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

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

* Core Contributor ♣ Project Lead

Abstract

*警告：该PDF由GPT-Academic开源项目调用大语言模型+Latex翻译插件一键生成，版权归原文作者所有。
翻译内容可靠性无保障，请仔细鉴别并以原文为准。

项目Github地址: https://github.com/binary-husky/gpt_academic/。

项目在线体验地址: <https://chatpaper.org>。

当前大语言模型: gpt-3.5-turbo, 当前语言模型温度设定: 1。为了防止大语言模型的意外谬误产生扩散影响,
禁止移除或修改此警告。

大型多模型（LMMs）通过增加视觉理解等多感知技能来扩展大型语言模型（LLMs），以实现更强大的通用智能。在本文中，我们分析了最新的模型——GPT-4V(ision)（以下简称“GPT-4V”），以深化对LMM的理解。分析重点关注了GPT-4V可以执行的有趣任务，其中包含了用于探测其能力质量和通用性的测试样本，以及模型支持的输入和工作模式，以及有效的提示模型的方法。在我们探索GPT-4V的方法中，我们精心设计和组织了一系列有关各个领域和任务的定性样本。这些样本的观察结果表明，GPT-4V在处理任意交错的多模态输入方面具有前所未有的能力，并且其通用性使其成为一个强大的多模态综合智能系统。GPT-4V独特的能力在理解输入图像上绘制的视觉标记的同时，还能产生新的人机交互方法，例如视觉指引提示。我们通过深入讨论以基于GPT-4V的系统为基础的新兴应用场景和未来研究方向，对报告进行了总结。我们希望这一初步的探索能够激发未来对下一代多模态任务制定、利用和增强LMM解决现实问题的新方法以及对多模态基础模型的更好理解的研究。

目录

插图目录

1	Introduction	8
1.1	Motivation and Overview	8
1.2	Our Approach in Exploring GPT-4V	9
1.3	How to Read this Report?	10
2	GPT-4V’s Input Modes	10
2.1	Text-only Inputs	10
2.2	Single Image-text Pair	10
2.3	Interleaved Image-text Inputs	11
3	GPT-4V’s Working Modes and Prompting Techniques	13
3.1	Following Text Instructions	14
3.2	Visual Pointing and Visual Referring Prompting	14
3.3	Visual + Text Prompting	15
3.4	In-context Few-shot Learning	21
4	Vision-Language Capability	28
4.1	Image Description on Diverse Domains	28
4.2	Object Localization, Counting, and Dense Captioning	40
4.3	Multimodal Knowledge and Commonsense	45
4.4	Scene Text, Table, Chart, and Document Reasoning	51
4.5	Multilingual Multimodal Understanding	60
4.6	Coding Capability with Vision	65
5	Interaction with Humans: Visual Referring Prompting	68
5.1	Understand Pointing Inputs	68
5.2	Visual Referring Prompting	71
5.3	Generate Pointing Outputs	71
6	Temporal and Video Understanding	75
6.1	Multi-image Sequencing	75
6.2	Video Understanding	75
6.3	Visual Referring Prompting for Grounded Temporal Understanding	81
7	Abstract Visual Reasoning and Intelligence Quotient Test	82
7.1	Abstract Visual Stimuli	82

7.2	Discovery and Association of Parts and Objects	82
7.3	Wechsler Adult Intelligence Scale	82
7.4	Raven’s Progressive Matrices	82
8	Emotional Quotient Test	88
8.1	Read Emotion from Facial Expressions	88
8.2	Understand How Visual Content Arouses Emotions	89
8.3	Emotion Conditioned Output	91
9	Emerging Application Highlights	92
9.1	Spot the Difference	92
9.2	Industry	92
9.3	Medical	101
9.4	Auto Insurance	106
9.5	Customized Captioner	109
9.6	Image Generation	113
9.7	Embodied Agent	118
9.8	GUI Navigation	123
10	LMM Powered Agents	152
10.1	Multimodal Plugins	154
10.2	Multimodal Chains	154
10.3	Self-Reflection	154
10.4	Self-Consistency	155
10.5	Retrieval-Augmented LMMs	155
11	Conclusions	155
11.1	Summary and Conclusions	155
11.2	Towards Future LMMs	155

插图

1	Section 2.3: multi-image inputs and interleaved image-text pairs.	11
2	Section 3.1: following text instructions.	13
3	Section 3.1: constrained prompting.	16
4	Section 3.1: condition on good performance.	17
5	Section 3.2: different modes of visual pointing.	18
6	Section 3.2: visual pointing and visual referring prompting.	19
7	Section 3.3: multimodal example-grounded instruction.	20
8	Section 3.4: few-shot comparison example 1, Part I: zero-shot.	22
9	Section 3.4: few-shot comparison example 1, Part II: one-shot.	23
10	Section 3.4: few-shot comparison example 1, Part III: two-shot.	24
11	Section 3.4: few-shot comparison example 2, Part I: zero-shot.	25
12	Section 3.4: few-shot comparison example 2, Part II: one-shot.	26
13	Section 3.4: few-shot comparison example 2, Part III: two-shot.	27
14	Section 4.1: celebrity recognition and description.	30
15	Section 4.1: landmark recognition and description.	31
16	Section 4.1: landmark recognition and description.	32
17	Section 4.1: food recognition and description.	33
18	Section 4.1: medical image understanding.	34
19	Section 4.1: medical image understanding.	35
20	Section 4.1: logo recognition and description.	36
21	Section 4.1: in-the-wild logo recognition and description.	37
22	Section 4.1: scene understanding.	38
23	Section 4.1: counterfactual examples.	39
24	Section 4.2: spatial relationship understanding.	41
25	Section 4.2: object counting.	42
26	Section 4.2: object localization.	43
27	Section 4.2: dense captioning.	44
28	Section 4.3: joke and meme understanding.	46
29	Section 4.3: science and encyclopedia knowledge.	47
30	Section 4.3: science and encyclopedia knowledge.	48
31	Section 4.3: science and encyclopedia knowledge.	49
32	Section 4.3: multimodal commonsense.	50

33	Section 4.4: scene text recognition.	52
34	Section 4.4: visual math reasoning.	53
35	Section 4.4: flow chart understanding.	54
36	Section 4.4: flow chart understanding.	55
37	Section 4.4: chart understanding and reasoning.	56
38	Section 4.4: table understanding and reasoning.	57
39	Section 4.4: document understanding.	58
40	Section 4.4: long document understanding.	59
41	Section 4.5: multilingual image descriptions.	61
42	Section 4.5: multilingual scene text recognition.	62
43	Section 4.5: multilingual text recognition, translation, and description reasoning.	63
44	Section 4.5: multilingual multiculture understanding.	64
45	Section 4.6: generate LaTeX codes based on the hand-written input.	65
46	Section 4.6: reconstruct table in Markdown or LaTeX code.	66
47	Section 4.6: write codes to replicate the input figure.	67
48	Section 5.1: understand pointing inputs for grounded description.	69
49	Section 5.1: grounded description with text-format region coordinates.	70
50	Section 5.2: examples of visual referring prompting.	72
51	Section 5.2: examples of visual referring prompting.	73
52	Section 5.3: generate pointing outputs.	74
53	Section 6.1: video frames understanding.	76
54	Section 6.2: long-term temporal order reasoning.	77
55	Section 6.2: short-term temporal order reasoning.	78
56	Section 6.2: short-term and long-term temporal anticipation.	79
57	Section 6.2: temporal localization and reasoning.	80
58	Section 6.3: grounded temporal understanding.	81
59	Section 7.1: abstract visual stimuli.	83
60	Section 7.2: part-object association.	84
61	Section 7.3: WAIS IQ test.	85
62	Section 7.4: RPM IQ test, entire question page.	86
63	Section 7.4: RPM IQ test, processed sub-figures.	87
64	Section 8.1: read emotions from facial expressions.	88
65	Section 8.2: how different visual contents may arouse emotions.	89

66	Section 8.2: image aesthetics.	90
67	Section 8.3: emotion conditioned outputs.	91
68	Section 9.1: spot the differences.	94
69	Section 9.1: spot the differences.	95
70	Section 9.2: industrial defect detection.	96
71	Section 9.2: industrial defect detection.	97
72	Section 9.2: industrial defect detection.	98
73	Section 9.2: safety inspection.	99
74	Section 9.2: grocery checkout.	100
75	Section 9.3: radiology report generation on medical images.	102
76	Section 9.3: radiology report generation on medical images.	103
77	Section 9.3: radiology report generation on medical images.	104
78	Section 9.3: radiology report generation with diagnosis history.	105
79	Section 9.4: auto damage evaluation.	107
80	Section 9.4: insurance reporting.	108
81	Section 9.5: customized captioner for photo organization.	110
82	Section 9.5: customized captioner for photo organization.	111
83	Section 9.5: dense captioning with segmentation cut-outs.	112
84	Section 9.6: evaluation of images generated from texts.	114
85	Section 9.6: evaluation of images generated from texts.	115
86	Section 9.6: image editing.	116
87	Section 9.6: image editing.	117
88	Section 9.7: embodied agent using coffee machine.	119
89	Section 9.7: embodied agent using coffee machine.	120
90	Section 9.7: embodied agent indoor navigation.	121
91	Section 9.7: embodied agent indoor navigation.	122
92	Section 9.8: web browsing for Mapo Tofu recipe.	124
93	Section 9.8: web browsing for Mapo Tofu recipe.	125
94	Section 9.8: web browsing for Mapo Tofu recipe.	126
95	Section 9.8: web browsing for Mapo Tofu recipe.	127
96	Section 9.8: web browsing for Mapo Tofu recipe.	128
97	Section 9.8: web browsing for today's news.	129
98	Section 9.8: web browsing for today's news.	130

99	Section 9.8: web browsing for today's news.	131
100	Section 9.8: web browsing for today's news.	132
101	Section 9.8: web browsing for today's news.	133
102	Section 9.8: web browsing for today's news.	134
103	Section 9.8: online shopping.	135
104	Section 9.8: online shopping.	136
105	Section 9.8: online shopping.	137
106	Section 9.8: online shopping.	138
107	Section 9.8: online shopping.	139
108	Section 9.8: online shopping.	140
109	Section 9.8: online shopping.	141
110	Section 9.8: online shopping.	142
111	Section 9.8: online shopping.	143
112	Section 9.8: notification understanding.	144
113	Section 9.8: notification understanding.	145
114	Section 9.8: notification understanding.	146
115	Section 9.8: watching videos.	147
116	Section 9.8: watching videos.	148
117	Section 9.8: watching videos.	149
118	Section 9.8: watching videos.	150
119	Section 9.8: watching videos.	151
120	Section 10.1: multimodal plugins.	152
121	Section 10.2: multimodal chains.	153
122	Section 10.3: self-reflection.	156
123	Section 10.3: self-reflection.	157
124	Section 10.4: self-consistency.	158

1 Introduction

1.1 Motivation and Overview

大语言模型（LLMs）的突破性进展 [22, 94, 27, 10, 116, 50] 在各个领域和任务中展现出了显著的多功能性和能力。这一领域的下一个发展阶段 - 大型多模型（LMMs）旨在通过整合多感觉技能来扩展LLMs的能力，以实现更强的通用智能。鉴于视觉在人类感觉中的占主导地位 [30, 55]，许多LMM研究都始于对视觉能力的拓展。初步的研究探索要么微调视觉编码器以与冻结的预训练LLM对齐 [118, 6, 67, 52, 39, 12, 45, 150, 75, 32, 139]，要么使用视觉语言模型将视觉输入转换为LLMs可以理解的文本描述 [142, 134, 124, 51, 106, 135]。然而，大多数现有模型 [12, 45, 150, 75, 32, 65] 在模型规模和数据规模上有限，可能限制了各种有趣能力的出现。因此，仍然不清楚基于最先进的LLMs（如GPT-4（无视觉） [94]和PaLM [27, 10]）开发的基于最新技术的LMMs的现状和新兴的多模态能力是什么。在本文中，我们报告了我们对（一个早期版本的）GPT-4V的初步探索，这是一个基于该领域的最先进LLM，带有视觉功能，并用大规模多模态数据进行训练。

我们对 GPT-4V 的探索是根据以下问题进行的。

1. GPT-4V的支持输入和工作模式是什么？多模态模型的通用性无疑要求系统能够处理各种不同输入模态的任意组合。GPT-4V在理解和处理输入图像、子图像、文本、场景文本和视觉指针的任意混合方面展现出了前所未有的能力。我们还展示了GPT-4V对LLMs中观察到的测试时技术的良好支持，包括指令跟随[96]、思维链[129, 63]、上下文少样本学习[22]等。
2. GPT-4V在不同领域和任务上的质量和通用性如何？我们对涵盖各种领域和任务的查询进行采样，以了解GPT-4V的能力，包括开放世界的视觉理解、视觉描述、多模态知识、常识、场景文字理解、文档推理、编码、时间推理、抽象推理、情感理解等等。在许多被实验涵盖的领域中，GPT-4V展现出令人印象深刻的接近人类水平的能力。
3. GPT-4V有哪些有效的使用和提示方法？GPT-4V在理解像素空间的编辑方面非常强大，例如对输入图像上绘制的视觉指针和场景文本的理解。受到这种能力的启发，我们讨论了通过直接编辑输入图像来指导感兴趣的任务的“视觉引用提示”。视觉引用提示可以与其他图像和文本提示无缝结合使用，为指导和示例演示提供了细致入微的界面。
4. 有哪些有前途的未来方向？鉴于 GPT-4V在各个领域和任务上的强大能力，我们探讨了多模态学习的下一步发展方向，更广泛地探讨了人工智能的未来研究方向。我们将思考和探索分为两个视角，即着眼于新兴应用场景的前景和基于 GPT-4V的系统的未来研究方向。我们提出了初步的探索，以鼓舞未来的研究。

在上述问题的指导下，我们全面组织和列出了我们的定性结果的探索。报告中包含了最少的定量基准结果，而主要是一些有趣的定性例子。尽管不够严谨，但这种设计可以提供更全面的分析，涵盖广泛的领域、任务、工作模式和提示技术，在一个固定的能力范围内。我们相信这个有组织的探索集合将会激发未来关于新兴应用、下一代多模式任务制定和发展先进的基于LMM的智能系统的研究。

1.2 Our Approach in Exploring GPT-4V

本报告的目标

评估系统的标准方法是通过将其与一系列精心设计的数据集进行基准测试，每个数据集代表特定领域和任务。挑战之一是现有的某些基准可能不再适用于评估LMMs。例如，与图像字幕基准数据集中的真实标注相比，LMMs的图像字幕输出更丰富，包含更详细的描述信息[24]。此外，关于GPT-4V的大规模预训练缺乏公开信息，这可能违反了某些现有数据集的训练-测试设置，并使这些基准结果失去效力。因此，将评估局限于现有的基准和指标可能会无意中狭窄了对 GPT-4V 的评估范围。

开发全面的下一代评估任务和基准列表将是理想的解决方案，但由于需要付出巨大努力，我们将其留作未来的工作。

作为定量基准测试的替代，本文着重利用定性结果来展示 GPT-4V 的新功能和潜在应用场景。我们的目标是发现并预览 GPT-4V 可能具备的功能，尽管这些新功能可能尚不完全可靠。我们希望这些探索的成果能够激发未来研究，建立下一代多模态任务的定量基准，更新现有的基准，进一步提高模型性能和系统可靠性，以及促进新兴应用场景的创新。在接下来的内容中，我们将深入讨论探索 GPT-4V 的核心设计。

样本选择指南

本报告的重点是通过展示定性结果来展示 GPT-4V 的潜在能力，而不是提供全面的定量基准结果。这自然引出了展示示例的可靠性问题。本报告中展示的示例可能需要进行仔细的指令调整才能放大 GPT-4V 相应的能力。需要注意的是，某些复杂情况可能只适用于特定设计的提示。因此，所展示的能力可能在不同样本中表现不一致。本报告的首要目标不是仅展示可靠的功能，而是向读者提供一份我们所发现的 GPT-4V 潜在能力的清单，这些潜力可能在一些不成功的尝试之后被忽视。

样本选择以防止仅通过训练进行记忆

定性报告中的一项基本设计考虑是区分模型真正的能力和仅仅记忆训练样本的回应或根据指令和上下文示例的提示进行猜测。我们仔细控制输入提示中的图像和文本，以防止在 GPT-4V 训练期间出现这些信息。我们从头开始生成原始文本查询，并尽量使用在线不可访问或时间戳在2023年4月之后的图像。我们将指出不符合此标准的特定样本，例如，故意使用特定视觉-语言数据集的样本。除了保证样本未被观察到外，我们还将理性化查询纳入过程中。这些查询旨在探究模型的推理过程，从而验证 GPT-4V 是否具备预期的能力。

默认工作模式

正如详细介绍的第3节所述，GPT-4V 在不同的工作模式下都能有效运行，包括使用指令进行零样本学习、上下文少样本学习等。其中，本报告主要关注零样本指令调整，而不是上下文少样本学习。这样设计是为了防止上下文示例中的潜在信息泄漏。虽然上下文少样本示例可以提高性能和可靠性，但它们并不一致地产生新的能力。因此，我们将零样本设定为演示的默认工作模式，并减少使用上下文示例以最小化示例对评估能力的影响。

1.3 How to Read this Report?

本报告介绍了计算机视觉和视觉-语言多模态领域的研究人员对 GPT-4V 进行的探索。主要面向相关学科的同行研究人员，他们希望对LMM的能力有一个定性印象，并了解其与传统的视觉-语言模型的区别。此外，本报告还准备给那些AI或计算机科学可能超出其专业领域的专业人士，以帮助他们理解LMM在其特定专业领域中如何提高自己的能力。

在本报告中，我们围绕我们的探索指导的四个核心问题进行了概述。

1. *GPT-4V*支持哪些输入和工作模式？第 2 节总结了 GPT-4V的支持输入，并概述了它们对应的使用情况。基于灵活的交织图像文本输入，第 3 节讨论了 GPT-4V的不同工作模式，如指导调整、环境学习和其他新出现的用法。该部分涵盖了 GPT-4V的新颖用途和提示方式，旨在为后续各节对 GPT-4V的使用提供全面的概述。
2. GPT-4V在不同领域和任务上的质量和通用性如何？探讨这个问题占据了报告的很大一部分。第 4节提供了包括图像描述、不同领域的图像识别、密集视觉理解、多模态知识、常识、场景文本理解、文档推理等在内的广泛分析,并且我们还提出了一些新颖和有趣的能力。第 6节研究了GPT-4V在时间、动作和视频理解方面的能力。第 7节探讨了抽象视觉理解和推理能力，而第 8节则涵盖了情感理解和情感判断能力。
3. 如何有效地使用和引导*GPT-4V*？我们从第 3节中的工作模式和引导方法介绍开始讨论这个问题。在第 5节中，我们强调了一种新颖的推广技术，即视觉指示引导，它在输入图像上绘制视觉指针和场景文本以引导GPT-4V。我们在给定的示例中展示了灵活的引导方法，例如指令和示范的组合。
4. 有哪些前景良好的未来方向？第 9 节侧重于 GPT-4V所支持的新型用例。我们希望这些初始示例能够激发未来的工作，设计新的任务设置并提出严格的基准。第 10 节设想了基于 GPT-4V可以构建的强大未来系统，例如多模态插件、多模态链、自我反思、自我一致性和检索增强的LMMs等等。

除了这个概述和目录，我们还包括了一个图例列表。这个列表列举了报告中详细描述的定义性示例，作为帮助读者导航到他们感兴趣的情景的附加工具。

2 GPT-4V’s Input Modes

本节总结了GPT-4V对支持的输入的概述，即作为一种仅支持文本输入的单模型语言模型，可选择地接收单个图像-文本对，也可选择地接收多个图像输入的交错图像-文本对。接下来，我们将重点介绍这些不同输入模式下的代表性用例。

2.1 Text-only Inputs

GPT-4V的强大语言能力使其成为一个有效的单模态语言模型 [35, 101, 22]，适用于纯文本输入。通过仅使用文本进行输入和输出，GPT-4V 能够执行各种语言和编码任务。我们建议读者参阅GPT-4技术报告 [94]，以获取关于GPT-4V的语言和编码能力的全面和深入的分析，以及GPT-4（无视觉）的比较。

2.2 Single Image-text Pair


GPT-4V，最新的大型多模态模型，接受图像和文本作为输入来生成文本输出。与现有的通用视觉语言模型（[8, 77, 69, 7, 66, 115, 113, 148, 25, 78, 42, 70, 54, 61, 68, 125, 26, 133, 38, 6,

121, 43, 37, 151, 65]) 一致, GPT-4V 可以接受单个图像-文本对或单个图像作为输入, 执行各种视觉和视觉语言任务, 如图像识别 [34], 目标定位 [146], 图像描述 [24], 视觉问答 [11], 视觉对话 [33], 密集描述 [59]等。我们指出, 图像文本对中的文本可以被用作指令, 例如对于描述图像的图像描述任务, 或作为查询输入, 例如在视觉问答中的问题。通过与之之前的技术相比, GPT-4V 在性能和泛化能力方面表现出显著增强, 体现了其卓越的智能水平。关于其在各个领域的多模态能力的全面分析详见第 4 节。

Sec. 2.3 Interleaved Image-text Inputs


Interleaved Image-text Pairs

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

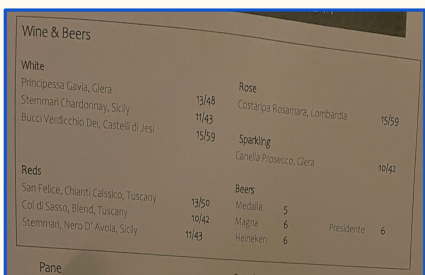


GPT-4V:
In the first receipt, you paid \$3.72 for tax. The information is located at the bottom of the receipt. In the second receipt, you paid \$42.23 for tax. The information is located near the bottom of the receipt, under the "Sales Tax" line. In the third receipt, you paid \$4.58 for tax. The information is located at the bottom of the receipt, under the "TAX" line. In total, you paid \$50.53 in tax across all three receipts.

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



Here is a photo of the menu.



GPT-4V:
According to the menu, the price for a Magna beer is \$6. Since there are two Magna beers on the table, the total cost for the beers would be \$12.

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

2.3 Interleaved Image-text Inputs

GPT-4V 的普适性进一步增强, 是因为它具备处理灵活交错的图像-文本输入的能力。交错的图像-文本输入可以是视觉为中心的, 例如多张图像配以简短的问题或指导; 也可以是以文本为中心的, 例如包含两张插入图像的长网页; 还可以是图像和文本的平衡混合。这种混合输入模式为广泛的应用提供了灵活性。例如, 它可以计算跨多张收据图像的总税款, 如图 1 所示。它还可以处理多个输入图像, 并提取查询信息。GPT-4V 还可以有效地关联交错的图像-文本输入中的信息, 例如在菜单上找到啤酒价格, 计算啤酒数量, 并返回总费用, 如图 1 所示。除了直接应用之外, 处理交错的图像-文本输入还作为上下文少样本学习和其他高级

测试时提示技术的基本组成部分，进一步提升了 GPT-4V 的普适性。我们将在下一节，第 3 节中展示这些有趣的新用法。

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

Sec. 3.1 Following Text Instructions

Following Text Instructions

Prompt:

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



GPT-4V:

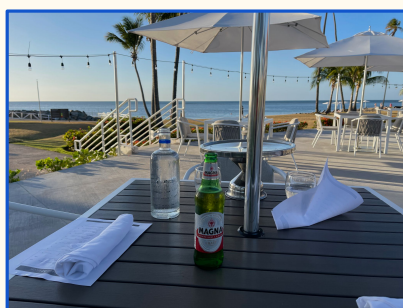
Three: Seaside dining table

Six: Dining table overlooking ocean with beer

Nine: Outdoor restaurant table with ocean view and a beer

Prompt:

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



GPT-4V:

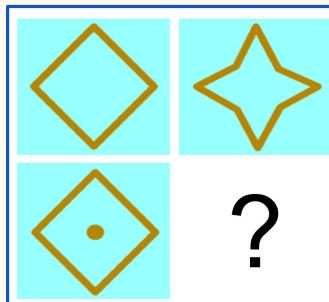
'B': Beachside bar boasts beautiful blue backdrop.

'D': Dining deck displays delightful drinks, distant docks.

'T': Tabletop terrace teeming with tropical tranquility.

Prompt:

What would the missing image look like?

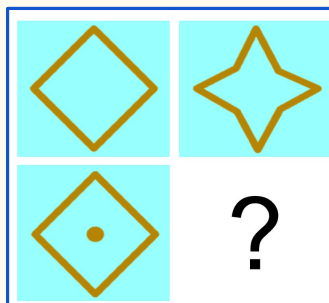


GPT-4V:

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

Prompt:

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



GPT-4V:

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

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

3.1 Following Text Instructions

GPT-4V的一个独特优势在于其普适性，在一定程度上通过其强大的文本指令理解和遵循能力实现 [96, 91, 127, 104]。指令为任意视觉语言应用场景下定义和自定义所需的输出文本提供了一种自然的方式。图 2展示了带有句子长度和使用词语约束的图像描述示例。另外，在输入方面，GPT-4V 可以通过理解详细的指令来执行具有挑战性的任务，例如通过在中间步骤上提供指令来使GPT-4V 更好地解释抽象推理问题。从指令中学习新任务展示了适应各种未见应用和任务的巨大潜力，具体在第 9 节详细说明。与最近的研究 [6, 12, 45, 150, 75, 32] 一致，本节讨论的指令主要以文本格式为主，提供了感兴趣任务的语言描述。我们将在第 3.3 节中讨论GPT-4V 遵循多模式示例指令的独特能力。

此外，我们展示了文本指令在定义GPT-4V 响应中扮演的重要角色，使用了LLM文献中采用的两种技术 [2, 149]，即 (i) “有限指示” 以使GPT-4V 以特定格式作出回应；和 (ii) “以好的表现为条件”，明确要求GPT-4V 表现良好。

有限指示。 在图 3 中，我们提示GPT-4V 阅读图像中的文本，并以特定的JSON格式返回信息。尽管GPT-4V 在从驾驶执照中提取相关信息时出现了一些错误，但响应仍受到文本指令中指定的JSON格式的约束。在第 9 节中，我们在某些应用场景中利用这种技术。

以好的表现为条件。 对于LLM而言，一个观察是LLM不想获得成功 [9]，而是想模仿训练集中各种性能质量的情况。如果用户希望模型在给定任务中成功，用户应明确要求，这已被证明对提高LLM性能有用 [149]。在LLM的背景下，我们也有类似的观察。在图 4 中，我们比较了模型对不同文本指令的响应情况。我们首先给出一个简单明确的提示：“数一下图像中的苹果的数量。” 然而，GPT-4V 错误地数了图像中总共有12个苹果。为了改善其性能，我们尝试了来自 [63] 的LLM零-shot思路链条，通过添加短语“让我们逐步思考”。尽管GPT-4V 的预测步骤大致正确，但对最终的计数并不是很有帮助，因为它仍然得出了错误答案“12个苹果”。接下来，我们将指令修改为“让我们一行一行地数苹果”，这更与视觉输入相关。虽然GPT-4V 提供的总数是正确的，但在计算第二/第三行时出现了错误。当我们进一步扩展指令为“首先数一下有多少行苹果，然后数每行的苹果，最后将它们相加得到总数”时，最终答案与正确答案偏离得更远（15个对比11个）。最后，我们按照 [149] 中“让我们以逐步的方式解决问题，以确保我们有正确的答案”来设计提示：“你是一位数物品的专家。让我们逐行数一下下面图像中的苹果，以确保我们有正确的答案。” 我们提示中的第一句话要求GPT-4V 扮演一个数物品的专家的角色，第二句话明确要求GPT-4V 取得成功。通过这个设计，GPT-4V 成功地返回了每行和总数的正确答案。在整篇论文中，我们在各种场景中采用了这种技术以获得更好的性能。

3.2 Visual Pointing and Visual Referring Prompting

指点是人与人之间互动的基本方面 [84]。为了提供一个可比较的交互渠道，研究了各种形式的“指点”来指代任意的空间兴趣区域。例如，如图 5 所示，“指点”可以表示为数值空间坐标，如盒子坐标和图像裁剪，或者是叠加在图像像素上的视觉标记，如箭头、盒子、圆圈和手绘。我们观察到，在理解直接在图像上绘制的视觉指示器方面，GPT-4V 特别强大。鉴于在图像上绘制的灵活性，这种能力可以作为未来人机交互中的一种自然方法 [85, 110, 150]。为此，我们探索了一种名为视觉指代提示的新的提示方法，人们可以通过编辑输入图像的像素空间来指定所需的目标，例如绘制视觉指示器或手写场景文本。如图 6 所示，视觉指代提示通过编辑图像像素而不是传统的文本提示来执行感兴趣的任務。例如，它可以是一个简单的基于图像上指定对象的描述，同时保持对整体图像上下文的理解，如图 6 (1,2)

所示。视觉指代提示还可以实现其他新的用例，例如将指定的对象与场景文本中的索引关联起来（图 6 (3)），或者解答靠近查询边缘或角度的问题（图 6 (4)）。第 5 节将更详细地讨论视觉指代提示。

3.3 Visual + Text Prompting

视觉引用提示可以与其他图像文本提示平滑地配合使用，呈现出一个简洁地代表感兴趣问题的细致界面。图 7 展示了两个示例，展示了 GPT-4V 的提示灵活性，特别是在整合不同输入格式和将指导与示例无缝混合的能力方面的熟练程度。GPT-4V 的泛化性和灵活性使其能够像人类一样理解多模态指令，并具备适应未知任务的前所未有的能力。

集成的多模态指令输入。 现有模型通常对交织的图像文本输入的格式有隐含约束，例如，上下文少样本学习要求图像-文本对与查询输入共享相似的格式。相比之下，GPT-4V 展现了处理任意混合图像、子图像、文本、场景文本和视觉指针的泛化能力。例如，在图 7 中展示的“添加一行”模式，可以通过在矩阵图像中用圆圈指向第一列（如子图(1)所示），或在图像内联显示子图像（如子图(2)所示）来说明。同样，在输入查询方面，可以采用将问题作为场景文本呈现的大型图像（如子图(1)所示），或以文本和子图像的混合形式发送（如子图(2)所示）。相比于 GPT-4V 的灵活性，现有的多模态模型在如何结合图像和文本以及它们可以处理的图像数量方面受到严格限制，从而对模型的能力和通用性施加了限制。

多模态示例引导指令。 除了支持更灵活的输入格式之外，与指令模式和上下文少样本学习相比，GPT-4V 的通用性还开启了更有效的任务展示方式。指令跟随技术 [96, 91, 127, 104] 最初是为自然语言处理任务提出的，它直观地关注于纯文本格式的任务指导。文本指导与视觉查询输入之间的关联较弱，可能无法提供清晰的任务演示。而上下文少样本学习 [22, 118, 6] 提供了包含图像和文本的测试示例，但这些示例必须与推理查询的格式完全一致，使其复杂且冗长。此外，上下文示例通常与指令分开使用，需要模型推断任务目标，从而损害了演示的效果。相比之下，GPT-4V 能够理解多模态指令的能力使得任务演示能够与相应的上下文示例建立关联，从而更有效地说明感兴趣的任務。例如，在图 7 中，将“在第一列找到模式”的指令基于示例演示中的关键步骤（即在(1)中圈出的模式和(2)中对应的子图）简化了学习过程，并提高了模型的性能。这种方法也反映了人类的学习过程，其中包含抽象指令与直观例子的配对。

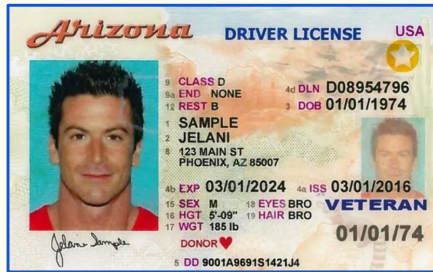
Sec. 3.1 Following Text Instructions

Constrained Prompting

Prompt:

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

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



GPT-4V:

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

Prompt:

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

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



GPT-4V:

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

Prompt:

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

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



GPT-4V:

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

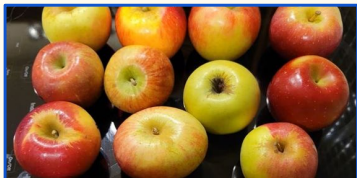
图 3: Constrained prompting to return in JSON format. Images are example IDs for samples. Red highlights the wrong answer. Check Section 3.1 for detailed discussions.

Sec. 3.1 Following Text Instructions

Condition on Good Performance

Prompt:

Count the number of apples in the image.



GPT-4V:

There are 12 apples in the image.

Prompt:

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



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

GPT-4V:

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

Step 2: Start counting the apples one by one.

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

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

Answer: There are 12 apples in the image.

Prompt:

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



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

GPT-4V:

First row: 4 apples

Second row: 3 apples

Third row: 4 apples

Total number of apples in the image: 11 apples

Prompt:

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



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

GPT-4V:

There are 3 rows of apples in the image.

The top row has 5 apples.

The middle row has 4 apples.

The bottom row has 6 apples.

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

Prompt:

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



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

GPT-4V:

Row 1: 4 apples

Row 2: 4 apples

Row 3: 3 apples

Total number of apples in the image: 11 apples

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



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

Sec. 3.2 Visual Referring Prompting

Visual Pointing and Visual Referring Prompting

(1) Prompt:

Describe the pointed region in the image.

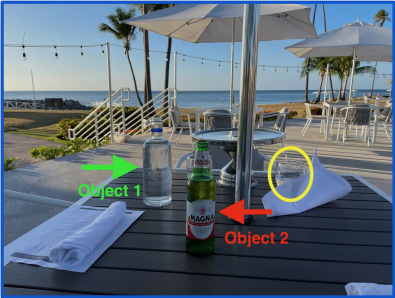


GPT-4V:

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

(3) Prompt:

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



GPT-4V:

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

(2) Prompt:

Describe the pointed region in the image.

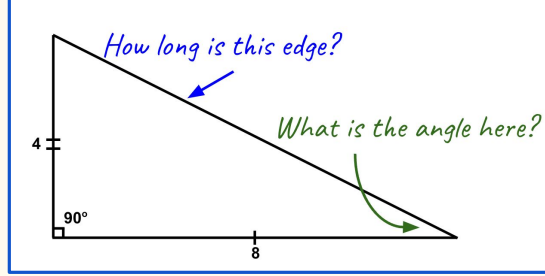
Method	Validation set								Test set							
	in.		near.		out.		overall		in.		near.		out.		overall	
	C	S	C	S	C	S	C	S	C	S	C	S	C	S	C	S
OSCAR	85.4	11.9	84.0	11.7	80.3	10.0	83.4	11.4	84.8	12.1	82.1	11.5	73.8	9.7	80.9	11.3
Human	84.4	14.3	85.0	14.3	95.7	14.0	87.1	14.2	80.6	15.0	84.6	14.7	91.6	14.2	85.3	14.6
VIVO	92.2	12.9	87.8	12.6	87.5	11.5	88.3	12.4	89.0	12.9	87.8	12.6	80.1	11.1	86.6	12.4
VinVL	103.7	13.7	95.6	13.4	83.8	11.9	94.3	13.1	98.0	13.6	95.2	13.4	78.0	11.5	92.5	13.1
UFO	103.9	14.5	95.5	13.8	83.5	12.3	94.3	13.6	98.9	14.3	94.7	13.9	77.9	12.1	92.3	13.6
mPLUG	-	-	-	-	-	-	114.8	14.8	-	-	-	-	-	-	-	-
SimVLM	113.7	-	110.9	-	115.2	-	115.2	-	113.7	-	110.9	-	115.2	-	115.2	-
LEMOM	118.0	15.4	116.3	15.1	120.2	14.5	117.3	15.0	112.8	15.2	115.5	15.1	110.1	13.7	114.3	14.9
UniversalCap	123.2	15.0	121.5	15.3	123.4	14.4	122.1	15.0	118.9	15.4	120.6	15.3	114.3	14.1	119.3	15.1
CoCa	-	-	-	-	-	-	122.4	15.5	-	-	-	-	-	-	120.6	15.5
GIT _B	100.7	13.8	97.7	13.5	89.6	12.5	96.6	13.4	-	-	-	-	-	-	-	-
GIT _L	107.7	14.9	107.8	14.5	102.5	13.7	106.9	14.4	-	-	-	-	-	-	-	-
GIT	129.8	16.3	124.1	16.0	127.1	15.7	125.5	16.0	122.4	16.2	123.9	16.0	122.0	15.7	123.4	15.9
GIT2	126.9	16.1	125.8	16.2	130.6	15.8	126.9	16.1	124.2	16.4	125.5	16.1	122.3	15.6	124.8	16.1

GPT-4V:

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

(4) Prompt:

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



GPT-4V:

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

$4^2 + 8^2 = c^2$ $16 + 64 = c^2$ $80 = c^2$ $\sqrt{80} = c$ $8.94 \approx c$
Therefore, the length of the blue arrowed edge is approximately 8.94.

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

$\tan\theta = 4/8$ $\tan\theta = 1/2$ $\theta = \tan^{-1}(1/2)$ $\theta \approx 26.57^\circ$
Therefore, the angle at the green arrow is approximately 26.57° .

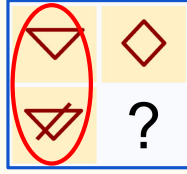
图 6: GPT-4V demonstrates the unique capability of understanding visual pointing directly overlaid on images. Based on such capability, we explore visual referring prompting that edits input image pixels (e.g., drawing visual pointers and scene texts) to prompt the task of interest. Check Section 3.2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

Sec. 3.3 Visual + Text Prompting

Multimodal Example-grounded Instruction

(1) Prompt:

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



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



with a line through it, looking like



Answer the question in the following figure.

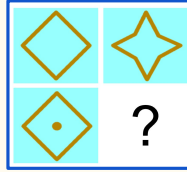


Figure out which figure of a to f that should logically replace the missing figure.

What would be the next image to complete the matrix?

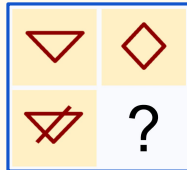


GPT-4V:

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

(2) Prompt:

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



For example,



Therefore, adds a line to image 2



will be a diamond with a line through it.

Prompt Continued

1.



2.



3.



4.

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

a.



b.



c.



d.



e.



f.



GPT-4V:

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

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

3.4 In-context Few-shot Learning

在-Large Language Models (LLMs)中观察到了另一种引人注目的新能力，即上下文Few-shot学习 [22, 36, 128, 31]。也就是说，在推理过程中，LLMs可以通过在输入查询之前添加一些上下文示例来生成期望的输出，而无需参数更新。这些示例与输入查询具有相同的格式，并且作为演示，用于说明所期望的输出。类似的能力最近也在多模态模型中观察到 [118, 6, 52, 39, 144]，其中查询输入是格式化的图像-文本对。与指令调整相辅相成，上下文学习通过在测试时间提供具有相同格式的上下文示例来“教授”模型执行新任务。我们通过一些引人入胜的示例展示了GPT-4V的上下文Few-shot学习能力。我们强调，在某些情况下，有足够数量的示例的上下文Few-shot学习变得至关重要，特别是在零射或一射指令方法不足的情况下。图 8 -10 探讨了涉及读取速度计的具有挑战性的情景。在图 8中，展示了GPT-4V在来自视频的速度计图像的截图上的零射表现。尽管经过多次尝试以零射方式提示GPT-4V，它仍然无法准确地读取图像中显示的当前速度。它生成的预测值（22/30/40 mph）与实际人工读数“约为9 mph”相差很大。即使使用了1射上下文示例，如图 9所示，使用不同的示例（图 9a）或相似的示例（图 9b），GPT-4V仍然无法准确地定位黄色指针左侧和右侧的两个数字。相比之下，图 10 表明当提供2个上下文示例时，一个与查询图像相似，另一个与之不同，GPT-4V成功预测速度读数为“约9 mph”，通过识别指针接近10 mph但还不到那个位置。

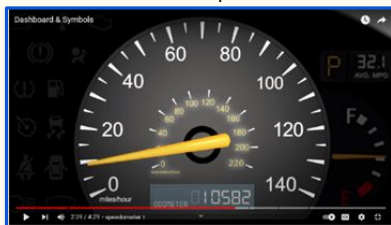
图 11 -13 说明了复杂线图的零射、1射和2射推理性能之间的比较。我们在这里探讨的示例具有很高的困难级别，因为它涉及多次跳跃推理。要回答“在图表中，哪一年的六月份平均汽油价格最高”，需要经过至少四个步骤：(i) 找到x轴上的六月份，(ii) 比较每条线在六月份的数据点，(iii) 确定具有最高值的线的颜色，和 (iv) 将颜色与顶部的对应年份相匹配的图例。在任何这些步骤中失败都会导致错误的预测。如图 11所示，即使以零射方式提示“文本指令，逐步思考”，GPT-4V仍然无法正确地将颜色与图例中的年份相关联。此外，它还受到图表中\$3.32的突出显示的汽油价格的干扰。同样，在图 12中，虽然GPT-4V在阅读图例方面有所改进（与零射相比，更正了2021和2022年的对应颜色），但它仍然坚持认为2023年是六月份平均汽油价格最高的年份，尽管这个图表只包括数据点到2023年1月17日。然而，当我们在图 13 中引入另一个上下文示例时，GPT-4V终于得出了正确答案（2022年），并提供了解释其推理过程的中间步骤，类似于上下文示例中展示的演示。这些概念验证示例生动地展示了在语境下少样本学习对于提高LMM性能的日益重要性。这种方法成为了一种可行的替代微调的选择，类似于对LLM的研究观察 [22, 36, 128, 31]。尽管在语境下少样本学习在提高LMM性能方面具有重要意义，但我们在本报告中限制其使用，以防止潜在的信息泄露或来自语境示例的不良提示。我们也留待未来研究对少样本学习的收益进行定量评估。

Sec. 3.4 In-context Few-shot Learning

Zero-shot

Prompt:

What is the read of the speed meter?



GPT-4V:

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

Zero-shot (think step-by-step)

Prompt:

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



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

GPT-4V:

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

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

Zero-shot (w/ text instructions)

Prompt:

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



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

GPT-4V:

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

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

Prompt:

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



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

GPT-4V:

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

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

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

Sec. 3.4 In-context Few-shot Learning

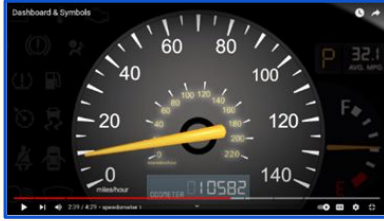
In-context Few-shot (1-shot)

Prompt:

What is the read of the speed meter?



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



GPT-4V:

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

(a)

Sec. 3.4 In-context Few-shot Learning

In-context Few-shot (1-shot)

Prompt:

What is the read of the speed meter?



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



GPT-4V:

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

(b)

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

Sec. 3.4 In-context Few-shot Learning

In-context Few-shot (2-shot)

Prompt:

What is the read of the speed meter?



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



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



GPT-4V:

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

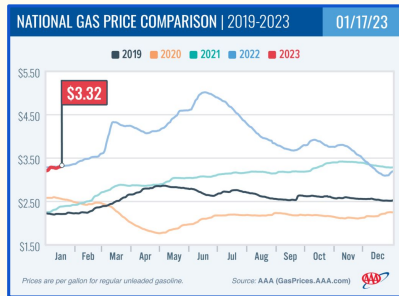
图 10: Two-shot performance under the challenging scenario of reading a speed meter. GPT-4V now can read the speed accurately. Green highlights the correct answer. Check Section 3.4 for detailed discussions.

Sec. 3.4 In-context Few-shot Learning

Zero-shot

Prompt:

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



GPT-4V:

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

Zero-shot (think step-by-step)

Prompt:

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



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

GPT-4V:

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

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

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

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

Zero-shot (w/ text instructions)

Prompt:

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



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

GPT-4V:

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

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

Prompt:

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



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

GPT-4V:

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

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

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

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

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

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

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

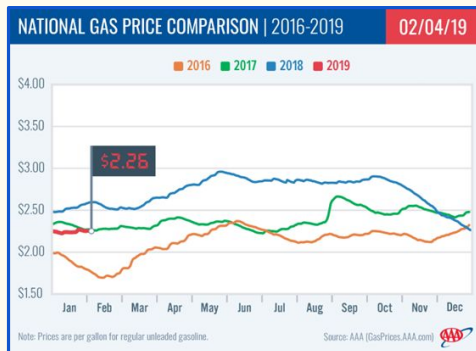
图 11: Zero-shot performance under the challenging scenario of reading a line plot. GPT-4V fails to answer the question even with different ways of ZS prompting. Red highlights the wrong answer. Check Section 3.4 for detailed discussions. 25

Sec. 3.4 In-context Few-shot Learning

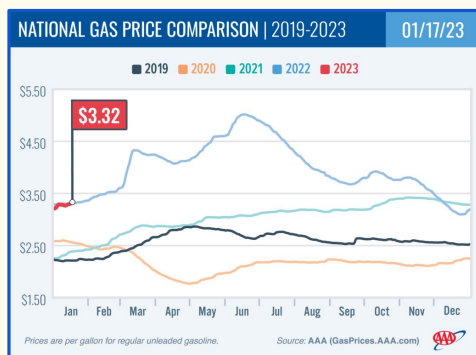
In-context Few-shot (1-shot)

Prompt:

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



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



GPT-4V:

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

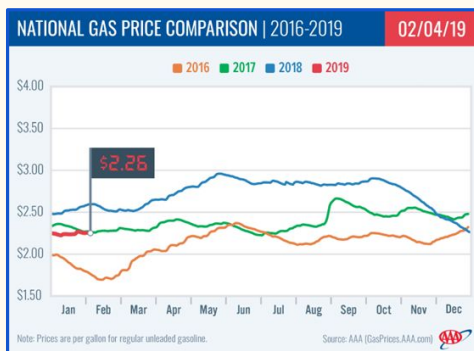
图 12: One-shot (or prompting with multimodal example instruction) performance under the challenging scenario of reading a line plot. GPT-4V still fails with 1-shot in-context example. Red highlights the wrong answer. Check Section 3.4 for detailed discussions.

Sec. 3.4 In-context Few-shot Learning

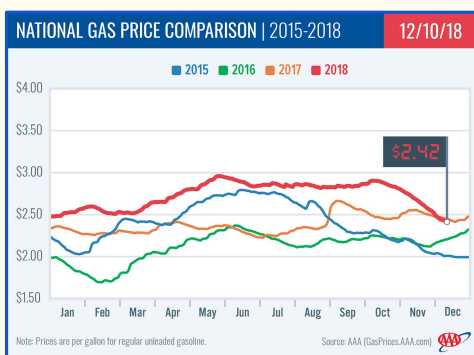
In-context Few-shot (2-shot)

Prompt:

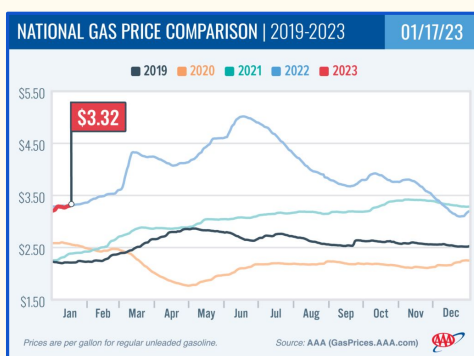
In the graph, which year has the highest average gas price for the month of June?



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



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



GPT-4V:

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

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

4 Vision-Language Capability

理解和描述视觉信息在人类认知中起着至关重要的作用。在本节中，我们将探讨如何利用 GPT-4V 来理解和解释视觉世界。我们将从研究模型在通用视觉字幕生成中的能力开始。

接下来，在第 4.2 节中，我们将探讨 GPT-4V 在更高级任务中的应用，例如空间关系分析、物体定位、物体计数和密集字幕生成。在第 4.3 节中，我们将深入研究模型的多模态知识和常识推理能力，并探讨模型是否能够理解不同类型信息之间的上下文和关系。

此外，在第 4.4 节中，我们将评估模型从各种来源中提取和分析信息的能力，包括场景文本、表格、图表和文档。在第 4.5 节中，我们将探索 GPT-4V 在多语言情境中理解和生成描述的能力。最后，在第 4.6 节中，我们将研究模型对视觉信息的编码能力，探索其在所选示例中执行任务的能力。

4.1 Image Description on Diverse Domains

我们通过提供一个“图像-文字对”作为输入来评估模型的能力和泛化性。我们要求 GPT-4V 生成自然语言描述，涵盖以下几个主题。

名人识别。 人类外貌识别 [46, 76] 是一个具有挑战性的任务，因为它具有内在的变异性。为了评估 GPT-4V 辨认和描述名人的能力，我们进行了一个实验，通过提供一个文本提示“描述图像”，连同一个人的图像作为输入。在图 14 的顶行，我们观察到 GPT-4V 可以准确识别这八位名人，尽管他们拥有不同的背景和领域。此外，当我们提供一个更具体的询问，“图像中的人物是谁，正在做什么？”如图 14 的底行所示，GPT-4V 可以理解到美国现任总统正在 2023 年的 G7 峰会上发表演讲。这展示了模型的泛化能力，以及处理新颖场景（例如 2023 年的 G7 峰会）的能力，即使这些情况不在其训练数据中。

地标识别。 地标的外貌因视点的变化、光照条件、遮挡和季节变化等因素产生了显著的变异。在这些变化条件下识别地标要求模型具备良好的泛化性，能够处理广泛的视觉外貌 [145, 4]。在实验中，我们使用一个简明的文本提示“描述图像中的地标”来测试模型的能力。如图 15-16 所示，GPT-4V 为每个测试图像生成了准确且开放的描述。例如，它可以准确识别出位于华盛顿州西雅图的太空针塔，并理解该塔是为了 1962 年的世界博览会而建，现已成为城市的象征。我们对其他测试照片也有类似的观察。生成的描述超越了简单的标签或通用词语，提供了生动而详细的叙述，捕捉到了地标的本质。

食物识别。 识别食物或菜肴是一项有趣的任务 [19, 90]，但由于外貌上的广泛变化以及其他物体或重叠成分引起的潜在遮挡，可能会具有挑战性。在我们的实验中，我们使用一个简明的文本提示，询问系统“描述菜肴的名称”，进行测试。图 17 展示了 GPT-4V 对各种菜肴的准确识别。此外，GPT-4V 可以有效地捕捉图像中的复杂细节，识别出菜肴中的特定成分、装饰品或烹饪技巧。

医学图像理解。 医学图像，如 X 光和 CT 扫描，可能会因患者群体和成像设备而存在较大变异性。此外，解读这些图像的视觉内容需要专业知识。通过提供提示“描述图像”，我们评估 GPT-4V 的表现。结果显示，GPT-4V 可以识别出给定 X 光中的牙齿和颌骨。此外，当我们提示“这张 X 光片上有需要拔除的智齿吗？”时，GPT-4V 可以根据视觉上下文进行推理，并解释说颌骨底部左右两侧的智齿没有完全长出牙床，这可能是需要拔除的原因。我们还对其他医学图像进行了测试，如图 19 所示。对于这些实验，我们使用了提示语“出了什么问题？”或“看看 CT 扫描，告诉我有什么问题。”观察结果表明，GPT-4V 可以识别出一般情况，如 Jones 骨折。它还可以根据肺部 CT 扫描指出潜在问题。这些实验展示了 GPT-4V 对医学图像的基本理解能力。我们在第 9.3 节中讨论了 GPT-4V 在医学领域的应用。

Logo识别

我们检验了GPT-4V在Logo识别方面的能力。在图 20 中，我们通过给出文本提示“描述图像”来开始实验。GPT-4V准确地识别出图像中所示的三个Logo。然后，我们提出了一个更具体的问题：“详细描述这些Logo”，GPT-4V分别提供了有关每个Logo的设计、风格和表征的详细描述。我们将评估扩展到一个更具挑战性的实际场景中，如图 21 所示，我们实验了可能部分遮挡、扭曲或位于杂乱背景中的Logo。我们使用文本提示“详细描述图像和Logo”进行*in-the-wild*实验。如图 21 所示，GPT-4V在理解困难场景中的Logo方面表现出很强的能力。值得注意的是，GPT-4V也可以对新的或新出现的Logo和图标提供描述，例如最近发布的Microsoft 365 Copilot。

场景理解

场景理解 [72, 29, 147] 是计算机视觉中的一个重要任务。我们通过提供一个简单的查询“描述图像”，来检验该模型的能力。在图 22 中，GPT-4V能够描述道路、车辆的位置和颜色。它还可以读取标示并注意到该道路的限速。

逆向例子

我们根据 [74] 随机选择逆向例子进行实验。在图 23 中，我们观察到当面临误导性的问题或指令时，GPT-4V可以正确地描述图像的内容。

Sec. 4.1 Image Description on Diverse Domains

Celebrity Recognition and Description

Prompt:

Describe the image

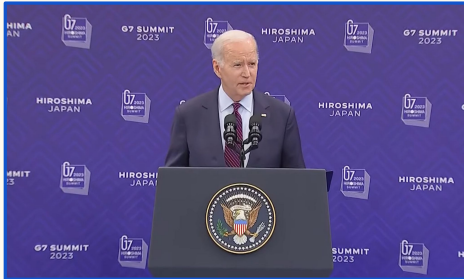


GPT-4V:

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

Prompt:

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

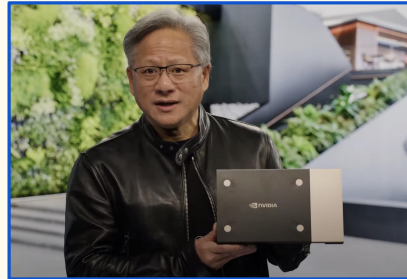


GPT-4V:

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

Prompt:

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



GPT-4V:

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

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

Sec. 4.1 Image Description on Diverse Domains

Landmark Recognition and Description

Prompt:

Describe the landmark in the image.



GPT-4V:

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

Prompt:

Describe the landmark in the image.



GPT-4V:

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

Prompt:

Describe the landmark in the image.

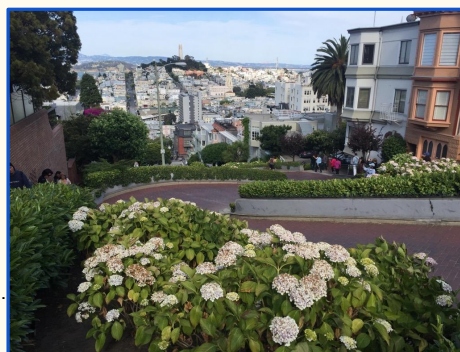


GPT-4V:

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

Prompt:

Describe the landmark in the image.



GPT-4V:

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.

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

Sec. 4.1 Image Description on Diverse Domains

Landmark Recognition and Description

Prompt:

Describe the landmark in the image.



GPT-4V:

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

Prompt:

Describe the landmark in the image.



GPT-4V:

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

Prompt:

Describe the landmark in the image.



GPT-4V:

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

Prompt:

Describe the landmark in the image.



GPT-4V:

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

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

Sec. 4.1 Image Description on Diverse Domains

Food Recognition and Description

Prompt:

Describe the name of the dish.



GPT-4V:

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

Prompt:

Describe the name of the dish.



GPT-4V:

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

Prompt:

Describe the name of the dish.



GPT-4V:

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

Prompt:

Describe the name of the dish..



GPT-4V:

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

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

Sec. 4.1 Image Description on Diverse Domains

Medical Image Understanding

Prompt:

Describe the image



GPT-4V:

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

Prompt:

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



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

GPT-4V:

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

Prompt:

Which wisdom tooth is hardest to remove?



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

GPT-4V:

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

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

Sec. 4.1 Image Description on Diverse Domains

Medical Image Understanding

Prompt:

What's wrong?



GPT-4V:

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

Prompt:

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



GPT-4V:

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

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

Sec. 4.1 Image Description on Diverse Domains

Logo Recognition and Description

Prompt:

Describe the image



GPT-4V:

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

Prompt:

Describe the logos in details



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

GPT-4V:

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

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

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

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

Sec. 4.1 Image Description on Diverse Domains

In-the-Wild Logo Recognition and Description

Prompt:

Describe both the image and logo in details



GPT-4V:

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

Prompt:

Describe both the image and logo in details



GPT-4V:

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

Prompt:

Describe both the image and logo in details



GPT-4V:

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

Prompt:

Describe both the image and logo in details



GPT-4V:

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

Prompt:

Describe both the image and logo in details

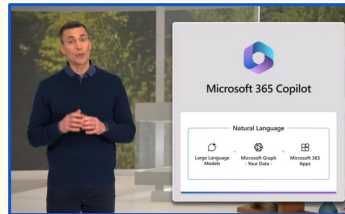


GPT-4V:

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

Prompt:

Describe both the image and logo in details



GPT-4V:

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

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

Sec. 4.1 Image Description on Diverse Domains

Scene Understanding

Prompt:

Describe the image

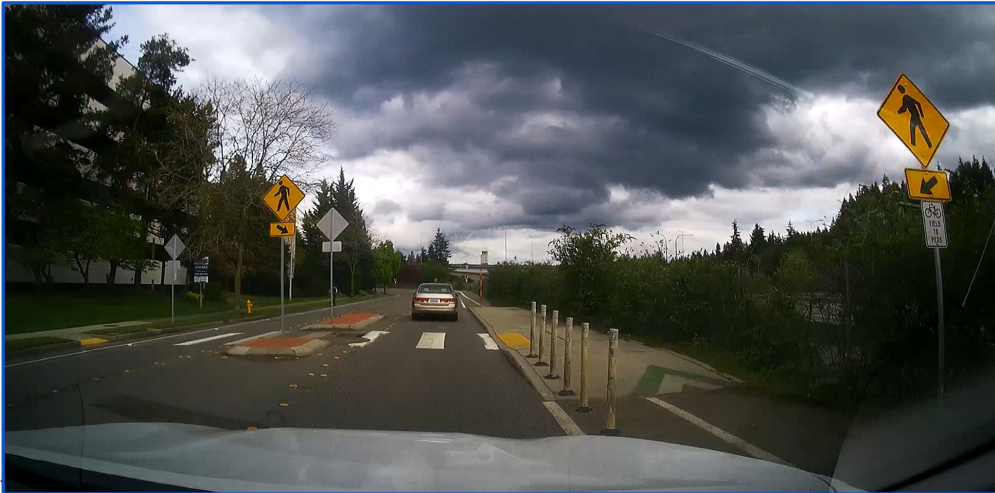


GPT-4V:

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

Prompt:

Describe the image



GPT-4V:

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

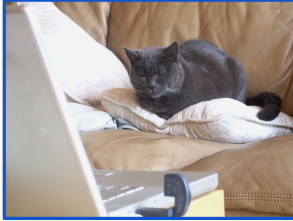
图 22: Results on scene understanding. GPT-4V is able to provide a detailed description regarding the scenes and objects. Check Section 4.1 for detailed discussions.

Sec. 4.1 Image Description on Diverse Domains

Counterfactual Examples

Prompt:

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



GPT-4V:

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

Prompt:

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



GPT-4V:

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

Prompt:

Can you spot the pink elephant playing in the corner?



GPT-4V:

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

Prompt:

Can you find the baseball player wearing a blue shirt?



GPT-4V:

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

Prompt:

Try to locate a large waterfall behind the trees.



GPT-4V:

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

Prompt:

Can you 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.

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

4.2 Object Localization, Counting, and Dense Captioning

空间关系理解

理解图像中人与物体的空间关系是视觉智能的重要方面 [58, 13]。在图 24 中，本模型展示了在此方面的出色能力。它可以识别图像中飞盘与人之间的空间关系。它还能够认识到人与汽车之间的空间关系，并指出摄像机的角度可能会影响他们的视觉大小。

物体计数

图 25 突出展示了本模型在物体计数方面的能力。在我们的实验中，我们使用文本提示“数一下图像中的X的数量”来评估其性能。结果表明，本模型可以成功地计算出图像中出现的物体的数量，如苹果、橙子和人。然而，当物体被遮挡或场景混乱时会遇到挑战，可能导致计数过程中的错误。在图 25 左下方，本模型识别出了12个人，但正确答案应为11个。这可能是由于本实验中使用的文本提示有限，需要进一步研究提示技术。

物体定位

物体定位 [146, 72, 48] 是计算机视觉领域的一个基本挑战。在我们的初步实验中，我们使用简单的文本提示“使用边界框定位图像中的每个人”。物体定位实验的初步结果如图 26 所示。结果表明，本模型能够以文本形式生成边界框坐标。然而，需要注意的是，生成的边界框坐标并不准确。当场景或背景相对简单且较少混乱时，我们可以观察到具有潜力的定位结果。需要进一步的提示技术来提升复杂和拥挤环境中的物体定位性能。

密集字幕生成

密集字幕生成 [59, 79] 涉及为给定图像中的每个感兴趣区域生成详细描述。在视觉语言领域中，这一先进任务通常需要一个复杂的系统，集成多个专家，如目标检测器、名人识别模型和图像字幕生成模型。为了探索本模型在密集字幕生成方面的能力，我们使用了如图 27 所示的一种指示性文本提示。结果是非常令人鼓舞的，本模型成功地定位并识别了图像中的个体，然后为每个科学家提供简洁的描述。

Sec. 4.2 Object Localization, Counting, and Dense Captioning

Spatial Relationship Understanding

Prompt:

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



GPT-4V:

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

Prompt:

Does the person bigger than the car?



GPT-4V:

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

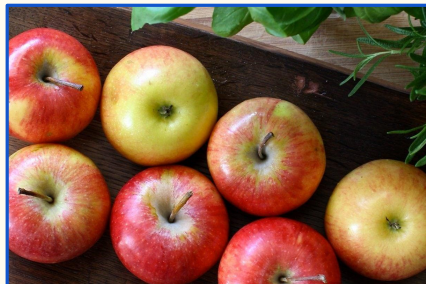
图 24: Results on spatial relationship understanding. GPT-4V recognizes the spatial relationship between the objects in the images. Check Section 4.2 for detailed discussions.

Sec. 4.2 Object Localization, Counting, and Dense Captioning

Object Counting

Prompt:

Count the number of apples in the image.



GPT-4V:

There are 7 apples in the image.

Prompt:

Count the number of apples in the image.



GPT-4V:

There are 16 apples in the image.

Prompt:

Count the number of oranges in the image.



GPT-4V:

There are 3 oranges in the image.

Prompt:

Count the number of people in the image.



GPT-4V:

There are 3 people in the image.

Prompt:

Count the number of people in the image.



GPT-4V:

There are 4 people in the image.

Prompt:

Count the number of people in the image.



GPT-4V:

There are 12 people in the image.

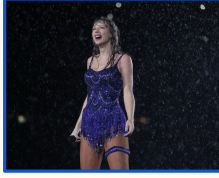
图 25: Results on object counting. GPT-4V is able to determine the quantity of the specified objects the image. Red highlights the wrong answer. Check Section 4.2 for detailed discussions.

Sec. 4.2 Object Localization, Counting, and Dense Captioning

Object Localization

Prompt:

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



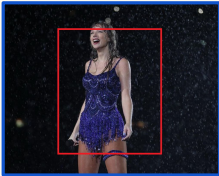
GPT-4V:

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

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

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

Visualization of Detected Bbox:



Prompt:

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



GPT-4V:

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

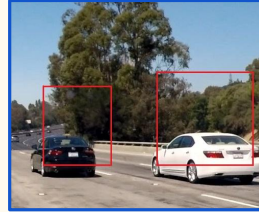
Bounding box for the black car:

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

Bounding box for the white car:

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

Visualization of Detected Bbox:



Prompt:

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



GPT-4V:

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

Bounding boxes for each person:

- Person 1 (leftmost): (x1, y1) = (22, 108), (x2, y2) = (170, 300)
- Person 2: (x1, y1) = (200, 80), (x2, y2) = (290, 270)
- Person 3: (x1, y1) = (350, 70), (x2, y2) = (460, 280)
- Person 4: (x1, y1) = (540, 80), (x2, y2) = (640, 270)
- Person 5 (rightmost): (x1, y1) = (760, 100), (x2, y2) = (920, 290)

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

Visualization of Detected Bbox:

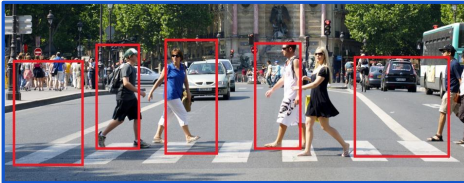


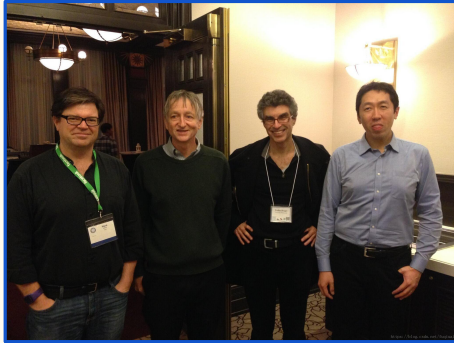
图 26: Results on object localization. GPT-4V is able to generate and approximate the bounding box coordinates for the specified objects in the image. When providing a simple text prompt only, the model may encounter challenges when dealing with more complex scenarios like object occlusions and cluttered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

Dense Captioning

Prompt:

Please follow the instructions

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



GPT-4V:

1. The size of the input image is 1600 x 1067 pixels.
2. Bounding box 1: $(x1, y1) = (120, 250), (x2, y2) = (420, 950)$ - Person 1
Bounding box 2: $(x1, y1) = (500, 240), (x2, y2) = (800, 940)$ - Person 2
Bounding box 3: $(x1, y1) = (880, 230), (x2, y2) = (1180, 930)$ - Person 3
Bounding box 4: $(x1, y1) = (1260, 220), (x2, y2) = (1560, 920)$ - Person 4
3. Person 1: Yann LeCun
Person 2: Geoffrey Hinton
Person 3: Yoshua Bengio
Person 4: Andrew Ng
4. Caption for bounding box 1: Yann LeCun, computer scientist and one of the founding fathers of artificial intelligence.
Caption for bounding box 2: Geoffrey Hinton, computer scientist and cognitive psychologist, known for his work on artificial neural networks.
Caption for bounding box 3: Yoshua Bengio, computer scientist and co-founder of the Montreal Institute for Learning Algorithms.
Caption for bounding box 4: Andrew Ng, computer scientist and entrepreneur, known for his work on deep learning and artificial intelligence.

Visualization of Dense Captioning:

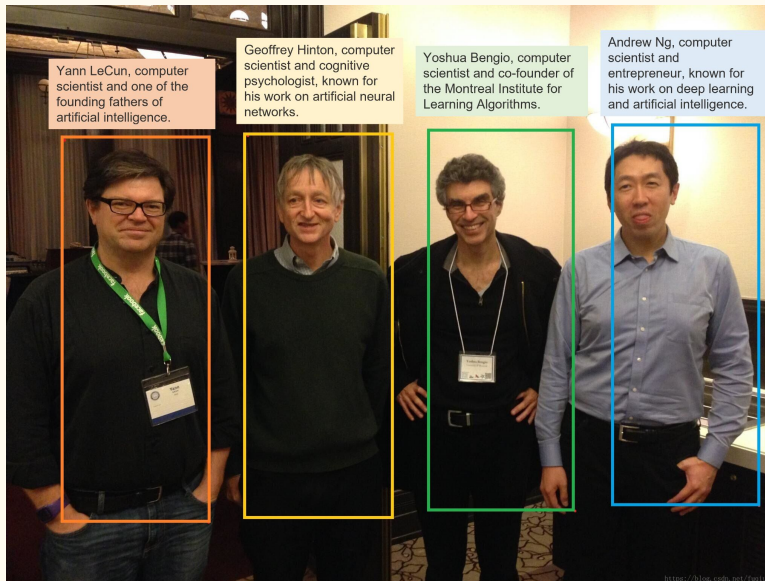


图 27: Results on dense captioning. GPT-4V follows the text prompt and successfully generates dense captions for the input image. Check Section 4.2 for detailed discussions.

4.3 Multimodal Knowledge and Commonsense

笑话和表情包

笑话和表情包经常涉及特定事件、流行文化或互联网趋势。理解这些参考需要对相关背景和文化知识有所了解。抓住视觉元素、它们与文本的关系和预期的幽默效果可能是一个复杂的任务 [94]。此外，表情包通常是用户生成的，使它们充满多样性和不断扩大。为了评估GPT-4V在这个领域的的能力，我们向GPT-4V输入了一对表情包和文本提示。示例文本提示包括“你能解释这个表情包吗？”和“图像有什么搞笑的地方？”图 28显示了示例结果。我们观察到，GPT-4V在从视觉和文本模态中收集信息，并理解嵌入在表情包中的幽默时具有显著的能力。

科学与知识

我们进一步研究了GPT-4V在涉及科学知识推理的任务中的能力 [80]。我们通过提供一个文本提示问题和相应的图像来进行实验。这些问题涵盖了广泛的主题，包括地理、物理、生物和地球科学。在图 29-31中，我们观察到，在视觉背景的基础上，GPT-4V能够正确回答科学问题。例如，在图 29的最后一行中，GPT-4V分别识别了样本A和样本B的平均粒子速度。通过考虑粒子速度、动能和温度之间的关系，GPT-4V正确回答了问题。另一个例子，在图 30的最后一行中，GPT-4V根据图中呈现的箭头确定了特定食物链中的生产者。此外，在图 31中展示的情况下，当我们提供了更具体的提示，例如“假设你是一位老师，请使用图来解释X”，我们观察到生成的答案采用了教学格式，并逐步解释了主题。

多模态常识

在图 32中，我们评估了GPT-4V在多模态常识推理方面的能力 [141, 49]。在我们的实验中，我们观察到，GPT-4V有效地利用图像中呈现的边界框作为视觉提示（例如，[person1]和[person2]）来识别个体执行的动作。如图 32中的第二个例子所示，基于[person1]和[person2]身着的礼服以及场景中的花卉装饰，可以推断他们正在参加婚礼仪式。此外，当我们提供一个更具体的输入提示，例如“假设你是一位侦探，你可以从视觉线索推断出什么？”，GPT-4V展示了辨别图像中许多微妙视觉线索的能力，并提供了一系列可能的假设。

Sec. 4.3 Multimodal Knowledge and Commonsense

Joke and Meme

Prompt:

Can you explain this meme?

Me: I'll do it at 8

Time: 8:05

Me: looks like I gotta wait till 9 now

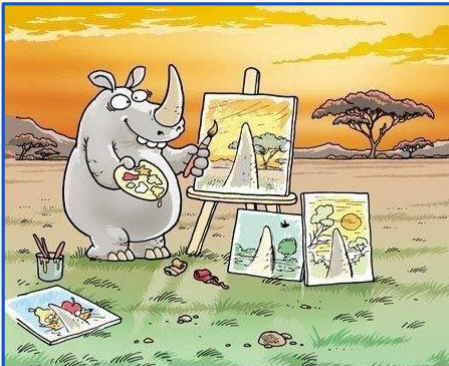


GPT-4V:

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

Prompt:

What is funny about this image?



**You see the world
as you are!**

GPT-4V:

There are a couple of things that people might find funny about this image:

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

Prompt:

Can you explain this meme?

My Handwriting In Exams

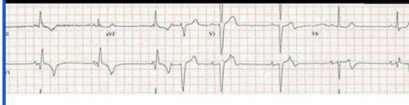
First two pages

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

Middle pages

Also with causing the melt down of
glaciers and making the ocean water
polluted and that makes the

Last two pages



GPT-4V:

This meme highlights the deterioration of a person's handwriting during an exam. The first two pages show neat and legible handwriting, the middle page shows slightly messier handwriting, and the last two pages show almost illegible handwriting.

The meme is poking fun at how, as an exam progresses, a person's handwriting tends to get worse as they become more tired and rushed to finish the exam. It also implies that the person's focus and attention to detail decrease towards the end of the exam.

Prompt:

What is funny about this image?



GPT-4V:

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

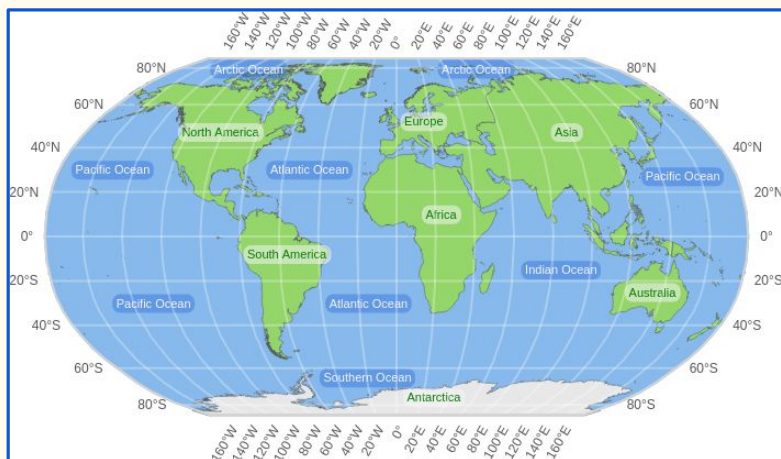
图 28: Results on joke and meme understanding. GPT-4V demonstrates the impressive capability to comprehend the humor embedded within memes. Check Section 4.3 for detailed discussions.

Sec. 4.3 Multimodal Knowledge and Commonsense

Science and Knowledge

Prompt:

Which of these oceans does the prime meridian intersect?



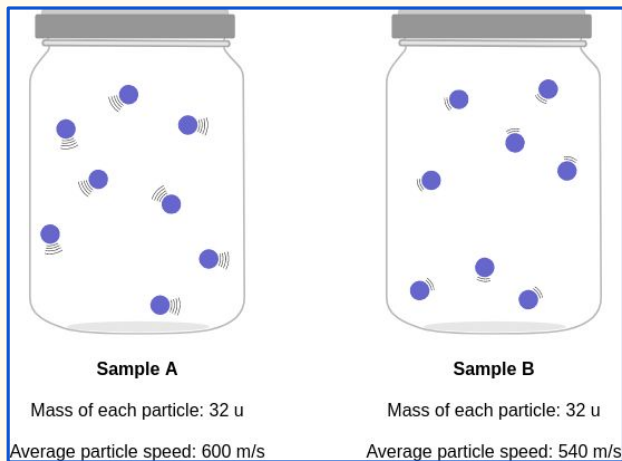
GPT-4V:

The prime meridian intersects the following oceans:

- Atlantic Ocean
- Southern Ocean
- Arctic Ocean

Prompt:

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



GPT-4V:

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

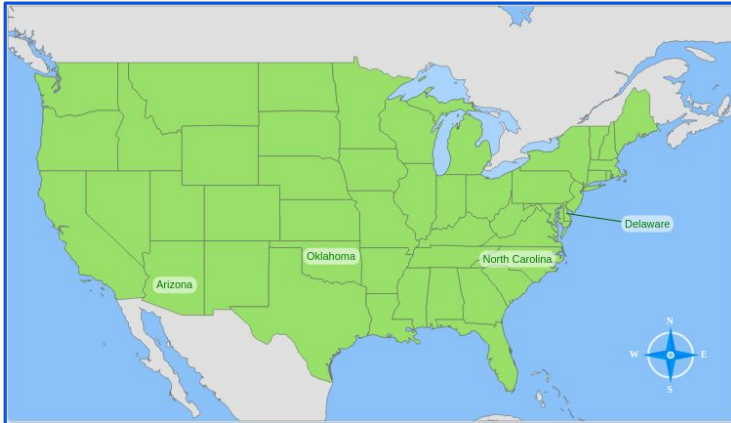
图 29: Results on answering science questions. GPT-4V can understand the question textually and visually, and gather necessary information to answer the question. Check Section 4.3 for detailed discussions.

Sec. 4.3 Multimodal Knowledge and Commonsense

Science and Knowledge

Prompt:

Which of these states is farthest north?

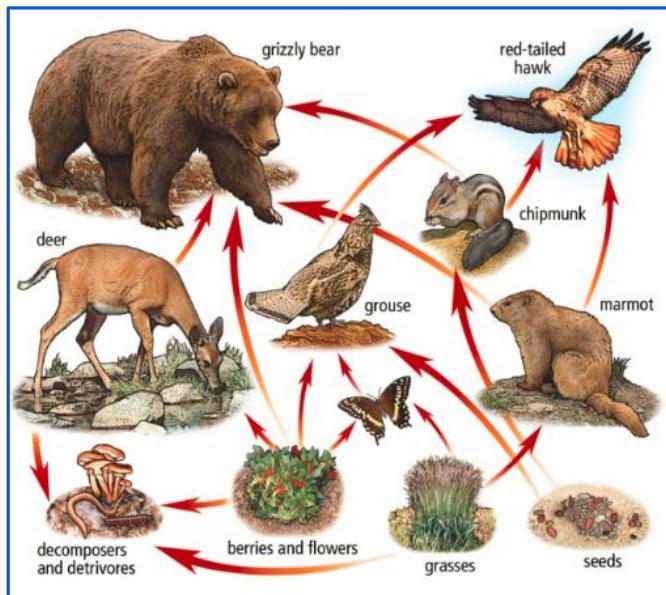


GPT-4V:

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

Prompt:

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



GPT-4V:

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

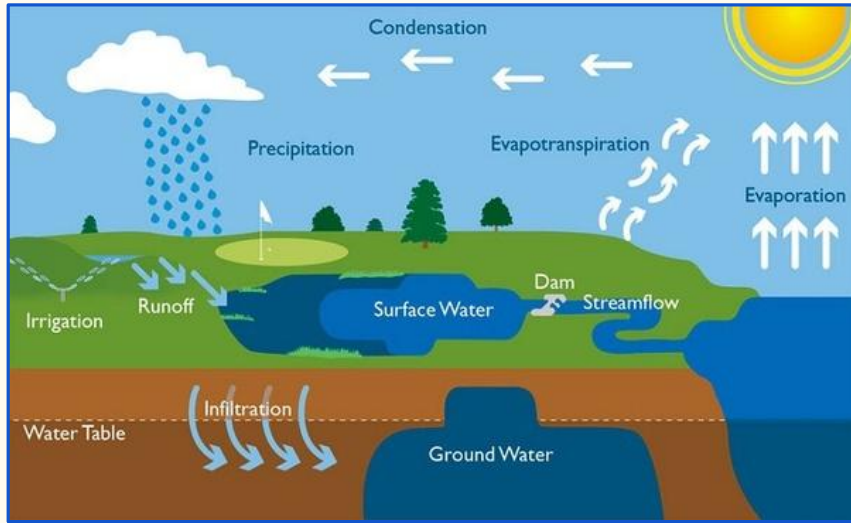
图 30: Results on answering science questions. GPT-4V can understand the question textually and visually, and gather necessary information to answer the question. Check Section 4.3 for detailed discussions.

Sec. 4.3 Multimodal Knowledge and Commonsense

Science and Knowledge

Prompt:

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



GPT-4V:

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

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

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

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

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

Sec. 4.3 Multimodal Knowledge and Commonsense

Multimodal Commonsense

Prompt:

What is [person3] doing?

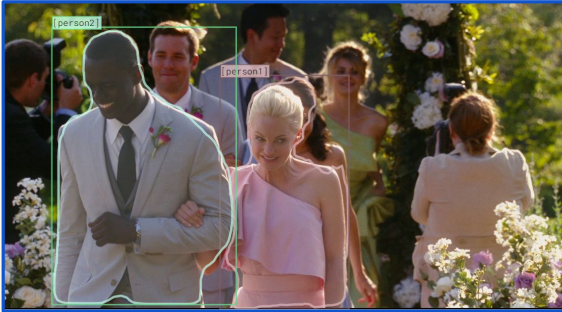


GPT-4V:

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

Prompt:

What are [person1] and [person2] doing?

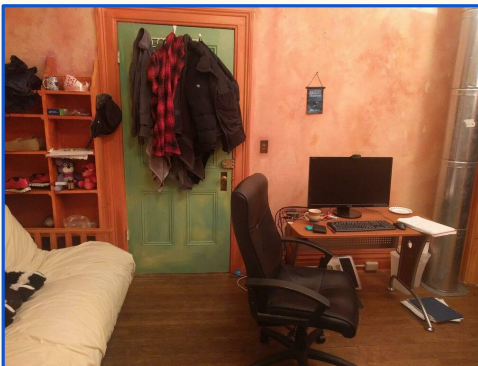


GPT-4V:

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

Prompt:

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



GPT-4V:

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

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

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

4.4 Scene Text, Table, Chart, and Document Reasoning

场景文本识别。 在图像中阅读和理解场景文本是视觉语言中的一个重要任务 [111–113, 16]。在我们的实验中，我们研究了 GPT-4V 识别场景文本的能力，利用了输入提示语“图像中的所有场景文本是什么？”图 33 展示了示例结果。我们观察到，GPT-4V 准确地识别了各种场景中的文本，包括手写和打印文本。在第 4.5 节中，我们展示了更多关于多语言场景的结果。

视觉数学推理。 在图 34 中，GPT-4V 展示了解决视觉数学问题的能力。在我们的实验中，我们观察到 GPT-4V 能够从图像中提取关键信息。例如，在图 34 中，GPT-4V 正确地识别出了直角三角形（或正交三角形）的存在，并确定 AB 为 4 个单位，BC 为 3 个单位。此外，我们注意到 GPT-4V 倾向于以一种结构化的方式呈现解决方案，逐步解决问题，从而展示了其提供清晰解释的能力。

图表理解和推理。 我们进一步研究了 GPT-4V 在图表理解和推理方面的能力。图 35-37 展示了示例结果。在我们的初步探索中，GPT-4V 表现出了提供图表详细描述的能力。例如，在图 35 中，该模型正确地解释了从开始到结束的提案过程。在图 36 中，该模型不仅理解了给定流程图中的程序，还将详细信息转化为 Python 代码。在图 37 的底部一行中，GPT-4V 对 x 轴和 y 轴都有清晰的理解，解释了图表中呈现的关键洞察。此外，在我们的实验中，我们观察到 GPT-4V 可以根据图表回答问题。在图 37 的顶部一行中，GPT-4V 正确计算了不包括 Ford F150 在内的平均总加油成本。

表格理解和推理。 在图 38 中，我们展示了对表格理解和推理的初步研究结果。与图表实验的发现类似，GPT-4V 在理解表格细节以及推理和准确回答相关问题方面展现出了有希望的结果。

文档理解。 图 39 展示了在各种类型文档（如平面图、海报和考卷）上的结果。我们观察到 GPT-4V 展现了对文档的理解并提供了合理的回答。例如，在平面图中，它准确识别出第二间卧室的浴室位置。它还识别出中国菜“热干面”，并通过场景文本将其与武汉城关联起来。此外，GPT-4V 能够阅读考卷。它准确地使用 Markdown 重构了表格，并填入了正确的答案。我们在第 4.6 节中展示了更多关于其编码能力的探索。

在图 40 中，我们通过提供一个由多个页面组成的技术报告 [119] 来探讨更具挑战性的案例。在有限的探索中，GPT-4V 展示出了令人印象深刻的结果。它通过考虑多个页面的上下文正确描述了主要思想和他们提出的方法。但是，它偶尔可能会遗漏一些实现细节。请注意，数据集应包含 $1196+665=1861$ 个示例，提取特征应包括方向梯度直方图（HOG）。我们相信，探索更先进的提示技术，如逐步思考或采用上下文少样本方法，可以进一步提升该模型的性能。

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

Scene Text Recognition

Prompt:

What are all the scene text in the image?

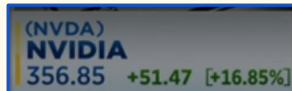


GPT-4V:

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

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image are:

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

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image are:

- "Royal"
- "London"

Prompt:

What are all the scene text in the image?



GPT-4V:

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

Prompt:

What are all the scene text in the image?



GPT-4V:

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

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image are:

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

Prompt:

What are all the scene text in the image?



GPT-4V:

There are three texts in the image:

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

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image are:

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

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image are:

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

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

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

Visual Math

Prompt:

What is x in the equation?

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

GPT-4V:

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

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

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

$$x + 3 = \pm 2$$

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

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

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

Prompt:

What is the value of d in the last equation?

Solve the following equations:

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

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

GPT-4V:

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

$$4x + 11 = 14$$

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

$$4x = 3$$

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

$$x = 3/4$$

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

$$-4 = 4d - 9$$

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

$$5 = 4d$$

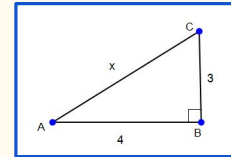
Finally, we can divide both sides by 4 to isolate the d variable on the right side of the equation. This gives us:

$$d = 5/4$$

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

Prompt:

What is the value of x in the image?



GPT-4V:

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

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

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

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

$$9 + 16 = c^2$$

$$25 = c^2$$

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

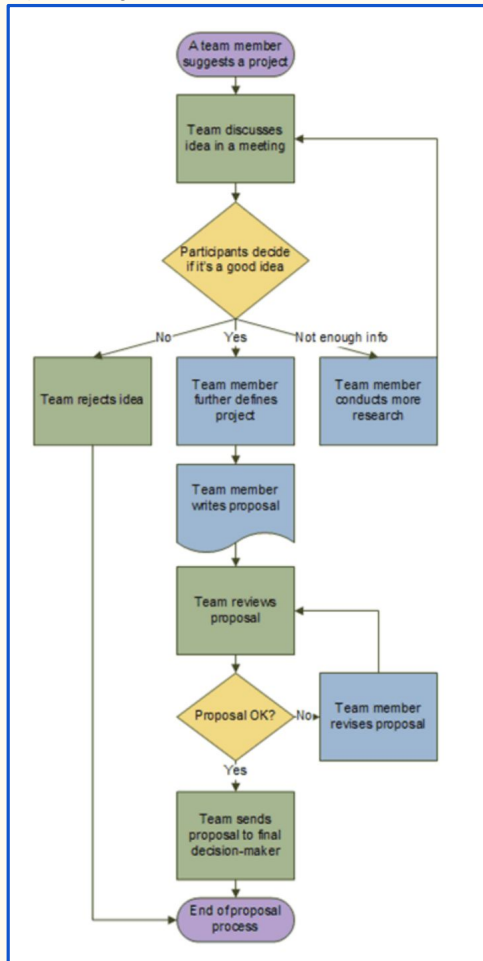
图 34: Results on visual math reasoning. GPT-4V is able to comprehend and solve visual math problems with a well-structured solution. Check Section 4.4 for detailed discussions.

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

Flow Chart Understanding

Prompt:

Explain the figure



GPT-4V:

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

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

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

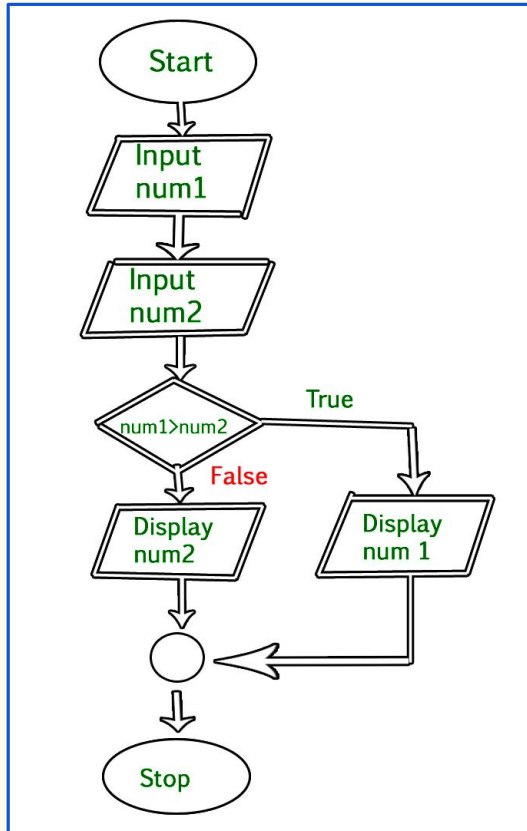
图 35: Results on flow chart understanding. GPT-4V correctly describes the proposal process in details. Check Section 4.4 for detailed discussions.

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

Flow Chart Understanding

Prompt:

Can you translate the flowchart to a python code?



GPT-4V:

Sure, here's the Python code:

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

Explanation:

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

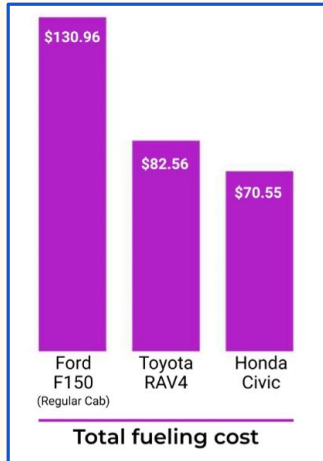
图 36: Results on flow chart understanding. GPT-4V is able to translate the flow chart to a python code. Check Section 4.4 for detailed discussions.

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

Chart Understanding and Reasoning

Prompt:

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



GPT-4V:

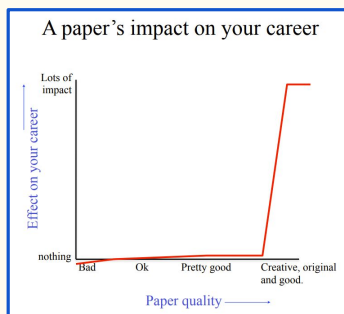
To find the average total fueling cost excluding the Ford F150, we need to add the total fueling costs of the Toyota RAV4 and Honda Civic and divide by 2 (since there are 2 vehicles in the calculation).

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

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

Prompt:

Explain the figure.



GPT-4V:

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

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

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

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

Table Understanding and Reasoning

Prompt:

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

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

GPT-4V:

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

Prompt:

Which country has the highest profit margin?

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

GPT-4V:

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

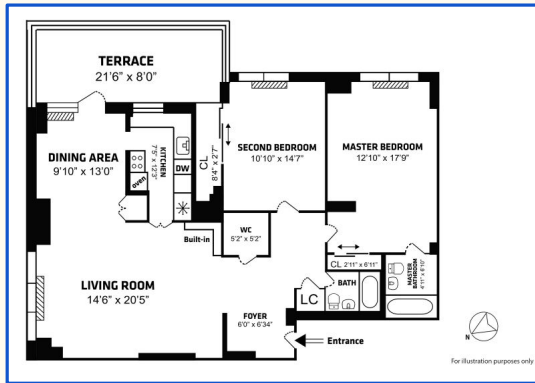
图 38: We observe GPT-4V can understand the details in the table, and answer related questions. Check Section 4.4 for detailed discussions.

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

Document Understanding

Prompt:

Describe the image.

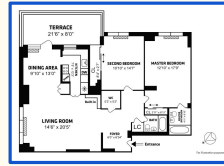


GPT-4V:

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

Prompt:

Where is the bathroom for the second bedroom?



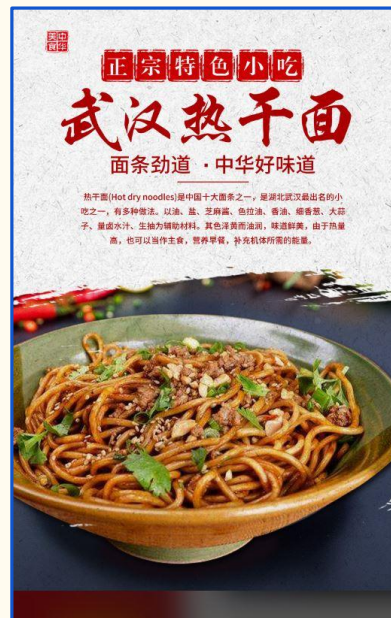
Note: Same query image as shown on left, resized to save space

GPT-4V:

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

Prompt:

Describe the poster. Which city is this dish from?



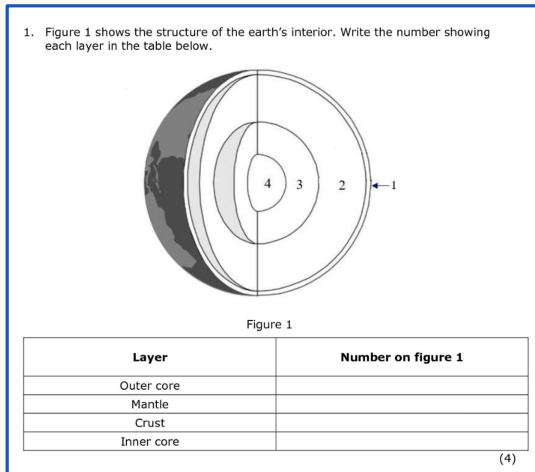
GPT-4V:

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

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

Prompt:

Answer the question in the image.



GPT-4V:

Layer | Number on figure 1

--- | ---

Outer core | 3

Mantle | 2

Crust | 1

Inner core | 4

Visualization of the generated results:

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

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

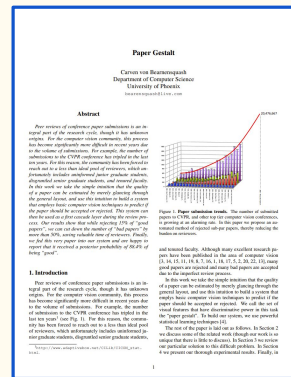
Sec. 4.4 Scene Text, Table, Chart, and Document Reasoning

Document Understanding

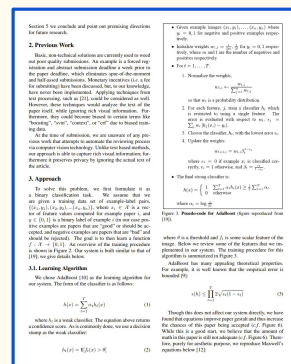
Prompt:

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

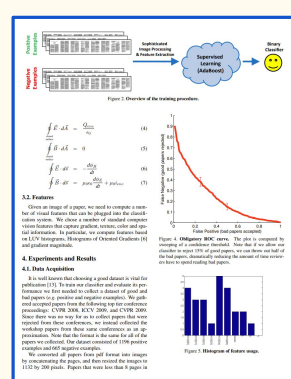
Page 1



Page 2

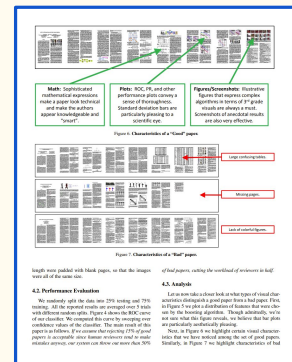


Page 3



Prompt Continued

Page 4



Page 5



GPT-4V:

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

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

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

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

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

图 40: Results on document understanding. GPT-4V reads a multi-page technical report, understands the content in each section, and provides a summary of the contribution of this technical report. Red highlights the wrong answer. Check Section 4.4 for detailed discussions.

4.5 Multilingual Multimodal Understanding

我们评估 GPT-4V 在理解多种语言和模态方面的能力。首先，我们通过评估没有场景文字的自然图像来探索这种能力，如图 41 所示。在图的第一行中，我们分别提供了中文、法语和捷克语的输入文本提示“描述图像”。GPT-4V 可以识别不同语言的输入文本提示，并生成相应语言的正确图像描述。在图 41 的第二行中，我们提供了英文的输入文本提示，并指定输出语言。GPT-4V 遵循指示，并生成所需语言的正确描述。在图 41 的底行中，我们提供了一个西班牙文的输入提示，让 GPT-4V 生成 20 种不同语言的图像描述。我们观察到 GPT-4V 可以处理不同语言的输入和输出文本。

此外，我们探讨了涉及多语言场景文字识别的场景，输入图像可能包含不同语言的场景文字。如图 42 所示，GPT-4V 能够正确识别和理解不同场景中的文字。如图 43 的前两行所示，我们观察到 GPT-4V 能够识别场景文字，并将其翻译成不同的语言。在图 43 的底行中，我们提供了一个使用加泰罗尼亚语的维基百科网站的截图，并指示 GPT-4V 用 20 种不同语言总结信息。GPT-4V 不仅可以识别加泰罗尼亚语中的文本，还能生成准确的摘要并将其翻译成不同语言。这展示了 GPT-4V 在理解和翻译多语言场景文字方面的能力。

我们还探索了多元文化理解的能力 [140, 73]。图 44 展示了这种场景下的示例结果。我们观察到，在给定的婚礼图像中，GPT-4V 能够理解文化细微差别，并生成合理的多语言描述。

在我们的探索中，我们发现 GPT-4V 能够无缝地理解和正确生成不同语言的描述，凸显了其在处理多样化语言环境中的多功能性。

Sec. 4.5 Multilingual Multimodal Understanding

Multilingual Image Description

Prompt:

描述這張圖片



GPT-4V:

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

Prompt:

Décrire l'image



GPT-4V:

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

Prompt:

Popište obrázek

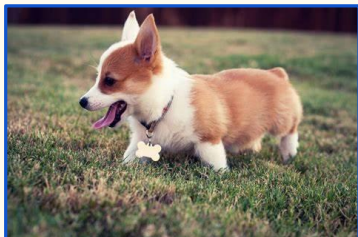


GPT-4V:

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

Prompt:

Describe the image in Chinese

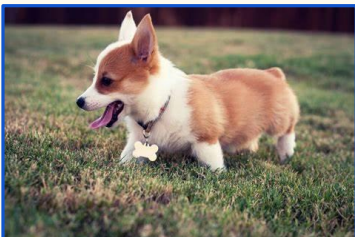


GPT-4V:

這是一隻小柯基犬，它在草地上走着，嘴巴張開，舌頭伸出來，好像很開心的樣子。它的身上有一條項圈，上面挂着一個骨头形狀的牌子。

Prompt:

Describe the image in French

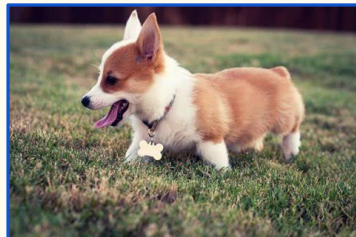


GPT-4V:

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

Prompt:

Describe the image in Czech



GPT-4V:

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

Prompt:

Describir la imagen en 20 idiomas diferentes



GPT-4V:

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

GPT-4V output continued

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

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

Sec. 4.5 Multilingual Multimodal Understanding

Multilingual Scene Text Recognition

Prompt:

What are all the scene text in the image?



GPT-4V:

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

Prompt:

What are all the scene text in the image?



GPT-4V:

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

Prompt:

What are all the scene text in the image?



GPT-4V:

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

Prompt:

What are all the scene text in the image?



GPT-4V:

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

Prompt:

What are all the scene text in the image?



GPT-4V:

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

Prompt:

What are all the scene text in the image?



GPT-4V:

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

Prompt:

What are all the scene text in the image?



GPT-4V:

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

Prompt:

What are all the scene text in the image?



GPT-4V:

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

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

Sec. 4.5 Multilingual Multimodal Understanding

Multilingual Text Recognition, Translation, and Description

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."

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

Prompt:

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

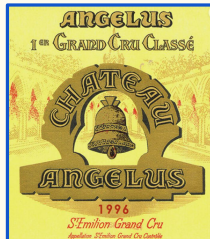


GPT-4V:

The text in the image is in Korean and translates to "Delicious food".

Prompt:

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



GPT-4V:

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

Prompt:

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

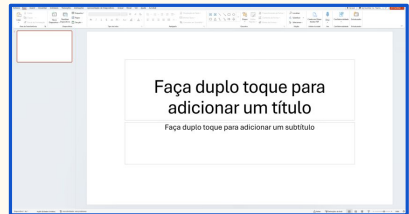


GPT-4V:

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

Prompt:

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:

Summarize the info in 20 languages



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.

GPT-4V output continued

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 알프스는 뉴질랜드의 산맥입니다.
11. Hindi: दक्षिणी एल्प्स न्यूजीलैंड का एक पर्वत श्रृंखला है।
12. Bengali: দক্ষিণ অ্যাল্পস নিউজিল্যান্ডের এই অধিক পর্বতমালা।
13. Arabic: الألب الجنوبية هي سلسلة جبال في نيوزيلندا.
14. Turkish: Güney Alpler Yeni Zelanda'da bir dağ silsilesidir.
15. Indonesian: Alpen Selatan adalah sebuah pegunungan di Selandia Baru.
16. Vietnamese: Dãy núi Alpes phía Nam là một dãy núi ở New Zealand.
17. Thai: เทือกเขาแอลป์ภาคใต้เป็นเทือกเขาในนิวซีแลนด์
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.

图 43: Results on multilingual text recognition, translation, and description. GPT-4V is able to recognize, translate and generate descriptions in different languages. Check Section 4.5 for detailed discussions.

Sec. 4.5 Multilingual Multimodal Understanding

Multilingual Multicultural Understanding

Prompt:

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



GPT-4V:

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

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

Prompt:

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



GPT-4V:

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

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

Prompt:

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



GPT-4V:

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

Amharic (Ethiopia's official language): ይህ ምስል በኢትዮጵያዊ ባህል ልብስ የለበሱት ሰጠብና ሰጠብን በገባር ይታያል። ሴት የቱ በገጽነት ያሉ አለባለባህር ልብሶች ለበሳሌት እና ወንድ የቱም በመሳሰሉ አለባለባህር ሴት ለበሳሌት።

Prompt:

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



GPT-4V:

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

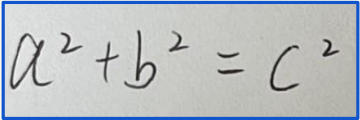
Hindi (local language in India): यह छवि एक नवविवाहित भारतीय जोड़े को एक चुंबन शेयर करते हुए दिखाती है। दुल्हन एक परंपरागत लाल साड़ी पहन रही है और दुल्हा एक लाल पगड़ी और सफेद सूट पहन रहा है। वे फूलों से सजे हुए एक मंच पर बैठे हैं।

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

4.6 Coding Capability with Vision

Sec. 4.6 Coding Capability with Vision

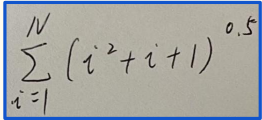
Prompt:



GPT-4V:

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

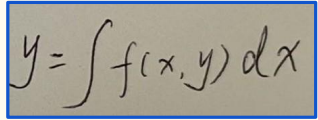
Prompt:



GPT-4V:

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

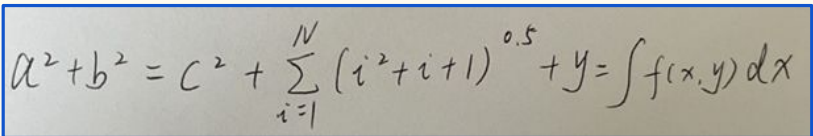
Prompt:



GPT-4V:

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

Prompt:



GPT-4V:

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

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

图 45展示了根据手写数学方程生成LaTeX代码的能力。这个功能可以帮助用户更高效地在LaTeX中编写方程。虽然模型无法为较长的方程生成代码，但它可以有效地处理较短的方程。通过将较长的方程分解为较短的组件，模型可以生成适当的代码。图 46进一步展示了GPT-4V如何将输入图片中的表格重构为Markdown/LaTeX代码。

图 47显示了使用Python、TikZ和SVG编写代码来复制输入图的示例。尽管生成的输出不是完全匹配的，但布局相似，代码可以很容易地修改以满足特定需求。

65

Sec. 4.6 Coding Capability with Vision

Prompt:

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

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

GPT-4V:

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

Visualization of Reconstructed Table (Markdown):

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

Prompt:

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

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

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

GPT-4V:

Here is the reconstructed table in latex:

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

Visualization of Reconstructed Table (LaTeX):

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

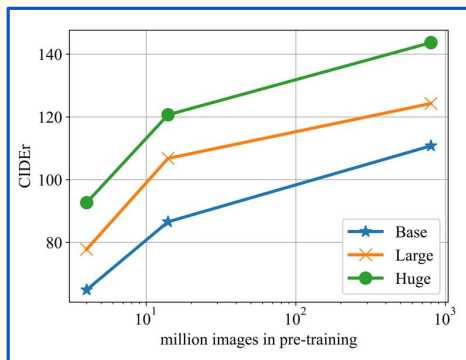
图 46: GPT-4V's capability to generate Markdown/LaTeX codes to reconstruct a table in the image.

Red highlights the errors in reconstruction. Check Section 4.6 for detailed discussions.

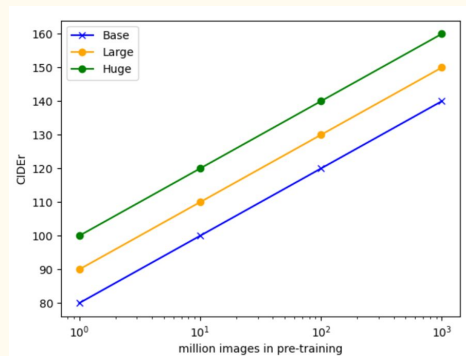
Sec. 4.6 Coding Capability with Vision

Prompt:

Generate python code to draw similar curves.

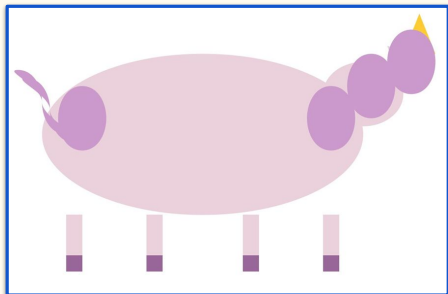


GPT-4V:

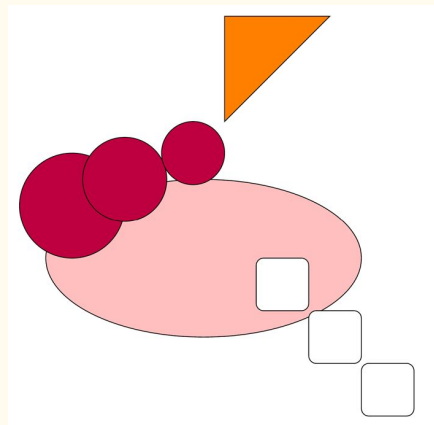


Prompt:

Generate the following image in tikz.

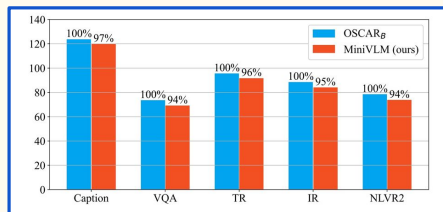


GPT-4V:

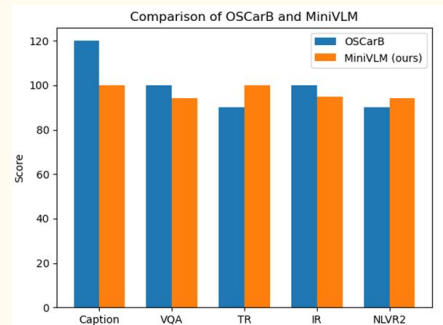


Prompt:

Write python codes to generate similar figures.



GPT-4V:



Prompt:

Write svg codes to generate following image.



GPT-4V:

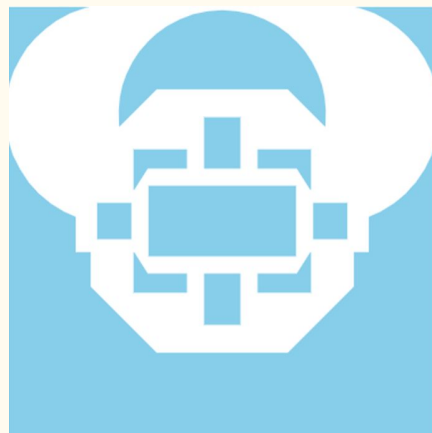


图 47: GPT-4V’s capability to write codes to replicate the input figure. We directly show the rendered figures by python/TikZ/SVG as GPT-4V’s response. The rendered figure is roughly aligned with the input figure, and the code can be easily adapted. GPT-4V Chart. Check Section 4.6 for detailed discussions.

5 Interaction with Humans: Visual Referring Prompting

指向特定空间位置是人机交互中的一项重要能力，尤其是在多模态系统中进行基于视觉的对话时。如 5.1 节所示，GPT-4V 能够很好地理解直接绘制在图像上的视觉指示。基于此观察，我们提出了一种名为“视觉引用提示”的新型模型交互方法。其核心思想是直接在图像像素空间中编辑绘制视觉指示或场景文本作为人类引用指令，如图 50 所示。我们将在 5.2 节详细介绍它的使用方法和优势。最后，我们在 5.3 节探索使 GPT-4V 生成视觉指针输出以与人类进行交互。这些视觉指针对人类和机器来说都是直观的，使它们成为人机交互的良好渠道。

5.1 Understand Pointing Inputs

如图 48 所示，GPT-4V 可以直接识别叠加在图像上的不同类型的视觉标记作为指针，例如圆形、方框和手绘。这种能力有助于 GPT-4V 生成具有基础描述的字幕，而这在传统的视觉语言模型中是一个已知的具有挑战性的问题 [121]，该问题是生成专注于特定感兴趣区域的视觉描述。稠密字幕方法 [59, 131] 使用裁剪的框或遮罩区域生成局部化描述，但常常忽略全局图像背景并产生次优的描述。视觉指示提供了一种自然的方式来指示感兴趣区域，同时保持全局图像背景。例如，左上方的示例侧重于提供指示的 Magna 啤酒的综合描述，同时提到了全局图像背景中啤酒瓶在桌子上的事实。

覆盖在图像上的视觉指示的直观替代方式是使用数值文本格式表示的区域坐标。如图 49 所示，GPT-4V 可以直接理解坐标，在没有额外的指令微调的情况下，通过文本指令实现空间引用的能力，而前期视觉语言模型需要进行指令微调 [122, 136]。尽管有着很好的能力，但我们要注意到我们当前的提示在空间上不够精确。例如，在图 49 的左上方示例中，GPT-4V 提到了周围的物体餐巾纸和水瓶，尽管只有啤酒瓶在区域 (0.47, 0.48, 0.55, 0.87) 中。总体而言，就我们的实验提示而言，与文本坐标相比，GPT-4V 被带有叠加视觉指示的提示更可靠地工作。这种独特的能力激励我们探索一种新的提示方法，即视觉引用提示。

Sec. 5.1 Understand Pointing Inputs

Grounded Description (overlaid visual pointers)

Prompt:

Describe the pointed region in the image.



GPT-4V:

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

Prompt:

Describe the pointed region in the image.



GPT-4V:

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

Prompt:

Describe the pointed region in the image.



GPT-4V:

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

Prompt:

Describe the pointed region in the image.



GPT-4V:

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

Prompt:

Describe the two pointed objects in the image.



GPT-4V:

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

Prompt:

Describe the pointed region in the image.



GPT-4V:

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

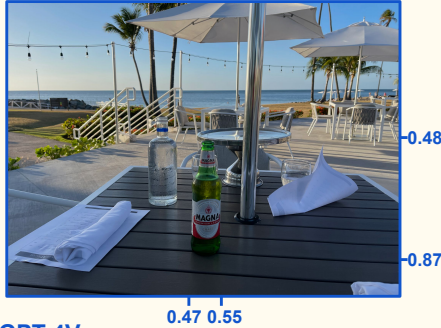
图 48: GPT-4V understands visual pointers directly overlaid on images. Conducting grounded description with both local and global visual information is one unique application scenario. Check Section 5.1 for detailed discussions.

Sec. 5.1 Understand Pointing Inputs

Grounded Description (text-format region coordinates)

Prompt:

Locate the region based on given coordinates and describe the contents in the region. Regions are represented by (x1,y1,x2,y2) coordinates. x1 x2 are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. y1 y2 are the top and bottom most positions, normalized into 0 to 1, where 0 is the top and 1 is the bottom. Describe the relative position of region (0.47, 0.48, 0.55, 0.87) in the image. Describe the contents in that relative position. No need to mention contents in other part of the image.



GPT-4V:

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

Prompt:

Locate the region based on given coordinates and describe the contents in the region. Regions are represented by (x1,y1,x2,y2) coordinates. x1 x2 are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. y1 y2 are the top and bottom most positions, normalized into 0 to 1, where 0 is the top and 1 is the bottom. Describe the relative position of region (0.01, 0.09, 0.29, 0.21) in the image. Describe the contents in that relative position. No need to mention contents in other part of the image.

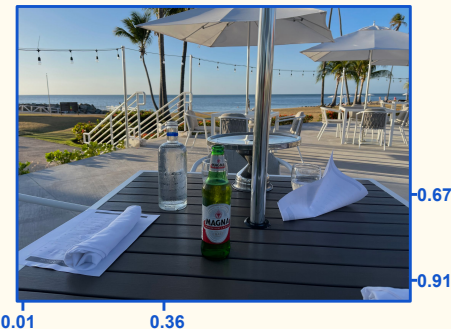


GPT-4V:

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

Prompt:

Locate the region based on given coordinates and describe the contents in the region. Regions are represented by (x1,y1,x2,y2) coordinates. x1 x2 are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. y1 y2 are the top and bottom most positions, normalized into 0 to 1, where 0 is the top and 1 is the bottom. Describe the relative position of region (0.01, 0.67, 0.36, 0.91) in the image. Describe the contents in that relative position. No need to mention contents in other part of the image.



GPT-4V:

The region (0.01, 0.67, 0.36, 0.91) in the image is located at the bottom left corner, covering approximately one third of the width and one fourth of the height of the image. In this region, there is a black table with a white napkin, a bottle of Magna beer, and a glass bottle of water. The table is in the foreground, while the background shows a part of the patio with white umbrellas and chairs.

Prompt:

Locate the region based on given coordinates and describe the contents in the region. Regions are represented by (x1,y1,x2,y2) coordinates. x1 x2 are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. y1 y2 are the top and bottom most positions, normalized into 0 to 1, where 0 is the top and 1 is the bottom. Describe the relative position of region (0.47, 0.48, 0.55, 0.87) in the image. Describe the contents in that relative position. No need to mention contents in other part of the image.



GPT-4V:

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

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

5.2 Visual Referring Prompting

受到GPT-4V在理解视觉指向和场景文本方面的强大能力的启发，我们探索了与GPT-4V交互的新方法，即视觉指称提示。与传统的文本提示技术不同，视觉指称提示是一种在输入图像的像素空间直接编辑的补充技术，用于人机交互。这种视觉提示可以提供一种更细致、全面的与图像的交互方式，潜在地释放模型的更多响应。例如，在图 50（1）中，GPT-4V自然地将箭头指向的对象与给定的对象索引关联起来，简化了其余的视觉推理和文本输出；在（2）中，GPT-4V理解图像上书写的问题并指向相应的边缘或角度，为基于图像的对话提供了细致的接口；在（3）中，人们可以指向图中任意区域，帮助GPT-4V更好地理解复杂的文档和图表；在（4）中，这个模式可以简洁地表示为一个箭头和场景文本“+点”，从而帮助GPT-4V预测下一个图像。与松散地与图像相关的文本提示相辅相成，视觉指称提示提供了一种新的交互方法，可以促进各种用例的实现，关于更多示例请参见图 51和第 9节的讨论。

5.3 Generate Pointing Outputs

第 5.1 节讨论了 GPT-4V 对人类生成的视觉指示的理解能力。一个自然的问题是：GPT-4V 是否能够生成自己的指示输出，从而促进人机交互中的闭环交互过程？

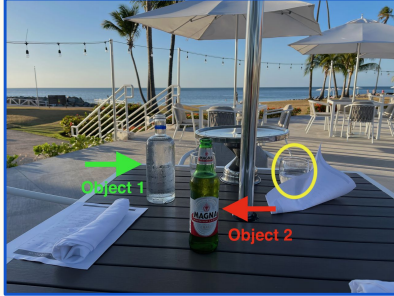
图 52 展示了通过让 GPT-4V 在文本格式中预测区域坐标来生成视觉指示输出。我们通过文本指代对象（例如“蓝色斯巴鲁SUV”的文本）或参考图像（例如“黑色奥迪轿车”的图像）来引导 GPT-4V。与对坐标输入的理解观察类似，模型对空间位置有大致的理解，但在实验使用的提示方面并不准确。例如，在图 52 的“绘制可视化”中，GPT-4V 可以大致识别查询中提到的蓝色SUV和黑色轿车，但它难以创建一个紧密贴合的边界框。我们观察到，在提示中包含例子引导指令有助于 GPT-4V 理解坐标的定义，并进而生成更好的指示输出。

生成的指示输出可能不完美地覆盖查询区域，但它们仍然为模型交互、解释和帮助多步骤视觉推理提供了有价值的工具。具体而言，人类可以解释这些指示输出以更好地理解 GPT-4V 的引用，或者 GPT-4V 本身可以根据先前的输出进行进一步推理。如图 52 底部所示，GPT-4V 能够解释自己所生成的指示，并通过图 48 中的提示提供了与现实场景相关的描述。此外，GPT-4V 自身的迭代指示生成和理解可以帮助处理复杂的多跳关联视觉推理任务。GPT-4V 能够拆解问题，为每个子步骤迭代生成不同的视觉标记，以便逐步聚焦于图像的不同区域，最终整理信息以得出最终答案。

Sec. 5.2 Visual Referring Prompting

(1) Prompt:

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

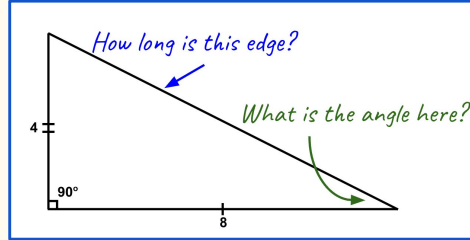


GPT-4V:

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

(2) Prompt:

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



GPT-4V:

To find the length of the blue arrowed edge, we can use the Pythagorean theorem since we have a right triangle. The theorem states that $a^2 + b^2 = c^2$, where a and b are the two legs of the right triangle and c is the hypotenuse. Plugging in the values we have, we get:
 $4^2 + 8^2 = c^2$ $16 + 64 = c^2$ $80 = c^2$ $\sqrt{80} = c$ $8.94 \approx c$
Therefore, the length of the blue arrowed edge is approximately 8.94.
To find the angle at the green arrow, we can use the tangent function since we have the opposite and adjacent sides of the angle. The tangent function states that $\tan\theta = \text{opposite/adjacent}$. Plugging in the values we have, we get:
 $\tan\theta = 4/8$ $\tan\theta = 1/2$ $\theta = \tan^{-1}(1/2)$ $\theta \approx 26.57^\circ$
Therefore, the angle at the green arrow is approximately 26.57°.

(3) Prompt:

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

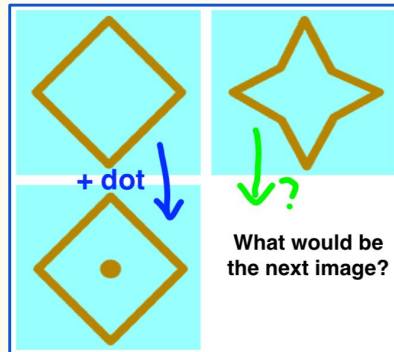
Exam	GPT-4	GPT-4 (no vision)
Uniform Bar Exam (MBE+MEE+MPT)	298 / 400 (~90th)	298 / 400 (~90th)
LSAT	163 (~88th)	161 (~83rd)
SAT Evidence-Based Reading & Writing	710 / 800 (~93rd)	710 / 800 (~93rd)
SAT Math	700 / 800 (~89th)	690 / 800 (~89th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)	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)

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.

(4) Prompt:

Answer questions in the following figure. Provide intermediate steps.



GPT-4V:

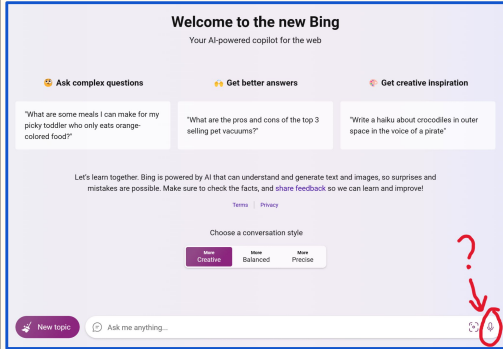
The next image in the sequence would be a star with a dot in the center. This is because the sequence appears to be following a pattern where the shape alternates between a diamond and a star, and a dot is added to the center of the shape every other image. Intermediate steps:
- First, we can observe that the sequence alternates between a diamond and a star shape.
- Next, we can see that a dot is added to the center of the shape every other image.
- Based on this pattern, we can predict that the next image will be a star with a dot in the center.

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

Sec. 5.2 Visual Referring Prompting

Prompt:

What is the icon used for?

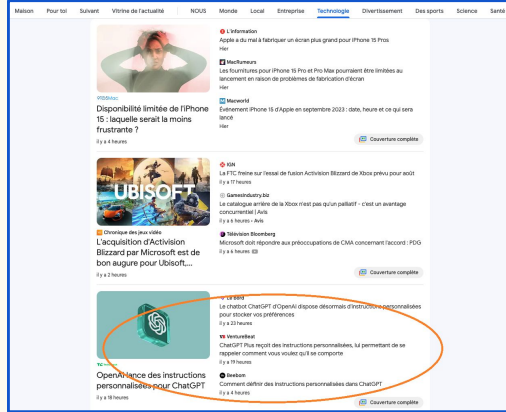


GPT-4V:

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

Prompt:

What does the pointed news talk about?

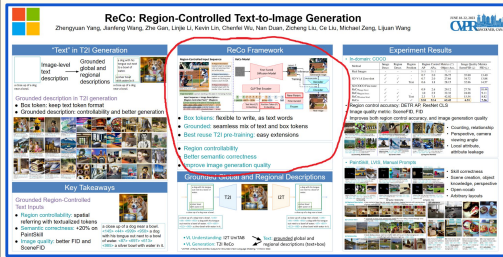


GPT-4V:

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

Prompt:

What does this part of the poster highlights?

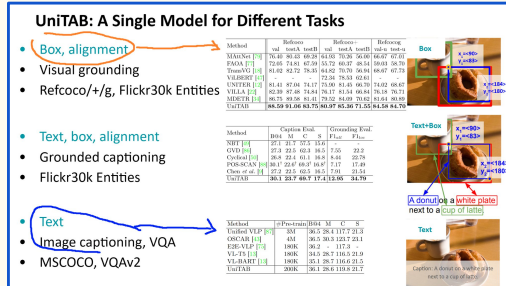


GPT-4V:

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

Prompt:

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



GPT-4V:

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

- 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.

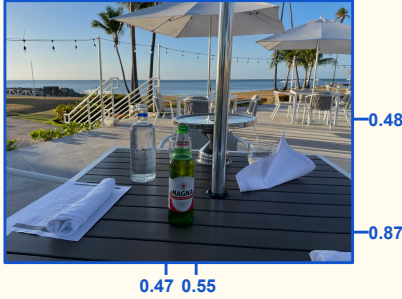
Overall, the two experiments are designed to evaluate the UniTAB model's performance on different tasks and datasets, and to demonstrate its ability to handle multiple tasks with a single model.

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

Sec. 5.3 Generate Pointing Outputs

Prompt:

Locate the referred object and represent the location of the region. Regions are represented by $(x1, y1, x2, y2)$ coordinates. $x1$ $x2$ are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. $y1$ $y2$ 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 $(x1, y1, x2, y2)$ coordinates for that region with the beer bottle is $(0.47, 0.48, 0.55, 0.87)$.



For the image below, what is the $(x1, y1, x2, y2)$ coordinates for the region with the blue Subaru SUV?



GPT-4V:

The $(x1, y1, x2, y2)$ coordinates for the region with the blue Subaru SUV would be $(0.68, 0.57, 0.95, 0.95)$.

Plot Visualization:

Refer to the query image below

Grounded Description

Prompt:

Describe the pointed region in the image.



GPT-4V:

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

Prompt:

Locate the referred object and represent the location of the region. Regions are represented by $(x1, y1, x2, y2)$ coordinates. $x1$ $x2$ are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. $y1$ $y2$ 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 $(x1, y1, x2, y2)$ coordinates for that region with the beer bottle is $(0.47, 0.48, 0.55, 0.87)$.



For the image below, what is the $(x1, y1, x2, y2)$ coordinates for the region with object



in it?



GPT-4V:

The $(x1, y1, x2, y2)$ coordinates for the region with the black Audi car in it would be approximately $(0.35, 0.55, 0.55, 0.85)$. This is based on the left and right most positions of the car being around 35% and 55% of the image width, respectively, and the top and bottom most positions being around 55% and 85% of the image height, respectively.

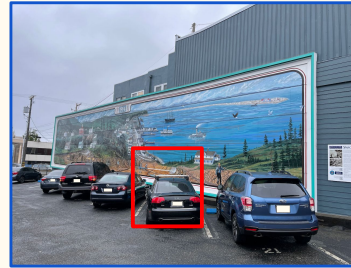
Plot Visualization:

Refer to the query image below

Grounded Description

Prompt:

Describe the pointed region in the image.



GPT-4V:

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

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

6 Temporal and Video Understanding

在这一部分中，我们讨论了时间和视频理解能力。尽管 GPT-4V 主要处理图像作为输入，但评估其对时间序列和视频内容的理解能力仍然是整体评估的关键方面。这是因为现实世界中的事件随时间推移而发生，人工智能系统理解这些动态过程的能力对于实际应用至关重要。例如，时间预测、时间排序、时间定位、时间推理和基于时间的理解等能力有助于评估模型在理解事件序列、预测未来事件和在一系列静态图像中对活动进行上下文分析方面的熟练程度。尽管其侧重于图像，GPT-4V 能够以类似于人类理解的方式理解视频和时间序列。为了增强 GPT-4V 等复杂人工智能模型的多功能性和适用性，对于其发展和完善，测试这一方面至关重要。在本节即将进行的实验中，我们将使用多个选定的视频帧作为输入，以测试模型在理解时间序列和视频内容方面的能力。

6.1 Multi-image Sequencing

在本小节中，我们展示了 GPT-4V 能够准确理解和分析视频帧序列。在逐帧分析中，GPT-4V 能够识别出活动发生的场景，提供更深入的上下文理解。如图 53 所示，该模型不仅局限于识别环境，还能准确解释视频中个体正在进行的动作。GPT-4V 能够理解各种人体姿势的序列和上下文，并智能地将其与正在进行的活动进行关联。通过理解除了识别之外的姿势变化，GPT-4V 能够从人体运动和动作的微妙之处获取含义。由于具有这种详细的理解水平，GPT-4V 能够捕捉视频中发生的事情的本质，提供丰富而细腻的见解，超越仅仅识别对象和场景的范畴。

6.2 Video Understanding

时间顺序。 时间顺序是时间常识的重要组成部分，并且是对 GPT-4V 的能力评估中的一个关键要素。这涉及为模型提供一系列打乱顺序的图像，并评估其辨别因果关系和时间进程的能力。要理解这种关系，需要能够在逻辑连贯和时间准确的方式下重新排序序列。图 54 展示了一个长期时间顺序的示例，GPT-4V 被呈现一系列打乱顺序的图像帧，描述了一个制作寿司的事件。尽管图像无序，GPT-4V 能有效地识别出事件，并确定寿司制作过程的正确时间顺序。此外，图 55 提供了一个短期时间顺序的示例。给定一个指定的操作，例如打开或关闭门，GPT-4V 展示了理解图像内容并确定事件的正确顺序的能力。这些示例突出了 GPT-4V 在时间常识方面的能力，强调其准确理解长期和短期序列的能力。

时间预测。 我们展示了 GPT-4V 在给定一组初始帧的情况下预测未来事件的能力。通过长期和短期的例子验证了 GPT-4V 预测未来事件的能力。图 56 的右侧展示了 GPT-4V 预测短期事件的能力，例如足球点球示例。给定前几帧，它准确地预测了射门者和守门员的典型下一步动作，这是由于其对游戏内在结构和规则的理解。此外，如图 56 的左侧所示，寿司制作过程展示了 GPT-4V 的长期预期能力。通过理解基于视觉线索的活动，GPT-4V 不仅能识别寿司制作的当前进展，还能准确预测后续步骤，展示了其解释和预测复杂多步骤过程的能力。这种长期和短期时间预测的结合使 GPT-4V 能够捕捉和理解具有不同时间结构和复杂度的活动。

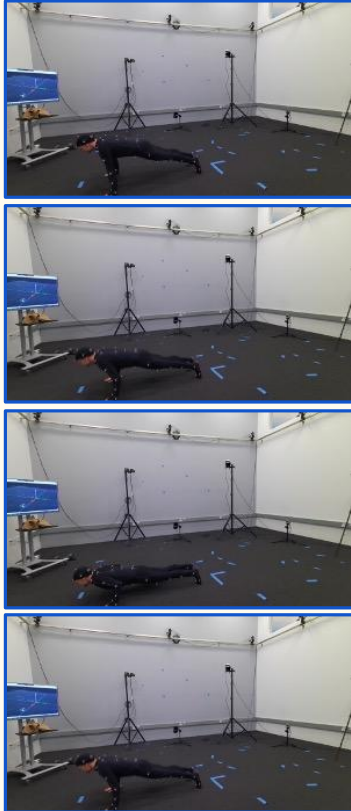
时态定位与推理。 图 57 展示了 GPT-4V 在时态定位与推理方面的能力。它准确地识别了球员击球的瞬间。此外，GPT-4V 通过推断守门员与球之间的关系来判断守门员是否成功挡住了球，展示了其对因果关系的理解能力。在给定的例子中，理解守门员是否能够挡住球不仅涉及到识别守门员和球的空间位置，还涉及到理解它们之间互动的动力学以及预测这种互动的结果。这展示了该模型在推理能力方面具有相当高的复杂性。

Sec. 6.1 Multi-image Sequencing

Action Recognition

Prompt:

Explain the images and the action the person is taking.



GPT-4V:

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

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

Sec. 6.2 Video Understanding

Temporal Ordering

Prompt:

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

a



b



c



Prompt Continued

d



e



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.

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

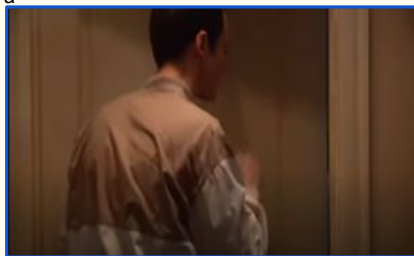
Sec. 6.2 Video Understanding

Temporal Ordering

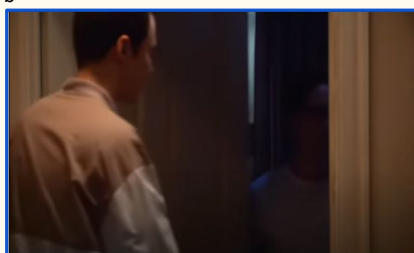
Prompt:

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

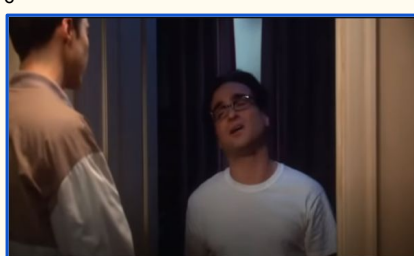
a



b



c



GPT-4V:

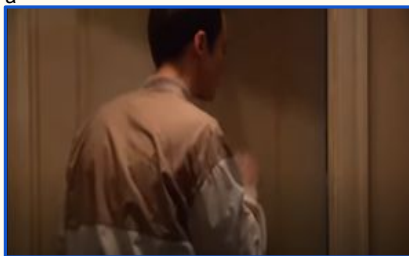
The correct order would be:

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

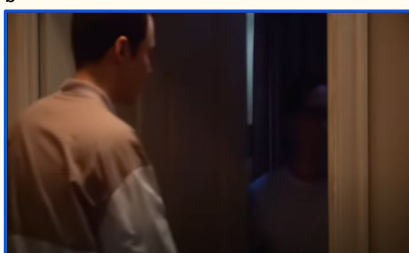
Prompt:

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

a



b



c



GPT-4V:

Correct order:

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

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

Sec. 6.2 Video Understanding

Temporal Anticipation

Prompt:

Predict what will happen next based on the images.



GPT-4V:

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

Prompt:

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



GPT-4V:

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

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

Sec. 6.2 Video Understanding

Temporal Localization and Reasoning

Prompt:

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

a



b



c



Prompt Continued

d



e



f



GPT-4V:

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

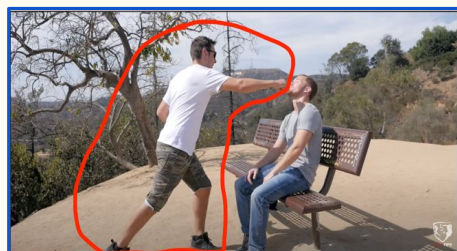
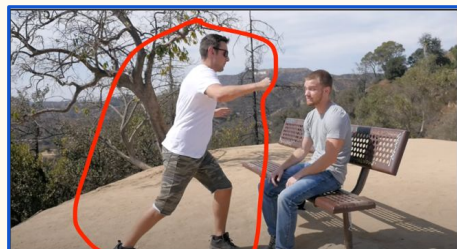
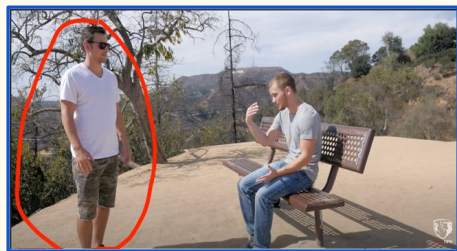
图 57: Temporal localization and reasoning: GPT-4V shows the capability in temporal localization by accurately identifying when the player strikes the ball. It also demonstrates cause-and-effect reasoning by determining whether the ball was blocked based on the goalkeeper-ball interaction. Check Section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

Sec. 6.3 Video Visual Referring Prompting

Grounded Temporal Understanding

Prompt:

Describe and summarize what the circled person doing?

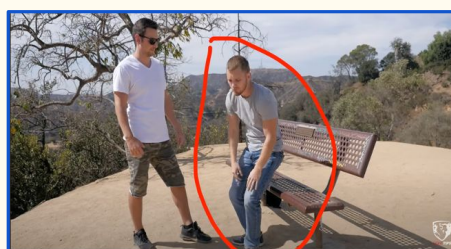
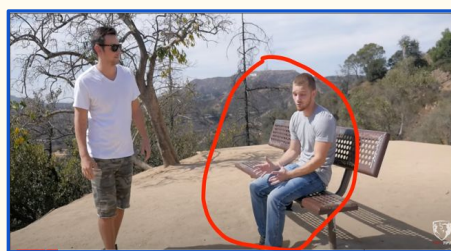


GPT-4V:

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

Prompt:

Summarize the activities of the person.



GPT-4V:

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

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

6.3 Visual Referring Prompting for Grounded Temporal Understanding

第5节展示了 GPT-4V 在视觉指向提示方面的能力。在本节中，我们的目标是通过测试视觉指向提示来扩展这种能力，以用于时间理解。这一进展提供了对视频理解任务的增强控制能力。

基于时间的理解。 基于时间的理解是 GPT-4V 能力的另一个关键方面，我们通过在一组图像帧中使用指示输入来探索这一能力。图 58 通过展示 GPT-4V 如何应用时间理解到特定感兴趣的人，即用圆圈标记的人，来说明这一点。GPT-4V 能够准确描述事件，并与相应的时间顺序保持一致，重点关注被圈出的个体的活动。此外，GPT-4V 还展示了对事件更精细的理解，能够识别出交互的性质。例如，GPT-4V 能够区分友好的互动和暴力事件，说明其不仅能够理解事件的时间流动，还能够解释互动发生的语气和性质。这表明 GPT-4V 具备处理和理解给定序列中复杂的时间和社交线索的能力，为其理解能力增加了更深层次的理解。

7 Abstract Visual Reasoning and Intelligence Quotient Test

理解和推理抽象的视觉刺激和符号是人类智能的一项基本能力。本节讨论了 GPT-4V 是否可以从视觉信号中抽象出语义，并能够完成不同类型的人类智商（IQ）测验。

7.1 Abstract Visual Stimuli

人类可以从抽象且常常含糊不清的视觉刺激中推断语义。图 59 探讨了让 GPT-4V 解释七巧板 [28, 92, 40, 56] 的能力。七巧板是一种传统的几何解谜游戏，由七个被称为“丸”的平面图形组成，这些图形被放在一起形成不重叠的形状。例如，GPT-4V 解释图 59 中的子图7 最能表现出飞翔的鹅，并对其他子图，例如4. 人或机器人，9. 船或帽子，以及 10. 狗或狐狸，提供了推理描述。此外，GPT-4V 还具备理解其他形式的抽象视觉图示的能力 [120, 15, 143]，如图 59 中的卡通人物的 ASCII 艺术以及图 61-62 的符号输入。

7.2 Discovery and Association of Parts and Objects

发现和关联物体的部分 [132, 41] 是另一个重要的抽象视觉推理能力。人类可以轻易地发现物体部分如何组成有语义意义的物体。图60 设计了示例来探测 GPT-4V 在关联物体部分方面的能力。在左边的例子中，我们要求 GPT-4V 根据其语义意义来定位物体的一个部分。在右边的例子中，GPT-4V 被要求将通过SAM [62]分割的物体部分关联起来。GPT-4V可以处理所有物体部分的图像并以有语义意义的方式将它们关联起来形成右下角所显示的男孩。

7.3 Wechsler Adult Intelligence Scale

第 7.1 节展示了 GPT-4V的抽象视觉理解能力。作为进一步的挑战，GPT-4V被要求完成不同的抽象推理任务，这些任务来自人类智商测试。Wechsler成人智力量表 (Wechsler Adult Intelligence Scale) [126] 被公认为是“黄金标准智商测试”之一，它旨在通过一系列子测试全面地衡量个体的认知能力。图 61 展示了每个子测试类别中代表性的问题和 GPT-4V的输出。GPT-4V在抽象推理、仅使用文本回答问题、符号化视觉输入和自然图像等方面显示出了潜力。例如，右下角的示例显示了 GPT-4V可以解释类比问题并找到最佳的鞋子比较。

7.4 Raven’s Progressive Matrices

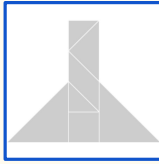
Raven’s Progressive Matrices (RPM) [102] 是另一种著名的非语言智力测验，旨在测量抽象推理和问题解决能力。该测试的设计旨在最小化语言、文化和正规教育对测试表现的影响，使其适用于测试人工智能模型 [15, 143, 52]。每个测试样本包含三个或八个图像，排列在2x2或3x3的矩阵中，其中一个图形缺失。目标是通过识别所提供样本中的模式，从多个候选图像中选择下一个图像。在我们的方法中，我们挑战 GPT-4V，通过将整个问题页作为单个图像发送，而不是将其转换为交错的图像-文本对，类似于人类处理智商测试的方法。如图 62所示，GPT-4V 可以生成合理的答案，而无需经过处理的文本描述或子图。然而，我们也注意到，将整个问题图像分解为交错的文本和子图，如图 63所示，确实简化了任务，并使 GPT-4V 能够产生更可靠的答案。

Sec. 7.1 Abstract Visual Stimuli

Prompt:

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

1.



2.



3.



4.



5.



6.



Prompt Continued

7.



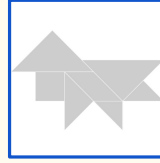
8.



9.



10.



GPT-4V:

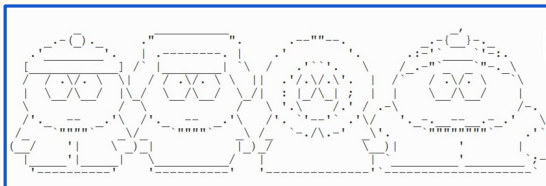
The tangram that best illustrates a flying goose is number 7. It has the triangular shape of a goose's body and wings, and the smaller triangle at the top could represent the head and neck.

The other figures could represent various things:

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

Prompt:

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



GPT-4V:

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

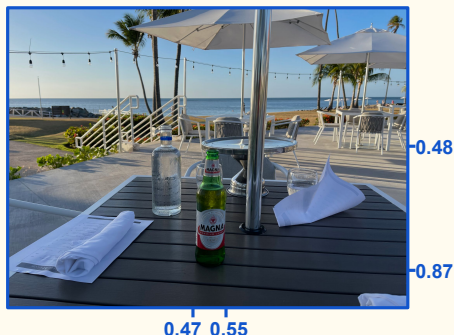
图 59: Understanding abstract visual stimuli such as tangram [56] and ASCII text art. Check Section 7.1 for detailed discussions.

Sec. 7.2 Association of Parts and Objects

Discovery and Association of Parts and Objects

Prompt:

A tangram is a geometric puzzle that consists of seven flat pieces to form shapes. Locate the referred object and represent the location of the region. Regions are represented by $(x1,y1,x2,y2)$ coordinates. $x1$ $x2$ are the left and right most positions, normalized into 0 to 1, where 0 is the left and 1 is the right. $y1$ $y2$ 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 $(x1,y1,x2,y2)$ 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 is the $(x1,y1,x2,y2)$ 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 $(x1,y1,x2,y2)$ coordinates for the region with the head of the bird would be approximately (0.42, 0.15, 0.58, 0.35).

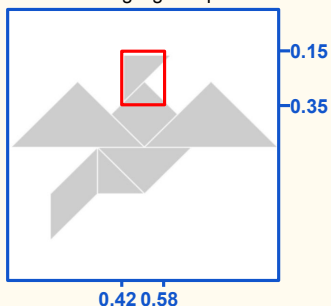
Plot Visualization:

Refer to the query image below

Grounded VQA

Prompt:

What is the highlighted part?

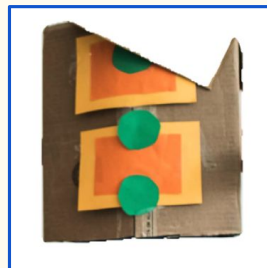


GPT-4V:

The highlighted part is the head of the goose.

Prompt:

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



GPT-4V:

The three visual parts are related in that they are all components that could be used to create a single object. In this case, the object is a boy. To combine these parts to form a boy, we can arrange them in the following way:

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

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

Plot Visualization:



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

Sec. 7.3 Wechsler Adult Intelligence Scale

Prompt:

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

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

GPT-4V:

d. Apple

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

Prompt:

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

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

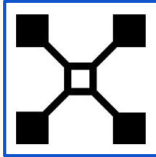
GPT-4V:

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

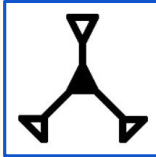
Prompt:

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

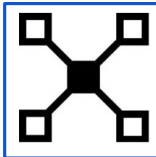
a.



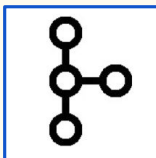
b.



c.



d.



e.



GPT-4V:

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

Prompt:

Which one of the five designs makes the best comparison?



a.



b.



c.



d.



e.



GPT-4V:

Answer: c. The black dress shoe

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

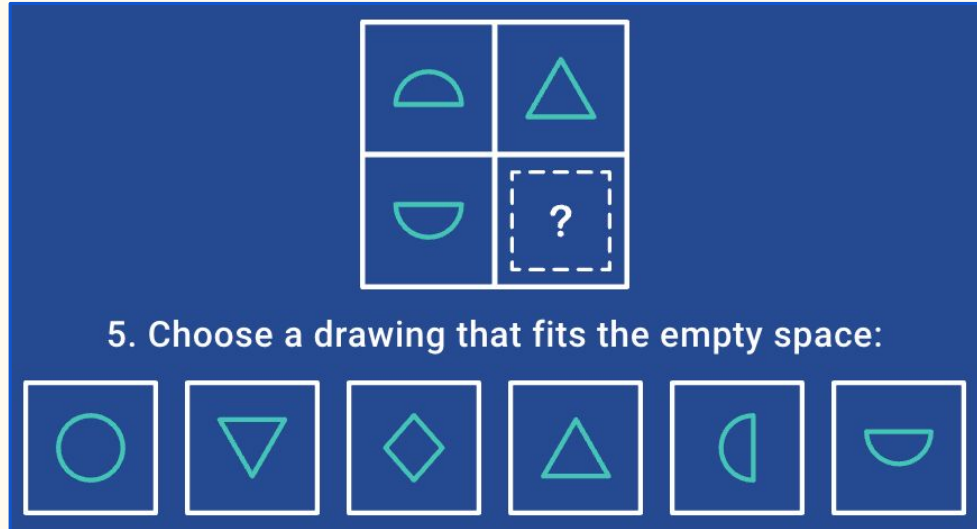
图 61: Example questions from the Wechsler Adult Intelligence Scale (WAIS) [126]. Check Section 7.3 for detailed discussions.

Sec. 7.4 Raven's Progressive Matrices

RPM: Entire Question Page

Prompt:

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



5. Choose a drawing that fits the empty space:

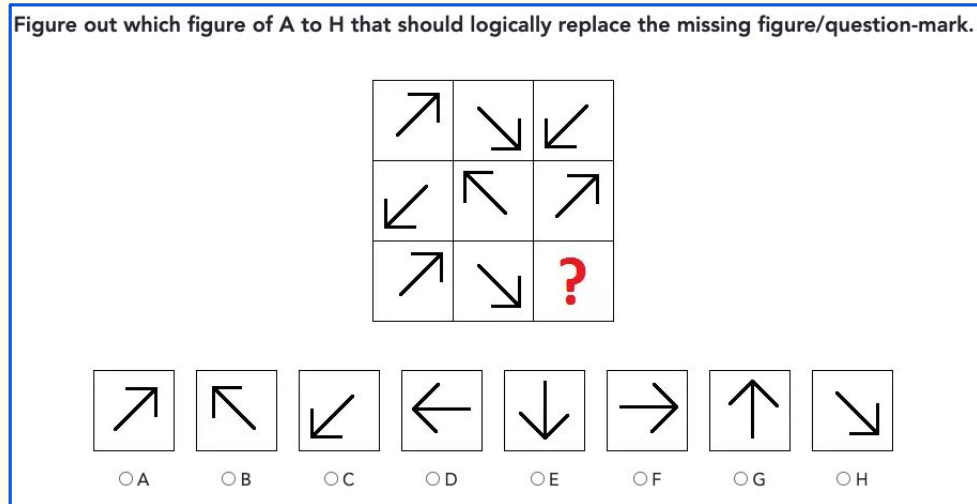
GPT-4V:

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

Prompt:

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

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



○ A ○ B ○ C ○ D ○ E ○ F ○ G ○ H

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.

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

Sec. 7.4 Raven's Progressive Matrices

RPM: Processed Sub-figures

Prompt:

1.



2.



3.



4.

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

a.



b.



c.



d.



e.



f.



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

Prompt Continued

1.



2.



3.



4.

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

a.



b.



c.



d.



e.



f.



GPT-4V:

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

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

8 Emotional Quotient Test

Sec. 8.1 Read Emotion from Facial Expressions

Prompt:

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

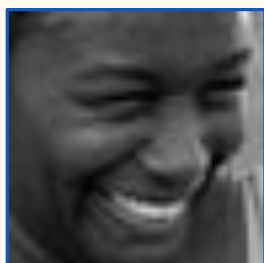


GPT-4V:

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

Prompt:

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



GPT-4V:

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

Prompt:

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



GPT-4V:

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

Prompt:

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



GPT-4V:

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

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

当与人类互动时，GPT-4V 具备同理心和情商（EQ）来理解和分享人类的情感是很重要的。受到人类EQ测试的定义的启发 [87, 86, 20]，我们检验了GPT-4V 在以下方面的能力：（1）从面部表情中识别和读取人类的情感，（2）理解不同视觉内容如何引发情感，以及（3）在所期望的情感和情绪态度条件下生成适当的文本输出。

8.1 Read Emotion from Facial Expressions

如图 64 所示，GPT-4V 能够可靠地通过面部表情识别和读取人们的情感。它还为观察到的视觉线索提供了合理的新理由，以便进行情感的解释，这表明对面部情感有很好的理解能力。

Sec. 8.2 How Visual Content Arouses Emotions

Prompt:

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



GPT-4V:

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

Prompt:

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



GPT-4V:

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

Prompt:

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



GPT-4V:

People might feel 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.

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.

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

8.2 Understand How Visual Content Arouses Emotions

接下来我们分析GPT-4V在视觉情感分析方面的能力，即理解人类在看到视觉内容后的情绪反应。这种能力对于来说非常关键，它能够预测视觉内容可能如何唤起人类情感并作出相应反应。如图65所示，GPT-4V可以根据语义内容和图像风格解读视觉情感，例如内容、愤怒、惊叹和恐惧。这些能力在家庭机器人等使用场景中至关重要。

Sec. 8.2 How Visual Content Arouses Emotions

Image Aesthetics

Prompt:

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

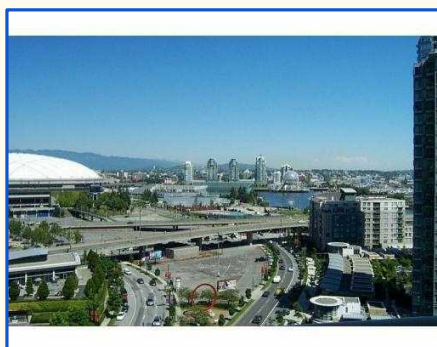


GPT-4V:

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.

Prompt:

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



GPT-4V:

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.

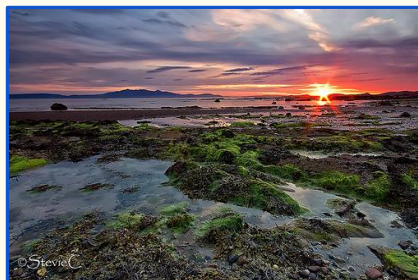
图 66: GPT-4V judges image aesthetics based on societal standards and norms. Check Section 8.2 for detailed discussions.

除了理解视觉情感，GPT-4V还能够与人类主观判断，如审美观点保持一致。图66展示了GPT-4V根据社会标准判断图像美学的示例。

Sec. 8.3 Emotion Conditioned Output

Prompt:

Describe the image in a humorous way.

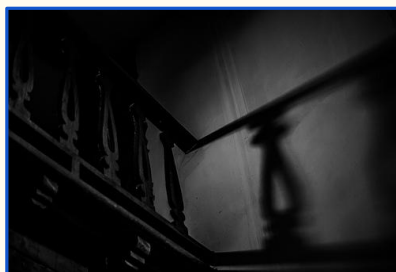


GPT-4V:

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

Prompt:

Describe the image to make people feel uneasy and anxious.



GPT-4V:

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

Prompt:

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

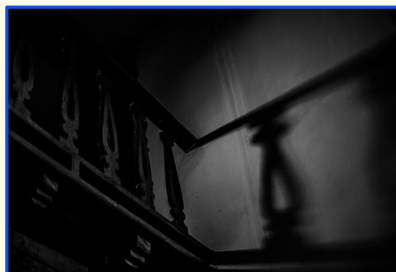


GPT-4V:

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

Prompt:

Describe the image to make people feel 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.

图 67: GPT-4V 生成适当的文本，基于感知到的或期望的情绪，使其沟通对人类舒适和有效。检查第 8.3 节以获取详细讨论。

8.3 Emotion Conditioned Output

基于感知到的情绪，GPT-4V 能够有效地生成与所需情绪相匹配的适当文本输出。例如，在图 67 中，GPT-4V 可以按照提示描述右边的恐怖图像，使其更加可怕或者令人安心。这展示了 GPT-4V 在实现情绪感知的人机交流方面的潜力。

9 Emerging Application Highlights

在本节中，我们展示了大量基于 GPT-4V 强大能力的高价值应用场景和新的用例，这些场景和用例原则上可以通过精心策划用于微调现有的视觉和语言（VL）模型的训练数据来实现。然而，我们强调 GPT-4V 的真正力量在于其能够毫不费力地立即使用。此外，我们还展示了 GPT-4V 如何无缝集成外部工具和插件，进一步拓展其潜力并实现更多创新和协作应用。

9.1 Spot the Difference

我们首先以一个由脑筋急转弯游戏“找出不同之处”启发的通用用例开始。在图 68-69 中，我们提供了两幅模型名称为“GPT-4V”的相似图片，其中一些区域存在微妙的差异。GPT-4V 的任务是确定这两个图像之间的所有不同之处。在这四个例子中，GPT-4V 成功地识别出了图像中不同的区域或组件。然而，它在为每个图像所描绘的内容提供准确的解释方面存在不足。为了深入了解 GPT-4V 的能力，让我们专注于图 68 中展示的第一个例子。尽管 GPT-4V 未能意识到差异在于头箍上剪刀的数量而不是头发的颜色，但它正确指出了两幅图像之间冠状饰品、裙子蝴蝶结和头发的差异。虽然 GPT-4V 在“找出不同之处”游戏中的预测不完美，但它比较两幅图像的内容的能力在现实应用中非常有价值，比如缺陷检测，我们将在下一节中探讨。

9.2 Industry

缺陷检测。 在制造业的发展历史中，计算机视觉技术发挥了关键作用。其中一个具体的应用场景是缺陷检测，这是制造过程中确保产品质量的必要步骤。及时检测并采取适当措施来发现故障或缺陷对于减少运营和与质量相关的成本至关重要。

在这个场景中，我们通过展示有缺陷产品的图片（如图 70-71）来展示 GPT-4V 的缺陷检测能力。对于现实生活中常见的产品（如图 70 中的榛子、织物、螺钉和汽车保险杠），GPT-4V 能够自信地识别出榛子/织物上的小洞、螺钉头上的剥落和汽车保险杠上的凹痕等缺陷。然而，对于不常见的产品图片（如图 70-71 中的金属零件）或外观有变化的产品（如图 71 中的药片），GPT-4V 可能会犹豫不决，甚至拒绝做出预测。图 71 中的一个有趣案例是汽车轮胎，图中观察到多个缺陷，包括轮胎上的污垢、轮毂外沿的损坏和轮胎磨损的迹象。但是，GPT-4V 仅集中在轮胎上的次要缺陷（轮胎上的污垢），没能提到轮毂外沿的主要缺陷（需要修复的损坏）。

鉴于 GPT-4V 在第 9.1 节的“找出差异”场景中的成功，我们探索了借鉴参考图像的理念，以说明无缺陷产品应该是什么样子，旨在改善图 71 中所示的故障案例。这个方法的结果如图 72 所示。通过包含参考图像并优化提示，GPT-4V 成功地识别出单图像缺陷检测中的所有三个故障案例。这些有希望的发现突出了 GPT-4V 在制造业缺陷检测方面的潜在高价值应用。

安全检查。 图 73 展示了个人防护装备（PPE）计数用于安全检查的探索。在像施工现场这样的工作环境中，不充分使用或未佩戴头盔、安全带和手套等 PPE，显著增加了与工作活动相关的风险水平。为了解决这个问题，计算机视觉技术被用作监测 PPE 合规性并及时识别任何安全规定违规的解决方案。以头盔为例，需要一个安全检查系统来准确检测和报告没有佩戴头盔的员工数量。

在图 73a 中，我们通过直接告知 GPT-4V 对佩戴头盔的人数进行计数来评估其性能。GPT-4V 给出了“佩戴头盔的 8 人”，与图像中显示的总人数相符，表明没有发现违规情况。显然，GPT-4V 未能检测到 3 个未佩戴头盔的人，从而危及其个人安全。对于 GPT-4V 来说，这个任

务提出了相当大的挑战，因为它涉及到在图像中检测人员，判断他们是否佩戴头盔，并计算未佩戴头盔的人数的最终值。

在图 73b 中，我们不再将原始包含所有 8 个人的图像提供给 GPT-4V，而是提供了外部人物检测器检测到的被裁剪的人物区域。这种方法将 PPE 计数的工作分为两个步骤：依靠现成的人物检测器进行人物检测，利用 GPT-4V 的强大的视觉推理能力和其处理交错的图像-文本输入的能力来识别安全问题。如图所示，GPT-4V 能够正确计算出未佩戴头盔的人员，同时展示了工具使用和分而治之的好处。

食品结账。 为了加快顾客结账流程并减轻员工工作量，自助结账机在沃尔玛、塔吉特和 CVS 等主要零售商中变得越来越受欢迎。然而，对于顾客来说，实际使用自助结账机可能会很令人沮丧。用户仍然需要寻找产品条形码或手动输入新鲜商品如苹果的代码，这可能是耗时的，尤其是对那些不熟悉该系统的人来说。在图 74 中，我们提供了一个简化的原型，展示了 GPT-4V 在无需用户干预的情况下实现自动自助结账系统的潜力。

当给出一张包含五个食品物品的购物篮的照片时，如图 74a 所示，GPT-4V 未能准确识别篮子中的产品。它错误地将草莓识别为覆盆子，将蟹肉酱误认为希腊酸奶，并且还放入了篮子中并不存在的三文鱼块。然而，如图 74b 所示，我们通过加入从零售网站检索到的食品产品的目录图像来改进提示。结果，GPT-4V 成功识别了篮子里的所有五个物品。这个成功的演示让自助结账系统能够从数据库中获取每个识别产品的价格。虽然这只是一个简单的例子，但它代表了自动自助结账系统迈向更高阶段的重要一步。进一步的研究和开发可以探索更复杂和真实的情境来完全自动化自助结账流程，使其更高效和方便。

Sec. 9.1 Spot the Difference

Prompt:

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

Image 1



Image 2



GPT-4V:

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

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

Prompt:

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

Image 1



Image 2



GPT-4V:

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

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

Sec. 9.1 Spot the Difference

Prompt:

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

Image 1



Image 2



GPT-4V:

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

Prompt:

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

Image 1

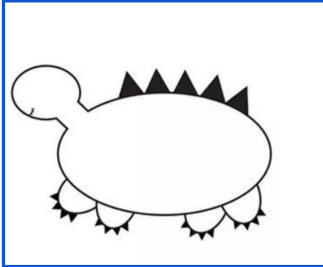
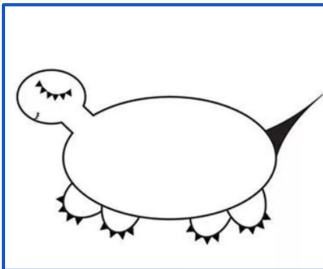


Image 2



GPT-4V:

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

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

Sec. 9.2 Industry

Defect Detection

Prompt:

What is wrong with the object in the image?



GPT-4V:

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

Prompt:

What is wrong with the object in the image?

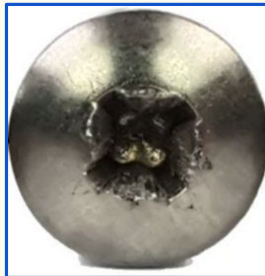


GPT-4V:

There is a small tear or hole in the fabric.

Prompt:

What is wrong with the object in the image?



GPT-4V:

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

Prompt:

What is wrong with the object in the image?

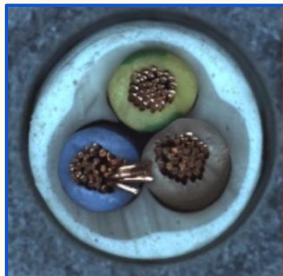


GPT-4V:

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

Prompt:

What is wrong with the object in the image?

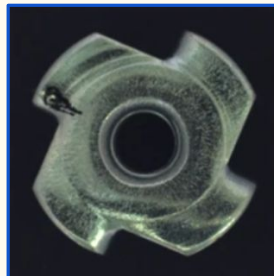


GPT-4V:

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

Prompt:

What is wrong with the object in the image?



GPT-4V:

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.

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

Sec. 9.2 Industry

Defect Detection

Prompt:

What is wrong with the object in the image?



GPT-4V:

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

Prompt:

What is wrong with the object in the image?



GPT-4V:

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

Prompt:

What is wrong with the object in the image?



GPT-4V:

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

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

Sec. 9.2 Industry

Defect Detection

Prompt:

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

Image 1



Image 2



GPT-4V:

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

Prompt:

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

Image 1



Image 2



GPT-4V:

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

Prompt:

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

Image 1



Image 2



GPT-4V:

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

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

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

Safety Inspection

Prompt:

How many people are wearing helmet?



GPT-4V:

There are 8 persons wearing helmets in the image.

(a)

Safety Inspection

Prompt:

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

Image 1



Image 2



Image 3



Image 4



Prompt Continued

Image 5



Image 6



Image 7



Image 8



GPT-4V:

Image 1: Not wearing a helmet
Image 2: Not wearing a helmet
Image 3: Wearing a helmet
Image 4: Wearing a helmet
Image 5: Wearing a helmet
Image 6: Wearing a helmet
Image 7: Wearing a helmet
Image 8: Not wearing a helmet

Summary: Out of 8 people, 5 are wearing helmets and 3 are not wearing helmets.

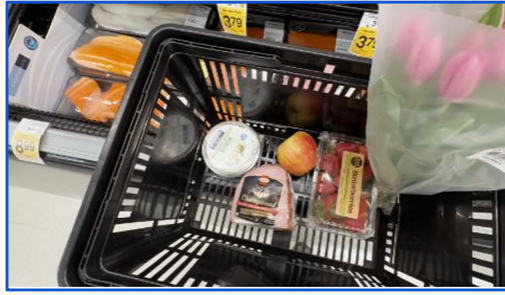
(b)

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

Grocery Checkout

Prompt:

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



GPT-4V:

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

(a)

Grocery Checkout

Prompt:

Here is a list of reference images:

Shrimp



Crab dip



Smoked Ham



Banana



Apple



Bread



Prompt Continued

Cheese



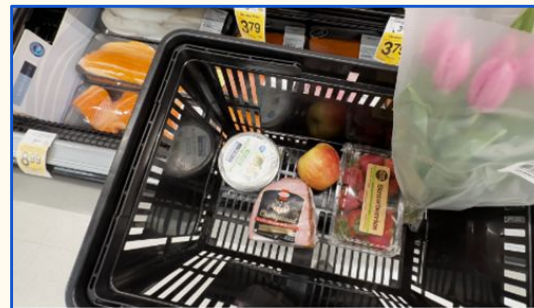
Strawberries



Pink Tulips



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



GPT-4V:

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

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

(b)

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

9.3 Medical

在第 4.1 节中，通过图 18-19，展示了 GPT-4V 在医学图像理解方面的有效性。此外，我们还对 GPT-4V 在放射学报告生成方面进行了详细调查，如图 75-78 所示。在这种情况下，我们向 GPT-4V 提供了各种医学图像，并要求它生成完整的放射学报告。由于评估生成报告的准确性需要领域知识，因此我们寻求了医疗专业人员的评估。

图 75 展示了两个准确的例子：一个涉及腹部 X 射线图像，另一个涉及右膝 MRI。在这两种情况下，GPT-4V 正确识别了研究对象并提供了准确的诊断。接下来是图 76，我们观察到 GPT-4V 为手/腕 X 射线图像生成的放射学报告。虽然 GPT-4V 成功地根据第一张 X 射线图像诊断出了推荐的处理方法，但它没有注意到第二张 X 射线图像中明显存在的远端桡骨骨折。尽管如此，生成的报告保持了高质量的格式，可以作为模板使用，从而减轻医疗专业人员在起草报告时的工作量。

在图 77 中，我们探讨了两个额外的例子，涉及胸部 CT 和脑部 MRI。在胸部 CT 的案例中，GPT-4V 错误地将提到的结节识别为左侧而不是右侧，并且虚构了测量结果。处理交错的图像-文本对的能力还允许 GPT-4V 引用先前的医学扫描和诊断历史，这在医疗专业人员的诊断过程中被证明是至关重要的 [14]。图 78 展示了从多个胸部 X 射线扫描中理解症状进展的示例 [57, 14]。这些示例揭示了 GPT-4V 作为放射学报告生成的 AI 助手的潜力。然而，关键是让医疗专业人员对生成的报告进行评估，以确保其正确性和准确性。

Radiology Report Generation

Prompt:

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



GPT-4V:

Radiology Report

Date: 03/23/2021

Exam: Abdominal x-ray

History: Patient presents with abdominal pain and distention

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

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

Evaluation from Professionals:

Correct study identification, correct diagnosis

Prompt:

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



GPT-4V:

Report:

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

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

Evaluation from Professionals:

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

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

Radiology Report Generation

Prompt:

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

Case:

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

**GPT-4V:**

Radiology Report:

Date: 01/01/2021

Exam: X-ray of the right wrist

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

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

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

Recommendation: Orthopedic consultation for management of the fracture.

Evaluation from Professionals:

Correct diagnosis and management

Prompt:

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

**GPT-4V:**

Report:

Date: 01/01/2021

Exam: X-ray of the right hand and wrist

Technique: AP and lateral views were obtained

Findings: The bones of the right hand and wrist appear intact and in proper alignment. There is no evidence of fracture or dislocation. The joint spaces are preserved and there is no evidence of degenerative changes. No soft tissue abnormalities are seen.

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

Evaluation from Professionals:

Missed the obvious distal radial fracture

Correct identification of the type of study

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

Radiology Report Generation

Prompt:

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



GPT-4V:

Report:

Date: 6/25/2021

Study Type: Chest CT

Subject: 42-year-old male

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

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

Evaluation from Professionals:

Correct study type and identification

Mentions nodule **but in the left rather than in the right.**

Gives measurements (hallucination)

Prompt:

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



GPT-4V:

Report:

Date: 01/01/2021

Modality: Magnetic Resonance Imaging (MRI)

Body Part: Brain

Technique: Axial T1-weighted

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

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

Evaluation from Professionals:

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

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

Sec. 9.3 Medical

Radiology Report Generation with Diagnosis History

Prompt:

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



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



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

GPT-4V:

Radiology Report

Date: 01/01/2021

Patient: John Doe

Exam: Chest X-ray

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

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

Annotations from Professionals:

Improving pleural effusion

Prompt:

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



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



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

GPT-4V:

Radiology Report

Date: 12/12/2021

Patient: Jane Doe

Exam: Chest X-ray

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

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

Annotations from Professionals:

Worsening pleural effusion

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

9.4 Auto Insurance

在本节中，我们探讨了模型名在汽车保险领域的另一个实际应用，具体聚焦于车祸报告。在这个背景下，我们可以进一步划分为两个不同的子类别：(i) 损伤评估和 (ii) 保险报告。前者涉及到准确识别和评估车辆所受损伤的重要任务，而后者不仅包括损伤识别，还包括识别图像中显示的车辆特定信息，如品牌、型号、车牌号码和其他相关细节。通过同时解决这两个方面，我们旨在展示模型名在汽车保险领域的全面能力。

损伤评估。 我们向模型名展示了图 79 中的车辆损伤图像，并用“想象您是一位专家，正在为汽车保险报告评估车祸中的车辆损伤。请评估下面图像中的损伤。”来提示它。模型名在准确识别和精确定位所有四个图像中所描绘的损伤方面展现出了出色的熟练度。此外，它还能够提供每个具体损伤实例的详细描述。在某些情况下，模型名甚至努力估计修复的潜在成本。

保险报告。 在损伤评估取得成功的基础上，我们调整了提示，要求模型名识别图像中所描绘车辆的品牌、型号和车牌号码，并以JSON格式返回所获取的信息。图 80 中的示例展示了这种能力。在这两个示例中，模型名尝试从图像中提取所有请求的细节。然而，值得注意的是，某些信息可能无法获取，例如修复的估计成本，或由于遮挡而难以识别，如第二个图像中的车牌号码。需要注意的是，真实的保险报告通常涉及捕捉车辆各个角度的多个图像，这种情景通常在互联网上不可公开访问。然而，图 79-80 中的示例生动地展示了模型名在车辆事故的保险报告流程自动化方面的潜力。

Sec. 9.4 Auto Insurance

Damage Evaluation

Prompt:

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

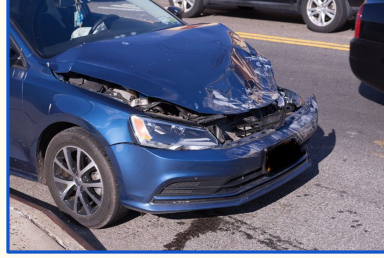


GPT-4V:

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

Prompt:

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



GPT-4V:

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

Prompt:

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



GPT-4V:

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

Prompt:

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



GPT-4V:

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

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

Sec. 9.4 Auto Insurance

Insurance Report Generation

Prompt:

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

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



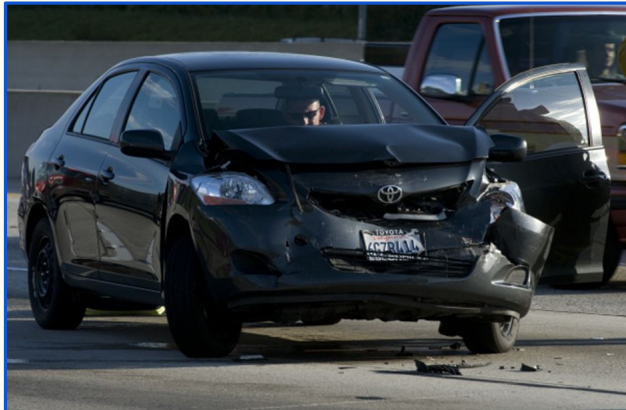
GPT-4V:

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

Prompt:

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

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



GPT-4V:

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

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

9.5 Customized Captioner

照片组织。 在这个场景中，让我们想象我们有一个家庭照片相册。我们展示了 GPT-4V 如何通过生成明确提到照片中每个家庭成员的名称的标题来增强相册。这种个性化的方法有助于更精确和定制的照片组织，如图 81-82 所示。通过为每个家庭成员提供可视提示和他们的姓名，GPT-4V 能够准确识别家庭成员（包括人、猫和狗）并生成详细的定制标题。将这样的标题存储在家庭相册的所有图片中有可能实现高度个性化的图像搜索。例如，用户可以搜索“琳达、科顿、马克斯、山姆和艾玛的一张家庭照片”，并轻松找到图 81 中显示的相应家庭照片，或者搜索“马克斯躺在山姆和艾玛之间”的照片，并找到图 82 的家庭照片。

带分割的密集字幕。 我们通过利用强大的分割模型 [62, 152, 153] 展示了 GPT-4V 在密集字幕中的增强性能。图 83 展示了通过将 SAM [62] 生成的物体分割图与提示相结合，提取每个感兴趣对象的更详细字幕的密集字幕结果。此外，我们将原始图像作为全局背景提供给 GPT-4V，并要求它尽可能详细地描述四个物体分割图，并与上下文图像引用结合起来。

结果显示，GPT-4V 能够为每个物体生成高度复杂的密集字幕，其中一些字幕还包含与上下文图像相关的引用。例如，在描述物体 3（一只青蛙）时，密集字幕提到了一只蛇在青蛙头上停留的特写镜头，尽管相应的物体 3 的分割图中没有蛇。类似地，当提到物体 4（一只乌龟）时，GPT-4V 从上下文图像中识别出乌龟漂浮在水中，从而进一步丰富了生成的字幕。

Sec. 9.5 Customized Captioner

Photo Organization

Prompt:

Here are the photos of my family members.

Linda



Cotton



Max



Sam



Emma



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



GPT-4V:

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

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

Sec. 9.5 Customized Captioner

Photo Organization

Prompt:

Here are the photos of my family members.

Linda



Cotton



Max



Sam



Emma



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



GPT-4V:

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

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

Sec. 9.5 Customized Captioner

Dense Captioning w/ Segmentation

Prompt:

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

Object 1



Object 2



Object 3



Object 4



Context image



GPT-4V:

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

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

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

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

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

9.6 Image Generation

在这一部分中，我们与另一个重要的多模态研究领域进行了联系：视觉合成。通过深入探索图像生成领域，我们探讨了 GPT-4V 在评估和提示等多个方面对该领域的贡献。

生成图像的评估。 在第 8.2 节的图 66 中展示了 GPT-4V 在评估图像审美能力方面的表现。在这里，我们展示了如何使用 GPT-4V 根据给定的文本-图像生成提示来评估生成的图像，这一过程是受 RL-Diffusion [17] 的启发。RL-Diffusion 利用 VL 模型 LLAVA [75] 来描述生成图像，然后使用 BERT [35] 计算提示和图像描述之间的文本相似性。所得的文本相似性分数作为强化学习（RL）引导扩散模型训练的反馈信号。值得注意的是，图 84 和图 85 生动展示了 GPT-4V 的单一模型如何有效地评估生成图像与提示之间的相似度。此外，GPT-4V 还提供了相似度分数减少的解释，这可能可用作改进图像生成的反馈。

在图 84 中，我们使用提示“图像中正在发生什么？请在1到10的评分范围内，评估图像与文本提示“一只鸚鵡开车”的相似度。”来评估图像相似性。对于最不相关的图像（一只海豚跃出水面），GPT-4V 给出了1分的评分，而将最相关的图像评分为9分。值得注意的是，图 84 中的最后三个图像是 RL-Diffusion 中逐步改进的生成结果，对应文本提示“一只鸚鵡开车”。GPT-4V 给这三个图像分配的评分（4 → 8 → 9）与改进过程相吻合。

图 85 展示了涉及蛋糕上的文本渲染的图像生成结果的评估。通过利用其强大的光学字符识别（OCR）能力，GPT-4V 准确地识别了生成图像中的渲染文本，例如“Azuze Research”、“ARAUIE”和“Azure Azure”，并将它们与文本提示要求“Azure Research”进行比较。

用于图像编辑的提示生成。 除了其出色的生成图像评估能力外，GPT-4V 还提供了一个有价值的功能，可以极大地增强图像编辑。通过生成或重写用于编辑的文本提示，GPT-4V 可以优化编辑后的图像，从而产生更具视觉吸引力的结果。图 86 演示了我们如何利用 GPT-4V 的力量生成专门用于图像编辑的文本提示。通过提供原始图像和描述所需编辑的文本要求，GPT-4V 为手头的任务生成了一个优化的提示。这个优化的提示考虑了图像的独特特征，确保后续的编辑过程是有根据、有效的。

此外，图 87 展示了另一个使用 GPT-4V 改进图像编辑的应用案例，通过改写编辑提示。通过考虑原始图像、初次提示和编辑后的图像，GPT-4V 可以生成一个改进的提示，其中包含了在之前的编辑过程中所做的改变。人们可以交替进行图 86-图 87 所描绘的过程，使得用户可以多次细化他们的编辑，直到实现满意的结果。因此，这个迭代过程有潜力显著提升编辑图像的整体质量，为用户在图像编辑中提供更多的控制权和创造自由。

Evaluation of Generated Images (Prompt-Image Alignment)

Prompt:

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



GPT-4V:

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

Prompt:

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

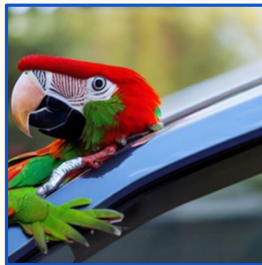


GPT-4V:

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

Prompt:

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



GPT-4V:

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

Prompt:

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



GPT-4V:

In the image, there 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.

Prompt:

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



GPT-4V:

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

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

Sec. 9.6 Image Generation

Evaluation of Generated Images (Prompt-Image Alignment)

Prompt:

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



GPT-4V:

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

Prompt:

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

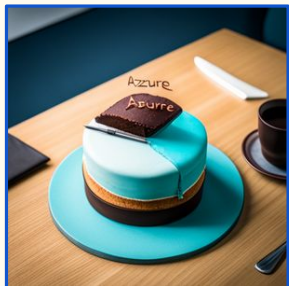


GPT-4V:

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

Prompt:

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



GPT-4V:

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

Prompt:

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



GPT-4V:

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

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

Sec. 9.6 Image Generation

Prompt Generation for Image Editing

Prompt:

This is the original image:



I want to make this image look like graphic novel cover. The prompt should be a single sentence in less than 15 words with descriptive words. For example: "have it look like a graphic novel". Based on the above prompt examples and the original image, help me to write a better prompt to an image editing model.

GPT-4V:

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

Instruct Pix2Pix:

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



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

Sec. 9.6 Image Generation

Prompt Generation for Image Editing

Prompt:

This is the original image:



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

The edited image I got is:



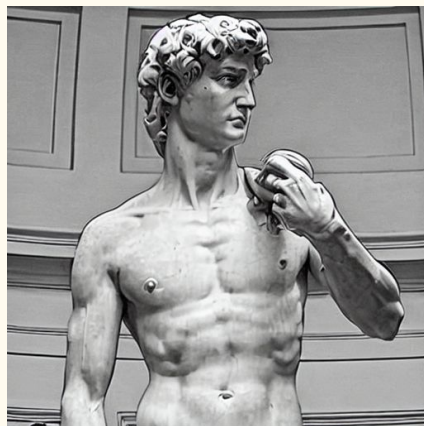
Tell me how I should change my prompt to make the edited image better.

GPT-4V:

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

Instruct Pix2Pix:

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



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



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

9.7 Embodied Agent

在本部分中，我们深入探讨了 GPT-4V 在具身人工智能中的令人兴奋的应用和影响，探讨了它如何将多模态对静态输入的理解与对动态环境中的物理交互进行桥接。为了提供一个具体的例证，让我们考虑 GPT-4V 扮演家庭机器人角色的场景。在这个背景下，我们看到它可以阅读菜单以操作家用电器（例如，咖啡机），并且能够在屋子里进行定向导航。

操作咖啡机。 想象一下，您刚刚购买了一台全新的咖啡机，而您值得欣喜的是，值得信赖的家庭机器人 GPT-4V 已经学会代替您操作它。在我们的实验中，我们向 GPT-4V 提供一张包含具有图示和文字的操作菜单的单张图片（图 88）。我们给 GPT-4V 的任务是，要识别出在咖啡机操作面板上对应于“8 盎司咖啡”选项的按钮。令人惊讶的是，GPT-4V 不仅准确地定位了“8 盎司咖啡”按钮，还成功识别出了“10 盎司咖啡”的按钮。然而，由于“6 盎司咖啡”选项在菜单和咖啡机本身上的位置相同，GPT-4V 错误地将电源按钮误认为是“6 盎司咖啡”的按钮。为了解决这个特定的失败案例，我们设计了一个解决方案，即将每个按钮的操作菜单隔离出来，并将它们同时呈现给 GPT-4V（图 89）。通过这种修订后的方法，GPT-4V 现在可以准确识别出“6 盎司咖啡”按钮的位置。

导航。

为了探索导航能力，我们使用 Redfin 虚拟房屋导览作为复制交互环境的手段，用于体验型智能体。目标是在任务导向的场景中评估 GPT-4V 的性能。为了说明这一点，我们给出了一个例子，如图 90-91 所示。首先，我们向 GPT-4V 提供了一个虚拟房屋导览的入口图像，从一个角落望向客厅。GPT-4V 的任务是“去厨房拿取冰箱里的物品”。我们的目的是促使 GPT-4V 预测随后的行动。

在第一步，如图 90 的前半部分所示，GPT-4V 预测了初始动作，建议“向右转并向前走到走廊。”这个预测是基于 GPT-4V 认为厨房很可能位于那个方向的假设。然后，我们手动执行了这个动作，通过可视化的房屋导览门户捕获了动作执行后的视图。这个视图然后用于提示 GPT-4V 下一步的行动，如图 90 的后半部分所示。重要的是要注意，在整个过程中，我们记录了上一次转向的信息，以为 GPT-4V 随后的行动提供上下文。

随着导航过程的展开，我们在第三次转向时成功到达了冰箱，如图 91 的后半部分所示的查询图像所示。GPT-4V 预测的最后一个动作是“向前前进并稍微向左，以与冰箱门对齐。然后使用我的机械臂打开冰箱门并取出所请求的物品。”这个决定性的动作标志着 GPT-4V 在这个任务导向的导航场景中的成功。

Sec. 9.7 Embodied Agent

Operating Machine

Prompt:

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

Menu:



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



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.

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

Sec. 9.7 Embodied Agent

Operating Machine

Prompt:

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

Menu 1:



Menu 2:



Menu 3:



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



GPT-4V:

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

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

Sec. 9.7 Embodied Agent

Navigation

Prompt:

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



GPT-4V:

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

Prompt:

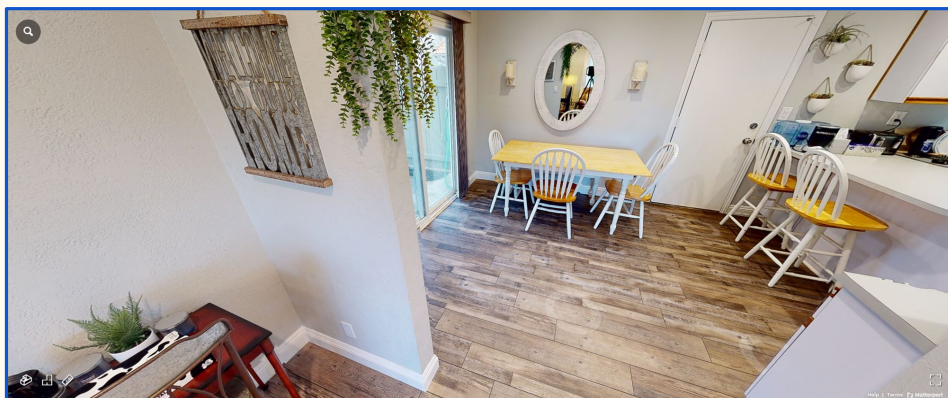
Imagine that you are a home robot, and is asked to go to the kitchen to fetch something from the fridge. At the previous turn, you were at the position shown in the below image:



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

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

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



GPT-4V:

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

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

Sec. 9.7 Embodied Agent

Navigation

Prompt:

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



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

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

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



GPT-4V:

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

Prompt:

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



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

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

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



GPT-4V:

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

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

9.8 GUI Navigation

除了在物理世界中导航外，本节展示了GPT-4V与计算机或智能手机的图形用户界面（GUI）进行交互和导航的能力。我们探索了 GPT-4V完成复杂任务的潜力，例如网页浏览，网上购物等。

网页浏览。 我们在任务导向的情境下评估了 GPT-4V在计算机 GUI 导航上的表现。模型会提供当前计算机屏幕的截图、导航的最终目标（例如找到烹饪食谱或阅读今日新闻）、可能的操作列表（例如移动鼠标、用鼠标点击图标或用键盘输入文本）作为输入。然后，模型被指示预测接下来的操作（参见图 92 中的完整提示示例）。根据模型的预测，我们手动执行预测的操作并捕获屏幕截图，作为下一轮 GPT-4V的输入。当预测的操作是移动鼠标时，特别指示 GPT-4V详细说明鼠标的位置。因此，预测的操作是有依据的，展示了在无人参与情况下自动化整个过程的潜力。

在图 92-96 中，GPT-4V预测了操作计算机 GUI 的合理操作，并最终完成了找到麻婆豆腐食谱并打印出一份食谱的目标（见图 95）。然后，我们向 GPT-4V提供了打印食谱的屏幕截图，并要求它尽可能详细地描述打印内容。如图 96 所示，GPT-4V能够识别打印内容中的细节，包括烹饪时间、配料清单、食谱作者、原始食谱链接等。图 97-102 展示了 GPT-4V如何通过 GUI 浏览网页以“阅读今日新闻”。尽管在图 100 中当尝试返回到前一个搜索结果页面以继续浏览更多新闻文章时存在一些小错误，GPT-4V仍然能够合理地执行导航并阅读两篇新闻文章。

网上购物。 图 103-111 展示了 GPT-4V如何在智能手机 GUI 上进行网上购物导航。类似地，我们向 GPT-4V提供当前手机屏幕的截图、可能的操作列表（例如用手指滑动到一个图标、用手指点击一个图标、向下滚动一个屏幕或用键盘输入文本）并要求它预测继续购买一个价格介于 50 到 100 美元之间的符合人体工程学的键盘所需的操作。GPT-4V预测打开亚马逊应用程序（图 103）、搜索人体工程学键盘（图 104）、打开筛选选项（图 105）、设置价格区间筛选为 50 到 100 美元（图 106）、显示筛选结果（图 107）、选择顶部搜索结果（图 108）、查看产品详情（图 109）、将产品添加到购物车（图 110）并最终去结账（图 111）。

通知理解。 通知是现代人机交互的重要组成部分。GPT-4V展示了解释通知内容并做出相应响应的能力。如图 112 所示，模型能够阅读并响应通知，例如在遇到西雅图的一个会议邀请时建议打开地图应用程序。它还能够计算机屏幕上有效地处理呼叫（图 113）和消息（图 114）通知。

观看视频。 除了网页浏览之外，视频是在线信息的关键来源之一。GPT-4V能够根据受欢迎的短视频的一系列截图描述视频内容。无论视频是否带有字幕覆盖（图 115和 116）还是没有（图 117， 118， 119），GPT-4V都能生成对视频内容有洞察力的描述，展示了它在为用户生成视频内容的自动生成字幕方面的潜力。

Sec. 9.8 GUI Navigation

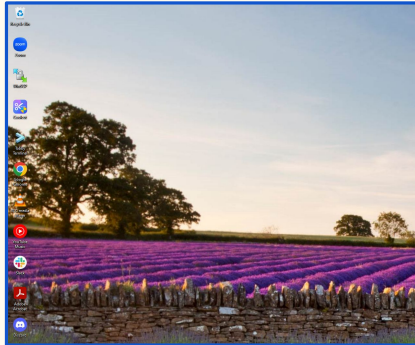
Web Browsing (Mapo Tofu Recipe)

Prompt:

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

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

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



GPT-4V:

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

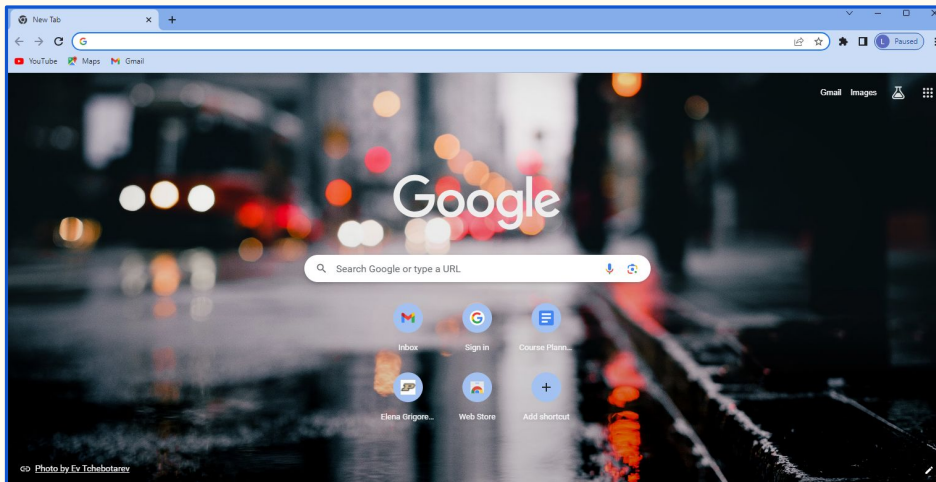
Prompt:

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

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

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

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



GPT-4V:

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

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

Sec. 9.8 GUI Navigation

Web Browsing (Mapo Tofu Recipe)

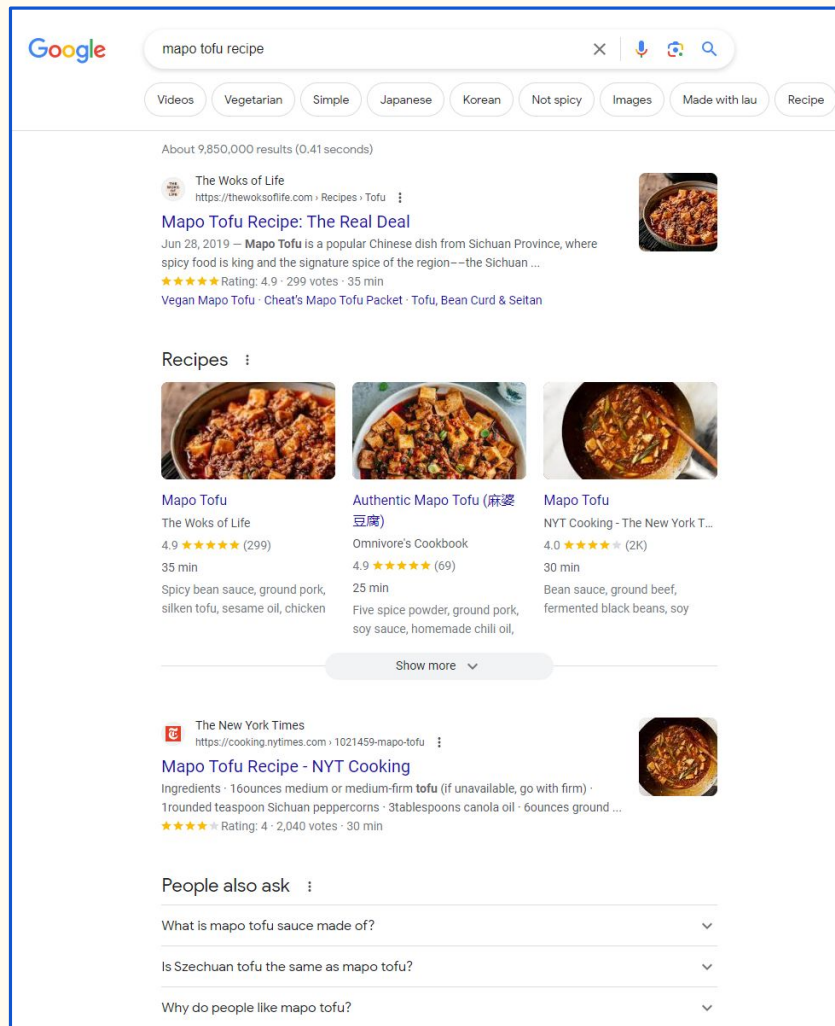
Prompt:

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

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

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

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



GPT-4V:

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

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

Sec. 9.8 GUI Navigation

Web Browsing (Mapo Tofu Recipe)

Prompt:

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

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

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

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



GPT-4V:

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

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

Sec. 9.8 GUI Navigation

Web Browsing (Mapo Tofu Recipe)

Prompt:

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

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

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

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

RECIPE ★★★★★ 4.89 from 299 votes

MAPO TOFU

This mapo tofu recipe is the true blue, authentic real deal —the spicy, tongue-numbing, rice-is-absolutely-not-optional mapo tofu that you get in the restaurants!

by: kaitlin course: tofu cuisine: chinese

SERVES: 6 servings tag or hover to scale

PREP: 10 minutes **COOK:** 25 minutes **TOTAL:** 35 minutes

INGREDIENTS

- ☐ ½ cup oil (divided)
- ☐ 1-2 fresh Thai bird chili peppers (thinly sliced)
- ☐ 6-8 [dried red chilies](#) (roughly chopped)
- ☐ 1/2- 1 1/2 tablespoons [Sichuan peppercorns](#) (powdered or finely ground, reserving 1/4 teaspoon for garnish at the end; if you want a milder flavor use 1/2 or 1 teaspoon ground Sichuan peppercorn)
- ☐ 3 tablespoons [ginger](#) (finely minced)
- ☐ 3 tablespoons [garlic](#) (finely minced)
- ☐ 8 ounces [ground pork](#)
- ☐ 1-2 tablespoons [spicy bean sauce](#) (depending on your desired salt/spice levels)
- ☐ 2/3 cup low sodium chicken broth (or water)
- ☐ 1 pound [silken tofu](#) (cut into 1 inch/2.5cm cubes)
- ☐ 1/4 cup water
- ☐ 1 1/2 teaspoons [cornstarch](#)
- ☐ 1/4 teaspoon [sesame oil](#) (optional)
- ☐ 1/4 teaspoon sugar (optional)
- ☐ 1 [scallion](#) (finely chopped)

US CUSTOMARY **METRIC**

INSTRUCTIONS

- First, we toast the chilies. If you have [homemade toasted chili oil](#), you can skip this step. Heat your wok or a small saucepan over low heat. Add half of the oil and throw in the fresh and dried peppers. Stir occasionally and

PRINT **P** **★ RATE**

Placeholder images near the

GPT-4V:

Next action: "I would move the mouse to the "Print Recipe" button located at the top right corner of the recipe section, next to the "Rate This Recipe" button. I would then click on the button to print out a copy of the detailed recipe for Mapo Tofu."

图 95: GPT-4V navigates through GUI to browse the web to search for the recipe of Mapo Tofu. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Web Browsing (Mapo Tofu Recipe)

Prompt:

Describe the image below as detailed as possible.

Mapo Tofu

This mapo tofu recipe is the true blue, authentic real deal—the spicy, tongue-numbing, rice-is-absolutely-not-optional mapo tofu that you get in the restaurants!

Prep Time	Cook Time	Total Time
10 mins	25 mins	35 mins

Course: Tofu Cuisine: Chinese Servings: 6 servings

Calories: 335kcal Author: Kaitlin

Ingredients

- ½ cup oil (divided)
- 1-2 fresh Thai bird chili peppers (thinly sliced)
- 6-8 dried red chilies (roughly chopped)
- 1/2- 1 1/2 tablespoons Sichuan peppercorns (powdered or finely ground, reserving 1/4 teaspoon for garnish at the end; if you want a milder flavor use 1/2 or 1 teaspoon ground Sichuan peppercorn)
- 3 tablespoons ginger (finely minced)
- 3 tablespoons garlic (finely minced)
- 8 ounces ground pork
- 1-2 tablespoons spicy bean sauce (depending on your desired salt/spice levels)
- 2/3 cup low sodium chicken broth (or water)
- 1 pound silken tofu (cut into 1 inch/2.5cm cubes)
- 1/4 cup water
- 1 1/2 teaspoons cornstarch
- 1/4 teaspoon sesame oil (optional)
- 1/4 teaspoon sugar (optional)
- 1 scallion (finely chopped)


Instructions

1. First, we toast the chilies. If you have homemade toasted chili oil, you can skip this step. Heat your wok or a small saucepan over low heat. Add half of the oil and throw in the fresh and dried peppers. Stir occasionally and heat until fragrant, about 5 minutes, ensuring that the peppers don't burn. Remove from heat and set aside.
2. Heat the remaining half of the oil in your wok over medium heat. Add the ginger. After 1 minute, add the garlic. Fry for another minute, and then turn up the heat to high and add the ground pork. Break up the meat and fry until it's cooked through. Add your ground Sichuan peppercorns and stir for about 15-30 seconds, taking care to not let it burn, as it will turn bitter if it does.
3. Add the spicy bean sauce to the mixture and stir it in well. Add the chicken broth to the wok and stir. Let this simmer for a minute or so. While that's happening, ready your tofu and combine the water and cornstarch in a small bowl.
4. Add the cornstarch mixture to your sauce and stir. Let it bubble away until the sauce starts to thicken. (If it gets too thick, splash in a little more water or chicken stock.)
5. Then add your chili oil from before—peppers and all! If you are using homemade chili oil, ONLY use the standing oil, as it's likely that you have salted it and you only want the oil, not additional salt. Stir the oil into the sauce, and add the tofu. Use your spatula to gently toss the tofu in the sauce. Let everything cook for 3-5 minutes. Add the sesame oil and sugar (if using) along with the scallions and stir until the scallions are just wilted.
6. Serve with a last sprinkle of Sichuan peppercorn powder as a garnish if desired.

Nutrition

Calories: 335kcal | Carbohydrates: 8g | Protein: 11g | Fat: 29g | Saturated Fat: 5g | Cholesterol: 27mg | Sodium: 126mg | Potassium: 340mg | Fiber: 1g | Sugar: 2g | Vitamin A: 165IU | Vitamin C: 3.5mg | Calcium: 47mg | Iron: 1.4mg

Mapo Tofu by The Woks of Life. Recipe URL: <https://thewoksoflife.com/ma-po-tofu-real-deal/>



GPT-4V:

The image is a recipe for Mapo Tofu, a Chinese dish. The recipe is presented in a structured format with a title, description, prep time, cook time, total time, course, cuisine, servings, and rating.

The title of the recipe is "Mapo Tofu" and it is written in bold black letters. Below the title, there is a description of the dish which states that it is the true blue, authentic real deal - the spicy, tongue-numbing, rice-is-absolutely-not-optional Mapo Tofu that you get in the restaurants.

The prep time for the dish is 10 minutes, the cook time is 25 minutes, and the total time is 35 minutes. The course is listed as "Main Course" and the cuisine is "Chinese". The recipe serves 6 people and has a rating of 4.69 stars out of 5.

Below this information, there is a list of ingredients required for the dish, including tofu, ground pork, Sichuan peppercorn, garlic, ginger, soy sauce, and cornstarch. The ingredients are listed in a bullet point format.

At the bottom of the image, there is nutritional information for the dish, including calories, carbohydrates, protein, fat, saturated fat, cholesterol, sodium, and calcium. The nutritional information is presented in a table format.

The image also includes a credit to the author of the recipe, Kaitlin, and a link to the original recipe on the website "The Woks of Life".

图 96: GPT-4V navigates through GUI to browse the web to search for the recipe of Mapo Tofu. As GPT-4V predicts to print out the recipe in the previous turn, we prompt it to read the screenshot of the printed recipe and summarize it. Red highlights the inaccurate description about the image. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

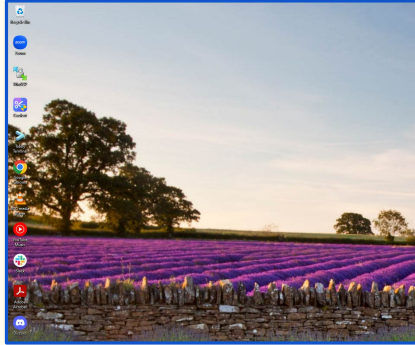
Web Browsing (Today's News)

Prompt:

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

You are asked to find what's in the news today.

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



GPT-4V:

My next move would be to move the mouse to the Google Chrome icon located on the left side of the screen, approximately 1/3 of the way down from the top. I would then click on the icon to open the web browser and search for news websites to find out what's in the news today.

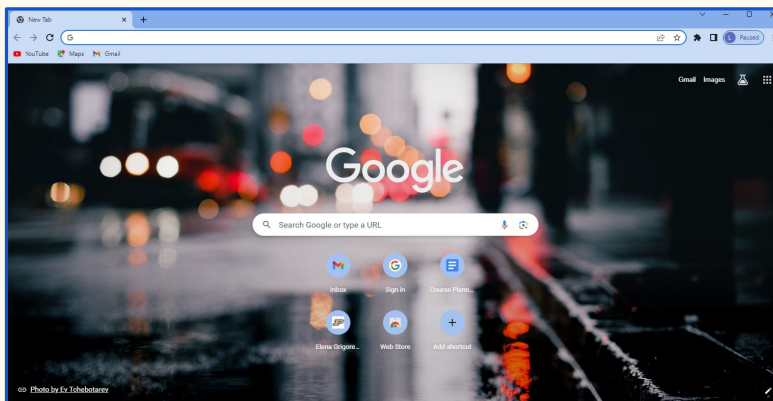
Prompt:

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

You are asked to find what's in the news today.

Previous action: "My next move would be to move the mouse to the Google Chrome icon located on the left side of the screen, approximately 1/3 of the way down from the top. I would then click on the icon to open the web browser and search for news websites to find out what's in the news today."

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



GPT-4V:

My immediate next step would be to move the mouse to the search bar located in the center of the screen, approximately 1/2 of the way down from the top. I would then click on the search bar and type in "news" and press the enter key on the keyboard to search for news websites.

图 97: GPT-4V navigates through GUI to browse the web to read today's news. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Web Browsing (Today's News)

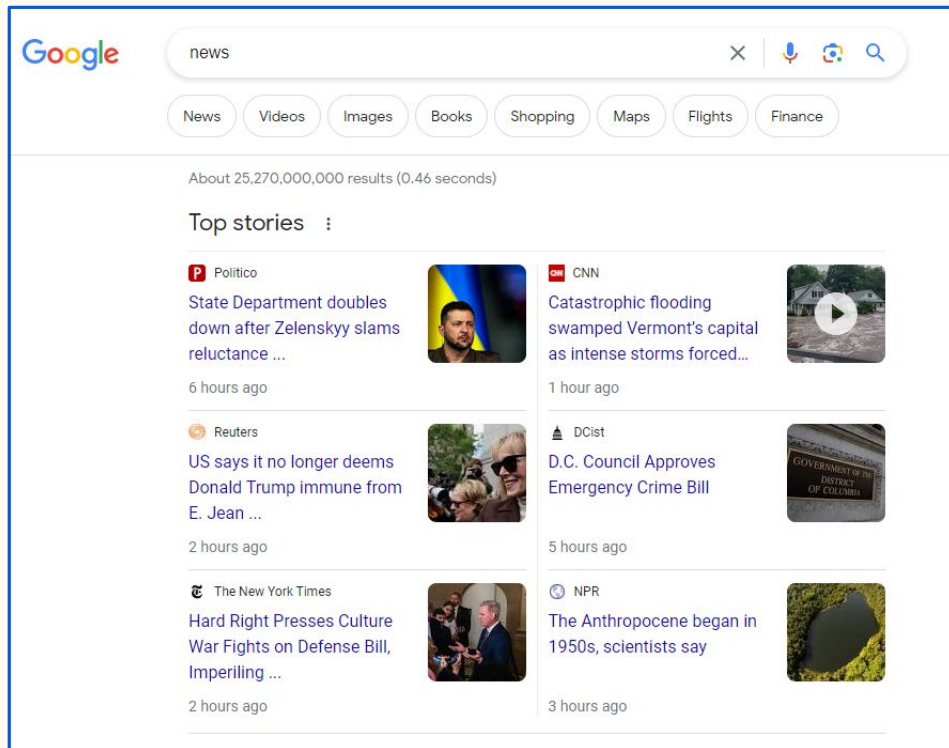
Prompt:

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

You are asked to find what's in the news today.

Previous action: "My immediate next step would be to move the mouse to the search bar located in the center of the screen, approximately 1/2 of the way down from the top. I would then click on the search bar and type in "news" and press the enter key on the keyboard to search for news websites. "

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



GPT-4V:

My immediate next step would be to move the mouse to the first search result, which is "State Department doubles down on Zelensky slams" located approximately 1/4 of the way down from the top of the screen and 1/8 of the way from the left side of the screen. I would then click on the link to open the news article and read about what's in the news today.

图 98: GPT-4V navigates through GUI to browse the web to read today's news. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Web Browsing (Today's News)

Prompt:


Read the article shown in the images below and summarize it.

POLITICO

FOREIGN AFFAIRS

State Department doubles down after Zelenskyy slams reluctance on Ukraine's entry to NATO

State Department press secretary Matthew Miller said the United States' hesitancy to see Ukraine join NATO was because of the ongoing war with Russia.



In a statement posted to Twitter on Tuesday, Ukrainian President Volodymyr Zelenskyy took issue with NATO's reluctance to commit to admitting his country | Francisco Seco/AP Photo

By LUCY HODGMAN
07/11/2023 04:18 PM EDT

The State Department on Tuesday defended NATO's hesitation to invite Ukraine into the alliance, even as Ukrainian President Volodymyr Zelenskyy voiced his frustration with how the negotiations have unfolded.

State Department press secretary Matthew Miller spoke to reporters on Tuesday from the Lithuanian capital, Vilnius, where NATO's two-day summit is being held this week. Miller cited several conditions that NATO had decided Ukraine must meet for an invitation to the alliance, including democratic reforms and the end of the war between Russia and Ukraine.

"We are ready to extend an invitation when conditions are met," Miller said. "There is an important reason why they are not becoming a member of NATO right now — because it would instantly put the United States in a shooting war with Russia."

Miller echoed the sentiment of President Joe Biden, who told CNN on Sunday that Ukraine could not be admitted into the alliance until the war with Russia had ended, adding that "if the war is going on, then we're all in war."

The U.S. has doubled down this week on military support of Ukraine in its war with Russia, agreeing to include controversial cluster munitions in its next arms package to Kyiv — a decision that Ukrainian Minister of Defense Oleksiy Reznikov said would be a "game changer" as the Ukrainian counteroffensive struggles to breach Russia's front lines.

In a summit communiqué released Tuesday, NATO leaders wrote that the alliance could not promise Ukraine membership until the war had ended and several democratic and security-sector reforms had been achieved.

"We will be in a position to extend an invitation to Ukraine to join the Alliance when Allies agree and conditions are met," the leaders wrote.

In a statement posted to Twitter on Tuesday, Zelenskyy took issue with NATO's reluctance to commit to admitting his country, and its unwillingness to provide an anticipated timeline for Ukraine's entry. He wrote that he did not see sufficient readiness on the part of NATO to integrate Ukraine into the alliance.

"It's unprecedented and absurd when time frame is not set neither for the invitation nor for Ukraine's membership," Zelenskyy wrote. "While at the same time vague wording about 'conditions' is added even for inviting Ukraine."

GPT-4V:

Summary:

The article discusses the statement made by Ukrainian President Volodymyr Zelenskyy on Tuesday, where he expressed his country's desire to join NATO. Zelenskyy stated that Ukraine cannot afford to go to war with Russia and that joining NATO would provide the necessary security and protection.

However, Matthew Miller, a spokesperson for NATO, stated that while the alliance is ready to extend an invitation to Ukraine, there are certain conditions that need to be met. These conditions include democratic and security-sector reforms. Miller also emphasized that Ukraine's entry into NATO is not a guarantee and that the decision ultimately lies with the members of the alliance.

The article also mentions that during a press conference on Tuesday, Zelenskyy reiterated his commitment to strengthening democracy in Ukraine, which is a necessary condition for NATO membership.

图 99: GPT-4V navigates through GUI to browse the web to read today's news. We prompt GPT-4V to read the screenshots of the first news article and summarize it. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Web Browsing (Today's News)

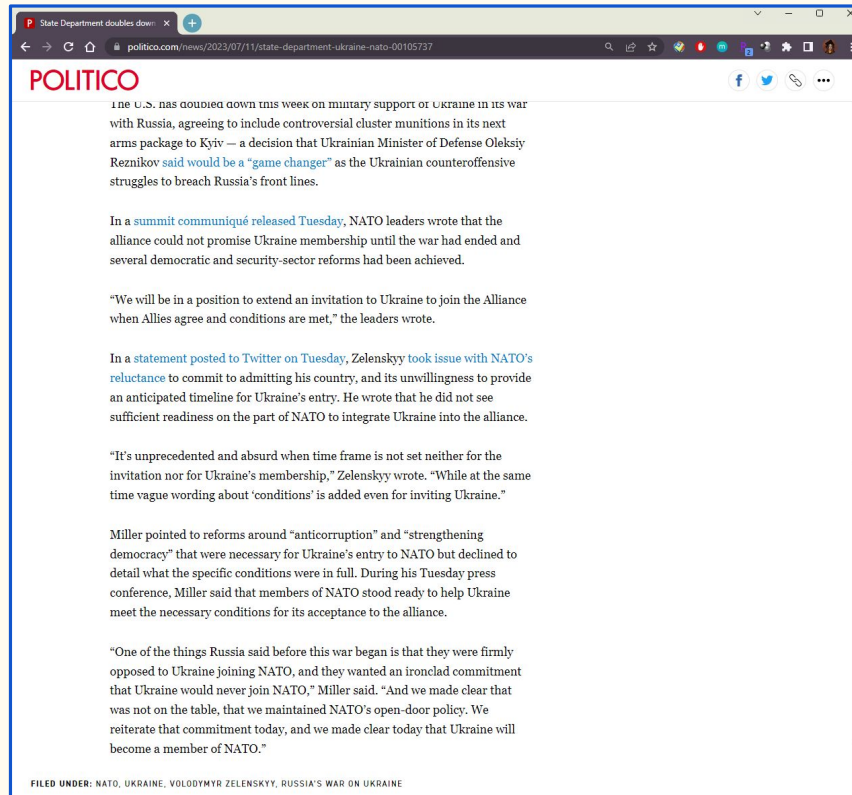
Prompt:

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

You are asked to find what's in the news today.

Previous action: "My immediate next step would be to move the mouse to the first search result, which is "State Department doubles down on Zelensky slams" located approximately 1/4 of the way down from the top of the screen and 1/8 of the way from the left side of the screen. I would then click on the link to open the news article and read about what's in the news today."

You have now completed reading "State Department doubles down on Zelensky slams". What is your immediate next step? If the action involves moving the mouse, describe the location as detailed as possible.



GPT-4V:

My immediate next step would be to move the mouse to the top right corner of the screen, where I can see the "X" icon to close the current tab. I would then click on the "X" icon to close the tab and return to the previous page to continue browsing for more news articles.

图 100: GPT-4V navigates through GUI to browse the web to read today's news. Upon finishing reading the first news article, GPT-4V predicts to close the tab and return to previous page to continue browsing more news articles (highlighted in blue). Red highlights the inaccurate action prediction. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Browsing the Web (Today's News)

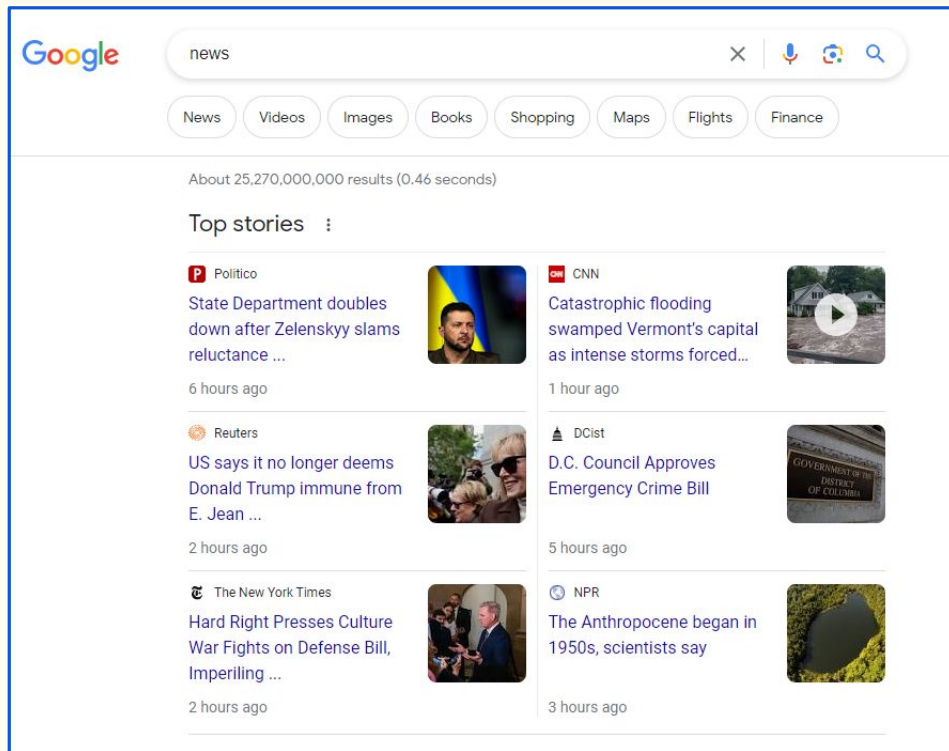
Prompt:

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

You are asked to find what's in the news today.

Previous action: "My immediate next step would be to move the mouse to the top right corner of the screen, where I can see the "X" icon to close the current tab. I would then click on the "X" icon to close the tab and return to the previous page to continue browsing for more news articles. "

You have now returned to the previous page after reading the first article "State Department doubles down on Zelensky slams". Below is what you see on the computer screen. Based on the image below, what is your immediate next step? If the action involves moving the mouse, describe the location as detailed as possible.



GPT-4V:

My immediate next step would be to move the mouse to the second article on the page, "Catastrophic flooding swamps Florida's panhandle as intense storms force evacuations". I would move the mouse to the left side of the article title, approximately 1/3 of the way down the page, and click on the article title to open it and read the news.

图 101: GPT-4V navigates through GUI to browse the web to read today's news. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.


Web Browsing (Today's News)

Prompt:

Read the article shown in the images below and summarize it.

Catastrophic flooding swamped Vermont's capital as intense storms forced evacuations and closures in Northeast

By Noran Salameh and Lauren Mascarenhas, CNN
Updated 10:18 PM EDT, Tue July 11, 2023



Watch how deadly flash flooding hit New York

02:08 Source: CNN

[CNN] — Intense rainstorms that lashed the Northeast turned creeks into rivers, forced evacuations and prompted officials in Vermont's capital, Montpelier, to close the downtown area.

"Make no mistake, the devastation and flooding we're experiencing across Vermont is historic and catastrophic," Gov. Phil Scott told reporters Tuesday.

Floodwaters in some areas "surpassed levels seen during Tropical Storm Irene," the governor said. Irene hit the United States as a hurricane in August 2011 and left entire communities submerged, killing more than 40 people in several Eastern states.

The current flooding claimed thousands of homes and businesses, Scott added.


In downtown Montpelier, authorities issued a travel ban that lasted through Tuesday afternoon. A boil water notice remains in place for the city, after officials warned the "extreme flooding" could contaminate the drinking water supply.

New Hampshire will send swift boat rescue crews and Black Hawk helicopters to assist with the response to the flooding in Vermont, Gov. Chris Sununu said. Teams from Connecticut, Massachusetts and North Carolina are currently in Vermont to provide assistance and others are en route, the Vermont Department of Public Safety said.

In Montpelier, the police department relocated its dispatch, police and fire operations due to heavy flooding in the basement of City Hall and the fire and police departments, Police Chief Eric Nordenson said Tuesday.

Authorities were also monitoring the Wrightsville Dam in Montpelier late Tuesday, after water levels earlier in the day inched toward the dam's spillway.

In an 8:30 p.m. update, Montpelier officials said the dam's water level was beginning to recede and was not expected to breach the spillway. On Wednesday, crews will begin clearing mud and debris from the city's roadways, building inspections will start in the downtown area and clean-up efforts will begin, officials said.



Floodwater surrounded buildings in downtown Montpelier, Vermont, on Tuesday.

Flood alerts affecting over 2 million people across parts of New England and Oklahoma expired Tuesday, but flood warnings remain in both regions.


The bodies of two children were recovered from Lake Overholser, a reservoir in Oklahoma City, after authorities said they were swept up Monday evening in strong currents following heavy rainfall in the city over the weekend.

The children, who authorities said were 10 and 11 years old, were among a group of four that was fishing in the area and entered the water.

Higher than levels reached during Hurricane Irene. The Winouski River at Montpelier rose nearly 14 feet Monday and passed major flood stage as the water continued to climb, threatening further flooding.

The Ball Mountain Dam and the Townshend Dam, in southwestern Vermont, were expected to overflow their spillways," the US Army Corps of Engineers warned Monday night. Warning of "severe flooding," the agency urged residents in threatened low-lying areas of nearby Vermont and New Hampshire communities to evacuate.

Rainfall is at 300% to 500% of normal



The aftermath of flooding in Main Street in Highland Falls, New York, is seen Monday.

Seven-day rainfall totals across much of the Northeast were already at 300% to 500% of normal levels, the Weather Prediction Center said Monday.


Widespread rainfall of 2 to 4 inches fell across the Northeast from eastern Pennsylvania and northern New Jersey into Vermont and New Hampshire. Isolated rainfall totals higher than 6 inches have been seen in several states.

In New York, a 35-year-old woman died after being swept away by floodwater as she tried to evacuate her Orange County home Sunday. The flooding has caused "easily tens of millions of dollars in damage," county Executive Steve Neuhaus said Monday.

Some areas in New York were hammered with more than 8 inches of rain within a 24-hour period.

The city and surrounding communities saw "record-breaking flooding" in two days, Montpelier officials said Tuesday afternoon. The city's downtown area remains flooded and is not safe for public travel, they added.

The city was hammered by a record-setting 5.29 inches of rainfall Monday, the National Weather Service in Burlington said. That's more than any other day on record, including when Irene dropped 5.27 inches of rain on the state capital on August 28, 2011.




RELATED ARTICLE
A perfect storm of ingredients is combining to create deadly flooding and record heat

"Irene had about a 12-hour duration of rain, and then it was over" the governor said. "This is different. We've had like 48 hours of steady rain," he said, adding more rain is expected in the coming days.

Vermont remains under a state of emergency, with water in downtown Montpelier running from knee to waist deep Tuesday, stranding residents in their homes and shutting down roads and businesses. The flooding forced evacuations and more than 100 rescues in the state, Urban Search and Rescue manager Mike Cannon told CNN.

"In many areas, the water conditions remain too dangerous for rescue by boat," state public safety commissioner Jennifer Morrison said Tuesday.



Want to stay updated on the latest space and science news?
We've got you.

Email address: Sign Me Up

By subscribing you agree to our [privacy policy](#).


"There are life-threatening situations that we are trying to identify and rescue," Morrison added.

President Joe Biden spoke with Scott and other officials about the flooding, the White House said Tuesday, adding it is coordinating with state and local officials and monitoring the flood impacts.

"The strong water currents pulled two boys under water while the other two safely made it to a concrete ledge," Oklahoma City Fire Department Capt. John Chenoweth said.

One resident saw water 'rising quickly'

As the rain lessens in Vermont, Scott cautioned against a false sense of complacency. "The water has to go somewhere, the reservoirs are filling up, and we have to determine the next phase, the next wave," he said.



Drove video taken over Londonderry, Vermont, on Monday shows the scale of the flooding.

Betsy Hart called 911 when the floodwater suddenly started rising fast at the basement of her Chester property in Windsor County.

"Water was rising quickly after being pretty tame most of the morning," Hart told CNN's Miguel Marquez. "All of a sudden, it was in the house."


Hart said she's never experienced flooding like what she saw Monday. "It was too close for comfort," she said.

"With Hurricane Irene, the water was raging like this, but it never really got to the house," she said, standing on a road near her home as water rushed nearby.

In the northern part of the state, a state of emergency was declared for the town of Long Lake due to severe flooding that destroyed roads and bridges, dented electric supply lines and forced some residents to evacuate. Long Lake town supervisor Clay Arsenault said in a declaration. A state of emergency is in effect for Hamilton County through Wednesday morning.

Long Lake officials also said non-essential travel in the town is not allowed at this time.

Water rushing into homes



A man carries belongings through floodwaters from a home in Bridgewater, Vermont, on Monday.

In Vermont, state Rep. Kelly Paglia said she woke up Monday morning to floodwater already at the front step of her Londonderry apartment. She and her son packed up their two cats and evacuated to higher ground.

"For people that were here during Irene, it feels like a very similar experience," she said. Hurricane Irene brought destructive flooding to the state in 2011, causing major infrastructure damage.

GPT-4V:

The article discusses the catastrophic flooding that occurred in Vermont's capital, Montpelier, on Tuesday, July 11, 2023. The flooding was caused by intense rainstorms that inundated the Northeast, leading to evacuations and closures in the area.

The floods were described as "unprecedented" and "historic" by officials, with some areas experiencing record-breaking rainfall. The National Weather Service issued a travel advisory, warning residents to avoid non-essential travel.

Vermont Governor Phil Scott declared a state of emergency, and New Hampshire sent rescue crews and Black Hawk helicopters to assist with the response. The Vermont Fire Department and other emergency services were also dispatched to the affected areas.

The floods caused extensive damage to homes and businesses, with some residents being forced to evacuate. A boil water notice was issued for Montpelier, and officials began clearing debris from roadways on Wednesday.

In Oklahoma City, two children were rescued from a reservoir after being swept away by floodwaters. The children, aged 10 and 11, were among a group of four that were fishing in the area.

In New York, Hurricane Irene caused 8 inches of rain within a 24-hour period, leading to major infrastructure damage. A state of emergency was declared for Long Lake Town Supervisor Clay Arsenault, and residents were urged to evacuate.

The article concludes with a quote from Dan Hancock, a resident of Chester, Vermont, who watched floodwater enter his basement. He expresses his frustration and helplessness, stating "Once the water goes down, we go back in, clean it up and move on. What can we do?"

Sec. 9.8 GUI Navigation

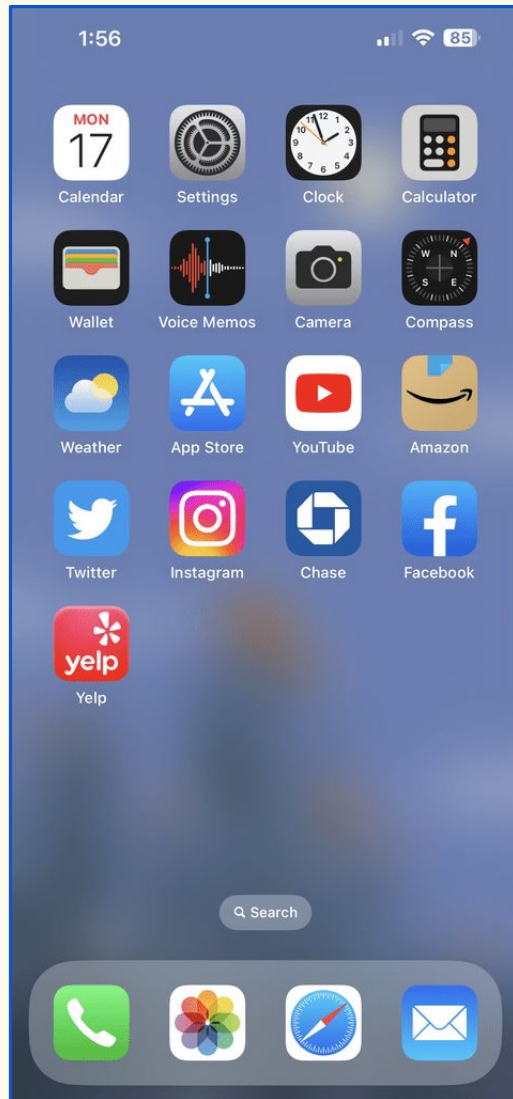
Online Shopping (Ergonomic Keyboard)

Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Below is what you see on the phone screen, predict your next move. If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My next move would be to move my finger to the Amazon icon, which is located in the third row and third column of the screen. I would then click on the icon to open the app and search for an ergonomic keyboard within my budget between \$50 and \$100.

图 103: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Red highlights the inaccurate location of the Amazon icon. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

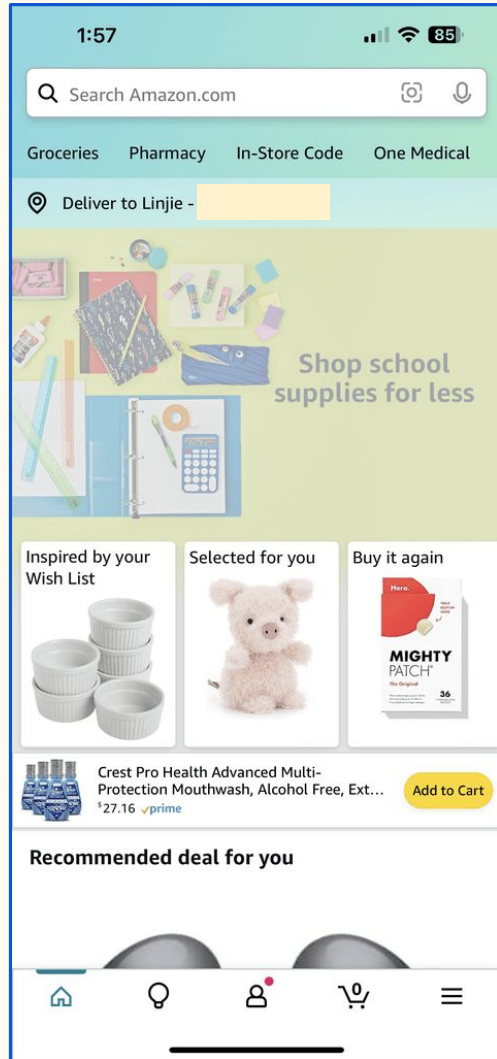
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: "My next move would be to move my finger to the Amazon icon, which is located in the third row and third column of the screen. I would then click on the icon to open the app and search for an ergonomic keyboard within my budget of \$50. "

You have now opened the Amazon app. Below is what you see on the phone screen. What will you do next? If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My next move would be to move my finger to the search bar, which is located at the top of the screen. I would then click on the search bar and type in "ergonomic keyboard" using the keyboard. After typing in the search term, I would click on the "Search" button, which is located on the bottom right corner of the keyboard, to initiate the search.

图 104: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

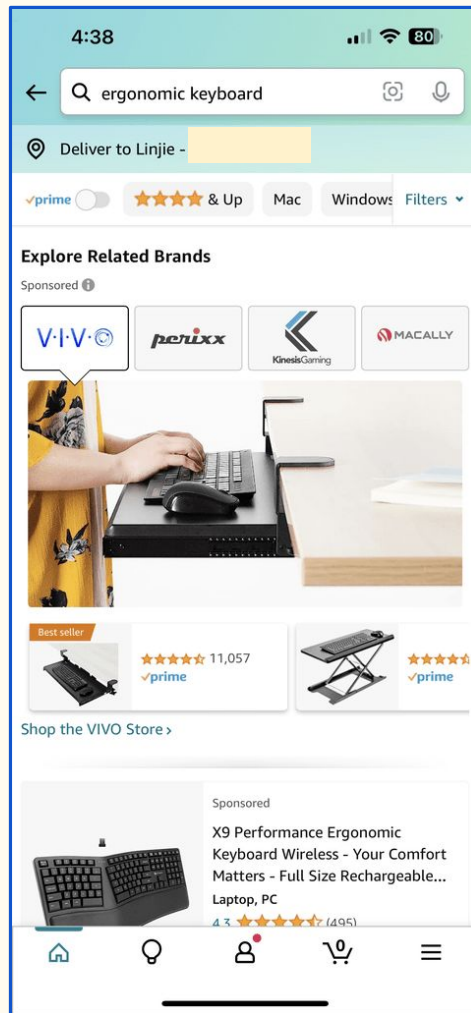
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: "My next move would be to move my finger to the search bar, which is located at the top of the screen. I would then click on the search bar and type in "ergonomic keyboard" using the keyboard. After typing in the search term, I would click on the "Search" button, which is located on the bottom right corner of the keyboard, to initiate the search. "

You have now completed the previous action as described above. Below is what you see on the phone screen. Please predict your immediate next move. If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My immediate next move would be to move my finger to the "Filters" button, which is located at the top right corner of the screen, next to the "Windows" button. I would then click on the "Filters" button to open the filter options and narrow down my search results to only show ergonomic keyboards within my budget range of \$50 to \$100.

图 105: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

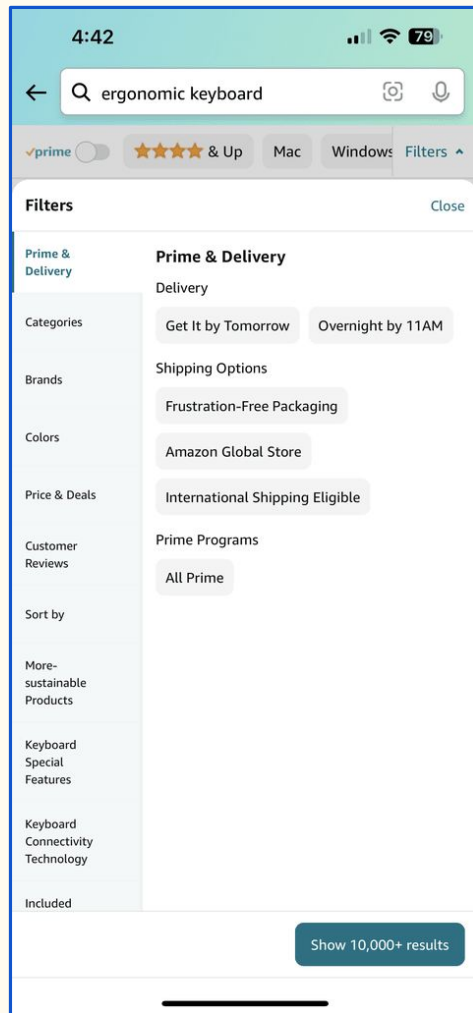
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: "My immediate next move would be to move my finger to the "Filters" button, which is located at the top right corner of the screen, next to the "Windows" button. I would then click on the "Filters" button to open the filter options and narrow down my search results to only show ergonomic keyboards within my budget range of \$50 to \$100. "

You have now completed the previous action as described above. Below is what you see on the phone screen. Please predict your immediate next move. If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My immediate next move would be to move my finger to the "Price & Deals" filter option, which is located under the "Colors" filter option and above the "Sort by" filter option. I would then click on the "Price & Deals" filter option to open the price range filter and set my minimum price to \$50 and my maximum price to \$100. This would further narrow down my search results to only show ergonomic keyboards within my budget range.

图 106: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

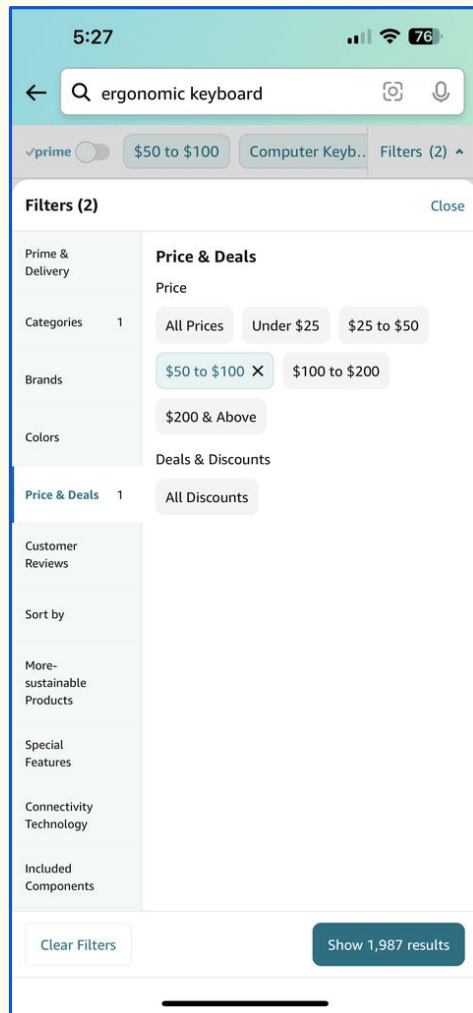
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: " My immediate next move would be to move my finger to the "Price & Deals" filter option, which is located under the "Colors" filter option and above the "Sort by" filter option. I would then click on the "Price & Deals" filter option to open the price range filter and set my minimum price to \$50 and my maximum price to \$100. This would further narrow down my search results to only show ergonomic keyboards within my budget range. "

You have now completed the previous action as described above. Below is what you see on the phone screen. What will you do next? If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My next move would be to move my finger to the "Show 1,987 results" button, which is located at the bottom of the screen, below the "Clear Filters" button and above the navigation bar. I would then click on the "Show 1,987 results" button to view the search results for ergonomic keyboards within my budget range.

图 107: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

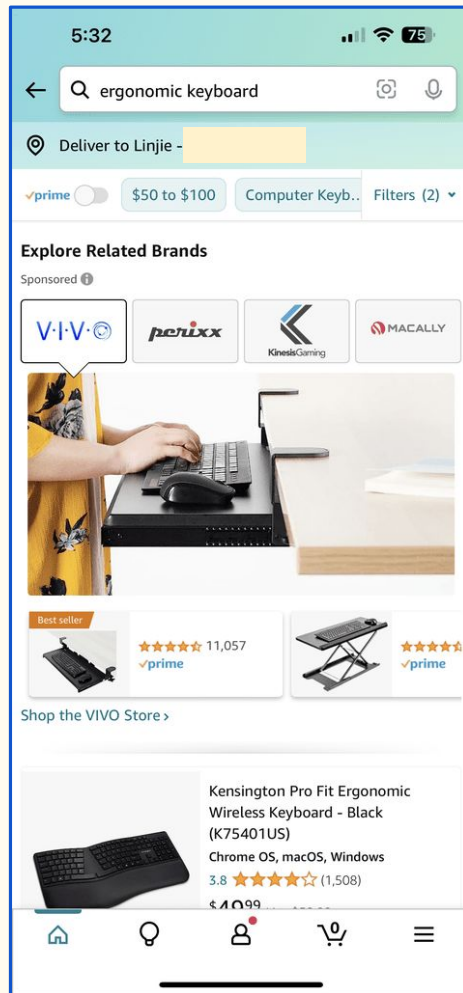
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: " My next move would be to move my finger to the "Show 1,987 results" button, which is located at the bottom of the screen, below the "Clear Filters" button and above the navigation bar. I would then click on the "Show 1,987 results" button to view the search results for ergonomic keyboards within my budget range. "

You have now completed the previous action as described above. Below is what you see on the phone screen. What will you do next? If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My next move would be to scroll down the screen to view the different options for ergonomic keyboards within my budget range. I would then move my finger to the "Kensington Pro Fit Ergonomic Wireless Keyboard - Black" option, which is located below the "Explore Related Brands" section and above the "VIVO" option. I would then click on the "Kensington Pro Fit Ergonomic Wireless Keyboard - Black" option to view more details about the product and potentially make a purchase.

图 108: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Red highlights the inaccurate location of the product option to be selected. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

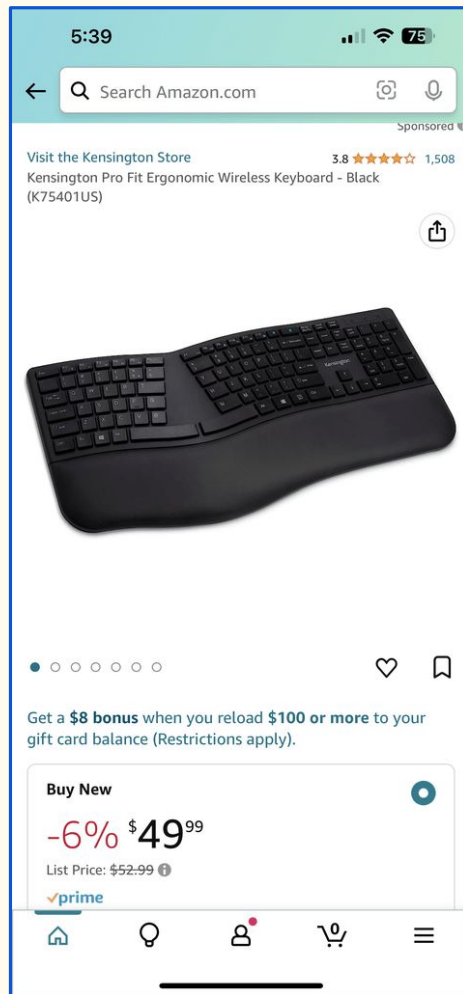
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: " My next move would be to scroll down the screen to view the different options for ergonomic keyboards within my budget range. I would then move my finger to the "Kensington Pro Fit Ergonomic Wireless Keyboard - Black" option, which is located below the "Explore Related Brands" section and above the "VIVO" option. I would then click on the "Kensington Pro Fit Ergonomic Wireless Keyboard - Black" option to view more details about the product and potentially make a purchase. "

You have now completed the previous action as described above. Below is what you see on the phone screen. What will you do next? If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My next move would be to scroll down the screen to view more details about the "Kensington Pro Fit Ergonomic Wireless Keyboard - Black" product. I would then move my finger to the "Buy New" button, which is located below the product image and above the "Get a \$8 bonus when you reload \$100 or more to your gift card balance" section. I would then click on the "Buy New" button to proceed with the purchase.

图 109: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Red highlights the inaccurate action prediction ("Buy New" is not a clickable button). Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

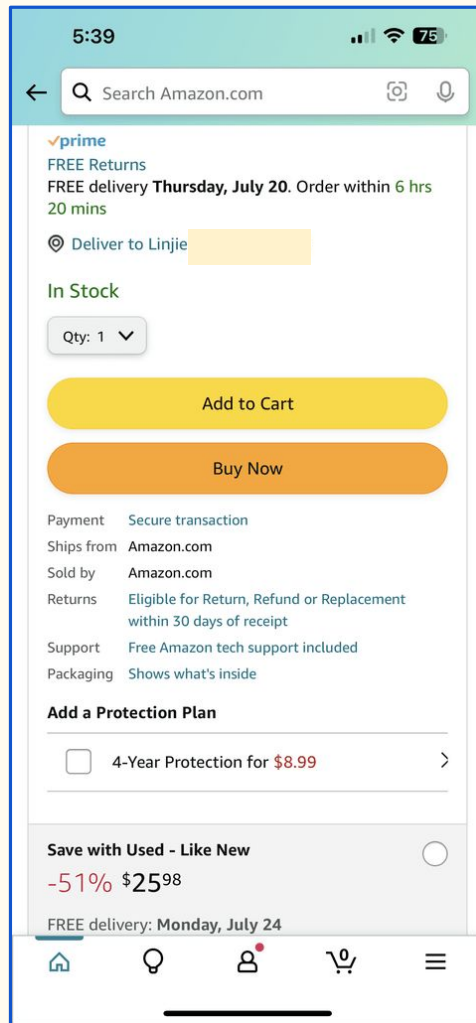
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: " My next move would be to scroll down the screen to view more details about the "Kensington Pro Fit Ergonomic Wireless Keyboard - Black" product. I would then move my finger to the "Buy New" button, which is located below the product image and above the "Get a \$8 bonus when you reload \$100 or more to your gift card balance" section. I would then click on the "Buy New" button to proceed with the purchase. "

You have now completed the previous action as described above. Below is what you see on the phone screen. What will you do next? If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My next move would be to move my finger to the "Add to Cart" button, which is located below the "Buy New" button and above the "Ships from Amazon" section. I would then click on the "Add to Cart" button to add the item to my shopping cart.

图 110: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Online Shopping (Ergonomic Keyboard)

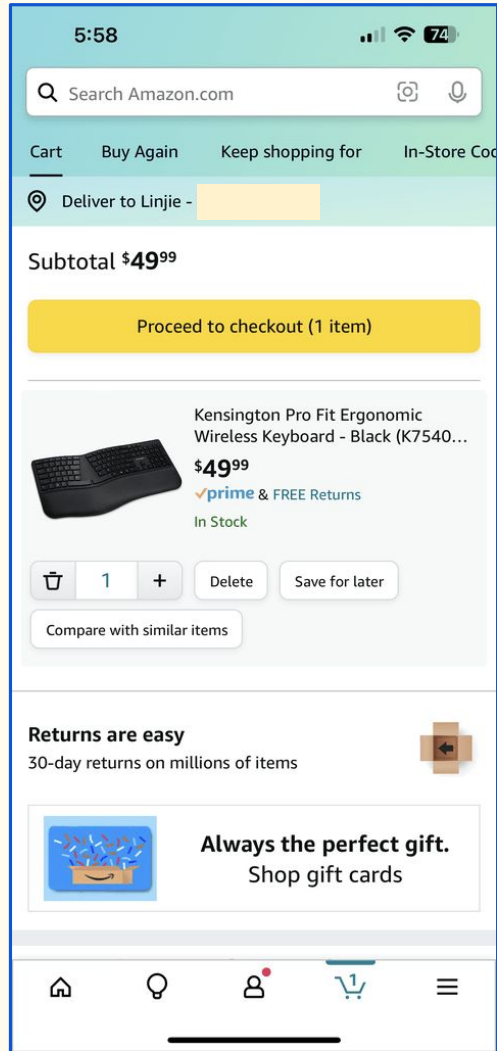
Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

You are asked to shop for an ergonomic keyboard, your budget is between \$50 and \$100.

Previous action: " My next move would be to move my finger to the "Add to Cart" button, which is located below the "Buy New" button and above the "Ships from Amazon" section. I would then click on the "Add to Cart" button to add the item to my shopping cart. "

You have now completed the previous action as described above. Below is what you see on the phone screen. What will you do next? If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

My next move would be to move my finger to the "Proceed to Checkout" button, which is located below the "Compare with similar items" section and above the "30-day returns" section. I would then click on the "Proceed to Checkout" button to proceed with the purchase.

图 111: GPT-4V navigates through GUI to shop for an ergonomic keyboard online. Blue highlights the predicted actions. Red highlights the inaccurate location of the "Proceed to checkout" button. Check Section 9.8 for detailed discussions.

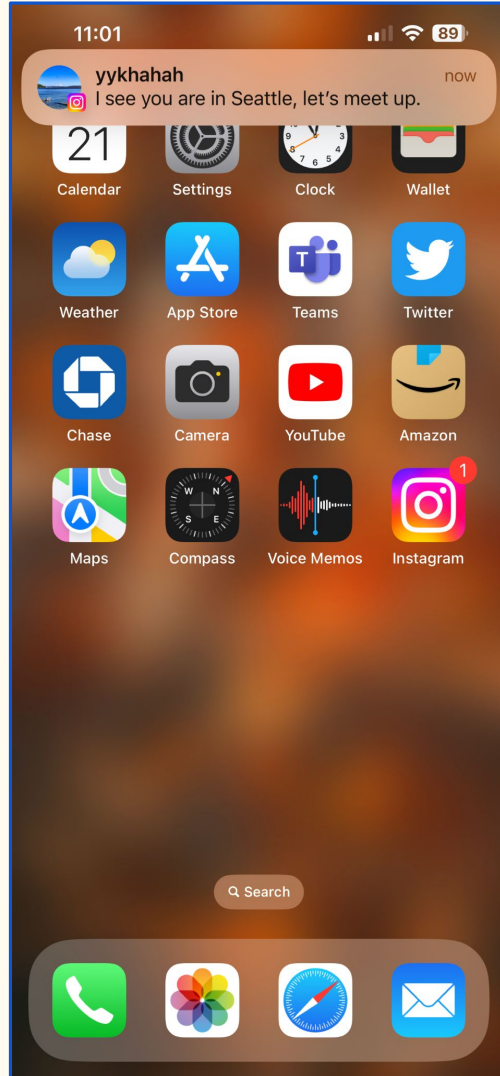
Sec. 9.8 GUI Navigation

Notification Understanding

Prompt:

Imagine that you are a robot operating a phone. Like how humans operate the phone, you can move your finger to an icon, click an icon with your finger, scroll down a screen, or type some texts with the keyboard.

Below is what you see on the phone screen. Describe the image and predict your next move. If the action involves moving your finger to an icon, describe the location of that icon as detailed as possible.



GPT-4V:

The image on the phone screen is the home screen of an iPhone. There are multiple app icons arranged in a grid-like pattern, including Calendar, Settings, Twitter, Amazon, Instagram, etc. At the top of the screen, there is a notification from a user named "yykhahah" that reads "I see you are in Seattle, let's meet up." At the bottom of the screen, there is a search bar.

Based on the notification, my next move would be to move my finger to the Maps app icon, which is located in the second row and the first column of the app grid. This will allow me to search for a location in Seattle and plan a meet up with the user.

图 112: Prompting GPT-4V to predict the action upon receiving a notification. GPT-4V can accurately recognize the notification and the corresponding content (highlighted in green). Blue highlights the predicted actions. Red highlights the inaccurate location of the Maps app icon. Check Section 9.8 for detailed discussions.

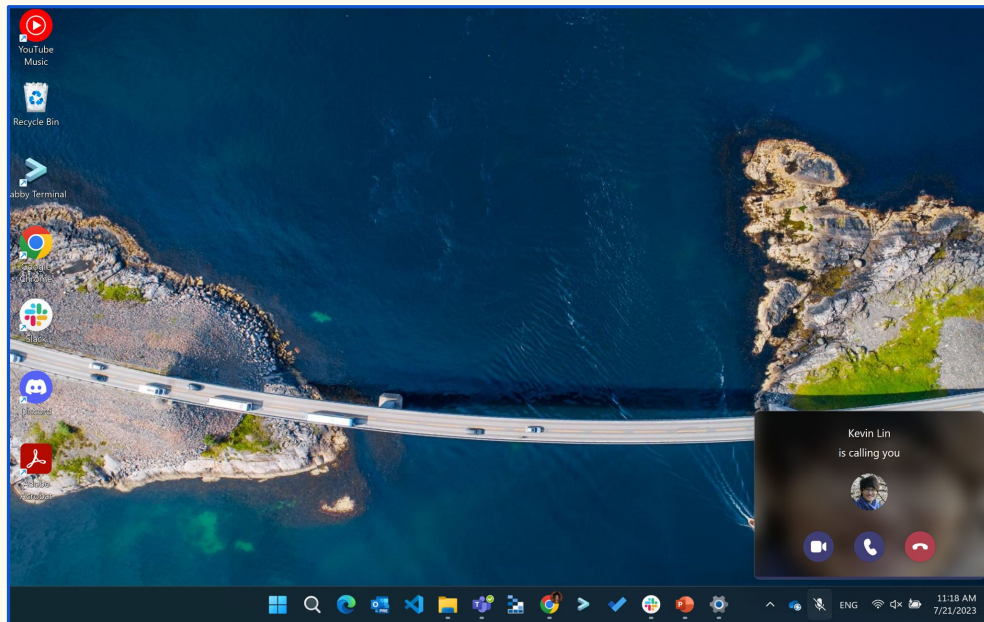
Sec. 9.8 GUI Navigation

Notification Understanding

Prompt:

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

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



GPT-4V:

The image on the computer screen is a desktop background of a bridge over a body of water. There are several icons on the desktop, including a YouTube icon, a Recycle Bin icon, and a Terminal icon. At the bottom of the screen, there is a taskbar with several application icons, such as Microsoft Edge and File Explorer. On the bottom right corner of the screen, there is a notification that says "Kevin Lin is calling you."

Based on the notification, my next move would be to move the mouse to the bottom right corner of the screen and click on the notification. This would open the call application and allow me to answer Kevin Lin's call.

图 113: Prompting GPT-4V to predict the action upon receiving a notification. GPT-4V can accurately recognize the notification and the corresponding content (highlighted in green). Blue highlights the predicted actions. Red highlights the inaccurate location of the Maps app icon. Check Section 9.8 for detailed discussions.

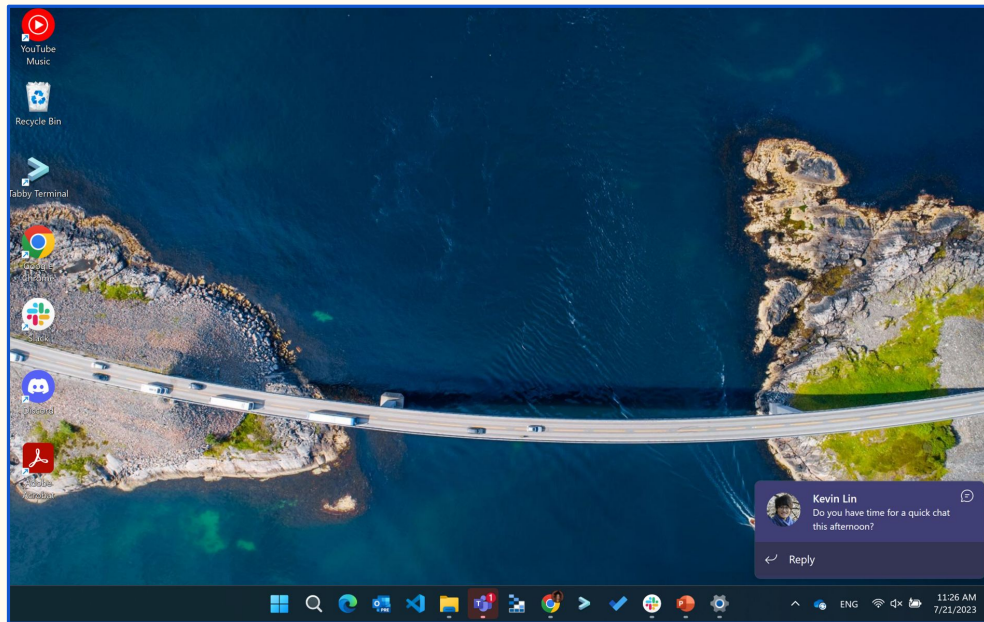
Sec. 9.8 GUI Navigation

Notification Understanding

Prompt:

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

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



GPT-4V:

The image on the computer screen is a desktop background of a beautiful aerial view of a bridge over the ocean. There are several icons on the left side of the screen, including the Recycle Bin, Terminal, and a few others. On the bottom right corner, there is a notification from a messaging app from a user named Kevin Lin asking if I am free for a chat this afternoon.

My next move would be to move the mouse to the bottom right corner of the screen and click on the notification from Kevin Lin to open the messaging app and respond to his message. I would move the mouse to the exact location of the notification, which is approximately 2 inches from the right edge of the screen and 1 inch from the bottom edge of the screen.

图 114: Prompting GPT-4V to predict the action upon receiving a notification. GPT-4V can accurately recognize the notification and the corresponding content (highlighted in green). Blue highlights the predicted actions. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Watching Videos

Prompt:

Explain the story shown in the images below.

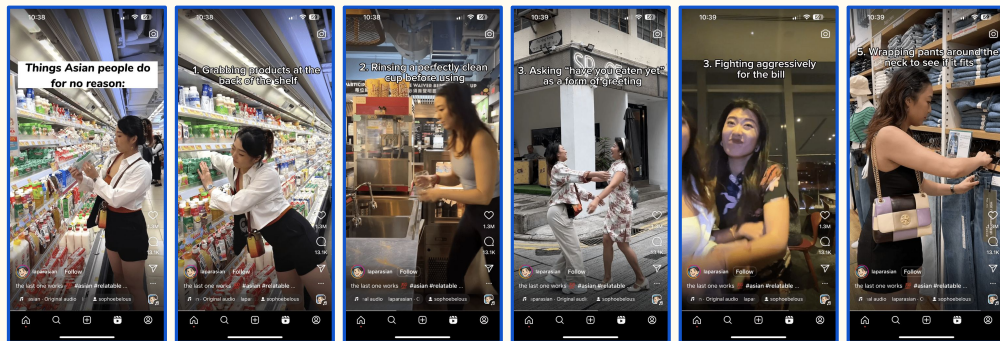


GPT-4V:

The images show a reel on Instagram, where a user is sharing a humorous post about how much they have spent on their dog. The first image shows a man holding his dog with the caption "When someone asks me how much I have spent on my dog". The next three images show a clip of Joe Biden saying "700 billion" followed by "700 billion and a trillion 300 million" and "700 billion and a trillion 300 million dollars". The user is jokingly exaggerating the amount they have spent on their dog by using the clip of Joe Biden saying a large sum of money.

Prompt:

Tell me about the video content, given the frames below.



GPT-4V:

The video content is a compilation of things that Asian people do for no reason. The video is posted on **TikTok** by user **laparasian** and has 1.3M views and 13.1K likes. The video is set to original audio by **sophrelabelous** and has the hashtag **#asian** and **#relatable**.

The video is broken down into 5 different scenes:

1. The first scene shows a girl grabbing products from the back of the shelf in a supermarket.
2. The second scene shows a girl rinsing a perfectly clean cup before using it.
3. The third scene shows two girls asking "have you eaten yet?" as a form of greeting.
4. The fourth scene shows two girls fighting aggressively for the bill in a restaurant.
5. The fifth scene shows a girl wrapping pants around her neck to see if it fits.

Overall, the video is a humorous take on some of the quirks and habits that are commonly seen in the Asian community.

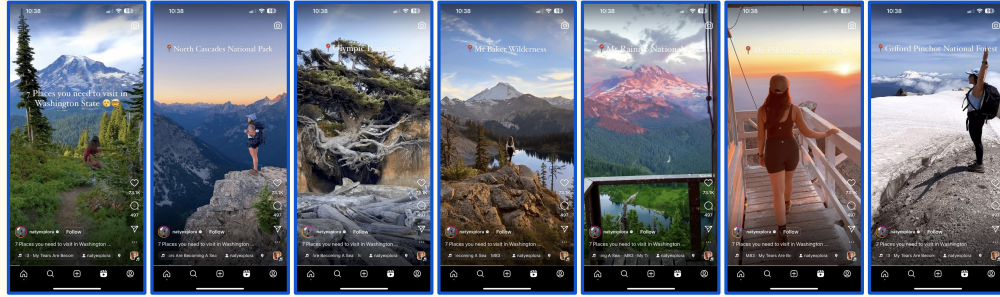
图 115: Prompting GPT-4V to watch web videos. We present GPT-4V the screenshot of the video frames following their temporal order in the original video. To save space, we illustrate the frames in a row, where the leftmost one is the first frame. Check Section 9.8 for detailed discussions. **Red** highlights the inaccurate descriptions about the video.

Sec. 9.8 GUI Navigation

Watching Videos

Prompt:

Transcribe the video content, given the frames below.



GPT-4V:

Hi everyone, it's natyexplora here and today I'm going to be sharing 7 places you need to visit in Washington State.

First up, we have North Cascades National Park. This park is filled with beautiful mountain views and hiking trails. It's a must-visit for any nature lover.

Next, we have the Olympic Peninsula. This area is home to the famous Hoh Rainforest and beautiful beaches. It's a great place to relax and enjoy the natural beauty of Washington.

Third on the list is Mt Baker Wilderness. This area is known for its stunning mountain views and hiking trails. It's a great place to escape the city and enjoy the great outdoors.

Next up, we have Mt Rainier National Park. This park is home to the highest mountain in Washington and is a great place for hiking and sightseeing.

Fifth on the list is Alpine Lakes Wilderness. This area is filled with beautiful alpine lakes and hiking trails. It's a great place to spend a day exploring and enjoying the natural beauty of Washington.

Next, we have Mt Pilchuck State Park. This park is known for its beautiful views and hiking trails. It's a great place to spend a day exploring and enjoying the natural beauty of Washington.

Last but not least, we have Gifford Pinchot National Forest. This forest is home to beautiful mountain views and hiking trails. It's a great place to escape the city and enjoy the great outdoors.

And that concludes my list of 7 places you need to visit in Washington State. I hope you enjoyed it and are inspired to explore more of this beautiful state.

Prompt:

Tell me about the video content, given the frames below.



GPT-4V:

The video content is about a corgi dog named Tofu who is working from home. The first frame shows Tofu sleeping on a bed with the caption "WFH Employees: 8:59 AM". The second frame shows Tofu sitting in front of a computer with multiple screens, with the caption "WFH Employees: 9:00 AM". The video is likely a humorous take on the work from home experience, highlighting how Tofu is just like any other employee who struggles to get out of bed in the morning but is ready to work once they're in front of their computer.

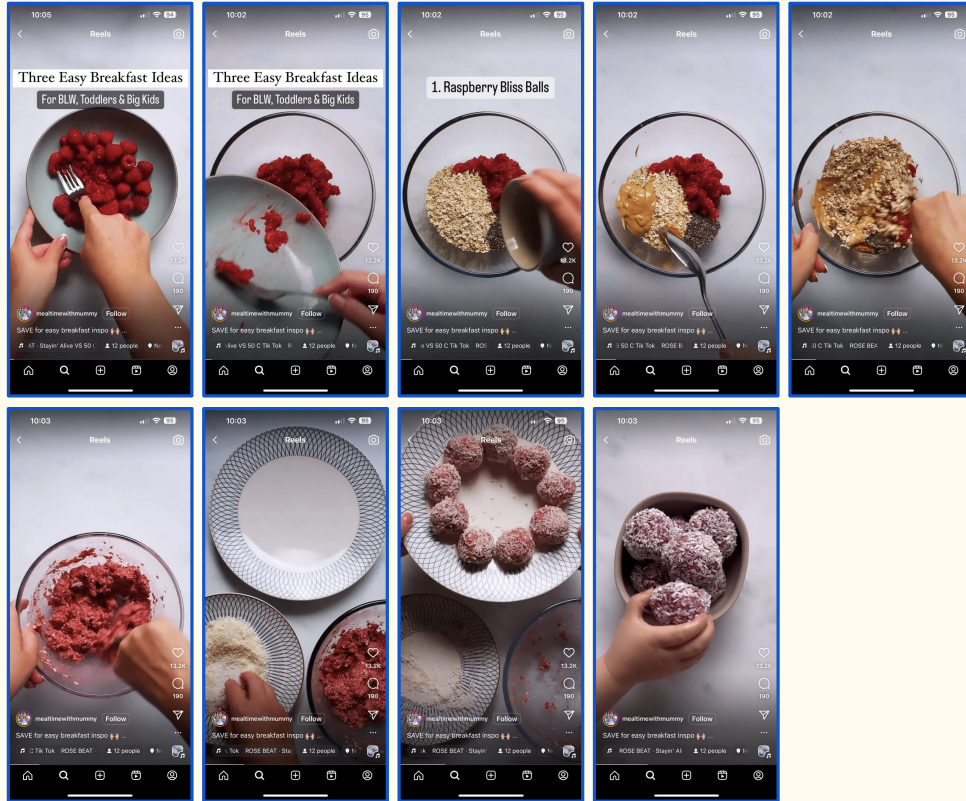
图 116: Watching web videos. We present GPT-4V the screenshot of the video frames following their temporal order in the original video. To save space, we illustrate the frames in a row, where the leftmost one is the first frame. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Watching Videos

Prompt:

Tell me about the video content, given the frames below. Please include as many details as possible.



GPT-4V:

The video content is a tutorial on how to make Raspberry Bliss Balls, a healthy breakfast option for toddlers and big kids. The video is posted by the Instagram account "mealtime with mummy" and has over 13k likes.

The video starts with a bowl of fresh raspberries and a fork, which is then mashed into a puree. The next frame shows a larger mixing bowl with oats, chia seeds, and peanut butter being added to the raspberry puree. The mixture is then stirred together until well combined.

The following frame shows a plate with shredded coconut on it, and the raspberry mixture being rolled into small balls and coated in the coconut. The final frames show the finished Raspberry Bliss Balls being placed in a bowl and ready to eat.

Throughout the video, there are text overlays with instructions and tips, making it easy for viewers to follow along and make the recipe themselves.

图 117: Watching web videos. We present GPT-4V the screenshot of the video frames following their temporal order in the original video. To save space, we illustrate frames 1-5 in the first row, and frames 6-9 in the second row. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Watching Videos

Prompt:

Tell me about the video content, given the frames below. Please include as many details as possible.



GPT-4V:

The video content is a recipe tutorial for apple and banana pancakes. The video is posted on Instagram Reels by the user "mealtime with mummy" and has 13.2K views.

In the first frame, we see a glass bowl with a mashed banana and an apple being added to it. The second frame shows the banana and apple being mixed together. The third frame shows an egg being added to the mixture. The fourth frame shows the mixture being stirred with a fork.

The fifth frame shows a bowl of oats being added to the mixture. The sixth frame shows the oats being mixed in with a spoon. The seventh frame shows the mixture being spooned into a frying pan. The eighth frame shows the pancakes being cooked in the frying pan.

The final frame shows the finished pancakes on a plate, served with sliced apples, bananas, and yogurt. The caption on the video reads "SAVE for easy breakfast inspo" and the video is set to the song "Stayin' Alive" by the Bee Gees.

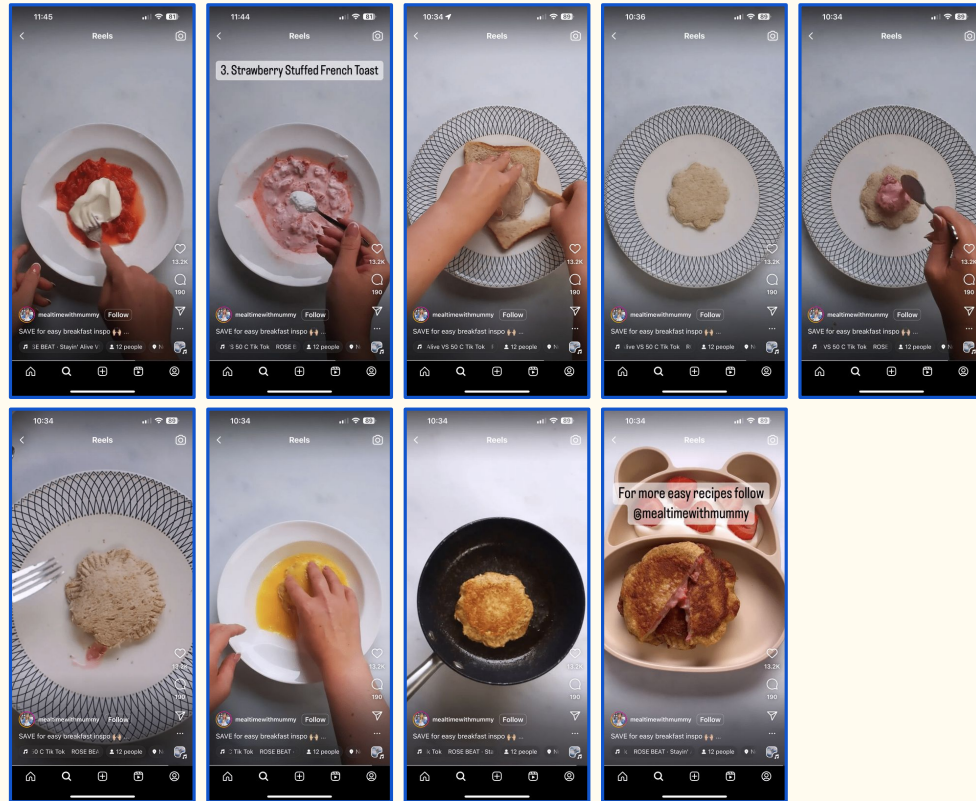
图 118: Watching web videos. We present GPT-4V the screenshot of the video frames following their temporal order in the original video. To save space, we illustrate frames 1-5 in the first row, and frames 6-9 in the second row. Check Section 9.8 for detailed discussions.

Sec. 9.8 GUI Navigation

Watching Videos

Prompt:

Tell me about the video content, given the 9 frames below. Please include as many details as possible.



GPT-4V:

The video content is a recipe tutorial for strawberry stuffed French toast. The video is posted on Instagram Reels by the user "mealtime with mummy". The video has 13.2K likes and 190 comments.

In the first frame, we see a white plate with a dollop of red sauce and a dollop of white cream. The second frame shows the plate with the red sauce and white cream mixed together. **The third frame shows a slice of bread with a knife spreading a mixture onto it.**

The fourth frame shows a plate with a circular piece of bread on it. The fifth frame shows a hand using a **knife** to spread a pink mixture onto the bread. The sixth frame shows the bread with the pink mixture covered with another piece of bread, making a sandwich.

The seventh frame shows a hand dipping the sandwich into a bowl of beaten eggs. The eighth frame shows the sandwich being fried in a pan. The final frame shows the finished strawberry stuffed French toast on a plate.

The video is set to the song "Stayin' Alive" by the Bee Gees. The caption reads "SAVE for easy breakfast inspo".

图 119: Watching web videos. We present GPT-4V the screenshot of the video frames following their temporal order in the original video. To save space, we illustrate frames 1-5 in the first row, and frames 6-9 in the second row. **Red** highlights the inaccurate descriptions about the video. Check Section 9.8 for detailed discussions.

10 LMM Powered Agents

在本节中，我们将讨论有可能进一步增强GPT-4V能力的潜在未来研究方向。讨论的重点是LLMs 中吸引人的用法如何扩展到多模态场景，并带来新的功能，例如多模态插件、多模态链、自我反思、自洽性以及检索增强的 LMMs 等等。在接下来的子节中，我们使用“人类生成”的示例来说明增强基于 GPT-4V的系统的潜在方法。

Sec. 10.1 Multimodal Plugins

Prompt:

Where is this photo taken?

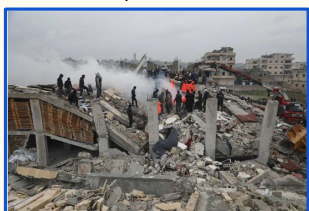


GPT-4V:

Sorry, I cannot answer this question accurately as no context or information is provided with the photo. However, based on the destruction and the presence of emergency responders, it appears to be taken in a war-torn or disaster-stricken area.

Prompt:

Where is this photo taken?



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

GPT-4V:

- I do not have that information.
- Enabling Bing Image Search Plugin.

Calling Bing Image Search...

- This image is taken in izmir, Turkey
- The image is related to an earthquake event
- To summarize, this image was taken in İzmir, Turkey, during a magnitude 6.6 earthquake in Turkey's Aegean region. The scene shows the aftermath of the disaster with people standing in a pile of rubble, a crane lifting a truck, and a red truck with a black front.

图 120: Illustration of using the Bing Image Search [89] plugin to enable GPT-4V with time-sensitive knowledge (bottom, highlighted in green). Note that the earthquake happened on February 6, 2023, which is after GPT-4V's training, thereby GPT-4V fails to identify the exact location without plugin (top). Check Section 10.1 for detailed discussions.

Sec. 10.2 Multimodal Chains

Prompt:

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



Thought 1

This question requires detecting people in the image.

Action 1

Assistant, detect the people shown in this image <ImagePath>

Observation 1

Person [4, 126, 72, 175]
 Person [63, 121, 88, 181]
 Person [30, 107, 56, 88]
 Person [223, 123, 43, 78]
 Person [178, 122, 52, 180]
 Person [322, 118, 41, 68]
 Person [268, 109, 55, 178]
 Person [357, 127, 88, 173]

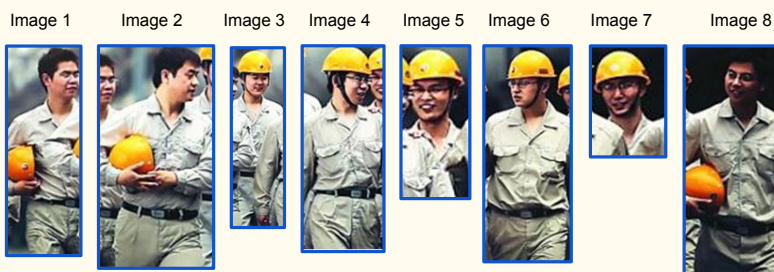
Thought 2

There are 8 people detected in this image.

Action 2

Assistant, crop the regions in this image specified by the bounding boxes <ImagePath>

Observation 2



GPT-4V:

Image 1: Not wearing a helmet
 Image 2: Not wearing a helmet
 Image 3: Wearing a helmet
 Image 4: Wearing a helmet
 Image 5: Wearing a helmet
 Image 6: Wearing a helmet
 Image 7: Wearing a helmet
 Image 8: Not wearing a helmet

Summary: Out of 8 people, 5 are wearing helmets and 3 are not wearing helmets.

图 121: Extending GPT-4V to multimodal chains with ReAct [138, 135] for PPE Counting scenario. Check Section 10.2 for detailed discussions.

10.1 Multimodal Plugins

在LLMs的背景下，插件 [93, 53, 5, 105, 82, 97] 在辅助LLMs进行各种任务时起着重要作用，例如访问最新信息、进行计算或利用第三方服务。这些插件主要设计用于处理自然语言输入或可解释为语言的输入，例如代码和数学方程。为了说明多模态插件的重要性，尤其是在LLMs的背景下，我们在图 120 中展示了必应图像搜索等多模态插件的意义。通过整合必应图像搜索插件，我们赋予 GPT-4V 获取与输入图像相关的及时知识的能力。在图的上半部分，我们展示了未使用必应图像搜索插件时 GPT-4V 的限制。它由于照片捕捉到2023年2月6日在土耳其和叙利亚边界发生的一次大地震之后的情况，无法准确回答“这张照片是在哪里拍摄的？”这是由于 GPT-4V 训练之后发生的情况。由于用当前信息不断重新训练模型可能具有很大的计算成本和开销，搜索引擎等插件被证明是模型获取最新信息的宝贵资源。在图 120 的下半部分，我们展示了配备必应图像搜索插件的 GPT-4V 的能力。它有效利用从插件中检索到的信息，实现了对土耳其伊兹密尔位置的准确识别。

10.2 Multimodal Chains

最近的研究广泛探索了使用语言和视觉模型（LLM）进行链式处理的方法（Yao等人，2022；Gao等人，2022；Trivedi等人，2022；Qin等人，2023）。这种方法不仅使用单个插件，而是建立了一种将LLMs与一组插件集成起来的系统范式，实现更先进的推理和交互。通过用图像字幕生成器、目标检测器或针对图像到文本生成和音频到文本转换的训练良好的模型等视觉/多模态专家替换仅使用语言的插件，可以构建具有强大多模态能力的LLMs链（Wu等人，2023；Yang等人，2023；Suris等人，2023；Shen等人，2023；Liang等人，2023；Lu等人，2023）。

然而，这些链中LLMs和插件之间的交互通常以文本格式进行。尽管插件可以接受多模态输入，但它们返回的结果是文本，以增强LLMs的知识。在图像合成/编辑方面存在一个明显的例外情况（Wu等人，2023），插件可以生成图像，但这些图像不会反馈到LLMs进行进一步分析或增强知识，因为LLMs只能处理基于语言的输入。

在图 121中，我们展示了如何将GPT-4V 扩展为支持与ReAct（Yao等人，2022；Yang等人，2023）结合使用的多模态链。该扩展使得链中的插件可以提供多模态信息，然后由GPT-4V 集体处理以实现诸如PPE计数等场景中的高级推理。图 121中展示的整体链式处理过程分为两轮思考、行动和观察。每一轮都涉及激活特定插件。在第一轮中，GPT-4V 推断出需要进行人员检测以计算戴安全帽的人数（思考1）。因此，它调用人员检测工具（行动1），并接收图像中每个检测到的人的边界框坐标（观察1）。进入第二轮，基于获取的边界框信息，GPT-4V 推断出图像中共有8个人（思考2）。然后，它使用图像裁剪工具根据各自的边界框坐标裁剪出每个人的单独图像（行动2）。最终的输出（观察2）包括8个带编号的图像，从图像1到图像8。接下来，GPT-4V 确定这些图像中的每个人是否戴着安全帽，并总结戴着安全帽的人的总数。

总的来说，将LLMs与多模态插件集成在一起，为增强推理和交互提供了新的可能性，充分发挥了语言和视觉能力的优势。多模态链的灵活性可以更全面地理解和分析多模态数据，并有可能在各种应用中提高性能。

10.3 Self-Reflection

图 122展示了自反思方法 [109, 83, 60]在改进图 47所示结果方面的应用。我们可以看到，自反思得到的结果与参考图像更加吻合。例如，在左侧，数据点的数量从4个被修正为3个，而在右侧，百分比被重新加回了柱状图上方。虽然结果仍然不完全相同，但明显能够看出，自

反思可以促进手动打磨。图 123展示了另一个例子，展示了自反思在改进文本到图像模型生成提示时的应用 [99]。

10.4 Self-Consistency

自一致性是一种解码策略，它聚合多个采样输出以产生最终答案，比如通过多数投票。Tree-of-Thoughts则是从边缘化到聚合最终答案的思想得到了延伸，它表明自一致性思想可以应用于中间思考过程，以提高LLM的推理性能。图 124展示了在计数问题上应用自一致性到GPT-4V 的使用。我们通过让GPT-4V 多次对同一图像进行计数来采样多个计数结果，可以通过进行多次运行（样本2-4）或重新表述输入文本指令（样本1,2）来实现。该示例使用简单的多数投票来聚合最终答案，得到了"4艘船"。关于自一致性LMMs的全面探索将留待未来的研究。

10.5 Retrieval-Augmented LMMs

检索增强的语言模型（Retrieval-Augmented LMMs）通过检索和整合相关信息到提示中，提升了文本生成的能力。当需要特定任务相关信息时，如高度专业化的专家领域的专业知识、可能与语言模型的记忆不同的最新信息以及因用户而异的可定制信息时，这种技术尤其有效。我们认为检索增强在语言模型中继续发挥着重要作用。图 74展示了检索增强的语言模型在帮助杂货结账时的一个示例。由于不同店铺的产品图像-文字-价格三元组不同，从店铺的数据库中检索并给出正确的结账信息会很有益。同样，在图 81的定制字幕场景中，我们想象系统可以自动从相册中检索家庭成员的照片，并实现定制字幕。

11 Conclusions

11.1 Summary and Conclusions

在本报告中，我们的主要重点是探索在各种应用场景下的 GPT-4V。研究结果揭示了其显著的能力，其中一些能力在现有方法中尚未得到研究或展示。虽然我们努力尽可能揭示这些能力，但我们承认我们的呈现可能并不详尽。然而，这份报告可以作为未来研究的参考，目标是探索 GPT-4V的其他应用，深化对线性混合模型的理解，并构建更强大的线性混合模型。

11.2 Towards Future LMMs

GPT模型的弱点和限制已在相关报告中广泛讨论 [94, 95, 23]。在本节中，我们简要介绍我们对未来研究方向的想法。

像GPT-1、GPT-2和GPT-3这样的模型主要作为文本输入-文本输出系统，只能处理自然语言。没有视觉的GPT-4展示了在文本理解和生成方面无与伦比的能力，而GPT-4V 则展现了强大的图像领域理解能力。

作为自然的发展，LMMs应该能够生成交织的图像-文本内容，比如生成包含文本和图像的生动教程，以实现全面的多模态内容理解和生成。此外，将其他模态，如视频、音频和其他传感器数据纳入LMMs中，将有助于扩展其能力。

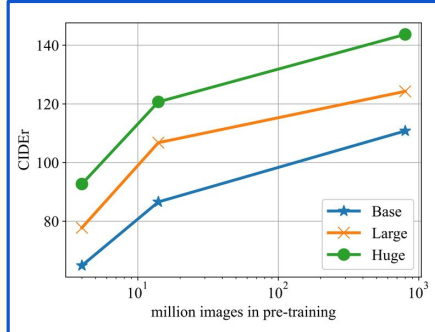
关于学习过程，当前的方法主要依赖于组织良好的数据，如图像标签或图像-文本数据集。然而，更通用的模型也许能够从各种来源学习，包括在线网络内容甚至真实世界的物理环境，以促进持续的自我演化。

Sec. 10.3 Self-Reflection

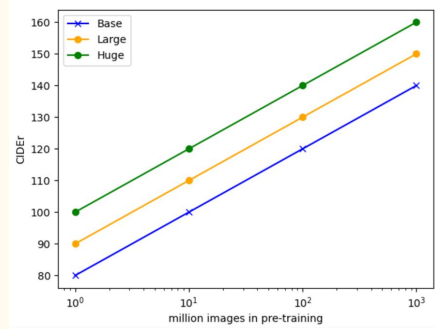
Self-Reflection to Improve Coding

Prompt:

Generate python code to draw similar curves.



GPT-4V:

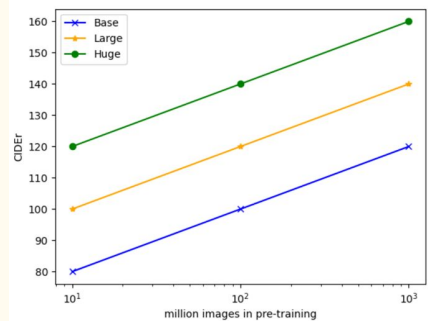


Prompt:

Can you double-check the code and align it better with the reference image?

Note: Code from the above GPT-4V response is omitted to save space

GPT-4V:

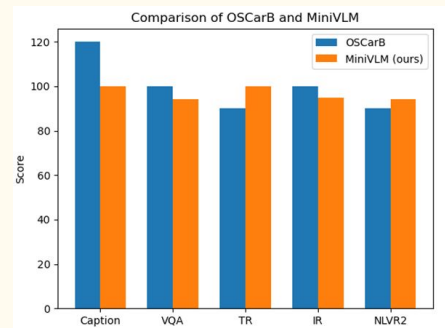


Prompt:

Write python codes to generate similar figures.



GPT-4V:



Prompt:

Can you double-check the code and align it better with the reference image?

Note: Code from the above GPT-4V response is omitted to save space

GPT-4V:

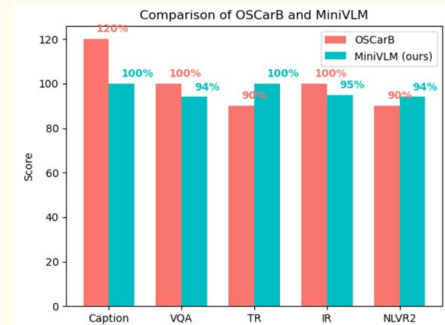


图 122: Illustration of using self-reflection to improve the code for figure drawing. Left: after reflection, the number of points in the curve aligns with the reference image. Right: the percentage is added to align with the reference image. Check Section 10.3 for detailed discussions.

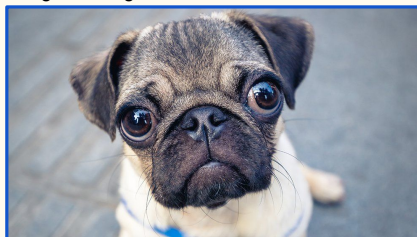
Sec. 10.3 Self-Reflection

Self-Reflection to Improve Text-to-Image Generation

Prompt:

Instruction: Given a user imagined IDEA of the scene, converting the IDEA into a self-contained sentence prompt that will be used to generate an image.

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

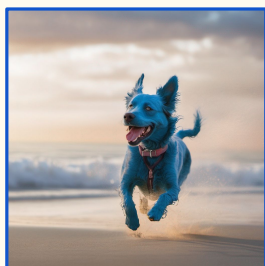


Based on the above information, I wrote a detailed prompt exactly about the IDEA follow the rules.

GPT-4V:

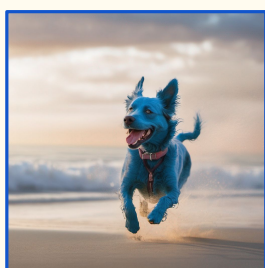
A blue dog running on the beach with a happy expression on its face.

SDXL:



Prompt:

I am iteratively improving the sentence prompt by looking at the images generated by an AI art generation model and find out what is different from the given IDEA. The IDEA is: photo of a dog ... beach. Generated sentence prompt for current round is: A blue dog running on the beach with a happy expression on its face. Corresponding image generated by the AI art generation model:



Prompt Continued

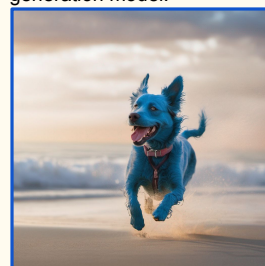
The generated image is not perfect. Write key reasons on how to further improve the image to better follow the user imagined IDEA of the scene.

GPT-4V:

The dog in the generated image does not look like the one in the given image. The dog in the given image is a pug, while the dog in the generated image is a different breed. To address this issue, we can modify the prompt to specifically mention that the dog should be a pug.

Prompt:

Instruction: Given a user imagined IDEA of the scene, converting the IDEA into a sentence prompt that will be used to generate an image. I am iteratively improving the sentence prompt by looking at the images generated by an AI art generation model and find out what is different from the given IDEA. The IDEA is: photo of a dog ... beach. Generated sentence prompt for current round is: A blue ... its face. Corresponding image generated by the AI art generation model:



The image can be improved because: The dog ... should be a pug. Based on the above information, to improve the image, I wrote a detailed prompt exactly about the IDEA follow the rules.

GPT-4V:

A blue pug running on the beach with a happy expression on its face, with a sunset and palm trees in the background.

SDXL:

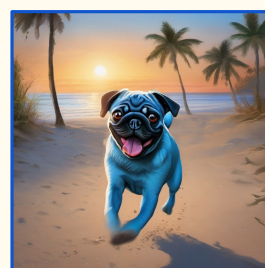


图 123: Illustration of using self-reflection to improve the generated text prompts for a text-to-image model SDXL [99]. GPT-4V reflects the error in the initial prompt that it does not mention the dog's breed, and makes the correct revision. Check Section 10.3 for detailed discussions.

Sec. 10.4 Self-Consistency

Self-Consistency for Counting

Sample 1

Prompt:

Count the number of boats in the image.



GPT-4V:

There are 6 boats in the image.

Sample 2

Prompt:

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



Note: Same query image as shown in Sample 1, resized to save space

GPT-4V:

There are a total of 4 boats in the image. 1) The first boat is on the left side of the image, docked at the pier. It is a larger boat with a tower on top. 2) The second boat is in the center of the image, also docked at the pier. It is a smaller boat with a white canopy on top. 3) The third boat is on the right side of the image, docked at the pier. It is a smaller boat with a white canopy on top, similar to the second boat. 4) The fourth boat is in the background, on the far right side of the image. It is a sailboat, and it appears to be sailing on the water.

Sample 3

Prompt:

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



Note: Same query image as shown in Sample 1, resized to save space

GPT-4V:

There are 4 boats in the image. 1) The first boat is on the left side of the image, docked at the pier. It is a larger boat with a tower on top. 2) The second boat is in the center of the image, also docked at the pier. It is a smaller boat with a white canopy on top. 3) The third boat is on the right side of the image, docked at the pier. It is a smaller boat with a white canopy on top, similar to the second boat. 4) The fourth boat is in the background, on the far right side of the image. It is partially obscured by the third boat, but appears to be a larger boat with a tower on top.

Sample 4

Prompt:

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



Note: Same query image as shown in Sample 1, resized to save space

GPT-4V:

There are 6 boats in the image. - 1 boat is on the left side of the image, docked at the pier with a tower on top - 2 boats are in the center of the image, docked at the pier side by side - 1 boat is on the right side of the image, docked at the pier with a cover on top - 2 boats are in the background on the right side of the image, docked at a different pier.

Self Consistency:

Sample 2: There are a total of 4 boats in the image. 1) The first boat is on the left side of the image

Sample 3: There are 4 boats in the image. 1) The first boat is on the left side of the image, docked

Sample 4: There are 6 boats in the image. - 1 boat is on the left side of the image, docked at the pier

[Marginalize out reasoning paths to aggregate final answers, e.g., majority vote] =>

There are 4 boats in the image.

图 124: Improve the counting reliability with self-consistency [123], which aggregates multiple counting results repeated on the *same* image. Check Section 10.4 for detailed discussions.

Acknowledgment

我们由衷感谢OpenAI提供我们早期使用他们出色工具的机会。我们要特别感谢Misha Bilenko提供的宝贵指导和支持。我们还要真诚感谢微软的同事们对我们的见解和观点，特别感谢John Montgomery、Marco Casalaina、Gregory Buehrer、Nguyen Bach、Gopi Kumar、

Luis Vargas、Kun Wu、Meenaz Merchant、Jianfeng Gao、Matt Lungren、Sheela Agarwal、Yumao Lu、Thomas Soemo、Fisayo Okikiolu、Ce Liu、Michael Zeng、Faisal Ahmed、Ehsan Azarnasab和Lin Liang提供的建设性反馈。我们还要感谢Yingkai Yu在GUI导航方面提供的截图帮助。

参考文献

- [1] Deepfloyd if. <https://github.com/deep-floyd/IF>, 2023.
- [2] Guidance. <https://github.com/microsoft/guidance/>, 2023.
- [3] Midjourney. <https://www.midjourney.com/>, 2023.
- [4] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011.
- [5] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.
- [6] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [7] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. In *EMNLP*, 2019.
- [8] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.
- [9] Karpathy Andrej. State of gpt. <https://karpathy.ai/stateofgpt.pdf>, 2023.
- [10] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.
- [11] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015.
- [12] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo, March 2023.
- [13] Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. Are elephants bigger than butterflies? reasoning about sizes of objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.
- [14] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023.

- [15] David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *International conference on machine learning*, pages 511–520. PMLR, 2018.
- [16] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluís Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, 2019.
- [17] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2023.
- [18] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.
- [19] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101—mining discriminative components with random forests. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014.
- [20] Marc A Brackett and Peter Salovey. Measuring emotional intelligence with the mayer-salovey-caruso emotional intelligence test (msceit). *Psicothema*, 18:34–41, 2006.
- [21] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023.
- [22] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.
- [23] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.
- [24] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.
- [25] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. In *ECCV*, 2020.
- [26] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021.
- [27] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [28] Herbert H Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986.

- [29] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.
- [30] Tom Cornsweet. *Visual perception*. Academic press, 2012.
- [31] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*, 2022.
- [32] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.
- [33] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.
- [34] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.
- [36] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.
- [37] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *Advances in Neural Information Processing Systems*.
- [38] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022.
- [39] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023.
- [40] Alicia Fasquel, Angèle Brunellière, and Dominique Knutsen. A modified procedure for naming 332 pictures and collecting norms: Using tangram pictures in psycholinguistic studies. *Behavior Research Methods*, pages 1–23, 2022.
- [41] Samir Yitzhak Gadre, Kiana Ehsani, and Shuran Song. Act the part: Learning interaction strategies for articulated object part discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15752–15761, 2021.

- [42] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020.
- [43] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022.
- [44] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR, 2023.
- [45] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans, 2023.
- [46] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016.
- [47] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.
- [48] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.
- [49] Jack *Hessel, Jena D *Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. The Abduction of Sherlock Holmes: A Dataset for Visual Abductive Reasoning. In *ECCV*, 2022.
- [50] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [51] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023.
- [52] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.
- [53] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022.
- [54] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

- [55] Fabian Huttmacher. Why is there so much more research on vision than on any other sensory modality? *Frontiers in psychology*, 10:2246, 2019.
- [56] Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert Hawkins, and Yoav Artzi. Abstract visual reasoning with tangram shapes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 582–601, 2022.
- [57] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
- [58] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.
- [59] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Denscap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016.
- [60] Geunwoo Kim, Pierre Baldi, and Stephen McAleer. Language models can solve computer tasks. *arXiv preprint arXiv:2303.17491*, 2023.
- [61] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.
- [62] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023.
- [63] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.
- [64] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [65] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 2023.
- [66] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020.
- [67] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

- [68] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.
- [69] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- [70] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.
- [71] Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, et al. Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis. *arXiv preprint arXiv:2303.16434*, 2023.
- [72] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
- [73] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [74] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.
- [75] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- [76] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- [77] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.
- [78] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020.
- [79] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018.
- [80] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

- [81] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*, 2023.
- [82] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022.
- [83] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.
- [84] Bertram F Malle, Louis J Moses, and Dare A Baldwin. *Intentions and intentionality: Foundations of social cognition*. MIT press, 2001.
- [85] Arjun Mani, Nobline Yoo, Will Hinthorn, and Olga Russakovsky. Point and ask: Incorporating pointing into visual question answering. *arXiv preprint arXiv:2011.13681*, 2020.
- [86] John D Mayer. Msceit: Mayer-salovey-caruso emotional intelligence test. *Toronto, Canada: Multi-Health Systems*, 2002.
- [87] John D Mayer, Richard D Roberts, and Sigal G Barsade. Human abilities: Emotional intelligence. *Annu. Rev. Psychol.*, 59:507–536, 2008.
- [88] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023.
- [89] Microsoft. Bing image search api. <https://www.microsoft.com/en-us/bing/apis/bing-image-search-api>, 2023.
- [90] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [91] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, 2022.
- [92] Tara Murfitt and Jan McAllister. The effect of production variables in monolog and dialog on comprehension by novel listeners. *Language and Speech*, 44(3):325–350, 2001.
- [93] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.
- [94] OpenAI. Gpt-4 technical report, 2023.
- [95] OpenAI. Gpt-4v(ision) system card. 2023.

- [96] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- [97] Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. Art: Automatic multi-step reasoning and tool-use for large language models. *arXiv preprint arXiv:2303.09014*, 2023.
- [98] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.
- [99] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [100] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. Tool learning with foundation models, 2023.
- [101] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- [102] John C Raven and JH Court. *Raven’s progressive matrices*. Western Psychological Services Los Angeles, 1938.
- [103] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
- [104] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2021.
- [105] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.
- [106] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *CVPR*, pages 14974–14983, 2023.

- [107] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023.
- [108] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.
- [109] Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*, 2023.
- [110] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. *arXiv preprint arXiv:2304.06712*, 2023.
- [111] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758, 2020.
- [112] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.
- [113] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vi-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019.
- [114] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.
- [115] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019.
- [116] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [117] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*, 2022.
- [118] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.
- [119] Carven Von Bearnensquash. Paper gestalt. *Secret Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2010.
- [120] Hong Wang, Xuan Luo, Weizhi Wang, and Xifeng Yan. Bot or human? detecting chatgpt imposters with a single question. *arXiv preprint arXiv:2305.06424*, 2023.
- [121] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*, 2022.

- [122] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.
- [123] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2022.
- [124] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. In *Advances in Neural Information Processing Systems*.
- [125] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022.
- [126] David Wechsler. Wais-r: Manual: Wechsler adult intelligence scale-revised. (*No Title*), 1981.
- [127] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *ICLR*, 2022.
- [128] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.
- [129] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- [130] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.
- [131] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022.
- [132] Zhenjia Xu, Zhijian Liu, Chen Sun, Kevin Murphy, William T Freeman, Joshua B Tenenbaum, and Jiajun Wu. Unsupervised discovery of parts, structure, and dynamics. In *International Conference on Learning Representations*, 2018.
- [133] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022.
- [134] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022.

- [135] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.
- [136] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255, 2023.
- [137] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.
- [138] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2022.
- [139] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- [140] Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. Broaden the vision: Geo-diverse visual commonsense reasoning. *arXiv preprint arXiv:2109.06860*, 2021.
- [141] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019.
- [142] Andy Zeng, Maria Attarian, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aweek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, et al. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*, 2022.
- [143] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327, 2019.
- [144] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.
- [145] Yan-Tao Zheng, Ming Zhao, Yang Song, Hartwig Adam, Ulrich Buddemeier, Alessandro Bissacco, Fernando Brucher, Tat-Seng Chua, and Hartmut Neven. Tour the world: building a web-scale landmark recognition engine. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1085–1092. IEEE, 2009.
- [146] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.
- [147] Bolei Zhou, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *Journal of Vision*, 17(10):296–296, 2017.

- [148] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, 2020.
- [149] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, 2022.
- [150] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
- [151] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023.
- [152] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized decoding for pixel, image and language. 2022.
- [153] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023.