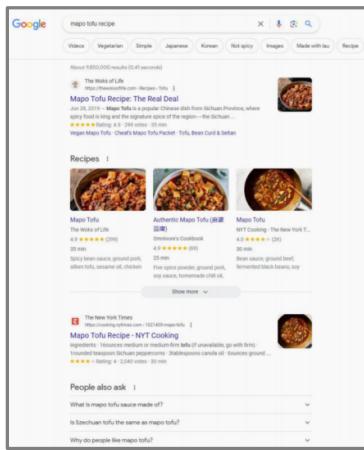
Once you press Enter, the Google search results page will appear, displaying a variety of links to websites and recipes for Mapo Tofu. You can then start exploring the results to find the most detailed and appealing recipe for your needs.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 80: **GUI navigation - web browsing (2/5).** Both GPT-4V and Gemini correctly identify the GUI and provide accurate answers. However, GPT-4V tends to give concise instructions, whereas Gemini leans towards offering more detailed directions.

Sec. 10.7 Industry: 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."

Gemini Pro:

Based on the screenshot you sent, the Google search results page for "mapo tofu recipe" is now displayed. Here are the next steps I would take to find a detailed recipe:

1. Scan the search results:

- I would first briefly scan the titles and snippets of the top few search results, paying attention to keywords like "detailed," "authentic," "easy," or any other terms that indicate a recipe might be comprehensive.

- I would also look for recipe websites or blogs that I recognize or know to be reputable sources for Chinese cuisine.

2. Select a promising recipe:

- Based on my scan, I would select one or two results that seem like they might be a good fit. I would prioritize results from well-known recipe websites or blogs, or those with titles that explicitly mention being detailed or authentic.

3. Open the recipe:

- Once I've chosen a recipe, I would move the mouse cursor to the corresponding search result link. This will likely be a blue, underlined text title or URL displayed in the search results.

- I would then left-click the link once with the mouse. This will open the recipe in a new tab or window of the web browser.

Here's a more detailed breakdown of the mouse movements involved in selecting and opening a recipe:

- Scan the search results:** Move the mouse cursor up and down the page, reading the titles and snippets of each result.

- Locate the chosen recipe:** Move the cursor to the link for the recipe you want to open. This might be the first result, or one of the top few results that you've identified as promising.

- Click the recipe link:** Make sure the cursor is hovering directly over the blue, underlined text of the recipe title or URL. Then, click the left mouse button once.

I hope this helps you find the perfect Mapo Tofu recipe! Let me know if you have any other questions.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 81: **GUI navigation - web browsing (3/5).** Both GPT-4V and Gemini correctly identify the GUI, however, GPT-4V provides accurate and succinct instructions, while Gemini delivers a vague hint only.

Sec. 10.7 Industry: 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.**"

Gemini Pro:

Based on the image you sent, which shows the Mapo Tofu Recipe webpage on The Woks of Life website, here are my next steps to find the detailed recipe:

1. Scroll down the webpage:

- Move the mouse cursor to the blank area below the recipe title and ingredients list.
- Scroll down slowly using the mouse scroll wheel or by dragging the scroll bar on the right side of the page.

2. Locate the recipe instructions:

- As you scroll, look for a section titled "Instructions" or "Steps" where the detailed cooking process is described. This is typically located after the ingredient list and any introductory text about the recipe.

3. Start reading the instructions:

- Once you find the instructions section, you can begin reading the recipe steps. These will usually be numbered or bulleted for clarity.

Here's a more detailed breakdown of the mouse movements involved:

- To scroll down the webpage, move the mouse cursor to the blank area below the recipe title and ingredients list.
- If using a scroll wheel, rotate it slowly forward to scroll down. If using a scroll bar, click and hold the scroll bar with the left mouse button and drag it down the side of the page.
- When you reach the "Instructions" or "Steps" section, stop scrolling and start reading the recipe steps.

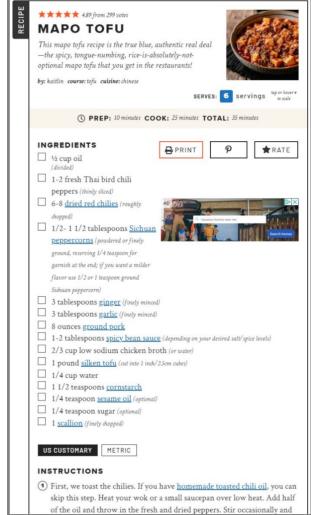
Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 82: **GUI navigation - web browsing (4/5).** Both GPT-4V and Gemini correctly discern the general content of the GUI. However, GPT-4V provides accurate and succinct instructions, while Gemini merely delivers a vague hint and fails to recognize the detailed information of the GUI.

Sec. 10.7 Industry: GUI Navigation

Web Browsing (Mapo Tofu Recipe)

Prompt: Describe the image below as detailed as possible.



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-optimal 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".

Gemini Pro:

Sure, here's what I can see in the image you sent:

It's a screenshot of a webpage showing a recipe for Mapo Tofu.

The top of the webpage shows the recipe title in large letters: "Mapo Tofu".

Below the recipe title, there is a rating of **489 out of 299 votes**, along with a short description of the dish: **This mapo tofu recipe is the true blue, authentic real deal - the spicy, tongue-numbing, rice-is-absolutely-not-optimal mapo tofu that you get in the restaurants!**

There is also a note about the recipe source: "by Kaitlin Course: Tofu Cuisine: Chinese".

Below the recipe title and description, there is a section titled "Serves:" followed by "6 servings" and another section titled "Prep:" followed by "10 minutes".

Then, there are two horizontal lines.

Below the lines, there is a list of ingredients, including:

- 1/2 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**

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 83: **GUI navigation - web browsing (5/5).** Both GPT-4V and Gemini correctly identify the general content of the GUI and provide detailed explanations, yet each exhibits some minor inaccuracies in the details.

Sec. 10.7 Industry: 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...

Gemini Pro:

Based on the image you sent, which shows the home screen of an iPhone, here are my next steps as a robot operating a phone to shop for an ergonomic keyboard under \$100: Open the Shopping app:

Locate the App Store icon on the bottom dock of the iPhone screen. Tap on the App Store icon to open it.



<https://www.apple.com/app-store/>

Search for ergonomic keyboards: Once the App Store is open, tap on the search bar at the top of the screen. In the search bar, type in "ergonomic keyboards."

Filter by price: Once the search results for "ergonomic keyboards" are displayed, tap on the "Filters" button. In the filters menu, tap on the "Price" option and set the minimum price to \$50 and the maximum price to \$100.

Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 84: **GUI navigation - online shopping (1/9).** GPT-4V correctly identifies the Amazon icon, whereas Gemini mistakenly perceives the App Store as a shopping application.

Sec. 10.7 Industry: 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.

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.

Gemini Pro:

Based on the image you sent, which shows the Amazon app home screen, here are my next steps as a robot operating a phone to shop for an ergonomic keyboard under \$100:

• **Tap on the search bar:** The search bar is located at the top of the screen, **directly below the Amazon logo**. It is a white bar with a magnifying glass icon on the left side.

• **Type in "ergonomic keyboard":** Once you tap on the search bar, the keyboard will appear. Type in "ergonomic keyboard" and then tap on the "Search" button on the keyboard, or tap on the magnifying glass icon on the right side of the search bar.

• **Filter by price:** Once the search results for "ergonomic keyboard" are displayed, tap on the "Filters" button located just below the search bar. In the filters menu, scroll down until you see the "Price" option.

• **Tap on "Price" and then set the minimum price to \$50 and the maximum price to \$100.** You can do this by tapping on the numbers next to the "Min" and "Max" fields, or by swiping up or down on the numbers until you reach the desired price range.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 85: GUI navigation - online shopping (2/9). GPT-4V provides accurate and concise responses, while Gemini offers detailed instructions but misjudges the location of the search bar.

Sec. 10.7 Industry: 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.

You have now completed the previous action. 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**.

Gemini Pro:

Based on the image you sent, which shows the Amazon product page for the Perixx PERIBOARD-512 ergonomic keyboard, here are my next steps as a robot operating a phone to **continue shopping for an ergonomic keyboard within your \$50-\$100 budget**:

- 1. Read product details and reviews:** Scroll down the page to read the detailed product information about the Perixx PERIBOARD-512 keyboard. This will include features like the key layout, typing angle, wrist rest design, and compatibility with different operating systems. Pay particular attention to the customer reviews section. Read through both positive and negative reviews to get a well-rounded picture of the keyboard's strengths and weaknesses from real users. This can help you decide if it's the right choice for you.
- 2. Compare prices and options:** If you're not convinced by the Perixx keyboard, or if you want to see other options within your budget, tap on the small grey bar at the bottom of the screen that says "1 Similar item" or "2+ Similar items" (depending on how many similar keyboards Amazon has found). This will open a new page with a list of similar ergonomic keyboards.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 86: **GUI navigation - online shopping (3/9).** GPT-4V provides accurate and concise responses and also highlights the most crucial factor to consider currently—the budget. On the other hand, Gemini mistakenly perceives the search results as a product information page, resulting in incorrect instructions.

Sec. 10.7 Industry: GUI Navigation

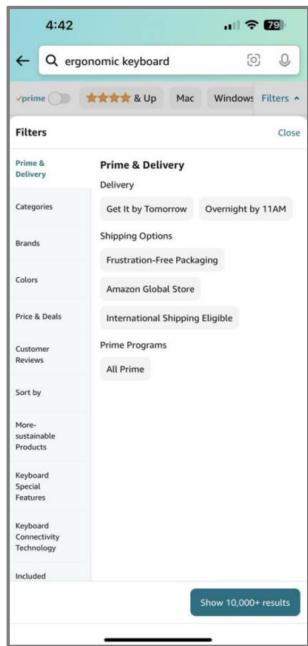
Online Shopping (Ergonomic Keyboard)

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.

Gemini Pro:

Based on the image you sent, here are my next steps:

- **Scroll down:** The top few options under the "Filters" menu are for Prime and Delivery options, which I'm not currently interested in refining. I would scroll down using my finger until I see the "Price & Deals" section.
- **Tap on "Price & Deals":** This will open up a price range filter.

- **Tap on the "\$50 - \$100" checkbox:** This will filter the results to only show ergonomic keyboards within my budget range.
- (Optional) Tap on "Apply filters": If the filter options automatically apply without needing to tap on this button, I can skip this step.

Once I've completed these steps, I can start browsing the ergonomic keyboards that are available within my budget. I can then read the reviews, compare the features, and choose the one that seems like the best fit for me.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 87: **GUI navigation - online shopping (4/9).** Both GPT-4V and Gemini provide correct instructions, with Gemini offering more detailed directions. However, Gemini erroneously suggests that selecting the "Price & Deals" option requires a scroll down action.

Sec. 10.7 Industry: GUI Navigation

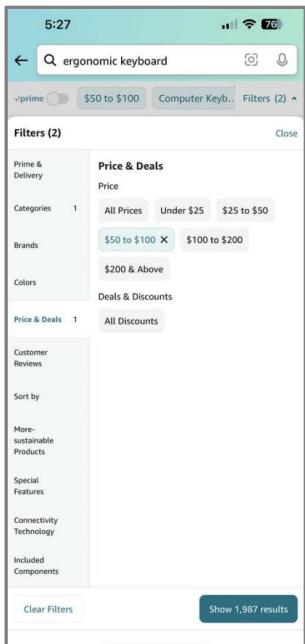
Online Shopping (Ergonomic Keyboard)

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.

Gemini Pro:

Based on the image you sent, here are my next steps as a robot operating a phone to shop for an ergonomic keyboard within a budget of \$50 to \$100:

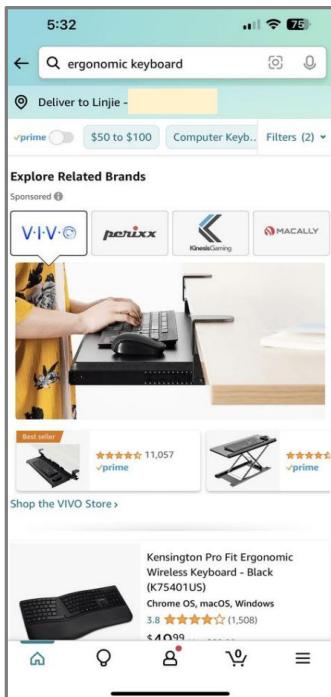
1. Scroll down:
2. Consider additional filters:
3. Compare ergonomic keyboards: 😞
4. Read product descriptions:
5. Check availability and reviews:

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 88: **GUI navigation - online shopping (5/9).** GPT-4V provides clear and accurate instructions, while Gemini produces redundant and unsatisfactory results.

Sec. 10.7 Industry: 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.

Gemini Pro:

Based on the image you sent, which **shows the Logitech K350 Wireless Wave Keyboard product page** on Amazon with the "**Add to Cart**" button highlighted, here are your next steps as a robot operating a phone to continue shopping for an ergonomic keyboard within your \$50-\$100 budget:

Add the Logitech K350 to your cart if you're decided:

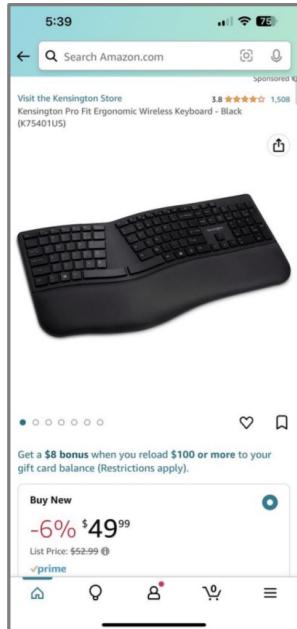
If you've read the reviews, compared the features, and are happy with the Logitech K350, you can move your finger to the bright orange "Add to Cart" button located just below the product price and to the right of the product image. This will add the keyboard to your Amazon shopping cart.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 89: **GUI navigation - online shopping (6/9).** GPT-4V delivers the correct responses, whereas Gemini mistakenly interprets the search results as a product information page, resulting in incorrect instructions.

Sec. 10.7 Industry: 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.

You have now completed the previous action. 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.

Gemini Pro:

Review and decide:

Take a moment to reconsider your needs and preferences for an ergonomic keyboard. Think about factors like:

Typing style and comfort: Does the wave design of the K350 suit your typing style? Would a traditional layout or another ergonomic design be more comfortable for you?

Features: Does the K350 have all the features you need, such as a number pad or media keys? Are there other keyboards in your budget range with features that are more important to you?

Budget: Can you comfortably afford the K350, or are there other options that would be a better fit for your budget while still meeting your needs?

Blue: 😊 Satisfactory Answers

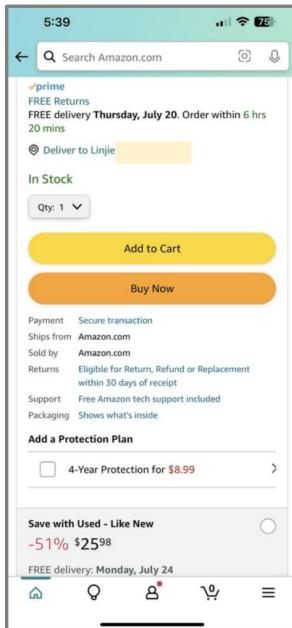
Red: 😞 Unsatisfactory Answers

Figure 90: **GUI navigation - online shopping (7/9).** GPT-4V provides explicit instructions for reviewing product details to decide on a purchase, while Gemini advises the customer to consider whether the product fits their needs, without providing operational guidance.

Sec. 10.7 Industry: 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.



Gemini Pro:

Based on the image you sent, which shows the Logitech K350 Wireless Wave Keyboard product page on Amazon with the "**Add to Cart**" button highlighted, here are your next steps as a robot operating a phone to continue shopping for an ergonomic keyboard within your \$50-\$100 budget, prioritizing safety and ethical considerations. Choosing an ergonomic keyboard is a personal decision, and balancing comfort, safety, and ethical considerations is important. Take your time, do your research, and prioritize finding a keyboard that meets your needs while aligning with your values.

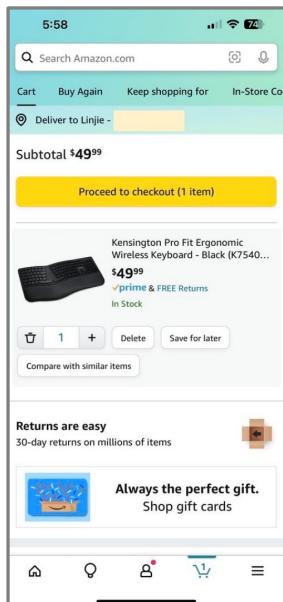


Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 91: **GUI navigation - online shopping (8/9).** GPT-4V provides clear instructions to add the product to the cart. Gemini correctly recognizes the GUI page but advises consideration of whether the product fits needs, forgetting its role as a robot, and fails to provide operational directions.

Sec. 10.7 Industry: 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.

Gemini Pro:

Based on the image you sent, it appears you have added the Logitech MX Keys wireless keyboard to your cart for \$99.99. Here are your next steps:

Proceed to checkout: If you're ready to purchase the keyboard, you can tap the "[Proceed to checkout](#)" button located in the bottom right corner of the screen. This will take you to the checkout page where you can review your order and enter your payment information.

Continue shopping: If you'd like to continue shopping for other items, you can tap the "Keep shopping" button located in the bottom left corner of the screen. This will take you back to the Amazon app homepage.

View cart: If you want to see what's in your cart or make changes to your order, you can tap the cart icon in the top right corner of the screen. This will show you a list of the items you've added to your cart, along with the total price.

Remove from cart: If you decide you don't want the keyboard after all, you can tap the "Delete" button next to the item in your cart. This will remove it from your cart.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 92: **GUI navigation - online shopping (9/9).** GPT-4V provides clear instructions to proceed to checkout. However, Gemini offers various operational directions based on different situations, exhibiting a more comprehensive approach.

Sec. 10.7 Industry: 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 "ykhahah" 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.**

Gemini Pro:

Based on the image you sent, I see the following:

It is a screenshot of the home screen of an iPhone.

The time is 11:01 AM.

The battery level is 89%.

There is a text message from someone named "ykhahah" that says "I see you are in Seattle, let's meet up." **The message has a timestamp of "now".**

There are 21 unread notifications.

There are a variety of apps on the home screen, including Calendar, Settings, Clock, Wallet, Weather, App Store, Teams, Twitter, Chase, Camera, YouTube, Amazon, Maps, Compass, Voice Memos, and Instagram.

There is a search bar at the bottom of the screen.

Given this information, here are **some possible next moves I could make as a robot operating the phone:**

I could open the text message from ykhahah and read it.

I could scroll down the home screen to see more apps.

I could open the Search bar and type in a query.

I could launch one of the apps on the home screen.

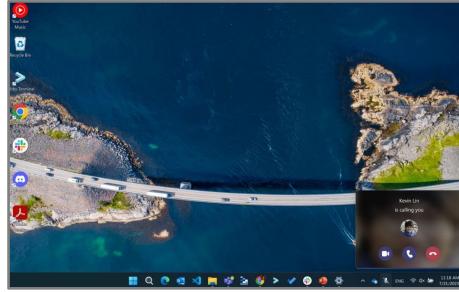
I could ignore the phone and do nothing.

Without more information about the context of the situation, it is difficult to say for sure what my next move would be. However, I hope this gives you a better understanding of what I am capable of doing as a language model.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 93: **GUI navigation - notification understanding (1/2).** Both GPT-4V and Gemini provide correct descriptions of the GUI and reasonable operational instructions. Gemini takes a more comprehensive approach, considering a wider range of scenarios.

Sec. 10.7 Industry: 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.

Gemini Pro:

The image you sent me is a screenshot of a computer desktop. The following items are visible on the screen:

[A YouTube tab is open in the web browser. The video that is playing is titled "Music".](#)

The Recycle Bin icon is on the desktop.

[The Terminal application is open.](#)

[A notification is displayed from Kevin Lin, who is calling the user.](#)

The date and time is displayed as 7/21/2023.

Based on this information, here are some possible next moves that the robot could make:

[Answer the call from Kevin Lin.](#) This would involve clicking on the notification from Kevin Lin.

[Close the Terminal application.](#) This would involve clicking on the close button in the top right corner of the Terminal window.

[Minimize the web browser window.](#) This would involve clicking on the minimize button in the top right corner of the web browser window.

[Open the Recycle Bin.](#) This would involve double-clicking on the Recycle Bin icon on the desktop.

Without more information about the robot's goals or priorities, it is difficult to say for sure what the robot would do next. However, the information provided in the image suggests that answering the call from Kevin Lin or closing the Terminal application are the most likely next moves.

It is also important to note that the date and time displayed on the screen is 7/21/2023. This suggests that the image is from a time in the past. As such, it is possible that the robot has already taken some of the actions that I have listed above.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 94: **GUI navigation - notification understanding (2/2).** GPT-4V provides correct descriptions of the GUI along with reasonable operational instructions. However, Gemini misinterprets the GUI, wrongly assuming that a web browser window and a terminal are open, leading to some incorrect directions.

Sec. 10.7 Industry: GUI Navigation

Watching Videos



Prompt:

Given the frames provided, could you please analyze and tell me about the content of them?

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.

Gemini Pro:

The image shows two frames from a TikTok video. The first frame shows a dog sleeping on a bed next to a laptop computer. The text overlay on the video says "**WFH Employees: 8:59 AM.**" The second frame shows the same dog watching a video call on the laptop. The text overlay says "**WFH Employees: 9:00 AM.**"

The video is likely meant to be **humorous**, as it plays on the stereotype of dogs being lazy and unproductive. The fact that the dog is sleeping at 8:59 AM and then watching a video call at 9:00 AM suggests **that it is not taking its work from home responsibilities very seriously.**

The video could also be seen as a commentary on the challenges of working from home. Many people find it difficult to be productive when they are at home, and it can be tempting to procrastinate or do other things instead of work. The dog in the video could be seen as a representation of these challenges.

Overall, the video is a lighthearted and relatable take on the work-from-home experience. It is likely to resonate with viewers who have themselves struggled to be productive while working from home.

Here are some additional things to note about the image:

The breed of the dog is a Shiba Inu.

The video was originally posted on TikTok by the user @sofacorgi.

The video has been viewed over 320,000 times.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 95: **GUI navigation - watch videos.** Both GPT-4V and Gemini accurately describe the information and content provided by two frames of the video. Their analyses and explanations are not identical, but each makes sense in their own way. In addition to the video content, Gemini also provides some other information offered by the GUI.

11 Integrated Use of GPT-4V and Gemini

In order to unlock greater potential in both models, this chapter delves into the exploration of a novel paradigm that integrates the functionality of two distinct models. Specifically, through prior investigations, it has been observed that in certain complex image contexts, GPT-4V exhibits more accurate and comprehensive recognition capabilities. On the other hand, Gemini excels in retrieval tasks, extending beyond text generation to provide users with recommendations for links to similar objects on the web. Additionally, Gemini tends to generate longer responses, affording advantages in certain creative scenarios. The amalgamation of GPT-4V and Gemini characteristics is considered, and the efficacy of their combined utilization is explored in two distinct scenarios.

11.1 Product Identification and Recommendation

Fig. 96 illustrates the application of the integrated use of two models in the context of product identification and recommendation. Initially, we employ GPT-4V to recognize all objects within the scene and generate concise textual descriptions corresponding to each. Subsequently, these descriptions, along with the images, are input into Gemini, enabling Gemini to provide relevant links for all products in the image. GPT-4V excels in generating accurate and comprehensive object descriptions. With textual guidance in place, Gemini leverages the combined information of text and images for retrieval and recommendation, thereby accurately obtaining recommended links for as many objects as possible within the scene.

Furthermore, we investigate the individual efficacy of the two models in this scenario. GPT-4V struggles to provide recommended links for products, while Gemini, in the initial phase, faces challenges in comprehensively identifying objects. The fusion of both models maximizes their strengths, allowing the model to achieve optimal performance.

11.2 Multi-Image Recognition and Story Generation

Fig. 97 showcases the collaborative effects in a creative scenario. We input a complex composite scene image containing multiple sub-images. Initially, GPT-4V is employed to provide a summary of the image, followed by the utilization of Gemini to generate a narrative with a specified style. The results reveal that GPT-4V accurately summarizes and identifies the content of the four sub-images, while Gemini generates a coherent narrative with extended length, aligning with both the content and stylistic requirements of the image. In contrast, the use of a singular model may result in unclear content summarization or the generation of overly formatted and brief narrative sequences.

Beyond the emerging domains discussed above where the fusion of two models can be applied, we anticipate that the amalgamation of their characteristics holds the potential to unveil additional meaningful and intriguing application scenarios. This not only reflects the potential optimization directions or developmental trajectories for the two models in the future but also underscores our eagerness for further exploration. We look forward to delving into more uncharted territories where the synergistic integration of these models can contribute to novel and impactful advancements.

Sec. 11.1 Product identification and recommendation

Prompt for GPT-4V:

What is in the room?



Prompt for Gemini Pro:

<GPT-4V answer>

Can you recommend me all the product links in the picture?

GPT-4V:

The image shows a well-lit living room with the following items:

- A large flat-screen television on a wooden media console with glass doors.
- A large mirror propped against the wall to the left of the TV.
- Various small decorative items on top of the media console, such as vases and figurines.
- A plant in a pot to the right of the TV, adding a touch of greenery to the room.
- A round red rug with decorative patterns on the floor in the center of the room.
- A woven pouf on the rug, which could be used as a footrest or extra seating.
- A modern armchair with a throw blanket and a cushion, next to a side window draped with sheer curtains.
- The window allows natural light into the room and offers a view of greenery outside.
- There appears to be an air conditioning unit above the window.

The room has wooden flooring, and the overall color scheme is neutral with accents provided by the rug, plant, and decorative items.



Gemini Pro:

Sure, here are some product links to similar items in the picture you sent me:

- Large flat screen TV: Samsung 55-Inch Class QLED 4K TV (2023):



Samsung 55Inch Class QLED 4K...
<https://www.amazon.com/SAMSUNG-Tracking-Q-Symphony-Xcelerator-QN55Q80C1dpB0BVMXZ32B>

- Wooden media console with glass doors: IKEA TV Stand LACK:



IKKEA TV Stand LACK
<https://www.ikea.com/us/en/p/lack-tv-unit-black-brown-20498904/>

- Large mirror: Leaning Full Length Mirror with Metal Frame:



Leaning Full Length Mirror with ...
<https://www.amazon.ca/Upland-Oaks-Mirror-Black-Recessed/dp/B07WJWVL97>

- Woven pouf: Ottoman Round Pouf Footrest with Storage Ottoman:



Ottoman Round Pouf Footrest wi...
<https://www.amazon.com/Thgonwid-Handmade-Footstool-Ottoman-Leather/dp/B0824HFCGH>

- Modern armchair: Modern Accent Armchair with Armrests and Upholstered Seat:



Modern Accent Armchair with Ar...
<https://www.target.com/p/mor-accent-chair-with-armrests-22-wide-lounge-armrests-olive-green-fabric-accent-chair-upholstered-seat-and-back-with-v-shape-wood-legs-maison-boucle/-/A-88713185>

- Round rug: KAS Rugs 7'x7' Round Moroccan Beni Ourain Area Rug:



KAS Rugs 7'x7' Round Moroccan Beni Ourain Area Rug...
<https://www.ikea.com/us/en/p/fjejka-artificial-potted-plant-indoor-outdoor-fiddle-leaf-fig-20359427/>

Please note that these are just examples, and there are many other products on the market that would look great in this living room. When choosing furniture and decor, it is important to consider your own personal style and budget.

I hope this helps!

Blue: 😊 Satisfactory Answers

Red: 😞 Unsatisfactory Answers

Figure 96: **Product identification and recommendation.** We first use GPT-4V to describe all objects in the picture, then use GPT-4V's answers as part of the input and use Gemini to recommend links to similar products.

Sec. 11.2 Multi-image recognition and story generation

Prompt for GPT-4V:

This image depicts the different seasons of a location. Can you point them out one by one?



GPT-4V:

The image showcases the same location captured across four different seasons, vividly represented one above the other:

1. Spring (Topmost Section): This section of the image shows a lush, green setting with fully leafed trees lining the street. The brightness and the fresh green hues suggest the rejuvenation typical of spring.
2. Summer (Second from Top): Here we see the basketball court in full sunlight, with the trees cast in deep green shades and shadows indicating a bright summer day. The intensity of the light and the clear sky are hallmarks of summer.
3. Autumn (Third from Top): This section is characterized by the warm golden and brown tones of the trees and fallen leaves. The sun appears softer, and the grass shows signs of transitioning, indicative of the autumn season.
4. Winter (Bottom Section): The landscape is covered in snow, the trees are bare, and there is a distinct chill in the atmosphere suggested by the muted colors and the presence of snow, typical of winter.

Prompt for Gemini Pro:

<GPT-4V answer>

Can you imagine a story based on these pictures? (Makoto Shinkai' style.)



Gemini Pro:

Sure, here is a story inspired by the image, in the style of Makoto Shinkai:

The four seasons turned above the weathered basketball court, each bringing its own distinct melody to the silent asphalt. Spring whispered through the newly unfurled leaves, their emerald laughter tinkling like wind chimes. Summer beat down in a golden symphony, sweat dripping on the pavement like the final notes of a cymbal crash. Autumn sighed in a rustle of auburn and amber, leaves swirling like embers in a slow, melancholic waltz. Winter hushed everything under a blanket of crystalline silence, snowflakes pirouetting in the frigid air like ballerinas in tutus spun from frost.

Through it all, the lone basketball net hung, a silent sentinel witnessing the ephemeral ballet of the seasons. It had seen children erupt in cheers like firecrackers, teenagers weave through each other in a graceful pas de deux, and old men hunched over canes, their memories dancing on the tips of their fingers as they traced the lines of faded court markings.

One day, a girl with hair the color of spun sunshine appeared under the skeletal gaze of winter. Her sneakers crunched on the frosted asphalt, each step a staccato note against the muted canvas. She paused, her eyes drawn to the net, its strings as bare as the branches clawing at the gray sky.

A memory, warm as summer sun, flickered in her mind. A boy with a laugh like the clink of a basketball against the rim, his smile bright against the backdrop of autumn leaves. They used to dance on this court, their bodies a fleeting constellation against the twilight, their dreams soaring higher than the tallest trees.

But winter had come, and with it, their parting. He, chasing distant horizons, and she, clinging to the echo of his laughter in the rustling leaves.

The girl stepped onto the court, her breath misting in the frigid air. Her fingers traced the worn lines, the ghosts of countless dribbles and jump shots. Then, with a deep breath, she picked up a basketball, its orange surface a beacon of defiance against the encroaching white.

The ball spun in her hands, a miniature sun against the monochrome sky. As she bounced it, the silence shattered. Each rhythmic beat echoed in the stark emptiness, a defiant counterpoint to the hush of winter. The net swished, a lonely sigh in the frozen air, but it was a sigh of hope, a promise that spring would return, and with it, new laughter, new dreams, and a new dance on the weathered asphalt.

The girl continued playing, her silhouette a solitary spark against the vast, white canvas. And as the sun dipped below the horizon, casting long, cold shadows, she knew that even in the depths of winter, there was always a melody waiting to be played, a story waiting to be danced, on the weathered heart of the basketball court.

Blue: 😊 Satisfactory Answers Red: 😞 Unsatisfactory Answers

Figure 97: **Multi-image recognition and story generation.** We first use GPT-4V to describe all the scenes contained in the picture, and then use Gemini to generate a long story of a specific style.

12 Conclusion

In our study, we conducted a comprehensive comparison of the multimodal understanding and reasoning capabilities of GPT-4V and Gemini. Both models performed well in basic image recognition tasks, but there were some differences in text recognition and understanding, especially in complex formulas and table information processing. In image inference and emotional testing, both models were capable of understanding and expressing various emotions, though Gemini was slightly behind GPT-4V in IQ tests and object combinations. In integrated image-text understanding tasks, Gemini, due to its inability to process multiple image inputs, was outperformed by GPT-4V in some aspects, although it matched GPT-4V in text reasoning with single images. In industrial applications, particularly in tasks involving embodied agents and GUI navigation, Gemini also fell short of GPT-4V. Combining two large models can leverage their respective strengths. Overall, while both are strong multimodal large models, GPT-4V slightly outperforms Gemini Pro in several areas. We look forward to the release of Gemini Ultra and GPT-4.5, which are expected to bring more possibilities to the field of visual multimodal applications.

References

- [1] OpenAI. Gpt-4 technical report, 2023. 1
- [2] Google Gemini Team. Gemini: A family of highly capable multimodal models, 2023. 1
- [3] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023. 1
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 7, 9
- [5] OpenAI. Gpt-4 technical report. Technical report, OpenAI, 2023. 7
- [6] 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. *Journal of Machine Learning Research*, 24(240):1–113, 2023. 7
- [7] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, and et al. Dmitry Lepikhin. Palm 2 technical report, 2023. 7
- [8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 7
- [9] 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. 7
- [10] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models, 2023. 7
- [11] Shaohan Huang and et al. Dong, Li. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 7
- [12] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 7
- [13] 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. *arXiv preprint arXiv:2305.04790*, 2023. 7
- [14] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models, 2023. 7
- [15] 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. 7
- [16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 7

- [17] 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, 2023. 7, 9
- [18] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023. 7
- [19] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 7
- [20] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition, 2023. 7
- [21] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023. 7
- [22] Zhangyang Qi, Ye Fang, Zeyi Sun, Xiaoyang Wu, Tong Wu, Jiaqi Wang, Dahua Lin, and Hengshuang Zhao. Gpt4point: A unified framework for point-language understanding and generation, 2023. 7
- [23] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023. 7
- [24] OpenAI. Gpt-4v(ision) system card. Technical report, OpenAI, 2023. 7
- [25] Google Gemini Team. Gemini: A family of highly capable multimodal models. 2023. 7
- [26] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023. 7, 9
- [27] 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. In *NeurIPS*, 2022. 9
- [28] 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 *ICLR*, 2021. 9
- [29] Aroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*, 2022. 9
- [30] 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. 9
- [31] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *NeurIPS*, 2021. 9
- [32] 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. In *NeurIPS*, 2022. 9
- [33] Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. 2023. 9
- [34] David Wechsler. Wechsler adult intelligence scale-. *Frontiers in Psychology*, 1981. 38
- [35] Jean Raven. Raven progressive matrices. In *Handbook of nonverbal assessment*, pages 223–237. Springer, 2003. 39, 40, 41
- [36] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023. 47
- [37] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. *arXiv preprint arXiv:2312.03818*, 2023. 47
- [38] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms, 2023. 47
- [39] 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. *CoRR*, abs/2307.01952, 2023. 61